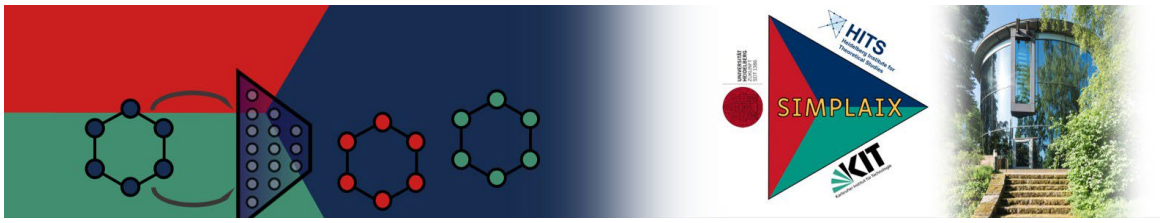

2nd SIMPLAIX Workshop on "Machine Learning for Multiscale Molecular Modeling"

15 - 17 May 2024
Studio Villa Bosch Heidelberg, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg
https://simplaix-workshop2024.h-its.org/

Abstract Book

Scientific Organizing Committee: Daniel Sucerquia, Rostislav Fedorov, Rebecca Wade, (HITS); David Hoffman, Pascal Friederich, Markus Elstner, (KIT)

Administrative Team: Britta Schläger, Angela Michel, Christina Blach, Marisa de Sá Almeida (HITS)

Organizers:

HITS Heidelberg Institute for Theoretical Studies

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

KIT Karlsruher Institut für Technologie

Sponsors:

Klaus Tschira Stiftung gemeinnützige GmbH

KTS

GRK 2450

Program

Wednesday, May 15, 2024

- *Registration*: 13:00
- Opening: 14:00
    - Speakers: Tilmann Gneiting, Rebecca Wade
- **Session 1:**
    - **Session Chair:** Frauke Gräter, HITS
        - Talk 1: 14:15 - Mark E. Tuckerman - "Synergizing enhanced sampling and machine learning strategies in molecular simulation for representing and deploying high-dimensional free energy surfaces and learning reaction coordinates"
        - Short presentation: 14:55 - Andres Escorcia - "Multi-scale simulations reveal a Cys-His-Water-Glu catalytic network in Cezanne-2"
        - Short presentation: 15:15 – Roman Remme – “Deep Learning meets Density Functional Theory”
- *Coffee break*: 15:45
- **Session 2:**
    - **Session Chair:** Tristan Bereau, Heidelberg University
        - Talk 1: 16:30 - Matteo Degiacomi - "Learning (from) protein dynamics"
        - Talk 2: 17:10 - Roberto Covino - "Molecular Free Energies, Rates, and Mechanisms from Machine learning-guided Path Sampling Simulations"
        - Short presentation: 17:50 - Aleksander Durumeric - "Coarse-grained molecular dynamics for proteins with neural networks: Challenges and breakthroughs"
- *Photo* : Speakers
- *Invited speakers' dinner*: 19:30

Thursday, May 16, 2024

- **Session 3:**
    - **Session Chair:** Pascal Friederich, KIT
        - Talk 1: 09:00 - Johannes Margraf - "Science-Driven Chemical Machine Learning"
        - Talk 2: 09:40 - Geemi Wellawatte - "Explainable AI in Chemistry"
        - Short Talk: 10:20 - Luis Itza Vazquez Salazar - "Asparagus: a toolkit for automatic construction of machine-learned potential energy surfaces"
- *Coffee break*: 10:40
- **Session 4:**
    - **Session Chair:** Marcus Elstner, KIT
        - Talk 1: 11:20 - Lorenzo Cupellini - "Understanding electronic excitations in complex systems with machine learning"
        - Short Talk: 12:00 - Daniel Bultrini - "Mixed quantum-classical dynamics for near term quantum computers"
- Discussion round table: 12:20 (Chairs: Pascal Friederich, KIT / Fred Hamprecht, Heidelberg University)
- *Lunch*: 13:00
- *Photo* : All: 14:20

Thursday, May 16, 2024

- **Session 5:**
    - **Session Chair:** Rebecca Wade, HITS
        - Talk 1: 14:30 - Felice Lightstone - "Using the Multiscale Machine-Learned Modeling Infrastructure (MuMMI) to Explore Protein Conformational Paths for RAS-RAF Activation"
        - Talk 2: 15:10 - Edina Rosta - "Enhanced Sampling Simulations of Biomolecular Systems"
        - Short Talk: 15:50 - Rostislav Fedorov - " Exploration of redox properties in chemical space"
- *Coffee break*: 16:10
- **Session 6:**
    - **Session Chair:** Ullrich Köthe, Heidelberg University
        - Talk 1: 16:40 - Julia Westermayr - "Advancing excited-state simulations with machine learning"
        - Talk 2: 17:20 - Grant Rotskoff - "Accelerating Biomolecular Ensemble Sampling with Generative Neural Networks"
        - Short Talk: 18:00 - Stiv Llenga - "Spying On Molecules: Constructing Novel Compound Spaces via Molecular Triangulation"
- *Poster session*: 18:20
- *Workshop Dinner + Posters*: 19:00

Friday, May 17, 2024

- **Session 7:**
    - **Session Chair:** Fred Hamprecht, Heidelberg University
        - Talk 1: 09:00 - Boris Kozinsky - "Combining data, physics and machine learning for accelerating materials computations."
        - Talk 2: 09:40 - Andrea Volkamer - "Hybrid AI for Molecular Design"
        - Short Talk: 10:20 - Leif Seute - "Learning a State-of-the-Art MM Force Field"
- *Coffee break*: 10:40
- **Session 8:**
    - **Session Chair:** Anya Grynova, HITS
        - Talk 1: 11:20 - Alessandro Troisi - "Digital Discovery of Organic Electronics Materials"
        - Short Talk: 12:00 - Anita Ragyanszki - "Understanding the Origins of Life - A Machine learning approach to estimate reaction mechanisms of biotic precursors"
- Discussion round table and round up: 12:20 (Chairs: Rostislav Fedorov / Rebecca Wade, HITS)
- *Lunch*: 13:00
- *End*: 14:30

Talks

Synergizing enhanced sampling and machine learning strategies in molecular simulation for representing and deploying high-dimensional free energy surfaces and learning reaction coordinates

Prof. Mark E. Tuckerman (mark.tuckerman@nyu.edu)

Department of Physics, Courant Institute of Mathematical Sciences
New York University, USA

Rare-event sampling techniques and machine learning models have become integral tools in the theoretical and computational molecular sciences. This talk will focus on the use of machine learning and rare-event sampling strategies for sampling conformational spaces, generating and representing high-dimensional free-energy landscapes, and finding reaction coordinates that characterize transitions between different basins on these high-dimensional free energy surfaces. We will first review collective-variable based enhanced sampling techniques as implemented in a branch of the OpenMM package. We will then show how machine learning techniques can be leveraged to regress high-dimensional free energy surfaces via an active learning approach for the computation of observables and to perform dimensional reduction to learn reaction coordinates using a game theory approach known as SHAP analysis. Specific examples from biomolecular conformational sampling and transition mechanisms will be presented, which will include small protein and RNA sequences.

Learning (from) protein dynamics

Dr. Matteo Degiacomi (matteo.t.degiacomi@durham.ac.uk)

Department of Physics, Durham University, UK

Determining the different conformational states of a protein and the transition paths between them is key to fully understanding the relationship between biomolecular structure and function. I will discuss how a generative neural network (GNN) can learn a continuous conformational space representation from example structures produced by molecular dynamics simulations or experiments. I will then show how such representation, obtained via our open-source available software molearn, can be leveraged on to predict putative protein transition states, or to generate conformations useful in the context of flexible protein-protein or protein-ligand docking. Finally, I will demonstrate that transfer learning is possible, i.e., a GNN can learn features common to any protein.

Molecular Free Energies, Rates, and Mechanisms from Machine learning-guided Path Sampling Simulations

Prof. Roberto Covino (covino@fias.uni-frankfurt.de)

Frankfurt Institute for Advanced Studies, Germany

Exascale computing holds great opportunities for molecular simulations. However, to take full advantage of the new possibilities, we must learn how to focus computational power on discovering and sampling complex molecular events and extracting quantitative yet understandable representations of their mechanisms from enormous amounts of data. I will present AIMMD, a computational framework that integrates transition path theory, multi-scale modeling, advanced sampling, and deep learning to guide the sampling and extract quantitative mechanistic models. Our framework enables efficient sampling and estimates of free energy and rates characterizing complex molecular events. It adaptively and autonomously initializes and orchestrates thousands of simulations in a data-driven way and is thus suitable for massively parallel computing.

Science-Driven Chemical Machine Learning

Prof. Johannes Margraf (johannes.margraf@uni-bayreuth.de)

Faculty of Biology, Chemistry & Earth Sciences, University of Bayreuth, Germany

This talk discusses research towards the establishment of a science-driven approach to chemical machine learning (ML).[1] In many fields, ML is a fundamentally data-driven endeavor, meaning that specific databases and benchmark problems (i.e. big data) are at the center of methodological development. While this has certainly led to tremendous advances in recent years (e.g. in image generation and natural language processing), the full diversity and complexity of chemistry cannot be adequately represented by a few predefined databases. A major driver of this work is therefore the desire to build accurate, data-efficient models which do not require enormous reference datasets for training. This is because we want to be able to apply our methods to any problem of chemical interest, not just to those problems for which big data happens to be available. To this end, we explicitly incorporate chemical and physical information into the ML models[2] and integrate the data generation or selection process with the model training[3]. Several examples of this in the context of the atomistic simulation of energy materials will be discussed. [1] J.T. Margraf, Angew. Chemie, 62, e202219170 (2023). [2] K. Chen et al. Chem. Sci., 14, 4913-4922 (2023). [3] H. Jung et al. NPJ Comput. Mater., 9, 114 (2023).

Explainable AI in Chemistry

Dr. Geemi Wellawatte (geemi.wellawatte@epfl.ch)

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Explainable Artificial Intelligence (XAI) is an emerging field in AI that aims to address the opaque nature of machine learning models. Furthermore, it has been shown that XAI can be used to extract input-output relationships, making them a useful tool in chemistry to understand structure-property relationships.

Understanding electronic excitations in complex systems with machine learning

Dr. Lorenzo Cupellini (lorenzo.cupellini@unipi.it)

Department of Chemistry, University of Pisa, Italy

Photophysical processes of biological interest occur in condensed phase, where light absorption and excited-state dynamics are tuned by chromophore-environment interactions. Faithful quantum chemical (QM) modeling of these processes requires a multiscale QM/classical strategy [1], which should be coupled to a thorough sampling of the configurational ensemble. The computational cost of QM calculations however forces a compromise between statistical accuracy and feasibility. Machine learning (ML) techniques can successfully substitute QM calculations for excited-state properties and dynamics [2]. However, much less attention has been paid to environment effects [3]. Here I will outline ML strategies that can model the environment effect on excited states in light-harvesting pigment-protein complexes [4]. The obtained ML models not only can faithfully predict the effect of the environment on electronic excitations[5], but also assist the interpretation of solvatochromic effects [6]. [1] M. Nottoli, et al.,Annu. Rev. Phys. Chem. 2021, 72, 489 [2] J. Westermayr, P. Marquetand. Chem. Rev. 2021, 121, 9873 [3] M. S. Chen et al., J. Phys. Chem. Lett. 2020, 11, 7559 [4] E. Cignoni et al., J. Chem. Theory Comput. 2023, 19, 965 [5] E. Betti et al., doi:10.26434/chemrxiv-2024-0h765 [6] A. Arcidiacono et al., doi:10.26434/chemrxiv-2024-3zrx2

Using the Multiscale Machine-Learned Modeling Infrastructure (MuMMI) to Explore Protein Conformational Paths for RAS-RAF Activation

Prof. Felice C. Lightstone (lightstone1@llnl.gov)

Lawrence Livermore National Laboratory, Livermore, USA

Frequently, biological mechanisms, like RAS-RAF activation, start at the atomistic level and result in large protein conformational changes that span seconds of time. As of today, not one simulation method allows the study of atomic detail over long experimental timescales. Molecular dynamics (MD) simulations are able to provide detailed insights to many biological mechanisms but are limited to short time scales. Macro-scale simulations can cover experimental conditions and longer timescales but do not provide the necessary detail to directly observe phenomena such as protein association or signaling. Here, we present the massively parallel Multiscale Machine-Learned Modeling Infrastructure (MuMMI) which effectively couples different simulation scales. MuMMI uses machine learning (ML) to couple adjacent simulation scales. ML-guided selections from the coarser, faster models sample larger time scale motion and changes in local environment, creating a vast simulation ensemble at finer resolution. The more detailed, but slower models, sample progressively finer details. Online analysis and aggregation of the finer resolution simulations is then fed back into the coarser models progressively improving their fidelity. Given sufficient resources, the system will effectively converge the higher resolution models at the parameter space explored in the lower resolution models.

Enhanced Sampling Simulations of Biomolecular Systems

Prof. Edina Rosta (e.rosta@ucl.ac.uk)

Department of Physics and Astronomy, University College London, UK

Phosphate catalytic enzymes are essential and ubiquitous to all forms of life. While structures of these proteins are typically readily available, prediction and design of their function and activity is a key current challenge. Here we present computing intensive free energy calculation data and machine learning applications to predict catalytic activity for prototype examples including Ras [1]. Our work highlights the important role of coupled proton transfer steps in the catalytic mechanism using the finite-temperature string method. This allows us to use multiple collective variables that govern the reaction path. Identification of these collective variables in complex processes presents a major problem. We offer promising AI-driven algorithms to help identify essential reaction coordinates in biomolecular processes [2,3]. References [1] Berta, D.; Gehrke, S.; Nyíri, K.; Vértessy, B. G.; Rosta, E. Mechanism-Based Redesign of GAP to Activate Oncogenic Ras. JACS, 2023, 10.1021/jacs.3c04330. [2] Badaoui, M.; Buigues, P. J.; Berta, D.; Mandana, G. M.; Gu, H.; Földes, T.; Dickson, C. J.; Hornak, V.; Kato, M.; Molteni, C.; Parsons, S.; Rosta, E. Combined Free-Energy Calculation and Machine Learning Methods for Understanding Ligand Unbinding Kinetics. J. Chem. Theory Comput. 2022, 10.1021/acs.jctc.1c00924. [3] Buigues, PJ; Gehrke, S; Badaoui, M; Mandana, G. M.; Qi, T; Bottegoni, G; Rosta, E. Investigating the Unbinding of Muscarinic Antagonists from the Muscarinic 3 Receptor. bioRxiv; 2023, 10.1101/2023.01.03.522558.

Advancing excited-state simulations with machine learning

Prof. Julia Westermayr (julia.westermayr@uni-leipzig.de)

Wilhelm-Ostwald-Institute for Physical and Theoretical Chemistry
University of Leipzig, Germany

Understanding the basic principles that determine the electronic properties and chemical dynamics of molecules and materials is of paramount importance in various fields such as medicine, materials science, and energy production. This understanding further paves the way for the targeted development of innovative compounds. In this talk, I will discuss the potential of machine learning algorithms to accelerate the discovery of chemical reactions and the evaluation of molecular electronic properties in both ground and excited states, thereby improving our understanding of photochemical processes [1,2]. In addition, I will present the effectiveness of integrating deep neural networks with generative models to refine the design of molecules with specific properties [3] and the application of reinforcement learning to navigate the vast chemical structure space [4]. References [1] Julia Westermayr and Philipp Marquetand, Chem. Rev. 2020, 121(16), 9873-9926. [2] Julia Westermayr, Michael Gastegger, Dóra Vörös, Lisa Panzenboeck, Florian Joerg, Leticia González, and Philipp Marquetand, Nat. Chem. 2022, 14(8), 914-919. [3] Julia Westermayr, Joe Gilkes, Rhyan Barrett, Nat. Comput. Sci. 2023, 3(2), 139-148. [4] Rhyan Barrett and Julia Westermayr, J. Phys. Chem. Lett. 2024, 15, 349-356.

Accelerating Biomolecular Ensemble Sampling with Generative Neural Networks

Prof. Grant M. Rotskoff (rotskoff@stanford.edu)

Department of Chemistry, Stanford University, USA

Excitement at the prospect of using data-driven generative models to sample configurational ensembles of biomolecular systems stems from the extraordinary success of these models on a diverse set of high-dimensional sampling tasks. Unlike image generation or even the closely related problem of protein structure prediction, there are currently no data sources with sufficient breadth to parametrize generative models for conformational ensembles. To enable discovery, a fundamentally different approach to building generative models is required: models should be able to propose rare, albeit physical, conformations that may not arise in even the largest data sets. In this talk, I will describe several recently developed approaches to this problem and how to leverage inexpensive data to scale deployment.

Combining data, physics and machine learning for accelerating materials computations

Prof. Boris Kozinsky (bkoz@g.harvard.edu)

Harvard School of Engineering and Applied Sciences, Harvard University, USA

We pursue a multi-tier method development strategy in which machine learning (ML) algorithms are combined with exact physical symmetries and constraints to significantly accelerate computations of electronic structure and atomistic dynamics. To improve DFT functionals, we introduce non-local charge density and orbital-dependent ML models, called CIDER, that satisfy exact scaling constraints and learn exchange functionals. These models are orders of magnitude faster in self-consistent calculations for solids than hybrid functionals but similar in accuracy, and able to produce accurate defect levels and band gaps. On a different level, we accelerate molecular dynamics (MD) simulations by using machine learning to capture the potential energy surfaces obtained from quantum calculations. We developed NequIP and Allegro, the first deep equivariant neural network interatomic potential models, whose Euclidean symmetry-preserving layer architecture achieves state-of-the-art data efficiency and accuracy for simulating dynamics of molecules and materials. In parallel, we implement autonomous active learning of interactions in reactive systems, with the FLARE algorithm that constructs accurate and uncertainty-aware Bayesian force fields on-the-fly from a molecular dynamics simulation, using Gaussian process regression. These MD simulations are used to explore long-time dynamics of phase transformations and heterogeneous reactions.

Hybrid AI for Molecular Design

Prof. Andrea Volkamer (volkamer@cs.uni-saarland.de)

Data Driven Drug Design, Saarland University, Saarbruecken, Germany

In the realm of molecular design, the fusion of artificial intelligence (AI) with domain expertise presents a compelling avenue for addressing the inherent complexities in drug design. This holds particularly true for human protein kinases, where challenges such as drug promiscuity, resistance, and uncharted kinase territories persist, despite the wealth of available data comprising over 6,000 human kinase structures in the PDB and around 70 small molecule kinase inhibitors.

This presentation showcases open source approaches that integrate knowledge into AI frameworks, overcoming data scarcity issues and enhancing model generalization. Leveraging openly available kinase data, we demonstrate how hybrid AI and classical methods can generate new insights and foster community engagement. The TeachOpenCADD platform [1,2] serves as a versatile tool for orchestrating diverse computer-aided drug design (CADD) tasks exemplified on individual kinases. We introduce freely available resources to support kinase research, including: (i) KinFragLib for fragment-based kinase inhibitor design [3], (ii) KiSSim – a KLIFS-based kinase structural similarity fingerprint [4], and (iii) a comprehensive pipeline for assessing kinase similarity from various data perspectives [5].

Furthermore, ongoing projects in structure-informed machine learning (ML) for kinase inhibitor design and affinity prediction across kinases are outlined, underscoring the fusion of deep learning with domain expertise in hybrid AI approaches [6,7]. By bridging experimental insights with computational methodologies, these initiatives advance the frontier of kinase-focused drug discovery.

Digital Discovery of Organic Electronics Materials

Prof. Alessandro Troisi (a.troisi@liverpool.ac.uk)

Department of Chemistry, Liverpool University, UK

The activities of the group combining virtual screening, machine learning and physics based model will be reviewed with particular focus on the integration between the different approaches.

Multi-scale simulations reveal a Cys-His-Water-Glu catalytic network in Cezanne-2

Dr. Andrés M. Escorcia (escorcia_cabrera@mpi-magdeburg.mpg.de)

Max Planck Institute for Dynamics of Complex Technical Systems

Cezanne-2 (Cez2), a deubiquitinating enzyme, is a potential target for the development of drugs against cancer and neurological disorders. Therefore, understanding the catalytic mechanism of Cez2 at the molecular level is of great importance, as it can help in the rational design of inhibitors. We performed molecular dynamics (MD) simulations in explicit water to study the interaction of di-ubiquitin (diUb) substrate with Cez2 and thus get molecular insight into Cez2 catalysis. Following the generally accepted reaction mechanism of cysteine proteases, we modeled both the apo-enzyme and the diUb Cez2 complex considering two different charge states of the catalytic Cys/His dyad: i) both His367 and Cys210 with neutral charge, and ii) His367 positively charged and Cys210 negatively charged. Four independent MD simulations were performed for each system, using OpenMM and the CHARMM36 force field. The MD simulations provide information on the arrangement of the Cez2 catalytic core and enzyme-substrate-water interactions during catalysis. Configurations of diUb are identified which can undergo hydrolysis by Cez2. The reliability of these configurations was verified by QM/MM (B3LYP/TZVP//CHARMM36) optimizations. In contrast to most deubiquitinating enzymes, these configurations suggest that Cez2 executes catalysis via a Cys-His-Water-Glu catalytic network.

Coarse-grained molecular dynamics for proteins with neural networks: Challeromannges and breakthroughs

Dr. Aleksander Durumeric (alekepd@gmail.com)

Freie Universitaet Berlin

Neural network force-fields have enabled molecular dynamics (MD) simulations at unprecedented accuracy by efficiently emulating expensive ab initio calculations. However, these advances have not yet accelerated the long-timescale modelling of biomolecular complexes, where the computational cost of classical force-fields is difficult to reduce. One leading approach for adapting neural network force fields to this context focuses on creating force-fields at a reduced (i.e. coarse-grained) resolution. We here discuss how this task differs from that at the atomistic resolution and discuss recent advances by myself and colleagues which have brought the of idea of an accurate and extrapolative neural network protein coarse-grained force-fields within reach, with focus on the collection and processing of training data.

Asparagus: a toolkit for automatic construction of machine-learned potential energy surfaces

Luis Itza Vazquez Salazar (litzavazquezs@gmail.com)

Institute of Theoretical Physics, Heidelberg University

The construction of machine-learned potential energy surfaces (PES) is a complicated task that involves several steps from sampling to deployment, passing from the training of a model. Each of these steps is usually performed separately, requiring different expertise in each step, which makes the process inefficient and cumbersome for newcomers to a vibrating field such as atomistic machine learning (ML). In addition to the described problems, most of the efforts have been focused on improving current methods and models while overlooking the steps of sampling and data management. Asparagus is our solution to automatize the process of constructing an ML-PES. This is achieved by putting together all the steps required to create a PES in a single code. The aim of the code is to create a user-friendly experience without falling into a black-box approach. Asparagus is fully written in Python, following a modular architecture that allows users to modify and extend the code at their convenience. Additionally, it relies on state-of-the-art tools that make it easy to apply in multiple conditions. In this talk, I will review the capabilities of asparagus, the different implemented methods, and the philosophy behind them. Additionally, representative examples of applications of asparagus will be discussed.

Mixed quantum-classical dynamics for near term quantum computers

Daniel Bultrini (daniel.bultrini@pci.uni-heidelberg.de)

University of Heidelberg

Mixed quantum-classical dynamics is a set of methods often used to understand systems too complex to treat fully quantum mechanically. Many techniques exist for full quantum mechanical evolution on quantum computers, but mixed quantum-classical dynamics are less explored. We present a modular algorithm based on parameterized quantum circuits for general mixed quantum-classical dynamics where the quantum subsystem is coupled with the classical subsystem. The parameters and optimal circuit are found with a machine learning-based toolset. We test it on a modified Shin-Metiu model in the first quantization through Ehrenfest propagation. The Time-Dependent Variational Time Propagation algorithm performs well for short-time evolutions. It retains qualitative results for longer-time evolutions and can be used to model molecular evolution through various timescales.

Spying On Molecules: Constructing Novel Compound Spaces via Molecular Triangulation

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

HITS

Since the introduction of machine learning in chemistry, the notion of chemical space has evolved from encompassing almost everything that chemistry touches to a relatively well-defined concept: an infinitely dimensional space populated by an infinite number of compounds. Molecular representation defines not only the dimensionality and shape of the compound space, but also the relationships between various compounds within this space. Two key philosophies of exploring the chemical space exist: interpolating from specific datasets within confined areas or extrapolating across the entire space. The latter approach, by definition, remains elusive. Recently, our group has developed a novel technique for constructing the compound space using triangulation and trilateration, which are at the heart of technologies such as GPS and location tracking. In this presentation, we demonstrate the superiority of our method, called matrix of reference similarity (MRS), for performing machine learning on large newly developed and existing datasets and achieving high accuracy in predicting several common electronic properties of chemical compounds. The application of MRS as both a dimensionality reduction technique and an input for supervised or unsupervised machine learning models is exemplified not only in chemistry but also in other array-like inputs. We also discuss the computational benefits of this technique in terms of time, memory, and power.

Learning a State of the Art MM Force Field

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

HITS

Simulating large molecular systems over long timescales requires force fields that are both accurate and efficient. In the recent years, E(3) equivariant neural networks have lifted the tension between computational efficiency and accuracy of force fields, but they are still several orders of magnitude more expensive than classical molecular mechanics (MM) force fields. Here, we propose a novel machine learning architecture to predict MM parameters from the molecular graph, employing a graph attentional neural network and a transformer with symmetry-preserving positional encoding. The resulting force field outperforms established and other machine-learned MM force fields in terms of accuracy at the same computational efficiency and can be used in existing MM engines like GROMACS and OpenMM. Besides predicting energies and forces of small molecules, peptides, RNA and radicals at state-of-the-art MM accuracy, our force field is transferable to macromolecules. We show that it keeps large proteins stable and can fold small proteins in molecular dynamics simulations, setting the stage for protein simulations at an unprecedented level of accuracy, but with the same computational cost as established protein force fields.

Understanding the Origins of Life - A Machine learning approach to estimate reaction mechanisms of biotic precursors

Dr. Anita Ragyanszki (ragyanszki@zib.de)

Zuse Institute Berlin

Life as we know it is the result of billions of years of evolution; however, understanding how the very first organisms came into existence is a challenge that has yet to be solved. One theory states that components of these molecules may have formed in the interstellar medium (ISM) and been transported to Earth. The ISM, with its specific conditions, allows for molecular stability and the formation of biotic precursors that would otherwise be unlikely in Earth's prebiotic conditions. Understanding how these molecules formed in the ISM may be the key to determining how life began. The goal of this research is to develop a model for solving astrobiophysical problems by studying the formation mechanisms of biomolecules found in the ISM. Although such pathways have been studied individually, there has not yet been a comprehensive method to understand the complete reactions mechanisms. Several QM methods are available for finding transition states (TS) and energy barriers (E) of chemical reactions but are time-consuming and can hardly be applied to more complex systems. Our interest is to develop a machine learning approach to approximate TS, and E, requiring as input only estimates of geometry and energies of reactants and products.

Posters

Basis-set correction to GW correlation self-energies

Dominik Steinmetz (dominik.steinmetz@kit.edu)

KIT

The explicitly correlated calculation of the correlation self-energy is investigated by computing pair-energies within the dMP2 (direct Moeller Plesset pertubation theory) method using F12 Slater-type geminals. This contribution to GW correlation self-energies is calculated for small molecules and compared to results obtained without this correction for the use of a finite basis. The comparison was conducted using the aug-cc-pVXZ basis sets (X = D, T, Q, 5, 6, 7) in conjunction with different extrapolation schemes to the CBS (complete basis set) limit.

Stand № 1

Brownian dynamics for studying the protein-ligand association process in a crowded environment

Riccardo Beccaria (riccardo.beccaria@h-its.org)

Heidelberg Institute for Theoretical Studies and Heidelberg University

The association process of biomolecules is one of the regulatory principles for proteins' functionality. For this reason, correct and realistic modelling of this process is indispensable to obtain information on how the association takes place. In this work, we study the association process of protein-ligand complexes in crowded environments. For this purpose, we use Brownian dynamics simulations.

Stand № 2

CGSmiles; a line notation for representing molecules at different resolutions

Dr. Fabian Grünewald (fabian.gruenewald@h-its.org)

HITS

The Simplified Molecular-Input Line-Entry System (SMILES) has become an integral part of the theoretical molecular sciences since its inception in the 1980s. A SMILES string allows researchers to represent a molecule's connectivity and atomic composition in a single string format. The shortness of the notation and the fact that it is human-readable contributed to the success of SMILES. However, SMILES strings are often hardly (human) readable for large polymeric molecules. Thus, notations such as HELM, or BigSMILES have been proposed that break down larger molecules into fragments. For example, proteins may be represented by their sequence and amino-acid residues. This idea is similar to coarse-graining (CG), where molecules are simulated by grouping multiple atoms into effective interaction centers. A line notation, which can describe CG molecules at arbitrary resolution and their correspondence to the underlying all-atom structure would be highly desirable, especially for building large databases of CG molecules and potential use in machine-learning applications. Unfortunately, all current notations are tailored towards all-atom molecules and lack features required by CG models. To overcome these limitations, we have created the CGSmiles notation that builds on ideas from BigSMILES and HELM but allows researchers to represent an arbitrary number of resolutions in a single string. Together with the notation we developed a light-weight Python package that allows reading of CGSmiles strings and resolving them into NetworkX graphs allowing for an API that easily interfaces with any major ML package.

Stand № 3

Conditional Normalizing Flows for Active Learning of Coarse-Grained Molecular Representations

Henrik Schopmans (henrik.schopmans@kit.edu)

KIT - Institute of Theoretical Informatics

Efficient sampling of the Boltzmann distribution of molecular systems is a long-standing challenge. Recently, instead of generating long molecular dynamics simulations, generative machine learning methods such as normalizing flows have been used to learn the Boltzmann distribution directly, without samples. However, this approach is susceptible to mode collapse and thus often does not explore the full configurational space. In this work, we address this challenge by separating the problem into two levels, the fine-grained and coarse-grained degrees of freedom. A normalizing flow conditioned on the coarse-grained space yields a probabilistic connection between the two levels. To explore the configurational space, we employ coarse-grained simulations with active learning which allows us to update the flow and make all-atom potential energy evaluations only when necessary. Using alanine dipeptide as an example, we show that our methods obtain a speedup to molecular dynamics simulations of approximately 15.9 to 216.2 compared to the speedup of 4.5 of the current state-of-the-art machine learning approach.

Stand № 4

Covalent Organic Frameworks for Drug Delivery

Gregor Lauter (gregor.lauter@gmx.de)

The University of Birmingham

Covalent organic frameworks (COFs) are highly ordered porous two- or three-dimensional polymers consisting of regularly connected nodes and linkers and characterised with diverse topologies, compositions, and properties. Voids inside the frameworks can be constructed to the desired dimensions and can accommodate a variety of guests: photons and excitons, electrons and holes, ions and molecules, enabling their uses as catalysts, chemosensors, and nanocarriers for biologically active targets. In this work we focus on capture, transport, storage, and release of drugs in/by COFs – a research avenue promising targeted delayed-release delivery of therapeutics by customisable, responsive, and biocompatible materials. We aim to develop a computational approach for predicting the encapsulation and uptake of drug molecules within the cavities of COFs, leveraging their structural characteristics. Our approach combines force field (FF)-based Molecular Dynamics, Density Functional Theory, and Grand Canonical Monte Carlo simulations. To further analyse the non-covalent interactions (NCIs) between the host COFs and the guest molecules, we are also developing dedicated NCI descriptors and structural fingerprints. Overall, these tools will allow not only the design of novel materials but also the rapid pre-assessment of their efficacy in applications related to the sequestration of small molecules.

Stand № 5

Electronic and optical properties of Metal-organic frameworks (MOFs)

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

Karlsruhe Instute of Technology

A metal-organic framework (MOF) is a material comprised of metal atoms or metallic aggregates, e.g. metal-oxo nodes, with organic linkers that form crystalline networks. Owing to the diversity of the components, a large number of MOFs has been synthesized (> 100.000) or postulated (> 1 million). So far, due to their big pore size and their ability to interact with other molecules, MOFs have been used mainly for gas adsorption, separation, catalysis, gas storage, etc. The existing synthesis routes for MOF yield powdery products that have limited applications in optoelectronics, photonics or energy harvesting. Yet, recent reports point to promising electronic properties that originate from the specific assembly of organic linkers in the material. The existing limitations of MOF were partly overcome by Wöll and coworkers at KIT, who can synthesize highly oriented MOF supported by a substrate, called SURMOF (surface-anchored MOF). Designed assembly and chemical composition of organic linkers incorporated in a SURMOF allows tuning of the electronic properties of MOFs to unprecedented accuracy. The application of computational tools, in my host group among others, has allowed a better understanding of new phenomena and behavior of the metal-organic frameworks. However, the systematic study of new MOFs and their electronic and optical properties induced by different organic linkers is still necessary. This study, with the special focus on SURMOF-2 structures, will be centered in the research of new SURMOF materials by a multiscale computational approach and study the electronic and optical properties resulting from the specific assembly of the functional organic linkers. We will apply multiscale modeling techniques to understand the dynamic behavior of MOFs and properties at finite temperature.

Stand № 6

Elucidating the exciton transfer mechanism in LHCII through ML

Ebru Akkus (ebru.akkus@kit.edu)

Karlsruhe Institute of Technology

A deep understanding of the exciton transfer mechanism in the major antenna trimer Light Harvesting Complex II (LHCII) is necessary to comprehend the complex mechanism of photosynthesis in plants. For this, we will simulate LHCII classically and perform DFTB+ calculations to get crucial informations on the exciton transfer via the system's Hamiltonian which includes excitation energies (site energies) and interactions between chlorophylls (couplings) through the extensive trajectories. To obtain the exciton transfer dynamics, machine learning (ML) techniques will be utilized to speed up the most costly part of the study.

Stand № 7

Estimation of hydrogen atom transfer reaction barriers in peptides by learning full radical potential energy surface

Marlen Neubert (marlen.neubert@kit.edu)

Institute of Theoretical Informatics, KIT

Hydrogen atom transfer is an important step in a wide range of chemical and biological processes such as protein mechanics but its mechanistic pathway is not yet fully understood. Simulating these reaction dynamics poses a challenge for fully classical approaches which is why we use machine-learned potentials to learn a reactive force field in a data-driven way. Since this requires global information of the potential energy surface, we combine efficient data sampling strategies with state of the art graph neural networks. The resulting full radical potential energy surface enables accurate estimations of hydrogen atom transfer reaction barriers in peptides.

Stand № 8

Exciton Transfer Simulation in Light Harvesting Complexes FMO and LH2

David Hoffmann (david.hoffmann@kit.edu)

KIT

In the course of photosynthetic evolution, nature has developed efficient photosynthetic units. Light-harvesting complexes collect and transfer sunlight energy in the form of excitons with extremely high quantum efficiency. The motion of the exciton has been simulated using the non-adiabatic molecular dynamics (NAMD) method of surface hopping. The rate-limiting step, the calculation of the transfer Hamiltonian elements (energies and couplings), was accelerated by machine learning. In this way, a statistically meaningful number of trajectories could be generated to elucidate the dynamics of the exciton.

Stand № 9

Exploration of redox properties in chemical space

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

Heidelberg Institute for Theoretical Studies (Poster and Talk)

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 extensive OMEAD dataset, we achieved the lowest mean absolute error (MAE) among existing approaches reported in the literature, establishing our method as state-of-the-art. Additionally, we introduce ReSolved - our own Dataset of Reduction potentials in different solvents and extend MPNN's capability to generalize across different solvents, broadening its applicability in various chemical environments. Furthermore, we combined our MPNN approach with an evolutionary algorithm to explore the vast chemical space for potential good candidates. This improvement marks a crucial step towards accurately predicting redox potentials in diverse conditions, thereby greatly accelerating the discovery of new catalysts and materials for redox reactions. Ultimately, our method contributes to the development of more efficient and sustainable chemical processes by enabling rapid and versatile evaluations of redox potentials.

Stand № 10

Geometric 3D-Convolutional Neural Network for Discriminating Peptide-HLA Unbinding Kinetics

Matheus Ferraz (matheus@oncoimmunity.com)

NEC OncoImmunity AS (Oslo/Mannheim)

The accurate identification of immunogenic peptides presented by the Human Leukocyte Antigen (HLA) on cell surfaces is crucial for developing effective immunotherapy strategies for cancer and infectious diseases, including vaccines. Traditionally, the focus has been on predicting the binding affinity of a peptide-HLA complex as a key indicator of immunogenicity. However, recent studies indicate that binding stability, measured as residence time (the reciprocal of the dissociation rate constant, k_off), may be a more reliable predictor of a peptide's immunogenic potential. Yet, predicting unbinding kinetics presents significant challenges for multiscale modeling due to the long timescales involved, which surpass the capabilities of traditional molecular dynamics simulations. Moreover, current state-of-the-art sequence-based methods relying on machine learning encounter limitations due to biases in the data and their ability to generalize, particularly for under-researched HLA alleles. We propose that structure-based methods, incorporating physical and geometric principles, could enhance this generalizability in predicting binding kinetics. In this work, we employ structure-based geometric deep learning to integrate physical-chemical properties into volumetric grids. This approach effectively captures local geometric relationships using a convolutional network architecture to predict which peptides quickly dissociate from HLA molecules. To achieve this, we have analyzed peptide-HLA interactions involving both human and pathogen peptides presented by HLA-A*02:01 using detailed residue-level feature types, such as residue and contact features. To augment the data set, Brownian dynamics simulations were employed to generate an ensemble of structures. Our results underscore the potential of deep learning models in predicting unbinding kinetics trends within peptide-HLA systems.

Stand № 11

High-Dimensional Neural Networks as Reactive Potentials for ML/MM Simulations

Lukas Petersen (lukas.petersen@kit.edu)

Karlsruhe Intitute of Technology

High-Dimensional Neural Network Potentials(HDNNPs) have been widely successful in a range of applications. The second generation is still mainly used, but with the additional prediction of partial charges in the third and fourth generation a pathway has been opened to a QM/MM variant of these machine learning potentials. Pure machine learning potentials would include solvent/environment molecules in the training data in an indiscriminate manner, but when it comes to complex environments like proteins, this approach becomes unfeasible. The QM/MM approach would facilitate this change of environment immensely. That's why we employ the QM/MM approach to incorporate information about the environment in the model inputs. The electrostatic potential caused by the MM-zone, typically solvent or protein-backbone, is included to enable the network to take the environment into account, similar to the electrostatic embedding.

Stand № 12

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

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

HITS

When subjecting a protein chain to extreme pulling forces, bonds in the stretched backbone ultimately break. As a most simple assumption, a protein backbone can be considered as a serie of springs each of which carries the same force. 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. We propose a method to do so and show how it works in basic cases. Using this method, we created a large dataset of groups of three amino acids to study how it varies when the amino acids are different and how it changes depending on the neighbors. This information can be used to predict which bond is more likely to break first when external forces pull a larger molecule. We used this dataset to train a Neural Network to predict a rupture in stretched proteins.

Stand № 13

Investigating Ligand Binding of the Guanidine-II Riboswitch

Franziska Eble (franziska.eble@uni-konstanz.de)

Universität Konstanz

A common mechanism of gene expression regulation in bacteria are riboswitches, structured RNAs which can bind small molecules, thus turning expression of downstream genes on or off. We investigated ligand binding of the guanidine-II riboswitch, which forms two ligand binding pockets that are close together in the bound state. This proximity encouraged the synthesis and description of a high-affinity bivalent ligand which is able to bind both pockets simultaneously. In addition to in vivo applications, a bivalent ligand presents opportunities in silico; recent work simulating the aptamer dimer of the guanidine-II riboswitch and a C4-diguanidine ligand was able to show spontaneous ligand binding from a setup with only one side of the ligand bound to the riboswitch and describe conformational states along the binding path. But our understanding of the relationship between RNA conformation and binding probability is still limited. Therefore, we examined local descriptors to gain a more nuanced understanding of ligand binding. With an expanded simulation dataset, we were able to investigate the binding transition, estimate the relative stability of states with the help of hydrogen bonding interactions, and describe the interplay between solvent molecules, ions and the guanidinium group during binding and in the bound state.

Stand № 14

Investigating the interaction between an autophagy-related protein and a C2 domain

Eva Hermann (eva.hermann@uni-konstanz.de)

Universität Konstanz

Autophagy is a conserved degradation pathway for the recycling of cellular materials like aggregated proteins or damaged cell organelles. In this study, the newly discovered interaction between an autophagy-related (ATG) protein and a C2-domain protein is investigated by applying bioinformatics and computational chemistry methods. Possible conformations of the protein complex were predicted with deep-learning and docking approaches. In many of the best-ranked predicted models, variable loops of the C2-domain protein interact with the ATG protein close to conserved binding pockets. Subsequently, molecular dynamics simulations were run starting from different complex structures to evaluate the stability and occurrence of different protein-protein conformations and contact interfaces. Recurring states were extracted by dimensionality reduction into a low-dimensional map with the help of graph-based featurization. The characteristic interactions of the so-obtained clusters were analyzed. In conclusion, analysis indicates that loops of the C2-domain protein likely interact with the ATG protein close to conserved binding pockets.

Stand № 15

Investigation of effect of loop motion on the residence time of histamine-1-receptor antagonists by τRAMD

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

Heidelberg Institute for Theoretical Studies (HITS)

Over a third of all clinical drugs target G protein-coupled receptors (GPCRs). Histamine-1-receptor (H1R) is a member of the GPCR superfamily and is a validated target for the treatment of allergies and some forms of gastric acid related conditions. Recent research demonstrates that binding kinetic parameters, i.e. dissociation and association rate constants, can be more important than binding affinity as predictors of GPCR drug efficacy in non-equilibrium in vivo conditions. Thus, there is a growing interest in developing methods for computing the drug-target dissociation rate or its inverse - residence time (τ). Computing protein-drug dissociation rates is challenging because the dissociation occurs over a wide range of timescales, generally extending beyond those accessible to conventional molecular dynamics (MD) simulations. One method for computing the relative residence time of a drug bound to a protein is the τRAMD method which is based on MD and enhanced sampling by Random Acceleration Molecular Dynamics (RAMD). For some protein-ligand complexes, residence time is affected by the relatively slow motion of flexible loops. We modified the τRAMD protocol to account for the motion of a flexible loop and used it to compute the relative residence times of a set of antagonists of H1R. We find that the modified τRAMD protocol gives a better correlation of the computed residence times of the H1R antagonists with experimental values when compared to the standard τRAMD protocol. The τRAMD calculations also provide mechanistic insights into the H1R-ligand dissociation kinetics.

Stand № 16

KineticNet: Deep learning a transferable kinetic energy functional for orbital-free density functional theory

Roman Remme (roman.remme@iwr.uni-heidelberg.de) (Poster and Talk) and
Tobias Kaczun (tobias.kaczun@iwr.uni-heidelberg.de)

IWR Heidelberg

Orbital-free density functional theory (OF-DFT) holds promise to compute ground state molecular properties at minimal cost. However, it has been held back by our inability to compute the kinetic energy as a functional of electron density alone. Here, we set out to learn the kinetic energy functional from ground truth provided by the more expensive Kohn–Sham density functional theory. Such learning is confronted with two key challenges: Giving the model sufficient expressivity and spatial context while limiting the memory footprint to afford computations on a GPU and creating a sufficiently broad distribution of training data to enable iterative density optimization even when starting from a poor initial guess. In response, we introduce KineticNet, an equivariant deep neural network architecture based on point convolutions adapted to the prediction of quantities on molecular quadrature grids. Important contributions include convolution filters with sufficient spatial resolution in the vicinity of nuclear cusp, an atom-centric sparse but expressive architecture that relays information across multiple bond lengths, and a new strategy to generate varied training data by finding ground state densities in the face of perturbations by a random external potential. KineticNet achieves, for the first time, chemical accuracy of the learned functionals across input densities and geometries of tiny molecules. For two-electron systems, we additionally demonstrate OF-DFT density optimization with chemical accuracy.

Stand № 17

Machine Learning assisted prediction of Free Energy Surfaces of Phosphorylation Reactions

Dominik Hachenthal (uxqck@student.kit.edu)

Karlsruhe Institute of Technology

Phosphorylation reactions are quintessential for understanding life on a molecular level. They appear in numerous system like histidine kinase reactions. Our neural network approach within a delta-learning scheme learns and predicts on a variety of phosphorylation reactions in order to predict free energy surfaces.

Stand № 18

Modeling Exciton Transport in Organic Semi-conductors Using Machine Learned Hamiltonian and its Gradients

Dr. Farhad Ghalami (farhad.ghalami@kit.edu)

Karlsruhe Institute of Technology (KIT)

In this study, we present a multiscale method to simulate the propagation of Frenkel singlet excitons in Organic Semi-conductors(OSCs). The approach uses neural network models to train Frenkel-type Hamiltonian and its gradient, obtained by the Long-Range Correction version of Density Functional Tight-Binding with Self-Consistent Charges (LC-DFTB2). Our models accurately predict site energies, excitonic couplings, and corresponding gradients, essential for the non-adiabatic molecular dynamics simulations. Combined with Fewest Switches Surface Hopping (FSSH) algorithm, the method was applied to four representative OSCs: Anthracene (ANT), Pentacene (PEN), Perylenediimide (PDI), and Diindenoperylene (DIP). The simulated exciton diffusion constants align well with experimental and reported theoretical values, and offer valuable insights into exciton dynamics in OSCs.

Stand № 19

Multi-scale Simulation of Charge Transfer in Organic Semiconductors Using Trajectory Surface Hopping Approach

Sonali Garg (sonali.garg@kit.edu)

Karlsruhe Institute of Technology

In recent years, trajectory surface hopping methods have become increasingly popular for studying charge and exciton transport in organic semiconductors (OSCs). Our study focuses on investigation of the charge transport in different class of OSCs using non-adiabatic molecular dynamics simulation methods, particularly on the performance of the Fewest Switches Surface Hopping (FSSH) framework. Additionally, we utilize a fragment orbital approach to model the charge transfer Hamiltonian where, Density Functional Tight Bonding (DFTB) method used to estimate the Hamiltonian elements. It has been shown that, DFTB is 2-3 orders of magnitude faster than Density functional Theory (DFT) method1. We have computed the charge mobilities of different OSCs with a specific focus on π-extended carbazole based systems and halogenated tetraazaperopyrenes (TAPPs). In TAPPs we examined the effect of halogen groups on the charge transport properties. Through our comprehensive analysis, we contribute to the identification of efficient methods for calculating charge transport mobilities in OSCs, providing valuable insights for advancing the field of organic semiconductor research.

Stand № 20

Navigating Chemical Space: An Active Learning Strategy Using Multi-Level Coarse-Graining

Luis Walter (walter@thphys.uni-heidelberg.de)

Institute for Theoretical Physics, University Heidelberg

The chemical compound space is vast and therefore challenging to explore. To navigate this space, conventional methods directly screen extensive compound libraries. However, the size of the library or the resources available for screening limit this strategy. We explore a different approach to chemical space exploration using a multi-level coarse-graining method. Our method is based on the idea that transferable coarse-grained models can compress chemical space into different resolutions. Each level of coarse-graining represents the same region of chemical space but with varying levels of detail and therefore different numbers of total possible molecules. To discover target compounds, we first use an autoencoder to transform the discrete molecular space into a continuous latent space. We thus generate one latent space representation for each level of resolution. Within these latent spaces, we run a multi-level Bayesian optimization-based active learning cycle. In contrast to most multi-fidelity Bayesian optimization studies, we explore this approach for varying resolutions rather than varying runtime and accuracy. For the active learning cycle, we use molecular dynamics simulations to calculate binding free energies for the coarse-grained compounds. We use lower fidelity levels for the chemical space exploration and higher fidelity levels for the exploitation phase of our compound optimization. We investigate challenges related to latent space structure and the efficiency of the multi-fidelity approach.

Stand № 21

Navigating protein landscapes with a machine-learned transferable coarse-grained model

Dr. Nicholas Charron (charron.nicholas.e@gmail.com)

Zuse Institute Berlin

The most popular and universally predictive protein simulation models employ all-atom molecular dynamics (MD), but they come at extreme computational cost. The development of a universal, computationally efficient coarse-grained (CG) model with similar prediction performance has been a long-standing challenge. By combining recent deep learning methods with a large and diverse training set of all-atom protein simulations, we here develop a bottom-up CG force field with chemical transferability, which can be used for extrapolative molecular dynamics on new sequences not used during model parametrization. We demonstrate that the model successfully predicts folded structures, intermediates, metastable folded and unfolded basins, and the fluctuations of intrinsically disordered proteins while it is several orders of magnitude faster than an all-atom model. This showcases the feasibility of a universal and computationally efficient machine-learned CG model for proteins.

Stand № 22

Parameterizing the intrinsically disordered glycoprotein lubricin for the Martini 3 coarse-grained force field

Christina Goss (christina.goss@kit.edu)

HITS

Lubricin, an intrinsically disordered glycoprotein, plays a pivotal role in facilitating the low-friction in boundary lubrication of synovial joints. Consisting of two globular end domains and a mucin-like, disordered central domain, lubricins tripartite structure is known to be essential for its lubricating function. Notably the 11 different O-glycans that glycosylate the serine, threonine and proline rich central domain are necessary for lubricins low-friction behaviour. However, a comprehensive understanding of the contributions of all three domains is lacking. With approximately 1400 amino acids and 200 O-glycans, modeling complete lubricin proteins at an all-atom scale poses significant challenges. To address this, we parameterize the O-glycans for the Martini 3 coarse-grained force field, enabling a computational exploration of lubricins low-friction properties.

Stand № 23

Polytope Fusion: A Novel Approach to Multi-Objective Generative Molecular Design

Mohammad Sajjad Ghaemi (MohammadSajjad.Ghaemi@nrc-cnrc.gc.ca)

National Research Council Canada (NRC)

In the domain of molecular design, the search for optimal compounds with desired properties has long been a challenge. Leveraging the advancements in artificial intelligence (AI), particularly generative models, presents a promising perspective for addressing this challenge. We propose a novel approach to molecular design by integrating a multi-objective optimization framework with convex optimization techniques. Encoding each molecular design objective as a distinct polytope is central to the Polytope Fusion method. These polytopes, representing the multi-dimensional space of molecular properties, are fused to characterize a unified search space, providing a holistic view of the design landscape. By employing convex optimization within this framework, we aim to efficiently navigate the complex landscape of molecular space to identify structures that simultaneously satisfy multiple objectives. One of the critical properties of Polytope Fusion lies in its ability to resolve conflicting objectives inherent in molecular design tasks. By leveraging the intersecting regions of the polytopes, our method identifies solutions that strike an optimal balance between competing criteria, thus unlocking novel design possibilities that were previously unattainable. As such, convex optimization ensures convergence to optimal solutions, enhancing the reliability and robustness of the generated molecular candidates. Moreover, Polytope Fusion transcends traditional boundaries by offering a scalable and adaptable framework suitable for various applications beyond molecular design, including drug discovery, and materials science. This novel approach to multi-objective optimization opens new avenues for innovation and discovery, empowering researchers and practitioners to push the boundaries of molecular design. Furthermore, we discuss potential applications of our approach in drug discovery, materials science, and other domains requiring tailored molecular designs. Overall, our proposed multi-objective optimization framework offers a systematic and efficient approach to generative AI-based molecular design, providing researchers and practitioners with a valuable tool for accelerating the discovery of novel compounds with desirable properties.

Stand № 24

Predicting Phosphorylation Reactions using Deep Neural Networks: A Multi-Reaction Training Approach

Christian Schmidt (christian.schmidt@kit.edu)

Karlsruhe Institute of Technology

Phosphorylation reactions play a pivotal role in biochemical processes. These reactions can be simulated using a Quantum Mechanics/Molecular Mechanics (QM/MM) approach with Density Functional Tight Binding (DFTB). In this study, we introduce a novel approach where machine learning methods are applied to enhance accuracy. A key aspect of our methodology is the training of these networks on multiple reactions. This multi-reaction training approach allows us to capture a broader understanding of phosphorylation reactions, rather than focusing on a single reaction. The application of neural networks not only increases the accuracy of the predictions but also saves computational power. This approach could potentially contribute to more efficient and accurate biochemical simulations.

Stand № 25

Predicting Radical Migration in Collagen

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

HITS

Forces in collagen can lead to homolytic bond breakage, and thus radical formation. Previous ESR experiments have shown that those radicals often end up at dihydroxy-phenylalanine (DOPA) groups, but the exact reaction pathway through collagen is unknown. In this work, we want to examine the possible reactions the radical can undergo before ending up at DOPA. Besides hydrolysis, hydrogen atom transfer (HAT) is the most common type of reaction expected and is the focus of this work. To model the reaction pathway, graph neural networks are employed in order to predict activation energy barriers. For training the networks, training sets of structures just before a HAT reaction together with their associated energy barrier are crafted and used in a supervised learning scheme. The energy barriers are obtained using density functional theory (DFT) on structures along an estimated reaction coordinate. To correct this estimated reaction coordinate, DFT optimized structures are calculated on a subset of systems. These make up an additional, higher quality dataset, which is used in a transfer learning scheme together with the former dataset.

Stand № 26

Predicting the biology of small molecules via an integrated structural and chemoinformatic approach

Elena Frasnetti (elena.frasnetti01@universitadipavia.it)

University of Pavia

This work is focused on different Machine Learning (ML) classifiers that aim to distinguish between potential orthosteric and allosteric binders. Our approach integrates information on the chemical fingerprints of the ligands with descriptors that recapitulate ligand effects on protein functional motions. The latter are derived from Molecular Dynamics (MD) simulations of the target protein in complex with orthosteric or allosteric ligands. In this framework, we train and test different ML architectures, which are initially probed on the classification of orthosteric vs. allosteric ligands for Cyclin Dependent Kinases (CDKs). The results demonstrate that different ML methods can successfully partition allosteric vs. orthosteric effectors (although to different degrees). Next, we further test the models with FDA-approved CDK drugs, not included in the original data set, as well as ligands that target other kinases, to test the range of applicability of these models outside of the domain on which they were developed. Overall, the results show that enriching training dataset with chemical-physics based information on protein-ligand dynamic cross-talk can significantly expand the reach and applicability of approaches for the prediction and classification of the mode of action of small molecules.

Stand № 27

Proton Coupled Electron Transfer in Biomimetic Peptides

Katharina Spies (katharina.spies@kit.edu)

Karlsruhe Institute of Technology

Proton Coupled Electron Transfer (PCET) plays an important role in several biological processes involving the oxidation and reduction of aromatic residues such as tyrosine. Electron transfer is mediated along long-range transfer pathways, and insight into the environment that influences the mechanism of the PCET is important for understanding the enzymatic reactions. Since experimental and computational studies of complex protein systems are costly, time-consuming and difficult to manage, biomimetic peptides and proteins are used instead. Two-dimensional multi walker metadynamic simulations were performed with biasing potentials applied to both the proton and electron transfer reaction coordinates to obtain the free energy surface. For the latter, the bias on the atomic charges, coupled-perturbed equations were implemented in the density functional tight binding (DFTB) method. The methodology was tested on two types of biomimetic peptides. For the PSII-inspired β-hairpin peptides, the mechanism of the PCET reaction between a tyrosyl radical and either a histidine, tryptophan, or tyrosine residue was investigated. It was found that the geometry, in which the residues are arranged relative to each other strongly influences the electron transfer. In addition, a maquette of an α-helical radical was studied and we were able to show the influence of the protein and water environment on the transfer mechanism.

Stand № 28

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 № 29

Relativistic effective core potentials for strong magnetic fields and their implementation into the Turbomole program package

Anja Appenzeller (anja.appenzeller@kit.edu)

Karlsruher Institut für Technologie

Heavy elements in quantum chemical computations require the usage of relativistic effective core potentials (RECP) to account for major relativistic effects and to reduce the computational cost. However, investigations in strong magnetic fields lead to a one-component formalism with additional terms accounting for contribution from the magnetic field. Therefore, RECPs for magnetic fields haven't been available yet in quantum chemical program packages, consequently limiting magnetic field methods to lighter elements. The goal is to implement RECPs for magnetic fields within a one-component framework into the Turbomole program package. Firstly, an expansion of the program code was carried out to deal with complex algebra. The implemented method is intrinsically faster than the available two-component methods for magnetic fields. Further, one-electron integrals will be implemented, finally enabling RECPs in magnetic fields. Possible future applications are the computation of heavy elements in weaker magnetic fields, especially in the context of MCD spectra, and the calculations of berry charges.

Stand № 30

Tailoring Denoising Diffusion Probabilistic Models to Stochastic Thermodynamics

Daniel Nagel (nagel@thphys.uni-heidelberg.de)

Institute for Theoretical Physics, Heidelberg University, Heidelberg, Germany

Stochastic thermodynamics has proven useful in understanding the dynamics of complex systems in non-equilibrium states. In particular, entropy production is an important concept within this framework due to its close relation to the Hamiltonian of the system and its significant role in the Crooks fluctuation relation. However, its computational complexity due to its dependence on the estimation of time-dependent probability distributions limits its application to smaller systems. To address these challenges, machine learning techniques have emerged as valuable tools. While many of these techniques, in particular denoising diffusion probabilistic models (DDPMs),[1] are inspired by statistical mechanics, we aim here to strengthen this link. By rewriting the DDPM score in terms of stochastic thermodynamical quantities, we explore the potential to impose physical constraints within the machine learning model, thereby improving learning efficiency while maintaining consistency with nonequilibrium statistical mechanics. Our work represents a step toward an integrated framework that combines the strengths of machine learning and stochastic thermodynamics, offering a new perspective for studying complex systems on a larger scale with greater computational efficiency.

Stand № 31

The Martini Connoisseur

Linus Grünewald (linus.grunewald@rug.nl)

University of Groningen

The Martini Connoisseur, named after the popular coarse-grained force field, is a machine learning-based program that aims to generate mappings of molecules in a fully automated fashion. Traditionally, the construction of Martini mappings and bead assignments for molecules is a challenging task, which often profits from extensive experience. Additionally, the task itself is time-consuming and therefore frequently a bottleneck for automated high-throughput simulation pipelines. Previous protocols such as AutoMartini have focused on small molecules and typically show poor scaling to larger molecules or even polymers. With the Martini Connoisseur, we aim to bridge these gaps, allowing new users an easier entry into the world of Martini as well as enabling high-throughput pipelines that profit from the extensive Martini ecosystem.

Stand № 32

Theoretical Investigation of ethylene dimerization on Ni- and Cr-based NU-1000 MOF.

Nikita Matsokin (nikita.matsokin@kit.edu)

Karlsruhe Institute of Technology

Olefin oligomerization is an important chemical process for converting ethylene to higher C12-C18 hydrocarbons. Several studies have demonstrated the potential application of metalloorganic frameworks (MOFs) containing catalytically active transition metals. For example, MOF catalysts with Nickel showed catalytic activity of ethylene oligomerization with a non-selective mechanism. However, it would be interesting for industry to use transition metals that allow selective ethylene oligomerization. Some studies have shown the potential use of Cr2+ compounds, which have intrinsic selective oligomerization mechanisms. Therefore, we performed a systematic benchmark study at different levels of theory based on periodic and cluster models of the catalytic MOF-based center with Ni+2 and Cr+2 for the ethylene dimerization process as a first part of olefin oligomerization. Results at DFT level are compared with the results of calculations by CASSCF/NEVPT2 and DLPNO-CCSD(T) to determine justified thermodynamic values with “gold standard” methods for multi and single-determinant cases. This study provides a characterization of the catalytic processes on both metals to compare the thermodynamics of the process. These characterizations will be further used to build a kinetic model that can show advantages and disadvantages of using Cr and Ni for catalysis.

Stand № 33

Towards Linear Dimensionality Reduction in QM/MM: Exploring Collective Variables for Complex Systems

Julian Boeser (julian.boeser@kit.edu), Lena Eichinger (lena.eichinger@kit.edu)

Karlsruhe Institut of Technology

Conformational transitions are often connected to the chemical step being catalyzed and are highly coupled to enzymatic activity. However, describing these complex transitions with enhanced-sampling methods can be challenging due to the difficulty in identifying appropriate collective variables (CVs) that accurately capture the relevant degrees of freedom. In our work, we frequently encounter the interplay between different timescales, presenting numerous options for describing these dynamics and providing rich data for analysis. By employing machine learning (ML) algorithms, we aim to identify CVs that effectively guide enhanced-sampling simulations, such as metadynamics, facilitating efficient exploration of the intricate conformational landscapes. Beginning with dimensionality reduction, we aim to streamline our understanding of enzymatic conformational transitions. As a case study, we examine the conformational transition in histidine kinase, a crucial enzyme in the signalling system of bacteria, and in the catalytic disulfide shuffling reaction of glutaredoxin.

Stand № 34

TOWARDS PEPTIDE-BASED THERAPEUTICS AGAINST CARDIAC DISEASE - Prediction and simulation of the S100A1ct peptide with a membrane environment

Tommaso Bartoloni (tommaso.bartoloni@h-its.org)

HITS

S100A1ct is a peptide composed of an N-terminal tag and the C-terminal portion of the S100A1 Ca2+-binding protein. Both in-vitro and in-vivo experiments show the peptide to be a candidate for the treatment of heart failure. The peptide, was docked to SERCA2a and predicted to act as a transmembrane helix. This work aims to characterize stable conformations of S100A1ct in a simple membrane environment, to assess the possibility of a transmembrane state, and understand the mechanism of peptide insertion in a phospholipid bilayer, through the use of conventional molecular dynamics (MD) simulations and Gaussian-accelerated MD simulations. Such simulations showed two major S100A1ct conformations in the membrane: the peptide sitting in a single membrane leaflet was observed as the most favourable arrangement, while, when restricting the analysis to just the trajectories featuring a transmembrane conformation, a strong helicity was observed in the most hydrophobic section of the peptide. Another set of simulations showed a mechanism of insertion relying on the transient displacement of the phosphate heads, exposing hydrophobic patches into which some residues can dive and drive the insertion process.

Stand № 35

Towards predictors of electron transfer rates in the Cytochrome P450 family

Jonathan Teuffel (jonathan.teuffel@h-its.org)

Heidelberg Institute for Theoretical Studies

J. Teuffel (1,2,3), S. B. Han (1), S. Ber (1), G. Mukherjee (1,4), R.C. Wade (1,2,4,5) Members of the cytochrome P450 (CYP) family of proteins catalyse many important biochemical reactions and determine the metabolization rate of most drugs. Their net reaction requires the supply of external electrons from secondary redox proteins. Electron transfer (ET) between them takes place over large distances between cofactors making it a slow process. These slow rates render ET a bottleneck in the overall CYP-reaction, and it has been demonstrated to be rate-determining for some CYPs. The ET-rates and the interactions with redox proteins differ significantly between members of the CYP superfamily. Thus, we aim to investigate how sequence and structural features of the individual CYPs and their redox proteins affect the respective ET-rates. A multiscale simulation protocol has been developed for this purpose, ranging from Brownian- and molecular dynamics simulations down to quantum mechanics. Combining these methods allows us to model how different class II CYPs interact with their membrane environment and the reductase, and to compute system-specific ET-rates. Here, we present the application of this protocol to investigate structure-kinetics relationships for several CYP-reductase membrane-bound protein-protein complexes. The results of our simulations enable the effects on ET rates of residues at specific positions in the protein sequence to be evaluated.

Stand № 36

Trans-dimensional generative modelling for molecular liquids

Balint Mate (balint.mate@unige.ch)

University of Geneva/University of Heidelberg

We propose to use a trans-dimensional extension of denoising diffusion models to model the equilibrium distribution of molecular liquids.

Stand № 37

Understanding ProteinMPNN: in-depth analysis of a deep learning-based method for protein design

Vsevolod Viliuga (vsevolod.viliuga@gmail.com)

HITS

Over the past two years, the protein design field has undergone a revolution. With the development of powerful deep learning-based methods for structure prediction, such as AlphaFold2 (AF2), a myriad of differently architectured neural networks for protein backbone and sequence generation were released. A particularly popular tool is the message-passing neural network (MPNN) for protein sequence design called ProteinMPNN. Although ProteinMPNN significantly outperformed state-of-the-art competitors in generating sequences which adopt the desired protein fold, as predicted by AF2, ProteinMPNN-internal scoring metrics, along with the confidence score of AF2 (pLDDT), do not always correlate well with experimental success rate. Thus, every design campaign usually requires laborious experimental testing of many designed sequences ranging from a few hundreds to several thousands. Furthermore, although ProteinMPNN was shown to yield natural proteins and enzymes that are more thermostable and soluble, it remains enigmatic which are the precise properties the algorithm alters that cause these effects. We sought to dissect changes introduced by ProteinMPNN on both sequence and structure level upon redesigning poorly expressed natural enzymes with biotechnological relevance, and developed a computational workflow for automated analysis of biophysical properties of the designed variants. We derived a set of metrics that can be potentially used to select optimized variants. Relying on our metrics, we found recurring physico-chemical properties in the redesigned sequences compared to their wild-type versions. Moreover, we found significant differences in ProteinMPNN's performance depending on model weights and input sequence. We believe these findings will aid in illuminating the AI black box and lead to a better anticipation of advantages and limitations of ProteinMPNN.

Stand № 38

Variously Crosslinked Coarse Grained Collagen Fibrils Under Force

Johanna Buck (Johanna.buck@h-its.org)

HITS

Collagen is a prevalent protein in the animal world, specially mammals. It is abundant in the connective tissues and subjected to substantial mechanical forces. Cross-links play an essential role in determining the structural and mechanical integrity of collagen while their type and distribution varies with aging. To examine the mechanical properties of collagen, we employed a computational model based on the Martini3 coarse-grained force- field and simulated under force. With this model it is possible to simulate multi-million particle systems exceeding capabilities of atomistic simulations. This study shows an improved parametrization and a fundamental analysis of fibrillar coarse-grained models of collagen which capture similar force responses as the all-atom model.

Stand № 39

Author Index