Geometric Learning And Discovery

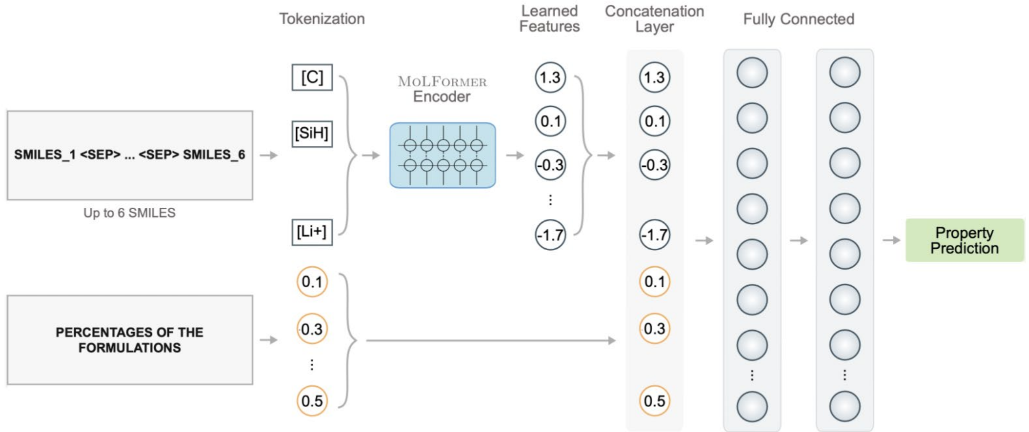
中国人民大學 RENMIN UNIVERSITY OF CHINA

Geometric Mixture Models for Electrolyte Conductivity Prediction

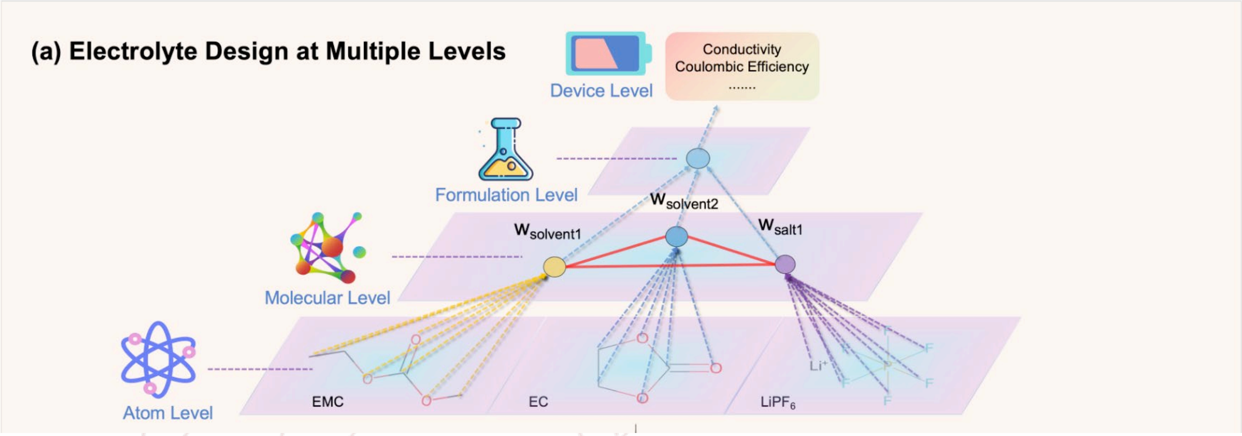
Anyi Li, Jiacheng Cen, Songyou Li, Mingze Li, Yang Yu, Wenbing Huang

2025.10

Representation of Mixture Systems



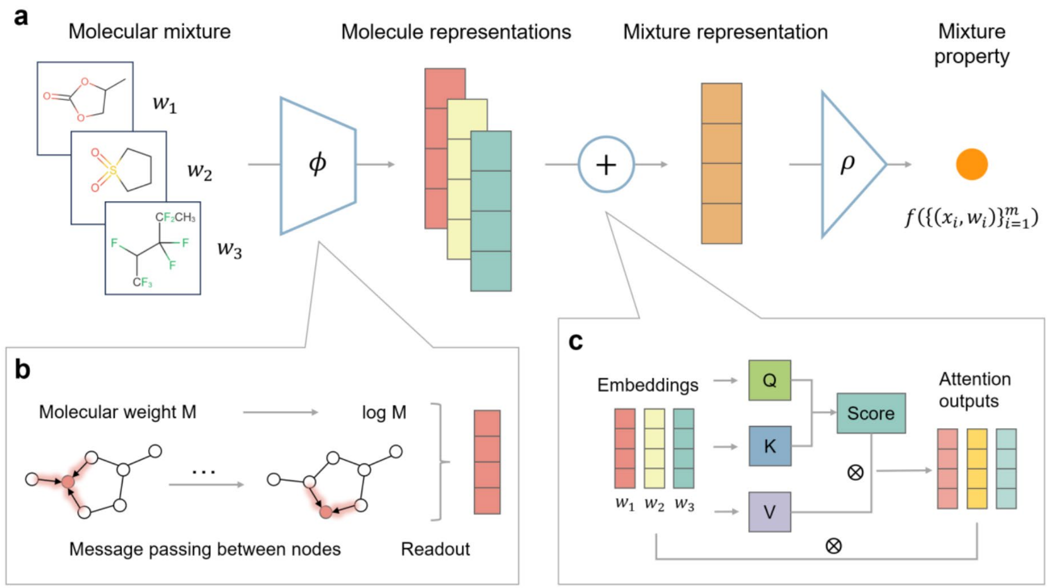
MM-MoLFormer[1] concatenates embedding and proportion of components, but **ignores the permutation invariance**.



Uni-ELF[3] uses a hierarchical graph representation, but geometric messages cannot pass between different components.

- Inputs:
  - A set of graphs $\{\mathcal{G}_m\}_{m=1}^{M}$, $\mathcal{G}_m = (\boldsymbol{H}_m, \vec{\boldsymbol{X}}_m, \vec{\boldsymbol{V}}_m, w_m)$.
  - A global environment condition c, i.e. temperature.
- Task: Scalar and vector information of mixture systems.

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



MolSets[2] respects permutation invariance but **discards 3D molecular geometry**.

Symmetry Constraints

- *Node-level* symmetry constraints
  - Permutation-equivariant in each graph

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

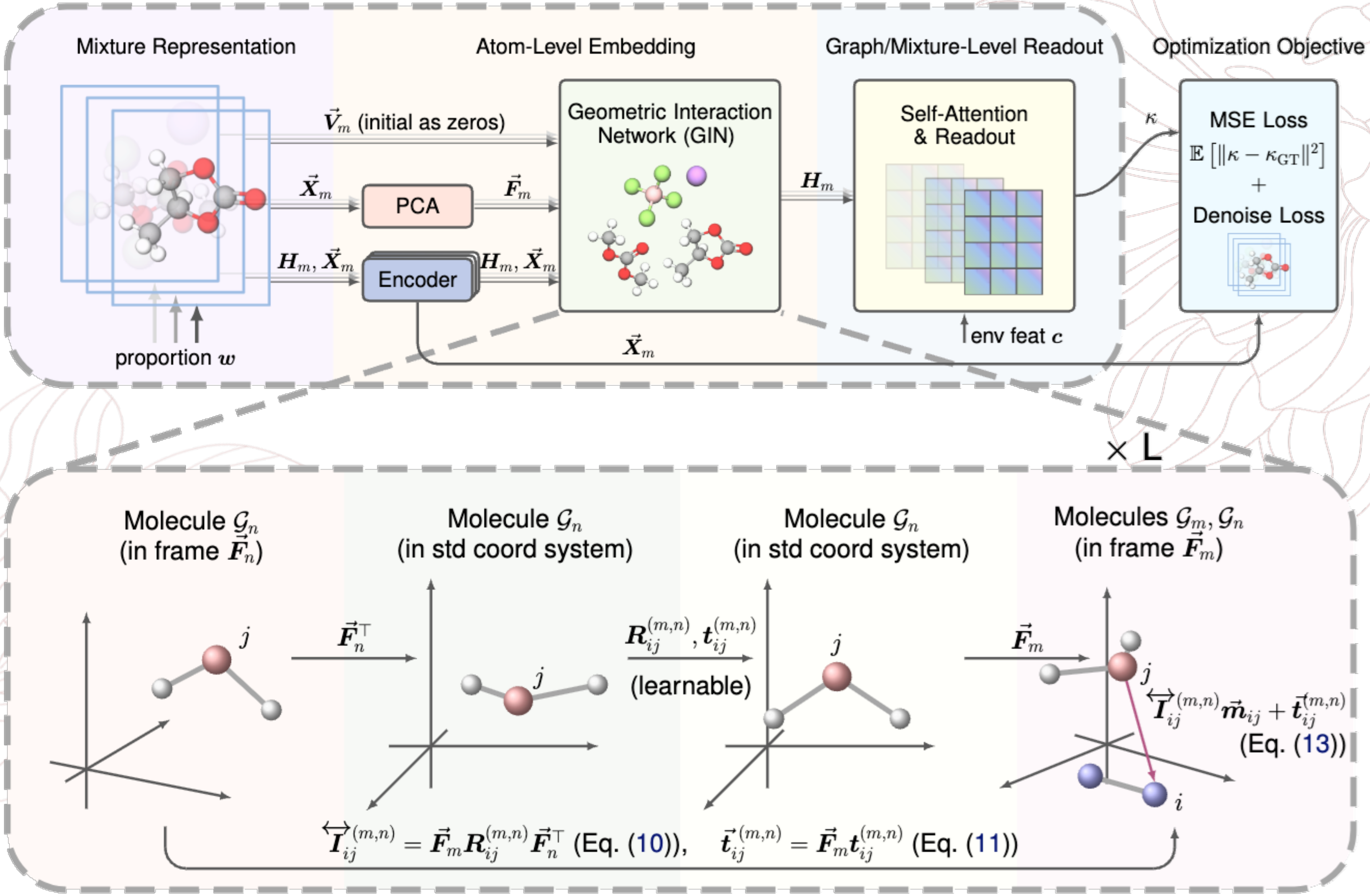
  - SE(3) equivariant on each graph

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

- *Graph-level* symmetry constraints
  - Permutation-equivariant with respect to the order of the graphs

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

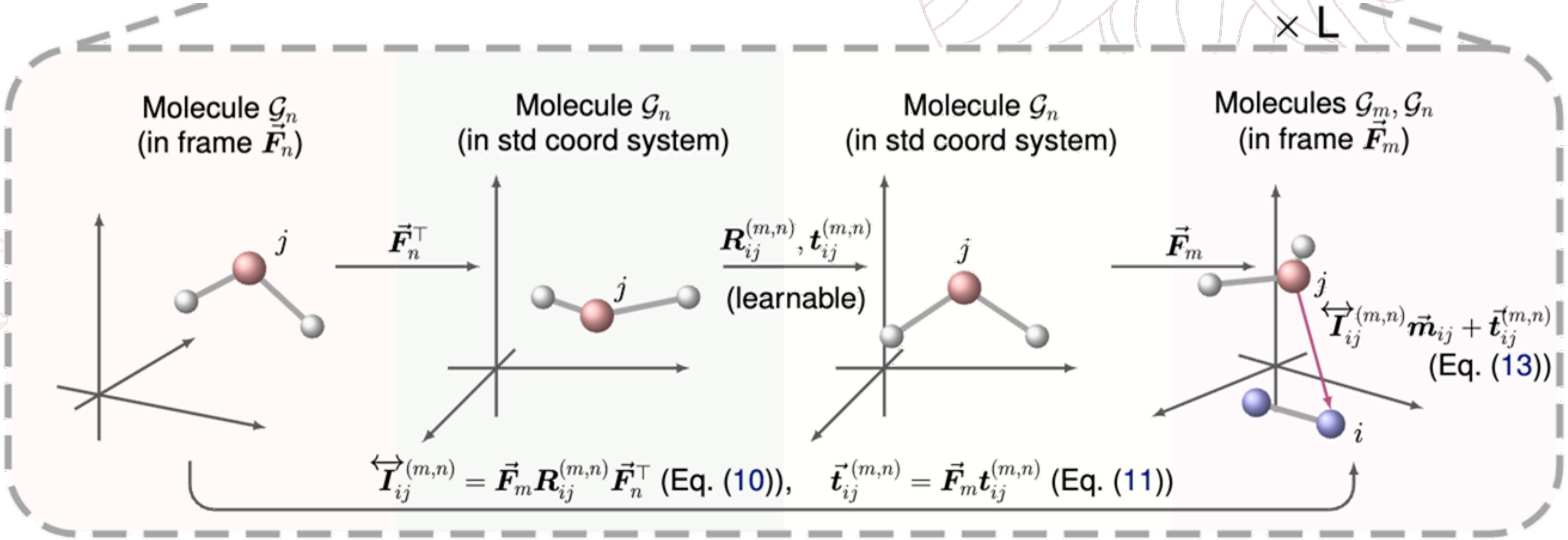
Pipeline



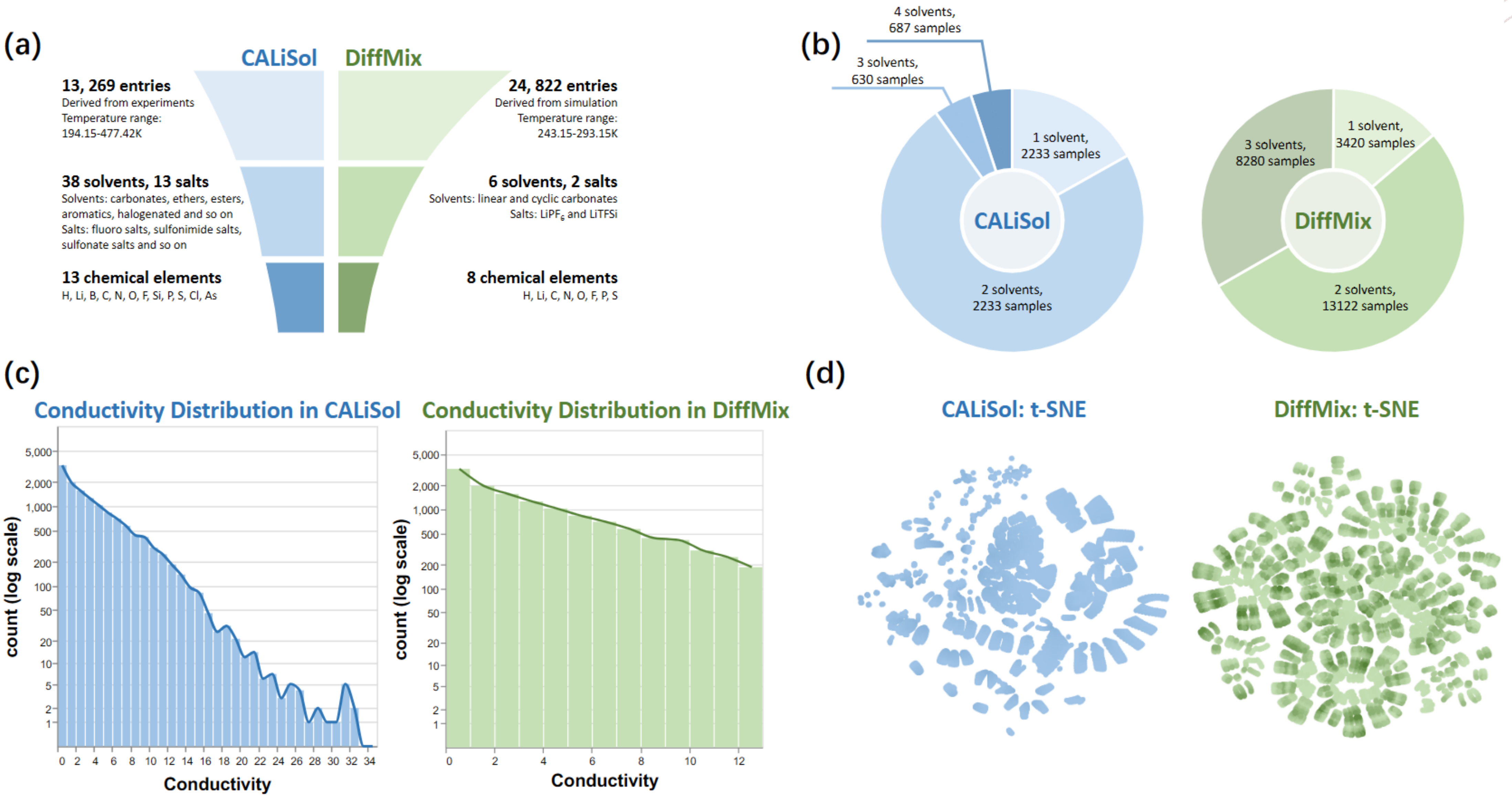
Mixture Representation
Atom-Level Embedding
Graph/Mixture-Level Readout
Optimization Objective
$\vec{V}_m$ (initial as zeros)
$\vec{X}_m$
PCA
$\vec{F}_m$
$H_m, \vec{X}_m$
Encoder
$H_m, \vec{X}_m$
Geometric Interaction Network (GIN)
H_m
Self-Attention & Readout
env feat c
κ
MSE Loss
$\mathbb{E}\left[\|\kappa - \kappa_{\mathrm{GT}}\|^2\right]$
+
Denoise Loss
proportion w
$\vec{X}_m$
$\times$ L
Molecule $\mathcal{G}_n$ (in frame $\vec{F}_n$)
Molecule $\mathcal{G}_n$ (in std coord system)
Molecule $\mathcal{G}_n$ (in std coord system)
Molecules $\mathcal{G}_m, \mathcal{G}_n$ (in frame $\vec{F}_m$)
$\vec{F}_n^\top$
$R_{ij}^{(m,n)}, t_{ij}^{(m,n)}$
(learnable)
$\vec{F}_m$
j
i
$\overleftrightarrow{I}_{ij}^{(m,n)} \vec{m}_{ij} + \vec{t}_{ij}^{(m,n)}$
(Eq. (13))
$\overleftrightarrow{I}_{ij}^{(m,n)} = \vec{F}_m R_{ij}^{(m,n)} \vec{F}_n^\top$ (Eq. (10)), $\vec{t}_{ij}^{(m,n)} = \vec{F}_m t_{ij}^{(m,n)}$ (Eq. (11))

Geometric Interaction Network (GIN)

- By **constructing local frames**, equivariant message can pass across molecules.
- Intermolecular Transformation:
  - $\boldsymbol{z}_{ij}^{(m,n)} = \sigma_{\text{inv}}\left(\boldsymbol{h}_i^{(m)}, \boldsymbol{h}_j^{(n)}, \left\|\vec{\boldsymbol{x}}_i^{(m)}\right\|, \left\|\vec{\boldsymbol{x}}_j^{(n)}\right\|, \left\|\vec{\boldsymbol{v}}_i^{(m)}\right\|, \left\|\vec{\boldsymbol{v}}_j^{(n)}\right\|\right).$
  - $\overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)} \coloneqq \vec{\boldsymbol{F}}_m \boldsymbol{R}_{ij}^{(m,n)} \vec{\boldsymbol{F}}_n^{\top}, \;\; \boldsymbol{R}_{ij}^{(m,n)} = \sigma_{\text{rot}}\left(\boldsymbol{z}_{ij}^{(m,n)}\right) \in \mathbb{R}^{3\times 3}, \;\; \vec{\boldsymbol{t}}_{ij}^{(m,n)} \coloneqq \vec{\boldsymbol{F}}_m \boldsymbol{t}_{ij}^{(m,n)}, \;\; \boldsymbol{t}_{ij}^{(m,n)} = \sigma_t\left(\boldsymbol{z}_{ij}^{(m,n)}\right) \in \mathbb{R}^3.$
- Message Construction:
  - $\boldsymbol{m}_{ij}^{(m,n)} = \sigma_{\text{msg}}\left(\boldsymbol{h}_i^{(m)}, \boldsymbol{h}_j^{(n)}, \left\|\vec{\boldsymbol{x}}_i^{(m)}\right\|, \left\|\vec{\boldsymbol{x}}_j^{(n)}\right\|, \left\|\vec{\boldsymbol{v}}_i^{(m)}\right\|, \left\|\vec{\boldsymbol{v}}_j^{(n)}\right\|, \left\|\vec{\boldsymbol{t}}_{ij}^{(m,n)}\right\|\right).$
  - $\vec{\boldsymbol{m}}_{ij}^{(m,n)} = \sigma_{\vec{x}}\left(\boldsymbol{m}_{ij}^{(m,n)}\right)\vec{\boldsymbol{x}}_j^{(n)} + \sigma_{\vec{v}}\left(\boldsymbol{m}_{ij}^{(m,n)}\right)\vec{\boldsymbol{v}}_j^{(n)}.$
- Aggregation and Update:
  - $\vec{\boldsymbol{m}}_{ij}^{(m)} = \overleftrightarrow{\boldsymbol{I}}_{ij}^{(m,n)} \vec{\boldsymbol{m}}_{ij}^{(m,n)} + \vec{\boldsymbol{t}}_{ij}^{(m,n)}.$



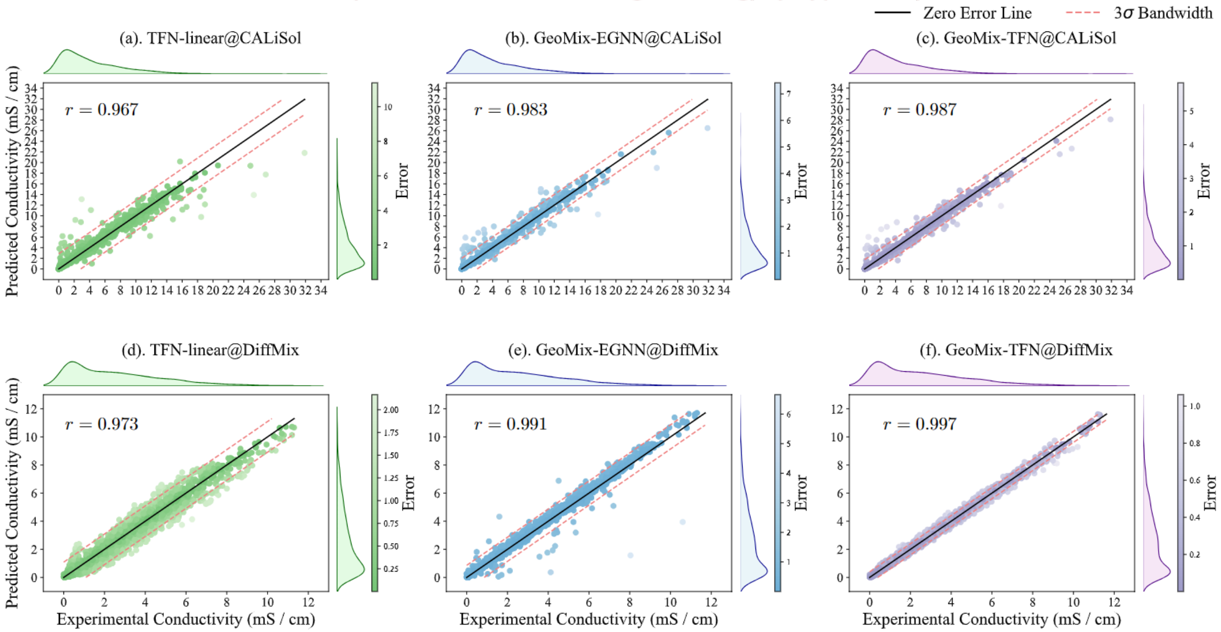
Dataset



(a)
CALiSol
DiffMix
13, 269 entries
Derived from experiments
Temperature range:
194.15-477.42K
38 solvents, 13 salts
Solvents: carbonates, ethers, esters, aromatics, halogenated and so on
Salts: fluoro salts, sulfonimide salts, sulfonate salts and so on
13 chemical elements
H, Li, B, C, N, O, F, Si, P, S, Cl, As
24, 822 entries
Derived from simulation
Temperature range:
243.15-293.15K
6 solvents, 2 salts
Solvents: linear and cyclic carbonates
Salts: $LiPF_6$ and LiTFSi
8 chemical elements
H, Li, C, N, O, F, P, S
(b)
4 solvents, 687 samples
3 solvents, 630 samples
1 solvent, 2233 samples
2 solvents, 2233 samples
CALiSol
1 solvent, 3420 samples
3 solvents, 8280 samples
2 solvents, 13122 samples
DiffMix
(c)
Conductivity Distribution in CALiSol
Conductivity Distribution in DiffMix
count (log scale)
Conductivity
(d)
CALiSol: t-SNE
DiffMix: t-SNE

Experiment

Models	CALiSol MSE ↓	CALiSol Pearson r ↑	DiffMix MSE ↓	DiffMix Pearson r ↑
MLP	3.657	0.906	1.363	0.874
MM-MoLFormer [12]	5.488	0.825	1.901	0.812
MolSets-Conv [13]	2.230	0.924	1.440	0.868
MolSets-SAGE [13]	2.751	0.909	0.708	0.937
EGNN-att [54]	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
GeoMix-EGNN	0.552	0.985	0.088	0.992
GeoMix-TFN	**0.432**	**0.987**	**0.035**	**0.997**



Ablation Study and Other Experiments

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

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**

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>

References

- [1] Soares, Eduardo, et al. "Capturing formulation design of battery electrolytes with chemical large language model." *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*. 2023.
- [2] Zhang, Hengrui, et al. "Molsets: Molecular graph deep sets learning for mixture property modeling." *arXiv preprint arXiv:2312.16473* (2023).
- [3] Zeng, Boshen, et al. "Uni-ELF: A Multi-Level Representation Learning Framework for Electrolyte Formulation Design." *arXiv preprint arXiv:2407.06152* (2024).