SOE 10: Data Analytics, Extreme Events, Nonlinear Stochastic Systems, and Networks (joint session DY/SOE)

Time: Wednesday 15:00–17:30 Location: ZEU 118

Invited Talk SOE 10.1 Wed 15:00 ZEU 118

I want it all and I want it now! — •Alexander K. Hartmann — University of Oldenburg, Germany

For every random process, all measurable quantities are described comprehensively through their probability distributions. Ideally, they would be obtained analytically, i.e., completely. Since most physical models are not accessible analytically, one has to perform numerical simulations. Usually this means one does many independent runs, allowing one to measure histograms. Since the number of repetitions is limited, maybe 10 million, correspondingly the distributions can be estimated in a range down to probabilities like 10^{-10}. But what if one wants to obtain the full distribution, in the spirit of obtaining all information? Thus, one desires to get the distribution down to the rare events, without waiting for a huge running time.

Here, we study rare events using a very general black-box method [1]. It is based on sampling vectors of random numbers within an artificial finite-temperature (Boltzmann) ensemble to access rare events and large deviations for almost arbitrary equilibrium and non-equilibrium processes. In this way, we obtain probabilities as small as 10^{-500} and smaller, hence (almost) the full distribution can be obtained in a reasonable amount of time. Examples are presented for applications to random graphs [2], traffic flow models, biological sequence alignment, particle diffusion, or calculation of partition functions [3].

[1] A.K. Hartmann, Phys. Rev. E **89**, 052103 (2014)
[2] A.K. Hartmann and M. Mézard, Phys. Rev. E **97**, 032128 (2018)
[3] A.K. Hartmann, Phys. Rev. Lett. **94**, 050601 (2005)

SOE 10.2 Wed 15:30 ZEU 118

Constructing accurate and data-efficient molecular force-fields with machine learning — •Igor Poltavskyi, Grégory Fonseca, Valentin Vassilev-Galindo, and Alexandre Tkatchenko — University of Luxembourg, Luxembourg

Employing machine learning (ML) force-fields (FF) is becoming a standard tool in modern computational physics and chemistry. Reproducing potential energy surfaces of any complexity, ML models extend our horizons far beyond the reach of *ab initio* calculations. One can already perform nanosecond-long molecular dynamics simulations for molecules containing up to a few tens of atoms on a coupled-cluster level of accuracy, providing invaluable information about subtle details of intra-molecular interactions [1,2]. Next challenges are constructing ML FFs to molecules with 1000s of atoms and describing far-from-equilibrium geometries without losing accuracy and efficiency. To reach these goals, we developed methods for optimizing reference datasets and partitioning the problem of training global FFs into parts. By minimizing the prediction error for subsets of molecular configurations obtained by clustering, we can build ML FFs equally applicable for the entire range of reference data. Dividing the configuration space into sub-domains by physical and chemical properties, training corresponding ML models, and combining them into one global model enables highly-accurate FFs for molecules containing hundreds of atoms.

[1] Sauceda *et al.*, J. Chem. Phys. **150**, 114102 (2019).
[2] Chmiela *et al.*, Nat. Commun., **9**(1), 3887 (2018).

SOE 10.3 Wed 15:45 ZEU 118

Interpretable Embeddings from Molecular Simulations Using Gaussian Mixture Variational Autoencoders — •Yasemin Bozkurt Varolgunes[1,2], Tristan Bereau[1], and Joseph F. Rudzinski[1] — [1]Max Planck Institute for Polymer Research, Mainz, Germany — [2]Koc University, Istanbul, Turkey

Extracting insight from the molecular simulations data requires the identification of a few collective variables (CVs) whose corresponding low-dimensional free-energy landscape (FEL) retains the essential features of the underlying system. Autoencoders are powerful tools for dimensionality reduction, as they naturally force an information bottleneck. While variational autoencoders (VAEs) ensure continuity of the embedding by assuming a Gaussian prior, this is at odds with the multi-basin FELs that typically arise from the identification of meaningful CVs. Here, we incorporate this physical intuition into the prior by employing a Gaussian mixture variational autoencoder (GMVAE), which encourages the separation of metastable states within the embedding. The GMVAE performs dimensionality reduction and clustering within a single unified framework, and is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. We illustrate our approach on two toy models and a peptide, demonstrating the anti-clustering effect of the prior relative to standard VAEs. The resulting embeddings stand as appropriate representations for constructing Markov state models, highlighting the transferability of the dimensionality reduction from static equilibrium properties to dynamics.

SOE 10.4 Wed 16:00 ZEU 118

The entropy of the longest increasing subsequences: typical and extreme sequences — Phil Krabbe[1], •Hendrik Schawe[1,2], and Alexander K. Hartmann[1] — [1]Carl von Ossietzky Universität Oldenburg, Germany — [2]Laboratoire de Physique Théorique et Modèlisation, Université de Cergy-Pontoise, France

Consider a game, where you get a sequence of n numbers. Your objective is to circle the maximum amount of numbers such that each circled number is larger than all circled numbers to their left. To circle the maximum amount numbers, one can calculate the *longest increasing subsequence* (LIS). If the sequence of numbers is a random permutation, this problem is remarkably well studied and for the length L, or in our game the number of circles, not only the mean value, but the whole distribution is known [1,2]. In recent time it was shown that this problem is equivalent to certain surface growth and ballistic deposition models, which led to a large interest from physicists.

Note that the LIS is not unique, there are possibly multiple ways to circle L numbers. While this degeneracy M is expected to increase exponentially with the sequence length n [1], we introduce an algorithm to count the number of degenerate LIS and sample uniformly from all LIS of a given sequence. Especially, we obtain the distribution $P(M)$ down into its far tails with probabilities smaller than 10^{-100} using sophisticated Markov chain sampling methods [3].

[1] D. Romik, The Surprising Mathematics of Longest Increasing Subsequences (2015); [2] J. Börjes, H. Schawe, A. K. Hartmann, Phys. Rev. E 99 (4), 042104 (2019); [3] A.K. Hartmann, EPJB 84, 627 (2011)

15 min. break.

SOE 10.5 Wed 16:30 ZEU 118

Large-deviation simulation of height distribution for the KPZ equation: dependence on initial conditions and morphologie of extreme configurations — •Alexander K. Hartmann[1], Pierre Le Doussal[2], Alexandre Krajenbrink[2], Baruch Meerson[3], and Pavel Sasorov[4] — [1]University of Oldenburg, Germany — [2]Ecole Normale Supérieure, Paris, France — [3]Hebrew University of Jerusalem, Israel — [4]Keldysh Institute of Applied Mathematics, Moscow, Russia

The distribution of relative free energies H of directed polymers in disordered media is studied, which is in the KPZ universality class. We study the distribution at large temperatures, corresponding to short times in KPZ. Using a statistical mechanics-based *large-deviation approach*, the distribution can be obtained over a large range of the support, down to a probability density as small as 10^{-1000} [1]. We compare with analytical predictions for different types of initial conditions and for full as well as for half space [2]. A very good agreement is found for $H < 0$ and a strong convergence is visible for $H > 0$. Furthermore, we study the morphology of atypical fluctuations [3], compare with analytical results from the *optimal fluctuation method*, and find again a good agreement.

[1] A.K. Hartmann, P. Le Doussal, S.N. Majumdar, A. Rosso and G. Schehr, Europhys. Lett. 121, 67004 (2018).
[2] A.K. Hartmann, A. Krajenbrink, and P. Le Doussal, preprint arXiv:1909.0384.
[2] A.K. Hartmann, B. Meerson, and P. Sasorov, arXiv: 1907.05677.

SOE 10.6 Wed 16:45 ZEU 118

Machine Learning on temperature fluctuations in health and disease — •Jens Karschau, Sona Michlíková, Daniel Kotik, Sebastian Starke, Steffen Löck, and Damian McLeod — OncoRay, HZDR, TU Dresden, Dresden, Germany

Rendering disease diagnoses from measurements is a highly complex

task. Clinicians train for many years in order to identify pathological events from patient data. Exemplarily, medical expert knowledge recognises subtle differences between normal and tumor-looking features. Today, machine learning (ML) allows us to support not only the clinical decision maker during classification; it also has potential to promptly warn self-monitored individuals. We developed an RNN model that learns on time series temperature data of up to 120 days to detect cancer features in mice. It successfully bins particular days into either tumor vs. no-tumor days. Using out-of-sample data from the same or a different cohort, the model successfully classifies with an accuracy and AUC of up to 0.80. The dynamic time warping dissimilarity measure applied to different days indicates that oscillation patterns contain distinctive features that the RNN model learns. We hypothesise that the model learns features based on oscillatory behaviour at the 150 min time scale: the so-called 'ultradian' rhythm. The double benefit from our method is: (a) it uses non-invasive measurements to classify the disease state and (b) it could be deployed for applications in future on-line monitoring of data from wearable devices. Our next efforts are testing human data to deliver actionable insights in disease control and decision support.

SOE 10.7 Wed 17:00 ZEU 118

Using data crawling and flexible semantic data models to enable sustainable research data management — •Alexander Schlemmer[1,2,3], Ulrich Parlitz[1,2,4], and Stefan Luther[1,2,4,5] — [1]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany — [2]German Center for Cardiovascular Research (DZHK), Partner Site Göttingen, Germany — [3]IndiScale GmbH, Göttingen, Germany — [4]Institute for the Dynamics of Complex Systems, Georg-August-Universität Göttingen, Germany — [5]Institute of Pharmacology and Toxicology, University Medical Center Göttingen, Germany

Despite the significant advances in computation power and information technology of the last decades, scientific data management is still lacking widespread adoption in many scientific communities. The absence of standardized workflows and corresponding tools significantly impedes complete transparency and reproducibility of research results.

We discuss key concepts to remove omnipresent barriers in scientific data management. Specifically, data crawling strategies and flexible semantic data models are highlighted. With examples using our open source software CaosDB (https://doi.org/10.3390/data4020083) we show how these concepts can be practically applied in order to achieve sustainable research data management.

SOE 10.8 Wed 17:15 ZEU 118

Non-Markovian barrier crossing with two-time-scale memory is dominated by the faster memory component — •Julian Kappler, Victor B. Hinrichsen, and Roland R. Netz — Freie Universität Berlin, Fachbereich Physik, Berlin, Germany

We investigate non-Markovian barrier-crossing kinetics of a massive particle in one dimension in the presence of a memory function that is the sum of two exponentials with different memory times. Our Langevin simulations for the special case where both exponentials contribute equally to the total friction show that the barrier-crossing time becomes independent of the longer memory time if at least one of the two memory times is larger than the intrinsic diffusion time. When we associate memory effects with coupled degrees of freedom that are orthogonal to a one-dimensional reaction coordinate, this counterintuitive result shows that the faster orthogonal degrees of freedom dominate barrier-crossing kinetics in the non-Markovian limit and that the slower orthogonal degrees become negligible, quite contrary to the standard time-scale separation assumption. We construct a crossover formula for the barrier crossing time that is valid for general multi-exponential memory kernels. This formula can be used to estimate barrier-crossing times for general memory functions for high friction, i.e. in the overdamped regime, as well as for low friction, i.e. in the inertial regime.