# PhD Project - Start Fall 2026

## Generative Modeling for Efficient Exploration of Reactive Molecular Configuration Spaces

**Supervisors**
Primary: Marylou Gabrié (Physics Department & Center for Data Science, ENS – PSL)
Co-advisor: Guillaume Stirnemann (Chemistry Department, ENS – PSL)

## 1. Context

**Funding and environment** — This PhD will be conducted within the PR[AI]RIE-PSAI initiative, at the interface between machine learning, statistical physics, and computational chemistry. It will benefit from the highly interdisciplinary research environment of ENS Paris, where close collaborations between physicists, chemists, and machine learning researchers enable the development of novel approaches to scientific machine learning. The PhD candidate with expertise in machine learning will work in close interaction with a second PhD student with complementary expertise in computational chemistry.

**Scientific context** — Machine-learned interatomic potentials (MLIPs) have recently emerged as a powerful paradigm for approximating quantum mechanical energies and forces at a fraction of the computational cost, enabling molecular simulations at unprecedented scales. Considerable progress has been achieved in recent years in the design of model architectures and local atomic descriptors, leading to highly accurate representations of molecular interactions.

As a result, the primary limitation of MLIPs no longer lies in their expressivity, but rather in the construction of their **training datasets**. While large and diverse databases of stable structures are available for materials science applications, the situation is markedly different for **chemical reactivity**. Training sets for reactive systems must include transition-state configurations and rare-event structures, which are intrinsically unstable and require identifying appropriate reaction coordinates.

The identification and efficient sampling of these reaction coordinates therefore constitute a central challenge. In practice, current approaches rely predominantly on molecular dynamics (MD) simulations, often combined with active learning strategies. However, MD-based sampling suffers from slow decorrelation, leading to inefficient exploration of configuration space, and may become unstable when the underlying MLIP is not yet sufficiently accurate. These limitations result in datasets that may lack diversity or reliability, ultimately constraining the performance of the learned potentials.

Recent advances in **generative modeling**, including normalizing flows, diffusion models, and flow matching methods, offer a promising alternative paradigm. These approaches aim at directly learning and sampling from complex, high-dimensional distributions and have demonstrated remarkable success in other domains. However, their application to molecular systems—particularly in the context of reactive processes—remains largely unexplored. Adapting these models to respect physical constraints and to efficiently capture rare but essential configurations poses significant challenges.

This project aims to develop **generative approaches tailored to reactive molecular systems**, combining tools from machine learning, statistical physics, and molecular simulation.

## 2. Objectives and Scientific Roadmap

The central objective of this PhD is to design **efficient, physically grounded, and scalable generative methods** for sampling molecular configurations relevant to chemical reactions, and to understand their potential in the construction of training datasets for MLIPs.

The research program is structured along three complementary axes.

### Axis 1: Coupled Learning of MLIPs and Generative Models on Toy Reactive Systems

The first axis focuses on the development of active learning strategies in which a machine-learned interatomic potential (MLIP) and a generative model are trained jointly. The central idea is to move beyond standard molecular-dynamics-based exploration by leveraging generative models to enhance the diversity of configurations used to train the MLIP, building on previous work from the supervisors' groups [1, 2].

This axis will be developed progressively on simplified yet representative reactive systems. We will first consider low-dimensional benchmark potentials, such as the Müller–Brown model, before moving to more realistic reactions in implicit solvent. These settings provide controlled environments to study rare-event sampling and the interplay between exploration and learning.

A key objective will be to understand how the interaction between the learned energy model and the generative model can improve the exploration of configuration space. In particular, we will investigate how generative models can help access poorly sampled but chemically relevant regions, and under which conditions such approaches outperform standard sampling strategies based on molecular dynamics.

Beyond methodological development, this axis aims to identify regimes in which generative-assisted exploration improves decorrelation, enhances coverage of reactive coordinates, and reduces the number of expensive reference calculations required to train reliable MLIPs.

### Axis 2: Hybrid Generation and Relaxation Strategies for Solvated Systems

The second axis addresses the extension of these approaches to more realistic molecular systems, and in particular to **reactions occuring in solution**. It builds on the observation that purely generative models may struggle to capture fine-scale physical consistency in high-dimensional settings, while simulation-based methods remain inefficient for exploration.

We will therefore investigate hybrid strategies in which configurations generated in a reduced representation are combined with a subsequent relaxation step to recover physically consistent configurations in the full system. In this framework, solvent degrees of freedom are not modeled directly by the generative component, but are incorporated during the reconstruction stage, for instance through guided simulation techniques inspired by previous work [5, 3, 4].

While normalizing flows provide a natural starting point, alternative generative approaches such as **diffusion models** and **flow matching methods** will also be explored. A central question will be how to balance generative exploration and physical reconstruction as system complexity increases, and how this impacts the diversity and reliability of sampled configurations.

This axis provides the methodological foundation required to extend the approaches developed in Axis 1 to realistic reactive systems.

### Axis 3: Applications to Reactive Systems and MLIP Training

Building on the methodologies developed in Axes 1 and 2, the third axis focuses on their application to realistic chemical systems. The approach will follow a progressive strategy, moving from simplified models to increasingly complex and chemically relevant reactions.

We will evaluate the impact of generative and hybrid sampling strategies on MLIP training, with particular attention to accuracy, stability, and data efficiency. Comparisons with standard active learning approaches will be carried out to assess improvements in exploration and robustness.

An important objective will be to identify regimes in which generative approaches provide the largest benefits, for instance in systems characterized by rare events or highly multi-modal distributions. These studies will provide practical guidelines for integrating generative models into molecular simulation workflows and highlight their potential as a new paradigm for exploring complex physical systems.

## 3. Timeline and Candidate Profile

**Timeline** — The first year will focus on acquiring the necessary background in generative modeling and molecular simulation, as well as developing initial implementations on simplified systems. The second year will be dedicated to the design and analysis of hybrid strategies, combining theoretical insights and numerical experiments. The third year will concentrate on applications to realistic reactive systems and on the integration of the developed methods into MLIP training pipelines.

**Candidate profile** — The project requires a strong background in machine learning and a solid interest in statistical physics and/or computational chemistry. The ideal candidate should have strong programming skills, in particular in Python and modern machine learning frameworks (e.g., PyTorch or JAX). A strong interest in interdisciplinary research is essential, as well as the ability to navigate between theoretical concepts and practical implementations.

**Non-discrimination, openness and transparency** — All partners of PR[AI]RIE-PSAI are committed to supporting and promoting equality, diversity, and inclusion within their communities. We encourage applications from candidates with diverse backgrounds, and we ensure that selection is carried out through an open and transparent recruitment process.

## Application Process

Candidates interested in this PhD position should submit the following materials:

- A detailed **curriculum vitae**, including academic background and relevant experience;
- A **one-page motivation letter**, describing the candidate's interest in the proposed research topic, their scientific ambitions, and the relevance of their background to the project;
- Copies of **academic transcripts and diplomas**.

Applications should be sent directly to the supervisors at the following email addresses: `marylou.gabrie@ens.psl.eu`, `guillaume.stirnemann@ens.psl.eu`.

The deadline for applications is May 1st, and candidates should note that the selection process will be conducted in two phases, with results communicated between the end of May and mid-June 2026.

## References

[1] Rolf David, Miguel de la Puente, Axel Gomez, Olaia Anton, Guillaume Stirnemann, and Damien Laage. ArcaNN: Automated enhanced sampling generation of training sets for chemically reactive machine learning interatomic potentials. *Digital Discovery*, 4(1):54–72, January 2025.

[2] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, March 2022.

[3] C. Schönle, M. Gabrié, T. Lelièvre, and G. Stoltz. Sampling metastable systems using collective variables and Jarzynski–Crooks paths. *Journal of Computational Physics*, 527:113806, April 2025.

[4] Christoph Schönle, Davide Carbone, Marylou Gabrié, Tony Lelièvre, and Gabriel Stoltz. Efficient Monte-Carlo sampling of metastable systems using non-local collective variable updates, December 2025.

[5] Samuel Tamagnone, Alessandro Laio, and Marylou Gabrié. Coarse-Grained Molecular Dynamics with Normalizing Flows. *Journal of Chemical Theory and Computation*, September 2024.