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

Ronast Subedi[*]
Florida State University
rs22ce@fsu.edu

Lu Wei[*]
Stony Brook University
lu.wei.1@stonybrook.edu

Wenhan Gao[*]
Stony Brook University
wenhan.gao@stonybrook.edu

Shayok Chakraborty[†]
Florida State University
shayok@cs.fsu.edu

Yi Liu[†]
Stony Brook University
yi.liu.4@stonybrook.edu

Abstract

Molecular learning is pivotal in many real-world applications, such as drug discovery. Supervised learning requires heavy human annotation, which is particularly challenging for molecular data; *e.g.*, the commonly used density functional theory (DFT) is highly computationally expensive. Active learning (AL) automatically queries labels for the most informative samples, thereby remarkably alleviating the annotation hurdle. In this paper, we present a principled AL paradigm for molecular learning, where we treat molecules as 3D molecular graphs. Specifically, we propose a new diversity sampling method to eliminate mutual redundancy built on distributions of 3D geometries. We first propose a set of new 3D graph isometries for 3D graph isomorphism analysis. Our method is provably at least as expressive as the Geometric Weisfeiler-Lehman (GWL) test. The moments of the distributions of the associated geometries are then extracted for efficient diversity computing. To ensure our AL paradigm selects samples with maximal uncertainties, we carefully design a Bayesian geometric graph neural network to compute uncertainties specifically for 3D molecular graphs. We pose active sampling as a quadratic programming (QP) problem using the proposed components. Experimental results demonstrate the effectiveness of our AL paradigm, as well as the proposed diversity and uncertainty methods. The code is publicly available at `https://github.com/sronast/al_3dgraph`.

1 Introduction

Molecular representation learning is essential for various real-world applications, such as molecular design, drug discovery, material design, *etc.*. In recent studies, molecules have been formulated as 3D graphs, based on the evidence that 3D spatial information is crucial to determine the properties of molecules [Liu et al., 2019, Townshend et al., 2019, Axelrod and Gomez-Bombarelli, 2020]. Generally, in a 3D graph, atoms are represented as nodes, each associated with Cartesian coordinates in 3D space. A predefined cut-off distance can be used as a threshold to determine if there is an edge between two nodes in the 3D graph. With the advance of deep learning, 3D graph neural networks (GNNs) have been developed to learn from 3D molecular graph data [Thomas et al., 2018, Schütt et al., 2017, Satorras et al., 2021, Gasteiger et al., 2020c, Liu et al., 2021, 2022, Wang et al., 2022, Liao and Smidt, 2022, Zhou et al., 2022, Yan et al., 2022, Wang et al., 2023, Lin et al., 2023, Zhang et al., 2023]. These models are data-hungry and necessitate a large amount of annotated training data

[*] Equal contribution.

[†] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to attain good performance. However, annotation usually consumes excessive manpower, which is particularly challenging for molecules, *e.g.*, the commonly used density functional theory (DFT) for molecular energy computing [Hohenberg and Kohn, 1964] is very expensive, inducing a complexity of $O(n_e^3)$, where n_e is the number of electrons. As a concrete example, DFT can be hundreds of thousands of times slower than a reasonably good GNN for inference [Gilmer et al., 2017].

Active Learning (AL) algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data [Settles, 2009, Ren et al., 2021]. This tremendously reduces the human annotation effort, as only the few samples identified by the algorithm need to be labeled manually. Further, since the deep network gets trained on the representative samples from the underlying data population, it typically depicts better generalization capability than a passive learner, where the training data are selected at random. Deep AL has been used with remarkable success in various applications, such as computer vision [Yoo and Kweon, 2019, Sinha et al., 2019], natural language processing [Zhang et al., 2022], medical diagnosis [Blanch et al., 2017], chemistry [Smith et al., 2018], and anomaly detection [Pimentel et al., 2020] among others. There are a few AL applications for 3D GNNs [Smith et al., 2021, van der Oord et al., 2023]; however, these works do not specifically account for 3D geometric information. The 3D geometry of molecules is crucial for determining molecular properties, but it introduces unique challenges in designing effective AL schemes. Currently, a principled AL algorithm for 3D molecular graphs is still lacking.

In this paper, we propose a principled AL paradigm for 3D molecular graphs. We formulate a criterion based on uncertainty and diversity, which ensures that the queried molecules are those where the graph learning model has maximal uncertainty about the labels, and that are also mutually diverse to avoid duplicate sample queries. In particular, diversity computing for 3D graphs is challenging and the *difficulties are twofold*. Firstly, the AL pipeline requires computing the difference between any two 3D molecular graphs, which could have different planar (2D) molecules (entangling different atom numbers, *etc.*), in most cases. Secondly, the 3D shape (geometry) of a 3D graph should be captured completely for expressive geometric representations and accurate diversity computation.

To tackle these challenges, we propose a novel diversity sampling method for 3D molecular graphs based on distributions of important 3D geometries. We propose a set of new 3D graph isometries for geometric modeling, which produces geometric representations that are at least as powerful as the Geometric Weisfeiler-Leman (GWL) test [Joshi et al., 2023] in distinguishing 3D graph geometries. This indicates our approach sets an upper bound on the expressive power of any existing 3D GNN models. Hence, the geometries derived from our geometric modeling method (*e.g.*, reference distances, triangles) can be used for accurate diversity computing. To compare any two 3D molecules (with different planar graphs), the moments of the distributions of the derived geometries are extracted for final diversity computing of 3D graphs. In addition, to ensure our AL paradigm selects samples with maximal uncertainties, we carefully design a Bayesian geometric GNN specifically for 3D graph uncertainty computing. Our method is shown to be effective and efficient based on a set of ground approximations. With our novel components, we pose the sample selection as a quadratic programming (QP) problem and implement a fast QP solver to identify exemplar molecules to be annotated. Our method is easy to implement and can be applied in conjunction with any 3D GNN architecture.

Overall, our proposed AL paradigm incorporates both diversity and uncertainty for 3D molecular graphs. The diversity component, driven by proposed geometric isometries, captures diverse chemical properties from geometries. The uncertainty component leverages chemical contexts, such as atom types, as node features, enhancing the model's ability to identify and learn from uncertain chemical interactions. By considering both, our method represents a powerful AL paradigm for 3D molecular graphs. We conduct extensive experiments, and the results demonstrate the effectiveness of the proposed diversity and uncertainty methods as well as the overall AL paradigm.

Our contributions are summarized below. (i) We propose a principled AL paradigm to alleviate the annotation hurdle of 3D molecular graphs. We employ diversity and uncertainty measures to select the most informative subset for AL. (ii) We introduce a novel diversity component for 3D molecular graphs. Investigating geometric graph isomorphism, we introduce a *model-agnostic* geometric modeling method, which is provably at least as expressive as the GWL test. Our method can significantly enhance the accuracy of diversity computing for 3D molecular graphs. (iii) Our proposed graph isometries set the theoretical upper bound to the expressive power of all existing 3D GNNs, and thus can serve as the new gold standard to test the expressiveness of various 3D

GNNs. (iv) Rooted in Bayesian inference, we develop an effective and efficient pipeline to compute uncertainties for 3D molecular graphs. (v) Our framework significantly outperforms mainstream AL baselines, achieving remarkable efficiency owing to the cheap complexity of $O(N^2)$ as well as the implementation of a fast QP solver.

2 Methods

2.1 Diversity Computing for 3D Molecular Graphs

In molecular AL tasks, diversity sampling is important for eliminating redundancy, thereby wisely leveraging the annotation budget. The model's capability of capturing the 3D shape diversity among molecules is crucial for informed sampling. A particular challenge is that a diversity measure for two 3D molecules with different planar graphs is indispensable. Methods for diversity measures for 3D molecules with the same planar graph have been developed [Kumar and Zhang, 2018, Kearnes et al., 2016, Gfeller et al., 2013], but a diversity method for two 3D molecules with different planar graphs (entailing different atoms, *etc*) is demanding. Inspired by the USR method [Ballester and Richards, 2007], we propose a novel solution to achieve the goal from the distribution perspective. Generally, we develop a set of new *isometries* for expressive representations of 3D molecular graphs, after which the distributions of geometries associated with the isometries are obtained for diversity computing.

2.1.1 Isometries of 3D Molecular Graphs

As the first step, we introduce a set of new *isometries* as a basis, aiming at expressive representations of 3D graphs. As we focus on 3D geometry of molecules in this section, for simplicity, we use 3D point clouds to illustrate our ideas. Let $A = \{a_1, a_2, ..., a_n\}$ and $B = \{f(a_1), f(a_2), ..., f(a_n)\}$ be two sets representing 3D point clouds. Here, each a_i in A is associated with a positional vector $\boldsymbol{a_i} = (x_{a_i}, y_{a_i}, z_{a_i})$ in 3D space. f denotes a bijective mapping between A and B. Then, similarly, each point $f(a_i)$ in B is associated with a positional vector $\boldsymbol{f(a_i)} = (x_{f(a_i)}, y_{f(a_i)}, z_{f(a_i)})$.

Two 3D point clouds, A and B, are said to be $E(3)$-isomorphic, if there exists $\gamma \in E(3)$ such that $A = \gamma B$. We further choose or compute a consistent reference point (*e.g.*, centroid) for each point cloud, denoted as r_1 and r_2, respectively. Without loss of generality, we use a_{far} to denote the farthest point from the reference point in point cloud A. Below, we will define three levels of isometries, each of which fulfills an isometric mapping between A and B. To satisfy *any* isometry, there needs to exist a bijective function $f : A \to B$, such that $\boldsymbol{h_{f(a)}} = \boldsymbol{h_a}$ for any node $a \in A$. Here, $\boldsymbol{h_{f(a)}}$ and $\boldsymbol{h_a}$ denote the node feature vectors for $f(a)$ and a, respectively.

Reference Distance Isometry: If there exists a collection of global group elements $\gamma_i \in E(3)$, such that $(r_2, f(a_i)) = (\gamma_i r_1, \gamma_i a_i)$ for each point $a_i \in A$, A is reference distance isometric to B.

Reference distance isometry involves the Euclidean distance between any atom in the molecule and the predefined reference point.

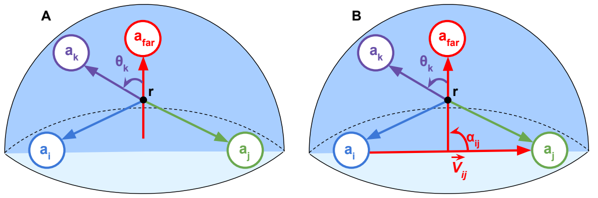

Figure 1: The illustrations of encoding the molecular triangular and cross-angular isometries

Triangular Isometry: If there exists a collection of global group elements $\gamma_i \in E(3)$, such that $(r_2, f(a_{\text{far}}), f(a_i)) = (\gamma_i r_1, \gamma_i a_{\text{far}}, \gamma_i a_i)$ for each point $a_i \in A$, A is triangular isometric to B.

With reference point r, we define the reference vector $\boldsymbol{v}_0$ as r pointing to the farthest point a_{far} in a 3D molecule. Based on reference distance isometry, triangular isometry further involves the angle between $\boldsymbol{v}_0$ and other vectors pointing from r to any other point in the molecule, computed as $\theta_k = \cos^{-1}\left(\frac{\boldsymbol{v}_0 \cdot \boldsymbol{v}_k}{\|\boldsymbol{v}_0\| \|\boldsymbol{v}_k\|}\right)$, where $\boldsymbol{v}_k$ denotes vectors originating from r and directed towards k^{th} atoms in the molecule. The process is illustrated in part A of Fig. 1. For a molecule with N nodes, we compute $N - 1$ angles. Essentially, such angles provide insights into the spatial arrangement of atoms with respect to the pre-assigned reference vector.

Cross-angle Isometry: If there exists a collection of global group elements $\gamma_{ij} \in E(3)$, such that $(r_2, f(a_j), f(a_i)) = (\gamma_{ij} r_1, \gamma_{ij} a_j, \gamma_{ij} a_i)$, $\forall a_i, a_j \in A$ $(i \neq j)$, A is cross-angle isometric to B.

Beyond the angles in triangular isometry as well as based on reference distance isometry, cross-angular isometry further considers angles formed by any two atoms in the molecule with respect to the reference vector as above. Specifically, for every pair of atoms i and j, a vector $\boldsymbol{v}_{ij}$ is formed from i to j. With the reference vector $\boldsymbol{v}_0$, the cross angle is computed as $\alpha_{ij} = \cos^{-1}\left(\frac{\boldsymbol{v}_0 \cdot \boldsymbol{v}_{ij}}{\|\boldsymbol{v}_0\| \|\boldsymbol{v}_{ij}\|}\right)$. This approach, as depicted in part B of Fig. 1, essentially reflects cross-angle information globally. For a molecule with N nodes, we compute $N(N-1)/2$ cross angles with the complexity of $O(N^2)$.

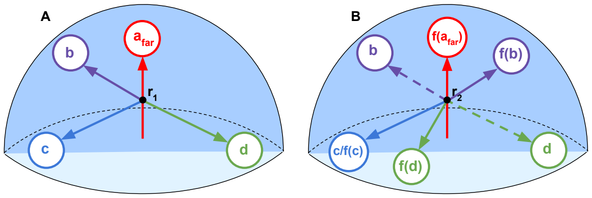

Figure 2: A and B are triangular isometric but not cross-angular isometric. The angles $\angle br_1 a_{far}$, $\angle cr_1 a_{far}$, and $\angle dr_1 a_{far}$ in structure A are equal to the angles $\angle f(b) r_2 f(a_{far})$, $\angle f(c) r_2 f(a_{far})$, and $\angle f(d) r_2 f(a_{far})$ in structure B, respectively. However, the cross angle $\angle dr_1 c$ is not equal to the cross angle $\angle f(d) r_2 f(c)$.

Next, we propose **Theorem 1** to indicate the relationship between these three isometries as below.

Theorem 1. *If A and B are triangular isometric, then A and B are reference distance isometric; If A and B are cross-angle isometric, then A and B are triangular isometric.*

The proof of **Theorem** 1 can be found in Appendix A.1. Generally, we define three levels of isometries for graph isomorphism. *Reference distance isometry* ensures that the Euclidean distance between each point and a predefined reference point is consistent in two different point clouds. *Triangular isometry* further manifests the spatial arrangement of atoms referring to the pre-assigned pivot. Built on *triangular isometry*, *cross-angular isometry* then reflects the pair-wise global information. An illustrative example for *triangular isometry* and *cross-angular isometry* is also given in Fig. 2. Clearly, cross-angular isometry represents the strictest isometry among the three. In the following Sec. 2.1.2, we show that a designed geometric representation based on *cross-angular isometry* can exhibit great expressive power.

2.1.2 Expressive Power of Our Geometric Representations

In this section, we aim to formally elucidate the expressive power of a geometric representation (GR) based on our developed isometries in Sec. 2.1.1. Naturally, we formulate GR_{ours} as a set containing all reference distances, triangles, and cross angles in a 3D graph.

We explore the Geometric Weisfeiler-Leman (GWL) test [Joshi et al., 2023], and then leverage GWL to illustrate the expressiveness power of our model. GWL test is an extension of the classic WL Test, enhancing its capabilities by incorporating both the topological structure of the graph and the geometric attributes of its vertices. Such an integration allows the GWL test especially apt for evaluating all 3D graph representation methods. Similar to the regular WL test, GWL test imposes an upper bound to the expressive power of 3D GNNs, *i.e.*, if GWL test fails to distinguish two 3D graphs, then all existing 3D GNNs would also fail. See details of the GWL test in Appendix A.2.

Proposition 1. *GR_{ours} is at least as expressive as the GWL test. In other words, GR_{ours} suffices to distinguish any non-isomorphic molecular structures that are distinguishable by any 3D GNN.*

The proof of **Proposition** 1 can be found in Appendix A.1. In conclusion, the molecular geometric representation GR_{ours} developed in this work has the greater expressive power than the GWL test, which indicates our diversity sampling method is accurate enough to capture the 3D shape diversity among different molecules. Notably, as mentioned before, GWL test sets the upper bound to the expresiveness of any existing 3D GNNs. ***Apparently, our geometric representation GR_{ours} is provably at least as powerful as any existing 3D GNN for learning geometric features.*** Essentially, the three isometries associated with GR_{ours} define expressiveness at different levels. For example, as only considering distance information, a well pretrained SchNet is upper bounded by reference distance isometry (but not triangular isometry or cross-angular isometry); as a more powerful model than SchNet, a well pretrained DimeNet is upper bounded by triangular isometry (but not cross-angular isometry). Additionally, learning accurate geometric representations requires a perfectly pretrained 3D GNN model, which is hard to guarantee in practice. ***Our isomorphy study provides a deterministic and model-agnostic diversity component for 3D graphs, avoiding the need of a 'perfectly' pretrained 3D GNN model, as well as achieving a theoretically guaranteed upper bound of the expressiveness of all existing 3D GNN models.***

2.1.3 Final Distributional Representations

Based on the isomorphy study in Sec. 2.1.1, we obtain our geometric representation method GR_{ours} and prove GR_{ours} possesses greater expressive power than any existing 3D GNN models in Sec.

2.1.2. In this section, we aim to extract the distributions of the ***entangled three geometries in GR_{ours}, including reference distances, triangles, and cross angles***, for diversity computing. Fortunately, we have the theorem [Hall, 1983] implying that the sequence of translated moments can be used to determine the original distribution. Following the USR work [Ballester and Richards, 2007], for each of the three aforementioned geometries, we also use four reference points to reflect the "translated" geometries; those are, the centroid (denoted as ctd) computed by the mean position of all the atoms in the 3D molecule, the point closest to the centroid (denoted as cst), the point farthest from the centroid (denoted as fct), and the point farthest from fct (denoted as ftf). For each reference point, we use a set of moments, including mean, variance, skewness, and kurtosis, which describe a distribution from different angles, *e.g.*, skewness indicates the asymmetry and kurtosis describes the tailedness of a distribution. Detailed formulae for these moments can be found in the Appendix A.3. Notably, we compute these translated moments for all three entangled geometries as above. Eventually, we obtain summarized representations of distributions over geometries of 3D graphs, capturing essential characteristics of a molecule's shape.

We use cross angles as an example to describe the final distributional vector. For a molecule with N atoms, as shown in Fig. 1, we can obtain a set of cross angles $[\alpha_{ij}^{\text{ref}}]_{i\neq j,0<i,j<N}^{N(N-1)/2}$ for a reference point (*e.g.*, ctd). After applying statistical moments as an approximation, we can obtain a 4-dimensional vector $\overrightarrow{M_{\text{ref}}^{\text{ca}}} = [m_{\text{ref}}^{\text{ca}}, v_{\text{ref}}^{\text{ca}}, s_{\text{ref}}^{\text{ca}}, k_{\text{ref}}^{\text{ca}}]$, where the four elements denote the mean, variance, skewness, and kurtosis for this reference point, respectively. We perform a similar process for all four reference points mentioned above. By doing this, we can obtain four 4-dimensional vectors including $\overrightarrow{M_{\text{ctd}}^{\text{ca}}}$, $\overrightarrow{M_{\text{cst}}^{\text{ca}}}$, $\overrightarrow{M_{\text{fct}}^{\text{ca}}}$, and $\overrightarrow{M_{\text{ftf}}^{\text{ca}}}$, which are then concatenated together, resulting in the final 16-dimensional vector to represent the distribution of cross angles. We repeat the similar process for reference distances and triangles, and then all three corresponding 16-dimensional vectors are further concatenated as a 48-dimensional distributional vector to represent the geometric information of the input molecule. The 48-dimensional distributional vectors are then used to compute the diversity matrix. For any two molecules n_1 and n_2 in the dataset with N molecules, we perform the inner product on their distributional vectors to achieve the similarity, and then use $1-$ similarity to obtain the final value $D_{n_1 n_2}$ as the diversity measure between them. Finally, a matrix $D \in \Re^{N\times N}$ is obtained, which contains the diversity between every pair of molecules.

Comparing Our Method to Traditional Structural Descriptors. Our method generates a 48-dimensional vector that encodes the geometric structure of a molecule. This representation is both equivariant to roto-translations and invariant to atomic permutations as the statistical quantities remain unchanged under such transformations. In contrast, Smooth Overlap of Atomic Positions (SOAP) [Bartók et al., 2013, De et al., 2016, Jäger et al., 2018] generates atom-wise vectors that capture local atomic environments by employing spherical harmonics and radial basis functions. While SOAP is also equivariant to roto-translations, it is not invariant to atomic permutations. On the other hand, Atomic Cluster Expansion (ACE) [Drautz, 2019] uses a systematic expansion to describe interactions of varying orders (*e.g.*, two-body, three-body interactions). However, ACE is less of a traditional descriptor compared to our method and SOAP; it is designed to provide a complete and systematic representation of atomic interactions by focusing on higher-order expansions (e.g., two-body, three-body interactions). This makes ACE more comprehensive in capturing the physical interactions within a system, but less suited for producing a fixed-dimensional, flexible descriptor. Unlike our method and SOAP, which generate more compact and adaptable descriptors, ACE emphasizes thorough expansions, making it less ideal for tasks requiring flexible, low-dimensional representations that can adapt easily to the active learning scheme. An empirical comparison between our method and the approach that uses SOAP will be provided, highlighting the effectiveness of our method in capturing molecular geometries.

2.2 Uncertainty Computing for 3D Molecular Graphs

In Sec. 2.1, we develop an effective method for diversity computing among different 3D molecular graphs. In addition to selecting diverse molecules, it is important to select molecules where the model has maximal prediction uncertainty about the labels, so as to append maximal information to the model. Uncertainty qualification is well-studied in planar graph analysis [Hirschfeld et al., 2020], but an effective paradigm for 3D molecular graphs is currently lacking. Additionally, existing methods, such as Bayesian neural networks (BNNs) [Lampinen and Vehtari, 2001, Titterington, 2004, Goan and Fookes, 2020] and deep model ensemble methods [Lakshminarayanan et al., 2017, Huang et al., 2017], are excessively computationally expensive, limiting their capacity in 3D graph analyses. In a concurrent work [Thaler et al., 2024] on active learning for partial charge prediction of metal-organic frameworks, a dropout Monte Carlo scheme has been proposed to lessen these issues.

In this work, we develop an effective and efficient method, known as Bayesian geometric graph neural network (BGGNN), that takes a 3D graph as input and produces the demanding properties as well as uncertainty values, *e.g.*, mean and variance. Formally, a 3D graph is represented as $\mathbf{G} = (V, E, P)$, where V denotes the set of vertices (atoms), E denotes the set of edges (bonds), and P denotes the set of Cartesian coordinates for all atoms. A 3D molecular graph is associated with a set of properties, denoted as $\mathbf{O}$. Recently, researchers have developed 3D GNNs, such as SchNet [Schütt et al., 2017], DimeNet [Gasteiger et al., 2020b], SphereNet [Liu et al., 2022], and GemNet [Gasteiger et al., 2021], for 3D graph representation learning. The likelihood of a 3D GNN can be represented as $p_{\text{3DGNN}}(\mathbf{O} \mid \mathbf{G}, \mathbf{w})$, where 3DGNN indicates any existing 3D GNN and $\mathbf{w}$ denotes the set of parameters of the used 3D GNN. We also use $p_{\text{3DGNN}}(\mathbf{w})$ to represent the prior distribution for the parameters. Assume we collect a new input and output pair, denoted as $\mathbf{g}^*$ and $\mathbf{o}^*$. Then based on the conventional Bayesian theorem, Bayesian inference for this new output $\mathbf{o}^*$ is given by

$$p_{\text{3DGNN}}\left(\mathbf{o}^* \mid \mathbf{g}^*, \mathbf{G}, \mathbf{O}\right) = \int_{\mathbb{R}^n} p_{\text{3DGNN}}\left(\mathbf{o}^* \mid \mathbf{g}^*, \mathbf{w}\right) p_{\text{3DGNN}}(\mathbf{w} \mid \mathbf{G}, \mathbf{O}) d\mathbf{w}, \tag{1}$$

where $\mathbb{R}^n$ is the whole space of n parameters in 3DGNN. It's infeasible to perform the above integration on $\mathbb{R}^n$ due to prohibitive computational cost. To tackle this, the variational inference method is introduced to approximate $p_{\text{3DGNN}}(\mathbf{O} \mid \mathbf{G}, \mathbf{w})$ with the parameterized $q_\theta(\mathbf{w})$ through minimizing the Kullback-Leibler (KL) divergence between these two distributions. After applying Bayesian theorem once more, the minimization objective becomes

$$\mathcal{L}_{\text{VI}}(\theta) = -\int_{\mathbb{R}^n} q_\theta(\mathbf{w}) \log p_{\text{3DGNN}}(\mathbf{O} \mid \mathbf{G}, \mathbf{w}) d\mathbf{w} + \text{KL}\left(q_\theta(\mathbf{w}) \| p_{\text{3DGNN}}(\mathbf{w})\right), \tag{2}$$

To completely avoid the integration over the whole parameter space, the MC-dropout method [Gal and Ghahramani, 2016, Srivastava et al., 2014] is further used in our BGGNN. Specifically, it employes the Monte-Carlo estimator [Gal et al., 2016, Gal and Ghahramani, 2016] to approximate the integration by performing summation over the sampled models. In practice, researchers implement an MC-dropout network by using dropout as the network's regularization[Gal and Ghahramani, 2016]. Following this, we propose to insert dropout layers after the linear layers in our used 3DGNN as an effective yet efficient estimation of Bayesian inference.

Now as we have obtained the variational predictive distribution of a new output with $q_\theta(\mathbf{w})$, we can easily compute the predictive mean and variance of this distribution. For the molecular property prediction tasks, after we sample N outputs from the same input, the heteroscedastic predictive uncertainty is then given by

$$\widehat{\sigma^2}\left(\mathbf{o}^* \mid \mathbf{g}^*\right) = \frac{1}{N} \sum_{n=1}^{N} \left(\hat{\mathbf{o}}_n^*\right)^2 - \left(\frac{1}{N} \sum_{n=1}^{N} \hat{\mathbf{o}}_n^*\right)^2 + \frac{1}{N} \sum_{n=1}^{N} \widehat{\sigma}_n^2, \tag{3}$$

where $\hat{\mathbf{o}}_n^*$ is the n^{th} sampled output and $\widehat{\sigma}_n^2$ is the variance that is the same among all the data samples. By doing this, we can obtain an uncertainty value (variance) for each molecule. Additionally, built on a 3D GNN, our BGGNN can faithfully produce a set of molecular properties $\mathbf{O}$.

Practically, any of the existing 3D GNN can be used as the backbone network for property prediction and uncertainty computing. In this study, we employ SphereNet [Liu et al., 2022] as our 3DGNN, owing to its great power in incorporating 3D geometric information. We apply dropout layers onto the linear layers of SphereNet for Bayesian inference in our BGGNN. To allow more accurate AL selections, we particularly employ the concrete dropout with a learnable dropout rate [Gal et al., 2017] in our BGGNN. Overall, our method is shown to be an effective and efficient paradigm for 3D graph uncertainty computing, as further empirically demonstrated in Sec. 4.

2.3 Active Sampling

$$\begin{aligned} \max_{z} \quad & z^\top r + \lambda z^\top D z \\ s.t. \quad & \sum_{i=1}^{N} z_i = k \\ & z_i \in \{0, 1\}, \forall i, \end{aligned} \tag{4}$$

A schematic diagram of our active sampling framework is depicted in Fig. 6 and described in A.4 in Appendix. Specifically, in Sec. 2.1, we obtain the matrix $D \in \Re^{N \times N}$ containing the mutual diversity between every pair of unlabeled molecules, where N is the number of unlabeled molecules. In Sec. 2.2, we employ our designed BGGNN to achieve the vector $r \in \Re^{N \times 1}$ quantifying the prediction uncertainty score of each unlabeled molecule. In the AL setting, our objective is to select a batch of k unlabeled molecules (k is a pre-defined query batch size) with high prediction uncertainty and high mutual diversity among them. Let $z \in \{0, 1\}^{N \times 1}$ be a binary vector with N entries which denotes whether the unlabeled molecule x_i will be included in the batch ($z_i = 1$) or not ($z_i = 0$). The molecule selection can thus be posed as the following optimization problem as in Eq. (4), where λ is a weight parameter governing the relative importance of the two terms. This is a standard quadratic programming (QP) problem; we relax the integer constraints into continuous constraints and solve the problem using an off-the-shelf QP solver. In this work, we employ the widely used Operator Splitting Quadratic Program (OSQP) [Stellato et al., 2020] to solve the QP problem in Eq. (4). We then apply a greedy approach to project the continuous solution back to the binary space, where the k highest entries of the continuous solution vector are set to 1 and the remaining to 0. Such an approach is commonly used to convert continuous solutions obtained from a QP solver to binary solutions in AL [Chattopadhyay et al., 2013, Wang and Ye, 2013]. To accelerate the optimization, we implement a solution to execute the problem in the GPU (instead of the CPU) using the parallel implementation of the alternating direction method of multipliers, as detailed in Schubiger et al. [2020]. Notably, the predictions in the main tasks (*e.g.*, molecular properties) are produced by our BGGNN built on SphereNet as in Sec. 2.2.

3 Related Work

3.1 Active Learning

AL is a well-researched problem in the machine learning community [Settles, 2009]. There exist two commonly used strategies for AL sampling. Uncertainty based sampling queries unlabeled samples with the highest prediction uncertainties for annotation. Diversity/representativeness based sampling aims to select the subset that can well represent the entire data distribution. A full review of the two AL sampling methods is provided in Appendix A.5.

3.2 Molecular Shape Similarity

Molecular shape similarity plays a pivotal role in drug discovery and virtual screening of compounds [Kumar and Zhang, 2018, Murgueitio et al., 2012, Shang et al., 2017]. Methods predominantly fall into several categories [Kumar and Zhang, 2018], including descriptor-based methods [Schreyer and Blundell, 2012, Cannon et al., 2008, Li et al., 2016, Armstrong et al., 2009, Zhou et al., 2010], atom-centered Gaussian-based methods [Haque and Pande, 2010, de Lima and Nascimento, 2013, Yan et al., 2013], surface-based methods [Hofbauer et al., 2004, Mavridis et al., 2007, Cai et al., 2012, Karaboga et al., 2013, Venkatraman et al., 2009, Sael et al., 2008], *etc*. Descriptor-based methods are notably represented by the Ultrafast Shape Recognition (USR) algorithm [Ballester and Richards, 2007], which uses statistical moments of the distance distribution to characterize molecular shapes. Gaussian overlay-based methods, with ROCS [Rush et al., 2005, Hawkins et al., 2007] being the most commonly used one, evaluate the maximum volume overlap between two molecules. Surface-based methods typically employ shape signatures [Zauhar et al., 2013] or shape histograms to delineate molecular surfaces for shape similarity assessment. Despite the progress, a principled and theoretically ground similarity method for 3D molecular graphs is currently lacking.

4 Experiments

4.1 Experimental Setup

Implementation Details: We use two mainstream 3D GNNs SphereNet [Liu et al., 2022] and DimeNet^{++} [Gasteiger et al., 2020a] as the backbone models of our BGGNN. We directly use the optimal network configurations from the original papers for both backbone models. We train the network for 200 epochs, unless otherwise specified. We use the *Adam Optimizer* with an initial learning rate 5×10^{-4} and scale it by a factor of 0.5 every 15 epochs.

Data and Active Learning Setup: We first perform experiments on the QM9 benchmark dataset. Since SphereNet is more stable and incorporates more 3D information, we conduct experiments on ***mu, alpha, homo, and lumo*** for SphereNet, and ***mu and lumo*** for DimeNet^{++}. These properties have continuous values, making the prediction problem a regression task. We randomly divide the training set of $110,000$ molecules into three splits of size $25,000$ each. From each split, we randomly select $5,000$ molecules as the initial labeled set and the remaining $20,000$ molecules as the unlabeled set. In each AL iteration, we query $1,500$ molecules from the unlabeled set, which are labeled and appended to the labeled set. The model's performance is evaluated on a held-out validation set containing $10,000$ molecules. We save the best-performing model on the validation set and report its performance on the test set containing $10,831$ molecules. The process is repeated for 7 AL iterations, which is taken as the stopping criterion. The final results are averaged over the three splits to rule out the effects of randomness. λ in Eq. 4 is taken as 1. The Mean Absolute Error (MAE) is used as the evaluation metric. In addition, to study the generalizability of our framework to more geometric data, we also conduct experiments to predict atomic forces for ***Aspirin*** in MD17 using our framework.

Comparison Baselines: We use four classic AL methods as baselines: *Random Sampling*, *Coreset* [Sener and Savarese, 2018], *Learning Loss* [Yoo and Kweon, 2019], and *Evidential Uncertainty* [Beluch et al., 2018, Amini et al., 2020]. *Random Sampling* is the default comparison baseline in AL research. *Coreset* and *Learning Loss* are two extensively used deep active learning algorithms for regression applications. *Evidential Uncertainty* is also a commonly used technique to quantify uncertainty for molecular property prediction and was hence included as a comparison baseline. Note some existing studies [Kulichenko et al., 2023, Gusev et al., 2023, Craig and García-Melchor, 2021] have applied AL to molecule research and chemistry. However, these works focus on 2D molecules without considering 3D geometry, which is the focus of our work. Additionally, the techniques used in existing studies can arguably fall into the aforementioned AL categories. Hence, we think comparing with these classic AL methods is sufficient to demonstrate the superiority of our pipeline.

4.2 Active Learning Performance

The active learning performance with SphereNet is depicted in Fig. 3. In each graph, the x-axis denotes the iteration number and the y-axis denotes the MAE on the test set. Our analysis revealed that *Evidential Uncertainty* depicted the worst performance and furnished significantly high error values for all four properties, which obscured the difference in performance among the other methods in the plots. For better interpretation and understanding, we exclude the *Evidential Uncertainty* method from the plots here and present the results with this baseline in Sec. A.6 of the Appendix. The other baseline methods depict more or less similar performance, with *Coreset* marginally outperforming the other baselines. Our method comprehensively outperforms all the baselines. At any given AL iteration, it consistently attains a lower MAE compared to all the baselines.

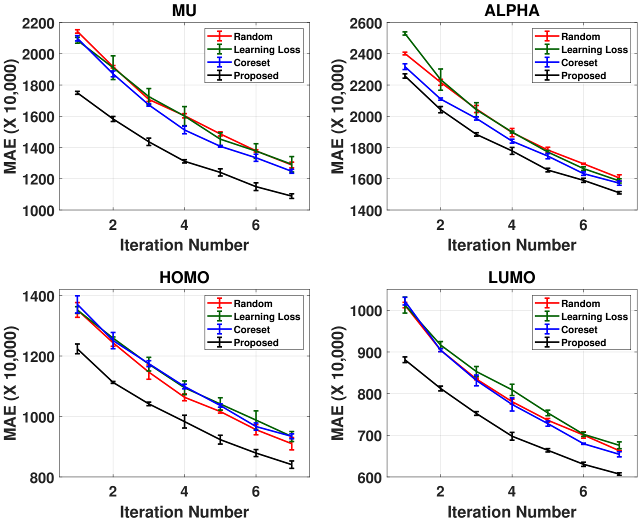

Figure 3: Active learning performance results with SphereNet. The graphs show the mean (averaged over 3 runs) and the errorbars for all the methods. We plot the MAE values from the first iteration onwards, to focus on the comparative performance of the methods after they start selecting samples using AL. Best viewed in color.

Table 1: The table shows the p-values obtained using paired t-test between the results our method against each of the baselines for all the properties studied. *Here, L. Loss refers to Learning Loss.*

Properties	Baselines			
	Random	L. Loss	Coreset	Evidential
mu	7.54×10^{-6}	5.09×10^{-5}	1.51×10^{-4}	2.19×10^{-7}
alpha	1.06×10^{-5}	8.14×10^{-4}	4.27×10^{-5}	2.72×10^{-4}
homo	2.26×10^{-5}	8.36×10^{-7}	4.23×10^{-6}	1.71×10^{-8}
lumo	4.48×10^{-5}	1.25×10^{-5}	3.12×10^{-4}	2.39×10^{-6}

We also conducted statistical tests of significance using paired t-test to assess whether the improvement in performance achieved by our method is statistically significant. For this purpose, we compared the average MAE achieved by our method against each of the baselines individually. The results are reported in Table 1; each entry in the table denotes the p-value of the paired t-test between our method and the corresponding baseline (denoted in the columns) for the property studied (denoted in the rows). From the table, we note that the improvement in performance achieved by our method is statistically significant ($p < 0.05$) compared to all the baselines, consistently for all the four properties studied. These results unanimously corroborate the promise and potential of the proposed active sampling method to tremendously reduce the annotation cost in inducing a robust 3D graph neural network for molecular property prediction.

In addition, to study the robustness of our framework to the underlying network architecture and generalizability to the underlying geometric graph data, we have the following results: ***1. To study the***

robustness of our framework to the underlying network architecture, results on the ***mu and lumo*** properties of the QM9 dataset using DimeNet^{++} [Gasteiger et al., 2020a] as the backbone model are presented in Section A.7 of the Appendix due to space constraints. The results depict a similar pattern as Figure 3, with the proposed method consistently outperforming all the baselines for both the properties. A paired t-test, presented in Table 3 revealed that the performance improvement achieved by our framework is statistically significant. ***2. To study the generalizability of our framework to the underlying geometric graph data***, results on predicting atomic forces for ***Aspirin*** molecules in the MD17 benchmark dataset [Chmiela et al., 2017b] using our framework are depicted in Section A.8 of the Appendix and further corroborate the potential of our framework.

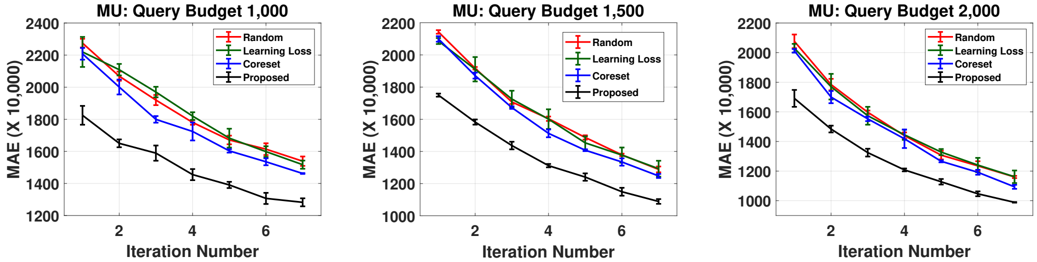

Figure 4: Study of query budget on the active learning performance. The graphs show the mean (averaged over 3 runs) and the errorbars for all the methods. The results with budget 1500 are the same as the those presented in Figure 3 and are included here for comparison. Best viewed in color.

4.3 Study of Query Budget

The goal of this experiment is to study the effect of query budget (batch size) on the AL performance. The results on the ***mu*** property with SphereNet for query budgets $1,000$, $1,500$ and $2,000$ are depicted in Fig. 4. Since *Evidential Uncertainty* depicted much worse performance than all the methods, it was excluded from this comparison. Our framework once again outperforms all the baselines consistently for all the query budgets. As before, we conducted a paired t-test and the results are presented in Appendix A.9. From the p-values, we conclude that the error values furnished by our method are statistically significantly better ($p < 0.05$) than all the baselines, consistently for all the query budgets. These results are particularly significant from a practical standpoint as the available query budget in a real-world application is dependent on time, resources, and other constraints.

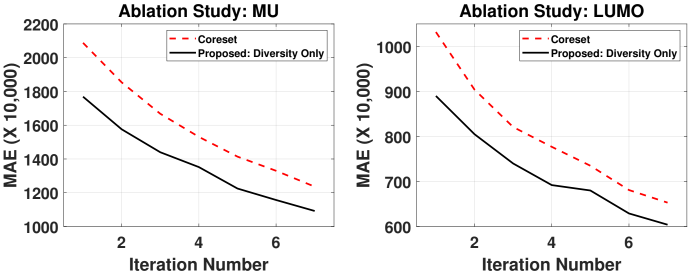

Figure 5: Ablation study results on the ***mu*** and ***lumo*** properties with SphereNet. Best viewed in color.

4.4 Ablation Studies

We conduct ablation studies to examine the power of our diversity computing method, as it is our primary contribution in this research. We perform experiments on the ***mu*** and ***lumo*** properties with SphereNet from two aspects. Firstly, we compare our framework with only the diversity term in Eq. 4 against *Coreset*, the state-of-the-art diversity-based AL technique. The results are reported in Fig. 5, from which we note that the diversity component of our framework consistently furnishes much lower MAE values than *Coreset* over all the AL iterations, for both properties. Secondly, we also conducted experiments where we compared the performance of our overall framework (using both uncertainty and diversity) against the baseline where only the uncertainty term in Eq. (4) was used for active sampling. The results revealed that removing the diversity term adversely affected the performance of our framework. A paired t-test revealed that the improvement in performance achieved by our diversity component is statistically significant ($p < 0.05$) for both these properties ($p = 0.0001$ for ***mu*** and $p = 0.04$ for ***lumo***). These results show the effectiveness of the proposed diversity metric for AL framework to train a 3D GNN for molecular property prediction. Additionally, we examine the individual impact of diversity and uncertainty components in Appendix A.10. We also compare our proposed diversity component with the SOAP-based diversity, and test our method against BatchBALD [Kirsch et al., 2019], a greedy clustering-based Bayesian uncertainty approach in Appendix A.10.

4.5 Computation Time Analysis

In this experiment, we analyze the computation time of all the methods studied in this paper. The average time taken to query a batch of unlabeled samples and update the SphereNet model (one active learning iteration) are shown in Table 2. For fair comparison, all the methods were run on the same NVIDIA RTX A4500 20GB GPU.

Table 2: Average (± std) time (minutes) taken by each method for sample selection and training the SphereNet model (one iteration of AL). *Here, L. Loss refers to Learning Loss.*

Random	L. Loss	Coreset	Evidential	Ours
53 ± 4.5	56 ± 2.1	127 ± 3.5	56 ± 2.3	64.9 ± 7.5

The computation time of our framework is much less than *Coreset*, which needs to solve a mixed integer programming (MIP) problem. The other three methods have similar computation time, as they don't involve iterative algorithms. Our method takes **only slightly more** time than them, owing to the implementation of a faster QP solver as mentioned in Sec. 2.3, as well as our vectorized implementation to enable the use of GPUs to perform diversity matrix computation. The performance studies in Sec. 4.2 show that our framework is much more accurate than these baselines, and the ablation studies in Sec. 4.4 indicate both the diversity and uncertainty components are necessary to form a QP problem. Given the large margin of performance improvement, we think the efficiency of our method is acceptable.

5 Conclusion, Limitations, Future Work, and Broader Impacts

We present a principled active learning framework with the goal of reducing the annotation cost for learning from 3D molecules represented as 3D graphs. The sample selection is posed as a QP problem, which selects samples with high mutual diversity and high uncertainty. Novel diversity and uncertainty components are proposed for 3D graphs, with strong empirical results presented.

We present a model-agnostic diversity component for 3D graphs, and our method is provably at least as powerful as any existing 3D GNN for learning geometric information. Even though our method can set the upper bound of the accuracy of diversity sampling for 3D molecules, it remains unexplored if such an advantage can be incorporated into 3D GNN models for diversity sampling. For example, molecular similarity might be incorporated into 3D GNNs to achieve comparable AL performance. Moreover, our experimental studies focus on small molecules in this work.

As part of future work, we plan to apply our methods to problems where much more accurate but expensive annotation is required, such as computing molecular systems' ground states using the Schrödinger equation. DFT calculations are widely used but still involve approximations, as Schrödinger equation is prohibitively expensive and its use is limited in very small molecules. Our AL pipeline is anticipated to unleash greater potential in such extreme-scale applications. Additionally, given AL needs several interactions with each requiring the model is well-trained, we test our methods on the commonly used but medium-scale QM9 and MD17 datasets in this work. Even though we think the empirical studies are sufficient to support our theory, we still plan to test the scalability of our methods on large-scale molecule datasets, such as OC20 [Chanussot et al., 2021], in the future.

This work facilitates a new avenue in graph analysis by effective and efficient representation of 3D geometric information, thereby dramatically advancing graph learning and mining. Our methods can reduce the annotation cost for molecular data and also have the potential in a broad set of scientific data types, such as materials and proteins, facilitating various disciplines including basic biology, material science, and quantum chemistry. This work is anticipated to have strong impacts on drug discovery and material design by enabling low-cost representation learning. Any positive and negative societal impact associated with those applications and domains can be applied to our methods.

6 Acknowledgment

This work of Y. Liu has used the computational equipment supported by the US Army Research Office under the award W911NF-20-10159. Y. Liu and W. Gao would like to thank Xiaolin Li at Stony Brook University for providing this computational equipment. The work of S. Chakraborty is partially supported by the National Science Foundation under the award IIS-2143424 (NSF CAREER Award).

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

M Stuart Armstrong, Garrett M Morris, Paul W Finn, Raman Sharma, and W Graham Richards. Molecular similarity including chirality. *Journal of Molecular Graphics and Modelling*, 28(4): 368–370, 2009.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations (ICLR)*, 2020.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom: Energy-annotated molecular conformations for property prediction and molecular generation. *arXiv preprint arXiv:2006.05531*, 2020.

Pedro J Ballester and W Graham Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of computational chemistry*, 28(10):1711–1723, 2007.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013. doi: 10.1103/PhysRevB.87.184115. URL `https://link.aps.org/doi/10.1103/PhysRevB.87.184115`.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

Marc Gorriz Blanch, Xavier Giro I Nieto, Axel Carlier, and Emmanuel Faure. Cost-effective active learning for melanoma segmentation. In *31st Conference on Machine Learning for Health: Workshop at NIPS 2017 (ML4H 2017)*, pages 1–5, 2017.

Felix Buchert, Nassir Navab, and Seong Tae Kim. Toward label-efficient neural network training: Diversity-based sampling in semi-supervised active learning. *IEEE Access*, 11:5193–5205, 2023.

Chaoqian Cai, Jiayu Gong, Xiaofeng Liu, Hualiang Jiang, Daqi Gao, and Honglin Li. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. *Journal of molecular modeling*, 18:1597–1610, 2012.

Edward O Cannon, Florian Nigsch, and John BO Mitchell. A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chemistry Central Journal*, 2:1–9, 2008.

Shayok Chakraborty, Vineeth Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1945–1958, 2015.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International Conference on Machine Learning (ICML)*, 2013.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017a.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017b. doi: 10.1126/sciadv.1603015. URL `https://www.science.org/doi/abs/10.1126/sciadv.1603015`.

Michael John Craig and Max García-Melchor. Applying active learning to the screening of molecular oxygen evolution catalysts. *Molecules*, 26(21):6362, 2021.

Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016. doi: 10.1039/C6CP00415F. URL `http://dx.doi.org/10.1039/C6CP00415F`.

Luis Antônio C Vaz de Lima and Alessandro S Nascimento. Molshacs: a free and open source tool for ligand similarity identification based on gaussian descriptors. *European journal of medicinal chemistry*, 59:296–303, 2013.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B*, 99:014104, Jan 2019. doi: 10.1103/PhysRevB.99.014104. URL `https://link.aps.org/doi/10.1103/PhysRevB.99.014104`.

Yoav Freund, Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

Yarin Gal et al. Uncertainty in deep learning, 2016.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020a.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020b.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020c.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

David Gfeller, Olivier Michielin, and Vincent Zoete. Shaping the interaction landscape of bioactive molecules. *Bioinformatics*, 29(23):3073–3079, 2013.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Query by committee made real. In *Neural Information Processing Systems (NeurIPS)*, 2005.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87, 2020.

Yuhong Guo and Russell Greiner. Optimistic active learning using mutual information. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. *Advances in neural information processing systems*, 20, 2007.

Filipp Gusev, Evgeny Gutkin, Maria G Kurnikova, and Olexandr Isayev. Active learning guided drug design lead optimization based on relative binding free energy modeling. *Journal of Chemical Information and Modeling*, 63(2):583–594, 2023.

Peter Hall. A distribution is completely determined by its translated moments. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 62(3):355–359, 1983.

Imran S Haque and Vijay S Pande. accelerating parallel evaluations of rocs. *Journal of computational chemistry*, 31(1):117–132, 2010.

Paul CD Hawkins, A Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *Journal of medicinal chemistry*, 50(1):74–82, 2007.

Hideitsu Hino. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*, 2020.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020.

Christian Hofbauer, Hans Lohninger, and András Aszódi. Surfcomp: a novel graph-based approach to molecular surface comparison. *Journal of chemical information and computer sciences*, 44(3): 837–847, 2004.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

Steven HOI, Rong JIN, Jianke ZHU, and Michael R LYU. Semi-supervised SVM batch mode active learning for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *ACM International Conference on World Wide Web*, 2006.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Lauri Himanen, and Adam S Foster. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, 4(1):37, 2018.

Chaitanya K Joshi, Cristian Bodnar, Simon V Mathis, Taco Cohen, and Pietro Lio. On the expressive power of geometric graph neural networks. *arXiv preprint arXiv:2301.09308*, 2023.

Arnaud S Karaboga, Florent Petronin, Gino Marchetti, Michel Souchet, and Bernard Maigret. Benchmarking of hpcc: a novel 3d molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *Journal of Molecular Graphics and Modelling*, 41:20–30, 2013.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30: 595–608, 2016.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Maksim Kulichenko, Kipton Barros, Nicholas Lubbers, Ying Wai Li, Richard Messerly, Sergei Tretiak, Justin S Smith, and Benjamin Nebgen. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science*, 3(3):230–239, 2023.

Ashutosh Kumar and Kam YJ Zhang. Advances in the development of shape similarity methods and their application in drug discovery. *Frontiers in chemistry*, 6:315, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Jouko Lampinen and Aki Vehtari. Bayesian approach for neural networks—review and case studies. *Neural networks*, 14(3):257–274, 2001.

Hongjian Li, Kwong-S Leung, Man-H Wong, and Pedro J Ballester. Usr-vs: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic acids research*, 44(W1):W436–W441, 2016.

Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 859–866, 2013.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, and Shuiwang Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, Keqiang Yan, Haoran Liu, Cong Fu, Bora M Oztekin, Xuan Zhang, and Shuiwang Ji. DIG: A turnkey library for diving into graph deep learning research. *Journal of Machine Learning Research*, 22(240):1–9, 2021. URL `http://jmlr.org/papers/v22/21-0343.html`.

Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in Neural Information Processing Systems*, 32:8464–8476, 2019.

Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations*, 2022.

Lazaros Mavridis, Brian D Hudson, and David W Ritchie. Toward high throughput 3d virtual screening using spherical harmonic surface representations. *Journal of chemical information and modeling*, 47(5):1787–1796, 2007.

Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

Manuela S Murgueitio, Sandra Santos-Sierra, and Gerhard Wolber. Discovery of novel tlr modulators by molecular modeling and virtual screening. *Journal of Cheminformatics*, 4:1–1, 2012.

Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Thomas S Rush, J Andrew Grant, Lidia Mosyak, and Anthony Nicholls. A shape-based 3-d scaffold hopping method and its application to a bacterial protein- protein interaction. *Journal of medicinal chemistry*, 48(5):1489–1495, 2005.

Lee Sael, David La, Bin Li, Raif Rustamov, and Daisuke Kihara. Rapid comparison of properties on protein surface. *Proteins: Structure, function, and bioinformatics*, 73(1):1–10, 2008.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

Adrian M Schreyer and Tom Blundell. Usrcat: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of cheminformatics*, 4(1):1–12, 2012.

Michel Schubiger, Goran Banjac, and John Lygeros. GPU acceleration of admm for large-scale quadratic programming. *Journal of Parallel and Distributed Computing*, 144:55–67, 2020.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Burr Settles. Active learning literature survey. 2009.

Jinling Shang, Xi Dai, Yecheng Li, Marco Pistolozzi, and Ling Wang. Hybridsim-vs: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. *Bioinformatics*, 33(21):3480–3481, 2017.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

Justin S Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*, 148(24), 2018.

Justin S Smith, Benjamin Nebgen, Nithin Mathew, Jie Chen, Nicholas Lubbers, Leonid Burakovsky, Sergei Tretiak, Hai Ah Nam, Timothy Germann, Saryu Fensin, et al. Automated discovery of a robust interatomic potential for aluminum. *Nature communications*, 12(1):1257, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020. doi: 10.1007/s12532-020-00179-2. URL `https://doi.org/10.1007/s12532-020-00179-2`.

Stephan Thaler, Felix Mayr, Siby Thomas, Alessio Gagliardi, and Julija Zavadlav. Active learning graph neural networks for partial charge prediction of metal-organic frameworks via dropout monte carlo. *npj Computational Materials*, 10(1):86, 2024.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

D Michael Titterington. Bayesian methods for neural networks and related models. *Statistical science*, pages 128–139, 2004.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)*, 2:45–66, 2001.

Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32:15642–15651, 2019.

Cas van der Oord, Matthias Sachs, Dávid Péter Kovács, Christoph Ortner, and Gábor Csányi. Hyperactive learning for data-driven interatomic potentials. *npj Computational Materials*, 9(1): 168, 2023.

Vishwesh Venkatraman, Padmasini Ramji Chakravarthy, and Daisuke Kihara. Application of 3d zernike descriptors to shape-based ligand similarity searching. *Journal of cheminformatics*, 1(1): 1–19, 2009.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2014.

Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuiwang Ji. ComENet: Towards complete and efficient message passing for 3D molecular graphs. In *The 36th Annual Conference on Neural Information Processing Systems*, pages 650–664, 2022.

Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3D graph networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=9X-hgLDLYkQ`.

Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *The 36th Annual Conference on Neural Information Processing Systems*, pages 15066–15080, 2022.

Xin Yan, Jiabo Li, Zhihong Liu, Minghao Zheng, Hu Ge, and Jun Xu. Enhancing molecular shape comparison by weighted gaussian functions. *Journal of chemical information and modeling*, 53(8): 1967–1978, 2013.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019.

Randy J Zauhar, Eleonora Gianti, and William J Welsh. Fragment-based shape signatures: a new tool for virtual screening and drug discovery. *Journal of Computer-Aided Molecular Design*, 27: 1009–1036, 2013.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, YuQing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, 2022.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2022.

Ting Zhou, Karine Lafleur, and Amedeo Caflisch. Complementing ultrafast shape recognition with an optical isomerism descriptor. *Journal of Molecular Graphics and Modelling*, 29(3):443–449, 2010.

Jia-Jie Zhu and José Bento. Generative adversarial active learning. In *arXiv:1702.07956*, 2017.

A Appendix

A.1 Proofs of the Theorems and Propositions

Theorem 1. *If A and B are triangular isometric, then A and B are reference distance isometric; If A and B are cross-angle isometric, then A and B are triangular isometric.*

Proof. Suppose that A and B are triangular isometric, then there exists a collection of global group elements $\gamma_i \in SE(3)$ such that

$$(r_2, f\left(a_{\text{far}}\right), f(a_i)) = (\gamma_i r_1, \gamma_i a_{\text{far}}, \gamma_i a_i), \quad \forall a_i \in A.$$

It follows immediately that for each point $a_i \in A$, $(r_2, f(a_i)) = (\gamma_i r_1, \gamma_i a_i)$ also holds. Thus, if A and B are triangular isometric, then A and B are reference distance isometric.

Suppose that A and B are cross-angular isometric, then there exists a collection of global group elements $\gamma_{ij} \in SE(3)$ such that

$$(r_2, f(a_j), f(a_i)) = (\gamma_{ij} r_1, \gamma_{ij} a_j, \gamma_{ij} a_i), \quad \forall a_i, a_j \in A, i \neq j.$$

By fixing a_j to be a_{far} and the corresponding γ_{ij} to be γ_i, it follows immediately that for each point $a_i \in A$, there exists a collection of global group elements $\gamma_i \in SE(3)$ such that $(r_2, f\left(a_{\text{far}}\right), f(a_i)) = (\gamma_i r_1, \gamma_i a_{\text{far}}, \gamma_i a_i)$. Thus, if A and B are cross-angular isometric, then A and B are triangular isometric.

□

Proposition 1. *GR_{ours} is at least as expressive as the GWL test. In other words, GR_{ours} suffices to distinguish any non-isomorphic molecular structures that are distinguishable by any 3D GNN.*

Proof. We prove the case when the reference points r_1 and r_2 are the centroids of the point clouds A and B, respectively. The proof for other choices of reference points follows analogously. First, we will show that ζ, which is the function that gives us the geometric representation GR_{ours} given a point cloud, is $E(3)$-orbit injective.

Without loss of generality, assume that the centroids of these two point clouds are at the origin. Otherwise, they can be fixed by a translation in $\mathbb{T}(3) \cong E(3)/O(3)$. For simplicity, we denote the point $f(a_i)$ as b_i, and bold symbol represents the corresponding vectors. Suppose that GR_{ours} is the same for both points clouds A and B, that is to say, we have the following conditions:

$$||\boldsymbol{a_i}|| = ||\boldsymbol{b_i}||, \ \forall i \in \mathbb{N}_{\leq n} \tag{5}$$

$$\frac{\langle \boldsymbol{a}_{\text{far}}, \boldsymbol{a_i} \rangle}{\|\boldsymbol{a}_{\text{far}}\| \cdot \|\boldsymbol{a_i}\|} = \frac{\langle \boldsymbol{b}_{\text{far}}, \boldsymbol{b_i} \rangle}{\|\boldsymbol{b}_{\text{far}}\| \cdot \|\boldsymbol{b_i}\|}, \ \forall i \in \mathbb{N}_{\leq n} \tag{6}$$

$$\frac{\langle \boldsymbol{a}_{\text{far}}, \boldsymbol{a_i} - \boldsymbol{a_j} \rangle}{\|\boldsymbol{a}_{\text{far}}\| \cdot \|\boldsymbol{a_i} - \boldsymbol{a_j}\|} = \frac{\langle \boldsymbol{b}_{\text{far}}, \boldsymbol{b_i} - \boldsymbol{b_j} \rangle}{\|\boldsymbol{b}_{\text{far}}\| \cdot \|\boldsymbol{b_i} - \boldsymbol{b_j}\|}, \ \forall i, j \in \mathbb{N}_{\leq n}, i \neq j. \tag{7}$$

It follows from (5) and (6) that for any $k \in \mathbb{N}_{\leq n}$,

$$\langle \boldsymbol{a}_{\text{far}}, \boldsymbol{a_k} \rangle = \langle \boldsymbol{b}_{\text{far}}, \boldsymbol{b_k} \rangle.$$

Then, for all $i, j \in \mathbb{N}_{\leq n}, i \neq j$,

$$\langle \boldsymbol{a}_{\text{far}}, \boldsymbol{a_i} - \boldsymbol{a_j} \rangle = \langle \boldsymbol{b}_{\text{far}}, \boldsymbol{b_i} - \boldsymbol{b_j} \rangle.$$

Thus, it is clear from (7) that all the pair-wise distances are the same, i.e., $\|\boldsymbol{a_i} - \boldsymbol{a_j}\| = \|\boldsymbol{b_i} - \boldsymbol{b_j}\|$ for all $i, j \in \mathbb{N}_{\leq n}, i \neq j$. Thus

$$\|\boldsymbol{a_i} - \boldsymbol{a_j}\|^2 = \|\boldsymbol{a_i}\|^2 - 2\langle \boldsymbol{a_i}, \boldsymbol{a_j} \rangle + \|\boldsymbol{a_j}\|^2$$
$$\|\boldsymbol{b_i} - \boldsymbol{b_j}\|^2 = \|\boldsymbol{b_i}\|^2 - 2\langle \boldsymbol{b_i}, \boldsymbol{b_j} \rangle + \|\boldsymbol{b_j}\|^2$$

It follows that $\langle \boldsymbol{a_i}, \boldsymbol{a_j} \rangle = \langle \boldsymbol{b_i}, \boldsymbol{b_j} \rangle$ from (5).

It is safe to assume that $(\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n})$ spans $\mathbb{E}^3$, otherwise the proof is trivial when all points are co-planer. Without loss of generality, let $(\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3})$ be a basis for $\mathbb{E}^3$. It is easy to see that $(\boldsymbol{b_1}, \boldsymbol{b_2}, \boldsymbol{b_3})$ is also a basis for $\mathbb{E}^3$.

Let X and Y denote the matrices whose columns are $(\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3})$ and $(\boldsymbol{b_1}, \boldsymbol{b_2}, \boldsymbol{b_3})$, respectively. Let G denote the associated Gram matrix, i.e. $G = X^T X = Y^T Y$, then G is symmetric and positive semi-definite. Moreover, as both X and Y are full-rank, there exist orthogonal matrices Q_X, Q_Y and upper triangular matrices R_X, R_Y such that

$$\begin{cases} X = Q_X R_X \\ Y = Q_Y R_Y \end{cases}$$

then

$$\begin{cases} G = X^\top X = R_X^\top R_X \\ G = Y^\top Y = R_Y^\top R_Y \end{cases}$$

The form above follows the pattern of Cholesky decomposition. As G is symmetric and positive semi-definite, the Cholesky decomposition is unique. Thus, $R_X = R_Y$. Thus, $X = Q_X Q_Y^{-1} Y$, where $Q_X Q_Y^{-1}$ is an orthogonal matrix. Thus, there exists $g \in O(3)$ such that $gX = Y$. If $n \leq 3$, this completes the proof.

When $n \geq 4$, for any $k \geq 4$, $\boldsymbol{a_k} = \sum_{i=1}^{3} c_i \boldsymbol{a_i}$, where $\{c_i\}_{i=1}^{3}$ are uniquely determined by $c_i = \langle \boldsymbol{a_k}, \boldsymbol{a_i} \rangle$. Then, $g\boldsymbol{a_k} = g\left(\sum_{i=1}^{3} c_i \boldsymbol{a_i}\right) = \sum_{i=1}^{3} c_i \left(g\boldsymbol{a_i}\right) = \sum_{i=1}^{3} c_i \boldsymbol{b_i} = \boldsymbol{b_k}$.

Now, without loss of generality, we can conclude that if $\zeta(A) = \zeta(B)$, then there exists $g \in E(3)$ such that $gA = B$. As we have an injective map, our method is naturally at least as expressive as the GWL test for $E(3)$ isomorphism. As a result, our method surpasses all existing 3D GNNs in terms of distinguishing non-isomorphic point clouds.

□

A.2 Geometric Weisfeiler-Leman (GWL) Test

For the Geometric Weisfeiler-Leman (GWL) test, consider a graph $\mathcal{G}$ with its set of vertices represented as $\mathcal{V}(\mathcal{G})$ and its set of edges as $\mathcal{E}(\mathcal{G})$. A vertex in graph $\mathcal{G}$ is denoted by i, and $\mathcal{N}_i$ signifies the set of vertices adjacent to i. The color of vertex i at iteration t is given by $c_i^{(t)}$, and the geometric object for vertex i at iteration t is represented by $\boldsymbol{g}_i^{(t)}$.

The procedure for the GWL test is as follows:

1. **Initialization**: Each vertex i is assigned an initial color $c_i^{(0)}$ and a geometric object $\boldsymbol{g}_i^{(0)}$, typically based on its local property or geometric attributes.
2. **Iterative Aggregation**: For each iteration $t \geq 1$, the geometric object of each vertex i is updated to aggregate geometric information from its t-hop neighborhood, represented as $\boldsymbol{g}_i^{(t)}$, which includes the colors and geometric objects from the previous iteration of vertex i and its neighbors.
3. **Color Update**: The color of each vertex i at iteration t is computed by aggregating the geometric information around vertex i using a $\mathfrak{G}$-orbit injective and $\mathfrak{G}$-invariant function, denoted by I-HASH, i.e., $c_i^{(t)} := \mathrm{I}^{-\mathrm{HASH}^{(t)}}\left(\boldsymbol{g}_i^{(t)}\right)$.
4. **Termination**: The procedure terminates when colors do not change from the previous iteration or a predetermined maximum number of iterations is reached.
5. **Graph Comparison**: Finally, two geometric graphs $\mathcal{G}$ and $\mathcal{H}$ are geometrically non-isomorphic if there exists some iteration t for which the sets of colors of their vertices are not equal, i.e., $\left\{\left\{c_i^{(t)} \mid i \in \mathcal{V}(\mathcal{G})\right\}\right\} \neq \left\{\left\{c_i^{(t)} \mid i \in \mathcal{V}(\mathcal{H})\right\}\right\}$.

A.3 Statistical Moments

The equations that we used for calculating four moments are as follows.

The **mean**, often referred to as the average, represents the sum of all data points divided by the number of data points and is given by

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{8}$$

Variance measures the spread or dispersion of a dataset and is defined as

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \text{Mean})^2}{n-1}. \tag{9}$$

Skewness gauges the asymmetry of a dataset's distribution. Here we sightly change its definition to be positive for convenience as

$$\text{Skewness} = \frac{\sum_{i=1}^{n}|x_i - \text{Mean}|^3/n}{\{\sum_{i=1}^{n}(x_i - \text{Mean})^2/(n-1)\}^{3/2}}. \tag{10}$$

Kurtosis assesses the "tailedness" of a dataset's distribution as

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \text{Mean})^4/n}{\{\sum_{i=1}^{n}(x_i - \text{Mean})^2/(n-1)\}^{2}}. \tag{11}$$

A.4 Schematic Diagram of our Framework

A schematic diagram of our active sampling framework is depicted in Fig. 6. We are given a labeled training set L, an unlabeled set U and a query budget k for each active learning iteration. The SphereNet model is first trained on the labeled set L. In the second step, the trained model is applied on the unlabeled set to compute a prediction uncertainty of each unlabeled molecule, which is used to populate the uncertainty vector r; the diversity matrix D is also computed in this step where $D(i, j)$ is the diversity between unlabeled molecules x_i and x_j. Next, the QP problem is solved to select k unlabeled molecules for annotation. These molecules are removed from the unlabeled set U and appended to the labeled set L. The active sampling process is continued iteratively until some stopping criterion is satisfied (taken as 7 iterations in our work).

Note that, computing the diversity matrix D in Step 3 needs to be executed just once for the whole process. Once we have the initial D, as more and more samples are queried through AL, we keep deleting the corresponding rows and columns from D to derive the updated matrix.

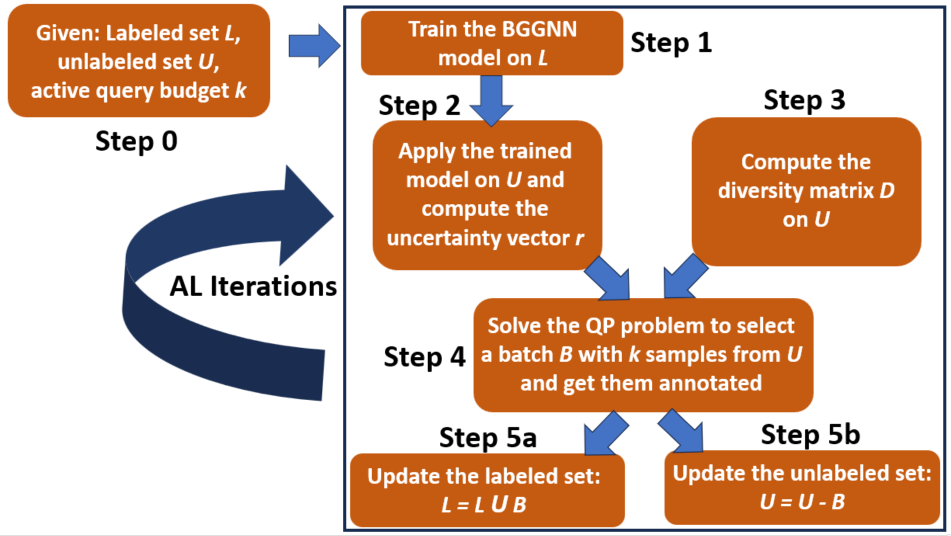
Given: Labeled set L, unlabeled set U, active query budget k
Step 0
Train the BGGNN model on L
Step 1
Step 2
Apply the trained model on U and compute the uncertainty vector r
Step 3
Compute the diversity matrix D on U
AL Iterations
Step 4
Solve the QP problem to select a batch B with k samples from U and get them annotated
Step 5a
Update the labeled set: L = L U B
Step 5b
Update the unlabeled set: U = U - B

Figure 6: Schematic diagram of the proposed active learning framework.

A.5 Related Work for Active Learning

Active Learning (AL) is a well-researched problem in the machine learning community [Settles, 2009]. Uncertainty sampling is an important strategy for AL, where unlabeled samples with the highest prediction uncertainties are queried for annotation. Several techniques have been explored to compute the uncertainty, such as Shannon's entropy [Guo and Schuurmans, 2007, Li and Guo, 2013], the distance of a sample from the separating hyperplane for SVM classifiers [Tong and Koller, 2001], the disagreement among a committee of classifiers regarding the label of a sample [Freund et al.,

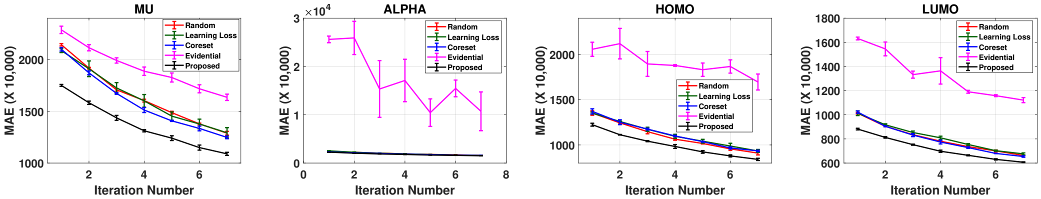

Figure 7: Active learning performance results. The graphs show the mean (averaged over 3 runs) and the errorbars for all the methods. Best viewed in color.

1997, Gilad-Bachrach et al., 2005], among others [Hoi et al., 2006, HOI et al., 2008, Guo and Greiner, 2007, Freytag et al., 2014]. With the advent of deep learning, Deep AL has attracted significant research attention [Hino, 2020, Ren et al., 2021], Entropy-based methods are developed as well [Wang and Shang, 2014, Ranganathan et al., 2017]. Yoo and Kweon [2019] cascaded a task-agnostic loss learning module that queries samples with the highest predicted loss values. Huang et al. [2021] proposed a strategy based on temporal output discrepancy. Techniques based on adversarial training have also been explored [Sinha et al., 2019, Mayer and Timofte, 2020, Zhang et al., 2020, Zhu and Bento, 2017]. Bayesian neural networks (BNNs) [Lampinen and Vehtari, 2001, Titterington, 2004, Goan and Fookes, 2020] and deep model ensemble [Lakshminarayanan et al., 2017, Huang et al., 2017] generally achieve superior performance but may induce excessive computational cost.

Diversity/representativeness based AL sampling has also been exploited. A core-set sampling technique proposed by **?** queries a batch of samples such that a model trained on the queried subset is competitive for the remaining data samples. Diversity sampling has also been exploited in the context of Bayesian neural networks [Kirsch et al., 2019]. Buchert et al. [2023] uses diversity sampling, together with self-supervised representation learning to select an informative seed set for AL. Combinations of uncertainty/diversity/representativeness-based criteria have also been used as query functions in AL research [Chakraborty et al., 2015, Wu et al., 2022, Ash et al., 2020].

A.6 Results with the Evidential Uncertainty Baseline

The active learning performance results on the four properties studied (***mu, alpha, homo, and lumo***) are depicted in Fig. 7. As mentioned in Sec. 4.2, we note that *Evidential Uncertainty* depicts significantly high error values than the other methods, for all the four properties.

A.7 Performance using the DimeNet^{++} Backbone

The objective of this experiment is to study the performance of our framework with DimeNet^{++} [Gasteiger et al., 2020a] as the backbone model of our active learning approach. We use the ***mu and lumo*** properties from the QM9 dataset in this experiment. We use the same experimental setup as detailed in Section 4.1 of the paper. The results are depicted in Figure 8. The proposed framework consistently outperforms all the baselines at each AL iteration across both the datasets.

The results of the statistical tests of significance are reported in Table 3. Each entry in the table denotes the p-value of the paired t-test between our method and the corresponding baseline (denoted in the columns) for the property studied (denoted in the rows). We note that the performance improvement achieved by our method is statistically significant ($p < 0.05$) compared to all the baselines for both the properties. These results corroborate the robustness of our framework to the underlying GNN backbone.

Table 3: The table shows the p-values obtained using paired t-test between the results our method against each of the baselines for the ***mu and lumo*** properties, using Dimenet^{++} backbone.

Properties	Baselines		
	Random	Learning Loss	Coreset
mu	1.56×10^{-6}	1.87×10^{-4}	$1.25{\times}10^{-4}$
lumo	8.55×10^{-4}	4.24×10^{-4}	1.16×10^{-2}

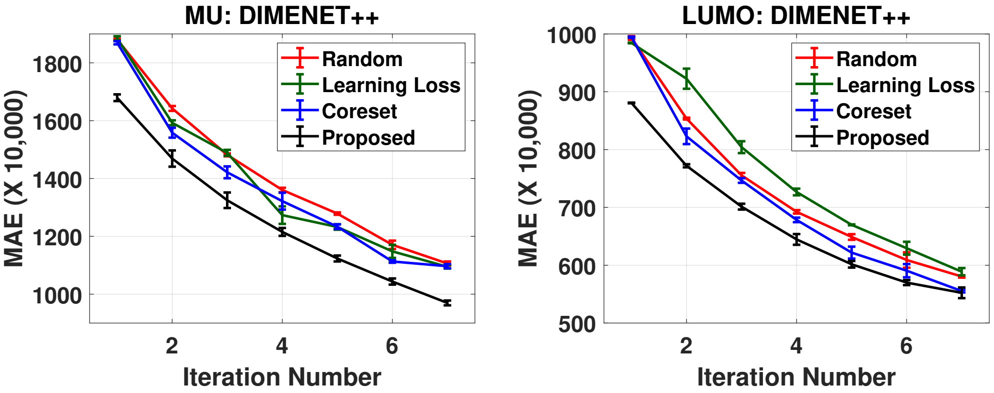

Figure 8: Study of our framework using the DimeNet^{++} backbone. The graphs show the mean (averaged over 3 runs) and the errorbars for all the methods. Best viewed in color.

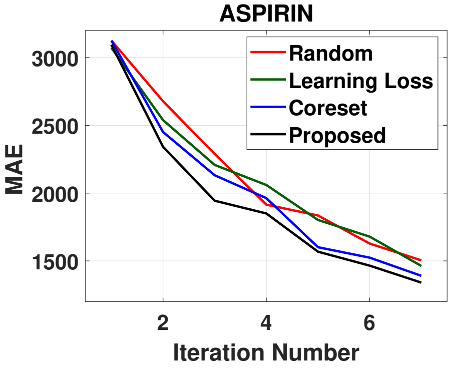

Figure 9: Study of our framework on Aspirin molecules using the SphereNet backbone. Best viewed in color.

A.8 Generalization: Performance on the MD17 Dataset

The research of 3D molecular learning is new, and there are only a few reliable benchmark datasets for 3D molecules (containing atom types as well as XYZ coordinates for all atoms for each molecule). To test the generalization ability of our proposed method, we study the performance of our framework on the MD17 dataset [Chmiela et al., 2017a]. QM9 consists of molecules in equilibrium, while MD17 contains several thermalized (*i.e.*, non-equilibrium, slightly moving) molecular systems. Additionally, QM9 contains various quantum properties for molecules, like the important *homo* and *lomo* orbitals. MD17 is for dynamic system simulation, thus it contains labels for both the energy and atomic forces. In summary, we test our methods on molecule systems in both equilibrium and non-equilibrium, covering various quantum properties and molecular dynamics tasks.

We used 300 samples as the initial training set, 700 samples as the unlabeled set and $1,000$ test samples; we used 100 as the batch size and conducted 7 iterations of active learning. For this dataset, we train the network for 500 epochs. The results on ***Aspirin*** molecules using the SphereNet as the backbone of our GNN are depicted in Fig. 9. Our framework once again depicts promising performance and attains the lowest MAE values across all the AL iterations compared to all the baselines. These results further demonstrate the promise and potential of our method for scientific applications.

A.9 Statistical Tests of Significance for the Query Budget Experiment

Table 4 reports the results of the statistical tests of significance for the study of query budget (presented in Sec. 4.3). Each entry in the table denotes the p-value of the paired t-test between our method

Table 4: The table shows the p-values obtained using paired t-test between the results our method against each of the baselines for the ***mu*** property for query budgets $1,000$, $1,500$ and $2,000$.

Budget	Baselines			
	Random	Learning Loss	Coreset	Evidential
1000	7.58×10^{-6}	1.05×10^{-5}	$5.32{\times}10^{-5}$	2.46×10^{-10}
1500	7.54×10^{-6}	5.09×10^{-5}	1.51×10^{-4}	2.19×10^{-7}
2000	7.90×10^{-5}	1.74×10^{-5}	1.94×10^{-4}	1.77×10^{-8}

and the corresponding baseline (denoted in the columns) for the query budget (denoted in the rows) for the ***mu*** property. From the table, we note that the improvement in performance achieved by our method is statistically significant ($p < 0.05$) compared to all the baselines, consistently for all the query budgets.

A.10 Addtional Ablation Studies

The Individual Impact of Diversity and Uncertainty Components. In Fig. 10, we present the result on the individual impact of the diversity and uncertainty components. Our proposed method outperforms the individual use of diversity or uncertainty alone. The key to this outperformance lies in our method's dual focus on both geometric importance and chemical contexts. Moreover, it can be observed that the diversity component alone shows strong performance; it is only slightly less effective than our method because it captures the geometries of molecules, which are fundamental in distinguishing different molecules with different properties. On top of this, we also conduct statistical tests to conclude that the improvement of our method is significant compared to only diversity or only uncertainty in Table 5.

Table 5: The table shows the p-values obtained using paired t-test between the result of our method against uncertainty only and diversity only components in ablation study for ***mu*** and ***lumo*** prediction.

Properties	Components	
	Uncertainty Only	Diversity Only
mu	$1.24{\times}10^{-5}$	$1.82{\times}10^{-4}$
lumo	$7.46{\times}10^{-6}$	$2.26{\times}10^{-4}$

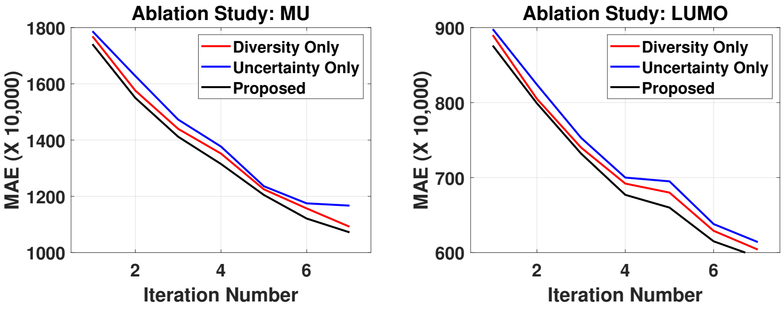

Figure 10: Ablation study results studying the individual impact of uncertainty and diversity on the ***mu*** and ***lumo*** properties with SphereNet. Best viewed in color.

A Comparison between Our Diversity Component and SOAP. We include a comparison between our method and one that uses a well-known geometric descriptor in chemistry, the SOAP descriptor [Bartók et al., 2013]. SOAP produces descriptors that characterize local atomic environments using spherical harmonics and radial basis functions. It incorporates both geometric information and elemental (species) details. Table 6 presents the results on the QM9 dataset for two important properties, *mu* and *lumo*. These results demonstrate that our diversity component clearly outperforms SOAP. Consequently, our overall method, which combines both diversity and uncertainty, also surpasses the performance of the SOAP descriptors. The p-values in Table 7 further illustrate that our proposed method significantly improves the selection strategy compared to SOAP. This improvement can be attributed to the more localized nature of the SOAP descriptors and the inability to maintain permutation invariance.

Table 6: The table shows a comparison between our proposed descriptor and the SOAP descriptor for the properties ***mu*** and ***lumo***, using SphereNet as the backbone.

Iteration	*mu*				*lumo*			
	SOAP D	Our D	SOAP B	Our B	SOAP D	Our D	SOAP B	Our B
1	2154	**1769**	2057	**1741**	1016	**890**	1013	**876**
2	1901	**1576**	1877	**1550**	921	**805**	902	**799**
3	1732	**1440**	1721	**1412**	839	**740**	838	**732**
4	1701	**1352**	1539	**1315**	797	**692**	791	**677**
5	1587	**1225**	1456	**1205**	759	**680**	744	**660**
6	1414	**1157**	1345	**1121**	699	**629**	711	**615**
7	1322	**1092**	1280	**1072**	667	**604**	681	**594**

Abbreviations: D means using diversity Only; B means using both uncertainty + diversity.
The results from the method with superior performance are highlighted in bold.

Table 7: The table shows the p-values obtained from a paired t-test comparing the results of our method against those of SOAP for the properties ***mu*** and *lumo*, using SphereNet as the backbone.

	mu	*lumo*
p-value	4.20×10^{-6}	2.51×10^{-6}

A Comparison between our method and BatchBALD. We investigate the impact of our quadratic programming formulation compared to the greedy, clustering-based Bayesian uncertainty baseline, BatchBALD [Kirsch et al., 2019], which selects a diverse batch of samples by maximizing mutual information. Our method provides a more structured approach to uncertainty, particularly tailored for 3D molecular data. The results, presented in Fig. 11, demonstrate the effectiveness of our approach. Additionally, we conducted statistical tests, as shown in Table 8, which confirm that the improvement of our method over the baselines is statistically significant. This outperformance can be attributed to the components specifically designed for 3D molecular graphs.

Table 8: The p-values obtained using paired t-test between the results our method against each of the baselines for all the properties studied. *Here, L. Loss refers to Learning Loss.*

Properties	Baselines				
	Random	L. Loss	Coreset	Evidential	BatchBALD
mu	7.54×10^{-6}	5.09×10^{-5}	1.51×10^{-4}	2.19×10^{-7}	7.20×10^{-5}
alpha	1.06×10^{-5}	8.14×10^{-4}	4.27×10^{-5}	2.72×10^{-4}	4.86×10^{-6}
homo	2.26×10^{-5}	8.36×10^{-7}	4.23×10^{-6}	1.71×10^{-8}	1.54×10^{-4}
lumo	4.48×10^{-5}	1.25×10^{-5}	3.12×10^{-4}	2.39×10^{-6}	3.96×10^{-5}

A.11 Licenses for Existing Assets

We list all the licenses for existing assets in Table 9.

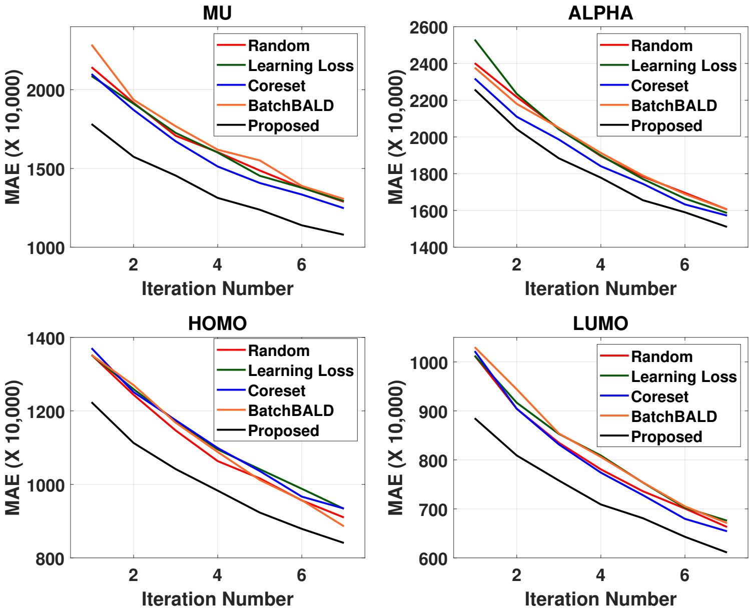

Figure 11: Active learning performance results with SphereNet on QM9 Dataset. Best viewed in color.

Table 9: Assets, Licenses, and Descriptions

Asset	License	Description
SphereNet [Liu et al., 2022]	GNU General Public License v3.0	GNN Model
DimeNet^{++} [Gasteiger et al., 2020a]	Hippocratic License v2.1	GNN Model
QM9 [Ramakrishnan et al., 2014]	CC BY-NC-SA 4.0 International License	Benchmark Dataset
MD17 [Ramakrishnan et al., 2014]	CC BY-NC-SA 4.0 International License	Benchmark Dataset
Coreset [Sener and Savarese, 2018]	MIT License	Active Learning Scheme
Learning Loss [Yoo and Kweon, 2019]	N/A	Active Learning Scheme
Evidential Uncertainty [Beluch et al., 2018, Amini et al., 2020]	N/A, Apache-2.0 License	Active Learning Scheme

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our novelty, contributions, and scope are accurately supported theoretically and empirically.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a separate Limitation paragraph at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proofs are complete without strong assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The used dataset, baseline models, and implementation details are provided in great detail in the main paper and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the code and the link is provided in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setup details are provided in both the main paper and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Most results are averaged over the three splits to rule out the effects of randomness. Both means and error bars are given.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information on the used GPUs is provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include a separate paragraph to discuss the potential societal impacts of our research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The licenses for all the data and models we used are explicitly mentioned in Table 9. In certain cases where no official implementation was released, we implemented our own; thus, the license is not applicable, and we note *N/A* in the table.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.