Generative Deep Learning for Atomistically Engineered Materials: Synergistic Integration of Molecular Dynamics Simulations, Experiments and Data Augmentation

PhD proposal

Key-words Atomistic Simulation, Molecular Dynamics (MD), Generative Deep Learning, Data Augmentation.

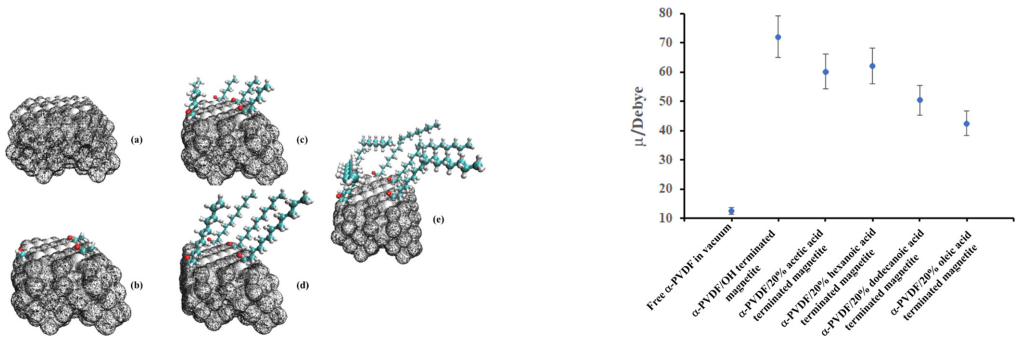
Intellectual Merit The emphasis in materials science has been on developing new materials with multiple functionalities through nanostructuration. The nanostructural nature of these materials, while rendering them versatile, presents a challenge to comprehensively understanding their behavior. Internationally, competition is intense, especially in developing devices to explore and exploit unusual physical aspects of nanostructures. Understanding the nanostructure-macroscopic behavior relationship is crucial for translating scientific findings into useful engineering systems.

A multi-scale description of the structure is the currently favored approach. However, predicting material properties requires precise values of relevant microstructural parameters, typically generated from advanced experimental characterization techniques. These data are often scarce or collected intensively on a narrow set of materials, making it challenging to construct a rationale for developing materials with different targeted performances.

To address this, the scientific community has pivoted towards atomistic simulation to investigate experimentally explored cases, aiming to understand observed phenomena and extrapolate findings to other configurations. Both approaches remain costly and require multiple iterations. In the contemporary realm of artificial intelligence, machine learning models offer an alternative, identifying influential parameters for targeted properties or uncovering hidden trends to predict new compositions' behavior. However, these techniques require large databases for efficient training, which are often lacking. Generative data augmentation may be a viable option if physically controllable and interpretable.

Context, Methodology and Thesis Objectives Research on nanostructured materials underscores the imperative to predict and comprehend their macroscopic properties for enhanced product design and performance optimization. Predicting behavior prior to experimental development mitigates the cost of new materials development. Atomistic simulation, owing to nanostructured aspects, represents the most prevalent methodology. Although providing valuable insights, the substantial computational cost constitutes a significant limitation. Consequently, while atomistic simulations yield valuable data, they are confined to specific cases and may not be readily applicable to other scenarios. An illustrative example, see Figure 1 is our recent findings on atomistic-driven multifunctionality in nanoparticle-reinforced polymers (1)

In this study, the nanoparticle (NP) surface chemistry is a critical parameter in polarizing the polymeric chain surrounding the NP, thereby modulating local polarization and piezoelectric properties of the composite materials in addition to its mechanical properties(2). Optimization of these



(a) Iron oxide slabs with different surfaces (b) Polymer dipole moment

Figure 1: example of the effect of iron oxide surface chemistry on the polymer polarization (1))

properties is essential for designing efficient nanostructured materials for energy harvesting, storage, or generation. Although significant, these findings are limited to specific NP compositions, surface chemistries, and polymeric materials. Generalizing these results to a broader range of NP surface chemistries and polymeric materials presents an opportunity to develop novel materials. Due to computational costs, the exploration of all possible parameters is precluded; similarly, experimental attempts are excluded for the same reasons. Deep learning algorithms that learn from available data and extrapolate to new cases may serve as a viable alternative (3).

Deep learning (4; 5), a subset of artificial intelligence, extracts patterns from data and makes predictions based on training inputs. While significant advancements have been made in deep learning techniques, the availability of large and representative training datasets remains crucial for model performance. Models trained on limited or non-representative datasets tend to generalize poorly. To address this, data augmentation techniques enhance model resilience and accuracy, particularly with small or unrepresentative datasets. Traditional methods involve simple transformations of existing samples to expand dataset variability. In contrast, deep learning-based augmentation employs generative models to synthesize new data points. Among these, Generative Adversarial Networks (GANs) and autoencoders are widely used. GANs function through a competitive process involving two neural networks: a generator that produces realistic synthetic data and a discriminator that distinguishes between authentic and generated samples. This adversarial training enables the generation of high-fidelity synthetic data. Autoencoders compress input data into a lower-dimensional latent space and then reconstruct it. By manipulating latent representations, autoencoders can generate novel data samples, contributing to dataset diversity and improving model training outcomes.

The primary objective is to advance the development of nanostructured materials with tailored properties through novel approaches. This thesis proposes an integrated, data-driven approach to expedite the development of advanced nanostructured materials. Using machine learning-driven data augmentation—specifically GANs, VAEs, and hybrid architectures—we address the constraints of limited datasets in materials science. This strategy complements existing experimental and atomistic modeling efforts, allowing robust predictions of material behavior across scales. It reduces time and cost associated with iterative experimentation and simulation. Moving forward, deeper validation of the synthetic data's physical relevance—via experiments and atomistic simulations—will be crucial.

Candidate profile This thesis suits students with strong backgrounds in materials science/computational chemistry or applied math and machine learning, willing to venture outside their narrow specialization. Student selection will be based first and foremost on their motivation to work on a multidisciplinary topic in a pluri-disciplinary team.

References

[1] M. Sahihi, A. Jaramillo-Botero, W. A. Goddard, and F. Bedoui, "Interfacial interactions in a model composite material: Insights into $\alpha \rightarrow \beta$ phase transition of the magnetite reinforced poly(vinylidene fluoride) systems by all-atom molecular dynamics simulation," *The Journal of Physical Chemistry C*, vol. 125, no. 39, pp. 21 635–21 644, 2021.

[2] F. Bedoui, M. Sahihi, A. Jaramillo-Botero, and W. A. Goddard, "Enhancing multifunctionality: Optimal properties of iron-oxide-reinforced polyvinylidene difluoride unveiled through full atom molecular dynamics simulations," *Langmuir*, vol. 40, no. 15, pp. 8067–8073, 2024, pMID: 38557046.

[3] J. Zhang, D. Chen, Y. Xia, Y.-P. Huang, X. Lin, X. Han, N. Ni, Z. Wang, F. Yu, L. Yang, Y. I. Yang, and Y. Q. Gao, "Artificial intelligence enhanced molecular simulations," *Journal of Chemical Theory and Computation*, vol. 19, no. 14, pp. 4338–4350, 2023, pMID: 37358079.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] M. Bouhadida, A. Mazzi, M. Brovchenko, T. Vinchon, M. Z. Alaya, W. Monange, and F. Trompier, "Neutron spectrum unfolding using two architectures of convolutional neural networks," *Nuclear Engineering and Technology*, vol. 55, no. 6, pp. 2276–2282, 2023.