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՝	…
	…
Առաջատար կազմակերպություն՝	…

Ատենախոսության պաշտպանությունը կկայանա 2025թ. հուլիսի …-ին, ժ. 15:00-ին 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 …-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:	…
	…
Leading organization:	…

The dissertation defence will take place on July …, 2025; at 15:00, at the Specialized Council 037 «Informatics» at the Institute of 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 …, 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. 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.

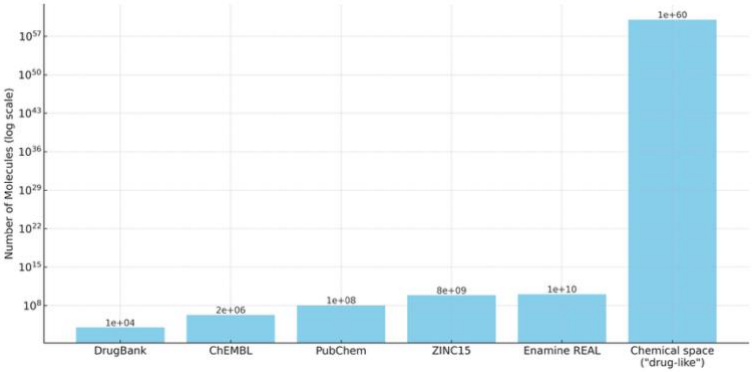

Figure 1: Chemical space ($\sim 10^{60}$ molecules) compared to known molecular datasets

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, juxtaposed with the hyper-dimensionality 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 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. **Experimental measurements** provide the ultimate ground truth but are often low-throughput, expensive, and may not be feasible for all properties or for vast numbers of compounds. At the other end of the spectrum, **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 entities or quantum mechanical phenomena. **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. Within this category, methods based on solving approximations to the Schrödinger equation [8], such as Density Functional Theory (DFT) [9], have become workhorses. While ab initio methods provide highly reliable data, they are the most computationally intensive. *This thesis focuses on leveraging ab initio methods, specifically DFT, for generating high-quality reference data due to their foundational accuracy, while simultaneously addressing the associated computational challenges.*

Generating accurate ab initio reference data is computationally demanding. Maximizing the utility of these expensive calculations for machine learning (ML) requires diverse datasets, avoiding redundant molecular structures. Consequently, quantitatively assessing molecular uniqueness is paramount, with molecular similarity metrics being essential for effective data curation and active learning strategies. In active learning, accurate similarity measures maximize information gain and crucially prevent wasteful re-labeling of conformations that are structurally identical or highly similar, especially when considering molecular symmetry. Robust structural comparison is thus key to efficiently building diverse, informative datasets and ensuring computational resources target genuinely novel structural information.

Challenges in Ab Initio Methods and Their Implementation

The fundamental properties and behavior of any given molecule are ultimately governed by the principles of quantum mechanics. The time-independent Schrödinger equation[8], $H\Psi = E\Psi$, 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 reactivity. 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:

- **Choice of Approximation Level:** The selection of an appropriate level of theory is a critical first step and involves choosing both a method (e.g., a specific DFT functional) and a basis set. DFT itself encompasses a vast array of exchange-correlation functionals (e.g., ωB97X [10], ωB97X-D [11]), each with different strengths and weaknesses for particular properties or molecular classes; no single functional is universally optimal. Similarly, the choice of basis set (e.g., Pople-style like 6-31G(d) [12] or specialized sets like def2-TZVP [13]) significantly impacts accuracy and computational cost, with larger, more flexible basis sets generally yielding more accurate results but at a substantially higher computational price. An inappropriate combination of functional and basis set can lead to inaccurate results or unmanageable computational demands.

- **Computational Cost: The Time-Accuracy Trade-off:** Ab initio calculations, particularly DFT, are computationally intensive, with costs typically scaling as a power (often N^3 to N^4, or higher for more sophisticated methods) of the number of basis functions, which correlates with molecular size. Consequently, pursuing higher accuracy—through larger basis sets, more advanced functionals, or post-Hartree-Fock [14] methods—invariably leads to a significant increase in computation time. Researchers must constantly navigate this trade-off, balancing the desired level of accuracy against available computational resources and project timelines.

- **Software and Tool Availability:** A variety of academic and commercial software packages are available for performing quantum chemical calculations, including well-known examples such as Gaussian [15], ORCA [16], Q-Chem [17], NWChem [18], and Psi4 [19]. Each package possesses its own distinct advantages, limitations, range of supported theoretical methods, parallelization efficiencies, and user interface paradigms. Effective utilization of these tools necessitates a degree of familiarity with the specific chosen software, encompassing the intricacies of input file preparation, the ability to correctly interpret complex output files, and the capacity to troubleshoot common computational issues and error conditions.

- **Resource Requirements:** Beyond the significant CPU time, ab initio calculations can impose substantial demands on other system resources, notably Random Access Memory (RAM) and disk storage. These requirements tend to escalate rapidly with the size of the molecule being studied and the extensiveness of the basis set employed.

Addressing these practical challenges is essential for any research endeavor that relies on generating or utilizing data from ab initio calculations. The complexity and resource intensity of these methods directly motivate the need for efficient data generation strategies, such as active learning and distributed computing, as explored in this thesis.

Challenges in Existing Quantum Mechanical Datasets

The efficacy of ML models in chemistry is profoundly 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:

- **ANI-1** [20] dataset, while large, is restricted to small organic molecules with only H, C, N, and O atoms (max 8 heavy atoms), limiting its chemical space representativeness for broader applications.

- **ANI-2x** [21] dataset, the successor of ANI-1, expanded atom types and provided multiple levels of theory but still predominantly features relatively small molecules.

- **NablaDFT** [22] dataset offers broader chemical space coverage than ANI-1 and includes more atom types, but its initial version's conformations were generated via RDKit without MD, and it was derived from the MOSES dataset, potentially not optimized for conformational analysis; NablaDFT 2.0 has started to address this by including relaxation trajectories.

- **GEOM** [23] dataset is extensive but contains molecules from QM9 and experimental sources, with the DFT level of theory for some subsets considered less accurate by some compared to ANI or NablaDFT.

- **QM9** [24] dataset is limited by providing at most one conformation per molecule, insufficient scaffold diversity, and significant train-test scaffold overlap, which can lead to data leakage and overoptimistic model evaluations.

- Other datasets like **MPCONF196** [25], **Transition1x** [26], **MD17** [27], and **MD22** [28] are highly specialized, focusing on specific areas like peptides, reaction pathways, or MD trajectories for a very limited number of systems, and thus do not offer the broad, diverse conformational energy landscapes required for general-purpose ML model development.

These limitations underscore the critical need for novel approaches to dataset curation. Specifically, there is a demand for methods that can generate large, diverse, and high-quality datasets covering therapeutically relevant chemical space, such as molecules from the ENAMINE database (ENAMINE Ltd., n.d.) [6]. 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. The bottleneck, therefore, is not just the computational cost of individual calculations, but the strategic generation and curation of datasets that are truly fit for the purpose of training robust and widely applicable ML models.

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, but its naive application can be misleading for symmetric molecules. Symmetry-corrected RMSD (SC-RMSD) addresses this by finding the optimal atomic mapping that minimizes the RMSD, effectively solving a graph isomorphism problem. As the graph isomorphism problem itself is known to not have a polynomial-time solution for general graphs (though practical algorithms exist for many specific graph classes, including molecular graphs), there is no universally efficient tool that entirely solves this problem for all molecular structures, particularly in the face of combinatorial explosion for larger or highly symmetric systems. Two variants of this metric are widely used in computational chemistry; however, tools for calculating both encounter significant issues.

- The first variant focuses on determining the SC-RMSD value given a fixed orientation of the two molecules, primarily solving the atom-mapping or graph isomorphism problem. Tools like DockRMSD [29] provide lightweight implementation, but suffer from high failure rates, segmentation faults, and intractably long runtimes for certain molecular topologies. SpyRMSD [30], relying on general graph libraries like NetworkX, offers flexibility but at the cost of efficiency, often proving "prohibitively slow" for practical applications and having limited support for nuanced chemical features like bond-type variations. These issues highlight that even without the added complexity of optimal superposition, the core atom-mapping step for symmetric molecules remains a significant hurdle for existing tools, impacting their reliability and throughput.

- The second variant, often referred to as minimized SC-RMSD, combines the symmetry-corrected atom mapping with the simultaneous optimization of the relative translation and rotation of one molecule with respect to the other to achieve the lowest possible SC-RMSD value. This is the more common requirement in structural alignment tasks. Obrms, from the OpenBabel suite [31], is a widely used tool that addresses this; however, it relies on iterative approaches or heuristics to explore the isomorphism space in conjunction with superposition algorithms. This can become computationally expensive, particularly for molecules with high degrees of symmetry. Another approach is introduced by tools like **pubchem_3d_align** [32] (integrated within RDKit [33]), which may use alternative strategies such as pharmacophore alignment or other feature-based methods to guide the superposition and similarity calculation. While potentially faster, these methods can introduce accuracy issues if the chosen features do not fully capture the relevant structural details or if the heuristic search is incomplete, leading to suboptimal alignments and RMSD values.

These documented issues across different approaches, ranging from outright crashes and incorrect outputs to prohibitive computational times or potential inaccuracies, highlight an unmet need for

improved SC-RMSD tools. Furthermore, the absence of a dedicated dataset that systematically tests RMSD performance on symmetrical structures means that evaluations often rely on ad hoc collections of molecules or focus on only a few specific chemotypes, failing to capture the breadth of real-world symmetry challenges. Robust and efficient SC-RMSD is critical for data acquisition pipelines to ensure that only unique structural information is subjected to expensive labeling, thereby maximizing the utility of computational resources.

2. Aim of the Work

This thesis addresses key challenges in computational molecular science, focusing on high quality data generation and structural analysis. The five main objectives are:

- Design and validation of a platform for large-scale DFT calculations via volunteer computing.
- Design active learning framework for efficient and informative molecular conformation selection.
- Generate and publish large-scale datasets of molecular energies, with a focus on diverse drug-like compounds.
- Develop machine learning models for accurate prediction of conformational energies using the new datasets.
- Develop fast and accurate tool for symmetry-corrected RMSD calculation.
- Generate and publish comprehensive benchmark dataset for comparison of symmetry-corrected RMSD calculation tools.

3. 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 and force 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, rigorously 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 across a wider swath of chemical space, thereby promoting innovation and reproducibility in the field.

- **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 datasets 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 these datasets 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.

4. Approbation of the Work

The key findings and methodologies developed in this dissertation were 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.

Publications

All results presented in this thesis are original and have been published in both local and international journals. The core findings are documented in 3 scientific articles. Additionally, 2 open-source datasets developed in this work have been published on Zenodo. A complete list of articles and datasets is provided at the end of the Synopsis.

5. Structure and Scope of Work

The dissertation consists of 5 chapters and a list of used literature. The thesis is written in 100 pages and has 100 literature references. The thesis contains 20 figures and 10 tables.

The thesis is organized as follows:

- *Chapter 1* serves as an introduction. It describes the problem, the main challenges in field of computational chemistry that ML models face and the aim of the thesis.
- *Chapter 2* introduces **SDDF** volunteer computing platform, its architecture and design choices.
- *Chapter 3.1* summarizes our research of **GNN**s on conformational energy prediction task on a pre-selected dataset.
- *Chapter 3.2* introduces **Active Learning** framework design, proposed methods for molecular sampling and comparison of approaches.
- *Chapter 3.3* introduces **Active Learning** framework extension for MD-driven conformational sampling, showcasing effects of conformational sampling on molecular dynamics stability.
- *Chapter 3.4* summarizes the datasets and models, generated and developed in this work. Also reporting current state of **SDDF** platform, ongoing projects and data bank content.
- *Chapter 4.1* introduces **FlashRMSD** tool for SC-RMSD calculation.
- *Chapter 4.2* introduces the comprehensive benchmark analysis between SC-RMSD calculation tools and discussion of case studies.
- *Chapter 4.3* introduces **FlashRMSD** extension for minimized SC-RMSD calculation. Discussion of case studies and comparison against widely acknowledged tools and their approaches.
- Finally, *Chapter 5* concludes the thesis with a summary of the contributions made.

6. The Main Results of the Work

The following points summarize the key contributions and findings:

1. **Platform for large-scale DFT calculations via volunteer computing:** 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 [34], [35], public-resource and peer-to-peer systems [36], [37], enterprise solutions [38], and versatile volunteer frameworks like BOINC [39], 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 such as fine-grained task management for molecular computations, efficiency for quantum chemistry methods, and streamlined handling of complex data for machine learning applications. SDDF was conceived to directly address this niche, offering a tailored solution for quantum chemistry research powered by public volunteers.

   Returning to SDDF's specifics, in the case of conformational energy for an average-sized molecule, a single-core machine is expected to calculate the property in about 10 minutes. The result of each task is a dictionary with property names as keys and respective calculated values.

Users can select the projects to which they want to contribute calculations, and they will receive computational tasks only from those projects. Otherwise, the platform assigns tasks from randomly selected projects.

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 and Client system, enabling volunteer nodes to retrieve molecular structures and submit results. Supporting these components are several scheduled 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.

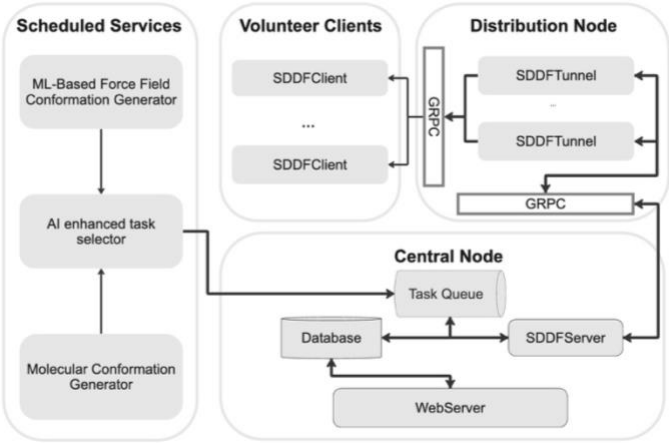

Figure 2: The architecture of the distributed computing system.

2. **Active learning framework for efficient and informative molecular conformation selection:** 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. The selection is performed based on an ensemble of ML models, which are used to determine the most challenging conformations among the generated set of conformations. The target property of the selected most challenging conformations is calculated using DFT. In addition, the selected conformations are used as initial points for MD calculations and the intermediate structures from the calculated trajectories are also labeled via DFT. All newly labeled examples are incorporated into the dataset and used to re-train the ML ensemble. Workflows for the molecule conformational energy dataset creation are illustrated in **Figures 3, 4**.

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.

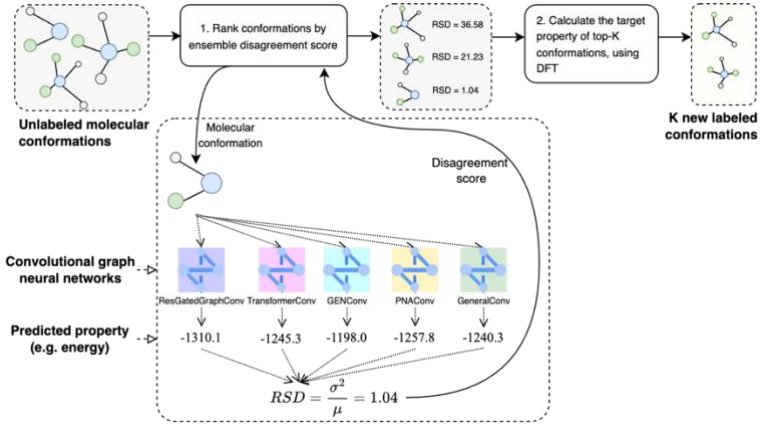

Figure 3: The labeling workflow of the SDDF: Model disagreement-based approach

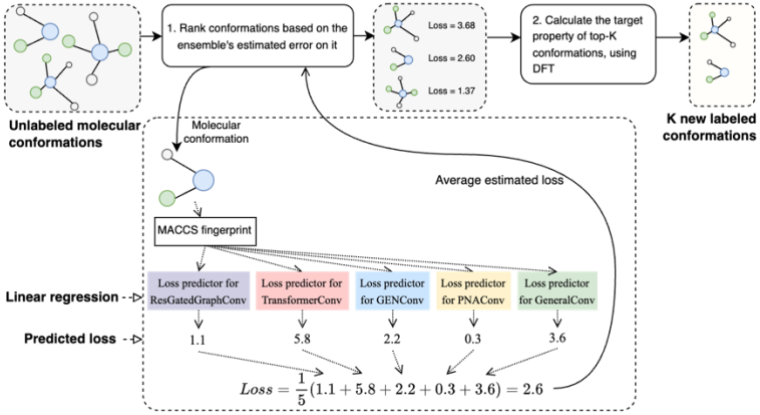

Figure 4: The labeling workflow of the SDDF: ML-based loss prediction approach

We use an ensemble of ML-based predictors where each predictor is a model trained separately as a regression problem that gets the molecular conformation graph as input and outputs an energy prediction. The nodes of the input graph are the molecule's atoms, and its adjacency matrix is constructed based on the bonds and distances between atoms (we considered an atom pair as adjacent if they have a bond or their distance is below a threshold value).

We performed initial model selection by training and evaluating 33 different graph convolutional neural network (GCNN) and Point Cloud architectures implemented in PyTorch Geometric [40] for the conformational energy prediction task. Based on the evaluation results we selected the 5 models with the best mean absolute error (MAE) scores on the validation set:

GeneralConv, PNAConv, GENConv, TransformerConv and ResGatedGraphConv models, as implemented in PyTorch Geometric. We further improved the models' performance by employing Point Pair Features for bonded atoms.

3. **Generation and Publication of Large-Scale Datasets of Molecular Energies:** A central outcome of this research is the generation and public dissemination of substantial, novel datasets of DFT-calculated molecular energies and their corresponding conformations. These datasets were specifically curated to address the limitations of existing public resources, focusing on chemical diversity and relevance to drug discovery by sourcing molecules primarily from the ENAMINE REAL database. The conformational space for these molecules was explored using a multi-pronged approach: initial conformer generation with RDKit (ETKDGv3 algorithm [41]) and OpenBabel, 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.

   All quantum chemical calculations for energy labeling were performed using the Psi4 toolkit at the ωB97X/6-31G(d) level of theory, a choice informed by its balance of accuracy and computational cost, and its precedent in established datasets like ANI. 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 [42], 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 [43]) 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. The resulting SDDF benchmark dataset demonstrates superior scaffold diversity compared to many existing datasets like QM9, ANI-1, and NablaDFT, and includes molecules of varying sizes, more representative of those encountered in drug discovery projects. These datasets are publicly available via Zenodo, providing a valuable resource for the broader scientific community.

4. **Development of Machine Learning Models for Accurate Prediction of Conformational Energies:** 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. 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) [44], which proved beneficial for model performance.

   Extensive experimentation with input features, including pre-trained Uni-Mol features, led to the selection of the current feature set for optimal balance of accuracy and inference speed. 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 [45] in terms of both RMSE and MAE, particularly for molecules containing bromine (which ANI-2x does not support) 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. **Development of a Fast and Accurate Tool for Symmetry-Corrected RMSD Calculation:** Addressing the critical need for reliable and efficient structural comparison, particularly for symmetric molecules, a novel software tool named **FlashRMSD** was developed. The motivation for **FlashRMSD** stemmed from documented limitations in existing open-source tools: spyRMSD's inefficiency due to reliance on general graph libraries, DockRMSD's high failure rates and restrictive file format support, and obrms's 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 (exhaustive permutation search) calculation for validation, inclusion/exclusion of hydrogen atoms, strict enforcement of bond order matching during atom mapping, verbose diagnostic output, atom-to-atom assignment reporting, cross-RMSD calculation (all-pairs RMSD within a single file), and multi-query input support.

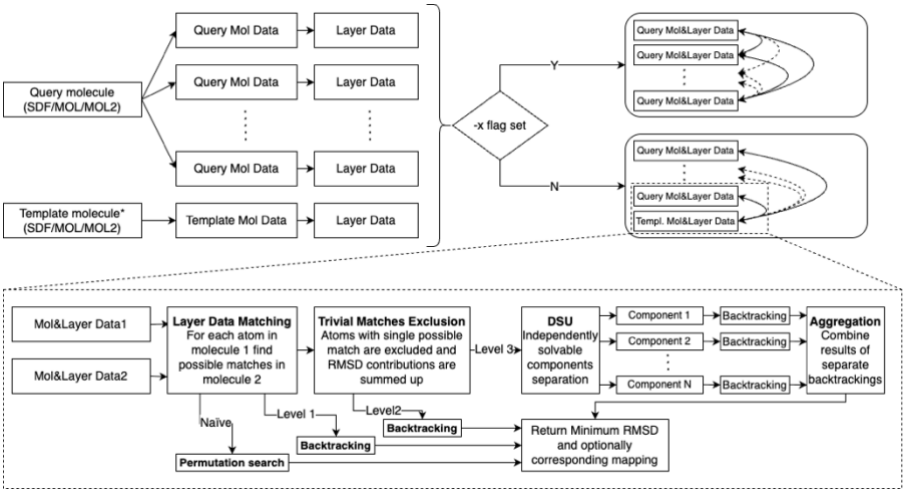

Figure 5: Flowchart of the FlashRMSD algorithm

The core of **FlashRMSD** employs an efficient two-stage algorithmic approach **(Figure 5)**. The first stage involves the generation of atom descriptors: each atom is featurized using a sorted array of values derived from a breadth-first traversal of the molecular graph rooted at that atom,

encoding periodic table numbers and graph distances (Layer Data), which are then hashed for rapid comparison. 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 levels of sophistication:

- *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 DockRMSDExt (an enhanced version of DockRMSD for fairer comparison [29]) 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). It also exhibited superior reliability, successfully processing many challenging cases (like CCD/PE3, CCD/330, CCD/60C, and BIRD/PRDCC_900031) where other tools failed or timed out.

The source code for **FlashRMSD** is publicly available on GitHub at https://github.com/altunyanv/FlashRMSD.

6. **Generation and Publication of a Comprehensive Benchmark Dataset for SC-RMSD Tools:** 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 was constructed using molecules from two primary, structurally diverse sources: the Chemical Component Dictionary (CCD) [46] and the Biologically Interesting molecule Reference Dictionary (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. Preprocessing included 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, resulting in a final set of **45,706** unique molecular structures. For each of these structures, up to nine docked conformations were generated using **SMINA** [47] (a fork of AutoDock Vina) against the HIV-1 protease target (PDB ID: 1EBY), 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 (5 to 244), a typical range of 3 to 6 distinct atom types, and a broad distribution of molecular symmetries as quantified by automorphism counts computed using nauty&Traces [48].

   This dataset, publicly available via Zenodo, along with the defined benchmark protocols, provides a robust platform for current and future assessment of SC-RMSD tools.

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).

3. Altunyan, V., *FlashRMSD: An Effective Approach for Symmetry-Corrected RMSD Calculation with Extensive Benchmark Analysis*. Mathematical Problems of Computer Science, **63**, 9-16.

List of published datasets

1. Altunyan, V., Ghukasyan, T., Bughdaryan, A., Aghajanyan, T., Smbatyan, K., Papoian, G., & Petrosyan, G. (2024). *SDDF Energy Dataset (**2024-Q3**)* [Data set]. Zenodo.

2. Altunyan, V., Ghukasyan, T., Bughdaryan, A., Aghajanyan, T., Smbatyan, K., Papoian, G., & Petrosyan, G. (2025). *SDDF Energy Dataset (**2025-Q1**)* [Data set]. Zenodo.

3. Altunyan, V. (2025). *Benchmark Dataset for Symmetry-Corrected RMSD Tools (FlashRMSD Study) (**1.0.0**)* [Data set]. Zenodo.

References

[1] J.-L. Reymond, "The chemical space project," *Acc Chem Res*, vol. 48, no. 3, pp. 722–730, 2015, doi: 10.1021/ar500432k.

[2] C. Knox *et al.*, "DrugBank 6.0: the DrugBank Knowledgebase for 2024," *Nucleic Acids Res*, vol. 52, no. D1, pp. D1265–D1275, Jan. 2024, doi: 10.1093/nar/gkad976.

[3] M. F. Adasme *et al.*, "The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Res*, vol. 51, no. D1, pp. D1401–D1413, Jan. 2023, doi: 10.1093/nar/gkad1004.

[4] S. Kim *et al.*, "PubChem 2023 update," *Nucleic Acids Res*, vol. 51, no. D1, pp. D1373–D1380, Jan. 2023, doi: 10.1093/nar/gkac956.

[5] T. Sterling and J. J. Irwin, "ZINC 15 – Ligand Discovery for Everyone," *J Chem Inf Model*, vol. 55, no. 11, pp. 2324–2337, Nov. 2015, doi: 10.1021/acs.jcim.5b00559.

[6] ENAMINE Ltd., "ENAMINE REAL Database."

[7] J. A. Pople, "Development of ab initio methods in quantum chemistry," *Rev Mod Phys*, vol. 71, no. 5, pp. 1267–1274, Oct. 1999, doi: 10.1103/RevModPhys.71.1267.

[8] E. Schrödinger, "Quantisierung als Eigenwertproblem (Erste Mitteilung)," *Ann Phys*, vol. 384, no. 4, pp. 361–376, 1926, doi: 10.1002/andp.19263840404.

[9] W. Kohn, A. D. Becke, and R. G. Parr, "A perspective on density functional theory," *J Phys Chem*, vol. 100, no. 31, pp. 12974–12980, 1996, doi: 10.1021/jp960669l.

[10] J.-D. Chai and M. Head-Gordon, "Systematic optimization of long-range corrected hybrid density functionals," *J Chem Phys*, vol. 128, no. 8, p. 84106, 2008, doi: 10.1063/1.2834918.

[11] J.-D. Chai and M. Head-Gordon, "Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections," *Physical Chemistry Chemical Physics*, vol. 10, no. 44, pp. 6615–6620, 2008, doi: 10.1039/B810189B.

[12] P. C. Hariharan and J. A. Pople, “The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies,” *Theor Chim Acta*, vol. 28, no. 3, pp. 213–222, 1973, doi: 10.1007/BF00533485.

[13] F. Weigend and R. Ahlrichs, “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy,” *Physical Chemistry Chemical Physics*, vol. 7, no. 18, pp. 3297–3305, 2005, doi: 10.1039/B508541A.

[14] C. C. J. Roothaan, “New Developments in Molecular Orbital Theory,” *Rev Mod Phys*, vol. 23, no. 2, pp. 69–89, Apr. 1951, doi: 10.1103/RevModPhys.23.69.

[15] M. J. Frisch *et al.*, “Gaussian 16, Revision C.01,” 2016, *Gaussian, Inc., Wallingford CT*.

[16] F. Neese, “Software update: the ORCA program system – Version 5.0,” *Wiley Interdiscip Rev Comput Mol Sci*, vol. 12, no. 5, p. e1606, 2022, doi: 10.1002/wcms.1606.

[17] E. Epifanovsky *et al.*, “Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package,” *J Chem Phys*, vol. 155, no. 8, p. 84801, 2021, doi: 10.1063/5.0055522.

[18] E. Aprà *et al.*, “NWChem: Past, present, and future,” *J Chem Phys*, vol. 152, no. 18, p. 184102, 2020, doi: 10.1063/5.0004997.

[19] D. G. A. Smith *et al.*, “Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry,” *J Chem Phys*, vol. 152, no. 18, p. 184108, 2020, doi: 10.1063/5.0006002.

[20] J. S. Smith, O. Isayev, and A. E. Roitberg, “ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules,” *Sci Data*, vol. 4, p. 170193, 2017, doi: 10.1038/sdata.2017.193.

[21] K. Huddleston *et al.*, “ANI-2x Release,” 2023, *Zenodo*. doi: 10.5281/zenodo.10108942.

[22] K. Khrabrov *et al.*, “nablaDFT: Large-Scale Conformational Energy and Hamiltonian Prediction benchmark and dataset,” *Physical Chemistry Chemical Physics*, vol. 24, no. 42, pp. 25853–25863, 2022, doi: 10.1039/d2cp03966d.

[23] S. Axelrod and R. Gómez-Bombarelli, “GEOM, energy-annotated molecular conformations for property prediction and molecular generation,” *Sci Data*, vol. 9, no. 1, p. 185, 2022, doi: 10.1038/s41597-022-01288-4.

[24] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, “Quantum chemistry structures and properties of 134 kilo molecules,” *Sci Data*, vol. 1, p. 140022, 2014, doi: 10.1038/sdata.2014.22.

[25] J. Řezáč, D. B\’\im, O. Gutten, and L. Rul\’\išek, “Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set,” *J Chem Theory Comput*, vol. 14, no. 3, pp. 1254–1266, 2018, doi: 10.1021/acs.jctc.7b01074.

[26] M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther, “Transition1x-a dataset for building generalizable reactive machine learning potentials,” *Sci Data*, vol. 9, no. 1, p. 779, 2022, doi: 10.1038/s41597-022-01879-5.

[27] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, “Machine learning of accurate energy-conserving molecular force fields,” *Sci Adv*, vol. 3, no. 5, p. e1603015, 2017, doi: 10.1126/sciadv.1603015.

[28] S. Chmiela *et al.*, “Accurate global machine learning force fields for molecules with hundreds of atoms,” *Sci Adv*, vol. 9, no. 2, p. eadf0873, 2023, doi: 10.1126/sciadv.adf0873.

[29] 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, doi: 10.1186/s13321-019-0362-7.

[30] R. Meli and P. C. Biggin, “spyrmsd: symmetry-corrected RMSD calculations in Python,” *J Cheminform*, vol. 12, p. 49, 2020, doi: 10.1186/s13321-020-00455-2.

[31] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J Cheminform*, vol. 3, p. 33, 2011, doi: 10.1186/1758-2946-3-33.

[32] National Center for Biotechnology Information, "PubChem-Align3D," 2025, *U.S. National Library of Medicine*. [Online]. Available: https://github.com/ncbi/pubchem-align3d

[33] RDKit UGM Organizers and Contributors, "RDKit: Open-source cheminformatics."

[34] I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit," *The International Journal of Supercomputer Applications and High Performance Computing*, vol. 11, no. 2, pp. 115–128, 1997, doi: 10.1177/109434209701100205.

[35] I. Foster, "The Grid: A new infrastructure for 21st century science," *Phys Today*, vol. 55, no. 2, pp. 42–47, 2002, doi: 10.1063/1.1457275.

[36] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, "SETI@home: an experiment in public-resource computing," *Commun ACM*, vol. 45, no. 11, pp. 56–61, 2002, doi: 10.1145/581571.581573.

[37] D. Brookshier, D. Govoni, N. Krishnan, and J. C. Soto, *JXTA: Java P2P Programming*. Sams Publishing, 2002.

[38] A. A. Chien, B. Calder, S. Elbert, and K. Bhatia, "Entropia: architecture and performance of an enterprise desktop grid system," *J Parallel Distrib Comput*, vol. 63, no. 5, pp. 597–610, 2003, doi: 10.1016/S0743-7315(03)00031-6.

[39] D. P. Anderson, "BOINC: a platform for volunteer computing," *J Grid Comput*, vol. 18, no. 1, pp. 99–122, 2020, doi: 10.1007/s10723-019-09497-9.

[40] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," *ArXiv*, 2019.

[41] S. Wang, J. Witek, G. A. Landrum, and S. Riniker, "Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences," *J Chem Inf Model*, vol. 60, no. 4, pp. 2044–2058, 2020, doi: 10.1021/acs.jcim.0c00025.

[42] P. Tosco, N. Stiefl, and G. A. Landrum, "Bringing the MMFF force field to the RDKit: implementation and validation," *J Cheminform*, vol. 6, p. 37, 2014, doi: 10.1186/1758-2946-6-37.

[43] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. Molecular frameworks," *J Med Chem*, vol. 39, no. 15, pp. 2887–2893, 1996, doi: 10.1021/jm9602928.

[44] H. Deng, T. Birdal, and S. Ilic, "PPFNet: Global Context Aware Local Features for Robust 3D Point Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2018, pp. 195–205.

[45] M. Rezaee, S. Ekrami, and S. M. Hashemianzadeh, "Comparing ANI-2x, ANI-1ccx neural networks, force field, and DFT methods for predicting conformational potential energy of organic molecules," *Sci Rep*, vol. 14, no. 1, p. 11791, May 2024, doi: 10.1038/s41598-024-62684-3.

[46] J. D. Westbrook, C. Shao, Z. Feng, M. Zhuravleva, S. Velankar, and J. Young, "The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank," *Bioinformatics*, vol. 31, no. 8, pp. 1274–1278, Apr. 2015, doi: 10.1093/bioinformatics/btu789.

[47] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, "Empirical scoring with smina from the CSAR 2011 benchmarking exercise," *J Chem Inf Model*, vol. 53, no. 8, pp. 1893–1904, 2013, doi: 10.1021/ci300604z.

[48] B. D. McKay and A. Piperno, "Practical Graph Isomorphism, II," *J Symb Comput*, vol. 60, pp. 94–112, 2014, doi: 10.1016/j.jsc.2013.09.003.

Ամփոփում

Վահագն Նորիկի Ալթունյան

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

Աշխատանքը նվիրված է հաշվողական քիմիայում առկա առանցքային մարտահրավերներին՝ մասնավորապես, բարձ որակի մոլեկուլային տվյալների լայնածավալ գեներացմանը և մոլեկուլային կառուցվածքների ճշգրիտ վերլուծությանը: Հետազոտությունը կենտրոնանում է արհեստական բանականության վրա հիմված նոր մոտեցումների, կամավորական հաշվարկների հարթակի, ինչպես նաև մոլեկուլային համեմատության կատարելագործված գործիքների մշակման ու վավերացման վրա:

Աշխատանքի հիմնական նպատակներն են՝

- Նախագծել կամավորական ռեսուրսներով հաշվարկների վրա հիմված հարթակ քվանտային քիմիական հաշվարկների իրականացման համար:
- Մշակել ակտիվ ուսուցման ալգորիթմ՝ մոլեկուլային կառուցվածքների արդյունավետ և տեղեկատվական ընտրության համար:
- Ստեղծել և հրապարակել մոլեկուլային էներգիաների լայնածավալ տվյալների բազաներ՝ կենտրոնանալով **դեղանման** (**“drug-like”**) մոլեկուլների բազմության վրա:
- Մշակել մեքենայական ուսուցման մոդելներ՝ նոր տվյալների բազաների հիման վրա մոլեկուլային էներգիաների ճշգրիտ կանխատեսման համար:
- Մշակել արագագործ և ճշգրիտ գործիք՝ սիմետրիայով ճշգրտված միջին քառակուսային շեղման (SC-RMSD) հաշվարկման համար:
- Գեներացնել և հրապարակել մոլեկուլային կառուցվածքների տվյալների բազա՝ սիմետրիայով ճշգրտված միջին քառակուսային շեղում հաշվարկող գործիքների համեմատության համար:

Ատենախոսության **առաջին գլխում** ներկայացվել են հետազոտության արդիականությունը, հաշվողական քիմիայի բնագավառում մեքենայական ուսուցման մոդելների առջև ծառացած հիմնական մարտահրավերները և աշխատանքի նպատակները:

Երկրորդ գլուխը նվիրված է **SDDF (Smart Distributed Data Factory)** կամավորական ռեսուրսներով հաշվարկների հարթակի ներկայացմանը, դրա ճարտարապետությանը և նախագծման հիմնական սկզբունքներին:

Երրորդ գլխում ներկայացվում է ակտիվ ուսուցման համակարգը: Մասնավորապես դիտարկվում է գրաֆային նեյրոնային ցանցերի (GNN) կիրառելիությունը մոլեկուլային

էներգիայի կանխատեսման համար, որից հետո ներկայացնում է ակտիվ ուսուցման համակարգը՝ նոր կառուցվածքների ընտրման տարբեր մեթոդների և դրանց համեմատության հետ միասին: Այս գլխում նաև քննարկվում է մոլեկուլային դինամիկայի (MD) վրա հիմնված կառուցվածքների ընտրության մոտեցումը և դրա ազդեցությունը MD-ի կայունության վրա: Գլուխն ամփոփվում է աշխատանքի ընթացքում գեներացված տվյալների բազաների, մշակված մոդելների, ինչպես նաև SDDF հարթակի ընթացիկ վիճակի և հեռանկարների ներկայացմամբ:

Չորրորդ գլուխը կենտրոնանում է սիմետրիայով ճշգրտված RMSD (SC-RMSD) հաշվարկման խնդիրների վրա: Այն ներկայացնում է **FlashRMSD** գործիքը, որը նախատեսված է SC-RMSD-ի արդյունավետ հաշվարկման համար: Այս գլխում իրականացվում է SC-RMSD հաշվարկող գործիքների համապարփակ համեմատական վերլուծություն՝ քննարկելով նաև առանձին դեպքեր: Բացի այդ, ներկայացվում է **FlashRMSD** գործիքի պարզ ընդլայնումը՝ մինիմիզացված SC-RMSD-ի հաշվարկման համար, որը համեմատվում է լայնորեն կիրառվող այլ գործիքների և մոտեցումների հետ:

Հինգերորդ գլխում ամփոփվում են ատենախոսության հիմնական արդյունքները և կատարված ներդրումները:

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

- **SDDF հարթակ:** Ակտիվ ուսուցման և կամավորական ռեսուրսներով հաշվարկների նորարարական ինտեգրում՝ հատուկ հարմարեցված DFT-ի վրա հիմնված մոլեկուլային տվյալների գեներացման համար:

- **Ակտիվ ուսուցում:** Տարբեր ճարտարապետություններով GNN մոդելների համախմբի **(GeneralConv, PNAConv, GENConv, TransformerConv, ResGatedGraphConv)** ներդրում՝ մեքենայական ուսուցման և մոլեկուլային դինամիկայի վրա հիմնված մոտեցումների հետ համատեղ՝ նոր մոլեկուլային կառուցվածքների ընտրության համար:

- **Նոր տվյալների բազաներ և մոդելներ:** ENAMINE տվյալների բազայից ստացված բարձրորակ մոլեկուլային կառուցվածքների և էներգիաների տվյալների բազաների և դրանց հիման վրա մարզված ճշգրիտ մեքենայական ուսուցման մոդելների հրապարակում:

- **FlashRMSD գործիք:** Սիմետրիայով ճշգրտված RMSD-ի հաշվարկման նոր ալգորիթմ, որն օգտագործում է համապարփակ ատոմային նկարագրիչներ (descriptors), ինչպես նաև հետընթաց որոնման (backtracking) և էտման (pruning) մոտեցումներ՝ հուսալիություն և բարձր արագագործություն ապահովելու համար:

- **SC-RMSD մետրիկայի համեմատական վերլուծություն:** CCD/BIRD տվյալների բազաների վրա հիմնված, մոլեկուլային կառուցվածքների համապարփակ բազա՝ գործիքների արագագործությունը և հուսալիությունը գնահատելու համար:

Заключение

Алтунян Ваагн Норикович

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

Работа посвящена ключевым проблемам вычислительной химии, в частности, широкомасштабной генерации высококачественных молекулярных данных и точному анализу молекулярных структур. Исследование сосредоточено на разработке и валидации новых подходов, основанных на искусственном интеллекте, платформы для добровольных вычислений, а также усовершенствованных инструментов для молекулярного сравнения.

Основные цели работы:

- Спроектировать платформу на основе добровольных вычислений для проведения квантово-химических расчетов.
- Разработать алгоритм активного обучения для эффективного и информативного отбора молекулярных структур.
- Создать и опубликовать широкомасштабные базы данных молекулярных энергий, уделяя особое внимание разнообразию фармакоподобных (“drug-like”) молекул.
- Разработать модели машинного обучения для точного предсказания молекулярных энергий на основе новых баз данных.
- Разработать быстрый и точный инструмент для расчета среднеквадратичного отклонения с поправкой на симметрию (SC-RMSD).
- Сгенерировать и опубликовать базу данных молекулярных структур для сравнения инструментов, вычисляющих среднеквадратичное отклонение с поправкой на симметрию.

В **первой главе** диссертации представлены актуальность исследования, основные проблемы в области вычислительной химии, с которыми сталкиваются модели машинного обучения, и цели работы.

Вторая глава посвящена представлению платформы добровольных вычислений **SDDF (Smart Distributed Data Factory)**, ее архитектуре и основным принципам проектирования.

В **третьей главе** представляется система активного обучения. В частности, рассматривается применимость графовых нейронных сетей (GNN) для предсказания молекулярной энергии, после чего представляется система активного обучения вместе с различными методами отбора новых структур и их сравнением. В этой главе также обсуждается метод отбора структур на основе молекулярной динамики (МД) и его

влияние на стабильность МД. Глава завершается представлением сгенерированных в ходе работы баз данных, разработанных моделей, а также текущего состояния и перспектив платформы SDDF.

Четвертая глава фокусируется на задачах вычисления RMSD с поправкой на симметрию (SC-RMSD). В ней представлен инструмент FlashRMSD, предназначенный для эффективного вычисления SC-RMSD. В этой главе проводится всесторонний сравнительный анализ инструментов для вычисления SC-RMSD с обсуждением также отдельных случаев. Кроме того, представлено простое расширение инструмента FlashRMSD для вычисления минимизированного SC-RMSD, которое сравнивается с широко используемыми другими инструментами и подходами.

В **пятой главе** обобщаются основные результаты диссертации и сделанные вклады.

Наиболее важные положения работы, содержащие научную новизну, следующие:

- **Платформа SDDF:** Инновационная интеграция активного обучения и добровольных вычислений, специально адаптированная для генерации молекулярных данных на основе DFT.
- **Активное Обучение:** Внедрение ансамбля GNN моделей с различными архитектурами (GeneralConv, PNAConv, GENConv, TransformerConv, ResGatedGraphConv) совместно с подходами на основе машинного обучения и молекулярной динамики для отбора новых молекулярных структур.
- **Новые Базы Данных и Модели:** Публикация высококачественных баз данных молекулярных структур и энергий, полученных из базы данных ENAMINE, а также точных моделей машинного обучения, обученных на их основе.
- **Инструмент FlashRMSD:** Новый алгоритм для вычисления RMSD с поправкой на симметрию, использующий комплексные атомарные дескрипторы, а также подходы обратного отслеживания (backtracking) и отсечения (pruning) для обеспечения надежности и высокой производительности.
- **Сравнительный Анализ Метрики SC-RMSD:** Комплексная база молекулярных структур на основе баз данных CCD/BIRD для оценки надежности инструментов и случаев их сбоя.