Workshop “Uncertainty Quantification in Molecular Simulation”

Abstracts

https://indico3.mpi-magdeburg.mpg.de/e/uqms

Contribution ID: **25** Type: **Talk**

Error estimates in molecular dynamics

Thursday, 29 August 2024 13:00 (1 hour)

I will provide a brief introduction to molecular dynamics (the computational implementation of the theory of statistical physics) and relate it to Bayesian inference, as these are two situations where sampling a high dimensional probability measure is required. Average properties for these two applications are typically obtained through ergodic averages of discretizations of certain stochastic differential equations. I will provide an introduction to the most popular stochastic dynamics to this end and their numerical analysis – in particular error estimates on the timestep discretization bias, and estimates on the statistical error.

Primary author: STOLTZ, Gabriel (CERMICS, Ecole des Ponts & Project-team MATHERIALS, Inria Paris)

Presenter: STOLTZ, Gabriel (CERMICS, Ecole des Ponts & Project-team MATHERIALS, Inria Paris)

Contribution ID: **19** Type: **Talk**

Utilizing Cramers-van Mises distance for the Global Sensitivity Analysis of Monte Carlo Models

Thursday, 29 August 2024 14:00 (30 minutes)

An important aspect in the interpretation of a molecular simulation model is to quantify which parameter uncertainties have the most influence on the simulation results. We present an approach to such Global Sensitivity Analysis (GSA) on basis of the Cramers-von Mises distance. Unlike revalent approaches to GSA it combines the following properties: i) it is equally suited for deterministic as well as stochastic model outputs, ii) it is free of gradients, and iii) it can be estimated from (almost) any numerical quadrature. Using Low-Discrepency Sequences for quadrature and a prototypical first-principles kinetic Monte Carlo models, we examine the performance of the approach. We find that the approach converges in a modest number of quadrature points. Furthermore, it seems to be robust against even extreme relative noise. These properties make the method particularly suited for expensive (kinetic) Monte Carlo models, because only a relatively small number of rather inaccurate Monte Carlo estimates is needed.

Primary authors: DORTAJ, Sina (Fritz Haber Institute of the MPS); MATERA, Sebastian (Fritz Haber Institute of the MPS)

Presenter: MATERA, Sebastian (Fritz Haber Institute of the MPS)

Contribution ID: **20** Type: **Talk**

Effective dynamics for discrete-in-time stochastic processes described by transfer operators

Thursday, 29 August 2024 14:30 (30 minutes)

Collective variables (CVs) play an important role in understanding the dynamics of high-dimensional metastable molecular dynamics. Given a set of CVs, effective dynamics of diffusion processes have been constructed using conditional expectations and their properties have been studied in previous works. In this talk, we extend the definition of effective dynamics to discrete-in-time Markov processes X_n. In particular, we show that the transition density of the effective dynamics solves a relative entropy minimization problem from certain family of densities to the transition density of X_n. We also show that many transfer operator-based data-driven numerical approaches essentially learn quantities of the effective dynamics. Finally, we discuss how our theoretical analysis can be converted into a numerical algorithm for identifying CVs. This is a joint work with Christof Schütte.

Primary author: ZHANG, Wei (Zuse Institute Berlin)

Co-author: SCHUETTE, Christof (Zuse Institute Berlin)

Presenter: ZHANG, Wei (Zuse Institute Berlin)

Contribution ID: **26**

Type: **Talk**

Bayesian Inverse Problems and Their Use in Molecular Dynamics

Thursday, 29 August 2024 15:30 (1 hour)

First, we will briefly introduce the basics of statistical inversion, where, in its most basic form, the goal is to study how to estimate model parameters from data. We will introduce mathematical concepts and computational tools for systematically treating these inverse problems in a Bayesian framework, including assessing how uncertainties affect the solution. In the second part, we will discuss an exemplary application of the Bayesian framework, which is pertinent to molecular dynamics. Specifically, we consider Bayesian inference for diffusion processes based on data available as (a collection of) stochastic trajectories or statistics like expectation values or probability distributions.

Primary author: KRUMSCHEID, Sebastian (KIT)

Presenter: KRUMSCHEID, Sebastian (KIT)

Contribution ID: **17**

Type: **Talk**

Evaluating uncertainty estimations of Gaussian process regression-based machine learning interatomic potentials

Thursday, 29 August 2024 16:30 (30 minutes)

Machine learning interatomic potentials (MLIPs) are machine learning (ML) models that map molecular configurations to corresponding energies and potentially forces, replacing highly accurate but expensive quantum chemical calculations. Quantum chemical calculations are carried out to calculate reference energies for only a few molecular configurations. When the ML model predicts an unseen sample, it introduces a new error, which highly depends on how different the sample is from the training data. Therefore, it is desirable to provide an uncertainty estimation for every model prediction. On the one hand, it is important to know the range of magnitude of a potential error of a prediction; on the other hand, it can be utilized in active learning (AL) schemes to identify samples for which the model is uncertain and add them as training samples. For MLIPs based on Gaussian process regression (GPR), the most commonly used uncertainty estimation is the variance of the predictive distribution provided by the GPR model. Besides that, ensemble-based uncertainties are possible. Although these GPR uncertainty measures have been applied to AL, it has rarely been studied how well or whether they correlate with the actual error, and it is not always clear whether AL strategies provide an actual improvement over simply adding samples randomly. In our study, we consider GPR models with Coulomb representations as well as Smooth Overlap of Atomic Positions (SOAP) representations as inputs. Our aim is to evaluate how the GPR variance and ensemble-based uncertainties correlate with the actual error. Furthermore, we want to find out how much additional information can be extracted from a pool of candidate samples, compared to selecting samples randomly, when we utilize the different uncertainty measures to iteratively add the most uncertain samples as training samples. This AL scheme is called uncertainty sampling. For different datasets and uncertainty estimations, we observed substantial differences in how error and uncertainties correlate. We find that, no matter if the correlation of an uncertainty estimation closely aligns with the theoretical expectations or not, the amount of additional information that can be extracted from the candidate pool via uncertainty sampling is highly limited and often close to zero.

Primary author: HOLZENKAMP, Matthias (Uni Wuppertal)

Co-author: ZASPEL, Peter (Uni Wuppertal)

Presenter: HOLZENKAMP, Matthias (Uni Wuppertal)

Contribution ID: **18** Type: **Talk**

Molecular simulations and experiments for the diffusion of hydrogen in brine

Thursday, 29 August 2024 17:00 (30 minutes)

The global shift towards carbon-neutral energy systems has heightened the need for efficient and secure storage solutions for renewable energies, with hydrogen (H_2) storage in deep saline aquifers emerging as a viable option for large-scale storage with high flexibility in terms of the number of annual storage cycles and the stored gas volume. This study aims to advance our understanding of hydrogen storage by investigating the diffusion coefficient of pure hydrogen and its mixtures with carbon dioxide in brine. This study uses experimental conditions similar to the Ketzin site in Brandenburg, Germany.

Molecular dynamics (MD) simulations were employed to predict the diffusion coefficients of hydrogen in chloride brine containing various cations (Na^+, K^+, Ca^{2+}) under a range of conditions (pressures from 1 to 218 atm and temperatures from 298 to 648 K). These simulations demonstrated that hydrogen diffusivity is significantly influenced by temperature, pressure, and ionic composition. The Arrhenius behavior observed for temperature dependence showed limitations at higher temperatures (≥ 400 K), indicating the need for more complex modeling approaches. The study utilized the TIP4P/2005 model for water, a two-site model for hydrogen, and the OPLS model for chloride ions, with simulations performed in the LAMMPS software. Experimental investigations complemented the simulations, employing a dual-chamber system to measure hydrogen diffusion through various rock samples, including Bentheimer sandstone, Werra rock salt, and Opalinus clay. The measured diffusion coefficients ranged from 10^{-8} to 10^{-9} m^2/s, with wetted samples showing greater hydrogen retention due to pore water saturation and microcrack closure from recrystallization in rock salt.

Our finding highlights the essential role of MD simulations in providing detailed insights into hydrogen diffusion in brine for a wide range of operating conditions, significantly contributing to the understanding and development of hydrogen storage mechanisms in geological formations. Integrating molecular simulations with experimental data offers a robust validation of the computational models in predicting hydrogen diffusion in saline aquifers.

Primary author: SINGH, Mrityunjay (GFZ Potsdam)

Co-authors: STRAUCH, Bettina (GFZ Potsdam); SCHMIDT-HATTENBERGER, Cornelia (GFZ Potsdam); SASS, Ingo (GFZ Potsdam); ZIMMER, Martin (GFZ Potsdam); PILZ, Peter (GFZ Potsdam)

Presenter: SINGH, Mrityunjay (GFZ Potsdam)

Contribution ID: **27** Type: **Talk**

Misspecification uncertainty in deterministic surrogate models

Friday, 30 August 2024 09:00 (1 hour)

Surrogate models approximate the action of expensive scientific calculations, saving time, energy and cost. Surrogate model parameters are typically determined by minimising the negative log likelihood, or empirical loss. However, as the loss ignores model misspecification, Bayesian parameter uncertainties are largely epistemic and thus severe underestimates, vanishing in the large-data (over-parametrised) limit. A “true”Bayesian regression scheme should minimise the generalisation error, for which the expected loss is a Gibbs-Bogoliubov-Jensen upper bound. Whilst typically intractable, for the special case of deterministic calculations (no aleatoric uncertainty), we derive a condition any minimiser of the generalisation error must obey and design a simple ansatz which can be variational minimised[1]. The final result gives provably superior generalisation ability over Bayesian regression, and is extremely efficient in both training and evaluation for high dimensional linear models, giving accurate prediction and very useful bounding of test errors. Importantly, model prediction errors are directly related to model parameter uncertainties, essential to capture the correlations present when propagating uncertainty through multi-scale simulations[2].

[1] https://arxiv.org/abs/2402.01810v3 (with Danny Perez, LANL)
[2] https://arxiv.org/abs/2407.02414 (with Ivan Maliyov and Petr Grigorev, CINaM/CNRS)

Primary author: SWINBURNE D., Thomas (CNRS)

Presenter: SWINBURNE D., Thomas (CNRS)

Contribution ID: **21**

Type: **Talk**

Implicit differentiation of atomic minima for uncertainty quantification and inverse problems

Friday, 30 August 2024 10:00 (30 minutes)

Interatomic potentials are essential to go beyond *ab initio* size limitations, but simulation results depend sensitively on potential parameters. Parameter dependence is typically explored through repeating simulations with resampled parameters, a forward-only approach that becomes prohibitively expensive for the high parameter dimension of modern machine learning potentials.

In this talk, I will present an analytical scheme to forward- and back-propagate potential parameter variation through energy minimization. This is achieved via the implicit derivative of an implicit function defined at a fixed point, such as an energy minimum. The implicit derivative gives an analytic expansion of minimum energy and structure as functions of potential parameters, used for high-throughput uncertainty quantification in forward propagation and solution of challenging inverse problems in backpropagation.

I will discuss how the implicit derivative can be efficiently evaluated for large atomic systems, in particular, using a sparse operator approach to compute the implicit derivative implemented in both automatic differentiation (AD) and non-AD frameworks. Our implementation in the LAMMPS code has minimal memory usage and excellent scalability, allowing implicit derivative evaluation for systems of arbitrary size[1].

I will show how the implicit derivative can be used to expand the scope of atomic simulation methods. In forward propagation, the implicit expansion has sufficient accuracy to replace thousands of energy minimizations with a single calculation. This enables high-throughput uncertainty quantification and exploration of model phenomenology that is not possible with existing methods. In backpropagation, the implicit derivative allows us to 'fine-tune'interatomic potentials and target subtle solute-induced defect reconstruction, a key feature in understanding plasticity and irradiation damage in bcc metals.

[1] I. Maliyov, P. Grigorev, T.D. Swinburne, arXiv:2407.02414, 2024

Primary authors: MALIYOV, Ivan (Aix-Marseille Université, CNRS, CINaM, France); GRIGOREV, Petr (Aix-Marseille Université, CNRS, CINaM, France); SWINBURNE, Thomas (Aix-Marseille Université, CNRS, CINaM, France)

Presenter: MALIYOV, Ivan (Aix-Marseille Université, CNRS, CINaM, France)

Contribution ID: **22**

Type: **Talk**

Shallow Ensembles: Simple, Scalable and Size-Extensive Uncertainty Estimates for Atomistic Modelling

Friday, 30 August 2024 11:00 (30 minutes)

Statistical learning algorithms provide a generally-applicable framework to sidestep time-consuming experiments, or accurate physics-based modeling, but they introduce a further source of error on top of the intrinsic limitations of the experimental or theoretical setup. Uncertainty estimation is essential to quantify this error, and make application of data-centric approaches more trustworthy. To ensure that uncertainty quantification is used widely, one should aim for algorithms that are reasonably accurate, but also easy to implement and apply. In particular, including uncertainty quantification on top of an existing model should be straightforward, and add minimal computational overhead. Furthermore, it should be easy to process the outputs of one or more machine-learning models, propagating uncertainty over further computational steps. We compare several well-established uncertainty quantification frameworks against these requirements, and propose a practical approach, which we dub shallow ensemble propagation, that provides a good compromise between ease of use and accuracy. We present applications to the field of atomistic machine learning for chemistry and materials, which provides striking examples of the importance of using a formulation that allows to propagate errors without making strong assumptions on the correlations between different predictions of the model.

Primary authors: KELLNER, Matthias (EPFL); CERIOTTI, Michele (EPFL)

Presenter: KELLNER, Matthias (EPFL)

Contribution ID: **28**

Type: **Talk**

Learning Kinetically Consistent Coarse Grained Dynamics via Kernel-based Approximation of Koopman Generator

Friday, 30 August 2024 11:30 (30 minutes)

Accurately detecting the coarse-grained coordinates is an essential task in the model discovery of complex systems, such as molecular dynamics. In many cases, such systems cannot be approximated properly with deterministic dynamics. An extension of Extended Dynamic Mode Decomposition (EDMD) has been introduced in [Klus et al., Physica D (2020)] to approximate the Koopman generator for the identification of stochastic dynamical systems. However, the selection of basis functions upon which the generator is approximated is a non-trivial task and needs to be done manually. By taking advantage of kernel methods introduced in [Klus, Entropy (2020)], we develop a kernel-based data-driven method to approximate the Koompan generator of a dynamical system. This method allows us to identify stochastic differential equations governing the coarse-grained model of a high-dimensional system. Dominant dynamics and metastabilities of the system in the reduced-order space, furthermore, can be obtained by the eigen-decomposition of the coarse-grained generator.

We numerically investigate the method using toy models and molecular dynamics problems and demonstrate that the results are thermodynamically and kinetically consistent with the full model.

Presenter: NATEGHI, Vahid

Contribution ID: **23** Type: **Poster**

Shallow ensembles for practical uncertainty quantification in equivariant neural network potentials

Machine-learning interatomic potentials (MLPs) efficiently approximate the potential energy surface given by an underlying first-principles method, for example density-functional theory, enabling simulations at previously unachievable time and length scales. However, for practical applications, it is critical to know when predictions can be relied upon, and when additional data is needed. We evaluate the recently proposed shallow ensembles [1] approach to uncertainty quantification for the state-of-the-art So3krates MLP [2]. As a prototypical usecase, we choose molecular dynamics simulations of CuI, where dynamic formation of interstitials prevents ahead-of-time training.

1: Kellner and Ceriotti, Mach. Learn. Sci. Technol. **5** (2024)

2: Frank, Unke, Müller, and Chmiela, arXiv:2309.15126 (2024)

Primary author: LANGER, Marcel F. (EPFL)

Co-authors: KNOOP, Florian; KELLNER, Matthias (EPFL)

Presenter: LANGER, Marcel F. (EPFL)

Contribution ID: **24** Type: **Poster**

Predicting hydrogen atom transfer energy barriers using Gaussian process regression

Predicting reaction barriers for arbitrary atomic configurations based on only a limited set of density functional theory (DFT) calculations would render the simulation of reactions within complex materials highly efficient. We propose Gaussian process regression (GPR) as a method of choice if DFT calculations are limited to hundreds or thousands of barrier calculations. For the case of hydrogen atom transfer (HAT), we obtain a mean absolute error of 3.23 kcal/mol using SOAP descriptors. We assess the uncertainty of HAT barrier predictions using the predictive distributions obtained directly from GPR as well as from an ensemble of a graph neural network-based model. Especially in the low-data regime, we find that GPR outperforms the latter with respect to various proper scoring rules. We suggest GPR as a valuable tool for an approximate but data-efficient model of chemical reactivity in a complex and highly variable environment.

Primary author: ULANOV, Evgeni (Heidelberg Institute for Theoretical Studies)

Co-authors: QADIR, Ghulam A. (Heidelberg Institute for Theoretical Studies); RIEDMILLER, Kai (Heidelberg Institute for Theoretical Studies); FRIEDERICH, Pascal (Karlsruhe Institute of Technology); GRÄTER, Frauke (Heidelberg Institute for Theoretical Studies, Heidelberg University)

Presenter: ULANOV, Evgeni (Heidelberg Institute for Theoretical Studies)