ABSTRACT

The integration of machine learning (ML) into the chemical sciences has catalyzed a paradigm shift, transforming chemistry into a data-driven discipline capable of accelerating discoveries and uncovering patterns beyond the reach of traditional methods. This thesis seeks to advance the field of chemical ML by addressing foundational challenges in data generation, model design, and benchmarking for molecular property prediction and reaction thermochemistry.

The cornerstone of this work is the RGD1 datasetthe worlds largest computationally generated reaction thermochemistry datasetwhich provides unparalleled chemical diversity and accuracy with over 176,000 reactions. This dataset was systematically curated using graphically defined elementary reaction steps and comprehensive conformational sampling. Despite its transformative potential, challenges such as incomplete conformational sampling and unintended transition states remain, underscoring the need for enhanced data curation and filtering techniques.

Building on RGD1, this thesis introduces Edge-featured Graph Attention Networks (EGAT), a novel graph-based architecture tailored for chemical property prediction. EGAT leverages attention mechanisms to capture both local and global chemical features, demonstrating competitive or superior performance on diverse datasets, including experimental (ESOL, FreeSolv, Lipophilicity) and computational (PCQMv2, QM9) benchmarks. The architectures flexibility is exemplified in its multi-task learning applications, where shared latent representations enable simultaneous prediction of multiple quantum chemical properties, achieving parameter efficiency and robust generalization.

The thesis also explores the scalability and generalizability of ML models through multi-task learning and globally parameterized frameworks. On the QM9 dataset, multi-task EGAT models efficiently balance shared and task-specific learning, mitigating the risk of negative transfer while improving prediction accuracy across heterogeneous quantum properties. Additionally, a globally parameterized model trained on incomplete and heterogeneous data highlights the potential and limitations of unifying chemical property prediction. This

approach underscores the importance of targeted data augmentation, loss normalization, and uncertainty-aware training to address data sparsity and improve model robustness.

Through systematic benchmarking and rigorous evaluation, this thesis provides a comprehensive framework for advancing ML in the chemical sciences. The findings emphasize the importance of integrating data diversity, architectural flexibility, and uncertainty quantification to develop scalable, transferable ML frameworks. These contributions not only enhance predictive capabilities but also pave the way for accelerated molecular and reaction-based discovery in the broader chemical landscape.