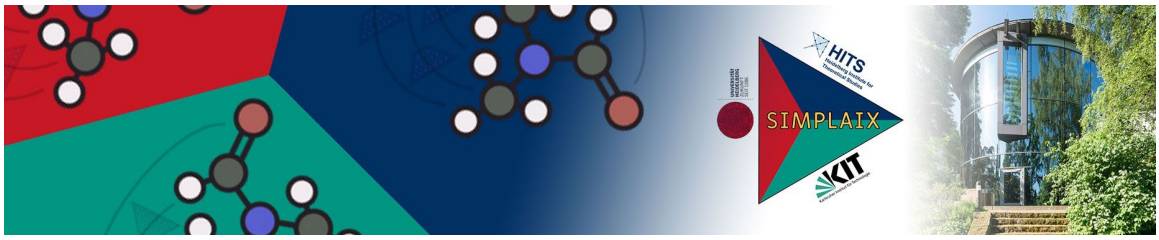

SIMPLAIX Workshop on "Machine Learning for Multiscale Molecular Modeling"

2 - 4 May 2023
Studio Villa Bosch Heidelberg, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg
https://simplaix-workshop2023.h-its.org/

Abstract Book

Organizing Committee: Rebecca Wade, Rostislav Fedorov, Frauke Gräter, Ganna (Anya) Gryn'ova (HITS); Fred Hamprecht, Andreas Dreuw, Ulrich Köthe (Heidelberg University); Pascal Friederich, Markus Elstner (KIT)

Organizers:

HITS Heidelberg Institute for Theoretical Studies

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

KIT Karlsruher Institut für Technologie

Sponsor:

Klaus Tschira Stiftung gemeinnützige GmbH

KTS

Program

Tuesday, May 2, 2023

13:00 - Registration
14:00 - Opening (Rebecca Wade, Tilmann Gneiting)

Session 1 (Chair: Frauke Gräter)

14:15 - Olexandr Isayev (Department of Chemistry, Carnegie Mellon University, Pittsburgh PA, USA)
Discovering chemistry with a transferable reactive machine learning potential (25 min talk + 15 min Q&A)
14:55 - Tristan Bereau (University of Amsterdam, Netherlands)
The role of coarse-graining in molecular discovery (25 min talk + 15 min Q&A)
15:40 - Gerhard Stock (Institute of Physics, Univ. Freiburg, Germany)
Correlation-based feature selection to identify functional dynamics in proteins (15 min talk + 5 min Q&A)

16:00 - Coffee Break

Session 2 (Chair: Prof. Dr. Andreas Dreuw)

16:30 - Sereina Riniker (ETH Zürich, Switzerland)
Learning Physical Interactions for Molecular Dynamics Simulations (25 min talk + 15 min Q&A)
17:10 - Yousung Yung (Seoul National University, S. Korea)
Machine-Enabled Chemical Structure-Property-Synthesizability Predictions (25 min talk + 15 min Q&A)
17:50 - Kirill Zinovjev (Universidad de Valencia, Spain)
Electrostatic embedding of Machine Learning potentials (15 min talk + 5 min Q&A)

18:15 - Group Photo (Speakers)

Wednesday, May 3, 2023

Session 3 (Chair: Rebecca Wade)

09:00 - Chris Oostenbrink (University of Natural Resources and Life Sciences, Vienna (BOKU), Austria)
Buffer region neural networks: A buffered scheme for polarizable QM/MM simulations with machine-learning (25 min talk + 15 min Q&A)
09:40 - Jochen Blumberger (University College London, UK)
Electron and Proton Transfer in Functional Materials from Neural Network Potentials (25 min talk + 15 min Q&A)
10:20 - Giorgia Brancolini (CNR, Modena, Italy)
Combining Neural Networks, Enhanced Sampling simulations and FRET technique to study the Structure and Dynamics of of Intrinsically Disordered Proteins of Therapeutic Value (15 min talk + 5 min Q&A)

10:40 - Coffee Break

Session 4 (Chair: Fred A. Hamprecht)

11:20 - Rocio Mercado (Chalmers University of Technology, Göteborg, Sweden)
Deep generative models for biomolecular engineering (25 min talk + 15 min Q&A)
12:00 - Tarak Karmakar (Indian Institute of Technology, Delhi, India)
Modelling monolayer protected atomically precise nanoclusters - self-assembly and interactions with biomolecules (15 min talk + 5 min Q&A)
12:20 - Roundtable Discussion (40 min)

13:00 - **Poster Session** & Lunch

14:20 - Group Photo (All participants)

Session 5 (Chair: Ullrich Köthe)

14:30 - Markus Lill (University of Basel, Switzerland)
Fusion of Deep Learning and Molecular Modelling for Drug Design Applications (25 min talk + 15 min Q&A)
15:10 - Christine Peter (Department of Chemistry, University of Konstanz, Germany)
Machine learning in biomolecular simulations: from characterizing conformational free energy landscapes to scale (25 min talk + 15 min Q&A)
15:50 - Jeffrey Vanhuffel (Technische Universität Darmstadt, Germany)
Recruiting Soft Actor-Critic Agents for de-novo discovery of covalent ligands for modulation of transient pockets (15 min talk + 5 min Q&A)

16:10 - Coffee Break

Session 6 (Chair: Ganna (Anya) Gryn'ova)

16:40 - Philip Schwaller (EPFL, Switzerland)
AI-Accelerated Organic Synthesis (25 min talk + 15 min Q&A)
17:20 - Renana Poranne (Technion, Haifa, Israel)
New Representations Enable Interpretable Machine and Deep-Learning for Polycyclic Aromatic Systems (25 min talk + 15 min Q&A)
18:00 - Vivin Vinod (Constructor University, Bremen, Germany)
Multi-Fidelity Machine Learning for Quantum Chemistry (15 min talk + 5 min Q&A)

19:30 - Workshop Dinner Restaurant 'S' Kastanie, Elisabethenweg 1 Heidelberg (leave the Villa Bosch Studio at 19:00)

Thursday, May 4, 2023

Session 7 (Chair: Marcus Elstner)

09:00 - Heather Kulik (Massachusetts Institute of Technology, Boston, USA)
Machine learning tools for discovery in open shell transition metal chemistry (25 min talk + 15 min Q&A)
09:40 - Kai Riedmiller (HITS)
Predicting reaction barriers of hydrogen atom transfer in proteins (15 + 5 min Q&A)
10:00 - Robert Strothmann (Fritz-Haber-Institut, Max-Planck-Gesellschaft, Berlin)
Machine Learning Assisted Photoswitch Design: A Multi-Property Optimization Perspective (15 min talk + 5 min Q&A)

10:20 - Coffee Break

Session 8 (Chair: Pascal Friederich)

10:50 - Mario Barbatti (Aix Marseille University, CNRS, ICR, France)
Nonadiabatic dynamics in the long timescale: the next challenge in computational photochemistry (25 min talk + 15 min Q&A)
11:30 - John Gardner (University of Oxford, UK)
Synthetic Data for Atomistic Machine Learning (15 + 5 min Q&A)
11:50 – Roundtable Discussion and Round up (40 min)

12:30 - Lunch

Talks

Nonadiabatic dynamics in the long timescale: the next challenge in computational photochemistry

Prof. Dr. Mario Barbatti (mario.barbatti@univ-amu.fr)

Aix Marseille University, CNRS, ICR

Nonadiabatic dynamics simulations in the long timescale (longer than 100,000 integration timesteps) are the next challenge in computational photochemistry.[1] In this talk, we will explore the scope of what we expect from methods to run such simulations: they should work in full nuclear dimensionality, be general enough to tackle any molecule, and not require unrealistic computational resources. We will examine the main methodological challenges we should venture into to advance the field, including the computational costs of the electronic structure calculations, stability of the integration methods, the accuracy of the nonadiabatic dynamics algorithms, and software optimization. Based on simulations designed to shed light on these issues, we show how machine learning may be a crucial element for long-timescale dynamics, either as a surrogate for electronic structure calculations or aiding the parameterization of model Hamiltonians. We also show that conventional methods for integrating classical equations should be adequate for extended simulations up to 1 ns and that surface hopping agrees semi-quantitatively with wavepacket propagation in the weak-coupling regime. Finally, we describe our optimization of the Newton-X program to reduce computational overheads in data processing and storage.

[1] Mukherjee et al. Philos Trans R Soc A 2022, 380, 20200382. (DOI: 10.1098/rsta-2020-0382)

The role of coarse-graining in molecular discovery

Dr. Tristan Bereau (t.bereau@uva.nl)

University of Amsterdam

Advanced statistical methods are rapidly impregnating many scientific fields, offering new perspectives on long-standing problems. In materials science, data-driven methods are already bearing fruit in various disciplines, such as hard condensed matter or inorganic chemistry, while comparatively little has happened in soft matter. I will describe how we use multiscale simulations to leverage data-driven methods in soft matter. We aim at establishing structure-property relationships for complex thermodynamic processes across the chemical space of small molecules. Akin to screening experiments, we devise a high-throughput coarse-grained simulation framework. Coarse-graining is an appealing screening strategy for two main reasons: it significantly reduces the size of chemical space and it can suggest a low-dimensional representation of the structure-property relationship. To illustrate these aspects, I will focus on a complex biomolecular system: the selective binding of small molecules to cardiolipin in mitochondrial membranes. A multiscale compound search helps us identify clear design rules for highly selective molecules. It also eases the identification of compounds for experimentation in vitro and in vivo.

Electron and Proton Transfer in Functional Materials from Neural Network Potentials

Prof. Dr. Jochen Blumberger (j.blumberger@ucl.ac.uk)

University College London

Machine learning approaches have changed the way we carry out molecular computations and the field of electron and proton transfer is no exception. I will present two examples where ML approaches have helped us to accelerate and/or improve the accuracy of traditional computational chemistry methods. In the first example we show how committee neural networks (c-NNPs) can be used to accelerate free energy calculations at ab-initio molecular dynamics-level by 3-4 orders of magnitude with little training data and negligible loss in accuracy. Applications to the calculation of pKa values at transition metal oxide/liquid water interfaces, where classical force fields struggle, will be presented. In the second example I will show how the c-NNP approach can be used to estimate electronic Hamiltonians, specifically electronic coupling matrix elements, in a site or diabatic basis, for simulation of charge transport in molecular materials. We find that pure ML methods are inferior to our ``traditional" physics-based methodology but they can be used to improve the latter in a delta-ML scheme. Open questions and challenges will be discussed.

Combining Neural Networks, Enhanced Sampling simulations and FRET technique to study the Structure and Dynamics of Intrinsically Disordered Proteins of Therapeutic Value

Dr. Giorgia Brancolini (giorgia.brancolini@nano.cnr.it)

CNR, Modena, Italy

Intrinsically disordered proteins (IDPs) are abundant in cells and have central roles in protein-protein and protein-ligand interaction. Many are involved in cancer, aging and neurodegenerative diseases. Their structure and dynamics is intimately related to their interactions with binding partners. Because IDPs are inherently flexible and do not have a single conformation, structural ensembles offer more useful representations than individual conformations. The aim of this work is to use Enhanced Molecular Dynamics simulations in conjunction with Neural Network (EncoderMap) and FRET technique to achieve a deeper understanding of the structure and dynamics of a specific IDPs, namely Heat Shock Protein B8 (HSPB8) and its mutant K141E involved in neurodegenerative diseases.1 The workflow is applied to the proteins conformational ensembles in solution and at different salt concentrations to analyze the effect of ionic strength. HSPB8 variants are also studied in the presence of paroxetine, a small molecule found in antidepressant drugs that was shown to have a high affinity for HSPB8 and to partially restore the chaperone activity in the mutated K141E variant. These studies provide the first 3D structural characterization of HSPB8 and reveal the effects of the pathogenic K141E mutation on its conformational ensembles offering the possibility of rationalize it.

Also featured as poster on the stand № 35

Synthetic Data for Atomistic Machine Learning

John Gardner (gardner.john97@gmail.com)

University of Oxford, UK

Chemical structures obtained from MD simulations driven by existing machine-learned potentials are relatively inexpensive to generate as compared to using as compared to first-principles methods. We have created and open-sourced a large (22.9 million atom) and synthetic dataset of carbon structures. In this contribution, I present a series experiments we have performed using this dataset [1]: - We compare the ability of various regression model classes (neural networks, deep kernel learning and gaussian process regression) to predict the synthetic, local energies as labelled by the ML potential. - We use the dataset to explore various hyperparameter settings for GPR in the limit of large amounts of training data. - We investigate the learning of these synthetic energies as a pre-training task for neural-network models. This exhibits positive transfer when learning to predict per-cell energies as labelled by DFT. - We combine supervised and unsupervised learning to guide the creation of chemical maps by extracting hidden layer representations from NN models.

[1] Synthetic data enable experiments in atomistic machine learning, J.L.A. Gardner, Z. Faure Beaulieu and V.L. Deringer, 2022. https://doi.org/10.48550/arXiv.2211.16443.

Discovering chemistry with a transferable reactive machine learning potential

Prof. Olexandr Isayev (olexandr@cmu.edu)

Department of Chemistry, Carnegie Mellon University, Pittsburgh PA

Deep learning is revolutionizing many areas of science and technology, particularly in natural language processing, speech recognition, and computer vision. In this talk, we will provide an overview of the latest developments in machine learning and AI methods and applications to the problem of drug discovery and molecular design at Isayev's Lab at CMU. We identify several areas where existing methods have the potential to accelerate computational chemistry research and disrupt more traditional approaches. Recently, we developed a general reactive ML potential through unbiased active learning with a nanoreactor molecular dynamics-inspired sampler. The resulting potential (ANI-1nr) is then applied to study five distinct condensed-phase reactive chemistry problems: carbon solid-phase nucleation, graphene ring formation from acetylene, biofuel additives, combustion of methane, and the spontaneous formation of glycine from early earth small molecules. In all studies, ANI-1nr closely matches experiments and/or previous studies using traditional model chemistry methods without needing to be refit for each application, which enables high-throughput in silico reactive chemistry experimentation.

Machine-Enabled Chemical Structure-Property-Synthesizability Predictions

Prof. Yousung Jung (yousung@gmail.com)

Seoul National University

This Discovery of new molecules and materials with desired properties is a practical goal of chemical research. A promising way to significantly accelerate the latter process is to incorporate all available knowledge and data to plan the synthesis of the next materials. In this talk, I will present several directions to use informatics and machine learning to efficiently explore chemical space. I will first describe methods of machine learning for fast and reliable predictions of materials and molecular properties. With these tools in place for property evaluation, I will then present a few generative frameworks that we have recently developed to allow the inverse design of molecules and materials with optimal target properties, either in the compositional space or structural space. One general challenge in digital discovery is that many of the molecules and materials that are computationally designed are often discarded in the laboratories since they are not synthesizable. I will thus lastly spend some time to talk about the synthesizability of molecules and materials, either by predicting the synthesis pathways (retrosynthesis) or chemical reactivity. Several challenges and opportunities that lie ahead for further developments of accelerated chemical platform will be discussed.

Modelling monolayer protected atomically precise nanoclusters - self-assembly and interactions with biomolecules

Prof. Tarak Karmakar (tkarmakar@iitd.ac.in)

Indian Institute of Technology, Delhi

Monolayer ligand Protected atomically precise nano-Clusters (MPCs) have gained enormous popularity due to their unique properties and versatile applications in bio-imaging, as sensors and effective drug carriers.[1] Understanding the structure and dynamics of these molecule-like nanoclusters is of paramount interest in designing new MPCs, tuning their physical properties, and expanding their applicability in diverse areas.[2] Using molecular dynamics simulations augmented with enhanced sampling methods, we have investigated the self-assembly of MPCs in solutions and the dynamics of MPCs in solid phases.3 Furthermore, to explore the capability of MPCs as drug carriers, we studied their interactions with an anticancerous peptide-based drug, Melittin, and subsequently, we investigated the permeation process of a MEL-bound MPC through a tumor cell membrane.[4,5] In my presentation, I will discuss results related to these aspects of MPCs and their bio-applications. Although all simulation models in our study so far are described atomistically, our future works would be on developing accurate coarse-grained models for MPCs, which would allow us to increase the system size and sample the long timescale processes.

References: (1) Jin, R.; Zeng, C.; Zhou, M.; Chen, Y. Atomically Precise Colloidal Metal Nanoclusters and Nanoparticles: Fundamentals and Opportunities. Chemical Reviews 2016, 116, 10346–10413. (2) Malola, S.; H ¨akkinen, H. Prospects and challenges for computer simulations of monolayer-protected metal clusters. Nature Communications 2021, 12 (3) Tiwari, V; Karmakar, T. Understanding Molecular Aggregation of Ligand-protected Atomically-Precise Metal Nanoclusters, 2023 (submitted) (4) Tiwari, V; Garg, S.; Karmakar, T. Insights into the Interactions of Peptides with Monolayer-Protected Metal Nanoclusters, ACS Applied Biomaterials, 2023 (5) Tiwari, V; Garg, S.; Karmakar, T. Monolayer-Protected Metal Nanoclusters as Effective Drug Carriers, 2023 (in preparation)

Machine learning tools for discovery in open shell transition metal chemistry

Prof. Heather Kulik (hjkulik@mit.edu)

Massachusetts Institute of Technology

I will discuss our efforts to use machine learning (ML) to accelerate the computational tailoring and design of transition metal complexes and metal-organic framework (MOF) materials for catalysis. One limitation in a challenging materials space such as open shell, 3d transition metal chemistry is that ML models and ML-accelerated high-throughput screening traditionally rely on density functional theory (DFT) for data generation, but DFT is both computationally demanding and prone to errors that limit its accuracy in predicting new open shell transition metal complexes. I will describe how we have leveraged consensus in DFT screening as well as gone beyond DFT by developing a "recommender" that achieves higher level accuracy to overcome some of these challenges. I will also describe how we have paired these ML models with efficient global optimization to accelerate searches for candidate materials by 1000 fold. Finally, time permitting, I will discuss our efforts in direct learning of experimental data as a way to bypass limitations of simulation in the prediction of metal-organic framework stability.

Fusion of Deep Learning and Molecular Modelling for Drug Design Applications

Prof. Dr. Markus Lill (markus.lill@unibas.ch)

University of Basel

Deep learning has started to play a significant role in many scientific fields. In drug discovery, deep learning will have an increasing impact in the near future, showcased by the use of recently pioneered deep neural network approaches for protein-structure prediction, synthesis prediction, and de novo molecular design. In this presentation, I will delve into our efforts to integrate physicochemical models and our current understanding of protein-ligand interactions with deep neural network techniques. A novel multiscale approaches for molecular docking will be discussed. Using deep learning in the pose-generation phase makes time-consuming sequential search algorithms obsolete. Initial coarse-graining of the protein binding site with full-atomistic reconstruction of the docked complex facilitates the incorporation of protein flexibility. This talk will also explore the development of novel free energy calculation methods. A combination of targeted free energy perturbation theory with normalizing flows overcomes the time-consuming stratification process of standard free energy calculations. Incorporation of dummy atoms extends those ideas to the relative free energy calculation of molecules with different number of atoms.

Deep generative models for biomolecular engineering

Prof. Rocío Mercado (rocom@chalmers.se)

Chalmers University of Technology, Gothenburg

AI is transforming our approach to biomolecular engineering. This includes the development of generative and predictive tools that can learn from biochemical data, such as molecular structures, chemical reactions, and biomedical data. While AI can be applied to a range of molecular engineering tasks, one ideal area is de novo molecular design. De novo design is the concept of designing molecules with desired properties from scratch to minimize experimental screening and is poised to allow scientists to traverse chemical space more efficiently in search of optimal molecules by delegating error-prone decisions to computers. In drug development, de novo design methods can aid medicinal chemists in the design and selection of drug candidates, with the added advantage that they can learn from datasets of billions of molecules in minutes and be constantly updated with new data. Deep molecular generative models are a particular approach to de novo design; they use deep neural networks to generate new molecules in silico, and work by proposing node and edge modifications to an initial graph structure to generate compounds predicted to achieve a specific property profile. Such models can be applied to the engineering of a range of therapeutic modalities, from small molecules to proteins. In this talk, I will discuss the development of deep generative models for various bimolecular engineering tasks relevant to early-stage drug discovery. These include: a generative model for synthesizability-constrained molecular design, a reinforcement learning framework for molecular graph optimization, and recent applications from our group to the design of large modalities for targeted protein degradation.

Buffer region neural networks: A buffered scheme for polarizable QM/MM simulations with machine-learning

Prof. Dr. Chris Oostenbrink (chris.oostenbrink@boku.ac.at)

University of Natural Resources and Life Sciences, Vienna (BOKU)

In hybrid quantum mechanics / molecular mechanics (QM/MM) approaches, the molecular system is partitioned into regions that are treated at different levels of theory. At the interfaces between these regions, artifacts may occur. We have recently introduced a buffered embedding scheme, in which a buffer region between the inner (QM) and outer (MM) region is defined for which the interactions are computed both at the QM and MM level. This comes at the cost of introducing a second QM-calculation at every timestep of the simulation. The use of neural networks to describe molecular potential energies, allows for an elegant solution to this problem. We train a neural network directly on the difference between the two QM calculations, ensuring that the network reproduces the QM-interactions of the inner region, with itself and with the buffer region as well as the polarization of the buffer region due to the inner region. Any remaining artifacts largely cancel in the trained differences and are far removed from the inner region of interest. The use of the Buffer Region Neural Network (BuRNN) approach, furthermore, allows us to apply alchemical free-energy calculations at the QM-level of theory.

Machine learning in biomolecular simulations: from characterizing conformational free energy landscapes to scale bridging

Prof. Dr. Christine Peter (christine.peter@uni-konstanz.de)

Department of Chemistry, University of Konstanz

Enhanced sampling methods, multiscale approaches, and improved simulation models in combination with ever growing computational power have given us access to unprecedented system sizes and simulation times and have led to a massive increase in the amount of simulation data being produced. Thus, processing and analyzing exceedingly large high-dimensional data sets has become one of the major challenges. I will show how multiscale approaches in combination with advanced analysis methods can be used to investigate and characterize the structural variability of biomolecular systems, in particular multidomain proteins and protein conjugates. Modern machine learning approaches are utilized to identify, compare, and classify relevant conformational states, to provide insights into the decisive features hidden in these high dimensional simulation data and to guide their interpretation with respect to experiments. Using efficient dimensionality reduction techniques we obtain low dimensional representations of the sampling which can be interpreted as conformational free energy landscapes. These low dimensional representations enable us to assess the consistency of the sampling in different models, to go back and forth between simulation scales or compare the conformational behavior of different systems.

New Representations Enable Interpretable Machine and Deep-Learning for Polycyclic Aromatic Systems

Prof. Renana Poranne (rporanne@technion.ac.il)

Technion, Haifa, Israel

Polybenzenoid hydrocarbons (PBHs) – molecules made up of multiple benzene rings – are the quintessential polycyclic aromatic species. In addition to their importance for a variety of functionalities, these molecules serve as model systems for the much larger and more heterogeneous space of polycyclic aromatic systems and provide the opportunity to investigate the effect of annulation geometry on different molecular properties. The structure-property relationships of PBHs have both conceptual and practical implications; understanding them can enable design of new functional compounds and elucidation of reactivity in a broader context. We interrogated these compounds using a combination of traditional computational techniques, including characterization of their aromatic character in the S0 and T1 states (described with the NICS metric), their spin density in the T1 state, and their S0—T1 energy gaps. Regularities were revealed that allowed for simple and intuitive design guidelines to be defined.1 To verify these guidelines in a data-driven manner, we generated a new database – the COMPAS Project2– which contains the calculated structures and properties of all PBHs consisting of up to 11 rings. Further, we developed and implemented two types of molecular representation to enable machine- and deep-learning models to train on the new data: a) a text-based representation3 and b) a graph-based representation.4 In addition to their predictive ability, we demonstrate the interpretability of the models that is achieved when using these representations. The extracted insight in some cases confirms well-known “rules of thumb” and in other cases disproves common wisdom and sheds new light on this classical family of compounds. In addition to corroborating domain-experts’ interpretation, the different models also highlight additional relationships that are harder for the human eye to discern.

Predicting reaction barriers of hydrogen atom transfer in proteins

Kai Riedmiller (kai.riedmiller@h-its.org)

HITS

Hydrogen atom transfer (HAT) reactions are important reactions in many biological systems. As these reactions are hard to observe experimentally, it is of high interest to shed light on them using simulations. Here, we present a machine learning model for the prediction of activation energies of HAT reactions. As the inference speed is high, this model enables evaluations of many chemical situations in rapid succession. It is trained on energy barriers calculated using hybrid density functional theory. We built and evaluated the model in the context of HAT in Collagen, but the same workflow can also be applied to HAT reactions in other biological or synthetic polymers. The access to fast predictions of HAT energy barriers, when combined with molecular dynamics in a kinetic Monte-Carlo scheme, paves the way towards reactive simulations.

Also featured as poster on the stand № 28

Learning Physical Interactions for Molecular Dynamics Simulations

Prof. Dr. Sereina Riniker (sriniker@ethz.ch)

ETH Zürich

From simple clustering techniques to sophisticated neural networks, the use of machine learning has become a valuable tool in many fields of chemistry in the past decades. Here, we describe different ways in which we explore the use of machine learning (ML) for predict physical interactions between particles in molecular dynamics (MD) simulations in order to improve their accuracy. In classical MD simulations, the physical interactions between atoms are described with an empirical force field. This involves a large number of parameters for each molecule, which are fitted to quantum-mechanical (QM) or available experimental data. There is a need for more accurate and general force fields. In this context, we demonstrate how ML approaches can aid in force-field development, from multipole prediction to generalized parametrization. In the second part, we explore the use of ML for increasing the speed and accuracy of QM/MM MD simulations. Concepts such as Δ-learning and different network architectures are explored.

AI-Accelerated Organic Synthesis

Prof. Dr. Philippe Schwaller (philippe.schwaller@epfl.ch)

EPFL, Lausanne, Switzerland

In organic chemistry, we are currently witnessing a rise in artificial intelligence (AI) approaches, which show great potential for improving molecular designs, facilitating synthesis and accelerating the discovery of novel molecules. Based on an analogy between written language and organic chemistry, we built linguistics-inspired Transformer neural network models for chemical reaction prediction, synthesis planning, and the prediction of experimental actions. We extended the models to chemical reaction classification and fingerprints. By finding a mapping from discrete reactions to feature vectors, we enabled efficient chemical reaction space exploration. Intrigued by the remarkable performance of chemical language models, we discovered that the models capture how atoms rearrange during a reaction, without supervision or human labeling, leading to the development of the open-source atom-mapping tool RXNMapper (http://rxnmapper.ai/). During my talk, I will provide an overview of the different contributions that are at the base of this digital synthetic chemistry revolution.

Correlation-based feature selection to identify functional dynamics in proteins

Prof. Dr. Gerhard Stock (stock@physik.uni-freiburg.de)

Institute of Physics, Uni Freiburg

The statistical analysis of molecular dynamics simulations requires dimensionality reduction techniques, which yield a low-dimensional set of collective variables that in some sense describe the essential dynamics of the system. A crucial first step of such an analysis is the identification of suitable input coordinates or 'features', such as backbone dihedral angles and interresidue distances. To discriminate collective motions underlying functional dynamics from uncorrelated motions, the correlation matrix of the input coordinates is block-diagonalized using the Leiden community detection algorithm by a clustering method. This strategy avoids possible bias due to presumed functional observables and conformational states or variation principles that maximize variance or timescales. Applications include the functional motion of T4 lysozyme to demonstrate the successful identification of collective motion, the folding of villin headpiece to show how correlated motions elucidate the folding mechanism, and the allosteric communication in PDZ3 domain which is achieved by a fluctuating and cooperative network of interresidue contacts.

Machine Learning Assisted Photoswitch Design: A Multi-Property Optimization Perspective

Robert Strothmann (strothmann@fhi-berlin.mpg.de)

Fritz-Haber-Institut, MPG, Berlin

The sheer vastness of chemical spaces poses a daunting challenge to molecular discovery through high-throughput screening based on exhaustive sampling. Generative models (GMs) are an emerging machine learning (ML) approach that enables a more guided discovery. Implicitly learning chemical design rules from large reference data sets and suitable descriptors of a targeted functionality, GMs directly propose promising, yet diverse candidates. Here we explore the use of GMs for the design of novel molecular photoswitches. This class of molecules represents a highly challenging design task, since the switching mechanism depends on different properties of the electronic ground and excited state. A balance between partly competing design goals is desired. In a first step, large general molecular databases are used to train a GM to generate chemically valid photoswitches. In a second step, the creation process needs to be conditioned towards performant switching capabilities. In the absence of sufficient corresponding experimental reference data, this conditioning is based on synthetic first-principles data in an iterative distrubtion learning workflow. For that purpose computationally efficient descriptors are used in a multi-objective fashion to account for the desired key aspects of the switching process

Recruiting Soft Actor-Critic Agents for de-novo discovery of covalent ligands for modulation of transient pockets

Jeffrey Vanhuffel (jeffrey.vanhuffel@tu-darmstadt.de)

Technische Universität Darmstadt

A protein's activity depends on its different, possible conformational states/ensembles between which it may or may not switch over the course of its biological pathway. Some (main) states in this ensemble are more prevalent than other (transient) conformations but any may cause pathologies. All conformations can have one or more binding pockets that may be unique to that specific state that can be targeted by (non-)covalent small molecules in order to combat the corresponding illnesses. Reinforcement learning models have already been successfully employed in de-novo discovery of non-covalent ligands for known protein conformations. This talk will explore combining Molecular Dynamics simulations and a Deep Reinforcement Learning model, that generates de-novo COVALENT ligands as a first-of-its-kind, for discovery and targeting of UNKNOWN transient binding pockets.

Multi-Fidelity Machine Learning for Quantum Chemistry

Vivin Vinod (v.vinod@jacobs-university.de)

Constructor University Bremen

There has been a great deal of progress in machine learning methods for quantum chemistry but there still exists the challenge to the cost of generating the training dataset. One cost reduction method is the use of multi-fidelity machine learning. By using data from multiple accuracy or fidelity of quantum chemistry calculations, it is possible to create sub-models which deliver higher accuracy than the conventional single fidelity models. The method delivers low-cost high-accuracy models for quantum chemistry properties. For this specific example, it is implemented for the first excited state energies.

Electrostatic embedding of machine Learning potentials

Dr. Kirill Zinovjev (kirill.zinovjev@uv.es)

Universidad de Valencia, Spain

We present a version of electrostatic embedding scheme that allows to combine arbitrary ML potentials learned in vacuo with molecular mechanics forcefields. This allows to accelerate the state-of-the-art QM/MM simulations, such as those used to simulate enzymatic catalysis, by replacing the QM method with a cheaper ML potential. The scheme relies on physically motivated models of molecular electrostatics and polarizability allowing to use only a handful of reference atomic environments for learning. We validate the scheme by predicting single point embedding energies for SARS-CoV-2 protease complex with PF-00835231 inhibitor, resulting in a predicted embedding energy RMSE of 2 kcal/mol, compared to explicit DFT/MM calculations. Zinovjev K. Electrostatic Embedding of Machine Learning Potentials. J. Chem. Theory Comput. 2023, https://doi.org/10.1021/acs.jctc.2c00914

Posters

Data-driven identification and analysis of the glass transition in polymer melts

Atreyee Banerjee (banerjeea@mpip-mainz.mpg.de)

Max Planck Institute for Polymer Research. Mainz

On cooling, the dynamical properties of many polymer melts slow down exponentially, leading to a glassy state without any drastic change in static structure. We propose a data-driven approach, which utilises the high-resolution details accessible through the molecular dynamics simulation and considers the structural information of individual chains. It clearly identifies the glass transition temperature of polymer melts of semiflexible chains. By combining principal component analysis (PCA) and clustering, we identify glass transition temperature at the asymptotic limit even from relatively short-time trajectories, which just reach into the Rouse-like monomerdisplacement regime. We demonstrate that fluctuations captured by the principal component analysis reflect the change in a chain's behaviour: from conformational rearrangement above to small vibrations below the glass transition temperature. We demonstrate the generality of the approach by using different dimensionality reduction and clustering approaches. The method can be applied to a wide range of systems with microscopic/atomistic information. More recently we applied this methodology to all-atom acrylic paint systems. Our study reveals the explicit role of backbone and side chain residues to determine the glass transition temperature.

Stand № 1

Reduction pathway of glutaredoxin 1 investigated with QM/MM moleculardynamics using a neural network correction

Julian Boeser (julian.boeser@kit.edu)

Karlsruhe Institute of Technology (KIT)

Glutaredoxins are small enzymes that catalyze the oxidation and reduction of protein disulfide bonds by the thiol–disulfide exchange mechanism. The exact mechanisms are not yet fully known. In our study, we investigated a proposed mechanism for the reduction of the disulfide bond in the protein HMA4n by a mutated monothiol Homo sapiens glutaredoxin and the co-substrate glutathione. The free energy profile of each reaction was obtained with hybrid quantum mechanical/molecular mechanical metadynamics simulations. For an accurate description, we used semi-empirical density functional tight-binding method with specific reaction parameters fitted to B3LYP energies of the thiol–disulfide exchange. In addition we applied a machine learned energy correction that was trained on coupled-cluster single double perturbative triple [CCSD(T)]energies of thiol–disulfide exchanges. The computational cost of the ML correction is comparable to a DFTB calculation, but offers a potential for higher accuracy and greater flexibility for as somewhat increased computational cost. Due to the extensive phase space sampling, this approach includes environmental effects and the ML correction allows to describe correlation effects relevant for the thiol–disulfide exchange reaction, which most DFT-GGA functionals do not capture.

Stand № 2

Molecular simulations of H1R-ligand binding kinetics

Mislav Brajković (mislav.brajkovic@h-its.org)

Heidelberg Institute for Theoretical Studies (HITS)

Molecular binding kinetic parameters are often essential determinants of the function of molecules such as cytokines and hormones, as well of the efficacy of drugs in the non-equilibrium conditions of living organisms. Because of this, many different *in silico* methods are being developed to compute the molecular association (k_{on}) and dissociation (k_{off}) rate constants. Knowing the dissociation rate constant allows for the computation of a drug residence time ($\tau=1/k_{off}$) which is an important factor in drug design because it gives insight on how long will a particular drug exert its effect. A successful computational approach for computing the relative molecular residence times is the molecular dynamics-based τRAMD method. We used the τRAMD method to simulate the dissociation process and to compute the relative residence times (dissociation rates) of different antagonists of the histamine-1-receptor (H_1R) which is a validated target for the treatment of allergies and some forms of gastric acid related conditions. Relative residence times of compounds calculated with τRAMD show good correlation with the experimental values. The results demonstrate that τRAMD is adequate and accurate method for prediction of future drug candidates.

Stand № 3

KIMMDY 2.0 -- A Kinetic Monte Carlo Reactive Molecular Dynamics Framework

Jannik Buhr (jannik.buhr@h-its.org)

HITS

Forcefield based molecular dynamics simulations allowed us to reach biologically relevant timescales and system sizes. A fundamental limit of this molecular mechanics approach is a lack of reactivity. We present a framework for combining classical molecular dynamics simulations with a kinetic Monte Carlo approach to bridge timescales and allow reactions to occur within a simulation. It is implemented as a user-friendly, extensible python module based on the open-source high-performance molecular dynamics software suit GROMACS. This poster focusses on making the necessary changes to topologies and forcefield parameters in a modular way.

Stand № 4

Discovering Long-term Polymer Dynamics with Generative Deep Neural Networks

Gian-Michele Cherchi (g.m.cherchi@gmail.com)

Simatlab, University Clermont-Auvergne

When simulating soft-matter systems, polymer melts can be challenging because complicated sub-diffusive patterns arise when increasing the chain length; nevertheless, these regimes are transitory, and asymptotically one observes normal diffusion. Reaching these timescales remains expensive and one of the solutions often adopted is coarse-graining. In this work, the objective is using short-term non-markovian molecular dynamics trajectories to extrapolate long-term diffusion behaviour. We employ generative a Machine learning techniques to model the conditional distribution of single polymers' normal modes, in an omopolymer system. This gives one the chance to train a neural network with an additional Score-Based SDEs regularization, which enhances its learning capabilities. Centre of mass dynamics, displaying transient anomalous diffusion, is then modelled with a Generalized Langevin Equation having an integrable kernel, which, in the zero-mass limit, yields a solution corresponding to a set of stochastic differential equations involving single polymer normal modes. Future research perspectives include evaluating the generalization capabilities of the model conditional to different temperatures or chain lengths. An open problem remains understanding the effect of the score-based regularization and exploiting it to generate a higher fidelity dynamics.

Stand № 5

COMPUTATION OF UNBINDING RATES AND MECHANISMS IN PROTEIN-PROTEIN SYSTEMS

Giulia D'Arrigo[1], Daria B. Kokh[1], Ariane Nunes-Alves[1,2], Rebecca C. Wade[1,3]
(giulia.darrigo@h-its.org)

[1]Heidelberg Institute for Theoretical Studies, [2]Technical University of Berlin, [3]Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, and Interdisciplinary Center for Scientific Computing (IWR)

The dissociation rate, or its reciprocal, residence time, is crucial for determining the duration and the biological effect of biomolecular interactions. Its prediction is key for better understanding protein-protein interactions (PPIs), drug targets for numerous diseases, and for the design of high-affinity modulators. However, the conventional molecular dynamics simulation approaches are limited to a short timescale hampering the estimation of the residence time (typically ranging from minutes to hours). After the extensive application on protein-small molecule systems, the use of τ-RAMD (Random Acceleration Molecular Dynamics) for estimating protein-protein dissociation rates is here presented. With τ-RAMD, the unbinding event is observed within the nanosecond timescale thus enabling the fast computation of the relative residence time. We have assessed the methodology on a diverse set of protein-protein systems showing good agreement between the computed and experimental data. In addition, the combination of τ-RAMD with the MD-IFP (Interaction Fingerprint) analysis allowed the investigation of the dissociation process easing the detection of the molecular hot-spots to target to modulate specific interactions. Our results demonstrate the applicability of τ-RAMD for computing protein-protein dissociation rates and for being a valuable tool for modelling biomolecular kinetics, as well as for assisting in the design of PPIs modulators.

Stand № 6

High accuracy neural network predictions for sulfur and phosphor reactions in solution

Lena Eichinger, Christian Schmidt (lena.eichinger@kit.edu) ,(christian.schmidt@kit.edu)

Karlsruhe Institute of Technology (KIT)

Saving computational cost on QM/MM calculations without losing accuracy is made possible by the use of neural networks (NN). Highly polarizable atoms like sulfur or phosphor, due to their ability to access d-orbitals, pose a challenge for established NN. Our approach aims towards describing reactions like thiol-disulfide exchange reactions and phosphorylation in solution with high accuracy while maintaining minimal computational cost.

Stand № 7

Prediction of Fluorophore Excitation Energies in Complex Surroundings

Manuel Enns (manuel.enns@student.kit.edu)

KIT

Certain dyes show promising sensitivities to their surroundings resulting in a significant fluorescence shift. In order to study this shift we made a neural network capable of predicting the emission spectrum of a fluorophore given the geometry and the electrostatic potential. The model is trained in several solvents but should be able to work in different surroundings too. Therefore the model is tested by looking at a mutated glucose binding protein that undergoes a conformational change upon binding. Such a system could be used as a glucose measuring device.

Stand № 8

Exploration of redox properties in chemical space

Rostislav Fedorov (rostislav.fedorov@h-its.org)

Heidelberg Institute for Theoretical Studies (HITS)

Redox potential plays a crucial role in many applications, and accurately estimating it can be time-consuming and resource-intensive. In this study, we present a novel method for fast estimation of redox potential using message passing neural networks (MPNN). By training on an OMEAD dataset, we achieved the lowest mean absolute error (MAE) among existing approaches, reported in the literature, making our method state-of-the-art. Furthermore, we combined our MPNN approach with an evolutionary algorithm to explore the vast chemical space for potential good candidates. Our method has the potential to greatly accelerate the discovery of new catalysts and materials for redox reactions, ultimately contributing to the development of more efficient and sustainable chemical processes.

Stand № 9

Structural prediction of transmembrane peptide-protein interaction and application to a potential peptide therapeutic

Manuel Glaser [1,5], Michael Egger[2,3,5], Lukas Jarosch[1], Rafael Salazar[1], Tommaso Bartoloni[1,5], Julia Ritterhoff[2,3,5], Patrick Most[2,3,5], Rebecca C. Wade[1,4,5] (manuel.glaser@h-its.org) (tommaso.bartoloni@h-its.org)

[1]Heidelberg Institute for Theoretical Studies (HITS); [2]Division of Molecular and Translational Cardiology, Department of Medicine III, Heidelberg University Hospital; [3]German Center for Cardiovascular Research (DZHK); [4]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance and Interdisciplinary Center for Scientific Computing (IWR), [5]Informatics for Life (I4L) consortium, Klaus Tschira Foundation

Modulating membrane-embedded protein-protein interactions is an attractive point of drug application. Designing transmembrane-spanning peptides that can regulate transmembrane domain interactions and thus influence the function of membrane proteins, such as ATPase pumps, therefore holds promise for therapeutic applications. Efficient structure prediction of these interactions could support the design process, however, it is complicated by the presence of both an aqueous and a lipid bilayer environment in the respective systems. Correspondingly, there is no out-of-the-box approach to tackle the docking of potentially membrane-insertable peptides to transmembrane proteins. We addressed this docking problem by customizing a pipeline that employed global rigid-body docking (ClusPro) followed by semi-flexible refinement in a membrane environment (Rosetta MPDock), which we validated on experimentally determined complexes of the transmembrane calcium ATPase SERCA bound to transmembrane miniprotein regulators. We also applied the pipeline to generate models of the peptide drug candidate S100A1ct in complex with its target SERCA2a. S100A1ct is a peptide derived from the C-terminal helix of the calcium sensor protein S100A1 and conveys positive effects in cardiomyocytes, e.g., by increasing SERCA2a activity. Based on our results, we hypothesize that its mechanism of action could be explained by perturbating the binding of the SERCA miniprotein inhibitor phospholamban.

Stand № 10

Automated Parameterization for Reactive Molecular Dynamics Simulations

Eric Hartmann (eric.hartmann@h-its.org)

HITS

In molecular dynamics (MD), classical force fields have enabled remarkable insights into a wide range of molecular systems. In these force fields, atoms are assigned an atom type, and parameters are inherited from this atom type for the duration of the simulation. However, reactions or other events may lead to a change in the chemical environment of an atom, necessitating changes to its parameters. Several methods have been established to deal with changes in an atom's environment. Parameters can be changed dynamically or forces can be evaluated using an entirely different potential, for example neural network potentials. Here, we extend the recently developed hybrid Kinetic Monte Carlo/ MD scheme KIMMDY to automatically reparameterize the molecular system in between simulations. To achieve this, changes to the connectivity are detected by direct chemical perception, parameter changes are applied to the simulation files and a smooth transition scheme between the parameters is employed. This approach is extensible to further reaction mechanisms which can be added to KIMMDY by supplying a function (e.g. a neural network) that determines the rate for a reaction given a system configuration. Toy peptide systems are used to demonstrate the simulation of several consecutive reactions without intervention while maintaining highly accurate parameters. Thus, the impact of a given reaction on the relative probability of the ones that follow it can be studied. One application case is the study of mechanoradical migration pathways in load-bearing proteins like collagen.

Stand № 11

Exciton transfer simulations in light harvesting complexes accelerated by machine learning

David Hoffmann (david.hoffmann@kit.edu)

KIT, Institut für Physikalische Chemie, Theoretische Chemische Biologie

Nature has developed highly efficient photosynthetic units in the course of evolution. Light-harvesting (LH) complexes are responsible for collecting and transmitting the energy of sunlight in the form of excitons. Photoactive pigments are arranged within a protein framework, that ensures the specific alignment of the pigments leading to optimized energy transfer. Non-adiabatic molecular dynamics (NAMD) methods, such as trajectory surface hopping, can be used to simulate the transfer of excitons between different pigments. The motion of excitons results from the coupling of nuclear and electronic degrees of freedom. Due to the size of biological LH complexes, such simulations are extremely challenging. We aim to integrate machine learning techniques to replace costly quantum-chemical calculations. Here, we present simulations of exciton transfer in the light-harvesting complex II (LH2) of purple bacteria, which are examined in terms of underlying transfer mechanisms, (de)localization and transfer timescales. Neural network models are trained for the prediction of transfer Hamiltonian elements using reference data from semi-empirical time-dependent long-range corrected density functional tight binding (TD-LC-DFTB). For the prediction of excitation energies, the models take into account the specific environment in the form of the electrostatic potential induced on the individual atoms of the pigment molecules.

Stand № 12

Scalable Machine Learning for Open System Quantum Dynamics

Yannick Holtkamp (yholtkamp@constructor.university)

Constructor University Bremen

To understand many processes in nature involving, for example, energy or electron transfer, the theory of open quantum systems needs to be involved. A prime example in this direction is the transfer of energy from an absorbed photon to the reaction center within the framework of photosynthesis. For larger systems, an accurate description of the underlying quantum dynamics is still a formidable task and, hence, approaches employing machine learning techniques have been developed and tested to reduce the computational effort of accurate dissipative dynamics. A downside of previous machine learning methods is that they require numerical expensive training data for systems of the same size as the ones they will be employed for, making them unfeasible to use for larger systems where those calculations are still too expensive. Here, we will introduce a new method that is implemented as a machine-learned correction term to the so-called Numerical Integration of Schrödinger Equation (NISE). We will show that this term can be trained on data from small systems, where the computationally expensive methods are still feasible to compute. Then the NISE with the new machine learned correction can be used to determine the dissipative quantum dynamics for larger systems.

Stand № 13

Candidate free active learning for exited state energies

Matthias Holzenkamp (mholzenkamp@constructor.university)

Constructor University Bremen

To substitute the high number of expensive ab initio calculations to build a potential energy surface, machine learning models which are trained only on a few samples can be used. Active learning strategies allow choosing the training samples in a way as few samples as possible are needed to reach a certain accuracy. As Gaussian process regression does not only provide the posterior mean for the prediction but also the posterior variance, this can be used to find geometries with high uncertainty. We follow an approach where new training samples are found via maximization of the variance, weighted with a function that restricts the configuration space, in which to search, in a certain way.

Stand № 14

Anharmonic Correction to the Adsorption Free Energy by Machine Learned Force Field-based Thermodynamic Integration

Thanh-Nam Huynh (thanh-nam.huynh@kit.edu)

KIT

An accurate description of adsorption process is of paramount importance for understanding heterogeneous catalytic process. However, the current approaches of adsorption free energies calculations either provide insufficient entropic contributions or are very complex and/or computationally expensive. The harmonic approximation (HA), the most frequently used method thus far, is prone to significant deviations [1]. There is need in method that is able to recover the free energy contributed from a system's anharmonicity for a better description of free energy. Herewith, we present a new approach, in which machine learned force field (MLFF) is used in place of DFT in the most time-consuming step of λ-path thermodynamic integration (MLFF-based λ-TI) to calculate the anharmonic contribution to free energy. The validation test on ethane system shows excellent agreement with DFT λ-TI and semi-analytic results. The approach is performed on an adsorbing system of interest, namely OH@Pt(111). The computed anharmonic correction for this system gives a significant value of -0.250 eV, as high as roughly 12.5% of the harmonically approximated adsorption free energy. Therefore, the MLFF-based λ-TI method could be promising to reach accurate free energies of adsorption processes, which could then provide more insight into the reaction mechanism of surface reactions.

Stand № 15

Fragmentation-based molecular representation suitable for ML/DL applications

Stiv Llenga (stiv.llenga@h-its.org)

Heidelberg Institute for Theoretical Studies (HITS)

The chemical space of molecules is infinite, but chemists' ability to study every molecule in the universe is limited. The goal of machine learning (ML) in chemistry is to find patterns in this infinite space, discover new compounds, identify which compound or class of compounds has a specific property, is stable under certain conditions, etc. The manner in which the chemical space is built is crucial to accomplish these tasks and in turn depends strongly on molecular representation. The matrix of fragment similarity representation (MFSR) is a new ML-ready fragmentation-based technique for mapping the chemical space of compounds composed of specific building blocks. Most industrially and biologically relevant macromolecules are formed as a combination of finite building blocks (e.g., all proteins are a combination of just 20 aminoacids), and our technique can predict their properties in less than a fraction of a second and with the quantum-chemcial accuracy. In this study, MFSR is applied to two datasets of N-heteropolycycles, N-HPC1 and N-HPC1x, to predict, analyse, and rationalise their properties using unsupervised and supervised deep learning techniques. In contrast to other molecular representations, MFSR allows even the most entangled deep learning models to be decodable in a form that chemists can easily understand. This is because the input to an ML model used in MFSR is simply the similarity of a specific building block to the parent compounds multiplied by the property of the building block.

Stand № 16

Simulation at molecular and mesoscopic level and its validation through X-ray and neutron scattering

Arnab Majumdar (arnab.majumdar@hereon.de)

Helmholtz Zentrum Hereon, Garching, Germany

With the advent of new technologies, it has been possible to design new materials through theoretical analysis and simulation. However, these simulations have to be validated by experiments. In this work, we develop a method to validate the structural features in continuum simulations directly on the mesoscopic level with X-ray and neutron scattering experiments. The workflow to compare atomistic computer simulations to scattering patterns is well established: the scattering amplitude of individual atoms is summed, whose positions are obtained from the simulation. This approach fails for larger mesoscopic structures due to the unrealistic computation time required to generate the simulations and scattering pattern on a mesoscopic scale with atomic resolution. We developed a methodology that calculates scattering patterns from a continuum simulation like phase-field modeling, where the material description is continuous instead of a collection of atoms. The approach is validated with simple structures and gradually applied to more complex structures. The long-term goal is to use this technique for the simulation of hydrogen storage materials and validation of the simulations with scattering data.

Stand № 17

What does the biosynthetic gene cluster say? Understanding biosynthetic gene clusters with protein language models

Tatiana Malygina (merlettaia@gmail.com)

Helmholtz-Institute for Pharmaceutical Research Saarland, Saarbrücken

Many organisms, such as bacteria, fungi, and plants, produce intricate chemicals that are not needed for their growth and reproduction, and thus are called secondary metabolites or natural products (NPs). NPs are a rich source of drugs, with most antibiotics being derivatives of NPs. In a producer organism, NPs are synthesized by a set of enzymes encoded by genes that often lie near each other on the chromosome and are called a biosynthetic gene cluster (BGC). Despite the clinical importance that some NPs have, only a small number of naturally-occuring BGCs are explicitly described. From the natural language processing (NLP) point of view, in terms of the number of samples, the existing collections of BGC sequences can be considered a low-resource language corpus. A common approach to tackle such datasets is to transfer an existing model trained on a high-resource language dataset from one language domain to another using transfer learning. A natural high-resource dataset for BGCs would be all protein sequences. Several different protein language models (pLM) trained with large collections of sequences are available nowadays. In this work, we use them employing transfer learning to explore the meaningfulness of representations of BGCs regarding the chemistry of expressed NPs.

Stand № 18

Manifold Learning for Boltzmann Generators

Marcel Meyer (marcel.meyer@h-its.org)

Heidelberg Institute for Theoretical Studies

Molecular dynamics are used to sample equilibrium states but require large computational resources for complex systems. Generative models, such as Boltzmann generators promise to accelerate sampling, but creating generative models of sufficient quality remains challenging. Domain knowledge tells us that reasonable data points have to lie on a manifold of much lower dimension than the ambient space the data is embedded in. This manifold is determined by many constraints: Bond lengths have to be respected, chiralities are fixed, etc. Manifold learning flows, a recently developed class of generative models, have been developed with cases like this in mind, but have not yet been applied to molecular conformer generation. Here, we build manifold learning architectures that split the learning task in two: a first neural network learns the structure of the data manifold, representing each data point in a lower dimensional space. a second neural network learns to estimate the density on this manifold. If we use normalizing flows for both networks, we recover a generative model with much fewer parameters than a flow working in the ambient space directly.

Stand № 19

Machine-Learned Potentials for the Simulation of Hydrogen Atom Transfer

Marlen Neubert (marlen.neubert@kit.edu)

Institute of Theoretical Informatics, KIT

Computer simulations are becoming an important tool for solving complex chemistry, physics, and materials science problems. Machine-learned potentials are able to accelerate simulations of systems on an atomic level. They can therefore be seen as a computational microscope. The wide range of application areas includes biological systems such as proteins and processes therein. In this poster, we will present our research on the development of graph neural networks and active learning approaches for the training of machine-learned potentials, with application to describe chemical reactions, in particular hydrogen atom transfer reactions in collagen. Challenges include the systematic and automated generation of data, coupled with the training of machine learning models in an iterative way. Both neutral and saturated molecular systems as well as radicals need to be covered by the training data and modeled by the graph neural network. Exploration algorithms directly coupled with the active learning model can help to find relevant transition states and thus energy barriers. This project is part of the SIMPLAIX project (Subproject 1) as well as the HIDSS4Health graduate school.

Stand № 20

Quality In, Quality Out: Computing Accurate Redox Potential to Train Neural Networks

Anastasiia Nihei (niheianastasiia@gmail.com)

Heidelberg Institute for Theoretical Studies (HITS)

Redox (oxidation/reduction) potentials determine utility of organic molecules in practical applications ranging from rechargeable batteries to enzymatic catalysis. In this project, we establish an accurate ab initio procedure for computing redox potentials of organic molecules, which will subsequently be used in machine learning. We test the ability of several density functional theory methods – PBE0, M062X, and B3LYP – in conjunction with various continuum solvent models for acetonitrile within the framework of the Born-Haber thermodynamic cycle to reproduce experimentally measured potentials. Among the three methods, M062X best reproduces the experimental data (R2 = 0.935), reaching chemical accuracy (mean absolute error, MAE, equal to 0.21; for experimental measurements, MAE = 0.25). In this manner, highly accurate redox potentials can be computed and fed into a graph neural network, ensuring quality predictions for new systems.

Stand № 21

Electronic properties of modeled Spiropiran-based Metal-Organic Frameworks

Helmy Pacheco Hernández (helmy.hernandez@kit.edu)

Karlsruhe Institute of Technology

Materials that possess photoswitchable electronic properties and the ability to undergo reversible changes in conductance are highly desirable. Metal-organic framework (MOF) films, functionalized with photoresponsive linkers based on spiropyran, have demonstrated the potential to switch conduction by light with large on-off ratios. The current process for synthesizing MOF materials is laborious and could benefit significantly from the implementation of in silico molecular design. In this study, we have designed photoswitchable Metal-Organic Frameworks (MOFs) that incorporate spiropyran photoswitches at precise positions. By implementing multiscale modeling and automated workflow protocols, we have characterized four MOF based on Spiropyran linkers and explored their potential for photoswitching properties. Through ab initio calculations of the electronic coupling between molecules in the MOFs, we have demonstrated that lattice distances significantly affect the photoswitching of conduction between spiropyran- and merocyanine-based MOFs upon light absorption. Merocyanine exhibits higher molecular flexibility and additional intermolecular π-π interactions between linkers, resulting in an increase in electronic coupling. This increase leads to on-off ratios in the order of 10^4 for conduction switching. This study offers valuable information for designing smart materials with large switching conduction ratios.

Stand № 22

PepAIsim: combining AI and molecular simulations for anticancer peptide and peptidomimetic design

Giulia Paiardi[1,2], Elke Burgermeister E.[3], Rebecca C. Wade[1,2] (giulia.paiardi@h-its.org)

[1]Zentrum für Molekular Biologie (ZMBH) Heidelberg University; [2]Heidelberg Institute for Theoretical Studies (HITS); [3]Universitätsmedizin Mannheim, Medical Faculty Mannheim, Heidelberg University

Anticancer peptides (APs) represent a promising class of therapeutic molecules. The development and identification of APs are time-consuming and expensive in traditional wet-lab-based approaches. Computational studies of protein-peptide interactions can speed up the process. However, several shortcomings hamper their computational optimization. Peptides are more flexible than proteins, making it more difficult to properly map their conformational landscape, predict their interactions and function, and design mimetics. Therefore, we aim to implement a computational approach, combining machine learning (ML) with physics-based simulations to investigate protein-peptide interactions and predict their molecular mechanisms to ultimately design optimized anticancer peptides. As a case study, we focus on the potential treatment of gastrointestinal cancer with peptides that mimic the effect of the tumor suppressor Docking protein-1 (DOK1). Here, we are investigating the binding mode and mechanisms of DOK1-peptide, derived from the DOK1 protein, to the two human proteins HAKAI and PPARγ. Our results show that, despite sharing common binding residues, DOK1-peptide interacts with its target proteins via different binding modes. We identify direct mechanisms by which DOK1-peptide can prevent the above-mentioned protein-protein interactions. Our results lay the basis for the ML-driven optimization of DOK1 peptide derivatives toward new anticancer therapeutics.

Stand № 23

4th Generation HDNNP for QM/MM Calculations

Lukas Petersen (lukas.petersen@kit.edu)

Karlsruhe Institut of Technology

High-dimensional neural network potentials (HDNNPs), developed by Behler and Parinello, apply atom-centered symmetry functions (ACSFs) as descriptors for molecular systems. This approach becomes problematic in condensed-phase reactions due additional element types and sampling of the phase space. A QM/MM-like approach, where the reaction center and the environment are handled separately, would be desirable. This could be achieved by including the electrostatic potential caused by environment, typically solvent or protein-backbone, in the model by considering the electrostatic interaction between the two zones during the CENT scheme of the 4th Generation HDNNP. The model reaction we consider is the thiol-disulfide exchange, which occurs dynamically in proteins to form new disulfide bridges. This reaction is especially ambitious due to the high polarizability of the sulfur atoms.

Stand № 24

Unveiling patterns in nonadiabatic molecular dynamics data with machine learning: the ULaMDyn program

Max Pinheiro Junior (max.pinheiro-jr@univ-amu.fr)

Aix-Marseille University

Trajectory-based nonadiabatic molecular dynamics (NAMD) are as robust data generators. Owing to the high dimensionality of NAMD data, it is challenging to find key active coordinates driving the molecular system towards critical regions of the potential energy surfaces, where chemical transformations may occur. Also, the myriad of possible reaction pathways accessible in NAMD adds an extra layer of difficulty to the data exploration problem. Unsupervised machine learning (ML) can bring an automated solution for the in depth analysis of NAMD data, facilitating the interpretation and understanding of the underlying photo-dynamical processes. To contribute to this solution, we have developed the Unsupervised Learning Analysis of Molecular Dynamics (ULaMDyn) program that provides a complete data analysis pipeline, going from data curation to molecular representations, dimension reduction, and clustering analysis. The unsupervised learning methods implemented in ULaMDyn aim to surpass existing barriers for chemists to extract insights from NAMD simulations regardless of the complexity of the molecular system under study. In this work, I will present the theoretical aspects of unsupervised learning, showcasing applications of dimensionality reduction and clustering techniques for analyzing NAMD data. ULaMDyn will assist chemists in understanding photochemical phenomena without requiring prior knowledge of the underlying chemical reaction mechanisms.

Stand № 25

How collagen is designed to tame its radicals

Benedikt Rennekamp (benedikt.rennekamp@h-its.org)

Heidelberg Institute for Theoretical Studies (HITS)

Collagen is a force-bearing, hierarchical structural protein important to all connective tissue. In tendon collagen, high load even below macroscopic failure level creates mechanoradicals by homolytic bond scission, similar to polymers. The location and type of initial rupture sites critically decide on both the mechanical and chemical impact of these micro-ruptures on the tissue, but are yet to be explored. We here use scale-bridging simulations to determine breakage points in collagen: In regular Molecular Dynamics (MD) simulations, covalent bonds are predefined and reactions cannot occur. To circumvent these limitations, we present our previously developed reactive Kinetic Monte Carlo / Molecular Dynamics (KIMMDY) scheme. Here, bond rupture rates are calculated based on the interatomic distances in the MD simulation and then serve as an input for a Kinetic Monte Carlo step. Recently, we have improved upon its accuracy with new Bond Dissociation parameters obtained by high-level quantum mechanical calculations. We find collagen crosslinks, as opposed to the backbone, to harbor the weakest bonds, with one particular bond in trivalent crosslinks as the most dominant rupture site. We identify this bond as sacrificial, rupturing prior to other bonds while maintaining the material's integrity. Also, collagen's weak bonds funnel ruptures such that the potentially harmful mechanoradicals are readily stabilized. Our results suggest this unique failure mode of collagen to be tailored towards combatting an early onset of macroscopic failure and material ageing.

Stand № 26

Analysis of transient pockets in proteins using TRAPP

Jui-Hung Yuan, Sungho Bosco Han, Stefan Richter, Daria B. Kokh and Rebecca C. Wade
(stefan.richter@h-its.org)

Heidelberg Institute for Theoretical Studies

TRAnsient Pockets in Proteins (TRAPP) is a tool designed to aid the discovery of molecules that bind in transient or cryptic subpockets in proteins. TRAPP is not designed to identify all of a protein's binding pockets, but rather to trace changes in the spatial and physicochemical properties of a specified pocket in a protein that may arise due to the protein's flexibility. It includes tools designed to efficiently generate and analyse binding site motions and explore protein cavity dynamics. The pocket conformations can be characterized by their physicochemical and sequence properties as well as by druggability indices. The druggability indices have been derived using a logistic regression model and a convoluted neural network trained using the Non-Redundant Druggable and Less Druggable (NRDLD) dataset augmented by a PDBbind-based dataset (DaPB). These models allow the selection of frames in molecular dynamics trajectories with potentially druggable protein conformations. TRAPP is available as webserver (https://trapp.h-its.org) as well as standalone tool.

Yuan J, Han SB, Richter S, Wade RC, Kokh DB (2020). Druggability Assessment in TRAPP Using Machine Learning Approaches, J. Chem. Inf. Model. 60(3):1685-1699 Stank A, Kokh DB, Horn M, Sizikova E, Neil R, Panecka J, Richter S, Wade RC (2017). TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets., Nucleic Acids Research 45(W1):W325-W330 Kokh DB, Richter S, Henrich S, Czodrowski P, Rippmann F, Wade RC (2013). TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins, J. Chem. Inf. Model. 53(5):1235-1252 57

Stand № 27

Designing a machine-learned protein force field

Leif Seute (leif.seute@h-its.org)

Heidelberg Institute for Theoretical Studies

In traditional Molecular Mechanics (MM) force fields, a finite set of hand-crafted chemical perception rules is used to determine parameters for a given molecule. Building upon work from Chodera et al., we design a protein force field in which the hand-crafted rules are replaced by a machine learning approach using graph neural networks. Besides potentially increasing accuracy of legacy protein force fields, our approach allows customized fine-tuning of the force field, making it applicable to rather exotic molecules like protein radicals while keeping the computational efficiency of MM potentials.

Stand № 29

Machine Learning the Fluoride Ion Affinity of p-Block Element-based Lewis Acids

Lukas Sigmund (lukas.sigmund@aci.uni-heidelberg.de)

Ruprecht-Karls-Universität Heidelberg

The fluoride ion affinity (FIA) is among the most common scales to quantify the strength of Lewis acids in a global sense. It is usually obtained with quantum chemical computations. For a single accurate solution-phase FIA, it needs twelve separate calculations, which hampers the fast computational exploration of new Lewis acids. Therefore, we developed a statistical model for the prediction of FIAs. A dataset constituting 15 different p-block elements as Lewis acidic centers combined with a wide range of mono- and polydentate substituents was compiled. The dataset contains over 7000 datapoints and spans a FIA range of around 550 kJ/mol. With that, a gradient-boosted decision trees regressor was trained. The feature space was constructed only with atom connectivity graph-based data from the Lewis acid and its fluoride adduct. Any kind of 3D information was not included, which allows FIA predictions with SMILES strings or analogous structural encodings. The features were organized in atom shells stratified around the central Lewis acidic atom. The model predicts the FIA of unseen compounds from the dataset with a MAE of 13 kJ/mol (R^2=0.94). For literature-known test Lewis acids the accuracy is slightly reduced (MAE of 18 kJ/mol, R^2=0.91).

Stand № 30

How a Stretching Force Differently Destabilizes Chemical Bonds on a Protein Backbone

Daniel Sucerquia (daniel.sucerquia@h-its.org)

Heidelberg Institute for Theoretical Studies

When subjecting a protein chain to extreme pulling forces, bonds in the stretched backbone ultimately break. Predicting such ruptures can help to understand failure of protein materials. As a most simple assumption, a protein backbone can be considered as a series of harmonic springs each of which carries the same force, and they only differ in their thermodynamic stability. However, proteins are more complex than that and force will distribute across the various degrees of freedoms in the peptide, largely depending on the chemical environment. We here study the changes of energy stored in the degrees of freedom of molecules at quantum level of accuracy using JEDI (T. Stauch and A. Dreuw, Chem. Rev. 116, 2016). JEDI assesses the distribution of energies in stretched molecules using density functional theory and a harmonic approximation around optimized conformations. We so far have tested this method in chains of amino acids consisting of alanines, glycines, and prolines, and their combinations. We observe a linear increase in energies per degree of freedom including bonds, angles, and dihedrals, during stretching, and proline to show an energy distribution distinct from the other amino acids due to the ring structure. Data from QM and JEDI calculations of a large set of small peptides will aid to predict the energy distribution in larger systems using Machine Learning, for example for allowing bond rupture during classical Molecular Dynamics simulations.

Stand № 31

Multiscale simulation of Cytochrome P450 electron transfer complexes : The reduction of CYP17A1 and its implications for the regulation of human sex hormone biosynthesis

Jonathan Teuffel[1,2], Goutam Mukherjee[1,3,4], Sungho Bosco Han[1] and Rebecca C. Wade[1,2,3,4] (jonathan.teuffel@h-its.org)

[1]Heidelberg Institute for Theoretical Studies; [2]Faculty for Engineering Sciences, Heidelberg University; [3]Zentrum für Molekulare Biologie (ZMBH), Heidelberg University; [4]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University.

The cytochrome P450 17A1 (CYP17A1) is a central node in the steroid hormone synthesis metabolic network and is one of a subgroup of cytochrome P450 enzymes that requires interaction with two redox proteins: cytochrome b5 and NADPH cytochrome P450 oxidoreductase. We applied a multiscale simulation protocol to investigate how human cytochrome P450 17A1 (CYP17A1) receives electrons from these two different redox proteins upon protein-protein complex formation. Extensive molecular dynamics simulations were run, generating ensembles that were analysed to yield information on the structural rearrangements of the redox partners upon complex formation and on electron-transfer kinetics. Our simulations indicate that both redox proteins can transfer electrons at a similar rate but via different pathways that involve non-aromatic residues and the protein backbone. We also found that the binding modes of both reductases are altered upon embedding the complexes into a phospholipid bilayer and that the binding of the reductases in turn alters the orientation of CYP17A1 relative to the membrane plane. Our findings show how association to different redox proteins differentially impacts the active site accessibility and the activity of CYP17A1 through conformational rearrangements. Comparison of the computed electron transfer pathways with those for other cytochrome P450 enzymes will provide a basis for deriving a machine learning model to predict the sequence dependence of cytochrome P450 electron transfer kinetics.

Stand № 32

The COMPAS Project: Expanding into peri-Condensed Polybenzenoid Hydrocarbons

Alexandra Wahab (awahab@org.chem.ethz.ch)

ETH Zürich

Polycyclic aromatic systems are one of most prevalent types of compounds in nature and are important in many fields in chemistry. We recently introduced the COMPAS project: a COMputational database of Polycyclic Aromatic Systems, the first methodically curated database for such systems. The first installment of this project focused on cata-condensed polybenzenoid hydrocarbons (PBHs). In the current work, we present the second installment, where we explore the chemical space of peri-condensed PBHs. We constructed two data sets containing optimized ground-state structures and molecular properties: COMPAS-2x (~40k molecules with up to 11 rings, calculated with GFN2-xTB) and COMPAS-2D (~9k molecules with up to 10 rings, calculated at the CAM-B3LYP-D3BJ/aug-cc-pVDZ level). Herein, we describe the workflow of data generation, the data curation pipeline, and the information available within the data sets. We compare the two types of computations and detail the structure–property relationships revealed by the data analysis. The data and insights gained can drive rational design of novel functional polycyclic aromatic molecules with applications in, e.g., organic electronics, and can provide a basis for additional data-driven machine- and deep-learning studies in chemistry, which we previously showed with our first installment of the COMPAS database.

Stand № 33

Active Learning for Potential Energy Surface Simulation

Chen Zhou (chen.zhou@kit.edu)

Karlsruhe Institute of Technology

The accurate and efficient exploration of chemical space is essential for rational compound design and production in the field of chemical, materials and pharmaceutical industries. This has led to growing interest within the scientific community of utilization of machine learning in molecular dynamic simulations, which are often prohibitive for large molecules or long time scales with classic quantum calculation methods. Here we introduce an active learning workflow that efficiently simulates the potential energy surface of a 38-atom molecule with 6 excited electronic states. The workflow takes advantage of a fully connected neural network, enables both data and task parallelism on computer clusters with Message Passing Interface, and achieves high accuracy on energy/force prediction (R2 = 0.999 and 0.998 respectively). With the flexible architecture, we expect this active learning workflow to be readily extensible towards other oracles and machine learning models.

Stand № 34

Author Index