NEURAL INFORMATION PROCESSING SYSTEMS

FSU FLORIDA STATE UNIVERSITY

Stony Brook University DARE TO BE

Empowering Active Learning for 3D Molecular Graphs with Geometric Graph Isomorphism

Ronast Subedi[1*], Lu Wei[2*], Wenhan Gao[2*], Shayok Chakraborty[1], Yi Liu[2]

[1]Florida State University

[2]Stony Brook University

NeurIPS 2024

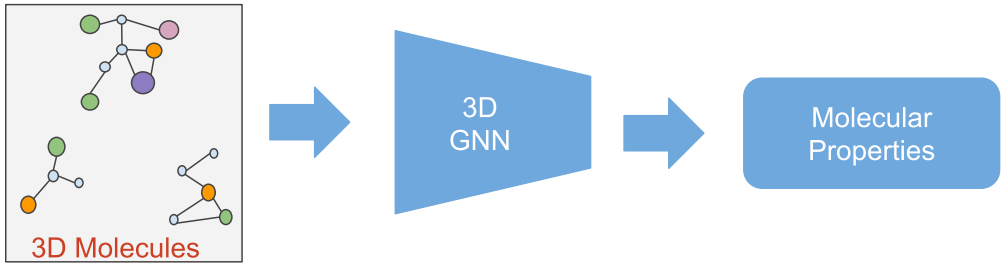
NEURAL INFORMATION PROCESSING SYSTEMS

FSU FLORIDA STATE UNIVERSITY

Stony Brook University DARE TO BE

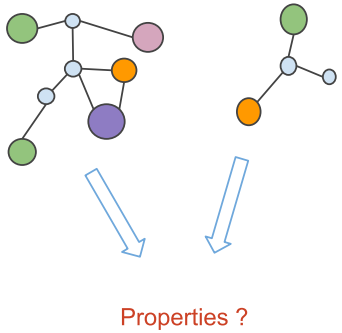
Introduction: 3D Molecular Learning

❑3D molecular learning learns 3D molecular representation to predict molecular properties



Challenges

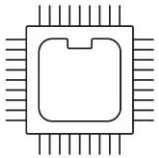
❑Annotating scientific data is difficult



Obtaining Annotations Require

Experts

High Cost



Extensive Computation

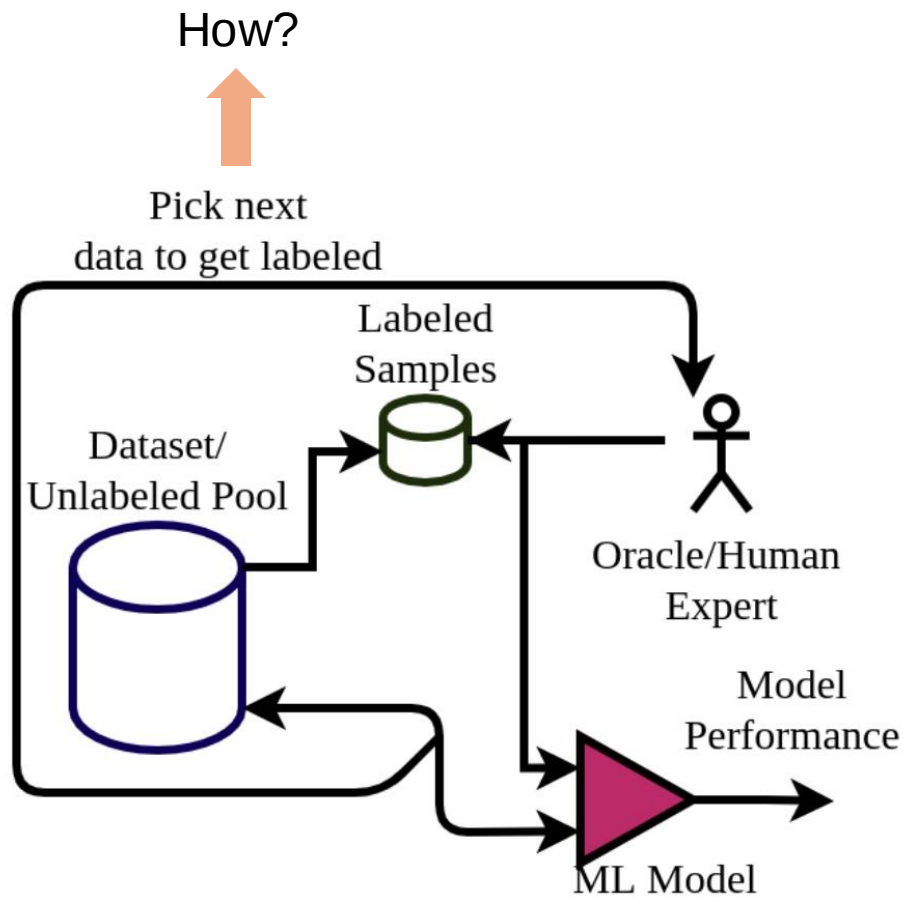
❑**Density Functional Theory (DFT)** can be used to label molecular energy, but it is very slow

NEURAL INFORMATION PROCESSING SYSTEMS

FSU FLORIDA STATE UNIVERSITY

Stony Brook University DARE TO BE

Introduction: Active Learning(AL) for 3D Molecular Graphs



Motivation

- *Pressing need for 3D molecular graphs*
- 3D geometric configuration is **crucial**
- ***Need to incorporate specialized knowledge of geometric configuration into AL!***

Method:

- Both **diversity** and **uncertainty** for 3D GNNs
- Diversity: How a 3D molecular graph is different from others
- Uncertainty: How the model is confident about a 3D molecular graph

Proposed Diversity Component

- Two 3D molecules can have **different number of atoms**, how to measure their diversity?
- Solution: **Study Geometric Graph Isomorphism for diversity.**
  - Three isometries: *reference distance*, *triangular*, and *cross-angle* are used as basis for expressive representation of 3D molecular graphs.
  - Expressive power: *reference distance* < *triangular* < *cross-angle*

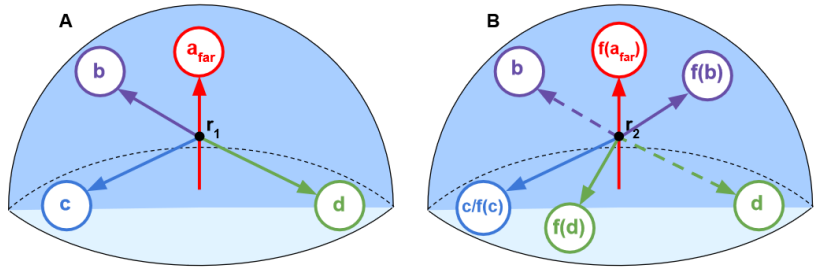
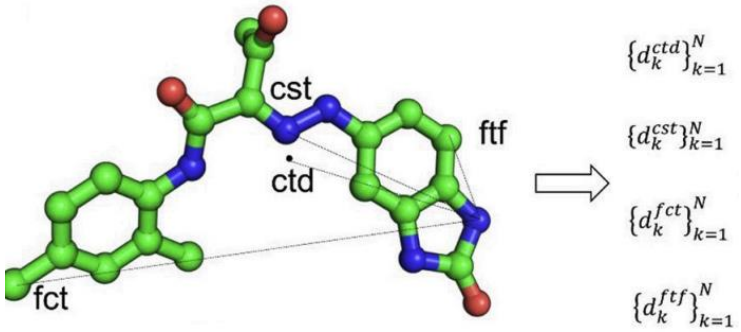
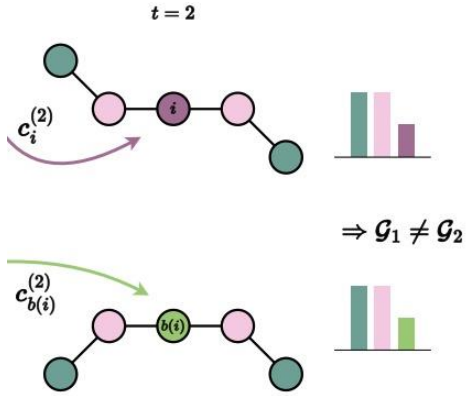


Figure 2: A and B are triangular isometric but not cross-angular isometric. The angles $\angle br_1a_{far}$, $\angle cr_1a_{far}$, and $\angle dr_1a_{far}$ in structure A are equal to the angles $\angle f(b)r_2f(a_{far})$, $\angle f(c)r_2f(a_{far})$, and $\angle f(d)r_2f(a_{far})$ in structure B, respectively. However, the cross angle $\angle dr_1c$ is not equal to the cross angle $\angle f(d)r_2f(c)$.

Proposed Diversity Component





- ❑ Encode the isometries into a geometric descriptor by sets of statistical moments.
- ❑ This geometric descriptor preserves Euclidean motion and permutation symmetries.
- ❑ Our method is **at least as expressive as the GWL test** (the descriptor suffices to distinguish any non-isomorphic molecular structures that are distinguishable by any 3D).

Uncertainty Component

- Select a batch of samples with high uncertainty values

- Computed by:

$$\widehat{\sigma^2}(o^* \mid g^*) = \frac{1}{N}\sum_{n=1}^{N}\left(\widehat{o_n^*}\right)^2 - \left(\frac{1}{N}\sum_{n=1}^{N}\widehat{o_n^*}\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\widehat{\sigma_n^2}$$

- $\widehat{\sigma^2}(o^* \mid g^*)$: Uncertainty of input g^*
- $\left(\widehat{o_n^*}\right)^2$ and $\left(\frac{1}{N}\sum_{n=1}^{N}\widehat{o_n^*}\right)^2$: Variance of all the sampled outputs for the input g^*
- $\frac{1}{N}\sum_{n=1}^{N}\widehat{\sigma_n^2}$: Variance that is same across all the data samples

o_n^* is the output of n^{th} Bayesian Geometric Graph Neural Network(BGGNN)

Proposed Framework

Active Sample Selection

Select samples the model is most uncertain about

Select a diverse set of samples

$$\max_z \quad z^\top r + \lambda z^\top D z$$
$$s.t. \quad \sum_{i=1}^{N} z_i = k$$
$$z_i \in \{0, 1\}, \forall i$$

Only k samples can be queried in one AL iteration

is a binary vector

- ❑ The problem is equivalent to standard Quadratic Programming(QP) optimization problem
- ❑ We relax the integer constraint into continuous constraints and solve it using GPU implementation of QP solver

Experimental Setup

Dataset

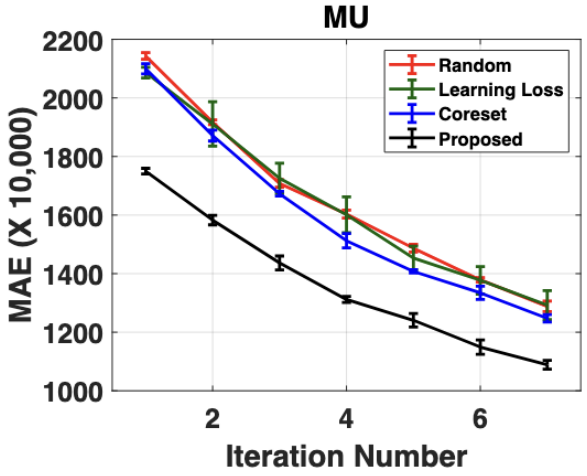
- A subset of ***QM9*** dataset
- Aspirin molecule of ***MD17*** dataset

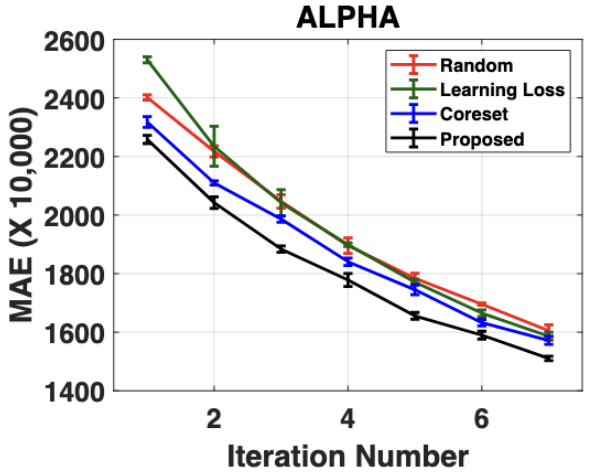
AL Comparison Baselines

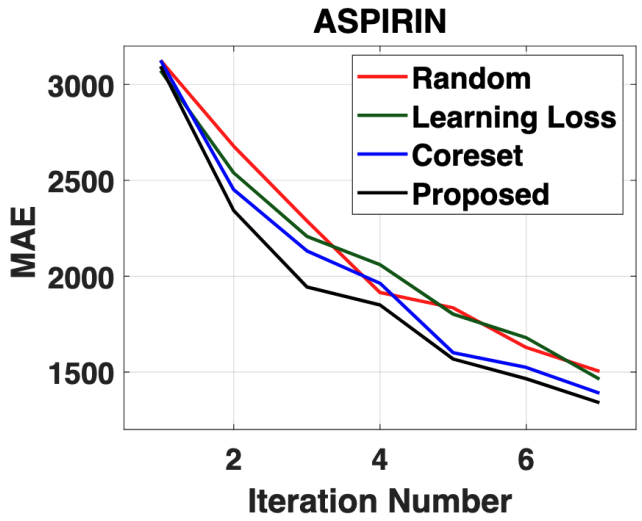
- Random
- Coreset
- Learning Loss

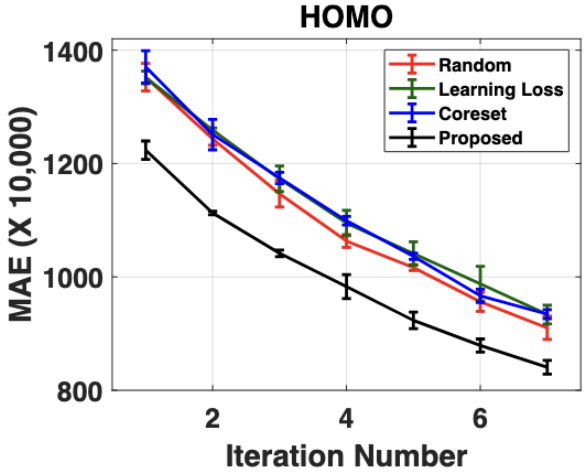
Dataset	QM9	MD17(Aspirin)
Train Size	25000	1000
Validation Size	10000	1000
Test Size	10831	1000

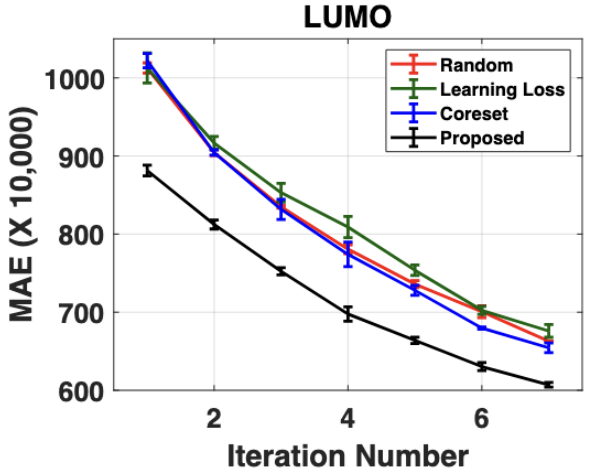
Results





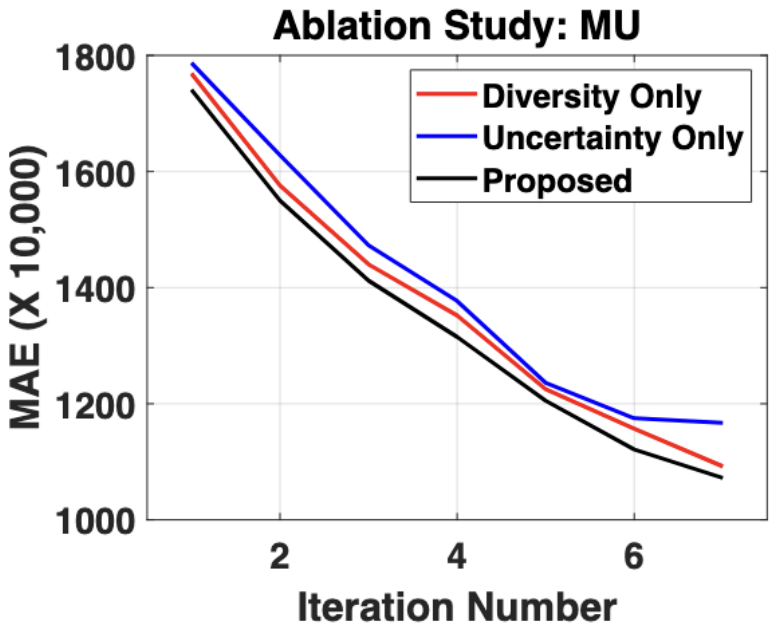


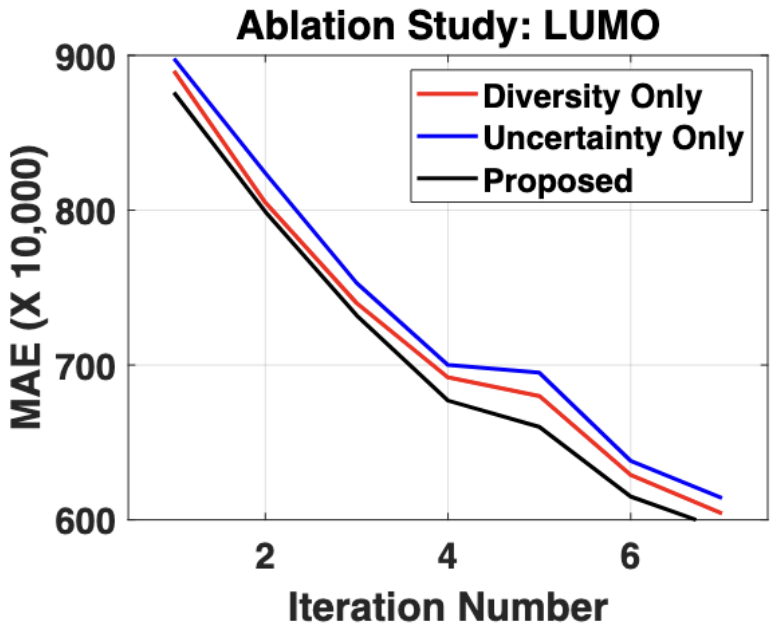




- The proposed method consistently outperforms the baselines by attaining the lowest ***MAE*** values in each AL iteration on ***4 properties of QM9 dataset(left)*** and ***Aspirin molecule of MD17 dataset(right)***

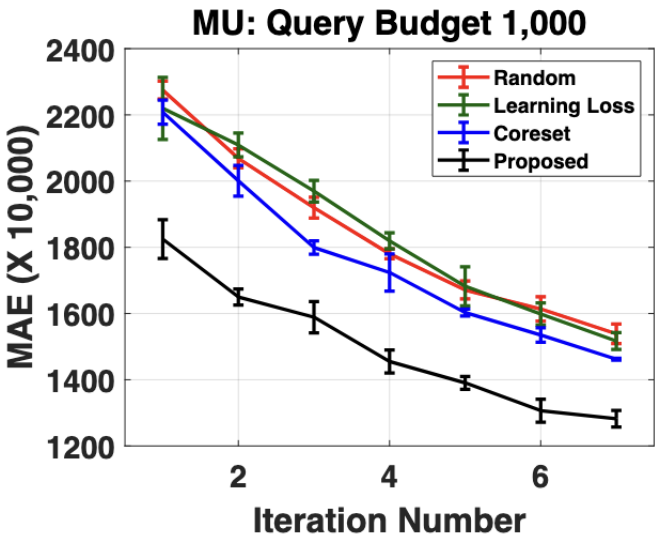
Results: Ablation Study

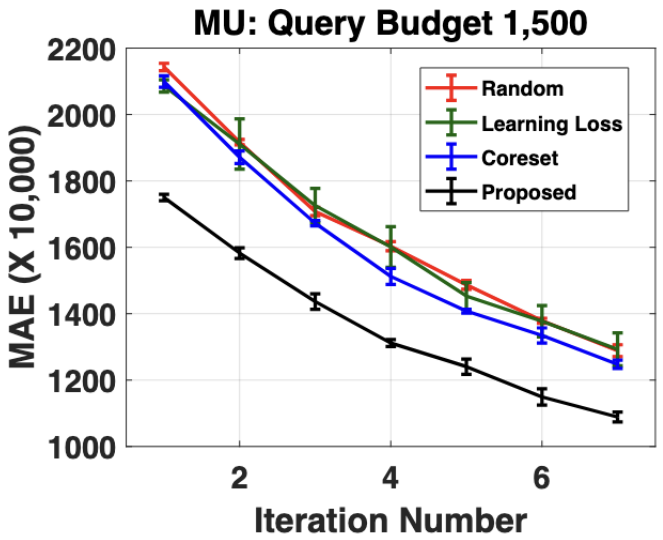


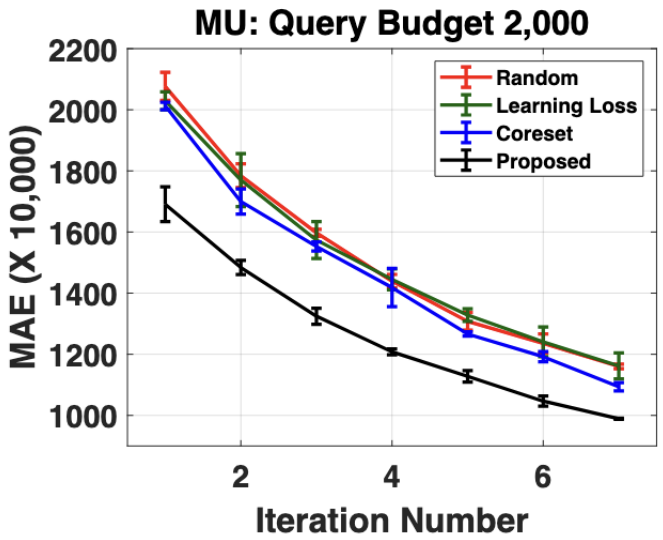


- Ablation study to observe the individual impact of uncertainty and diversity components on ***mu*** and ***lumo*** properties of ***QM9*** dataset

Results: Study on the effect of query size







❑ Study on the effect of different query size, k, in AL performance

Conclusion

- The work presents an Active Learning(AL) pipeline for informative data selection for 3D molecular graphs
- Novel diversity component based on the geometric representation of graph is proposed for AL
- Empirical study on medium-scaled QM9 and MD17 dataset demonstrates the effectiveness of our framework

Future Work

- Study the scalability of method on large scale molecular datasets, such as OC20

NEURAL INFORMATION PROCESSING SYSTEMS

FSU FLORIDA STATE UNIVERSITY

Stony Brook University

DARE TO BE

Thank You

Find our work at:

Git Repo

Ronast's Homepage

Wenhan's Homepage