Geometric Mixture Models for Electrolyte Conductivity Prediction

Anyi Li[1 2 3], **Jiacheng Cen**[1 2 3], **Songyou Li**[1 2 3], **Mingze Li**[1 2 3], **Yang Yu**[4] , **Wenbing Huang**[1 2 3*]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3] Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4] Hisun Pharm
{li_anyi, jiacc.cn, yangyubard}@outlook.com;
{songyou_li, hwenbing}@126.com; limingzetony@gmail.com;

Abstract

Accurate prediction of ionic conductivity in electrolyte systems is crucial for advancing numerous scientific and technological applications. While significant progress has been made, current research faces two fundamental challenges: (1) the lack of high-quality standardized benchmarks, and (2) inadequate modeling of geometric structure and intermolecular interactions in mixture systems. To address these limitations, we first reorganize and enhance the CALiSol and DiffMix electrolyte datasets by incorporating geometric graph representations of molecules. We then propose GeoMix, a novel geometry-aware framework that preserves Set-SE(3) equivariance—an essential but challenging property for mixture systems. At the heart of GeoMix lies the Geometric Interaction Network (GIN), an equivariant module specifically designed for intermolecular geometric message passing. Comprehensive experiments demonstrate that GeoMix consistently outperforms diverse baselines (including MLPs, GNNs, and geometric GNNs) across both datasets, validating the importance of cross-molecular geometric interactions and equivariant message passing for accurate property prediction. This work not only establishes new benchmarks for electrolyte research but also provides a general geometric learning framework that advances modeling of mixture systems in energy materials, pharmaceutical development, and beyond.

1 Introduction

Predicting the properties of multi-molecular mixture systems plays a central role in advancing a wide range of emerging technologies, such as next-generation energy storage, advanced materials design, and pharmaceutical formulation [1–4]. Electrolytes, as a prototypical example of these systems, typically consist of lithium salts and organic solvents of diverse types. Accurate prediction of electrolyte conductivity is a key research focus, as it directly determines critical battery performance metrics including charge transfer efficiency, power density, and operational stability [5; 6]. However, modeling such complex systems presents significant challenges, as it requires simultaneously accounting for both the internal structure of individual molecules and the intricate interactions among different molecules. Classical theoretical approaches—ranging from the Nernst equation to dilute solution models—either neglect explicit intermolecular forces or rely on extrapolations from infinitely dilute regimes, resulting in limited accuracy and poor universality for practical applications [7; 8].

Spurred by recent advancements in machine learning, particularly deep learning, researchers have increasingly adopted learning-based approaches to model electrolyte systems [9–11], achieving

*Wenbing Huang is the corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

significant progress over classical methods. As a data-driven paradigm, the development of predictive models for these systems hinges critically on high-quality and standardized benchmarks. However, the field still faces substantial challenges: most existing datasets are semi-structured, requiring labor-intensive manual curation from primary literature due to inconsistent formatting standards. This lack of accessible, well-curated benchmarks has severely hindered systematic comparisons and reproducibility across methodologies.

Another critical limitation in existing research stems from inadequate geometric characterization of mixture systems. Previous works such as MM-MoLFormer [12] and MolSets [13] represent molecules in electrolyte systems using SMILES sequences or topological graphs. However, these approaches neglect the geometric structure of molecules in 3D space, which significantly impacts the properties of entire systems [14–17]. While recent methods such as DiffMix [11] and Uni-ELF [18] have incorporated 3D structural information, they aggregate molecular information into global features and simulate intermolecular interactions through comparing these aggregated representations, thereby overlooking fine-grained and full-atom interactions across molecules. A naive solution would involve direct atom-level coordinate transfer between molecules, but this violates the fundamental symmetries of mixture systems. Crucially, any physically meaningful representation must maintain consistency under: 1) independent permutation of each molecule's atoms; 2) independent SE(3) transformations (rotation/translation) of each molecule's local coordinate system; and 3) permutation of molecules within the system. We formalize these requirements as *Set-SE(3) equivariance* (see Eqs. (2) to (4)), a challenging but essential property for developing accurate models of molecular interactions.

To address the above two challenges, we first curate and standardize two public datasets for electrolyte conductivity prediction: CALiSol [19] and DiffMix [11]. Inspired by the powerful expressiveness of geometric information [20–44], we augment the molecular representation by constructing geometric graphs where atoms are characterized by both invariant features (*e.g.*, atomic numbers) and equivariant features (*e.g.*, 3D coordinates), ensuring downstream models can fully exploit 3D structural information. We then propose **GeoMix**, a geometry-aware framework for mixture system modeling and property prediction. At its core is the Geometric Interaction Network (GIN), a novel equivariant module designed for intermolecular geometric message passing. The fundamental idea involves first creating a local coordinate frame for each molecule, then computing learnable coordinate transformations to align atom pairs from different local frames, and finally performing equivariant message passing over the transformed coordinates. Our code and dataset are available at Github[2].

In summary, this paper contributes to the following aspects:

- We curate and standardize two public datasets, CALiSol and DiffMix, covering a wide range of electrolyte mixture systems. The datasets are carefully processed with geometric graph construction, forming a benchmark for electrolyte conductivity prediction.
- We propose GeoMix, a novel geometric framework for modeling mixture systems. GeoMix maintains Set-SE(3) equivariance while capturing fine-grained geometric relationships between molecules, overcoming the limitations of existing methods.
- We compare GeoMix against diverse baseline models built upon MLPs, GNNs, and geometric GNNs, under consistent settings. GeoMix consistently outperforms all baselines on both datasets, demonstrating its effectiveness in modeling mixture systems.

2 Related Work

Early-Stage Methods Earliest approaches compute ionic conductivity by applying the Nernst–Einstein equation, which assumes that all ions diffuse independently and intermolecular interactions can be neglected. Under this framework, each ionic species contributes additively to the total conductivity. While effective at low concentrations, this framework breaks down in concentrated or strongly interacting systems, where correlated ionic motion becomes significant [8]. With advances in computational chemistry, first-principles methods such as Density Functional Theory (DFT) [45] and Machine Learning Force Fields (MLFFs) [10; 46] became available. These techniques accurately capture interactions between ions and molecules over short timescales, enabling more precise estimates of ion mobilities under realistic intermolecular forces [10; 47–49]. Despite their accuracy, classical and ab initio MD simulations require modeling every atom in the local system, leading

[2] https://github.com/GLAD-RUC/GeoMix

to substantial computational expense. Moreover, because conductivity predictions rely on multiple intermediate steps, accumulated numerical errors can offset the benefits of high precision.

Deep Learning Methods Recent deep learning frameworks [11–13; 18; 50; 51] aim to model mixture systems directly from their components and proportions. These methods can be roughly grouped into three categories: (1) *SMILES-based methods* [12; 50; 51] compute mixture embeddings by weighted summation over molecular embeddings derived from molecular text representations such as SMILES. However, these approaches ignore the influence of molecular geometry on both the representation and interactions within the mixture. (2) *Topology-based methods* such as MolSets [13] model intermolecular interactions using self-attention over molecular-level graph embeddings. While molecular topology is captured, spatial interactions are only implicitly encoded through attention mechanisms. (3) *Geometry-based methods* further leverage 3D molecular information. DiffMix [11] aggregates molecular embeddings using learned polynomial mixing coefficients, while incorporating some physical priors. Uni-ELF [18] combines microscopic descriptors such as radial distribution functions (RDFs) with learned geometric features to extract invariant intermolecular representations. However, it does not capture equivariant features between molecules. While most studies about equivariant models focus on molecular interactions within a single reference frame [52; 53], our proposed method explicitly models node-level geometric relationships across molecules through equivariant message passing. This approach enables fine-grained, geometry-aware interaction modeling at the atomic scale.

3 Method

In this section, we first formulate the regression task for mixture systems and discuss its associated symmetry constraints in § 3.1. We then introduce GeoMix, a novel geometric learning framework designed for both scalar (invariant) and vector (equivariant) property prediction, in § 3.2. Finally, we show a crucial component of GeoMix, GIN, in § 3.3. All proofs are provided in Appendix A.

3.1 Preliminaries

Notations We employ three-level representations to describe mixture systems in a hierarchical way. 1) **System-level**: A mixture system is a set of M molecules, *i.e.*, $\mathbb{S} = \{\mathcal{G}_m\}_{m=1}^{M}$, where the geometric graph $\mathcal{G}_m$ represents the m-th molecule in the system. The global environment vector $\boldsymbol{c}$ encodes physical conditions (*e.g.* temperature and pressure) and κ denotes the target electrolyte conductivity. 2) **Graph-level**: Each geometric graph $\mathcal{G}_m(\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m, \vec{\boldsymbol{V}}_m, w_m)$ contains N_m nodes. The quantities $\boldsymbol{H}_m \in \mathbb{R}^{H\times N_m}, \vec{\boldsymbol{X}}_m \in \mathbb{R}^{3\times N_m}$ denote node scalar features and node 3D coordinates, respectively; $\vec{\boldsymbol{V}}_m \in \mathbb{R}^{3\times N_m}$ refers to node geometric features, which are usually initialized as zeros or just node coordinates $\vec{\boldsymbol{X}}_m$; $0 < w_m < 1$ is the proportion of molecule $\mathcal{G}_m$ in the mixture, and the sum of all w_ms is equal to 1. All coordinates $\vec{\boldsymbol{X}}_m$ are centered at the origin beforehand, implying that the model should be translation-invariant. 3) **Node-level**: We denote by $\boldsymbol{h}_i^{(m)} \in \mathbb{R}^H, \vec{\boldsymbol{x}}_i^{(m)}, \vec{\boldsymbol{v}}_i^{(m)} \in \mathbb{R}^3$ the representations of the i-th node in $\mathcal{G}_m$.

Task Definition We aim to predict both scalar and geometric properties of the mixture system $\mathbb{S}$. Formally, our goal is to learn a mapping φ from system space to property space:

$$\varphi\colon (\{\mathcal{G}_m(\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m, \vec{\boldsymbol{V}}_m, w_m)\}_{m=1}^{M}, \boldsymbol{c}) \mapsto (\{\boldsymbol{H}'_m, \vec{\boldsymbol{V}}'_m\}_{m=1}^{M}, \kappa), \tag{1}$$

where $\boldsymbol{H}'_m$ and $\vec{\boldsymbol{V}}'_m$ respectively represent the predicted scalar and geometric properties of all nodes. These node-level properties can be easily aggregated via readout functions to derive graph-level or system-level property κ, such as the electrolyte conductivity.

Symmetries Symmetries are fundamental in the physical world and should be satisfied when modeling interactions between physical systems. In mixture systems, symmetries are more complicated than those in single-instance systems. To be specific, the symmetries of the function φ in Eq. (1) are divided into two parts: the *node-level* and the *graph-level* symmetry constraints.

For node-level symmetries, the output should be permutation-equivariant within each graph, namely,

$$\varphi\colon (\{\mathcal{G}_m(\boldsymbol{H}_m\boldsymbol{P}_m, \vec{\boldsymbol{X}}_m\boldsymbol{P}_m, \vec{\boldsymbol{V}}_m\boldsymbol{P}_m, w_m)\}_{m=1}^{M}, \boldsymbol{c}) \mapsto (\{\boldsymbol{H}'_m\boldsymbol{P}_m, \vec{\boldsymbol{V}}'_m\boldsymbol{P}_m\}_{m=1}^{M}, \kappa), \tag{2}$$

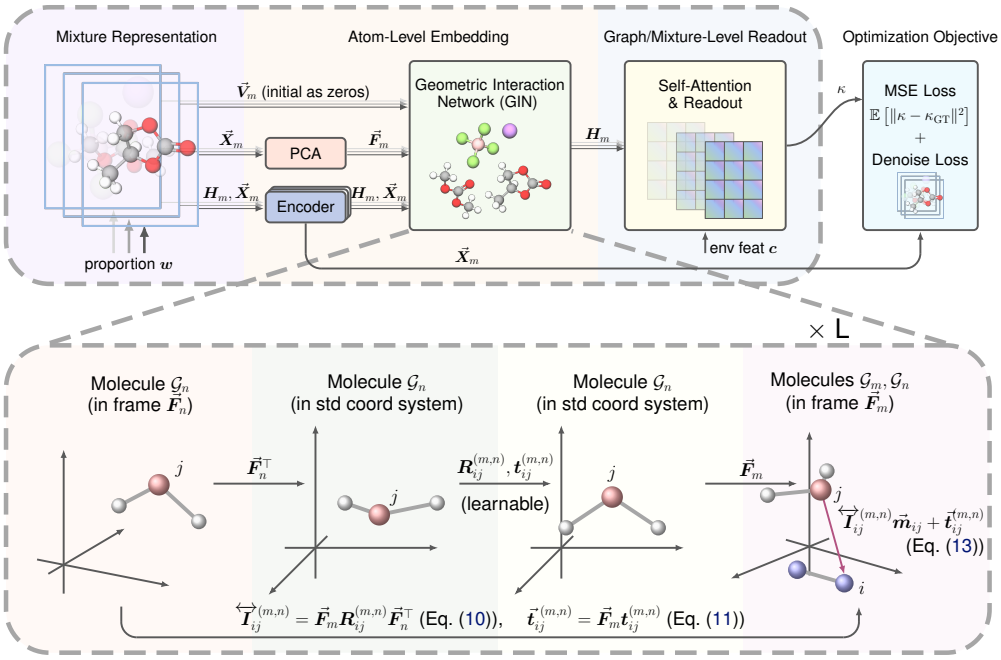

Figure 1: Overview of our GeoMix. By taking a set of molecules $\{\mathcal{G}_m\}_{m=1}^M$ with proportion $\boldsymbol{w} = [w_m]_M$ as input, it first constructs local frames $\vec{\boldsymbol{F}}_m$ via PCA, then applies an equivariant encoder to update $\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m$. Intermolecular message passing is performed via GIN, which learns transformations $\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)}$ and $\vec{\boldsymbol{t}}_{ij}^{(m,n)}$, which enable equivariant message passing across molecules. The scalar features $\boldsymbol{H}_m$ are finally aggregated along with the environment descriptor $\boldsymbol{c}$ for prediction.

where $\boldsymbol{P}_m \in \mathbb{R}^{N_m \times N_m}$ is a permutation matrix acting all nodes in graph $\mathcal{G}_m$. Besides, the output should also be rotation-equivariant on each graph, *i.e.*,

$$\varphi\colon (\{\mathcal{G}_m(\boldsymbol{H}_m, \boldsymbol{R}_m\vec{\boldsymbol{X}}_m, \boldsymbol{R}_m\vec{\boldsymbol{V}}_m, w_m)\}_{m=1}^M, \boldsymbol{c}) \mapsto (\{\boldsymbol{H}'_m, \boldsymbol{R}_m\vec{\boldsymbol{V}}'_m\}_{m=1}^M, \kappa), \tag{3}$$

where $\boldsymbol{R}_m \in \mathbb{R}^{3\times 3}$ is a rotation matrix acting on node coordinates and geometric features in $\mathcal{G}_m$. Translation equivariance is naturally achieved by centering each graph, and is therefore omitted.

For graph-level symmetries, the output should be equivariant with respect to the order of the graphs:

$$\varphi\colon (\{\mathcal{G}_{\sigma(m)}\}_{m=1}^M, \boldsymbol{c}) \mapsto (\{\boldsymbol{H}'_{\sigma(m)}, \vec{\boldsymbol{V}}'_{\sigma(m)}\}_{m=1}^M, \kappa), \tag{4}$$

where σ denotes any permutation acting on all graphs $\{\mathcal{G}_m\}_{m=1}^M$.

Note that the output property κ is invariant to any transformation. We call the two-level symmetries as **Set-SE(3) Equivariance** in this paper.

3.2 Model Architecture

As illustrated in Fig. 1, we design a general equivariant architecture that enables the update of both invariant and equivariant features through intramolecular and intermolecular message passing.

Proportion Embedding Each molecule in the mixture system is associated with a concentration value w_m. Prior studies [13; 50] often treat each molecule as having a fixed representation regardless of its concentration. In chemical thermodynamics, however, the activity of each molecule depends nonlinearly on its mole fraction and the activity coefficient, reflecting coupling between concentration and intermolecular forces [8]. To capture such context dependence, we concatenate each normalized proportion w_i directly into atomic-level features via feature concatenation.

Intramolecular Encoding We employ either EGNN [54] or TFN [55] as equivariant encoders to characterize molecular geometries through message passing within individual molecules. The same encoder with shared parameters is used for all solvent molecules, while a separate encoder is

employed for salts due to their distinct structural properties (organic salts versus inorganic solvents). In this stage, we directly update node coordinates $\vec{\boldsymbol{X}}_m$ by treating them as geometric features $\vec{\boldsymbol{V}}_m$.

Frame Construction The rotation equivariance described in Eq. (3) arises primarily because each molecule is represented in an independent local coordinate system. This inherent symmetry prevents direct transfer of coordinate information between different molecules. However, if we are able to derive the local frame of each molecule, we can conduct geometric message passing between different molecules by first transforming coordinate information into standard coordinate space, as depicted in Fig. 1. Here, we achieve this purpose based on Principal Component Analysis (PCA)[3]. In particular, we first compute the covariance $\mathrm{Cov}(\vec{\boldsymbol{X}}_m) = \frac{1}{N_m}\vec{\boldsymbol{X}}_m\vec{\boldsymbol{X}}_m^\top$, and then perform eigendecomposition $\mathrm{Cov}(\vec{\boldsymbol{X}}_m) = \vec{\boldsymbol{F}}_m\Lambda_m\vec{\boldsymbol{F}}_m^\top$ to obtain its orthogonal eigenvectors $\vec{\boldsymbol{F}}_m$ to form a local frame. If $\det(\vec{\boldsymbol{F}}_m) = -1$, we flip the third axis to ensure a right-handed coordinate system.

Intermolecular Interaction With the pre-constructed local frames $\{\vec{\boldsymbol{F}}_m\}_{m=1}^M$, we design a novel component GIN to simulate geometric interactions between different molecules. The details of GIN will be presented in § 3.3. In this module, the coordinates $\vec{\boldsymbol{X}}_m$ are fixed, the geometric features $\vec{\boldsymbol{V}}_m$ are initialized as zeros and will be updated throughout GIN.

Readout and Prediction Head The outputs of GIN are node-level features. For mixture property prediction tasks such as electrolyte conductivity prediction, we first average the node-level features and read them out as graph-level features, which are then modeled by a self-attention encoder without position encoding. The system-level feature is obtained by averaging these graph-level features, and finally, we concatenate with the global environment feature and pass the result through an MLP with an activation function to generate the prediction. These operations can be formulated as:

$$\boldsymbol{h}_m = \tfrac{1}{N_m}\mathbf{1}^\top\boldsymbol{H}_m, \quad y = \texttt{MLP}\left(\tfrac{1}{M}\mathbf{1}^\top\texttt{SelfAtt}\left(\bigoplus_{m=1}^M \boldsymbol{h}_m\right)\oplus\boldsymbol{c}\right), \tag{5}$$

where $\oplus$ represents the tensor concatenate operation, $y \in \mathbb{R}$ is the predicted conductivity.

Optimization Objective For invariant prediction tasks, we use Mean Squared Error (MSE) as the loss function. To take full advantage of the equivariant network, we also adopt the Noisy Nodes method [57]. Noisy Nodes method encourages the model to learn the configurations with lower energy, effectively regularizing the learning process. Specifically, our model additionally learns a denoising process of the encoder:

$$\{\mathcal{G}(\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m^{\text{noisy}})\}_{m=1}^M \mapsto \{\mathcal{G}(\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m^{\text{denoise}})\}_{m=1}^M, \tag{6}$$

where $\vec{\boldsymbol{X}}_m^{\text{noisy}} = \vec{\boldsymbol{X}}_m + \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$, and $\sigma = 0.3$ is taken for general molecular-level tasks. The auxiliary loss is calculated as:

$$\mathcal{L}_{\text{denoise}} = \textstyle\sum_{m=1}^M \|\vec{\boldsymbol{X}}_m^{\text{denoise}} - \vec{\boldsymbol{X}}_m\|_{\text{F}}, \tag{7}$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm. The optimization objective is the sum of two losses given by:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \gamma\cdot\mathcal{L}_{\text{denoise}}. \tag{8}$$

The implementation details can be found in Appendix B.

3.3 Geometric Interaction Network

In this section, we introduce GIN, a novel equivariant message-passing module to facilitate the geometric message passing between molecules using coordinate transformations, while promisingly ensuring Set-$\mathrm{SO}(3)$ equivariance.

Architecture Overview GIN takes as input a set of geometric graphs $\{\mathcal{G}_m(\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m, \vec{\boldsymbol{V}}_m)\}_{m=1}^M$ along with the corresponding local frames $\{\vec{\boldsymbol{F}}_m\}_{m=1}^M$. In each layer, the node features $\{\boldsymbol{H}_m\}_{m=1}^M$ and the geometry features $\{\vec{\boldsymbol{V}}_m\}_{m=1}^M$ are updated, while the coordinates $\{\vec{\boldsymbol{X}}_m\}_{m=1}^M$ are unchanged. To model the interaction from the source molecule $\mathcal{G}_m$ and the target one $\mathcal{G}_n$, GIN consists of three key components. 1) **Intermolecular Transformation**: Establishing a transformation within a

[3]Since frame construction via traditional PCA is not strictly equivariant owing to the non-unique axis-orientation [56], we use an improved, equivariance-preserving PCA variant (see Appendix B.8).

standard coordinate system based on the features of both the source and target molecules. 2) **Message Construction**: Computing the message in the source molecule's coordinate frame using the learned transformation, then converting it to the target molecule's frame. 3) **Aggregation and Update**: Aggregating the messages to update the node-level embeddings in the target molecule. For brevity, in our following formulation, $i \in \mathcal{G}_m$ consistently denotes the atom receiving message, while $j \in \mathcal{G}_n$ represents the atom sending message.

Intermolecular Transformation Since the two molecules exist in different reference frames, the message passing strategy used in traditional equivariant GNNs is no longer suitable. Consequently, we aim to design a transformation that facilitates equivariant message passing across distinct reference frames. For atoms $i \in \mathcal{G}_m$ and $j \in \mathcal{G}_n$, we first calculate a scalar $\boldsymbol{z}_{ij}^{(m,n)}$ to combine the invariant features in both atoms as follows:

$$\boldsymbol{z}_{ij}^{(m,n)} = \sigma_{\texttt{inv}}(\boldsymbol{h}_i^{(m)}, \boldsymbol{h}_j^{(n)}, \|\vec{\boldsymbol{x}}_i^{(m)}\|, \|\vec{\boldsymbol{x}}_j^{(n)}\|, \|\vec{\boldsymbol{v}}_i^{(m)}\|, \|\vec{\boldsymbol{v}}_j^{(n)}\|), \tag{9}$$

where $\sigma_{\texttt{inv}}$ is an MLP, and subsequent σ functions with different subscripts represent distinct MLPs. Based on $\boldsymbol{z}_{ij}^{(m,n)}$ and the frames $\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n$, we introduce a learnable matrix $\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)} \in \mathbb{R}^{3\times 3}$ to "rotate" the equivariant message from $j \in \mathcal{G}_n$ into the reference frame of $i \in \mathcal{G}_m$:

$$\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)} \coloneqq \vec{\boldsymbol{F}}_m \boldsymbol{R}_{ij}^{(m,n)} \vec{\boldsymbol{F}}_n^\top, \quad \boldsymbol{R}_{ij}^{(m,n)} = \sigma_{\texttt{rot}}(\boldsymbol{z}_{ij}^{(m,n)}) \in \mathbb{R}^{3\times 3}. \tag{10}$$

Similarly, we also design a learnable vector $\vec{\boldsymbol{t}}_{ij}^{(m,n)}$ to model the "translation":

$$\vec{\boldsymbol{t}}_{ij}^{(m,n)} \coloneqq \vec{\boldsymbol{F}}_m \boldsymbol{t}_{ij}^{(m,n)}, \quad \boldsymbol{t}_{ij}^{(m,n)} = \sigma_t(\boldsymbol{z}_{ij}^{(m,n)}) \in \mathbb{R}^3. \tag{11}$$

It is worth noting that the matrix corresponding to $\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)}$ does not represent a strict rotation, as the determinant of $\boldsymbol{R}_{ij}^{(m,n)}$ is not constrained to be 1. We also explore alternative constructions of $\boldsymbol{R}_{ij}^{(m,n)}$ with a unit determinant, by using quaternion [58] and 6D vector [59]. However, ablation studies (see § 4.4) reveal that these alternatives yield less favorable results. In fact, we can give the following theorem to explain the expressive power of Eq. (10).

Theorem 3.1 (Expressivity of Intermolecular Transformation Matrix)**.** *Given geometric graphs $\mathcal{G}_m$ and $\mathcal{G}_n$, with $\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n$ being respective frames, any matrix $\overleftrightarrow{\boldsymbol{I}}_{mn}$ satisfying* $\mathrm{SO}(3)$*-equivariance $\overleftrightarrow{\boldsymbol{I}}_{mn} \xmapsto{\boldsymbol{R}_m, \boldsymbol{R}_n \in \mathrm{SO}(3)} \boldsymbol{R}_m \overleftrightarrow{\boldsymbol{I}}_{mn} \boldsymbol{R}_n^\top$ must take the form like Eq. (10) as $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n) = \vec{\boldsymbol{F}}_m \boldsymbol{R} \vec{\boldsymbol{F}}_n^\top$.*

Message Construction The invariant message $\boldsymbol{m}_{ij}^{(m,n)}$ with geometric interaction is similar to the construction of $\boldsymbol{z}_{ij}^{(m,n)}$ in Eq. (9). The difference is that in order to align the invariant message and the equivariant message, we additionally introduce the modulus of the translation vector $\vec{\boldsymbol{t}}_{ij}^{(m,n)}$ in Eq. (11). The entire formula is given by:

$$\boldsymbol{m}_{ij}^{(m,n)} = \sigma_{\texttt{msg}}(\boldsymbol{h}_i^{(m)}, \boldsymbol{h}_j^{(n)}, \|\vec{\boldsymbol{x}}_i^{(m)}\|, \|\vec{\boldsymbol{x}}_j^{(n)}\|, \|\vec{\boldsymbol{v}}_i^{(m)}\|, \|\vec{\boldsymbol{v}}_j^{(n)}\|, \|\vec{\boldsymbol{t}}_{ij}^{(m,n)}\|). \tag{12}$$

We now compute cross-reference frame equivariant messages by two steps. First, the atom coordinate $\vec{\boldsymbol{x}}_j^{(n)}$ and the geometric feature $\vec{\boldsymbol{v}}_j^{(n)}$ are combined to construct the message $\boldsymbol{m}_{ij}^{(m,n)}$ in the reference frame of $\mathcal{G}_n$. Second, this message is transformed into the reference frame of $\mathcal{G}_m$ by applying the learnable transformation, *i.e.* the "rotation matrix" $\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)}$ and the "translation vector" $\vec{\boldsymbol{t}}_{ij}^{(m,n)}$, resulting in the updated message $\boldsymbol{m}_{ij}^{(m,n)}$ in the $\mathcal{G}_m$ frame:

$$\vec{\boldsymbol{m}}_{ij}^{(m)} = \overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)} \vec{\boldsymbol{m}}_{ij}^{(m,n)} + \vec{\boldsymbol{t}}_{ij}^{(m,n)}, \quad \vec{\boldsymbol{m}}_{ij}^{(m,n)} = \sigma_{\vec{\boldsymbol{x}}}(\boldsymbol{m}_{ij}^{(m,n)}) \vec{\boldsymbol{x}}_j^{(n)} + \sigma_{\vec{\boldsymbol{v}}}(\boldsymbol{m}_{ij}^{(m,n)}) \vec{\boldsymbol{v}}_j^{(n)}. \tag{13}$$

Aggregation and Update After devising the message, we proceed to aggregate them and update the atom features. Each atom is required to receive message from all atoms in every other molecule. To this end, we employ the average operator to aggregate both invariant and equivariant message:

$$\boldsymbol{m}_i^{(m)} = \frac{1}{N - N_m} \sum_{n \neq m} \sum_{j=1}^{N_n} \boldsymbol{m}_{ij}^{(m,n)}, \quad \vec{\boldsymbol{m}}_i^{(m)} = \frac{1}{N - N_m} \sum_{n \neq m} \sum_{j=1}^{N_n} \vec{\boldsymbol{m}}_{ij}^{(m)}, \tag{14}$$

where $N = \sum_{m=1}^{M} N_m$ denotes the number of all atoms in the system. Finally, we update the invariant feature $\boldsymbol{h}_i^{(m)}$ and equivariant feature $\vec{\boldsymbol{v}}_i^{(m)}$ as follows:

$$\boldsymbol{h}_i^{(m)} = \sigma_{\boldsymbol{h}}(\boldsymbol{h}_i^{(m)}, \boldsymbol{m}_i^{(m)}), \quad \vec{\boldsymbol{v}}_i^{(m)} = \vec{\boldsymbol{v}}_i^{(m)} + \vec{\boldsymbol{m}}_i^{(m)}. \tag{15}$$

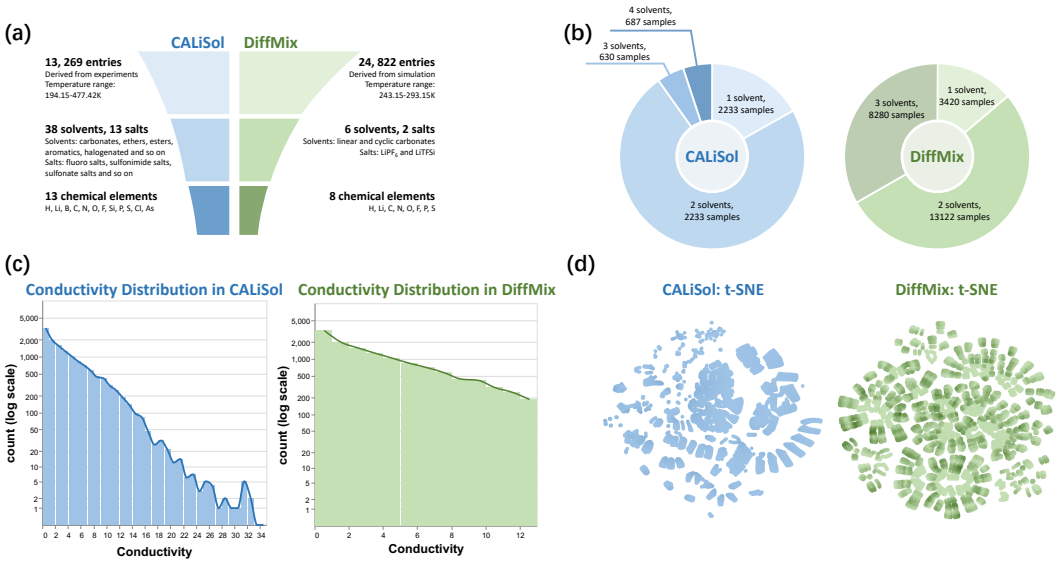

Figure 2: Overview of the CALiSol and DiffMix datasets. (a) Summary of dataset statistics. (b) Composition diversity in terms of the number of solvents used per sample. CALiSol exhibits a more diverse mixture design space, while DiffMix contains a greater number of samples with 2 or 3 solvents. (c) Log-scaled histograms of conductivity values indicate that CALiSol exhibits a broader range and a more pronounced long-tail distribution than DiffMix dataset. (d) The t-SNE visualization [60] reveals distinct structural characteristics in each dataset. CALiSol, derived from experimental measurements, exhibits greater diversity with many scattered outliers, reflecting heterogeneous sampling. In contrast, DiffMix shows a more uniform and clustered distribution, consistent with its simulation-based origin.

4 Experiments

In this section, we first introduce the two large-scale electrolyte datasets, CALiSol and DiffMix, covering their composition, property distributions, and geometric graph construction in § 4.1, and provide corresponding visualizations in Fig. 2. We then describe our experimental setup, including baseline models, GeoMix implementation, and evaluation metrics in § 4.2. The main results in § 4.3 highlight GeoMix's superior performance, and ablation studies in § 4.4 assess the impact of key components. Further details are provided in Appendix B.

4.1 Datasets

CALiSol [19] and DiffMix [11] Datasets These two datasets provide comprehensive resources for electrolyte conductivity analysis. CALiSol contains 13,269 experimentally measured ionic conductivity entries (after filtering 556 incomplete points) spanning 13 lithium salts and 38 solvents across carbonate, ether, and ester classes, with broad coverage of salt concentrations and temperatures to reflect experimental diversity. In contrast, the simulation-driven DiffMix dataset employs systematic discretization with 24,822 formulations focusing on two salts ($LiPF_6$, LiTFSI) and up to three solvents selected from six carbonate types, quantizing salt concentrations into seven levels and normalizing solvent mass fractions via coarse sampling. Both datasets provide metadata including solvent proportions and measurement conditions, with full parameter ranges detailed in Appendix B.

Geometric Graph Construction To facilitate geometric modeling, we further augmented the dataset with geometric molecular graph representations. We obtained molecular conformations from PubChem [61] when available. For molecules not found in PubChem, we generate 3D structures using RDKit [62] or download conformations from the Materials Project [63]. Each molecule is represented as a geometric graph, where atoms are nodes and edges are formed between atoms within a cutoff distance of 6 Å. Edge weights correspond to the Euclidean distances between atoms. Node features include atomic numbers and one-hot encodings of atom types.

Data Splitting Strategy For both CALiSol and DiffMix datasets, we adopt a standard random split of 70%/20%/10% into train, validation, and test sets, respectively. Splitting is performed

Table 1: Results of MSE and Pearson correlation coefficient on CALiSol dataset and DiffMix dataset. Bold values indicate the best performance, while underlined values indicate the second-best.

	Models	Symmetries	CALiSol		DiffMix	
			MSE ↓	Pearson r ↑	MSE ↓	Pearson r ↑
MLP-based	MLP	None	3.657	0.906	1.363	0.874
	MM-MoLFormer [12]		5.488	0.825	1.901	0.812
Topology-based	MolSets-Conv [13]	Permutation	2.230	0.924	1.440	0.868
	MolSets-SAGE [13]		2.751	0.909	0.708	0.937
Geometry-based	EGNN-att [54]	Set-SE(3)	2.666	0.908	0.752	0.930
	TFN-att [55]		1.808	0.946	0.804	0.921
	EGNN-linear [54]		1.461	0.951	0.195	0.988
	TFN-linear [55]		1.107	0.967	0.285	0.973
Geometry-based	GeoMix-EGNN	Set-SE(3)	<u>0.552</u>	<u>0.985</u>	<u>0.088</u>	<u>0.992</u>
	GeoMix-TFN		**0.432**	**0.987**	**0.035**	**0.997**

independently for CALiSol and DiffMix with a fixed seed to ensure reproducibility. No electrolyte formulation is shared across different subsets.

4.2 Experiment Setup

Baselines Previous studies such as MM-MoLFormer [12], MolSets [13], and Uni-ELF [18] have proposed individual models and conducted comparisons within limited baselines, but no prior work has provided a unified evaluation across modeling families using standardized datasets and metrics.

We compare with diverse baselines across architectural families: 1) **MLP-based**: We use a vanilla MLP taking the proportion vector as input, and MM-MoLFormer [12], which combines pretrained molecular embeddings with proportion and environment features. These models do not respect the symmetry constraints discussed in § 3.1 and exhibit limited generalization. 2) **Topology-based**: We adopt MolSets [13], a recent framework designed for molecular mixture modeling, and select GraphConv and GraphSAGE as representative backbones (denoted as MolSets-Conv and MolSets-SAGE). 3) **Geometry-based**: To account for spatial effects such as steric hindrance and solvation structure, we employ equivariant GNNs including EGNN [54] and TFN [55]. Each is paired with either a composition-weighted average (-linear) or a learnable attention-based aggregator (-att), yielding four geometric baselines: EGNN-att, TFN-att, EGNN-linear, and TFN-linear. The specific hyperparameters of the model are shown in Appendix B. Other recent geometry-based methods, such as DiffMix [11] and Uni-ELF [18], are not included in our baselines due to the lack of publicly available implementations.

Implementation For our GeoMix, we test it using two classical equivariant networks, EGNN and TFN, as the backbone for intramolecular embedding. Our method is consistent with the baseline in terms of hyperparameters such as the number of equivariant graph network layers and MLP layers. For other hyperparameters and other implementation details, please refer to Appendix B. We evaluate model performance using MSE and Pearson correlation coefficient (r) between the predicted and measured conductivity values.

4.3 Main Results

Table 1 summarizes the predictive performance of various baseline models and our proposed GeoMix on the CALiSol and DiffMix datasets. From these results, we observe the following: **1. Overall performance**: GeoMix consistently achieves the best performance across all metrics on both datasets, demonstrating its effectiveness in modeling molecular mixtures. **2. Advantage of geometry-based methods**: Geometry-based methods outperform both MLP-based and standard Topology-based baselines, highlighting the importance of incorporating molecular geometry into the learning process. **3. Benefit of higher-degree features**: Among geometric models, those using TFN as the backbone generally surpass those based on EGNN, suggesting that modeling higher-degree geometric features is particularly beneficial in mixture-related tasks. We further visualize the regression performance of some selected models in Fig. 3, which clearly illustrates the superiority of GeoMix-based methods in both accuracy and error consistency.

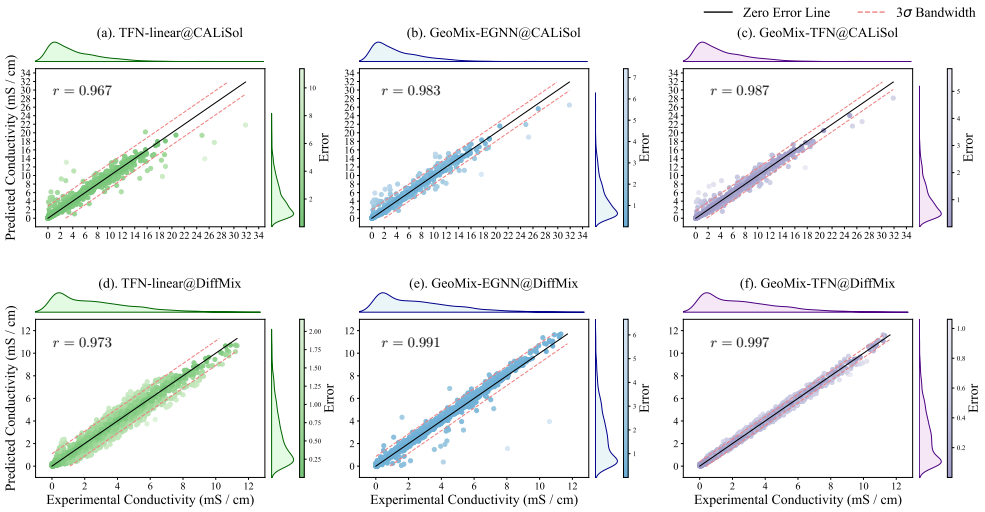

Figure 3: Regression plots for electrolyte conductivity prediction. (a–c) show results on the CALiSol dataset, using TFN-linear, GeoMix-EGNN, and GeoMix-TFN, respectively. (d–f) show corresponding results on the DiffMix dataset with the same model order.

4.4 Ablation Studies

To evaluate the individual contributions of our proposed components, we perform a series of ablation studies on the CALiSol dataset. All experiments are based on the GeoMix-EGNN model, which adopts the EGNN backbone due to its architectural simplicity and suitability for controlled analysis.

Proportion Embedding In the ablation variant denoted as "Multiply" in Table 2, we remove the proportion information from atomic features and instead apply scalar weighting to each per-molecule graph embedding before aggregation. This strategy mirrors traditional approaches that treat component ratios as external mixing coefficients rather than intrinsic modulators of molecular behavior. Table 2 reveals that embedding proportions into node features can effectively improve the effect of the model in modeling mixture systems.

Table 2: Ablations on CALiSol dataset.

CALiSol	MSE ↓	Pearson r ↑
GeoMix	0.552	0.985
Proportion Embedding		
Multiply	3.657	0.906
Transformation Matrix's Form		
Quaternion	0.702	0.981
6D vector	0.574	0.983
Graph-wise	0.662	0.980
Linear v.s. Attention		
GeoMix-linear	0.851	0.975
Noisy Nodes Loss		
w/o Noisy Nodes	1.213	0.969

The Form of the Transformation Matrix We conduct ablation studies to examine how different parameterizations of the transformation matrix $\boldsymbol{R}_{ij}^{(m,n)}$ affect model performance. We consider three alternative designs: 1) **Quaternion**: The rotation matrix is derived from a learned unit quaternion, ensuring orthogonality with $\det(\boldsymbol{R}_{ij}^{(m,n)}) = 1$ (*i.e.*, rigid transformation without scaling). 2) **6D Vector**: The rotation is parameterized by two learned vectors followed by Gram-Schmidt orthogonalization, providing a continuous and differentiable representation of SO(3) [59], also enforcing $\det(\boldsymbol{R}_{ij}^{(m,n)}) = 1$. 3) **Graph-wise**: Instead of learning node-pair-specific transformations, we learn a single transformation per graph pair by pooling node features. This enforces graph-level rigidity and ignores local structural flexibility. From the results in Table 2, we observe that: (1) Models with non-rigid, node-wise transformations (GeoMix, Quaternion, 6D vector) outperform the graph-rigid variant (Graph-wise), suggesting the importance of fine-grained local transformations. (2) Among non-rigid variants, GeoMix, which does not constrain the transformation matrix to be orthogonal or have unit determinant, achieves the best performance. This indicates that allowing scaling transformations further enhances flexibility and expressivity in modeling cross-graph geometry.

Graph-Level Readout As shown in Table 1, the simple linear average readout performs competitively among baseline models. To investigate its effect within our framework, we replace the original graph-level readout in GeoMix with the linear average strategy, resulting in an ablation

variant denoted as GeoMix-linear. As reported in Table 2, GeoMix-linear yields lower performance compared to the original GeoMix. This suggests that the full model benefits from the rich, attention-based aggregation mechanism, which we attribute to the GIN. The learned transformation-aware interactions provide expressive relational cues between molecules, enabling the self-attention module to generate more informative graph-level embeddings beyond simple averaging.

Noisy Nodes Loss We further evaluate the impact of the Noisy Nodes loss [57] by removing it from the training objective, denoted as "*w/o* Noisy Nodes" in Table 2. The results demonstrate that incorporating this regularization consistently improves the model's performance. We attribute this to its role in enhancing representation robustness by perturbing node features during training.

5 Conclusion

We presented GeoMix, a Set-SE(3) equivariant framework for predicting properties of molecular mixtures. GeoMix leverages equivariant graph neural networks to embed individual species and introduces a Geometric Interaction Network (GIN) to perform intermolecular message passing via pre-constructing local frames. To support systematic evaluation, we curated two large-scale datasets with geometric features, CALiSol and DiffMix, focused on electrolyte conductivity. Based on these datasets, we established a unified benchmark for electrolyte conductivity prediction. Experimental results demonstrate that GeoMix significantly outperforms existing baselines, highlighting the importance of geometry-aware modeling in mixture systems. Future work may extend this framework to other multi-component tasks in chemistry and materials science.

6 Limitations

Despite the promising results, this work has several limitations that warrant further investigation. First, the available datasets remain limited in both scope and diversity. In particular, no datasets currently provide molecular dynamics-based microscopic structural information, and there is a lack of conductivity datasets covering a broader range of chemistries, such as sodium-ion or potassium-ion electrolytes. Second, the encoder architecture adopted in GeoMix is relatively simple, leaving room for improvement.

In future work, we plan to incorporate richer descriptors and leverage recent advances in graph neural networks to enhance both the accuracy and the generalization ability of the model [64; 65]. We also aim to curate and release new benchmark datasets that include microscopic structural features and broader compositional coverage, thereby enabling more comprehensive evaluation and fostering further research in mixture property prediction.

Acknowledgement

This work was jointly supported by the following projects: the National Natural Science Foundation of China (No. 62376276); Beijing Nova Program (No. 20230484278); the Fundamental Research Funds for the Central Universities; the Research Funds of Renmin University of China (23XNKJ19); Public Computing Cloud, Renmin University of China.

References

[1] Yann L Müller and Anirudh Raju Natarajan. Constructing multicomponent cluster expansions with machine-learning and chemical embedding. *npj Computational Materials*, 11(1):60, 2025.

[2] Chen Shen, Siamak Attarian, Yixuan Zhang, Hongbin Zhang, Mark Asta, Izabela Szlufarska, and Dane Morgan. Supersalt: Equivariant neural network force fields for multicomponent molten salts system. *arXiv preprint arXiv:2412.19353*, 2024.

[3] Pin-Wen Guan. Differentiable thermodynamic modeling. *Scripta Materialia*, 207:114217, 2022.

[4] Fabian Jirasek, Robert Bamler, and Stephan Mandt. Hybridizing physical and data-driven prediction methods for physicochemical properties. *Chemical Communications*, 56(82):12407–12410, 2020.

[5] Natalia V Plechkova and Kenneth R Seddon. Applications of ionic liquids in the chemical industry. *Chemical Society Reviews*, 37(1):123–150, 2008.

[6] Chen Ling. A review of the recent progress in battery informatics. *npj Computational Materials*, 8(1):33, 2022.

[7] Max Planck. *Vorlesungen uber thermodynamik*. Leipzig, 1930.

[8] Carl H. Hamann, Andrew Hamnett, and Wolf Vielstich. *Electrochemistry*. Wiley, 2007.

[9] Yumin Zhang, Imanuel Bier, and Venkatasubramanian Viswanathan. Predicting electrolyte conductivity directly from molecular-level interactions. *ACS Energy Letters*, 7(11):4061–4070, 2022.

[10] Sheng Gong, Yumin Zhang, Zhenliang Mu, Zhichen Pu, Hongyi Wang, Xu Han, Zhiao Yu, Mengyi Chen, Tianze Zheng, Zhi Wang, et al. A predictive machine learning force-field framework for liquid electrolyte development. *Nature Machine Intelligence*, 7:543–552, 2025.

[11] Shang Zhu, Bharath Ramsundar, Emil Annevelink, Hongyi Lin, Adarsh Dave, Pin-Wen Guan, Kevin Gering, and Venkatasubramanian Viswanathan. Differentiable modeling and optimization of non-aqueous li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15:8649, 2024.

[12] Eduardo Soares, Vidushi Sharma, Emilio Vital Brazil, Renato Cerqueira, and Young-Hye Na. Capturing formulation design of battery electrolytes with chemical large language model. In *AI for Accelerated Materials Design-NeurIPS Workshop*, 2023.

[13] Hengrui Zhang, Tianxing Lai, Jie Chen, Arumugam Manthiram, James M Rondinelli, and Wei Chen. Learning molecular mixture property using chemistry-aware graph neural network. *PRX Energy*, 3(2):023006, 2024.

[14] Wenbing Huang and Jiacheng Cen. Geometric graph learning for drug design. *Deep Learning in Drug Design*, pages 133–151, 2026.

[15] Jiaqi Han, Jiacheng Cen, Liming Wu, Zongzhao Li, Xiangzhe Kong, Rui Jiao, Ziyang Yu, Tingyang Xu, Fandi Wu, Zihe Wang, et al. A survey of geometric graph neural networks: Data structures, models and applications. *Frontiers of Computer Science*, 19(11):1911375, 2025.

[16] Yuelin Zhang, Jiacheng Cen, Jiaqi Han, Zhiqiang Zhang, Jun Zhou, and Wenbing Huang. Improving equivariant graph neural networks on large geometric graphs via virtual nodes learning. In *Forty-first International Conference on Machine Learning*, 2024.

[17] Yuelin Zhang, Jiacheng Cen, Jiaqi Han, and Wenbing Huang. Fast and distributed equivariant graph neural networks by virtual node learning. *arXiv preprint arXiv:2506.19482*, 2025.

[18] Boshen Zeng, Sian Chen, Xinxin Liu, Changhong Chen, Bin Deng, Xiaoxu Wang, Zhifeng Gao, Yuzhi Zhang, Linfeng Zhang, et al. Uni-elf: A multi-level representation learning framework for electrolyte formulation design. *arXiv preprint arXiv:2407.06152*, 2024.

[19] Paolo de Blasio, Jonas Elsborg, Tejs Vegge, Eibar Flores, and Arghya Bhowmik. Calisol-23: Experimental electrolyte conductivity data for various li-salts and solvent combinations. *Scientific Data*, 11(1):750, 2024.

[20] Jiacheng Cen, Anyi Li, Ning Lin, Tingyang Xu, Yu Rong, Deli Zhao, Zihe Wang, and Wenbing Huang. Universally invariant learning in equivariant gnns. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Chaohao Yuan, Kangfei Zhao, Ercan Engin Kuruoglu, Liang Wang, Tingyang Xu, Wenbing Huang, Deli Zhao, Hong Cheng, and Yu Rong. A survey of graph transformers: Architectures, theories and applications. *arXiv preprint arXiv:2502.16533*, 2025.

[22] Zian Li, Xiyuan Wang, Yinan Huang, and Muhan Zhang. Is distance matrix enough for geometric deep learning? *Advances in Neural Information Processing Systems*, 36:37413–37447, 2023.

[23] Zian Li, Xiyuan Wang, Shijia Kang, and Muhan Zhang. On the completeness of invariant geometric deep learning models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. In *Forty-second International Conference on Machine Learning*, 2025.

[25] Shaoheng Yan, Zian Li, and Muhan Zhang. Georecon: Graph-level representation learning for 3d molecules via reconstruction-based pretraining. *arXiv preprint arXiv:2506.13174*, 2025.

[26] Zongzhao Li, Jiacheng Cen, Bing Su, Tingyang Xu, Yu Rong, Deli Zhao, and Wenbing Huang. Large language-geometry model: When LLM meets equivariance. In *Forty-second International Conference on Machine Learning*, 2025.

[27] Zongzhao Li, Jiacheng Cen, Wenbing Huang, Taifeng Wang, and Le Song. Size-generalizable rna structure evaluation by exploring hierarchical geometries. In *The Thirteenth International Conference on Learning Representations*, 2025.

[28] Fanmeng Wang, Hongteng Xu, Xi Chen, Shuqi Lu, Yuqing Deng, and Wenbing Huang. Mperformer: An se (3) transformer-based molecular perceptron. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2512–2522, 2023.

[29] Fanmeng Wang, Wentao Guo, Minjie Cheng, Shen Yuan, Hongteng Xu, and Zhifeng Gao. Mmpolymer: A multimodal multitask pretraining framework for polymer property prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, 2024.

[30] Fanmeng Wang, Wentao Guo, Qi Ou, Hongshuai Wang, Haitao Lin, Hongteng Xu, and Zhifeng Gao. Polyconf: Unlocking polymer conformation generation through hierarchical generative models. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Fanmeng Wang, Minjie Cheng, and Hongteng Xu. WGFormer: An SE(3)-transformer driven by wasserstein gradient flows for molecular ground-state conformation prediction. In *Forty-second International Conference on Machine Learning*, 2025.

[32] Angxiao Yue, Dixin Luo, and Hongteng Xu. A plug-and-play quaternion message-passing module for molecular conformation representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16633–16641, 2024.

[33] Angxiao Yue, Zichong Wang, and Hongteng Xu. Reqflow: Rectified quaternion flow for efficient and high-quality protein backbone generation. In *Forty-second International Conference on Machine Learning*, 2025.

[34] Yurou Liu, Jiahao Chen, Rui Jiao, Jiangmeng Li, Wenbing Huang, and Bing Su. DenoiseVAE: Learning molecule-adaptive noise distributions for denoising-based 3d molecular pre-training. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Fanglei Xue, Meihan Zhang, Shuqi Li, Xinyu Gao, James A Wohlschlegel, Wenbing Huang, Yi Yang, and Weixian Deng. Se (3)-equivariant ternary complex prediction towards target protein degradation. *Nature Communications*, 16(1):5514, 2025.

[36] Liming Wu, Zhichao Hou, Jirui Yuan, Yu Rong, and Wenbing Huang. Equivariant spatio-temporal attentive graph networks to simulate physical dynamics. *Advances in Neural Information Processing Systems*, 36, 2023.

[37] Liming Wu, Wenbing Huang, Rui Jiao, Jianxing Huang, Liwei Liu, Yipeng Zhou, Hao Sun, Yang Liu, Fuchun Sun, Yuxiang Ren, et al. Siamese foundation models for crystal structure prediction. *arXiv preprint arXiv:2503.10471*, 2025.

[38] Qi Li, Rui Jiao, Liming Wu, Tiannian Zhu, Wenbing Huang, Shifeng Jin, Yang Liu, Hongming Weng, and Xiaolong Chen. Powder diffraction crystal structure determination using generative models. *arXiv preprint arXiv:2409.04727*, 2024.

[39] Haitao Lin, Odin Zhang, Jia Xu, Yunfan Liu, Zheng Cheng, Lirong Wu, Yufei Huang, Zhifeng Gao, and Stan Z Li. Tokenizing electron cloud in protein-ligand interaction learning. *arXiv preprint arXiv:2505.19014*, 2025.

[40] Yang Liu, Zinan Zheng, Yu Rong, and Jia Li. Equivariant graph learning for high-density crowd trajectories modeling. *Transactions on Machine Learning Research*, 2024.

[41] Zinan Zheng, Yang Liu, Jia Li, Jianhua Yao, and Yu Rong. Relaxing continuous constraints of equivariant graph neural networks for broad physical dynamics learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4548–4558, 2024.

[42] Yang Liu, Zinan Zheng, Yu Rong, Deli Zhao, Hong Cheng, and Jia Li. Equivariant and invariant message passing for global subseasonal-to-seasonal forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1879–1890, 2025.

[43] Yang Liu, Jiashun Cheng, Haihong Zhao, Tingyang Xu, Peilin Zhao, Fugee Tsung, Jia Li, and Yu Rong. SEGNO: Generalizing equivariant graph neural networks with physical inductive biases. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Chaohao Yuan, Maoji Wen, Ercan Engin Kuruoğlu, Yang Liu, Jia Li, Tingyang Xu, Deli Zhao, Hong Cheng, and Yu Rong. Non-stationary equivariant graph neural networks for physical dynamics simulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[45] Timothy T Duignan, Shawn M Kathmann, Gregory K Schenter, and Christopher J Mundy. Toward a first-principles framework for predicting collective properties of electrolytes. *Accounts of Chemical Research*, 54(13):2833–2843, 2021.

[46] Ioan-Bogdan Magdău, Daniel J Arismendi-Arrieta, Holly E Smith, Clare P Grey, Kersti Hermansson, and Gábor Csányi. Machine learning force fields for molecular liquids: Ethylene carbonate/ethyl methyl carbonate binary solvent. *npj Computational Materials*, 9(1):146, 2023.

[47] Nan Yao, Xiang Chen, Zhong-Heng Fu, and Qiang Zhang. Applying classical, ab initio, and machine-learning molecular dynamics simulations to the liquid electrolyte for rechargeable batteries. *Chemical Reviews*, 122(12):10970–11021, 2022.

[48] Oleg Borodin, Guorong V Zhuang, Philip N Ross, and Kang Xu. Molecular dynamics simulations and experimental study of lithium ion transport in dilithium ethylene dicarbonate. *The Journal of Physical Chemistry C*, 117(15):7433–7444, 2013.

[49] Jose Manuel Vicent-Luna, Jose Manuel Ortiz-Roldan, Said Hamad, Ramon Tena-Zaera, Sofia Calero, and Juan Antonio Anta. Quantum and classical molecular dynamics of ionic liquid electrolytes for na/li-based batteries: molecular origins of the conductivity behavior. *ChemPhysChem*, 17(16):2473–2481, 2016.

[50] Indra Priyadarsini, Vidushi Sharma, Seiji Takeda, Akihiro Kishimoto, Lisa Hamada, and Hajime Shinohara. Improving performance prediction of electrolyte formulations with transformer-based molecular representation model. In *ICML Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024.

[51] Xinhe Li, Zhuoying Feng, Yezeng Chen, Weichen Dai, Zixu He, Yi Zhou, and Shuhong Jiao. Coeff-kans: A paradigm to address the electrolyte field with kans. *arXiv preprint arXiv:2407.20265*, 2024.

[52] Jiacheng Cen, Anyi Li, Ning Lin, Yuxiang Ren, Zihe Wang, and Wenbing Huang. Are high-degree representations really unnecessary in equivariant graph neural networks? *Advances in Neural Information Processing Systems*, 37:26238–26266, 2024.

[53] Sarp Aykent and Tian Xia. Gotennet: Rethinking efficient 3d equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*. PMLR, 2021.

[55] 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.

[56] Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022.

[57] Jonathan Godwin, Michael Schaarschmidt, Alexander L Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations*, 2022.

[58] Michiel Hazewinkel, Nadiya Gubareni, and Vladimir V Kirichenko. *Algebras, Rings and Modules: Volume 1*, volume 575. Springer Science & Business Media, 2006.

[59] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[61] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 2025.

[62] RDKit: Open-source cheminformatics. https://www.rdkit.org.

[63] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[64] Nofit Segal, Aviv Netanyahu, Kevin P Greenman, Pulkit Agrawal, and Rafael Gomez-Bombarelli. Known unknowns: Out-of-distribution property prediction in materials and molecules. *arXiv preprint arXiv:2502.05970*, 2025.

[65] Zihan Tan, Guancheng Wan, Wenke Huang, and Mang Ye. Fedssp: Federated graph learning with spectral knowledge and personalized preference. *Advances in Neural Information Processing Systems*, 37:34561–34581, 2024.

[66] Soledad Villar, David W Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. *Advances in Neural Information Processing Systems*, 34:28848–28863, 2021.

[67] Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Equivariant graph mechanics networks with constraints. In *International Conference on Learning Representations*, 2022.

[68] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Contents of Appendix

A Theoretical Details

A.1 Expressivity of Intermolecular Transformation Matrix

Lemma A.1. *For any $\mathrm{O}(n)$-equivariant function $\hat{f}(\vec{\boldsymbol{Z}})$, it must fall into the subspace spanned by the columns of $\vec{\boldsymbol{Z}}$, namely, there exists a function $s(\vec{\boldsymbol{Z}})$, satisfying $\hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}})$.*

Proof. The proof is given by [66]. Essentially, suppose $\vec{\boldsymbol{Z}}^{\perp}$ is the orthogonal complement of the column space of $\vec{\boldsymbol{Z}}$. Then there must exit functions $s(\vec{\boldsymbol{Z}})$ and $s^{\perp}(\vec{\boldsymbol{Z}})$, satisfying $\hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}}) + \vec{\boldsymbol{Z}}^{\perp}s^{\perp}(\vec{\boldsymbol{Z}})$. We can always find an orthogonal transformation $\boldsymbol{O}$ allowing $\boldsymbol{O}\vec{\boldsymbol{Z}} = \vec{\boldsymbol{Z}}$ while $\boldsymbol{O}\vec{\boldsymbol{Z}}^{\perp} = -\vec{\boldsymbol{Z}}^{\perp}$. With this transformation $\boldsymbol{O}$, we have $\hat{f}(\boldsymbol{O}\vec{\boldsymbol{Z}}) = \hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}}) + \vec{\boldsymbol{Z}}^{\perp}s^{\perp}(\vec{\boldsymbol{Z}})$, and $\boldsymbol{O}\hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}}) - \vec{\boldsymbol{Z}}^{\perp}s^{\perp}(\vec{\boldsymbol{Z}})$. The equivariance property of $\hat{f}$ implies $\vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}}) + \vec{\boldsymbol{Z}}^{\perp}s^{\perp}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}s(\vec{\boldsymbol{Z}}) - \vec{\boldsymbol{Z}}^{\perp}s^{\perp}(\vec{\boldsymbol{Z}})$, which derives $s^{\perp}(\vec{\boldsymbol{Z}}) = 0$. Hence, the proof is concluded. □

Lemma A.2. *If the $\mathrm{O}(n)$-equivariant function $\hat{f}(\vec{\boldsymbol{Z}})$ lies in the subspace spanned by the columns of $\vec{\boldsymbol{Z}}$, then there exists a function σ satisfying $\hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}\sigma(\vec{\boldsymbol{Z}}^{\top}\boldsymbol{Z})$.*

Proof. The proof is provided by Corollary 2 in [67]. The basic idea is that $\hat{f}(\vec{\boldsymbol{Z}})$ can be transformed to $\hat{f}(\vec{\boldsymbol{Z}}) = \vec{\boldsymbol{Z}}\eta(\vec{\boldsymbol{Z}})$ where $\eta(\vec{\boldsymbol{Z}})$ is $\mathrm{O}(n)$-invariant. According to Lemma 1-2 in [67], $\eta(\vec{\boldsymbol{Z}})$ must be written as $\eta(\vec{\boldsymbol{Z}}) = \sigma(\vec{\boldsymbol{Z}}^{\top}\vec{\boldsymbol{Z}})$, which completes the proof. □

Theorem 3.1 (Expressivity of Intermolecular Transformation Matrix)**.** *Given geometric graphs $\mathcal{G}_m$ and $\mathcal{G}_n$, with $\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n$ being respective frames, any matrix $\overleftrightarrow{\boldsymbol{I}}_{mn}$ satisfying $\mathrm{SO}(3)$-equivariance $\overleftrightarrow{\boldsymbol{I}}_{mn} \xmapsto{\boldsymbol{R}_m, \boldsymbol{R}_n \in \mathrm{SO}(3)} \boldsymbol{R}_m \overleftrightarrow{\boldsymbol{I}}_{mn} \boldsymbol{R}_n^{\top}$ must take the form like Eq. (10) as $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n) = \vec{\boldsymbol{F}}_m \boldsymbol{R} \vec{\boldsymbol{F}}_n^{\top}$.*

Proof. First, we want to show that, $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n))$ is equivalent to $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n))$. We know the Kronecker product $\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n$ contains 9×9 elements, which could be divided into 9 small matrices of 3×3, each corresponding to the product of a row in $\vec{\boldsymbol{F}}_m = [\vec{\boldsymbol{u}}_1; \vec{\boldsymbol{u}}_2; \vec{\boldsymbol{u}}_3]$ and a column in $\vec{\boldsymbol{F}}_n = [\vec{\boldsymbol{v}}_1^{\top}; \vec{\boldsymbol{v}}_2^{\top}; \vec{\boldsymbol{v}}_3^{\top}]$, *i.e.* $\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n = \left[\vec{\boldsymbol{u}}_i \otimes \vec{\boldsymbol{v}}_j^{\top}\right]$. Note $\vec{\boldsymbol{u}}_i \otimes \vec{\boldsymbol{v}}_j^{\top}$ is a rank-one matrix and $\vec{\boldsymbol{u}}_i, \vec{\boldsymbol{v}}_j$ are both unit vectors. Therefore, $\vec{\boldsymbol{u}}_i \otimes \vec{\boldsymbol{v}}_j^{\top}$ corresponds to two possibilities: $(\vec{\boldsymbol{u}}_i, \vec{\boldsymbol{v}}_j^{\top})$ or $(-\vec{\boldsymbol{u}}_i, -\vec{\boldsymbol{v}}_j^{\top})$. By analyzing the 9 submatrices, we can find that $\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n$ also corresponds to two possible decomposition, namely $(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n)$ and $(-\vec{\boldsymbol{F}}_m, -\vec{\boldsymbol{F}}_n)$. Note that during the construction, we agreed that $\det(\vec{\boldsymbol{F}}_m) = \det(\vec{\boldsymbol{F}}_n) = 1$, so there is a direct bijection between $\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n$ and $(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n)$, which means $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n))$ is equivalent to $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n))$.

Then, we use the vectorization operator $\mathrm{vec}(\cdot)$ to prove $\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n))$ must take the form:

$$\mathrm{vec}(\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m \otimes \vec{\boldsymbol{F}}_n)) = \mathrm{vec}(\vec{\boldsymbol{F}}_m \boldsymbol{R} \vec{\boldsymbol{F}}_n^{\top}) = (\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)\,\mathrm{vec}(\boldsymbol{R}). \tag{16}$$

Note that $\vec{\boldsymbol{F}}_n, \vec{\boldsymbol{F}}_m$ are constructed frames, which is full-rank. Then we find $\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m$ will be also full-rank, spanning the whole space $\mathbb{R}^9$ where $\mathrm{vec}(\overleftrightarrow{\boldsymbol{I}}_{mn}(\vec{\boldsymbol{F}}_m, \vec{\boldsymbol{F}}_n)$ lies in. According to Lemma A.2, we know $\mathrm{vec}(\overleftrightarrow{\boldsymbol{I}}_{mn})$ must satisfy:

$$\mathrm{vec}(\overleftrightarrow{\boldsymbol{I}}_{mn}) = (\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)\sigma((\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)^{\top}(\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)). \tag{17}$$

Moreover, we have such formula:

$$(\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)^{\top}(\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m) = (\vec{\boldsymbol{F}}_n^{\top}\vec{\boldsymbol{F}}_n) \otimes (\vec{\boldsymbol{F}}_m^{\top}\vec{\boldsymbol{F}}_m) = \boldsymbol{I}_{3\times 3} \otimes \boldsymbol{I}_{3\times 3} = \boldsymbol{I}_{9\times 9}, \tag{18}$$

which is a constant matrix. Thus $\sigma((\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)^{\top}(\vec{\boldsymbol{F}}_n \otimes \vec{\boldsymbol{F}}_m)) = \sigma(\boldsymbol{I}_{9\times 9})$ will be learnable parameters, which are equivalent to the $\mathrm{vec}(\boldsymbol{R})$ here. Up to this point, the matrix $\overleftrightarrow{\boldsymbol{I}}_{mn}$ appears to exhibit $\mathrm{O}(3)$-equivariance. However, it is crucial to observe that the frame constructions $\mathcal{F}_n$ and $\mathcal{F}_m$ are only $\mathrm{SO}(3)$-equivariant. Consequently, the overall transformation reduces to $\mathrm{SO}(3)$-equivariance. This completes the proof. □

A.2 Equivariance/Invariance of GeoMix

A.2.1 Enhancement of Equivariance/Invariance

In order to prove the equivariance/invariance of the entire GeoMix model, we need to strengthen the conditions.

Theorem A.3 (Enhancement of Equivariance/Invariance)**.** *If each module of GeoMix is equivariant/invariant, then the entire GeoMix will also be equivariant/invariant.*

Proof. Consider a sequence composed of functions $\{\phi_i : \mathcal{X}^{(i-1)} \to \mathcal{X}^{(i)}\}_{i=1}^{N}$ equivariant to a same group G, the equivariance lead to an interesting property that

$$\phi_N \circ \cdots \circ \phi_{i+1} \circ \rho_{\mathcal{X}^{(i)}}(g)\phi_i \circ \cdots \circ \phi_1 = \phi_N \circ \cdots \circ \phi_{j+1} \circ \rho_{\mathcal{X}^{(j)}}(g)\phi_j \circ \cdots \circ \phi_1,$$

holds for all $i, j = 1, 2, \ldots, N$ and $g \in G$, which means that the group elements g can be freely exchanged in the composite sequence of equivariant functions. In particular, if one of the equivariant functions (*e.g.* ϕ_k) is replaced by an invariant function, the group element g will be absorbed, that means

$$\phi_N \circ \cdots \circ \phi_k \circ \cdots \circ \phi_{i+1} \circ \rho_{\mathcal{X}^{(i)}}(g)\phi_i \circ \cdots \circ \phi_1 = \phi_N \circ \cdots \circ \phi_1.$$

holds for all $g \in G$ but only $i = 1, 2, \ldots, k$. Although $\phi_N \circ \cdots \circ \phi_k$ is still equivariant, because the group elements must be input starting from ϕ_1, the overall $\phi_N \circ \cdots \circ \phi_1$ is still an invariant function. □

Specifically, we need to prove the following:

1. $\mathrm{SO}(3)$-equivariance of local frame.
2. $\mathrm{SO}(3)$-equivariance of learnable transformations.

A.2.2 Equivariance of Local Frame

Theorem A.4 (Equivariance of Local Frame)**.** *The construction of local frame is* $\mathrm{SO}(3)$*-equivariance.*

Proof. We prove that the local frame $\vec{\boldsymbol{F}}_m$ is $\mathrm{SO}(3)$-equivariant with respect to the target graph $\mathcal{G}_m$. More formally, for any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{3\times3}$, the local frame should satisfy:

$$\boldsymbol{Q}\vec{\boldsymbol{F}}_m = \texttt{LocalFrame}(\boldsymbol{Q}\vec{\boldsymbol{X}}_m). \tag{19}$$

Formally, we have

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Q}\vec{\boldsymbol{X}}_m) &= \frac{1}{N_m}\boldsymbol{Q}\vec{\boldsymbol{X}}_m(\boldsymbol{Q}\vec{\boldsymbol{X}}_m)^\top = \boldsymbol{Q}\frac{1}{N_m}\vec{\boldsymbol{X}}_m\vec{\boldsymbol{X}}_m^\top\boldsymbol{Q}^\top \\
&= \boldsymbol{Q}\mathrm{Cov}(\vec{\boldsymbol{X}}_m)\boldsymbol{Q}^\top = \boldsymbol{Q}\vec{\boldsymbol{F}}_m\Lambda_m\vec{\boldsymbol{F}}_m^\top\boldsymbol{Q}^\top \\
&= (\boldsymbol{Q}\vec{\boldsymbol{F}}_m)\Lambda_m(\boldsymbol{Q}\vec{\boldsymbol{F}}_m)^\top.
\end{aligned} \tag{20}$$

Concluding we showed that an $\mathrm{SO}(3)$ transformation on the original coordinate matrix $\vec{\boldsymbol{X}}_m$ results in the same transformation on the output local frame $\vec{\boldsymbol{F}}_m$. □

A.2.3 Equivariance of Learnable Transformations

Theorem A.5 (Equivariance of Learnable Transformations)**.** *The construction of learnable transformations is* $\mathrm{SO}(3)$*-equivariance.*

Proof. We prove the message $\vec{\boldsymbol{m}}_{ij}^{(m)}$ and $\vec{\boldsymbol{m}}_{ij}^{(m,n)}$ in Eq. (13) is $\mathrm{SO}(3)$ equivariant on the target graph $\mathcal{G}_m$ and the source graph $\mathcal{G}_n$ respectively, that is, for any orthogonal matrix $\boldsymbol{Q}_m, \boldsymbol{Q}_n \in \mathbb{R}^{3\times3}$, the message should satisfy:

$$\begin{aligned}
\boldsymbol{Q}_n\vec{\boldsymbol{m}}_{ij}^{(m,n)} &= \texttt{MessageSource}(\boldsymbol{Q}_n\vec{\boldsymbol{X}}_n, \boldsymbol{Q}_n\vec{\boldsymbol{V}}_n), \\
\boldsymbol{Q}_m\vec{\boldsymbol{m}}_{ij}^{(m)} &= \texttt{MessageTarget}(\boldsymbol{Q}_m\vec{\boldsymbol{X}}_m).
\end{aligned} \tag{21}$$

It is easy to see $\boldsymbol{m}_{ij}^{(m,n)}$ is invariant on both the source and the target graph. So,

$$\begin{aligned}
&\sigma_{\vec{\boldsymbol{x}}}(\boldsymbol{m}_{ij}^{(m,n)})\boldsymbol{Q}_n\vec{\boldsymbol{x}}_j^{(n)} + \sigma_{\vec{\boldsymbol{v}}}(\boldsymbol{m}_{ij}^{(m,n)})\boldsymbol{Q}_n\vec{\boldsymbol{v}}_j^{(n)} \\
=&\boldsymbol{Q}_n\sigma_{\vec{\boldsymbol{x}}}(\boldsymbol{m}_{ij}^{(m,n)})\vec{\boldsymbol{x}}_j^{(n)} + \boldsymbol{Q}_n\sigma_{\vec{\boldsymbol{v}}}(\boldsymbol{m}_{ij}^{(m,n)})\vec{\boldsymbol{v}}_j^{(n)} \\
=&\boldsymbol{Q}_n\vec{\boldsymbol{m}}_{ij}^{(m,n)},
\end{aligned} \tag{22}$$

which is $\mathrm{SO}(3)$ equivariant on the source graph.

We have proven that the local frame $\vec{\boldsymbol{F}}_m$ is $\mathrm{SO}(3)$ equivariant on the target graph, so $\vec{\boldsymbol{F}}_n$ is $\mathrm{SO}(3)$ equivariant on the source graph. We have:

$$\begin{aligned}
&\boldsymbol{Q}_m\vec{\boldsymbol{F}}_m\boldsymbol{R}_{ij}^{(m,n)}(\boldsymbol{Q}_n\vec{\boldsymbol{F}}_n)^\top\boldsymbol{Q}_n\vec{\boldsymbol{m}}_{ij}^{(m,n)} + \boldsymbol{Q}_m\vec{\boldsymbol{F}}_m\boldsymbol{t}_{ij}^{(m,n)} \\
=&\boldsymbol{Q}_m\vec{\boldsymbol{F}}_m\boldsymbol{R}_{ij}^{(m,n)}\vec{\boldsymbol{F}}_n^\top\boldsymbol{Q}_n^\top\boldsymbol{Q}_n\vec{\boldsymbol{m}}_{ij}^{(m,n)} + \boldsymbol{Q}_m\vec{\boldsymbol{t}}_{ij}^{(m)} \\
=&\boldsymbol{Q}_m\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)}\vec{\boldsymbol{m}}_{ij}^{(m,n)} + \boldsymbol{Q}_m\vec{\boldsymbol{t}}_{ij}^{(m)} \\
=&\boldsymbol{Q}_m\vec{\boldsymbol{m}}_{ij}^{(m)},
\end{aligned} \tag{23}$$

which is $\mathrm{SO}(3)$ equivariant on the target graph. □

B More Experiment Details and Results

Our code and dataset are available at https://github.com/GLAD-RUC/GeoMix.

B.1 Dataset Details

In this part, we provide supplementary details and visualizations of the datasets introduced in the main text. These include the distribution of molecular species, property ranges, and composition coverage, as illustrated in Table 3.

Table 3: Summary statistics of the CALiSol [19] and DiffMix [11] datasets, including the number of samples, solvent and salt types, elemental coverage, maximum number of solvents per formulation, solute concentration ranges, and temperature ranges.

	CALiSol	**DiffMix**
Number of samples	13,269	24,822
Number of unique solvents	38	6
Solvent list	EC, PC, DMC, DEC, DME, DMSO, AN, MOEMC, TFP, EA, MA, FEC, DOL, 2-MeTHF, DMM, Freon 11, Methylene chloride, THF, Toluene, Sulfolane, 2-Glyme, 3-Glyme, 4-Glyme, 3-Me-2-Oxazolidinone, 3-MeSulfolane, Ethyldiglyme, DMF, Ethylbenzene, Ethylmonoglyme, Benzene, g-Butyrolactone, Cumene, Propylsulfone, Pseudocumeme, TEOS, m-Xylene, o-Xylene	EC, PC, FEC, EMC, DEC, DMC
Number of unique salts	13	2
Salt list	$LiPF_6$, $LiBF_4$, LiFSI, LiTDI, LiPDI, LiTFSI, $LiClO_4$, $LiAsF_6$, LiBOB, $LiCF_3SO_3$, LiBPFPB, LiBMB, $LiN(CF_3SO_2)_2$	$LiPF_6$, LiTFSI
Elements covered	H, Li, B, C, N, O, F, Si, P, S, Cl, As	H, Li, C, N, O, F, P, S
Maximum number of solvents per sample	4	3
Solute composition range	0.0286-2.37 mol/kg and 0.0771-4.00 mol/L	0.025-3.0 mol/kg [4]
Temperature range	194.15-477.42K	243.15K-293.15K

B.2 Implementation Details

In this part, we describe the implementation details. Each model architecture remains the same in both datasets.

- **MLP**: A 3-layer multilayer perceptron is used, with each hidden layer having 64 dimensions.
- **MM-MoLFormer**: We follow the settings in [12]. MM-MolFormer uses a pre-trained MoLFormer model [68], obtained from `https://ibm.ent.box.com/v/MoLFormer-data`. The molecular embeddings are concatenated with the corresponding proportion and temperature features, and then passed through a 2-layer MLP with hidden size 64 to generate the final prediction.
- **MolSets**: For both MolSets-Conv and MolSets-SAGE, we retain the original hyperparameter settings from [13], except that the hidden dimension of each graph neural network layer is increased to 64.
- **EGNN-att**: This variant employs a 4-layer EGNN backbone with 64 hidden dimensions per layer. The attention module takes 8 molecular embeddings as input, with 4 attention heads and 3 stacked attention layers. The resulting representation is concatenated with temperature T and salt concentration c, and then passed through a 3-layer MLP for prediction.
- **TFN-att**: TFN-att shares the same configuration as EGNN-att except for the backbone, which is replaced with a TFN. The TFN backbone uses a maximum angular momentum quantum number of $\max L = 2$, and each irreducible representation (irrep) has 8 channels.
- **EGNN-linear**: This model uses the same EGNN backbone as EGNN-att. Instead of an attention module, a simple non-learnable linear average (scatter) operation is applied to obtain the mixture representation
- **TFN-linear**: This model adopts the same TFN backbone as TFN-att, and replaces the attention module with a non-learnable linear average operation, as in EGNN-linear.
- **GeoMix-EGNN**: GeoMix-EGNN adopts a multi-channel EGNN as the backbone, with 8 equivariant channels. Other hyperparameters follow the EGNN-att setting. The GIN consists of 3 layers. The Noisy Nodes loss [57] is applied with hyperparameter $\gamma = 128$.
- **GeoMix-TFN**: GeoMix-TFN uses the same TFN backbone as TFN-att. The GIN also consists of 3 layers. The Noisy Nodes loss is applied with $\gamma = 128$, as in GeoMix-EGNN.

B.3 Training Settings

All models are trained using the Adam optimizer with a learning rate of 5×10^{-5}, weight decay of 1×10^{-12}, and a maximum of 500 training epochs. The default batch size is 1024, except for GeoMix models, which use a reduced batch size of 128 due to GPU memory constraints. We fix the random seed to 7 for all experiments to ensure reproducibility.

We run our methods mainly on A100 80GB GPUs.

B.4 OOD Evaluation across Conductivity.

In practical chemical research, it is often necessary to identify substances with high ionic conductivity based on prior knowledge of low-conductivity samples, which naturally poses an out-of-distribution (OOD) generalization challenge. However, most existing machine learning models struggle to generalize reliably under such distribution shifts. Here, we specifically consider an OOD split where low-conductivity samples are used for training and validation, while high-conductivity samples are reserved for testing.

Experiment Setup. To evaluate our model's robustness in this setting, we selected samples with conductivity ≤ 10 mS/cm as the training and validation sets, and those with conductivity > 10 mS/cm as the test set on the CALiSol dataset. The training and validation sets were split randomly in a 4:1 ratio. The high-conductivity test set contains 1,244 samples, ensuring that this split is statistically meaningful. Here, we evaluate only EGNN-linear and TFN-linear, the two best-performing models

[4]The salt concentrations are discretized into seven levels: {0.025, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0} molal. For each formulation, the solvent mass fractions vary from 0 to 1 in increments of 0.2 and are normalized to sum to one.

Table 4: Results of OOD evaluation across conductivity on CALiSol dataset. Bold values indicate the best performance, while underlined values indicate the second-best.

Models	MSE ↓	MAE ↓	Pearson r ↑	Spearman r ↑
EGNN-linear [54]	32.720	4.457	0.253	0.341
TFN-linear [55]	24.218	3.585	0.397	0.508
GeoMix-EGNN	**17.132**	**2.853**	**0.579**	<u>0.571</u>
GeoMix-TFN	<u>19.427</u>	<u>2.925</u>	<u>0.436</u>	**0.609**

in § 4.2, as baselines. During training, we observed that attention-based aggregation of graph-level information tended to overfit. To mitigate this, we applied a dropout rate of 0.1 to the attention layer and set the attention temperature to 2.0. All other settings remained unchanged.

Results. The results for this OOD split based on conductivity values are summarized in Table 4. Both baseline models, EGNN-linear and TFN-linear, exhibit limited generalization. Our proposed models, GeoMix-EGNN and GeoMix-TFN, show significantly lower MSE and higher correlation metrics compared to the baselines, demonstrating their effectiveness in capturing trends in this OOD setting. These results highlight that our approach improves predictive performance while enhancing OOD generalization under conductivity-based distribution shifts.

B.5 OOD Evaluation across Temperature.

Because the free diffusion rate of most materials at ambient temperature is low, it is common to perform molecular dynamics simulations at elevated temperatures and then extrapolate to lower temperatures using the Arrhenius empirical relation [8]. Therefore, we explored the consistency of model predictions with respect to temperature.

Experiment Setup. To empirically assess this, we performed an OOD generalization experiment where we trained the model on data with temperature ≤ 320 K (using a train/valid split of 4:1), and evaluated on samples with temperature > 320 K. All other settings remained unchanged.

Results. Results in Table 5 show that GeoMix achieves significantly better predictive performance on this challenging temperature extrapolation task compared to other baseline models, suggesting that the learned representations effectively capture physically meaningful temperature trends without requiring hard-coded constraints.

Table 5: Results of OOD evaluation across temperature on CALiSol dataset. Bold values indicate the best performance, while underlined values indicate the second-best.

Models	MSE ↓	MAE ↓	Pearson r ↑	Spearman r ↑
MLP	32.432	3.857	0.432	0.556
MM-MoLFormer [12]	22.935	2.837	0.512	0.560
MolSets-Conv [13]	9.930	2.196	0.839	0.845
MolSets-SAGE [13]	9.193	2.135	0.784	0.802
EGNN-att [54]	8.134	1.594	0.813	0.861
TFN-att [55]	6.516	1.383	0.853	0.887
EGNN-linear [54]	7.675	1.616	0.827	0.847
TFN-linear [55]	4.780	1.286	0.914	0.930
GeoMix-EGNN	<u>2.366</u>	**0.883**	<u>0.950</u>	**0.948**
GeoMix-TFN	**2.354**	<u>0.917</u>	**0.952**	<u>0.945</u>

B.6 Experiment of coordinate perturbation.

For complex molecular systems, the initial conformations used in modeling may not always be fully accurate. To assess the robustness of our model to perturbations in the input geometries, we conducted an additional robustness evaluation.

Experiment Setup. All other settings remained unchanged, except that Gaussian noise with a standard deviation of $\sigma = 0.3$ Å was added to the atomic coordinates in the test set to simulate conformational uncertainty.

Results. As shown in Table 6, most models—including GeoMix-EGNN—exhibit only a marginal performance drop, indicating stability under moderate geometric perturbations. Notably, models based on the TFN encoder show a much larger performance drop, which may stem from their reliance on higher-degree equivariant features (e.g., spherical harmonics) that could be sensitive to atomic coordinate perturbations.

Table 6: Results of coordinate perturbation on CALiSol dataset. Bold values indicate the best performance.

Models	MSE ↓	MAE ↓	Pearson r ↑	Spearman r ↑
MolSets-Conv [13]	1.326	0.703	0.960	0.965
MolSets-SAGE [13]	1.748	0.794	0.946	0.958
EGNN-att [54]	1.919	0.587	0.942	0.965
TFN-att [55]	2.294	0.760	0.929	0.954
EGNN-linear [54]	1.114	0.555	0.967	0.976
TFN-linear [55]	2.393	0.937	0.929	0.944
GeoMix-EGNN	**0.631**	**0.388**	**0.982**	**0.983**
GeoMix-TFN	4.517	1.324	0.859	0.886

B.7 Runtime Analysis on CALiSol dataset.

Results in Table 7 shows the runtime analysis comparing our methods with other main baselines on the CALiSol dataset.

Table 7: Runtime Analysis on CALiSol dataset.

Models	Inference Time ($\times 10^{-2}$ s)	Memory Usage (GB)	MSE ↓	Pearson r ↑
EGNN-linear [54]	4.41	2.20	1.461	0.951
TFN-linear [55]	24.13	5.91	1.107	0.967
EGNN-att [54]	23.40	2.22	2.666	0.908
TFN-att [55]	43.20	5.91	1.808	0.946
GeoMix-EGNN	34.04	25.15	0.552	0.985
GeoMix-TFN	53.64	45.18	0.432	0.987

B.8 Evaluation of Different Frame Construction Methods

As described in § 3.2, the local frame $\vec{\boldsymbol{F}}_m$ for each molecule is constructed via eigen-decomposition of the covariance matrix $\mathrm{Cov}(\vec{\boldsymbol{X}}_m)$. However, the axis orientation of $\vec{\boldsymbol{F}}_m$ obtained from standard PCA is not unique, as simultaneous sign flips of eigenvectors can lead to instability in local frame alignment. To alleviate this ambiguity, we additionally test a deterministic version of PCA by enumerating all eight possible axis-orientation combinations of the eigenbasis and selecting the one with the largest coordinate sum, i.e.,

$$\vec{\boldsymbol{F}}_m^* = \underset{\vec{\boldsymbol{F}}'_m \in \mathbb{F}(\vec{\boldsymbol{F}}_m)}{\operatorname{argmax}} \mathbf{1}^\top \vec{\boldsymbol{F}}'_m \mathbf{1}, \tag{24}$$

where $\mathbb{F}(\vec{\boldsymbol{F}}_m)$ denotes the set of all sign permutations of the three principal axes. This procedure ensures a consistent frame orientation across molecules.

Performance Comparison. Table 8 summarizes the predictive performance of different frame construction methods under the same experimental setup as described in § 4.2. The results show that the deterministic (strict) PCA achieves performance comparable to the standard PCA, though with slightly higher errors in some cases. According to [56], the standard PCA-based frame construction, while not strictly equivariant, remains reliable and effective for defining local coordinate systems in

molecular modeling. The reduced flexibility of the strictly equivariant variant may account for its marginally lower performance. In addition, the methods in [16; 17] can also be used as alternatives, which we will study in future work.

Table 8: Results on the CALiSol dataset. *FC Methods* denotes the frame construction methods. The *strict PCA* refers to the deterministic variant defined in Eq. (24).

Models	FC Methods	MSE ↓	Pearson r ↑
GeoMix-EGNN	PCA	0.552	0.985
GeoMix-TFN	PCA	0.432	0.987
GeoMix-EGNN	strict PCA	0.628	0.981
GeoMix-TFN	strict PCA	0.460	0.986

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: The abstract and introduction clearly summarize both the model design in § 3 and the experiment results mentioned in § 4, accurately reflecting the paper's key contributions and scope.

   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 provide all these in § 6.

   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: We provide all these in Appendix A.

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: We provide all these in Appendix B.

   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: Our code is available at https://github.com/GLAD-RUC/GeoMix.

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: We provide all these in Appendix B.

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: In Fig. 3, we visualize the prediction results along with 3σ error bands. This provides a visual indication of the uncertainty and variability of our model's predictions.

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: We provide all these in Appendix B.

   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: Our research conducted in the paper 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: Our work focuses on improving the predictive modeling of electrolyte conductivity, which may benefit the development of high-performance batteries and contribute positively to clean energy technologies and sustainability. We do not foresee immediate negative societal impacts or misuse risks associated with this foundational 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 risks.

    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: All existing assets used in our paper are used with citation. The CALiSol [19] dataset is available under a CC-BY license. The DiffMix [11] dataset is available under a CC-BY-NC-ND 4.0 license.

    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: [Yes]

    Justification: Our code is available at `https://github.com/GLAD-RUC/GeoMix`.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.