PHD PROJECT DESCRIPTION

(4000 characters max., including the aims and work plan to be published online)

Project title:

Reaction Coordinates in Molecular Systems: From Simulations to Thermodynamics and Kinetics via Spectral Machine Learning

1. Project goals

Reaction coordinates (RCs) are among the most important properties of physical and chemical systems, quantifying how they evolve through energy landscapes, yet they are virtually impossible to define for complex molecular systems. Recently, machine learning (ML) has been employed to directly define RCs from data generated by molecular dynamics (MD) simulations.

However, existing ML methods for constructing RCs from MD data are often purely data-driven and can violate fundamental physical principles such as detailed balance and Boltzmann statistics, leading to unphysical free energies and kinetic rates. This makes these RCs unusable in real-world scenarios, where simulations should computationally guide experiments.

The project goal is to develop a physics-informed ML method to construct RCs directly from MD simulation data, consistent with the principles of statistical mechanics. By embedding physical constraints into the learning framework, the method yields RCs that are both predictive and physically interpretable, enabling the reliable computation of free-energy landscapes and kinetic rates. The PhD candidate will work with a newly emerging framework of spectral ML (methods such as spectral and diffusion maps) that learns the spectral properties of complex systems using neural networks.

This project is at the interface of spectral methods, molecular simulation, and machine learning. It offers the student a unique opportunity to develop both ML and practical computational expertise (GPU-based training, MD) while working on a problem of direct physical significance.

2. Outline

The project will address the fundamental problem of identifying low-dimensional descriptors (reaction coordinates, RCs) that capture the essential dynamics of high-dimensional molecular systems. The approach will combine:

1. Data generation via molecular dynamics (MD) simulations of model and realistic systems (e.g., proteins).
2. Method development of a spectral ML architecture that respects statistical-mechanical constraints (e.g., detailed balance, Boltzmann statistics, committor probability), yielding RCs that are physically interpretable.
3. Thermodynamic and kinetic analysis using the learned RCs to extract free-energy profiles (potential of mean force) and kinetic rates (transition state theory).
4. Benchmarking and application against established RC discovery methods (e.g., diffusion maps, autoencoders) on systems of increasing complexity, culminating in a realistic molecular problem (e.g., protein folding, ligand binding, or chemical reactions).

3. Work plan

Year 1: Foundations and early research

- Literature review of RC identification methods and physics-informed ML
- Consolidation of ML methods using PyTorch and MD simulation tools
- Generation of benchmark MD datasets from simple model systems
- Study of statistical mechanics underlying reaction coordinates, free-energy landscapes, and kinetic rates

Year 2: Method development and validation

- Design and implementation of a prototype physics-informed ML method for RC construction
- Testing and validation on simple model systems with known analytical results
- Calculation of free-energy landscapes and kinetic rates from learned RCs

Year 3: Benchmarking and application

- Systematic comparison of the method to other ML techniques (autoencoders, diffusion maps)
- Training the method on a realistic physical, chemical, or biological system
- Knowledge extraction and physical interpretation of the generated results

Year 4: Dissemination and completion

- Publication of results in peer-reviewed journals
- Participation in international conferences
- Writing and defending PhD thesis
- Completion of required coursework and examinations

4. Literature (max. 7 listed as a suggestion for a PhD candidate preliminary study)

1. T. Gökdemir, J. Rydzewski, Machine Learning of Slow Collective Variables and Enhanced Sampling via Spatial Techniques, *Chem. Phys. Rev.* 6, 011304 (2025).
2. Noé, F. et al. Machine learning for molecular dynamics on long timescales. Annu. Rev. Phys. Chem. 75, 719–743 (2024).
3. J. Rydzewski, M. Chen, O. Valsson, Manifold Learning in Atomistic Simulations: A Conceptual Review, *Mach. Learn.: Sci. Technol.* 4, 031001 (2023).
4. L. Bonati et al. Enhanced Sampling in the Age of Machine Learning: Algorithms and Applications, Chem. Rev. 126, 1, 671–713 (2026).
5. J. Rydzewski, Spectral Map: Embedding Slow Kinetics in Collective Variables, *J. Phys. Chem. Lett.* 14, 5216 (2023).

5. Required initial knowledge and skills of the PhD candidate

- Programming: Fluency in Python; familiarity with scientific computing libraries (NumPy, SciPy) and deep learning frameworks (PyTorch).
- Machine Learning: Basic understanding of supervised/unsupervised learning, neural networks, and optimization.
- Physics/Chemistry: Undergraduate-level statistical mechanics and thermodynamics; basic knowledge of molecular

simulation concepts.
- Soft skills: Ability to work independently, scientific writing in English, and willingness to collaborate in an interdisciplinary environment.

1.1. Expected development of the PhD candidate's knowledge and skills

Upon completion, the candidate will possess an interdisciplinary profile combining expertise in ML/AI, molecular simulation, and statistical mechanics, equipping them for careers in computational chemistry/biology, data science, or academia.