ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Վահագն Նորիկի Ալթունյան

Մեքենայական ուսուցման և բաշխված հաշվարկային մոտեցումներ քվանտային քիմիական տվյալների ստեղծման և մոլեկուլային հատկությունների կանխատեսման համար

Ե.13.05 - «Մաթեմատիկական մոդելավորում, թվային մեթոդներ և ծրագրերի համալիրներ» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ - 2025

INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS OF THE NAS RA

Vahagn Norik Altunyan

Machine learning and distributed computing approaches for quantum chemistry-based data generation and molecular property prediction

SYNOPSIS

of the dissertation for obtaining a Ph.D. degree in Technical Sciences on specialty 05.13.05 “Mathematical modeling, digital methods and program complexes”

YEREVAN - 2025

Ատենախոսության թեման հաստատվել է ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում

Գիտական ղեկավար՝	Ֆիզ. մաթ. գիտ. թեկնածու Ա. Ն. Հարությունյան
Պաշտոնական ընդդիմախոսներ՝	Ֆիզ. մաթ. գիտ. դոկտոր Ա. Հ. Պողոսյան
	Տեխ. գիտ. թեկնածու Վ. Գ. Վարդանյան
Առաջատար կազմակերպություն՝	Հայ-ռուսական համալսարան

Ատենախոսության պաշտպանությունը կկայանա 2025թ. հուլիսի 8-ին, ժ. 13:00-ին ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող 037 մասնագիտական խորհրդի նիստում հետևյալ հասցեով' Երևան, 0014, Պ. Սևակի 1

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ ԻԱՊԻ գրադարանում:

Սեղմագիրն առաքված է 2025թ. հունիսի 7-ին:

Մասնագիտական խորհրդի գիտական
քարտուղար ֆիզ. մաթ. գիտ. դոկտոր՝ Մ. Ե. Հարությունյան

The topic of the dissertation was approved at the Institute for Informatics and Automation Problems of the NAS RA

Scientific supervisor:	Candidate of Phys-Math sciences A. N. Harutyunyan
Official opponents:	Doctor of Phys-Math sciences A. H. Poghosyan
	Candidate of Tech. sciences V. G. Vardanyan
Leading organization:	Russian-Armenian University

The dissertation defence will take place on July 8, 2025; at 13:00, at the Specialized Council 037 «Informatics» at the Institute for Informatics and Automation Problems of NAS RA. Address: Yerevan, 0014, P. Sevak 1.

The dissertation is available in the library of IIAP NAS RA.

The abstract is delivered on June 7, 2025.

Scientific Secretary of the Specialized Council, D.Ph.M.S. M. E. Haroutunian

1. Relevance of the Theme

The pursuit of novel molecules with tailored functionalities—be it for therapeutic intervention, advanced materials, or sustainable chemical processes—navigates an extraordinarily vast and complex landscape known as chemical space. This conceptual multidimensional space encompasses all theoretically possible molecular structures. Of particular interest for drug discovery is the subspace of "drug-like" molecules, which feature physicochemical properties that give them a higher likelihood of being therapeutically effective and orally bioavailable. Exploring this specific region is paramount for developing new medicines. Conservative estimates place the number of "drug-like" molecules alone on the order of 10^{60}[1], a figure of such magnitude that it underscores the impossibility of exhaustive experimental enumeration and characterization. Consequently, the rational exploration and exploitation of chemical space necessitate powerful computational and theoretical frameworks capable of predicting molecular properties and guiding the search for promising candidates. **Figure 1** illustrates the astronomical scale of chemical space in contrast to existing molecular databases[2][3][4][5][6], highlighting the vast opportunity space accessible only through efficient in silico strategies.

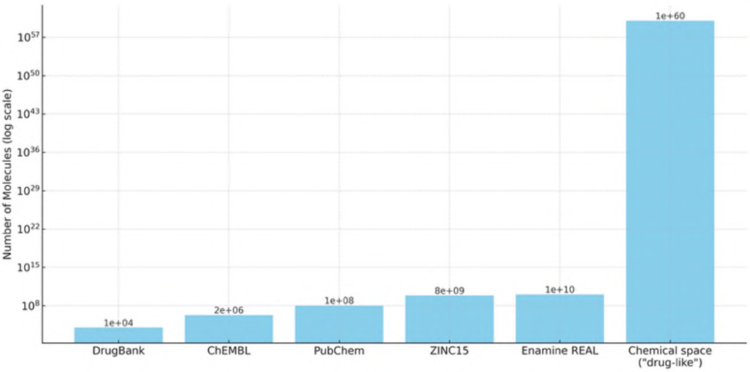

Figure 1: Size comparison between "drug-like" chemical space and known molecular datasets

The application of machine learning (ML) to chemical discovery, while holding immense promise, faces unique challenges not typically encountered in other data-rich disciplines like computer vision or natural language processing. In those fields, vast datasets are often readily available or can be generated at a relatively low cost per instance. In contrast, each data point in chemistry, a molecule annotated with its experimentally determined or accurately computed properties, can represent a significant investment of time, resources, and expert labor. This inherent data scarcity, combined with the immense complexity of chemical space, means that traditional ML approaches requiring voluminous training data often encounter limitations in terms of generalizability and predictive accuracy. The development of robust models for chemistry is therefore critically dependent on

[1] J.-L. Reymond, "The chemical space project," Acc Chem Res, vol. 48, no. 3, pp. 722–730, 2015,

[2] https://go.drugbank.com

[3] https://www.ebi.ac.uk/chembl

[4] https://pubchem.ncbi.nlm.nih.gov

[5] https://zinc15.docking.org

[6] https://enamine.net/compound-collections/real-compounds/real-database

strategies that can either maximize the information gained from limited, high-cost data or dramatically increase the efficiency of high-quality data generation.

The process of annotating molecular structures with relevant properties, can be approached through several distinct methodological tiers. On the one hand, **experimental measurements** are the gold standard, but are often low-throughput, expensive, and may not be feasible for all properties or for vast numbers of compounds. On the other hand, computational approaches (including empirical, semi-empirical, and ab initio methods) offer scalable alternatives, varying in complexity, accuracy, and generalizability. **Empirical methods**, such as classical molecular mechanics force fields, offer high computational speed but their accuracy and transferability can be limited, particularly for novel chemical compounds. **Semi-empirical methods** provide a compromise by incorporating some quantum mechanical approximations with empirical parameterization, offering improved accuracy over force fields at a greater computational cost. **Ab initio methods**[7], derived from first principles of quantum mechanics without empirical parameters, offer the highest potential for accuracy and generalizability. These methods fall within the broader domain of quantum chemistry, which applies quantum mechanical principles to understand and predict the behavior of atoms and molecules. Within this category, methods based on solving approximations to the Schrödinger equation, such as Density Functional Theory (DFT)[8], have become workhorses. While ab initio methods provide highly reliable data, they are the most computationally intensive.

This thesis focuses on ab initio methods, specifically DFT, for accurate molecular data generation.

Given the high computational cost of ab initio methods, it is essential to maximize the utility of each calculation. This necessitates the strategic curation of datasets to ensure that generated data provide representative and meaningful chemical coverage. Existing quantum chemical datasets often suffer from significant curation problems, such as a lack of scaffold (molecular core structures) diversity and a narrow focus on relatively small molecules. These issues lead to poor coverage of chemical space; thus, there is a clear need for new, carefully designed datasets that thoroughly explore and exploit chemical space to enable generalizable and high-quality molecular modeling.

In this context, maximizing information gain from newly generated data samples heavily depends on the diversity of molecular structures. When considering quantum chemical properties, the uniqueness of molecules can be assessed using structural similarity metrics. However, the computation of these metrics can be problematic for samples representing highly symmetrical molecular graphs. Robust structural comparison is thus key to efficiently building diverse, informative datasets and ensuring computational resources are directed towards genuinely novel structural information.

1.1. Challenges in Ab Initio Methods

The fundamental properties and behavior of any given molecule are governed by the principles of quantum mechanics. The time-independent Schrödinger equation, $H\Psi = E\Psi$, where H is the Hamiltonian operator, Ψ is the molecular wavefunction, and E is the energy of the system, provides the theoretical bedrock for understanding molecular structure and properties. While this equation offers a complete description, its exact solution is intractable for multi-electron systems, necessitating approximations. For performing ab initio calculations, several practical challenges arise:

[7] *J. A. Pople*, “Development of ab initio methods in quantum chemistry,” Rev Mod Phys, vol. 71, no. 5, pp. 1267–1274, 1999
[8] *W. Kohn, et al.*, “A perspective on density functional theory,” J Phys Chem, vol. 100, no. 31, pp. 12974–12980, 1996

- **The Cost-Accuracy Trade-off:** Ab initio calculations, particularly DFT, are computationally intensive, with costs typically scaling as a power (generally N^3 or N^4) of the number atoms in molecule. Furthermore, the selected level of approximation directly influences computational cost: achieving higher accuracy necessitates a higher computational cost.
- **Software and Tool Availability:** A variety of academic and commercial software packages are available for performing quantum chemical calculations. However, many packages require paid licenses, which can make them less practical or cost-prohibitive for large-scale data generation.
- **Resource Requirements:** Beyond significant CPU time, ab initio calculations can impose substantial demands on other system resources like Random Access Memory (RAM) and disk usage.

Addressing these practical challenges is essential for any research initiative that relies on generating or utilizing data from ab initio calculations.

1.2. Challenges in Existing Quantum Chemical Datasets

The efficacy of ML models in chemistry is dependent on the quality and diversity of training data, particularly for predicting quantum mechanical properties like conformational energy. However, many existing molecular datasets exhibit significant limitations that hinder the development of truly generalizable models capable of navigating the vastness of chemical space. Here are some common limitations of existing datasets:

- **Molecule Size:** Molecules in existing datasets are often chosen to be small to keep the computational cost of DFT calculations manageable.
- **Molecular Diversity:** Many datasets provide only a single 3D structure for each molecule. This ignores the inherent flexibility of molecules and prevents the model from learning how properties change as a molecule's shape changes.
- **Scaffold Diversity:** Structural diversity is often low, with datasets built from a small, repetitive set of scaffolds. This limits a model's ability to generalize to new or fundamentally different molecular architectures.
- **Data Splitting and Leakage:** When predefined train-test splits are provided, they often contain fundamental flaws. Molecular scaffolds in the test set can be similar to those in the training set, leading to overly optimistic performance metrics that fail to measure how a model will perform on genuinely novel data.

These limitations underscore the critical need for novel approaches to dataset curation. Furthermore, the adoption of rigorous train-test splitting methodologies, including scaffold-based separation augmented by similarity filtering, is essential to avoid overoptimistic model evaluation and ensure true generalization.

1.3. Challenges in Symmetry-Corrected RMSD Calculation Tools (SC-RMSD)

The accurate comparison of molecular structures, essential for identifying unique conformations and ensuring dataset diversity, is complicated by molecular symmetry. The Root Mean Square Deviation (RMSD) is the standard metric for molecular structural similarity assessment, but its naive application can be misleading for symmetric molecules. Symmetry-corrected RMSD (SC-RMSD) addresses this by finding the optimal atomic mapping **consistent with chemical equivalence** that minimizes the RMSD. This corresponds to an optimization problem over the space of graph isomorphisms, where node attributes are defined by atom types, and optionally, edge attributes by bond types. The solution thereby effectively accounts for molecular symmetry.

Two variants of this metric are widely used in computational chemistry:

- **SC-RMSD:** This variant is focused on fixed-oriantation SC-RMSD. Widely used in docking evaluations:

$$\boldsymbol{SC\text{-}RMSD} = \min_{\pi \in P} \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{i=1}^{N} \left\| \boldsymbol{v_i} - \boldsymbol{w_{\pi(i)}} \right\|^2}$$

- **Minimized SC-RMSD:** This variant is focused on combination of optimal atoms mapping and optimization of rotation and translation values for achieving minimal SC-RMSD. Mainly used for comparison of molecular structures without positional restrictions:

$$\boldsymbol{SC\text{-}RMSD_{min}} = \min_{\pi \in P, U, t} \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{i=1}^{N} \left\| \boldsymbol{v_i} - \left(\boldsymbol{U w_{\pi(i)}} + \boldsymbol{t} \right) \right\|^2}$$

In both equations, v_i and w_i denote coordinates of first and second conformations, P is the space of graph isomorphisms between first and second molecules, U is rotation matrix and t is translation vector.

Calculation of both metrics introduces computational problems such as:

- **Optimization over Graph Isomorphisms**: The calculation is fundamentally an optimization problem over graph isomorphisms space, which represents all possible chemically-equivalent mappings between the two molecular structures.
- **Exponential Search Space:** The size of this search space can grow exponentially with the number of atoms, making a full search computationally infeasible for many molecules.
- **Increased Complexity for Minimized SC-RMSD:** For the minimized SC-RMSD, the optimization space becomes even more complex as it must account for rotational and translational alignment in addition to the atomic mapping.

Despite several open-source tools providing solutions for these problems, they often introduce new issues related to calculation speed, accuracy, and reliability.

2. Aim of the Work

The primary scientific goal of this dissertation is to design and validate an end-to-end computational pipeline for the systematic generation of high-quality quantum chemical datasets and the prediction of molecular properties. The object of this research is molecular systems and their quantum chemical properties, specifically the relationship between molecular structure (including conformational variations) and measurable chemical properties. The subject of the research is defined as the development of computational methodologies for efficient and systematic exploration of chemical and conformational space to enable accurate prediction of molecular properties through high-quality quantum chemical datasets.

This work has several key aims:

- To formulate and implement a distributed computing infrastructure which is tailored for large-scale Density Functional Theory (DFT) calculations, addressing computational bottlenecks of ab initio methods through parallelized and democratized resource use.
- To develop a rigorous theoretical and algorithmic framework for the efficient selection of molecular conformations that optimizes chemical space exploration by incorporating:
    - Methodologies for sampling the molecular space that ensure diverse coverage, with relevance to specific target molecular properties.
    - Methodologies for sampling the conformational space that capture energetically relevant structural variations.
    - Systematic approaches for data diversification and redundancy mitigation that maximize the informational content of generated datasets and prevent unnecessary computational effort.

3. Contributions

To achieve the stated goals and address the outlined tasks, we employed methods from machine learning and distributed computing, alongside algorithmic techniques and specialized data structures.

The main contributions submitted for defense are as follows:

- **A Volunteer Computing Platform for Large-Scale DFT Calculations:** Designed, implemented, deployed, and validated the Smart Distributed Data Factory (SDDF)[9], a novel volunteer computing platform enabling large-scale DFT calculations by distributing tasks across a global network, significantly reducing the cost of ab initio data generation. SDDF features a modular architecture with AI-guided molecular sampling and has been successfully validated through a conformational energy dataset generation project.
- **An Active Learning Framework for Efficient Molecular Data Generation:** Developed and integrated within SDDF a new active learning (AL) method for molecular sampling that employs a heterogeneous ensemble of Graph Neural Network (GNN) models and novel selection strategies to efficiently identify informative molecular conformations for DFT labeling, thereby enhancing chemical space exploration. The selection strategies are based on using separate ML models to estimate the property prediction errors of the ensemble's GNNs on given samples and sampling conformations from the MD trajectories based on the force prediction disagreement among the GNNs.

[9] https://sddfactory.cloud

- **Large-Scale Quantum Chemistry Datasets:** Generated and publicly released substantial, high-quality dataset[10] of molecular conformational energies (approx. 2.75 million by May 2025) from the ENAMINE REAL database, annotated with ωB97X/6-31G(d) DFT energies. In addition to the dataset, the SDDF benchmark is provided, with rigorous scaffold-based train-test splits for robust model evaluation.
- **GNN-based Models for Conformational Energy Prediction:** Created and trained accurate GNN models for predicting molecular conformational energies using the SDDF-generated datasets. The GNN ensemble demonstrated strong predictive capabilities, outperforming established benchmarks, with models and code released to the community.
- **A High-Performance Method and Tool for Symmetry-Corrected RMSD Calculation:** Developed FlashRMSD, a novel method for SC-RMSD calculation that demonstrates superior speed and reliability compared to existing approaches, especially for highly symmetric molecules. In addition, a high-performance C-based command-line tool based on the method was implemented and released publicly[11].
- **A Comprehensive Benchmark Dataset for SC-RMSD Tool Evaluation:** Curated and published a comprehensive benchmark dataset[12] (from PDB's CCD and BIRD) for rigorous SC-RMSD tool evaluation. Both the benchmark and reference values are publicly available.

4. Scientific Novelty

The scientific novelty of this work is centered on several newly developed methods and systems that advance molecular data generation, modeling, and evaluation.

- The Smart Distributed Data Factory (SDDF) represents a novel volunteer computing platform specifically designed for AI-guided molecular datasets generation. Its modular architecture and integration of AI-guided molecular sampling distinguish it from existing distributed computing solutions. Machine learning models, generated by SDDF, outperformed existing state-of-the art models on some benchmarks. [2]
- A second novel contribution is the active learning framework integrated within SDDF. This framework combines novel selection strategies based on using separate ML models to estimate the property prediction errors of the ensemble's GNNs on given samples and sampling conformations from the MD trajectories based on the force prediction disagreement among the GNNs. This design enables more efficient identification of informative conformations for DFT labeling. [2]
- The dissertation also introduces FlashRMSD, a novel high-performance method for symmetry-corrected RMSD calculation. FlashRMSD offers substantial improvements in speed and reliability over existing tools, particularly for highly symmetric molecules. [3]

These innovations collectively advance the state of the art in computational molecular modeling by enabling smarter data acquisition and more efficient structural analysis.

[10] https://zenodo.org/records/15359529
[11] https://github.com/altunyanv/FlashRMSD
[12] https://zenodo.org/records/15097621

5. The Practical Significance of the Work

The methodologies, tools, and datasets developed in this thesis offer significant practical benefits across various domains of computational chemistry, drug discovery, and materials science. The key areas of impact include:

- **Facilitating Broader Molecular Data Generation:** The validated volunteer computing platform (SDDF) is not limited to conformational energies. Its architecture allows for the definition of new computational projects, enabling the community to leverage distributed resources for generating datasets of other crucial molecular properties (e.g., atomic charges, dipole moments, vibrational frequencies, reaction energies) that are also expensive to compute via DFT.
- **Advancing Neural Network Potentials for Molecular Dynamics:** The high-quality DFT energy data generated through this work serve as ideal training material for next-generation machine learning potentials (MLPs), also known as neural network potentials (NNPs). These MLPs can subsequently power molecular dynamics (MD) simulations with an accuracy approaching that of DFT but at a significantly reduced computational cost.
- **Enriching Community Resources with Open Datasets:** The public release of large-scale, curated datasets of molecular energies, particularly those derived from drug-discovery relevant libraries like ENAMINE and featuring diverse molecular sizes, directly addresses the critical issue of data scarcity in molecular ML. These open-source datasets provide invaluable resources for the broader scientific community to train more robust and generalizable predictive models, and enhance the performance of existing tools.
- **Improving Molecular Docking Accuracy and Efficiency:** Molecular docking simulations, a cornerstone of structure-based drug design, often generate numerous potential binding poses for a ligand within a receptor's active site. Accurate calculation of SC-RMSD enables the selection of structurally diverse conformations, reducing redundancy and improving the efficiency of downstream analyses, which are often computationally intensive.
- **Enhancing Reliability in Virtual Screening Pipelines:** Virtual screening aims to computationally evaluate vast chemical libraries to identify promising candidate molecules with potential biological activity. When structural similarity analysis or conformational assessment forms part of the screening cascade, the reliability and efficiency of SC-RMSD calculations become paramount. This is especially critical to avoid the propagation of errors that can arise from less reliable or computationally prohibitive symmetry correction and alignment methods in large-scale automated workflows, leading to more efficient and accurate hit identification.
- **Providing Standardized Benchmarks for Tool Validation:** The comprehensive benchmark dataset developed for the evaluation of SC-RMSD tools, serve as a vital, standardized resource for the cheminformatics community. Developers of new SC-RMSD algorithms or software can utilize this dataset to rigorously validate their methodologies, objectively compare performance metrics (accuracy, speed, failure rates) against established tools, and pinpoint areas requiring further improvement. This fosters a more systematic and transparent assessment of new computational instruments, thereby promoting continued innovation and raising the overall standard in the field of molecular structural comparison.

6. Integration of Results

1. SDDF platform was deployed and made openly accessible. The platform supports both volunteer-based computation and project proposals from external contributors, enabling collaborative large-scale data generation. Company DeepOrigin was an active contributor for the molecular energy calculation project as a volunteer user, and also proposed a new project for atomistic RESP charges calculation, which is now ongoing and has over 600 thousand labeled samples as of May 1 2025.
2. An open-source dataset comprising over 2.75 million molecular conformations paired with their conformational energies was released. As of the end of May, the dataset had been downloaded over 500 times, indicating significant interest and active use in research applications. This dataset is currently being used by company DeepOrigin, in development of various machine learning methods.
3. FlashRMSD open-source tool for symmetry-corrected RMSD (SC-RMSD) calculation was developed and published. The tool is currently being used by Laboratory of Computational Modeling of Biological Processes of Institute of Molecular Biology NAS RA and has been integrated into the molecular analysis toolkit of the company DeepOrigin.

7. Publications and Approbation of the Results

All results presented in this thesis are original and have been published in both local and international peer-reviewed journals. This includes one publication[2] in the Scopus-indexed journal Nature's "Scientific Reports" (Q1) and two publications[1, 3] in periodicals acceptable for the SSC. A complete list of articles and published datasets is provided at the end of the Synopsis.

Some of the results have been presented at the scientific conference *Current Issues in Computer Science and Applied Mathematics* (Yerevan, Armenia, April 28–30, 2025). Additionally, the research underwent internal review and discussion within the company DeepOrigin.

8. Structure and Scope of Work

The dissertation consists of 5 chapters and a list of used literature. The thesis is written in 107 pages and has 71 literature references. The thesis contains 24 figures and 20 tables.

The thesis is organized as follows:

Chapter 1. Serves as an introduction. It describes the main challenges in high quality molecular data generation, and the main objectives of the thesis.

Chapter 2. **Smart Distributed Data Factory**

Section 2.1 reviews existing volunteer computing platforms and the rationale for SDDF. *Section 2.2* introduces SDDF platform, its architecture and design choices. *Section 2.3* describes user-side computations, system requirements, task execution for energy calculations, and DFT method selection with resource analysis. *Section 2.4* summarizes this chapter.

Chapter 3. **Active Learning-Driven Molecular Energy Prediction and Data Generation**

Section 3.1 reviews related work in application of active learning for molecular data generation. *Section 3.2* covers initial dataset generation, and detailed discussion of GNN ensemble: the core of AL framework. In particular, it focuses on GNN models selection, detailed discussion of chosen models architectures, experiments for finding optimal feature representations, training procedure, and preformance comparisons. *Section 3.3* starts from introduction of the active learning framework design, and benchmark setup for framework validation. It introduces novel approaches for molecular sampling through ML-based loss predictors and conformational sampling through atomistic forces confidence evalutation among GNN ensemble models. *Section 3.4* summarizes generated datasets, model development, and the SDDF platform's status. *Section 3.5* summarizes this chapter.

Chapter 4: **FlashRMSD: High-Performance Symmetry-Corrected RMSD Calculation and Benchmarking**

Section 4.1 defines the SC-RMSD problem and its challenges. *Section 4.2* presents FlashRMSD, a method for SC-RMSD calculation, its features, and algorithmic design. *Section 4.3* details dataset curation for SC-RMSD tool benchmarking, benchmarking methodology, results, and comparative analysis. *Section 4.4* introduces FlashRMSD-M for minimized SC-RMSD calculation, its approach, performance, case studies, and evaluation summary. *Section 4.5* summarizes this chapter.

Finally, *Chapter 5* concludes the thesis with a summary of the contributions made.

9. The Main Results of the Work

The following points summarize the key contributions and findings:

1. **Volunteer Computing Platform for Large-Scale DFT Calculations:** We developed the SDDF (Smart Distributed Data Factory) platform, which provides a website (https://sddfactory.cloud) where volunteers can sign up and receive molecular conformations for DFT calculations on their personal computers. Each calculation task consists of a single conformation of a molecule and a property specifier indicating a set of properties to calculate.

   While distributed computing has a rich history, spanning academic grids, public-resource and peer-to-peer systems, enterprise solutions, and versatile volunteer frameworks like BOINC[13], these pioneering platforms often require extensive customization or are not optimally suited for the specific demands of accessible, volunteer-driven DFT calculations in chemistry. Challenges typically arise in areas of intergration of AI-guided task generation and efficient quantum chemistry calculations.

   The Smart Distributed Data Factory (SDDF) system is composed of interconnected components that collectively manage, distribute, and process computational chemistry tasks **(Figure 2)**. At its core, the Central Node includes a Task Queue for managing workloads, a Database for storing task-related data, and an SDDF Server that formulates and distributes computational tasks via gRPC. A Web Server, hosted with FastAPI and backed by MongoDB, enables external client interaction and visualizes volunteer contributions through a leaderboard interface. Complementing this, the Distribution Node contains an SDDF Tunnel, enabling volunteer nodes to retrieve molecular structures and submit results and offloading the Central Node from large number of requests. Supporting these components are several scheduled

[13] https://boinc.berkeley.edu

services: a conformation generator using RDKit or OpenBabel, a machine learning–based conformation generator that leverages energy model gradients for force estimation, and an AI-enhanced task selector that prioritizes challenging conformations for model improvement. Together, these modules ensure scalable, intelligent generation and distribution of high-quality molecular data.

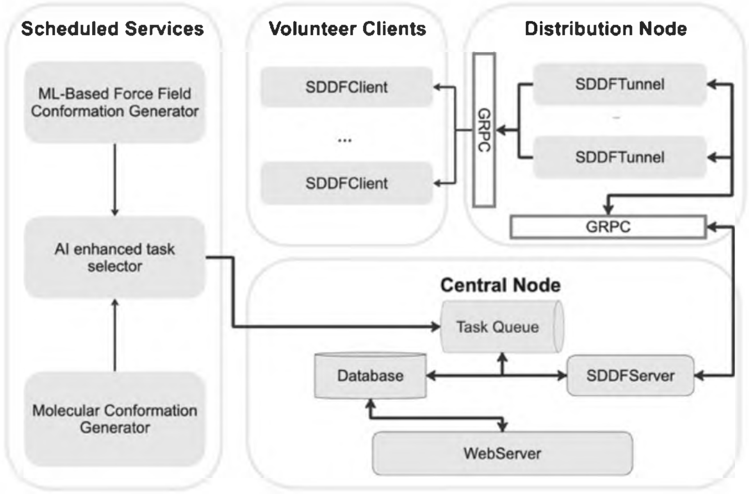

Figure 2: The architecture of the distributed computing system.

The platform's primary goal is to generate an extensive dataset of molecular conformational energies. To achieve this, we selected the ωB97X/6-31G(d) DFT theory level. This choice offers a pragmatic balance between accuracy and computational cost for volunteer contributors, a decision supported by benchmarks against higher-accuracy DFT methods. We chose the Psi4[14] toolkit for DFT calculations due to its open-source nature and broad CPU architecture support.

With these selections, calculating the conformational energy of an average-sized molecule (approximately 25 heavy atoms) on a single-core machine with minimal system requirements takes about 10 minutes. This runtime demonstrates the suitability of conformational energy calculations for the volunteer computing model: it's not so quick as to be ideal for cloud computing, nor so slow that it would disengage volunteers.

2. **Active Learning Framework for Efficient Molecular Data Generation:** SDDF implements an active learning framework to select molecules for labeling and addition to the dataset. The framework iteratively samples molecules from a large database in random fashion and generates multiple conformations for each molecule using RDKit and MD. At each iteration, a fraction of the generated conformations is selected and labeled, after which they are added to the dataset.

[14] https://psicode.org

Our AL framework leverages ensemble of Graph Neural Networks (GNN) for molecular conformation selection. For this, we investigated different strategies.

We introduced a novel approach utilizing machine learning predictors. These predictors are linear regression models that estimate the corresponding ensemble models' error for a given molecule. Each regression model takes the MACCS fingerprint of the molecule as input and is trained on the same dataset as the ensemble. The training process employs a regression approach, predicting the MAE between the ensemble model's prediction and the actual conformational energy. A high predicted error value indicates a challenging and thus valuable instance for data acquisition. The conformations are ranked based on this error (from highest to lowest), and the highest ranked are selected for DFT calculations and subsequent inclusion in the dataset. The predictor's relatively simple architecture (**Figure 3**) and its independence from 3D structural information enable rapid error estimation across a large molecular dataset.

Additionally, we explored an earlier strategy based on a disagreement score calculated from the predictions of the GNN ensemble (**Figure 4**). For SDDF, this score is based on the relative standard deviation of model predictions, quantifying the uncertainty or conflict among the ensemble's predictions for a given conformation. While this method has been validated in several previously published works, the machine learning-based approach ultimately provided a more efficient solution in our final implementation.

In order to train the ML ensemble, our platform labels a small initial dataset of randomly selected conformations, and then its constituent models are re-trained after each iteration of data selection and labeling.

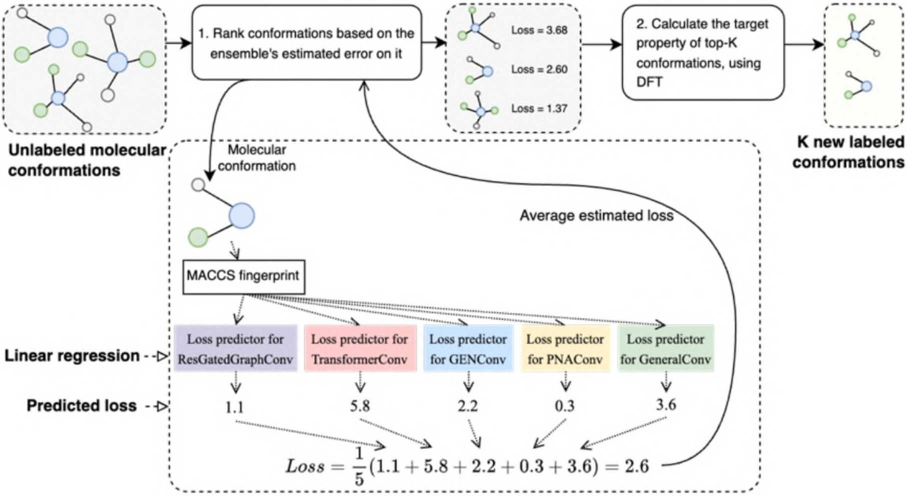
Unlabeled molecular conformations
1. Rank conformations based on the ensemble's estimated error on it
Loss = 3.68
Loss = 2.60
Loss = 1.37
2. Calculate the target property of top-K conformations, using DFT
K new labeled conformations
Molecular conformation
Average estimated loss
MACCS fingerprint
Linear regression
Loss predictor for ResGatedGraphConv
Loss predictor for TransformerConv
Loss predictor for GENConv
Loss predictor for PNAConv
Loss predictor for GeneralConv
Predicted loss
1.1
5.8
2.2
0.3
3.6
$Loss = \frac{1}{5}(1.1 + 5.8 + 2.2 + 0.3 + 3.6) = 2.6$

Figure 3: The labeling workflow of the SDDF: ML-based loss prediction approach

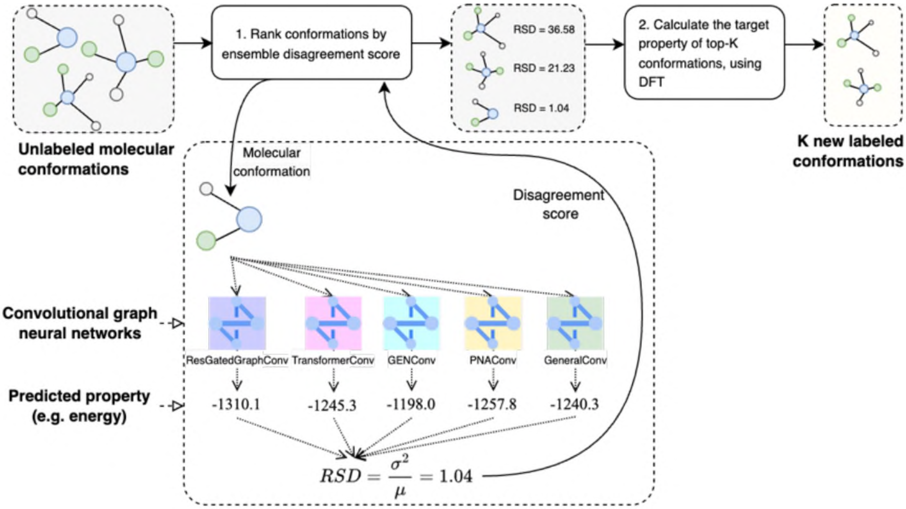
Unlabeled molecular conformations
1. Rank conformations by ensemble disagreement score
RSD = 36.58
RSD = 21.23
RSD = 1.04
2. Calculate the target property of top-K conformations, using DFT
K new labeled conformations
Molecular conformation
Disagreement score
Convolutional graph neural networks
ResGatedGraphConv
TransformerConv
GENConv
PNAConv
GeneralConv
Predicted property (e.g. energy)
-1310.1
-1245.3
-1198.0
-1257.8
-1240.3
$RSD = \frac{\sigma^2}{\mu} = 1.04$

Figure 4: The labeling workflow of the SDDF: Model disagreement-based approach

In addition to the active learning-based sampling of new data points, we also introduce novel approach for conformational sampling. We employ the top-performing models of SDDF ensemble to perform Langevin MD and generate new conformations for labeling. This approach uses a randomly selected small subset of the existing set of conformations, and generates the MD trajectories for each conformation in the subset. For the atoms in the generated conformations we obtain the forces using the gradient of the SDDF ensemble model's predicted energies. Selection of conformations is done based on *ConfidenceScore* metric. Conformations with lowest confidence score values are then selected for labeling.

$$ConfidenceScore = \frac{1}{N_{atoms}} \sum_{i=1}^{N_{atoms}} \min(\{\cos \angle(F_q^i, F_k^i) \mid q, k \in Ensemble, q \neq k\})$$

where F_m^i is the gradient of the m-th model's predicted energy with respect to the i-th atom's position. This selection approach showed much more positive impact on ensemble's ability to do MD than random selection of conformations from MD trajectories.

3. **Large-Scale Quantum Chemistry Datasets:** A central outcome of this thesis is generation and public release of substantial, high-quality dataset[4, 5] of molecular conformational energies (approx. 2.75 million by May 2025) from the ENAMINE REAL database, annotated with ωB97X/6-31G(d) DFT energies. This dataset was specifically curated to address the limitations of existing public resources, focusing on chemical diversity and relevance to drug discovery. The conformational space for these molecules was explored using a multi-pronged approach: initial conformer generation with RDKit (ETKDGv3 algorithm), optional geometry optimization using the MMFF94 force field, and further enrichment via Molecular Dynamics simulations driven by ML-derived forces as described in the active learning framework.

The full labeled dataset resulting from the SDDF project comprises **2,170,553** conformations, including **535,338** generated by RDKit, **1,151,936** by RDKit followed by MMFF94 optimization, and **483,279** generated via MD. A significant subset of this data has been meticulously prepared and released as a benchmark for training and evaluating energy prediction models. This benchmark dataset is characterized by a strict train-validation-test splitting methodology, which first applies a scaffold split (using the RDKit Bemis-Murcko framework) and then further refines the splits by applying a Tanimoto similarity filter (maximum 0.7 similarity between test and train scaffolds) to minimize data leakage and ensure a more realistic assessment of model generalization.

4. **GNN-based Models for Conformational Energy Prediction:** Leveraging the newly generated datasets, a suite of machine learning models for the accurate prediction of molecular conformational energies was developed and benchmarked. The core models are based on the five selected GCNN architectures (GeneralConv, PNAConv, GENConv, TransformerConv, and ResGatedGraphConv) that also form the ensemble within the SDDF active learning framework.

   These models take molecular conformation graphs as input, where nodes represent atoms and edges are defined by bonds and inter-atomic distances below a threshold. Several experiments were conducted for optimal node and edge representation choice. As a final choice of representations, node features consist of trainable embeddings for atom types, while edge features are a concatenation of embeddings for unique atom pairs, edge types (bond types or unspecified for non-bonded interactions), and an expanded version of rotation-invariant Point Pair Features (PPF-Diff variant), which proved beneficial for model performance.

   The models were trained using the Adam optimizer with a Mean Absolute Error (MAE) loss function, employing techniques such as dropout for regularization and a target energy shifting scheme based on estimated self-interaction atomic energies to facilitate learning.

   The performance of these individual models, as well as their ensembles (particularly an ensemble of the top three: PNAConv, ResGatedGraphConv, GENConv), was rigorously evaluated on the held-out SDDF test set. The results demonstrate that the SDDF-trained models, especially the ensemble, outperform the widely recognized ANI-2x ensemble in terms of both RMSE and MAE and generally show more stable error profiles across varying molecule sizes, unlike ANI-2x which exhibits noticeably higher MAE on molecules larger than its typical training distribution. The developed models and inference code are made publicly available, providing the community with accurate tools for energy prediction on diverse chemical structures.

5. **High-Performance Method and Tool for Symmetry-Corrected RMSD Calculation:** Addressing the critical need for reliable and efficient structural comparison, particularly for symmetric molecules, a novel, open-source tool named **FlashRMSD** was developed. The motivation for **FlashRMSD** stemmed from documented limitations in existing open-source tools[1]: spyRMSD's[15] inefficiency due to reliance on general graph libraries, DockRMSD's[16] high failure rates and restrictive file format support, and obrms's (part of Openbabel toolkit) potential overhead. **FlashRMSD** is designed for high performance and robustness, offering a comprehensive set of features including support for multiple molecular file formats (SDF, MOL, MOL2), handling of multi-conformer files, options for naive calculation for validation, inclusion/exclusion of hydrogen atoms, strict enforcement of bond type matching during atom

[15] R. Meli and P. C. Biggin, “spyrmsd: symmetry-corrected RMSD calculations in Python,” J Cheminform, vol. 12, p. 49, 2020

[16] E. W. Bell and Y. Zhang, “DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism,” J Cheminform, vol. 11, p. 40, 2019

mapping, optimal atom mapping reporting, cross-RMSD calculation (all-pairs RMSD within a single file), multi-query input support and minimized SC-RMSD calculation (≥*v1.1.0*).

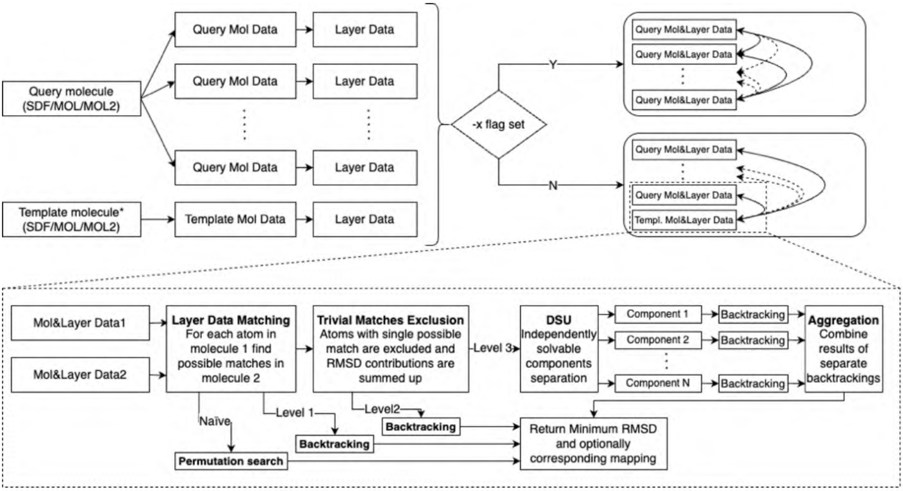

Figure 5: Flowchart of the FlashRMSD algorithm

The core of **FlashRMSD** employs an efficient two-stage algorithmic approach **(Figure 5)**. The first stage involves the generation of atom descriptors that effectively encode atom's neighborhood information, layer-by-layer, ensuring that chemically equivalent atoms have matching descriptors. This heavy featurization is particularly advantageous for cross-RMSD calculations. The second stage performs atom mapping via an optimized (early stopping based pruning) backtracking algorithm with multiple optimization levels:

- *Level 1* offers naive backtracking,
- *Level 2* resolves trivial one-to-one matches before backtracking,
- *Level 3* (default) further decomposes the problem by identifying and processing independent molecular blocks using a Disjoint Set Union (DSU) structure based on descriptor matches or bonding, significantly pruning the search space.

Extensive benchmarking demonstrated that **FlashRMSD** consistently outperforms existing tools like DockRMSD (for fair algorithmic comparison we enhanced this tool in terms of reliability) and obrms in terms of mean runtime for both cross-RMSD and all-to-all pairwise RMSD calculations, often by a significant margin (e.g., ~4 times faster in cross-RMSD).

Introduced in **v1.1.0**, the **FlashRMSD-M** mode (-m flag) calculates the minimized root-mean-square deviation (RMSD). It explores the full space of molecular isomorphisms, applying the Kabsch alignment algorithm to each potential atom mapping. This process guarantees the identification of the single best alignment that corresponds to the true minimum RMSD, a crucial feature for the accurate structural comparison of symmetric molecules.

Additional benchmark was conducted to capture each tool's runtime growth as number of symmetries in molecules grow. This benchmark was conducted both for SC-RMSD and minimized SC-RMSD calculations. In both cases **FlashRMSD** remained confident with higher number of symmetries (100K+), while other tools showed exponential runtime growth.

6. **Comprehensive Benchmark Dataset for SC-RMSD Tool Evaluation:** To facilitate rigorous and standardized evaluation of symmetry-corrected RMSD calculation tools, a comprehensive benchmark dataset was generated and published as part of the **FlashRMSD** study. This was motivated by the observation that existing evaluations often relied on ad-hoc molecule collections, failing to capture the full spectrum of symmetry challenges.

   The new benchmark dataset[6] was constructed using molecules from two primary, structurally diverse sources: the Chemical Component Dictionary[17] (CCD) and the Biologically Interesting molecule Reference Dictionary[18] (BIRD), both obtained from the RCSB Protein Data Bank (PDB). As of February 2024, this involved processing **45,622** molecules from CCD and **819** from BIRD. After preprocessing and initial conformer generation (primarily using RDKit's EmbedMolecule with MMFF94 optimization, and OpenBabel as a fallback) and filtering out molecules with fewer than five heavy atoms, **45,706** unique molecular structures were left. For each of these structures, up to nine docked conformations were generated using **SMINA** docking tool against the HIV-1 protease target (PDB ID: 1EBY[19]), chosen for its symmetrical dimeric structure and large, accommodating binding pocket. These conformations were saved in both multi-conformer and individual MOL2 and SDF files, creating a systematically organized dataset. Statistical analysis of the benchmark molecules revealed a wide range of heavy atom counts, a typical range of 3 to 6 distinct atom types, and a broad distribution of molecular symmetries as quantified by molecular graph automorphism counts computed using nauty&Traces[20].

[17] https://www.wwpdb.org/data/ccd
[18] https://www.wwpdb.org/data/bird
[19] https://www.rcsb.org/structure/1EBY
[20] https://pallini.di.uniroma1.it

List of author’s publications

[1] Altunyan, V., *Comparative Analysis of Symmetry-Corrected RMSD Calculation Tools in Molecular Docking*. Vestnik RAU, **1**, 25-36, (2024).

[2] Ghukasyan, T., Altunyan, V., Bughdaryan, A., Smbatyan, K., Aghajanyan, T., Papoian, G. A., & Petrosyan, G., *Smart Distributed Data Factory Volunteer Computing Platform for Active Learning-Driven Molecular Data Acquisition*. Sci Rep **15**, 7122 (2025). https://doi.org/10.1038/s41598-025-90981-6

[3] Altunyan, V., *FlashRMSD: An Effective Approach for Symmetry-Corrected RMSD Calculation with Extensive Benchmark Analysis*. Mathematical Problems of Computer Science, **63**, 81–101, (2025).

List of published datasets

[4] Altunyan, V., Ghukasyan, T., Bughdaryan, A., Aghajanyan, T., Smbatyan, K., Papoian, G., & Petrosyan, G. (2024). *SDDF Energy Dataset **(2024-Q3)*** [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15359529

[5] Altunyan, V., Ghukasyan, T., Bughdaryan, A., Aghajanyan, T., Smbatyan, K., Papoian, G., & Petrosyan, G. (2025). *SDDF Energy Dataset **(2025-Q1)*** [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14008357

[6] Altunyan, V. (2025). *Benchmark Dataset for Symmetry-Corrected RMSD Tools (FlashRMSD Study) **(1.0.0)*** [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15097621

Ամփոփում

Վահագն Նորիկի Ալթունյան

Մեքենայական ուսուցման և բաշխված հաշվարկային մոտեցումներ քվանտային քիմիական տվյալների ստեղծման և մոլեկուլային հատկությունների կանխատեսման համար

Աշխատանքը նվիրված է հաշվողական քիմիայի առանցքային խնդիրներից մեկին՝ բարձր ճշգրտության, բազմազան, մեքենայական ուսուցման (ՄՈւ) մոդելների արդյունավետ ուսուցմանը հարմարեցված, մոլեկուլային հատկությունների բազաների ստեղծմանը, ինչպես նաև այդ տվյալների հիման վրա որակյալ կանխատեսող մոդելների կառուցմանը:

Աշխատանքի **հիմնական նպատակներն** են՝

- Ստեղծել բաշխված հաշվարկների ենթակառուցվածք, որը հարմարեցված է քվանտային քիմիական հաշվարկների լայնամասշտաբ իրականացմանը:
- Մշակել տեսական և ալգորիթմական մեթոդներ, կառուցվող տվյալների բազաներում նոր մոլեկուլների և կառուցվածքների ներառման և քիմիական տարածության օպտիմալ հետազոտման համար: Սա ներառում է.
  - Մոլեկուլային տարածության նմուշառման մեթոդներ, որոնք ապահովում են բազմազան մոլեկուլների ընդգրկում, դիտարկվող մոլեկուլային հատկություններից կախված:
  - Կոնֆորմացիոն տարածության նմուշառման մեթոդներ, որոնք ընդգրկում են ներքին էներգիայի տեսանկյունից էական կառուցվածքային տատանումները:
  - Մոլեկուլների կառուցվածքային բազմազանեցման և ավելորդությունների նվազեցման համակարգված մոտեցումներ, որոնք առավելագույնի են հասցնում ստեղծված տվյալների տեղեկատվական արժեքը, ինչպես նաև կանխում են հաշվողական ռեսուրսների կրկնակի ծախսերը:

Աշխատանքի **գիտական նորույթ** պարունակող առավել կարևոր դրույթները հետևյալն են՝

- Ներկայացվել է SDDF հարթակը, որն ապահովում է ակտիվ ուսուցման և կամավորական ռեսուրսներով հաշվարկների նորարարական ինտեգրում: Համակարգը թեթև է և դինամիկ, ինչի շնորհիվ կարող է հեշտությամբ հարմարվել ապագա նոր նախագծերի հաշվողական պահանջներին:
- Ներկայացվել է ակտիվ ուսուցման համակարգը, մասնավորապես.
  - ներկայացվել է նոր մոտեցում մոլեկուլային տարածությունից նմուշառման համար, որը հիմնված է ՄՈւ մոդելներով հատկություն կանխատեսող մոդելի սխալանքի կանխատեսման վրա:

    - ներկայացվել է նոր մոտեցում կոնֆորմացիոն տարածությունից նմուշառման համար, որը հիմնված է GNN անսամբլի մոդելների միջև, կանխատեսվող ատոմական ուժերի փոխադարձ անհամաձայնության գնահատման վրա:

- Ներկայացվել են մոլեկուլային էներգիայի տվյալների նոր բազաներ՝ ներառելով ինչպես բուն տվյալների բազան, այնպես էլ խիստ կառուցվածքային բաժանման ենթարկված բազա՝ ապահովված ուսուցման-վալիդացիայի-թեստավորման(train-validation-test) բաժանումներով, որոնք երաշխավորում են խիստ կառուցվածքային տարբերություն ուսուցման և թեստավորման բաժանումների միջև: Նման բաժանումով տվյալների բազան թույլ է տալիս ստուգել էներգիա կանխատեսող մոդելների ընդհանրացվելու իրական հնարավորությունները: Նշված բազաների վրա մարզված ՄՈւ մոդելները, ցույց են տվել զգալի առավելություն մոլեկուլային էներգիա կանխատեսող առաջադեմ այլ մոդելների նկատմամբ:
- Ներկայացվել է **FlashRMSD** գործիքը, որն իրականացնում է սիմետրիայով ճշգրտված RMSD-ի հաշվարկման նոր ալգորիթմ: Այն օգտագործում է մեր կողմից սահմանված համապարփակ ատոմային նկարագրիչներ (descriptors), ինչպես նաև հետընթաց որոնման (backtracking) և էտման (pruning) նոր մոտեցումներ՝ ապահովելով հուսալիություն և բարձր արագագործություն: Գործիքի հետ մեկտեղ ներկայացվել է նաև տվյալների բազա, որը պարունակում է մոլեկուլային սիմետրիաների տեսանկյունից բազմազան մոլեկուլային կառուցվածքներ և որի հիման վրա կատարվել է համեմատություն գոյություն ունեցող նման գործիքների հետ:

Заключение

Алтунян Ваагн Норикович

Подходы машинного обучения и распределенных вычислений для генерации квантово-химических данных и предсказания молекулярных свойств

Работа посвящена одной из ключевых проблем вычислительной химии: созданию высокоточных, разнообразных баз данных молекулярных свойств, адаптированных для эффективного обучения моделей машинного обучения (МО), а также построению качественных предсказательных моделей на основе этих данных.

Основные цели работы:

- Создать инфраструктуру распределенных вычислений, адаптированную для крупномасштабного выполнения квантово-химических расчетов.
- Разработать теоретические и алгоритмические методы для включения новых молекул и структур в создаваемые базы данных и для оптимального исследования химического пространства. Это включает:
    - Методы выборки молекулярного пространства, обеспечивающие разнообразие молекул в зависимости от рассматриваемых молекулярных свойств.
    - Методы выборки конформационного пространства, охватывающие существенные структурные вариации с точки зрения внутренней энергии.
    - Систематические подходы к структурной диверсификации данных и снижению избыточности, максимизирующие информационную ценность данных и предотвращающие дублирование затрат вычислительных ресурсов.

Наиболее важные положения работы, содержащие **научную новизну**, следующие:

- Представлена платформа SDDF, обеспечивающая инновационную интеграцию активного обучения и распределённых вычислений на добровольных ресурсах. Система лёгкая и динамичная, легко адаптируемая под вычислительные требования будущих проектов.
- Представлена система активного обучения, в частности:
    - Новый подход к выборке из молекулярного пространства, основанный на прогнозировании ошибки предсказательных моделей МУ.
    - Новый подход к выборке из конформационного пространства, основанный на оценке расхождений предсказанных атомных сил между GNN ансамблями.
- Представлены новые базы данных молекулярной энергии, включающие как саму базу данных, так и базу со строгим структурным разделением с обеспеченными разделениями обучение-валидация-тест (train-validation-test), которые гарантируют строгое структурное различие между разделениями обучения и тестирования. База

данных с таким разделением позволяет проверить реальные возможности обобщения моделей предсказания энергии. МО модели, обученные на указанных базах, показали значительное преимущество по сравнению с другими передовыми моделями предсказания молекулярной энергии.

- Представлен инструмент FlashRMSD, реализующий новый алгоритм расчёта скорректированного по симметрии RMSD. Он использует разработанные нами комплексные атомные дескрипторы, а также новые подходы обратного поиска (backtracking) и обрезки (pruning), обеспечивая надёжность и высокую производительность. Вместе с инструментом представлена также база данных, содержащая разнообразные с точки зрения молекулярной симметрии молекулярные структуры, на основе которой проведено сравнение с существующими аналогичными инструментами.