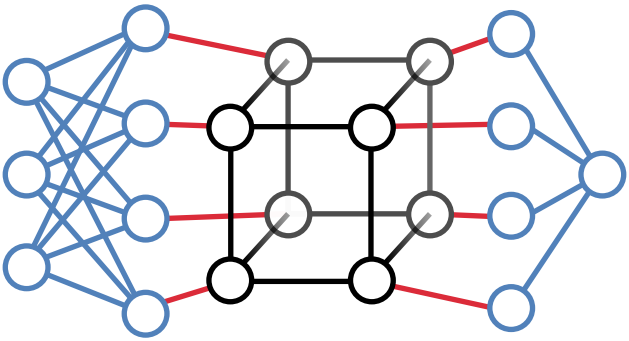
YOUNG RESEARCHER'S WORKSHOP
ON MACHINE LEARNING FOR MATERIALS

ML 4 M

2022



ABSTRACT BOOK

9-13 May 2022
SISSA Miramare Campus
Trieste, Italy

Contents

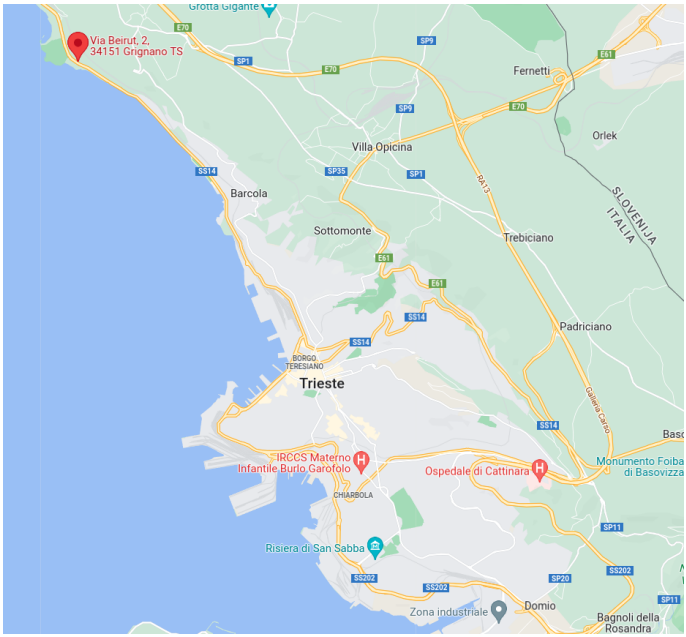
General

Organizers

Claudio Zeni, Scuola Internazionale Superiore di Studi Avanzati, Italy
Kevin Rossi, École polytechnique fédérale de Lausanne, Switzerland
Milica Todorović, Turku University, Finland
Patrick Rinke, Aalto University, Finland
Stefano de Gironcoli, Scuola Internazionale Superiore di Studi Avanzati, Italy

Venues

Main venue: SISSA Conference area Miramare Campus, Via Beirut 2, Grignano (IT). See also Map below.
The workshop location can be reached using the bus 6 passing through Trieste centre. The nearest stop to the conference's venue is SS 14 Centro Fisica, castello Miramare. Signs will be there to indicate the path from the bus stop to the conference venue.



Poster session: SISSA Main Building, Via Bonomea 256, Opicina (IT)
The poster session will be held on Wednesday, 11 May, 18.00-20.00h.
To reach SISSA Main Building, private buses have been booked and will leave SISSA Miramare Campus at 17.30. During the poster session, a warm dinner and drinks will be served. The posters will be displayed outdoor if weather allows. The "Theater room" is otherwise going to be used. Poster sizes are A0 portrait or A1 landscape.

Meals

Lunch will be provided free of charge on Tuesday, Wednesday, and Thursday.

Coffee breaks will also be provided free of charge.

A warm dinner will be served free of charge during the poster session.

The social dinner will be held on Thursday evening at Pizzeria da Pino, this meal is not subsidized.

Internet Access

An EDUROAM account is required to access the internet through the ICTP network at the Main Venue.

If you do not possess an EDUROAM account, you will receive an email from SISSA to create a guest account for the conference.

If you do not possess an EDUROAM account and you have not received an email to create a guest account, please contact the organizers.

Tutorials Lecture and Internet Access

A Google account is required to fully benefit from the Tutorial Lectures, as Google Colab will be mostly used as the tool to run calculations and examples.

Code of Conduct

We strive for making every attendee feel welcome and respected, regardless of gender identity, sexual orientation, disability, physical appearance, body size, race, nationality, religion, background, level of education, or socioeconomic status.

Do not hesitate to contact any of the organizers if you are the witness or the victim of any discrimination or harassment. Step up and speak out to stop any kind of inappropriate behaviour you witness. Bystander intervention creates safer communities and prevents harmful escalation.

Health and Safety Measures

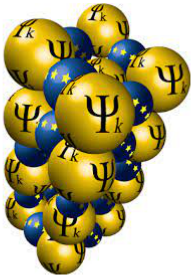
When you are indoor and you are not speaking, eating, or drinking, you are highly recommended to wear a mask covering your mouth and nose, according to the latest Italian laws on the matter. The use of a mask is mandatory on all Italian public transport such as bus, train, and planes.

If you feel unwell do not attend the conference in person and join the event online. Temperature checks will take place at the entrance of the conference venue, as per Italian Laws on the subject.

Every attendee should feel in a safe environment, make sure to agree and respect the right distance from others. If in doubt, a 1.5m distance is likely to ensure a lesser likelihood of COVID19 transmission.

Sponsors


CENTRE EUROPÉEN DE CALCUL ATOMIQUE ET MOLÉCULAIRE
cecam
SINCE 1969




MA PER SEGUIR VIRTUTE E CANOSCENZA
SISSA


ICTP


A!
Aalto University


WILEY


MAX
DRIVING
THE EXASCALE
TRANSITION

Programme

Start	End	Monday 09/05	Tuesday 10/05	Wednesday 11/05	Thursday 12/05	Friday 13/05
09:00	09:20		Anton Bochkarev	Aldo Glielmo	Francesca Grisoni	Zachary Ulissi
09:20	09:40					
09:40	10:00				Sebastiano Saccani	Jarno Laakso
10:00	10:20				Andrea Anelli	Jonathan Schmidt
10:20	10:40		Coffee Break	Coffee Break	Panel Discussion + Coffee	Panel Discussion + Coffee
10:40	11:00					
11:00	11:20		Milica Todorović	Franco Pellegrini		
11:20	11:40				Sintija Stevanoska	Jigyasa Nigam
11:40	12:00				Julia Westermeyer	Johannes Margraf
12:00	12:20					
12:20	12:40		Lunch Break	Lunch Break	Lunch Break	Remarks + Prizes
12:40	13:00					
13:00	13:20					
13:20	13:40	Registration + Remarks		Registration + Remarks		
13:40	14:00					
14:00	14:20	Felix Faber	Alessandro Laio	Boris Kozinsky	Lars Banko	
14:20	14:40					
14:40	15:00			Ilyes Batatia	I-J u Chen	
15:00	15:20			David P. Kovacs	Felix Arendt	
15:20	15:40	Coffee Break	Coffee Break	Coffee Break	Coffee Break	
15:40	16:00			Robin Winter	Rianne van der Berg	
16:00	16:20	Matthias Rupp	Kristoff Schütt	Tuan Le	Lopanitsyna Nataliya	
16:20	16:40			Yasemin B. Varolgünes	Zhi Li	
16:40	17:00			Sašo Džeroski	Núria López	
17:00	17:20					
17:20	17:40			Transport to SISSA		
17:40	18:00					
18:00	18:20			Poster Session + Dinner		
18:20	18:40					
18:40	19:00				Social Dinner	
19:00	19:20					
19:20	19:40					
19:40	20:00					
20:00	20:20					
20:20	20:40					
20:40	21:00					

Legend	
	Intro Lectures
	Tutorials
	Invited Talks
	Contributed Talks
	Industry Talks
	Food Breaks
	Other

Monday

13:00 - 13:45	Registration
13:45 - 14:00	Opening Remarks
14:00 - 15:20	**Felix Faber** Intro Lecture: Representing Materials; an Overview and Covering New Ground
15:20 - 16:00	Coffe Break
16:00 - 17:20	**Matthias Rupp** Intro Lecture: Introduction to Learning with Kernels

Representing Materials; an Overview and Covering New Ground

Felix Faber[1]

[1] *University of Cambridge, Cambridge, United Kingdom*

Many varieties of representations for materials have emerged over the past decade. They exist across a range of complexities and input spaces, from purely stoichiometry-based to representations that encode coordinates and internal degrees of freedom. These representations all have advantages and drawbacks, and using the right tool for the job is important for a successful outcome. I will give an overview of what I believe are the major categories of representations and will discuss their advantages, and drawbacks. I will then discuss a promising way to overcome the problem of innumerable search spaces when using coordinate-based representations to discover stable materials by coarse-graining crystal structures based on their possible symmetry operations.

Introduction to Learning with Kernels

Matthias Rupp[1]

[1] *University of Konstanz, Konstanz, Germany*

Kernel-based machine learning uses positive definite functions to systematically obtain nonlinear versions of linear algorithms such as ridge regression, Gaussian processes, support vector machines, principal component analysis, and many others. I will provide a brief introduction to learning with kernels, focusing on regression and interpolation of electronic-structure calculations as an application.

Tuesday

9:00 - 10:20	**Anton Bochkarev** Tutorial Lecture: Structural descriptors and atomic cluster expansion basis
10:20 - 11:00	Coffee Break
11:00 - 12:20	**Milica Todorović** Tutorial Lecture: Materials structure-property mapping with kernel-based methods
12:20 - 14:00	Lunch Break
14:00 - 15:20	**Alessandro Laio** Intro Lecture: Automatic topography of multidimensional probability densities
15:20 - 16:00	Coffee Break
16:00 - 17:20	**Kristoff Schütt** Intro Lecture: Neural networks for materials applications

Structural descriptors and atomic cluster expansion basis.

Anton Bochkarev[1]

[1] *Rhur Bochum Universitait, Bochum, Germany*

Most modern machine learning interatomic potentials (MLIP) assume that the system energy can be represented as a sum of atomic energy contributions which each depend on their local atomic environments. A descriptor maps the local atomic environment to a number. Here, we consider a few common descriptors and discuss the difference between descriptor and basis-based MLIPs with the focus on the atomic cluster expansion (ACE). We demonstrate a connection between the ACE basis and other common structural descriptors. Additionally, we will train a simple ACE potential and demonstrate how one can use it for materials simulations.

Materials structure-property mapping with kernel-based methods

Milica Todorović[1]

[1] *University of Turku, Turku, Finland*

Automatic topography of multidimensional probability densities

Alessandro Laio[1]

[1] *Scuola Internazionali di Studi Superiori Avanzati, Trieste, Italy*

Unsupervised methods in data analysis aim at obtaining a synthetic description of high-dimensional data landscapes, revealing their structure and their salient features. We will describe an approach for charting complex and heterogeneous data spaces, providing a topography of the high-dimensional probability density from which the data are harvested. We obtain information on the number and the height of the probability peaks, the depth of the "valleys" separating them, the relative location of the peaks and their hierarchical organisation. The topography is reconstructed by using an unsupervised variant of Density Peak clustering [1,2] exploiting a non-parametric density estimator [3], which automatically measures the density in the manifold containing the data [4]. Importantly, the density estimator provides an estimate of the error. This is a key feature, which allows distinguishing genuine probability peaks from density fluctuations due to finite sampling. We show that this approach allows identifying the Markov States explored during a protein folding molecular dynamic trajectory directly from the shape of the mulidimensional probability density, namely without exploiting any kinetic information [5].

[1] Science, 1492, vol 322 (2014)
[2] Inf. Sci., doi.org/10.1016/j.ins.2021.01.010 (2021)
[3] JCTC ,1206, vol 14 , (2018)
[4] Sci Rep. 12140, vol 7 (2017)
[5] JCTC 80, vol 1, (2020)

Neural networks for materials applications

Kristoff Schütt[1]

[1] *MPI, Berlin, Germany*

Neural networks have become a powerful tool to model potential energy surfaces and predict chemical properties. Starting from descriptor-based neural network approaches, this talk will focus on end-to-end learning of atomic representations and the incorporation of physical constraints. Finally, I will highlight some advanced applications of these techniques such as field-dependent potentials and generative neural networks.

Wednesday

09:00 - 10:20	**Aldo Glielmo** Tutorial Lecture: Data-manifold characterisation: Estimating intrinsic dimension, density, and density peaks with DADApy
10:20 - 11:00	Coffee Break
11:00 - 12:20	**Franco Pellegrini** Tutorial Lecture: Training neural network potentials with PANNA
12:20 - 13:40	Lunch + Registrations
13:40 - 14:00	Opening Remarks
14:00 - 14:40	**Boris Kozinsky** Symmetry and uncertainty-aware models of interatomic interactions for fast molecular dynamics
14:40 - 15:00	**Ilyes Batatia** A Unified Understanding of Equivariant Interatomic Potentials
15:00 - 15:20	**David Peter Kovacs** Atomic Cluster Expansion based Force Fields for Molecular Materials
15:20 - 15:40	Coffee Break
15:40 - 16:00	**Robin Winter** Unsupervised Learning of Group Invariant and Equivariant Representations and its Application to Molecular Conformations
16:20 - 16:40	**Tuan Le** Unsupervised Representation Learning on Molecular Conformations
16:20 - 16:40	**Yasemin Bozkurt Varolgünes** Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders
16:40 - 17:20	**Sašo Džeroski** Learning explainable models from complex data with applications in QSAR (quantitative structure-activity relationships) modelling
17:30 - 20:00	Poster session + Dinner

Data-manifold characterisation: Estimating intrinsic dimension, density, and density peaks with DADApy

Aldo Glielmo[1]

[1] *Banca d'Italia.*

Data in physics and chemistry come in the form of very high dimensional descriptors possessing hundreds of even thousands of components, but they typically lie on manifolds of much lower dimensionality and a rich set of hidden properties. In this tutorial, we will overview some common numerical techniques to analyse fundamental characteristics of data-manifold. We will cover methods to estimate the intrinsic dimension, the manifold density, and the peaks of such density by accompanying a brief theoretical explanation of the methods with "hands-on" sessions with applications on toy datasets. The tutorial will extensively use the DADApy package, available at https://github.com/sissa-data-science/DADApy.

Training neural network potentials with PANNA

Franco Pellegrini[1]

[1] *Scuola Internaziona di Studi Superiori Avanzati, Trieste, Italy.*

In this tutorial, we will present the basic ideas behind neural network potentials (NNP) and show you how to create one from scratch with our software PANNA. We will cover the whole pipeline: starting from ab initio simulations to compute local descriptors, packing them and feeding them to the NN training algorithm, monitoring and validating training and finally extracting the resulting NNP to be deployed in real life applications. The tutorial will be contained in a convenient notebook environment and will run on lightweight sample data that can be processed in real time on your laptop or on the cloud. However, all the tools presented and skills acquired will be directly translatable to a production environment.

Symmetry and uncertainty-aware models of interatomic interactions for fast molecular dynamics

Boris Kozinski [1]

[1] *Harvard University Camridge, United States of America*

Harnessing the accuracy of quantum mechanics to design complex materials requires a series of approximations to reach the desired length and time scales. I will describe our pursuit of the paradigm of "ex-machina" computations where data-driven approximations are automatically developed using machine learning algorithms and enable access to previously intractable systems. Non-parametric regression methods allow for learning of potential energy surfaces from expensive quantum calculations. To accelerate molecular dynamics calculations, we developed the Neural equivariant interatomic potential model (NequIP) based on tensor-valued symmetry-preserving layer architectures and used them to achieve state-of-the-art accuracy and training efficiency for simulating dynamics of molecules, liquids, heterogeneous catalysts, and ionic conductors [1]. In order to enable autonomous selection of the training set for reactive systems, we developed the FLARE adaptive closed-loop algorithm that constructs accurate and uncertainty-aware Bayesian force fields on-the-fly from a molecular dynamics simulation, using Gaussian process regression [2]. We demonstrate the performance of ML-accelerated MD simulations by studying 2D-to-3D transformations of layered quantum materials [3] and catalyst dynamics [4,5].

[1] S. Batzner et al, Arxiv:2101.03164 (2021)
[2] J. Vandermause, et al, NPJ Computational Materials, 6, 20 (2020)
[3] Y. Xie et al, NPJ Computational Materials, 7, 40 (2021)
[4] J. S. Lim, et al, JACS. 2020, 142, 37, 15907–15916 (2020)
[5] J. Vandermause et al, arXiv:2106.01949 (2021)

A Unified Understanding of Equivariant Interatomic Potentials

Ilyes Batatia [1]

[1] *ENS Paris-Saclay, Paris, France*

Equivariant interatomic potentials constitute the general class of symmetry-aware data-driven models learning potential energy surfaces from ab initio data. In this talk, I will demonstrate how the Atomic Cluster Expansion (ACE) framework can be extended to provide a unifying understanding of machine learning potentials including Message Passing Neural Networks (MPNNs) like SchNet and NequIP, Behler-Parinallo neural networks as well as the Gaussian Process regression-based SOAP-GAP approaches. The classification of a wide range of models in the new design space generated by this framework provides a novel tool to probe the key choices made by the different approaches. In particular, I will provide a systematic analysis of the design choices of NequIP, the state-of-the-art MPNN potential, and assess the key elements to its success. Finally, I will present BOTNet, a simplification of NequIP with an interpretable architecture reaching state-of-the-art accuracy on a set of challenging datasets.

Atomic Cluster Expansion based Force Fields for Molecular Materials

David Kovacs [1]

[1] *University of Cambridge, Cambridge, United Kingdom*

Machine Learning based force fields have revolutionised the modelling of materials at the atomistic scale. I will demonstrate how Atomic Cluster Expansion (ACE) can be used to build highly accurate and fast force fields. I will demonstrate that ACE force fields parametrised using regularised linear regression can compete in accuracy with most Gaussian Process and Neural Network based approaches. In particular, I will describe several applications of ACE to molecular systems where it shows excellent smooth and physical extrapolation to unseen parts of the Potential Energy Surface. I will also show examples of reactive simulations and strategies for including long-range interactions into Machine Learned force fields. Finally, I will briefly describe the multi-ACE framework, which unifies all equivariant machine learning interatomic potentials including Message Passing Neural Networks.

Unsupervised Learning of Group Invariant and Equivariant Representations and its Application to Molecular Conformations

Robin Winter[1]

[1] *Bayer, AG*

Equivariant neural networks, whose hidden features transform according to representations of a group G acting on the data, exhibit training efficiency and an improved generalisation performance. In this work, we extend group invariant and equivariant representation learning to the field of unsupervised deep learning. We propose a general learning strategy based on an encoder-decoder framework in which the latent representation is disentangled in an invariant term and an equivariant group action component. We describe explicitly our construction for rotations, translations and permutations and test the validity and the robustness of our approach in a variety of experiments with diverse data types employing different network architectures. Finally, we demonstrate how our proposed framework can be directly applied on molecular conformations expressed in Cartesian coordinates.

Unsupervised Representation Learning on Molecular Conformations

Tuan Le[1]

[1] *Bayer, AG*

We study the generation of 3D molecular conformations from a representation learning perspective by autoencoding conformation described as a point cloud into a fixed-sized latent embedding that serves as informative bottleneck to reconstruct the original conformation. To faithfully represent the underlying geometry of a conformation and perform an accurate reconstruction of its spatial coordinates after decoding, we implement an equivariant graph neural network whose intermediate feature representations change according to rigid transformations. We aim to achieve the generation of diverse conformations, by formulating the learning problem as a self-supervised task and include a conditional prior distribution on the conformation through the topological molecular graph.

Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders

Yasemin Bozkurt Varolgunes[1]

[1] *Max Planck Institute for Polymer Research, Germany*

Extracting insight from the enormous quantity of data generated from molecular simulations requires the identification of a small number of collective variables whose corresponding low-dimensional free-energy landscape retains the essential features of the underlying system. Data-driven techniques provide a systematic route to constructing this landscape, without the need for extensive a priori intuition into the relevant driving forces. In particular, autoencoders are powerful tools for dimensionality reduction, as they naturally force an information bottleneck and, thereby, a low-dimensional embedding of the essential features. While variational autoencoders ensure continuity of the embedding by assuming a unimodal Gaussian prior, this is at odds with the multi-basin free-energy landscapes that typically arise from the identification of meaningful collective variables. In this work, we incorporate this physical intuition into the prior by employing a Gaussian mixture variational autoencoder (GMVAE), which encourages the separation of metastable states within the embedding. The GMVAE performs dimensionality reduction and clustering within a single unified framework, and is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. We illustrate our approach on two toy models, alanine dipeptide, and a challenging disordered peptide ensemble, demonstrating the enhanced clustering effect of the GMVAE prior compared to standard VAEs. The resulting embeddings appear to be promising representations for constructing Markov state models, highlighting the transferability of the dimensionality reduction from static equilibrium properties to dynamics.

Learning explainable models from complex data with applications in QSAR (quantitative structure-activity relationships) modelling

Sašo Džeroski [1]

[1] *Jozef Stefan Institute, Ljubljana, Slovenia*

We will present our recent methods for machine learning of explainable models from complex data. This may involve structured inputs (e.g., descriptions of chemicals by graphs), as in relational classification, and/or structured outputs (e.g., multiple, possibly hierarchically structured targets), as in multi-target regression. Other forms of complexity that we consider involve partial annotations with target values (as in semi-supervised learning) and massive or streaming data (as in data streams).

We will also present two applications of these methods in the context of drug discovery and re-purposing. One of these concerns the re-purposing of drugs as host-directed antimicrobials for the treatment of tuberculosis and salmonella, while the other focuses on the discovery of compounds to prevent lung fibrosis. Our approaches are general and we are currently exploring their use in materials science.

Thursday

09:00 - 09:40	**Francesca Grisoni** Chemical language models for de novo molecule design
09:40 - 10:00	**Sebastiano Saccani** Industrial applications of generative machine learning methods
10:00 - 10:20	**Andrea Anelli** Exploring drugs' solid state landscape using atomistic machine learning
10:20 - 11:20	Panel discussion and coffee
11:20 - 11:40	**Sintija Stevanoska** Towards predicting corrosion inhibitors' performance with machine learning
11:40 - 12:20	**Julia Westermayr** Deep learning for excited states and molecular design
12:20 - 14:00	Lunch
14:00 - 14:40	**Lars Banko** Data-driven high-throughput experimentation using combinatorial material science methods and machine learning
14:40 - 15:00	**I-Ju Chen** Precise atom manipulation through deep reinforcement learning
15:00 - 15:20	**Felix Arendt** Evaluation of descriptors for property predictions of glasses by machine learning
15:20 - 15:40	Coffee break
15:40 - 16:00	**Rianne van den Berg** AI4science at Microsoft Research
16:00 - 16:20	**Lopanitsyna Nataliya** Alchemical machine learning for high entropy alloys
16:20 - 16:40	**Zhi Li** The phase diagram of iron up to Earth's inner core conditions
16:40 - 17:20	**Nuria Lopez** Machine Learning techniques in Heterogeneous Catalysis
19:00 - 20:40	Social Dinner, da Pino

Chemical language models for de novo molecule design

Francesca Grisoni [1]

[1] *Technical University of Eindhoven, Eindhoven, Netherlands*

Deep learning is fueling computer-aided molecule discovery.[1,2] Chemical language models (CLMs) are one of the most recent additions to the chemist's toolkit for AI-driven molecule design.[3–5] CLMs can be used to generate novel molecules in the form of strings (e.g., SMILES or amino-acid sequences) without relying on human-engineered assembly or enumeration rules. Thanks to such a 'rule-free' character, CLMs allow navigating the chemical space and generating focused chemical libraries.[6,7] In multiple instances, CLMs have shown able to learn "grammar" rules for molecule construction, and to implicitly capture "semantic" features, such as physicochemical properties, bioactivity, and chemical synthesizability.[3,7,8] This talk will illustrate some successes of CLMs for drug discovery, at the interface between method development and prospective application, e.g., the design of natural-product-inspired modulators of nuclear receptors,[9] and the combination with microfluidics-assisted synthesis.[10] Moreover, the talk will provide insights into recent advances towards automation of the design-make-test-analyze cycle.

[1] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. Drug Discov. Today 2018, 23 (6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

[2] Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. Expert Opin. Drug Discov. 2021, 16 (9), 949–959. https://doi.org/10.1080/17460441.2021.1909567.

[3] Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent. Sci. 2018, 4 (1), 120–131.

[4] Yuan, W.; Jiang, D.; Nambiar, D. K.; Liew, L. P.; Hay, M. P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.-T.; Tibshirani, R.; others. Chemical Space Mimicry for Drug Discovery. J. Chem. Inf. Model. 2017, 57 (4), 875–882.

[5] Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. Mol. Inform. 2018, 37 (1–2), 1700153. https://doi.org/10.1002/minf.201700153.

[6] Skinnider, M. A.; Stacey, R. G.; Wishart, D. S.; Foster, L. J. Chemical Language Models Enable Navigation in Sparsely Populated Chemical Space. Nat. Mach. Intell. 2021, 3 (9), 759–770. https://doi.org/10.1038/s42256-021-00368-1.

[7] Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative Molecular Design in Low Data Regimes. Nat. Mach. Intell. 2020, 2 (3), 171–180.

[8] Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. J. Chem. Inf. Model. 2020, 60 (3), 1175–1183. https://doi.org/10.1021/acs.jcim.9b00943.

[9] Moret, M.; Helmstädter, M.; Grisoni, F.; Schneider, G.; Merk, D. Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence. Angew. Chem. Int. Ed. 2021, 60 (35), 19477–19482. https://doi.org/10.1002/anie.202104405.

[10] Grisoni, F.; Huisman, B. J.; Button, A. L.; Moret, M.; Atz, K.; Merk, D.; Schn

Industrial applications of generative machine learning methods

Sebastiano Saccani [1]

[1] *AIndo*

Generative machine learning methods are a paradigm that is receiving a lot of investments from the biggest IT and industrial companies worldwide. We will tackle the relationship existing between self-supervised learning and generative models and showcase industrial applications that exploit such techniques.

Exploring drugs' solid state landscape using atomistic machine learning

Andrea Anelli [1]

[1] *Roche, Hoffman-La Roche, Basel, Switzerland*

Organic molecules can adopt different solid forms depending on their chemistry and external conditions, generating a polymorphic degree of freedom. Since the different solid forms have distinct physico-chemical behavior (e.g. solubility, compressibility etc. . .), it is paramount to know which crystalline patterns a molecule can aggregate into. A promising route to characterize the crystalline landscape of organic crystals is to perform quantum mechanical atomistic simulations of each potential polymorph, so as to obtain a much needed ranking of the observable crystals. While this routine is fully general and applicable to any class of compounds, it requires a substantial computational investment, allowing its use mostly for molecules in the later development stages. To overcome these costs, and render this approach truly high-throughput, we propose an infrastructure to speed our calculations by training a machine learning crystal ranker. By using our method, we report n-fold increases in performances, paving the way for crystal structure prediction to penetrate in the earlier development stages.

Towards predicting corrosion inhibitors' performance with machine learning

Sintija Stevanoska [1]

[1] *Lubjana, Slovenia*

Organic corrosion inhibitors are molecular substances used in relatively low concentrations that effectively reduce the rate of corrosion of metals and their alloys. It has been widely believed that corrosion inhibition is related to inhibitor molecular electronic properties, and one commonly utilized inference is that efficient inhibitors display a small HOMO–LUMO gap. However, studies that considered many corrosion inhibitors demonstrated the lack of correlation between a given individual molecular electronic property and corrosion inhibition.[1,2] For this reason, several studies utilized machine-learning (ML) approaches to generate predictive models for screening new inhibitors.[2,3] These studies showed that the usefulness of inhibitor molecular electronic properties as input features in ML models depends on the metallic material to be protected, i.e., they were found useful for pure Mg,[3] but not for Al alloys.[2] Herein, we apply several ML approaches for regression (including model trees, regression trees and random forests thereof, and multiple linear regression) to screen heterocyclic organic compounds as corrosion inhibitors of pure copper in 3 wt.% NaCl solution. Our molecular dataset consisted of 24 heterocyclic organic compounds, whose corrosion inhibition performance was determined experimentally.[1] We utilized two classes of input features (named class-1 and class-2). AlvaDesc 2.0[4] and MarvinSketch 21.9[5] were used to generate 3984 class-1 descriptors, including constitutional and topological parameters and molecular physicochemical variables. Preprocessing steps were performed in AlvaDesc 2.0 to reduce the total number of descriptors to a set of 356 information-rich variables. As for class-2, we utilized 31 different molecular electronic parameters[6]—such as ionization potential, electron affinity, electronegativity—calculated with the B3LYP density-functional-theory method. We built ML models by using either class-1, class-2, or class-1+class-2 input features. As the molecular dataset was small, we used leave-one-out cross-validation. Results indicate that neither class-1 nor class-2 input features (nor their combination) can provide highly predictive models (the best models have a cross-validated correlation coefficient of $r \sim 0.5$). This limited success implies that neither the molecular electronic properties nor the constitutional and topological parameters are useful descriptors for corrosion inhibition. The challenge is thus to find useful descriptors for corrosion inhibition. This is in line with our view that it is currently unknown what makes a good corrosion inhibitor in terms of physicochemical properties and how it differs from a bad inhibitor.

[1] A. Kokalj at al., Corros. Sci. 2021, 179, 108856.
[2] D. A. Winkler et al., Green Chem. 2014, 16, 3349–3357.
[3] C. Feiler et al., Corros. Sci. 2020, 163, 108245.
[4] A. Mauri, in Ecotoxicological QSARs (Ed.: K. Roy), Springer US, New York, NY, 2020, pp. 801–820.
[5] Calculator Plugins, Marvin 21.9, ChemAxon, 2021. http://www.chemaxon.com
[6] A. Kokalj et al., Corros. Sci. 2021, 193, 109900.

Deep learning for excited states and molecular design

Julia Westermayr [1], Joe Gilkes[1], Rhyan Barrett[1], Reinhard J. Maurer[1]

[1] *University of Warwick, Warwick, United Kingdom*

High-throughput screening of excited states in molecules can advance the search for functional organic molecules with potential importance for modern organic electronics. However, time-consuming experiments and high computational costs of accurate quantum chemical calculations severely limit the characterization of such systems and consequently targeted molecular design [1].

In this talk, we will show how deep learning can advance this research area by using an automated approach to molecular design that combines two deep learning techniques: The first is based on a physics-inspired deep learning model for high-throughput screening of optical and electronic properties with near experimental accuracy [2]. The second technique we use is generative deep learning, which allows us to generate novel molecules by learning the structural distribution of a dataset. In successive iterations, the output of the generative model is filtered with the spectroscopic deep learning model and then used to retrain the generative model with a bias. In this way, molecules with finely tuned optoelectronic properties can be efficiently generated [3].

[1] J. Westermayr and P. Marquetand, Chem. Rev., 121(16) 121, 16, 9873–9926 (2020).
[2] J. Westermayr and R. J. Maurer, Chem. Sci. 12, 10755-10764 (2021).
[3] J. Westermayr, J. Gilkes, R. Barrett, R. J. Maurer, unpublished (2022).

Data-driven high-throughput experimentation using combinatorial material science methods and machine learning

Lars Banko [1]

[1] *Ruhr Universitait Bochum, Bochum, Germany*

Machine learning offers an enormous potential to accelerate the discovery of new material systems and the optimization of existing ones. The applications are manifold, e.g., for the analysis of complex data sets, for the prediction of new materials and the active control of experiments. Combinatorial materials science allows relatively large data sets to be generated through parallel synthesis and high-throughput measurements. Increasingly complex chemical composition spaces, however, pose challenges to the current methodology. This talk will introduce some concepts and application examples of machine learning in materials science, covering synthesis-processing-microstructure-relationships, machine learning companion agents for X-ray diffraction data analysis, and an outlook towards autonomous experimentation.

Precise atom manipulation through deep reinforcement learning

I-Ju Chen [1]

[1] *Aalto University, Helsinki, Finland*

Atom manipulation in scanning tunneling microscopy has enabled the creation of artificial lattices for research on exotic quantum states and atomic-scale miniaturization of computational devices. The ability to autonomously arrange atomic structures with precision will enable the scaling up of nanoscale fabrication and expand the range of artificial structures. However, it is challenging to select manipulation parameters that can achieve atomic precision throughout extended operations due to spontaneous tip apex changes and the difficulty of modeling tip-atom interactions. Here we use deep reinforcement learning to control the real-world atom manipulation process. Several state-of-the-art reinforcement learning techniques are used jointly to boost data efficiency. The deep reinforcement learning agent learns to manipulate Ag adatoms on Ag(111) surfaces with optimal precision and is integrated with path planning algorithms to complete an autonomous and robust atomic assembly software toolkit. These results demonstrate that state-of-the-art deep reinforcement learning can offer effective solutions to real-world challenges in nanofabrication. In particular, we expect deep reinforcement learning to be a promising approach to discover manipulation parameters for novel surface/adsorbate combinations.

Evaluation of descriptors for property predictions of glasses by machine learning

Felix Arendt [1]

[1] *Friedrich Schiller University Jena, Jena, Germany*

The development of new glasses is often hampered by inefficient "trial-and-error" design discovery methods. As an alternative, machine learning (ML)-based approaches are becoming increasingly popular to accelerate the discovery and optimization of novel advanced materials. The most Subsequently, the glass structure is used as an additional source of information to improve the predictive performance of the ML models. To this end, high-throughput molecular dynamics simulations of glasses are performed and image recognition methods such as beta-variational autoencoders are used to extract descriptors related to structural information. This combined approach of ab initio descriptors and structural descriptors is expected accelerate the development of glasses with tailored properties. Subsequently, the glass structure is used as an additional source of information to improve the predictive performance of the ML models. To this end, high-throughput molecular dynamics simulations of glasses are performed and image recognition methods such as beta-variational autoencoders are used to extract descriptors related to structural information. This combined approach of ab initio descriptors and structural descriptors is expected accelerate the development of glasses with tailored properties. source of information for deriving ML models for glasses is their chemical composition. Here we evaluate different descriptors for ML models of various glass properties based on both compositional information as well as ab initio derived properties of their components. These ab initio descriptors can reproduce the performance of their purely compositional counterparts for a large, comprehensive glass dataset, and in some cases even improve upon it for a smaller dataset covering a variety of oxides, chalcogenides, and metallic glasses.

AI4science at Microsoft Research

Rianne van den Berg [1]

[1] *Microsoft Research Amsterdam, Netherlands*

Over the last 10 years deep learning has clearly disrupted fields like computer vision, speech recognition and language modeling. In this talk I will explain why we at Microsoft research think AI is going to disrupt molecular sciences, and in particular computational chemistry and physics. I will discuss the research areas that we are currently exploring in Cambridge UK and in our new lab in Amsterdam the Netherlands, ranging from drug discovery and material generation to computational catalysis.

Alchemical machine learning for high entropy alloys

Lopanitsyna Nataliya[1]

[1] *Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland*

High entropy alloys (HEAs) are a class of metallic materials composed of five or more principal elements[1]. Interest in HEAs has grown over the last decades due to their exceptional structural and mechanical properties. HEAs are particularly challenging for atomistic modeling. Conventional empirical forcefields are usually inaccurate for their complex compositions, but first-principles simulations are very demanding, limiting the sampling of disorder, and the time and length scales that are reachable.

Machine-learning models have emerged as a promising alternative, with the ability to deliver the accuracy of first principle methods with lower computational resources. However, the complexity of machine learning models grows rapidly with the number of different elements due to the unfavourable scaling of their associated feature space sizes, limiting the chemical diversity of the systems tackled thus far. The exponential increase in feature dimensionality poses two distinct challenges. The former is that memory and computational requirements to evaluate full feature vectors grows out of control with the chemical diversity. The latter is that problems with high dimensionality require large descriptive datasets to achieve transferability and low validation errors due to the interpolative nature of machine-learning models.

To address these points, first, we propose a PyTorch-based chemical embedding compression scheme to reduce the dimensionality of the feature space required for multi-component systems based on Willat et al framework[2]. Second, we demonstrate generated training set and the associated energy and forces computed at the ab initio level, including BCC and FCC phases. Finally, to showcase the effectiveness of our approach, we present a benchmark on the elpasolites dataset[3] and an application to our proposed HEA dataset, including phase stability prediction.

[1] Yeh, J.W., et al. "Nanostructured high entropy alloys with multiple principal elements: novel alloy design concepts and outcomes." Advanced engineering materials 6.5 (2004): 299-303.

[2] Willatt, Michael J., Félix Musil, and Michele Ceriotti. "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements." Physical Chemistry Chemical Physics 20.47 (2018): 29661-29668.

[3] Faber, Felix A., et al. "Machine learning energies of 2 million elpasolite (ABC2D6) crystals." Physical review letters 117.13 (2016): 135502.

The phase diagram of iron up to Earth's inner core conditions

Zhi Li [1]

[1] *ICTP, Trieste, Italy*

The primary chemical composition of Earth's inner core is known to be iron. However, it is still debated how the atoms in the solid iron are arranged. The possible candidates include body-centered cubic (bcc), hexagonal-close-packed (hcp), and face-centered cubic (fcc) structures. Here we developed a deep learning interatomic potential that allows attaining ab initio accuracy but with scaling that cannot be accessed by density functional theory. By coupling molecular dynamics simulation with the thermodynamic integration method to evaluate the Gibbs free energy, we construct the solid iron phase diagram up to 700 GPa. We also determine the physical properties like thermal conductivity of these phases at Earth's inner core conditions. Our results will provide new insight into explaining the seismic observations associated with the inner core, such as seismic anisotropy, strong attenuation, and low shear wave velocity.

Machine Learning techniques in Heterogeneous Catalysis

Nuria Lopez [1]

[1] *Institute of Chemical Research of Catalonia, Tarragona, Spain*

In the presentation I will review some of the uses of machine learning in Heterogeneous Catalysis that we have been developing in our group. First we will focuss on the determination of descriptors for complex reactivity and how these descriptors can be mapped into the traditional methodologies. Second I will use these as feed for performance equations, able at presenting genaral forms for the reactivity of metals through hybrid data (mixing experiments and theory). The last part of the talk will be devoted to "smart characterization techniques", in them using deep learning methodologies the most complex images from microscopes can be analyzed.

Friday

09:00 - 09:40	**Zachary Ulissi** Continued progress towards generalizable machine learning models in computational catalysis
09:40 - 10:00	**Jarno Laakso** Compositional Engineering of Perovskites for Solar Energy Applications with Machine Learning
10:00 - 10:20	**Jonathan Schmidt** Machine Learning Thermodynamic Stability of Materials
10:20 - 11:20	Panel discussion and coffee
11:20 - 11:40	**Jigyasa Nigam** Equivariant representations for machine learning molecular Hamiltonians
11:40 - 12:20	**Johannes Margraf** Predicting Molecular Properties through Machine Learned Energy Functionals
12:20 - 12:40	Closing remarks and prize announcement

Continued progress towards generalizable machine learning models in computational catalysis

Zachary Ulissi [1]

[1] *Carnegie Mellon, Pitsburg, United States of America*

Machine learning accelerated catalyst discovery efforts have seen much progress in the last few years. Datasets of computational calculations have improved, models to connect surface structure with electronic structure or adsorption energies have gotten more sophisticated, and active learning exploration strategies are becoming routine in discovery efforts. However, there are several large challenges that remain: to date, models have had trouble generalizing to new materials or reaction intermediates and applying these methods requires significant training. To address these challenges, I will briefly introduce the Open Catalyst Project and the Open Catalyst 2020 dataset, a collaborative project with Facebook AI Research to span surface composition, structure, and chemistry and enable a new generation of deep machine learning models for catalysis. I will then discuss initial results for state-of-the-art deep graph convolutional models and significant recent progress from others in the community, including academic and industrial AI labs. Innovation in these models is likely to improve models in related materials science areas. Finally, I will discuss current efforts and open challenges for deep graph networks and beyond in computational catalysis, and how to use the models/datasets or contribute new methods to the open leaderboard.

Compositional Engineering of Perovskites for Solar Energy Applications with Machine Learning

Jarno Laakso [1]

[1] *Aalto University, Helsinki, Finland*

Perovskites solar cells are a promising, emergent green energy technology. They have shown high power conversion efficiencies, rivaling traditional silicon based solar cells, but their commercialization has been hindered by toxicity and lack of stability. Studies of mixed perovskite materials have demonstrated that compositional engineering can effectively mitigate these problems [1], but the complexity of the perovskite materials space inhibits the search for a highly efficient, non-toxic, and stable solar cell material. We here facilitate this task with a machine learning (ML) accelerated computational approach.We developed a ML model that utilizes the many-body tensor representation for the perovskite atomic structure and kernel ridge regression to learn from density functional theory calculations. We trained the model on CsPb(Cl/Br)3 perovskite data with randomized Cl-Br configurations and structural variations, and used it to quickly predict energies, atomic forces, and stresses. We then employed the ML model in MC sampling combined with structural relaxations to gain access to low energy structures and compute the convex hull for CsPb(Cl/Br)3. This data-driven approach offers a pathway to the study of more complex perovskites and other alloy materials with quantum mechanical accuracy.

Machine Learning Thermodynamic Stability of Materials

Jonathan Schmidt [1]

[1] *Martin-Luther University Halle-Wittenberg, Germany*

In recent years machine learning methods have greatly accelerated the theoretical discovery of new stable materials. Models for the prediction of material properties have evolved from simple elemental descriptors and decision tree models, over convolutional networks to graph neural networks. Graph neural networks for crystal structures typically use the atomic positions and the atomic species as input. Unfortunately, this information is not available when predicting new materials, for which the precise geometrical information is unknown. Crystal-graph attention networks replace these precise bond distances with embeddings of graph distances. This allows for the application in high-throughput studies based on both compositions and crystal structure prototypes without using relaxed structures as input. With these techniques we have already scanned hundreds of ternary prototypes spanning a space of more than 200M compounds and identified thousands of theoretically stable materials.

Equivariant representations for machine learning molecular Hamiltonians

Jigyasa Nigam [1]

[1] *EPFL, Lausanne, Switzerland*

Most of the widely used machine learning schemes that have been successful in predicting chemical and material properties rely on a mathematical representation of atomic configurations that reflects the physical symmetries and locality of learning targets. A class of these structural descriptions is built as a hierarchy of correlations of atom centered densities[1] and is subsequently used to model corresponding atomic properties, or atomic contributions to a global observable. However, many quantum mechanical quantities, such as the effective single-particle Hamiltonian written on an atomic-orbital basis, are associated with multiple atom-centers, rendering the atom-centered approach inadequate to describe the additional degrees of freedom. We recently proposed an N-center representation[2] that extends the atom-centered framework to the case of such targets. Devising this family of multicenter and high-correlation order representations opens avenues for new classes of machine learning models that are fully equivariant and thus incorporate molecular symmetries, and that can serve to assist electronic structure calculations.

[1] J. Nigam, S. Pozdnyakov, M. Ceriotti, JCP 153,121101, 2020
[2] J. Nigam, M. Willatt, M. Ceriotti, JCP 156, 014115, 2022

Predicting Molecular Properties through Machine Learned Energy Functionals

Jhoannes Margraf [1]

[1] *Fritz-Haber-Institut der MPG, Berlin, Germany*

My recent work focuses on using machine learning (ML) to understand chemical phenomena (such as the nature of complex reaction networks)(1) and predict the properties of new molecules and materials(2). A major driver of this work is the desire to build accurate, data-efficient models, which do not require enormous reference datasets for training. This is because I want to be able to apply these methods to any problem of chemical interest, not just to those for which "big data" happens to be available. This is achieved by using strong physical priors, e.g. in the form of baseline models or by tightly integrating ML and electronic structure theory(3,4). In addition to being data-efficient, the latter approach also expands the applicability of ML in chemistry, by providing access to electronic properties and energies on an equal footing.

[1] Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J.T. Nat. Commun. 2020, 11, 5505.
[2] Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J.T. Chem. Sci. 2021, 12, 4536-4546.
[3] Margraf, J. T.; Reuter, K. Nat. Commun. 2021, 12, 344.
[4] Staacke, C.; Wengert, S.; Kunkel, C.; Csányi, G.; Reuter, K.; Margraf, J.T. Mach. Learn. Sci. Technol., in press.

Posters

Alaukik Saxena
A materials informatics framework to discover patterns in atom probe tomography data
Alena Vishina
High-throughput and Data-mining Search for Rare-earth-free Permanent Magnets
Alfonso Gallo Bueno
Unsupervised machine learning for detection of spurious structures
Amirhossein Naghdi Dorabati
Mechanical properties of crystalline W and Mo metals by Neural network interatomic potentials
Andreas Møller Slavensky
Searches for stable nano-sized silicates clusters in the interstellar medium
Antti Pihlajamäki
Orientation Adaptive Minimal Learning Machine for Atomic Forces
Avula Venkata Siva Nikhil
Prediction of Shear Viscosity of Complex Fluids Using Machine Learning Methods: Model Performance, Ranking, Interpolation, and Uncertainty Quantification
Carlo Maino
Excited State Machine Learning for Chromophores in Complex Environments
Cesare Malosso
Viscosity in water from first-principles and deep-neural-network simulations
Claudio Zeni
Exploring the robust extrapolation of high-dimensional machine learning potentials
Connor Allen
A Computationally Efficient Approach for Improving Phonon Representation in Machine Learned Potentials
Fabio Priante
Identifying Unknown Organic Molecules in Atomic Force Microscopy Images through Deep Generative Models
Hannes Kneiding
Machine Learning Quantum Properties of Transition Metal Complexes using Graph Neural Networks
Hyunwook Jung
Global optimization protocol for adsorbate geometries using an on-the-fly surrogate Gaussian Approximation Potential
Ibrahim Buba Garba
Temperature-dependent phonons from ab-initio simulations
Indranil Saha
Understanding Germanosilicate Hydrolytic (In)stability based on Germanium Distributions derived using Neural Network Potentials
Iuri Macocco
An intrinsic dimension estimator for discrete-metric spaces
Jean Baptiste Fankam
Theoretical investigation of the molecular structure, vibrational spectra, thermodynamic and nonlinear optical properties of 4, 5-dibromo-2, 7dinitro- fluorescein
Johannes Döhn
Computational Screening of Oxide Perovskites as Insertion Type Cathode Materials
Kevin Rossi
Data-driven modelling and characterization of Au nanoparticles melting
Khatereh Azizi
Unsupervised learning on solvation of molecules in water

Kunal Ghosh
Assessing the potential of active machine learning for curating molecular datasets
Kushal Ramakrishna
Physics-informed and data-driven molecular dynamics simulations of Iron under extreme conditions
Kyeonghyeon Nam
Ab Initio Thermodynamics for Surface Motifs of the M1 Selective Oxidation Catalyst
Lakshmi Shenoy
Developing a machine learning interatomic potential for simulation of fracture in irradiated $alpha$-iron
Lars Leon Schaaf
Machine Learned Force Field for Oxides as Catalysts
Lauri Kurki
A scanning probe microscopy study into the hydration of an adenine self-assembled monolayer
Lea Gašparič
Towards predicting corrosion inhibitors' performance with machine learning
Lei Zhang
Active Learning of Gaussian Approximation Potential: Application for Fracture in Iron
Lucas Lang
Δ-Machine Learning with Equivariant Graph Neural Networks
Mahmoud Attia
Multiscale Modeling of lithium diffusion and NMR properties in ceramics solid electrolytes for the new generation of solid state batteries
Mandana Safari
Correct Vibrational Properties of Polar Materials from Neural Network Interatomic Potentials
Mani Lokamani
Evolution of Single-Level-Model parameters in the Mechanically controllable Break Junctions
Manuel Kuchelmeister
Multi-fidelity machine learning to accelerate materials research
Marco Bertolini
Unsupervised Learning of Group Invariant and Equivariant Representations
Mohsen Sotoudeh
Understanding ion mobility mechanism through the descriptor and scaling relations in solid crystallines
Nikolaj Rønne
Global atomistic optimisation enhanced by local surrogate model
Nejc Hodnik
Using Machine Learning to Predict Activity from Synthesis Parameters for PtCu Alloy Fuel Cell Nanoparticulate Electrocatalysts
Niko Oinonen
Molecule graph reconstruction from Atomic Force Microscopy images with machine learning
Nore Stolte
Reactions of aqueous carbon at deep Earth pressure and temperature conditions
Ondrej Krejci
Density Functional Theory models simulating Kelvin Probe Force Microscopy with flexible tip apices
Pablo Sánchez-Palencia Vallejo
Exploring the configurational space of spinel phase (Sn1-xGex)3N4 solid solutions with machine learning
Riccardo Dal Molin
Development a DeePMD potential for monolayer WTe2
Romina Wild
Identifying informative distance measures in high dimensional feature spaces: Application to COVID-19 severity prediction

Ruggero Lot
PANNA: a comprehensive toolkit for creating neural network models for atomistic systems
Samare Rostami
Exploring the novel mixed TiO2/ZrO2 structures and interfaces from structure predictions based on charge equilibration via neural network technique.
Sebastian Havens
Investigation of the phase behaviour of embedded atom models of metals using nested sampling and coexistence simulations
Simone Di Cataldo
Mapping Superconductivity in High-Pressure Hydrides
Uroš Hribar
Modeling the relation between processing parameters and material properties in foamed glass production with machine learning
Valerio Briganti
Extending spectral neighbour analysis potentials with long range physics
Wojciech Stark
Investigation of nonadiabatic effects for H2 at Cu surfaces: A unified machine-learned electronic friction model for multiple facets
Xi Chen
Baysian Optimization Conformer Search of Molecular Adsorbates on a Gold-thiolate Cluster
Yonghyuk Lee
Ab Initio Thermodynamics for Surface Motifs of the M1 Selective Oxidation Catalyst
Yuxuan Yao
Modified active machine learning (AML) approach to explore the molecular design through the use of surrogate models for charge injection and transport descriptors.