A Appendix

A.1 Dataset licenses

We list the different licenses of the dataset curated in CHEMIXHUB below:

- Miscible Solvent: CC BY-NC 4.0
- IlThermo: CC BY 4.0
- NIST TRC SOURCE Zenodo archive: CC BY 4.0
- Drug solubility: CC BY 4.0
- Solid polymer electrolyte: MIT
- Motor Octane Number: CC BY 4.0
- Olfactory Similarity: CC BY 4.0

A.2 Additional statistics on molecules for each of the 11 tasks in CheMixHub

Table 6: **Additional statistics on molecules for each of the 11 tasks in CheMixHub**.

Dataset	Tasks	Avg. # Atoms/Mol	Max # Atoms/Mol	Min # Atoms/Mol	Avg # Fragments	Max # Fragments	Avg Molecular Weight	Avg # Rotatable Bonds	Avg Components Mixture
Miscible solvents	ρ ΔH_{mix} ΔH_{vap}	8.28±3.17	18	3	1.0±0.0	1	123.73±43.96	3.40±3.17	3.72±1.08
IlThermo	$\ln(\kappa)$ $\ln(\eta)$	15.80±9.28 17.33±10.73	77 62	1 1	1.76±0.54 1.85±0.59	4 4	250.91±145.56 280.30±174.57	5.12±6.13 5.76±6.51	2.21±0.41 2.40±0.49
NIST Viscosity	$\ln(\eta_{\text{NIST-full}})$ $\ln(\eta_{\text{NIST}})$	12.90±8.98 9.12±4.71	95 63	1 1	1.50±0.70 1.0±0.0	8 1	203.98±135.28 140.52±73.17	4.00±5.60 3.14±3.79	1.88±0.33 1.92±0.28
Drug solubility	$\ln(S)$	14.48±9.16	51	1	1.11±0.33	3	212.40±128.17	2.37±2.45	1.91±0.29
Solid Polymer Electrolyte	$\ln(\kappa)$	30.86±47.75	676	2	1.24±0.44	3	473.36±738.30	18.11±33.19	2.24±0.67
Olfactory mixtures	Perceptual similarity	9.53±3.43	21	3	1.0±0.0	1	135.67±45.03	2.72±2.29	13.30±10.51
Fuel mixtures	MON	7.93±1.94	12	2	1.0±0.0	1	110.66±26.19	1.71±1.69	5.69±14.24

A.3 IlThermo Dataset curation details

We use the ILTHERMOPY package to retrieve IlThermo entries, selecting entries that are either binary or ternary mixtures and corresponding to our property of choice (for the scope of this paper, we limit ourselves to viscosity and ionic conductivity properties) [26]. We remove mixture that exhibits multiple phases behavior and are not liquid at the indicated temperature. We apply a natural logarithm transformation to the viscosity and ionic conductivity values present in IlThermo to make the range of values easier to learn. We also constrain the pressure range to be near the standard value of 1 atm or 101.325 kPa by applying a $\pm$2kPa threshold on pressure values.

We then standardize the mixture composition metric to mole fraction by converting as many entries as possible into that format. Data points which have the mixture composition expressed using molarity are discarded, as the conversion would require making assumption about the component densities. Assuming a binary mixture of component A and B with a given mole ratio $r_{A:B} = \frac{n_A}{n_B}$ where n_A and n_B are the number of moles of A and B, respectively, the mole fractions χ_A and χ_B can be calculated using:

$$\chi_A = \frac{n_A}{n_A + n_B} = \frac{r_{A:B}}{r_{A:B} + 1} \tag{2}$$

$$\chi_B = 1 - \chi_A \tag{3}$$

Similarly, assuming a ternary mixture of component A, B and C with given mole ratios $r_{A:B} = \frac{n_A}{n_B}$ and $r_{A:C} = \frac{n_A}{n_C}$, the mole fractions χ_A, χ_B and χ_C can be retrieved using:

$$\chi_A = \frac{n_A}{n_A + n_B + n_C} = \frac{r_{A:B}}{r_{A:B} + \frac{r_{A:B}}{r_{A:C}} + 1} \tag{4}$$

$$\chi_B = \frac{n_B}{n_A + n_B + n_C} = \frac{\frac{1}{r_{A:B}}}{\frac{1}{r_{A:B}} + \frac{1}{r_{A:C}} + 1} \tag{5}$$

$$\chi_C = \frac{\frac{1}{r_{A:C}}}{\frac{1}{r_{A:B}} + \frac{1}{r_{A:C}} + 1} \tag{6}$$

Assuming a binary mixture of component A and B, and given the mass ratio $r_{A:B} = \frac{m_A}{m_B}$ where m_A and m_B are the mass of A and B in g, respectively and the molecular weights MW_A and MW_B, to retrieve the mole fractions χ_A and χ_B, we first calculate mass fractions γ_A and γ_B using:

$$\gamma_A = \frac{m_a}{m_a + m_b} = \frac{r_{A:B}}{r_{A:B} + 1} \tag{7}$$

$$\gamma_B = 1 - \gamma_A \tag{8}$$

then assuming $m_{tot} = m_A + m_B = 1$g, we use $m_A = \gamma_A m_{tot}$ and $m_B = \gamma_B m_{tot}$ to obtain

$$n_A = \frac{m_A}{MW_A} \tag{9}$$

$$n_B = \frac{m_B}{MW_B} \tag{10}$$

$$n_{tot} = n_A + n_B \tag{11}$$

$$\chi_A = \frac{n_A}{n_{tot}} \tag{12}$$

$$\chi_B = 1 - \chi_A \tag{13}$$

The same process is naturally extended for ternary mixtures, assuming $r_{C:B} = \frac{m_C}{m_B}$ and MW_C are given.

Assuming a binary mixture of component A and B, and given the molarity $M_A = \frac{n_A}{m_B}$ where m_B is the mass of B in kg and n_A the number of moles of A and the molecular weights M_A and M_B, to retrieve the mole fractions χ_A and χ_B, we assume $m_B = 1$kg so $M_A = n_A$ and use $n_B = \frac{m_B}{MW_B}$ to obtain:

$$\chi_A = \frac{n_A}{n_A + n_B} = \frac{M_A}{M_A + \frac{1000}{MW_B}} \tag{14}$$

$$\chi_B = 1 - \chi_A \tag{15}$$

where a factor of 1000 is introduced since M_A and M_B are expressed in g/mol. The same process is naturally extended for ternary mixtures, assuming $M_C = \frac{n_C}{m_B}$ and MW_C are given.

Assuming a binary mixture of component A and B, and given the weight fraction γ_A and the molecular weights M_A and M_B, to retrieve the mole fractions χ_A and χ_B, we assume $m_{tot} = m_A + m_B = 1$g and use $m_A = \gamma_A m_{tot}$ and $m_B = \gamma_B m_{tot}$ to obtain:

$$n_A = \frac{m_A}{MW_A} = \frac{\gamma_A}{MW_A} \tag{16}$$

$$n_B = \frac{m_B}{MW_B} = \frac{\gamma_B}{MW_B} = \frac{1 - \gamma_A}{MW_B} \tag{17}$$

$$\chi_A = \frac{n_A}{n_A + n_B} \tag{18}$$

$$\chi_B = 1 - \chi_A \tag{19}$$

The same process is naturally extended for ternary mixtures, assuming γ_C and MW_C are given.

A.4 Details of molecular graph representation

The GNN takes in molecular graphs derived from the SMILES representations of molecules. Each graph, written as $G = (U, V, E)$, consists of a special global vertex U connected to all other vertices V, and a set of edges E. The global vertex U encodes overall properties of the molecule and is initialized with 200 normalized RDKIT descriptors obtained from DESCRIPTASTORUS [29]. The atoms of the molecules are the vertices (nodes), with node vectors $V = \{v_i\}_{i=1}^{N_v}$ for a molecule with N_v atoms, where v_i are feature vectors encoding atomic properties. Covalent bonds between atoms are represented as edges $E = \{(e_k, r_k, s_k)\}_{k=1}^{N_e}$ for a molecule with N_e bonds, where e_k are feature vectors of edge properties, and $r_k, s_k \in [1, \ldots, N_v]$ are indices of the two atoms that the bond joins together. Note $r_k \neq s_k$, since bonds must be between two different atoms

The node features used in the molecular graph representation as input to the GNN are 85-dimensional one-hot encoding vectors, encoding categorical information about the atoms. The edge features encode the categorical information about the bonds as 14-dimensional one-hot encoding vectors. The molecular information for the features are shown in Table 7.

Table 7: Features for node and edge features of molecular graphs
All categories are one-hot encoded and stacked to give a singular bit vector. `UNK` stands for "unknown", and is a catch-all category.

Node features	Categories
Atomic number	1 (hydrogen) to 54 (iodine), `UNK`
Atom degree	0, 1, 2, 3, 4, 5, `UNK`
Formal charge	-2, -1, 0, 1, 2, `UNK`
Chirality	unspecified, CW, CCW, other, `UNK`
Number of hydrogens	0, 1, 2, 3, 4, 5, 6, 7, 8, `UNK`
Hybridization	sp, sp2, sp3, sp3d, sp3d2, `UNK`
Aromatic	True/False
Edge features	**Categories**
Bond type	single, double, triple, aromatic, `UNK`
Is conjugated	True/False
In ring	True/False
Stereo-configuration	none, Z, E, *cis*, *trans*, any, `UNK`

As mentioned in Section 3.3, polymers and salts are present in the dataset and this probes important modeling considerations when employing GNNs. For polymers, we decided to restrict our modeling consideration to passing their monomeric units to the GNN. For salts, we conducted a chemical analysis to determine the impact of modeling the cation and anion as one disconnected graph. The details of it can be found in Section A.12.

A.5 Compute resources details

All model training/validation was conducted on a single A100 40GB NVDIA GPU.

A.6 Training details

Each run was performed for 500 epochs using the Adam optimizer [30], with a batch size set to 1024. Early stopping was implemented with patience set to 100. Two different learning rates were used to train the models end-to-end, one for the molecular-level model and one for the rest of the model. The splits used are specified in Section 3.4, further details on hyperparameter tuning can be found in Section A.7.

A.7 Hyperparameter search

For each task, the search was performed using Weights & Biases [5] with the BOHB algorithm [20] and a budget of 160 runs. 80 runs were allocated to the GNN-based molecular representations and 80 to CLMs and descriptors runs. Each run was performed for 500 epochs with early stopping patience set to 100. The search was conducted using the first split of the 5-fold random CV splits (70/10/20 training/validation/test split). The search space is defined as follows

- Molecular featurization: ["custom molecular graphs", "molt5 embeddings", "rdkit2d normalized features"]

- General hyper-parameters:
  - Loss type: ["mae", "mse"]
  - Dropout rate: [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5]
  - Learning rate (molecular level): [8e-5, 5e-5, 1e-4, 5e-4, 8e-4, 1e-3, 5e-3, 1e-2]
  - Learning rate (mixture level and head): [8e-5, 5e-5, 1e-4, 5e-4, 8e-4, 1e-3, 5e-3, 1e-2]
- Molecular-level hyper-parameters:
  - Molecular context aggregation type: ["concatenate", "multiply", "film"]
  - FiLM layer activation function: ["sigmoid", "relu"]
- Mixture-level hyper-parameters:
  - Mixture interaction module: ["self attention", "deepset"]
  - MLP head in self-attention: ["True", "False"]
  - Embedding dimension: [32, 64, 96, 128]
  - Number of layers: [0, 1, 2, 3]
  - Aggregation type: ["mean", "max", "pna", "scaled pna", "attention", "set2set"]
  - Number of attention heads: [1, 4, 8, 16]
  - Output dimension: [96, 128, 256]
  - Mixture context aggregation type: ["concatenate", "film"]
  - FiLM layer activation function (mixture context): ["sigmoid", "relu"]
- Predictive head hyper-parameters:
  - Embedding dimension: [64, 128, 192, 256, 320]
  - Number of layers: [1, 2, 3]

For runs where the molecular featurization used GNNs, the following additional parameters were added to the search space:

- GNN hyper-parameters:
  - Embedding dimension: [64, 128, 192, 256, 320]
  - Number of layers: [2, 3, 4]

A.8 XGBOOST modeling

The XGBOOST model was given a maximum of 1,000 estimators and tree depth of 1,000 except for the NIST-full task, where a maximum of 250 estimators and a tree depth of 250 was used. To ensure the model does not overfit, we use the validation set for early stopping, with a patience of 25 epochs. The model is trained with mean squared error, with a learning rate of 0.01.

A.9 Additional metrics for performances across tasks

In addition to the MAE results reported in Section 4.2, we report the results compiled from the CV splits for all models evaluated in terms of Pearson correlation coefficient ρ and Kendall ranking coefficient τ in Table 8 and 9, respectively.

Table 8: **Model performances across CHEMIXHUB tasks** Reported Pearson correlation coefficient ρ (↑) on 5-fold random CV splits. The mean and standard deviation are reported.

Molecular rep.	Mixture rep.	Miscible Solvents			Drug Solubility	SPE	NIST-full
		ρ	ΔH_{mix}	ΔH_{vap}	$\ln(S)$	$\ln(\kappa)$	$\ln(\eta)$
GNN	Attention	0.948 ± 0.076	**0.974 ± 0.003**	0.998 ± 0.000	0.993 ± 0.001	0.970 ± 0.004	0.980 ± 0.002
	Deepsets	**0.999 ± 0.000**	0.974 ± 0.004	0.851 ± 0.296	0.996 ± 0.001	0.969 ± 0.010	0.981 ± 0.002
MolT5	XGB	0.992 ± 0.001	0.924 ± 0.005	0.987 ± 0.001	**0.999 ± 0.000**	0.976 ± 0.001	0.989 ± 0.001
	Attention	0.998 ± 0.001	0.976 ± 0.003	0.997 ± 0.004	0.992 ± 0.005	0.973 ± 0.001	0.975 ± 0.038
	Deepsets	0.997 ± 0.001	0.976 ± 0.003	**0.999 ± 0.000**	0.983 ± 0.003	0.967 ± 0.002	0.977 ± 0.002
RDKit	XGB	0.992 ± 0.000	0.945 ± 0.007	0.986 ± 0.001	**0.999 ± 0.000**	**0.977 ± 0.001**	**0.992 ± 0.000**
	Attention	0.997 ± 0.001	0.972 ± 0.003	0.991 ± 0.003	0.996 ± 0.001	0.947 ± 0.008	0.995 ± 0.000
	Deepsets	0.996 ± 0.001	0.954 ± 0.005	**0.999 ± 0.000**	0.973 ± 0.004	0.963 ± 0.003	0.970 ± 0.002

Molecular rep.	Mixture rep.	IlThermo		MON	NIST	Olfaction
		$\ln(\kappa)$	$\ln(\eta)$	MON	$\ln(\eta)$	Mixture similarity
GNN	Attention	0.986 ± 0.008	0.975 ± 0.012	0.360 ± 0.300	0.990 ± 0.002	0.447 ± 0.120
	Deepsets	0.989 ± 0.002	0.939 ± 0.060	0.820 ± 0.087	0.990 ± 0.002	0.132 ± 0.103
MolT5	XGB	**0.997 ± 0.000**	**0.995 ± 0.001**	0.860 ± 0.032	0.950 ± 0.003	0.432 ± 0.047
	Attention	0.988 ± 0.001	0.993 ± 0.003	0.893 ± 0.028	**0.991 ± 0.001**	**0.559 ± 0.040**
	Deepsets	0.993 ± 0.002	0.986 ± 0.002	0.880 ± 0.036	0.981 ± 0.001	0.548 ± 0.025
RDKit	XGB	**0.997 ± 0.001**	**0.995 ± 0.001**	**0.913 ± 0.019**	0.957 ± 0.003	0.476 ± 0.062
	Attention	0.987 ± 0.003	0.981 ± 0.003	0.197 ± 0.351	0.977 ± 0.024	0.056 ± 0.130
	Deepsets	0.991 ± 0.002	0.992 ± 0.001	0.752 ± 0.155	0.980 ± 0.002	-0.091 ± 0.050

Table 9: **Model performances across CHEMIXHUB tasks** Reported Kendall ranking coefficient τ (↑) on 5-fold random CV splits. The mean and standard deviation are reported.

Molecular rep.	Mixture rep.	Miscible Solvents			Drug Solubility	SPE	NIST-full
		ρ	ΔH_{mix}	ΔH_{vap}	$\ln(S)$	$\ln(\kappa)$	$\ln(\eta)$
GNN	Attention	0.910 ± 0.091	0.835 ± 0.004	0.969 ± 0.002	0.932 ± 0.003	0.868 ± 0.016	0.904 ± 0.007
	Deepsets	**0.973 ± 0.000**	0.833 ± 0.003	0.816 ± 0.318	0.949 ± 0.004	0.869 ± 0.023	0.905 ± 0.002
MolT5	XGB	0.924 ± 0.00	0.730 ± 0.008	0.897 ± 0.003	**0.978 ± 0.001**	**0.899 ± 0.003**	0.950 ± 0.000
	Attention	0.963 ± 0.006	**0.835 ± 0.002**	0.955 ± 0.034	0.935 ± 0.022	0.881 ± 0.003	0.956 ± 0.003
	Deepsets	0.966 ± 0.002	**0.835 ± 0.002**	**0.976 ± 0.001**	0.893 ± 0.010	0.861 ± 0.004	0.910 ± 0.004
RDKit	XGB	0.929 ± 0.002	0.773 ± 0.005	0.898 ± 0.003	**0.978 ± 0.001**	0.899 ± 0.004	**0.966 ± 0.000**
	Attention	0.961 ± 0.003	0.829 ± 0.003	0.944 ± 0.012	0.948 ± 0.006	0.840 ± 0.014	0.957 ± 0.003
	Deepsets	0.956 ± 0.001	0.788 ± 0.008	0.973 ± 0.002	0.856 ± 0.009	0.855 ± 0.007	0.921 ± 0.002

Molecular rep.	Mixture rep.	IlThermo		MON	NIST	Olfaction
		$\ln(\kappa)$	$\ln(\eta)$	MON	$\ln(\eta)$	Mixture similarity
GNN	Attention	0.923 ± 0.021	0.941 ± 0.029	0.348 ± 0.203	0.940 ± 0.007	0.312 ± 0.073
	Deepsets	0.930 ± 0.006	0.863 ± 0.103	0.687 ± 0.093	**0.942 ± 0.009**	0.166 ± 0.067
MolT5	XGB	**0.974 ± 0.000**	0.967 ± 0.001	0.756 ± 0.038	0.863 ± 0.002	0.319 ± 0.047
	Attention	0.930 ± 0.004	0.967 ± 0.010	0.768 ± 0.033	0.939 ± 0.001	0.377 ± 0.042
	Deepsets	0.942 ± 0.004	0.945 ± 0.002	0.714 ± 0.012	0.896 ± 0.001	**0.390 ± 0.011**
RDKit	XGB	0.973 ± 0.001	**0.968 ± 0.001**	**0.781 ± 0.029**	0.883 ± 0.003	0.342 ± 0.040
	Attention	0.924 ± 0.015	0.957 ± 0.000	0.164 ± 0.266	0.916 ± 0.029	0.036 ± 0.065
	Deepsets	0.928 ± 0.005	0.956 ± 0.002	0.583 ± 0.121	0.897 ± 0.005	-0.048 ± 0.035

A.10 Additional transfer-learning benchmark

We evaluated transfer learning capabilities of two models trained on different datasets and tasks: the best deep learning model trained on the Miscible Solvent ΔH_{vap} task and the other one trained on the Motor Octane Number (MON) task (according to Section 4.2). We compare these fine-tuned models to the best performing models for these tasks found in Section 4.2 (see Table 2).
We observe a simple fine-tuning approach of the best Deep Learning models for each task on another task from a different dataset does not yield good performance, especially compared to "in-dataset" finetuning results above, which could suggest the models are overfitting to their respective tasks.

An interesting experimental set up to further answer this questions would be to evaluate multi-task learning capabilities of these models across datasets, which should be easily implementable thanks to our unified framework.

Table 10: **Transfer learning capabilities of models across the Miscible Solvent (MS) ΔH_{vap} task and the MON task.** Metrics are reported on 5-fold random CV splits. The mean and standard deviation are reported. The best model statistics are taken from Section 4.2 and Appendix A.9.

Fine-tuning Dataset	Best model Original Dataset	Pearson ρ (↑)	MAE (↓)	Kendall τ (↑)
MON	MON	0.913 ± 0.019	4.570 ± 0.348	0.781 ± 0.029
	MS-ΔH_{vap}	0.160 ± 0.108	33.199 ± 1.606	0.144 ± 0.056
Miscible Solvent ΔH_{vap}	MS-ΔH_{vap}	0.999 ± 0.000	0.071 ± 0.002	0.976 ± 0.001
	MON	0.501 ± 0.095	1.582 ± 0.095	0.296 ± 0.067

A.11 Additional benchmark

Table 11: **DiffMix tasks summary.** *T* indicates temperature dependency. *Mole Fractions* indicates mole fractions availability. *Arrhenius relationship* indicates if the task can be modeled using the Arrhenius equation. *Exp.* indicates if the data was obtained from wet-lab experiments or simulations.

Tasks		Units	Datapoints	Max # Components	# Unique Mixtures	# Unique Molecules	Mixture Context	Mole Fractions	Arrhenius Relationship	Exp.
DiffMix	κ	mS/cm	24,822	4	82	8	T	✓	✓	✗
	ΔV	cm^3/mol	1069	2	28	25	T	✓	✓	✓
	H_m^E	kJ/mol	631	2	34	35	T	✓	✓	✓

DiffMix (3 tasks) Battery electrolytes—mixtures of salts and solvents—have been optimized to facilitate ion transport, prevent electron transfer, and stabilize electrode-electrolyte interfaces to produce energy-dense and durable battery systems [72, 21]. The DiffMix dataset is a collection of three tasks centered around thermodynamic and transport properties predictions of electrolytes originally gathered by Zhu et al. [81]. This data is under the CC BY-NC-ND 4.0 license, and we therefore cannot include it as part of our dataset.

- **Excess molar enthalpy** H_m^E: The excess molar enthalpy reflects changes in intermolecular interactions that occur during the mixing of different components [77]. It shows the non-ideality of the final solution and gives an explanation about enthalpic effects [49]. In particular, differences in molecular shape, size, and interaction types between components—along with variations in temperature and pressure—can lead to either an increase or a decrease in excess molar enthalpy [36, 63]. DiffMix dataset includes 631 H_m^E data points curated from literature, covering 34 unique mixtures composed of 35 organic compounds across varying compositions. We rescaled the original range of the DiffMix excess molar enthalpy task from J/mol to kJ/mol to avoid passing big values to the neural networks.
- **Excess molar volume** V_m^E: The excess molar volume represents the deviation from ideal mixing volume. It exhibits a non-linear dependence on mole fraction [10] and temperature [66]—often showing a U-shaped trend with concentration and a decrease in absolute values as temperature increases. At higher temperatures, the dependence may shift to an S-shaped profile, making accurate prediction particularly challenging [71]. DiffMix dataset includes 1069 V_m^E data points curated from literature, covering 28 unique mixtures composed of 25 organic compounds.
- **Ionic conductivity** κ: The ionic conductivity of the electrolyte is known as a key parameter to evaluate the performance of the solution in practical engineering applications. In the context of batteries, κ changes considerably with the change of the electrolyte concentration [78]. DiffMix dataset includes 24,822 mixtures of single-salt-ternary-solvent electrolyte solutions generated using Advanced Electrolyte Model [22], and covering arbitrary combinations of two unique salts and six organic carbonate solvents at different concentration.

Table 12: **Model performances across CHEMIXHUB tasks** on 5-fold random CV splits. The mean and standard deviation are reported.

(a) MAE (↓)

Molecular rep.	Mixture rep.	DiffMix		
		κ	V_m^E	H_m^E
GNN	Attention	0.205 ± 0.061	0.060 ± 0.004	0.029 ± 0.006
	Deepsets	0.306 ± 0.054	0.072 ± 0.004	0.062 ± 0.014
MolT5	XGB	0.059 ± 0.002	**0.042 ± 0.007**	0.042 ± 0.004
	Attention	0.167 ± 0.164	0.056 ± 0.005	0.023 ± 0.003
	Deepsets	**0.046 ± 0.006**	0.062 ± 0.005	**0.021 ± 0.002**
RDKit	XGB	0.050 ± 0.001	0.045 ± 0.00	0.045 ± 0.006
	Attention	0.168 ± 0.064	0.079 ± 0.008	0.251 ± 0.123
	Deepsets	0.110 ± 0.011	0.074 ± 0.005	0.090 ± 0.065

(b) Pearson ρ (↑)

Molecular rep.	Mixture rep.	DiffMix		
		κ	V_m^E	H_m^E
GNN	Attention	0.993 ± 0.004	**0.950 ± 0.005**	0.996 ± 0.004
	Deepsets	0.984 ± 0.007	0.946 ± 0.007	0.982 ± 0.006
MolT5	XGB	0.998 ± 0.000	0.933 ± 0.023	0.989 ± 0.003
	Attention	0.994 ± 0.010	0.950 ± 0.008	0.998 ± 0.001
	Deepsets	**1.000 ± 0.000**	0.949 ± 0.009	**0.998 ± 0.000**
RDKit	XGB	0.999 ± 0.000	0.932 ± 0.026	0.983 ± 0.010
	Attention	0.995 ± 0.003	0.944 ± 0.005	0.422 ± 0.378
	Deepsets	0.998 ± 0.000	0.945 ± 0.006	0.964 ± 0.052

(c) Kendall τ (↑)

Molecular rep.	Mixture rep.	DiffMix		
		κ	V_m^E	H_m^E
GNN	Attention	0.929 ± 0.019	0.873 ± 0.023	0.928 ± 0.028
	Deepsets	0.887 ± 0.013	0.863 ± 0.026	0.852 ± 0.031
MolT5	XGB	0.973 ± 0.002	**0.901 ± 0.025**	0.909 ± 0.026
	Attention	0.948 ± 0.039	0.890 ± 0.022	0.957 ± 0.005
	Deepsets	**0.983 ± 0.002**	0.881 ± 0.023	**0.957 ± 0.002**
RDKit	XGB	0.980 ± 0.001	0.900 ± 0.025	0.903 ± 0.012
	Attention	0.945 ± 0.015	0.838 ± 0.045	0.472 ± 0.257
	Deepsets	0.954 ± 0.006	0.850 ± 0.027	0.828 ± 0.098

A.12 Modeling salts

Salts are often present in mixtures, these are non-bonded small molecules that are found in the same environment as the molecule. To explore how to properly model salts we first look at if they contribute meaningfully to basic featurizations.

We constructed a 200-dimensional molecular embedding space using RDKIT 2D descriptors obtained from DESCRIPTASORUS [29], incorporating both salts and fragments for all the tasks in CHEMIXHUB. The number of unique salts is 824, and the number of fragments is 476. This space was projected into two dimensions using UMAP to visualize structural relationships, Figure 5. As shown in the UMAP plot, The resulting plot shows that salts (blue triangles) and fragments (orange circles) broadly co-localize, with many salts embedded near fragment clusters. To quantify these observations, we computed cosine distances between each salt and the fragment-only descriptor space. The resulting

distribution confirms that the vast majority of salts lie within a narrow cosine distance range centered around 0.04–0.05, with very few exceeding 0.1, Figure 6. In RDKIT descriptor space, such low distances imply near-identity in structural features. From these, we can observe that most salts appear to retain descriptor-level similarity to their constituent fragments. However, there is a subset which introduces structural changes significant enough to shift them away from the fragment space.

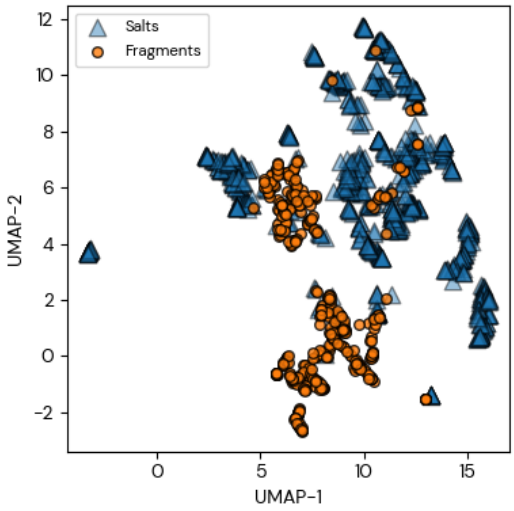

Figure 5: **The embedding space of salts and fragments in CHEMIXHUB**. UMAP projection of the combined RDKIT 2D descriptor space (200 dimensions) for salts and fragments. The embedding reveals well-defined structural clusters with apparent separation between salts and fragments, rather than overlap. Most salts appear in peripheral regions relative to the fragment clusters, suggesting distinct structural patterns at the descriptor level.

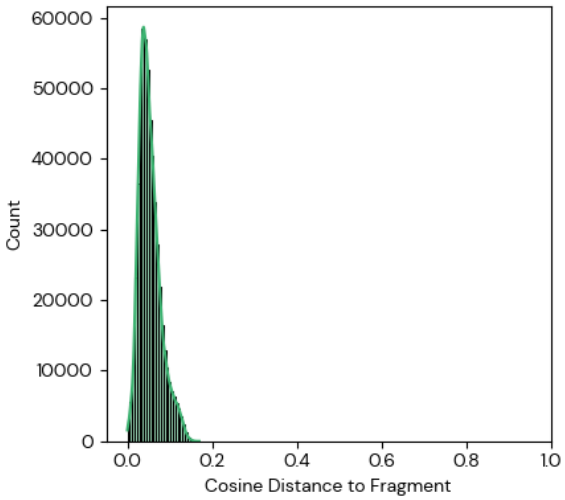

Figure 6: **Distribution of cosine distances**. The majority of salts fall within a tight cosine distance range (centered around 0.04–0.05), indicating strong structural similarity at the descriptor level. A smaller subset of salts shows higher distances, suggesting meaningful deviations from fragment-like chemistry.

Based on this analysis we conclude that most basic featurizations do not properly model salts. We think the best way to currently model salts is either as disconnected nodes in a graph. When using a GRAPHNETS architecture, these disconnected nodes get routed to the globals, so they are roughly equivalent to learnable salt-specific embeddings at the globals level of the graph.