Eugen Hruška, Ph.D.

Hruska-Lab.github.io
Charles University
0000-0001-5679-8419
@HruskaEugen
eugen-hruska
eugen.hruska@faf.cuni.cz

Publications

1 Chen, X., Li, P., **Hruska**, **E.**, & Liu, F. (2023). Δ-machine learning for quantum chemistry prediction of solution-phase molecular properties at the ground and excited states. *Phys. Chem. Chem. Phys.*, *25*(19), 13417–13428. https://doi.org/10.1039/D3CP00506B

Abstract: Due to the limitation of solvent models, quantum chemistry calculation of solution-phase molecular properties often deviates from experimental measurements. Recently, Δ-machine learning (Δ-ML) was shown to be a promising approach to correcting errors in the quantum chemistry calculation of solvated molecules. However, this approach's applicability to different molecular properties and its performance in various cases are still unknown. In this work, we tested the performance of Δ-ML in correcting redox potential and absorption energy calculations using four types of input descriptors and various ML methods. We sought to understand the dependence of Δ-ML performance on the property to predict the quantum chemistry method, the data set distribution/size, the type of input feature, and the feature selection techniques. We found that Δ-ML can effectively correct the errors in redox potentials calculated using density functional theory (DFT) and absorption energies calculated by time-dependent DFT. For both properties, the Δ-ML-corrected results showed less sensitivity to the DFT functional choice than the raw results. The optimal input descriptor depends on the property, regardless of the specific ML method used. The solvent–solute descriptor (SS) is the best for redox potential, whereas the combined molecular fingerprint (cFP) is the best for absorption energy. A detailed analysis of the feature space and the physical foundation of different descriptors well explained these observations. Feature selection did not further improve the Δ-ML performance. Finally, we analyzed the limitation of our Δ-ML solvent effect approach in data sets with molecules of varying degrees of electronic structure errors.

2 **Hruska**, **E.**, Gale, A., Huang, X., & Liu, F. (2022). AutoSolvate : A Toolkit for Automating Quantum Chemistry Design and Discovery of Solvated Molecules. *J. Chem. Phys.*, *156*(12). https://doi.org/10.1063/5.0084833

Abstract: The availability of large, high-quality data sets is crucial for artificial intelligence design and discovery in chemistry. Despite the essential roles of solvents in chemistry, the rapid computational data set generation of solution-phase molecular properties at the quantum mechanical level of theory was previously hampered by the complicated simulation procedure. Software toolkits that can automate the procedure to set up high-throughput explicit-solvent quantum chemistry (QC) calculations for arbitrary solutes and solvents in an open-source framework are still lacking. We developed AutoSolvate, an open-source toolkit to streamline the workflow for QC calculation of explicitly solvated molecules. It automates the solvated-structure generation, force field fitting, configuration sampling, and the final extraction of microsolvated cluster structures that QC packages can readily use to predict molecular properties of interest. AutoSolvate is available through both a command line interface and a graphical user interface, making it accessible to the broader scientific community. To improve the quality of the initial structures generated by AutoSolvate, we investigated the dependence of solute-solvent closeness on solute/solvent identities and trained a machine learning model to predict the closeness and guide initial structure generation. Finally, we tested the

capability of AutoSolvate for rapid data set curation by calculating the outer-sphere reorganization energy of a large data set of 166 redox couples, which demonstrated the promise of the AutoSolvate package for chemical discovery efforts.

3 **Hruska**, **E.**, Gale, A., & Liu, F. (2022). Bridging the experiment-calculation divide: Machine learning corrections to redox potential calculations in implicit and explicit solvent models. *J. Chem. Theory Comput.* 🔗`https://doi.org/10.1021/acs.jctc.1c01040`

Abstract: Prediction of redox potentials is essential for catalysis and energy storage. Although density functional theory (DFT) calculations have enabled rapid redox potential predictions for numerous compounds, prominent errors persist compared to experimental measurements. In this work, we develop machine learning (ML) models to reduce the errors of redox potential calculations in both implicit and explicit solvent models. Training and testing of the ML correction models are based on the diverse ROP313 dataset with experimental redox potentials measured for organic and organometallic compounds in a variety of solvents. For the implicit solvent approach, our ML models can reduce both the systematic bias and the number of outliers. ML corrected redox potentials also demonstrate less sensitivity to DFT functional choice. For the explicit solvent approach, we significantly reduce the computational costs by embedding the microsolvated cluster in implicit bulk solvent, obtaining converged redox potential results with a smaller solvation shell. This combined implicit-explicit solvent model, together with GPU-accelerated quantum chemistry methods, enabled rapid generation of a large dataset of explicit-solvent-calculated redox potentials for 165 organic compounds, allowing detailed investigation of the error sources in explicit solvent redox potential calculations.

4 **Hruska**, **E.**, Zhao, L., & Liu, F. (2022). Ground truth explanation dataset for chemical property prediction on molecular graphs. *Preprint*. 🔗`https://doi.org/10.26434/chemrxiv-2022-96slq-v2`

Abstract: Interpretation of chemistry on an atomic scale improves with explainable artificial intelligence (XAI). The parts of the molecule with the most significant influence on the chemical property of interest can be visualized with atomwise and bondwise attributions. Nonetheless, the attributions from different XAI methods regularly disagree substantially, causing uncertainty about which explanation is correct. To determine a ground truth for attributions, we define chemical operations which avoid alchemical steps or approximations and allow extracting one attribution per atom or bond from existing datasets of chemical properties. This general procedure allows for generating large datasets of ground truth attributions. The approach allowed us to create a ground truth explanation dataset with more than 5 million data points for the HOMO-LUMO gap chemical property. This open-source dataset of atomistic ground truth explanations may serve as a reference for XAI approaches.

5 Gale, A., **Hruska**, **E.**, & Liu, F. (2021). Quantum chemistry for molecules at extreme pressure on graphical processing units: Implementation of extreme-pressure polarizable continuum model. *J. Chem. Phys*, *154*, 244103. 🔗`https://doi.org/10.1063/5.0056480`

Abstract: Pressure plays essential roles in chemistry by altering structures and controlling chemical reactions. The extreme-pressure polarizable continuum model (XP-PCM) is an emerging method with an efficient quantum mechanical description of small- and medium-sized molecules at high pressure (on the order of GPa). However, its application to large molecular systems was previously hampered by a CPU computation bottleneck: the Pauli repulsion potential unique to XP-PCM requires the evaluation of a large number of electric field integrals, resulting in significant computational overhead compared to the gas-phase or standard-pressure polarizable continuum model calculations. Here, we exploit advances in graphical processing units (GPUs) to accelerate the XP-PCM-integral evaluations. This enables high-pressure quantum chemistry simulation of proteins that used to be computationally intractable. We benchmarked the performance using 18 small proteins in aqueous solutions. Using a single GPU, our method evaluates the XP-PCM free energy of a protein with over 500 atoms and 4000 basis functions within half an hour. The time taken by the XP-PCM-integral evaluation is typically 1% of the time taken for a gas-phase density functional theory (DFT) on the same system. The overall XP-PCM calculations require less computational effort than that for their gas-phase counterpart due to the improved convergence of self-consistent field iterations. Therefore, the description of the high-pressure effects with our GPU-accelerated XP-PCM is feasible for any molecule tractable for gas-phase DFT calculation. We have also validated the accuracy of our method on small molecules whose properties under high pressure are known from experiments or previous theoretical studies.

6 **Hruska, E.** (2020). *Adaptive sampling of conformational dynamics* [Doctoral dissertation, Rice University]. https://scholarship.rice.edu/handle/1911/108744

Abstract: At the core of our limited ability to understand many biophysical processes is the challenge of predicting the conformational dynamics of biomolecules. This challenge includes many open questions around the biophysical causes of many diseases or open questions in biophysics theory. Adaptive sampling is an approach to increase our ability to predict conformational dynamics. Adaptive sampling is a class of sampling strategies, where an ensemble of molecular dynamics trajectories is generated, where the starting points for the individual trajectories depend on the previously simulated trajectories. This approach will be investigated in this thesis. The application of adaptive sampling to biomolecules is one example of the more general problem of accurately sampling the time-dynamics of high-dimensional stochastic systems. The high-dimensionality, combined with a complex energy landscape, impede simpler approaches. Due to the broad scope of the general challenge, this Dissertation will focus only on improving the prediction of conformational dynamics for proteins. Many previous approaches to unravel this challenge have achieved significant improvements. In the case of proteins, the timescales where we can predict the conformational dynamics have increased by many orders of magnitudes to the millisecond scale. Despite the improvements, the current state-of-art can only predict the accurate behavior for small proteins. This illustrates the magnitude of the challenge. For most of the larger biomolecules, we are not able to simulate the precise behavior. This is not only caused by the several magnitudes longer timescales for these larger systems but also an order of magnitude larger sizes of these biomolecules. In this thesis, the adaptive sampling of conformational dynamics will be investigated in several steps. First, the prediction of the effectivity of different adaptive sampling strategies will be discussed. Due to significant stochasticity and protein-to-protein variation, the choice of adaptive sampling strategy is not apparent. The performance of different strategies for different goals varies as well. Second, to deepen our theoretical understanding of adaptive sampling strategies, an upper limit for the performance of any adaptive sampling strategy is developed. This theoretical upper limit allows us to understand the potential and limits of adaptive sampling. Third, adaptive sampling is heavily dependent on software due to the necessary thousands or millions of individual steps. All these steps have to be executed efficiently on a High-Performance Computer (HPC). Here we show the development of the software package ExTASY. This framework allows performing all the necessary steps in adaptive sampling while reducing the workload. The innovations of ExTASY are both the high-scalability and the modularity. The modularity allows for an easy change of the adaptive sampling strategies and better maintainability. ExTASY is reducing the entry barrier to utilizing adaptive sampling. Finally, the package ExTASY will be applied to show the results of adaptive sampling for several proteins. Future developments to extend the investigated approaches to longer timescales will be addressed. All the approaches mentioned above facilitate further advancements in predicting conformational dynamics of larger biomolecules.

7 **Hruska, E.**, Balasubramanian, V., Lee, H., Jha, S., & Clementi, C. (2020). Extensible and scalable adaptive sampling on supercomputers. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.0c00991

Abstract: The accurate sampling of protein dynamics is an ongoing challenge despite the utilization of high-performance computer (HPC) systems. Utilizing only "brute force"molecular dynamics (MD) simulations requires an unacceptably long time to solution. Adaptive sampling methods allow a more effective sampling of protein dynamics than standard MD simulations. Depending on the restarting strategy, the speed up can be more than 1 order of magnitude. One challenge limiting the utilization of adaptive sampling by domain experts is the relatively high complexity of efficiently running adaptive sampling on HPC systems. We discuss how the ExTASY framework can set up new adaptive sampling strategies and reliably execute resulting workflows at scale on HPC platforms. Here, the folding dynamics of four proteins are predicted with no a priori information.

8 **Hruska, E.**, Abella, J. R., Nüske, F., Kavraki, L. E., & Clementi, C. (2018). Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.*, *149*(24), 244119. https://doi.org/10.1063/1.5053582

Abstract: Adaptive sampling methods, often used in combination with Markov state models, are becoming increasingly popular for speeding up rare events in simulation such as molecular dynamics (MD) without biasing the system dynamics. Several adaptive sampling strategies have been proposed, but it is not clear which methods perform better for different physical systems. In this work, we present a systematic evaluation of selected adaptive sampling strategies on a wide selection of fast folding proteins. The adaptive sampling strategies were emulated using models constructed on already existing MD trajectories. We provide theoretical limits for the sampling speed-up and compare the performance of different strategies with and without using some a priori knowledge of the system. The results show that for different goals, different adaptive sampling strategies are optimal. In order to sample slow dynamical processes such as protein folding without a priori knowledge of the system, a strategy based on the identification of a set of metastable regions is consistently the most efficient, while a strategy based on the identification of microstates performs better if the goal is to explore newer regions of the conformational space. Interestingly, the maximum speed-up achievable for the adaptive sampling of slow processes increases for proteins with longer folding times, encouraging the application of these methods for the characterization of slower processes, beyond the fast-folding proteins considered here.

9 Balasubramanian, V., Bethune, I., Shkurti, A., Breitmoser, E., **Hruska, E.**, Clementi, C., Laughton, C., & Jha, S. (2016). Extasy: Scalable and flexible coupling of md simulations and advanced sampling techniques, 361–370. https://doi.org/10.1109/eScience.2016.7870921

Abstract: For many macromolecular systems the accurate sampling of the relevant regions on the potential energy surface cannot be obtained by a single, long Molecular Dynamics (MD) trajectory. New approaches are required to promote more efficient sampling. We present the design and implementation of the Extensible Toolkit for Advanced Sampling and analYsis (ExTASY) for building and executing advanced sampling workflows on HPC systems. ExTASY provides Python based "templated scripts" that interface to an interoperable and high-performance pilot-based run time system, which abstracts the complexity of managing multiple simulations. ExTASY supports the use of existing highly-optimised parallel MD code and their coupling to analysis tools based upon collective coordinates which do not require a priori knowledge of the system to bias. We describe two workflows which both couple large "ensembles" of relatively short MD simulations with analysis tools to automatically analyse the generated trajectories and identify molecular conformational structures that will be used on-the-fly as new starting points for further "simulation-analysis" iterations. One of the workflows leverages the Locally Scaled Diffusion Maps technique; the other makes use of Complementary Coordinates techniques to enhance sampling and generate start-points for the next generation of MD simulations. We show that the ExTASY tools have been deployed on a range of HPC systems including ARCHER (Cray CX30), Blue Waters (Cray XE6/XK7), and Stampede (Linux cluster), and that good strong scaling can be obtained up to 1000s of MD simulations, independent of the size of each simulation. We discuss how ExTASY can be easily extended or modified by end-users to build their own workflows, and ongoing work to improve the usability and robustness of ExTASY.

Bookchapter

2022 Quantum Chemistry in the Age of Machine Learning, 1st Edition, Elsevier, Chapter 6: Machine learning: An overview, **Eugen Hruska**, Fang Liu, Editor: Pavlo Dral, ISBN: 9780323900492