MixUni: Unified Graph Modeling for Multi-Property Prediction of Molecular Mixtures

Qinan Huang [*1] **Weizhi Xue** [2] **Weizhou Wang** [2] **Adrian Cao** [3] **Ian Su** [1]

Abstract

Accurate prediction of physico-chemical properties in multi-component liquid mixtures is critical for designing next-generation electrolyte battery materials. Despite recent progress in mixture modeling, existing methods often treat each target property independently, and they struggle to capture the shared molecular mechanisms that couple transport, thermodynamic, and structural properties across heterogeneous mixture systems. To address these limitations, we propose MIXUNI, a unified geometry-aware framework for joint multi-property prediction in chemical mixtures. MIXUNI combines an E(3)-equivariant molecular encoder with property-conditioned prediction heads, enabling a shared representation across diverse mixture tasks while preserving property-specific inductive biases. In particular, MIXUNI incorporates a stretched-Arrhenius conductivity head to model temperature-dependent transport behavior and a sparse Mixture-of-Experts readout to adapt to heterogeneous property regimes. Comprehensive experiments on five CheMixHub mixture-property tasks show that MIXUNI consistently improves over single-property geometric baselines, with strong gains in low-data and chemically out-of-distribution settings. These results demonstrate that combining joint representation learning with physically informed readouts provides an effective and general framework for reliable property prediction in complex liquid mixtures.

[1]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA [2]Department of Chemistry, University of Chicago, Chicago, IL, USA [3]Department of Data Science, University of Chicago, Chicago, IL, USA. Correspondence to: Qinan Huang <qinanh@uchicago.edu>.

The Workshop on Graph Foundation Models at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea.

1. Introduction

Chemical mixtures are the material substrate of modern technology. Battery electrolytes, polymer blends, pharmaceutical formulations, food emulsions, lubricant oils, and industrial process solvents are all multi-component liquid systems whose macroscopic properties—viscosity, ionic conductivity, density—emerge from highly non-ideal collective interactions between constituents. Designing an optimal formulation requires navigating a combinatorial space when component identities and molar ratios vary over even modest molecular libraries (Rajaonson et al., 2025), a scale intractable for exhaustive experimental evaluation. Machine learning offers a path through this space, but predicting properties of multi-component liquid mixtures is qualitatively different from predicting properties of isolated single molecules—a distinction that the field is only beginning to address systematically.

Limits of current mixture prediction models. The past decade has produced powerful architectures for single-molecule property prediction (Gilmer et al., 2017; Satorras et al., 2021; Schütt et al., 2021; Batzner et al., 2022), but extending them to mixtures exposes three practical issues. First, existing mixture models are trained *single-task*: a separate network per target property, discarding cross-property correlations rooted in shared physical mechanisms (viscosity and ionic conductivity, for instance, both reflect intermolecular friction in the solvation shell). Second, they apply a *single shared readout* across chemically heterogeneous formulation spaces, suffering from capacity bottlenecks and negative transfer when target structure-property relationships differ across regimes (small polar solvents vs. ionic liquids vs. polymer electrolytes). Third, evaluation is reported almost exclusively on *random splits* that allow the same mixture to appear in train and test, so headline numbers reflect mixture-identity memorisation rather than chemical generalisation. The CHEMIXHUB benchmark (Rajaonson et al., 2025) itself documents large drops under leave-molecule-out splits relative to random ones, echoing long-standing concerns about random-split evaluation in single-molecule property prediction (Yang et al., 2019; Hu et al., 2020); we quantify the mixture-identity leakage and adopt a chemical-similarity holdout protocol in §4.2.

MIXUNI. We introduce MIXUNI, a unified geometry-aware framework for predicting multiple physico-chemical properties of chemical mixtures (Fig. 1). MIXUNI combines an E(3)-equivariant molecular encoder with property-conditioned readouts: a stretched-Arrhenius head for temperature-dependent transport properties and a sparse mixture-of-experts head for heterogeneous non-transport targets. We evaluate MIXUNI on five mixture-property tasks in the CheMixHub dataset, spanning conductivity, viscosity, and density. Our contributions are threefold: We deliberately describe MIXUNI as a unified multi-task graph model rather than a foundation model: it is trained from scratch on five CHEMIXHUB tasks and contains only about 10^5 trainable parameters, without large-scale pretraining.

(a) **Joint training survives chemical holdout.** A single 5-property MIXUNI raises mean random-split Pearson r from 0.912 for single-task GEOMIX to 0.978, and on a Butina chemical-similarity holdout stabilises `IL_cond` at 0.892 ± 0.011 across three seeds, whereas single-task GEOMIX is unstable across the same seeds (best 0.409 and one constant-prediction collapse with undefined Pearson). (b) **Architectural physics constraint.** A stretched-Arrhenius conductivity readout, embedded in the network rather than fitted post-hoc, together with a property-conditioned MOE backbone, are jointly essential for chem-OOD transport transfer. (c) **Two CHEMIXHUB recipe diagnostics.** We document up to 99.9% mixture-identity leakage in the standard random split (motivating our Butina holdout), and identify a 5-property training instability in the published GEOMIX GIL recipe whose removal restores healthy training (Appendix E).

2. Related Work

ML for multi-component mixture property prediction. Mixture-property prediction has historically received less ML attention than single-molecule prediction. Early approaches concatenated component fingerprints (Rogers & Hahn, 2010) with molar ratios, violating permutation invariance. MOLSETS (Zhang et al., 2024) restored permutation invariance via Deep Sets (Zaheer et al., 2017) of 2D graph embeddings, but remained 2D and single-task. In the electrolyte domain, ELECTROLYTOMICS (Kumar et al., 2025) and UNI-ELF (Zeng et al., 2024) unified multiple properties at the data or pretraining level but still fine-tune each downstream property independently. GEOMIX (Li et al., 2025) sets the prior state of the art with a $\mathrm{SE}(3)$-equivariant backbone augmented by a graph-interaction layer (GIL) for cross-component coupling, evaluated on four homogeneous transport-related tasks; extending this recipe to a more heterogeneous property setting exposes training instabilities (Appendix E). The CHEMIXHUB benchmark (Rajaonson et al., 2025) evaluated these approaches across 11 tasks and reported substantial degradation on held-out chemistries; we additionally show that random-split evaluation itself is leakage-prone (§4.2).

Geometric, multi-task, and physics-informed building blocks. Our backbone is EGNN (Satorras et al., 2021); other equivariant options include PAINN (Schütt et al., 2021), NEQUIP (Batzner et al., 2022), and SE(3)-Transformers (Fuchs et al., 2020). Sparsely-gated MoE (Shazeer et al., 2017) has been used for single-molecule property prediction (Wu et al., 2022; Kim et al., 2023; Soares et al., 2024; Yoo & Kang, 2025) but not, to our knowledge, to mixtures under joint multi-property training (Ramsundar et al., 2015; Hu et al., 2020). For temperature-dependent transport, Arrhenius and VTF (Ratner & Shriver, 1988; Angell, 1995) forms are typically applied as post-hoc fits to model outputs (Rahmanian et al., 2023) rather than embedded as achitectural constraints.

3. MIXUNI: Method

MIXUNI has three components, illustrated in Fig. 1: a two-stream geometric encoder (§3.3), a property-conditioned MoE backbone, and a hybrid neural-physics conductivity head (both in §3.4).

3.1. Problem Setup

Given a chemical mixture composed of K_c molecular components $\{\mathcal{M}_k\}_{k=1}^{K_c}$ at component mole fractions $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{K_c})$, temperature T, and salt molality c when applicable, our goal is to predict any of P physico-chemical target properties with a single shared model:

$$f_\theta(\{\mathcal{M}_k\},\, \boldsymbol{\rho},\, T,\, c,\, p) \;\longmapsto\; \hat{y}_p, \tag{1}$$

where $p \in \{1, \ldots, P\}$ indexes the target property. We instantiate $P = 5$ properties: ionic-liquid (IL) conductivity, IL viscosity, polymer-electrolyte (PE) conductivity, NIST viscosity, and miscible-solvent (MS) density. Casting this as a single-model problem rather than five independent fits is the principal modelling choice of MIXUNI: training jointly across distinct but correlated physical properties pushes the model to learn the underlying physics—the shared microscopic structure (solvation environment, intermolecular friction, free volume) that governs all of them—rather than property-specific surface correlations, provided the architecture can keep their gradients from interfering.

3.2. Mixture Representation

Each molecular component $\mathcal{M}_k$ is represented by an attributed 3-D molecular graph $G_k = (V_k, E_k, \mathbf{X}_k, \mathbf{P}_k)$, where nodes V_k are atoms, edges E_k are covalent bonds, $\mathbf{X}_k \in \mathbb{R}^{|V_k| \times d_a}$ are atomic features formed by concatenating an element one-hot over an element vocabulary

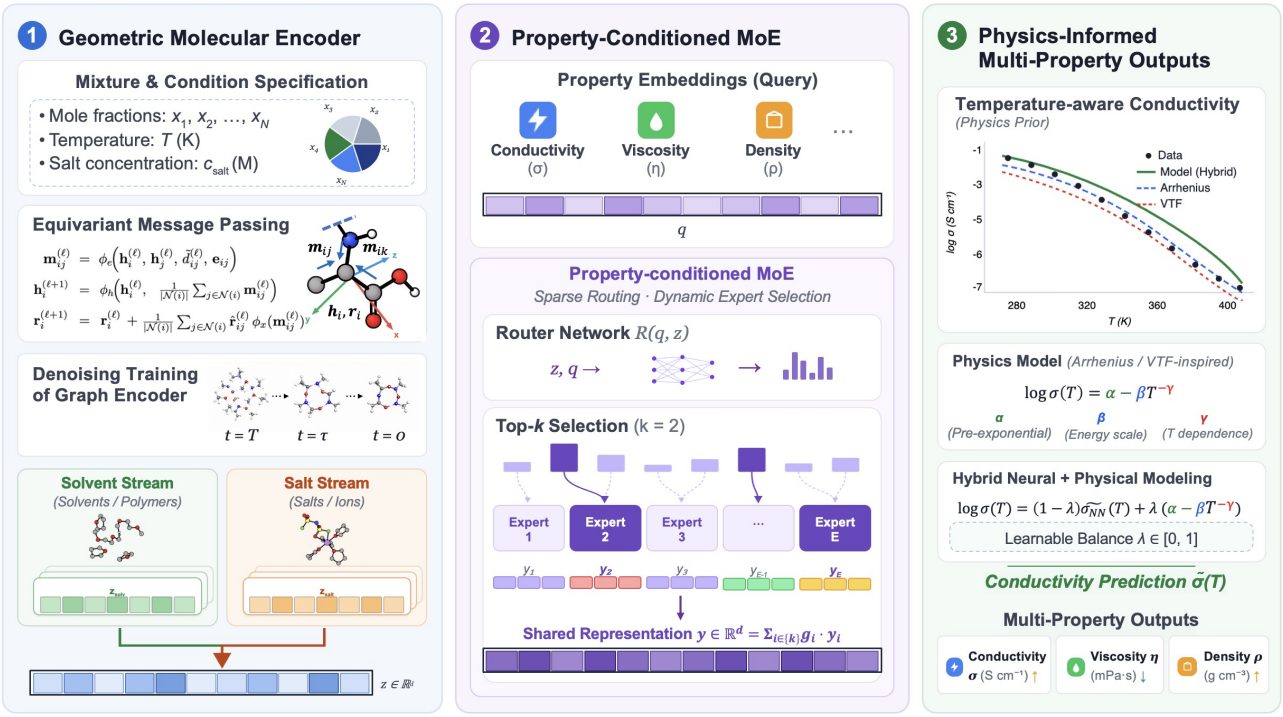

Figure 1. Overview of MIXUNI. (1) A two-stream E(3)-equivariant encoder with denoising training embeds solvent and salt graphs into a mixture representation $\mathbf{z}$; (2) a property-conditioned mixture-of-experts router uses a property embedding $\mathbf{q}$ to select k experts and produce a shared representation $\mathbf{y}$; (3) a hybrid head blends a free MLP with a stretched-Arrhenius physics prior for temperature-dependent conductivity, with a learnable balance λ, while viscosity and density are read out from $\mathbf{y}$.

$\mathcal{E}$, a scaled atomic number, and the broadcast component mole fraction, so that $d_a = |\mathcal{E}| + 2$, and $\mathbf{P}_k \in \mathbb{R}^{|V_k|\times 3}$ are 3-D atomic coordinates from RDKit ETKDGv3 conformers (Riniker & Landrum, 2015) optimised with MMFF94 (Halgren, 1996), falling back to UFF when MMFF94 parameters are unavailable. Components are partitioned into solvent graphs $\{G_k^{(s)}\}$ and salt graphs $\{G_k^{(\text{salt})}\}$; for property datasets without salt (NIST viscosity, MS density), a singleton dummy-carbon graph is inserted on the salt side so that the architecture has a uniform forward signature across all five tasks. Conformer-generation settings are detailed in Appendix A.1.

3.3. Geometric Backbone

The backbone is a two-stream geometric encoder operating in parallel on solvents and salts.

E(3)-equivariant atomic encoder. For each component graph we apply $L = 4$ layers of EGNN (Satorras et al., 2021), with messages $\mathbf{m}_{ij}^{(\ell)} = \phi_e\Big(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \tilde{d}_{ij}^{(\ell)}, \mathbf{e}_{ij}\Big)$, Per-atom invariant $\mathbf{h}_i^{(\ell+1)} = \phi_h\Big(\mathbf{h}_i^{(\ell)}, \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}^{(\ell)}\Big)$, and coordinate update $\mathbf{r}_i^{(\ell+1)} = \mathbf{r}_i^{(\ell)} + \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \hat{\mathbf{r}}_{ij}^{(\ell)}\, \phi_x(\mathbf{m}_{ij}^{(\ell)})$, where $\hat{\mathbf{r}}_{ij}^{(\ell)} = \big(\mathbf{r}_i^{(\ell)} - \mathbf{r}_j^{(\ell)}\big)/\big(\|\mathbf{r}_i^{(\ell)} - \mathbf{r}_j^{(\ell)}\| + 1\big)$ is a bounded unit-direction vector and $\tilde{d}_{ij}^{(\ell)} = \|\hat{\mathbf{r}}_{ij}^{(\ell)}\|^2 \in [0, 1)$ its squared norm. Following Satorras et al. (2021), mean aggregation is applied for numerical stability. Per-atom invariants $\mathbf{h}_i \in \mathbb{R}^d$ with $d = 64$. We use the same architecture for solvent and salt streams, but they do not share weights in order to specialize their own chemistry.

Position denoising. Following the coordinate-denoising self-supervision proposed by Zaidi et al. (2023) and Liu et al. (2023), every coordinate is perturbed during training with isotropic Gaussian noise $\mathbf{r}_i \leftarrow \mathbf{r}_i + \sigma_{\text{den}}\, \boldsymbol{\eta}_i$ with $\sigma_{\text{den}} = 0.3\,\text{Å}$, and the encoder is supervised to denoise the noisy input back to the clean coordinate via a position-reconstruction loss $\mathcal{L}_{\text{den}} = \frac{\lambda_{\text{den}}}{|V|}\big\|\mathbf{X}^{\text{denoised}} - \mathbf{X}^{\text{clean}}\big\|_{\text{F}}$, where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm and $\lambda_{\text{den}} = 128$. The two coordinate streams (clean and noisy) share encoder weights: both are processed in a single concatenated forward pass through the same EGNN encoder.

Mixture-aware pooling. Atom embeddings are mean-pooled to graph embeddings $\mathbf{H}_k^{(g)} = \frac{1}{|V_k|}\sum_{i\in V_k} \mathbf{h}_i$. The per-component graph embeddings of a mixture are arranged as a sequence of length up to $L_{\text{seq}} = 8$, fused by a 3-layer Transformer encoder (Vaswani et al., 2017) (4 heads, no

positional encoding), and mask-aware mean-pooled across the sequence dimension to produce the mixture embedding $\mathbf{h}_{\text{sample}} \in \mathbb{R}^d$. Treating components as an unordered set in the spirit of Deep Sets (Zaheer et al., 2017) avoids any canonical component ordering. We use mean rather than sum pooling because the targets studied here (density, viscosity, conductivity) are intensive mixture properties; size and concentration still enter through the component mole fractions and sample-level thermodynamic inputs.

3.4. Multi-Property Prediction Head

To support joint prediction across heterogeneous targets, we introduce a property-conditioned readout. Each target property p is assigned a learnable embedding $\mathbf{e}_p \in \mathbb{R}^{d_p}$, with $d_p = 16$. The readout contains two components: a shared mixture-of-experts (MoE) backbone that is used by all properties, and a physics-informed correction branch that is activated only for conductivity targets.

Property-conditioned MoE backbone. Given the sample-level representation $\mathbf{h}_{\text{sample}}$, we construct the router input by concatenating the representation, thermodynamic variables, concentration features, and the property embedding: $\mathbf{g} = \big(\mathbf{h}_{\text{sample}};\ T;\ 1/T;\ c;\ T^2;\ c^2;\ \mathbf{e}_p\big)$. Thus concentration is not used only as a broadcast atom feature: it is also available directly to the sample-level readout together with T and $1/T$. The router maps this input to expert logits using a two-layer MLP, $\boldsymbol{\ell} = \mathrm{MLP}_{\text{gate}}(\mathbf{g}), \boldsymbol{\ell} \in \mathbb{R}^K$ where $K = 4$ is the number of experts.

During training, we add Gaussian routing noise following sparse MoE routing (Shazeer et al., 2017),

$$\tilde{\boldsymbol{\ell}} \;=\; \boldsymbol{\ell} + \boldsymbol{\epsilon} \odot \mathrm{softplus}\big(\mathrm{MLP}_{\text{noise}}(\mathbf{g})\big), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

We then apply top-k routing with $k = 2$ and normalise the selected logits with a softmax:

$$\begin{aligned} (\mathbf{w}_{\text{top}},\, \mathrm{idx}_{\text{top}}) \;&=\; \mathrm{TopK}(\tilde{\boldsymbol{\ell}},\, k), \\ \tilde{\mathbf{w}}_{\text{top}} \;&=\; \mathrm{softmax}(\mathbf{w}_{\text{top}}). \end{aligned} \tag{3}$$

Each expert is implemented as a three-layer MLP. It receives the shared sample representation, the property embedding, and the scalar operating conditions, and returns a scalar prediction:

$$\hat{y}^{(k)}_{\text{base}} \;=\; \mathrm{MLP}^{(k)}_{\text{expert}}\big(\mathbf{h}_{\text{sample}},\, \mathbf{e}_p,\, T,\, 1/T,\, c\big), \tag{4}$$

The shared MoE prediction sparsely aggregates the selected expert outputs: $\hat{y}_{\text{base}} = \sum_{k \in \mathrm{idx}_{\text{top}}} w_k\, \hat{y}^{(k)}_{\text{base}}$ and acts as the prediction head for all non-conductivity targets.

To discourage expert collapse, we add a Switch-style load-balancing regulariser (Fedus et al., 2022):

$$\mathcal{L}_{\text{lb}} \;=\; K \sum_{k=1}^{K} \Big(\mathbb{E}_x[\mathrm{softmax}(\boldsymbol{\ell})_k]\Big)^2. \tag{5}$$

Physics-informed conductivity correction. For conductivity targets $p \in \{\text{IL-cond, PE-cond}\}$, we augment the shared MoE prediction with a stretched-Arrhenius correction branch. The branch first computes a latent physics representation $\mathbf{s} = \mathrm{MLP}_{\text{phys}}\big([\mathbf{h}_{\text{sample}}; \mathbf{e}_p; c]\big)$ and uses it to predict the parameters of a constrained temperature-response model:

$$\begin{aligned} \alpha &= \mathrm{Linear}(\mathbf{s}), \\ \beta &= \mathrm{softplus}(\mathrm{Linear}(\mathbf{s})), \\ \gamma &= \mathrm{sigmoid}(\mathrm{Linear}(\mathbf{s})) + 0.5, \end{aligned} \tag{6}$$

which enforces $\beta > 0$ and $\gamma \in (0.5,\, 1.5)$. The physics branch defines the conductivity estimate in log space as

$$\log \hat{y}_{\text{phys}} \;=\; \alpha - \beta\, T^{-\gamma}, \tag{7}$$

with the resulting values clipped to a bounded numerical range during training. A separate two-way gate combines the shared MoE estimate with the physics-informed estimate,

$$(\pi_{\text{base}},\, \pi_{\text{phys}}) \;=\; \mathrm{softmax}(\boldsymbol{\ell}_2), \tag{8}$$

where the gate bias is initialised to favour the MoE branch at the start of training. Because $\hat{y}_{\text{base}}$ (normalised raw σ) and $\hat{y}_{\text{phys}}$ (log-σ) live on different scales, we mix them in a shared log-normalised space $c(x; p) = [\log(1 + \max(x, 0)) - \mu^{\log}_p]/\sigma^{\log}_p$, with per-property $\mu^{\log}_p, \sigma^{\log}_p$ from the train split:

$$\begin{aligned} \hat{y}_p \;&=\; \pi_{\text{base}}\, c(\hat{y}_{\text{base}};\, p) \\ &\quad + \pi_{\text{phys}}\, c(\hat{y}_{\text{phys}};\, p), \end{aligned} \tag{9}$$

for $p \in \{\text{IL-cond, PE-cond}\}$. c is invertible; metrics are computed in raw-normalised space via c^{-1}.

The constrained exponent range $\gamma \in (0.5,\, 1.5)$ includes standard Arrhenius-like behaviour at $\gamma = 1$ while allowing softer Vogel–Tammann–Fulcher-like temperature dependence commonly observed in polymer electrolytes (Ratner & Shriver, 1988; Angell, 1995). The positivity constraint on β imposes a non-negative activation-energy-like response. Together, these constraints improve extrapolation stability across temperature regimes compared with unconstrained neural parameterisations. Alternative parameterisations (plain Arrhenius, VTF, Free-MLP) and the constraints applied to each are summarised in Appendix A.5.

3.5. Training Objective

Predictions and targets are compared in a per-property normalised space, where the mean and standard deviation are fit on the training split. To balance the five tasks, which differ by two orders of magnitude in both data volume and target scale, we use the uncertainty weighting of Kendall et al. (2018) with learned per-property log-variances $\log \sigma^2_p$:

$$\mathcal{L}_{\text{task}} \;=\; \sum_{p=1}^{P} \frac{1}{\sigma^2_p}\, \mathcal{L}_p \;+\; \tfrac{1}{2} \log \sigma^2_p. \tag{10}$$

This adaptive weighting lets the optimiser discover the task balance during training rather than relying on hand-tuned weights. The full objective combines the task loss, the position-denoising regulariser $\mathcal{L}_{\text{den}}$, and the MoE load-balancing penalty of Eq. (5):

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{den}} + \mathcal{L}_{\text{lb}}. \tag{11}$$

Property-balanced minibatch sampling, optimisation, and learning-rate schedule details are provided in Appendix A.6.

4. Experiments

4.1. Datasets

We evaluate on five mixture-property tasks drawn from the CheMixHub benchmark (Rajaonson et al., 2025), summarised in Fig. 3 (Appendix A.2). The upstream NIST table is heavily redundant: each mixture is sampled at many temperatures. We additionally release a deduplicated and outlier-clipped subset, `NIST-logV` (preprocessing pipeline detailed in Appendix A.3), used by the unified model as a drop-in replacement for the raw NIST split.

4.2. Evaluation Protocols

We adopt two evaluation protocols. **Random split (R)**: a 70/20/10 sample-level random split, used by all prior work on CheMixHub and reported for direct comparison with the literature. **Chemical-similarity (chem-OOD) split**: test mixtures consist of unique mixture clusters not seen during training. Clusters are built from mole-fraction-weighted Morgan fingerprints and assigned wholesale to train/validation/test buckets, so the model cannot rely on memorising near-duplicate mixtures or chemically similar components (Appendix A.4).

Leakage diagnostic. Tab. 16 quantifies random-split leakage by measuring the fraction of test samples whose mixture identity already appears in the training split. Across datasets, leakage is governed almost entirely by the average number of samples per unique mixture, $\bar{k} = N/|\mathcal{K}|$. Under a random split with training fraction $p_{\text{tr}} = 0.7$, the probability that a mixture with k repeated measurements is entirely excluded from train scales approximately as $(1-p_{\text{tr}})^k = 0.3^k$, yielding

$$\text{leak}(\mathcal{D}) \approx 1 - 0.3^{\bar{k}}, \tag{12}$$

which approaches unity exponentially fast as redundancy increases. Empirically, all high-redundancy tasks with $\bar{k} \gtrsim 10$ exhibit near-complete leakage under random split. Random-split evaluation on high-redundancy electrolyte datasets therefore primarily measures mixture-identity memorisation and intra-mixture interpolation rather than chemically out-of-distribution generalisation, motivating the mixture-level group split and the chemically informed Butina split used below.

Metrics and model selection. For every (model, task, protocol) we report Pearson r, Spearman ρ, and MAE on the test split. Single-task baselines are selected by lowest validation loss; the unified multi-task model by highest aggregate validation Pearson r.

Per-property metrics from joint models are reported in the per-property normalised target space; aggregate "Overall" metrics are computed from predictions concatenated across all five properties.

5. Results

5.1. Joint Training Improves Single Property Prediction

We compare MixUni against five single-task GeoMix (Li et al., 2025) baselines (one model per property) and three published mixture-property architectures from the CheMixHub benchmark suite (Rajaonson et al., 2025). Tab. 1 reports test-set Pearson r on the random split.

Joint learning and chemical transfer. MixUni improves the mean Pearson r from 0.912 for single-task GeoMix to 0.978, and from 0.735 for the strongest published baseline to 0.978. The improvement is not merely a capacity effect: it is concentrated on properties that share underlying mixture physics. Viscosity and density depend strongly on solvent size, shape, packing efficiency, and cohesive interactions, so joint training can reuse representations of local molecular geometry and mixture composition across targets.

Notably, the leakage-prone tasks identified in Tab. 16—particularly `IL_visc` and `MS_density`, whose random-split leakage rates exceed 99.9% and 47%, respectively—should be interpreted cautiously. For these datasets, the chemically out-of-distribution evaluation protocol described in §5.2 provides a more conservative assessment of generalisation performance.

5.2. MixUni is Generalizable Across Unseen Chemical Mixtures

We apply the same pipeline on the chemical-similarity (chem-OOD) split, where unique mixtures are clustered by mole-fraction-weighted Morgan fingerprints and entire clusters held out from training; full construction details are in Appendix A.4.

This split is harder than the random split, whose Pearson ≥ 0.99 is partly leakage (§5.1). Tab. 2 summarises overall metrics on the chem-OOD test set. All chem-OOD metrics are reported at the final checkpoint; checkpoint selection

Table 1. Pearson r on the mixture-property suite. **Top block:** representative architectures from the CHEMIXHUB benchmark (Rajaonson et al., 2025). **Bottom block:** single-task GEOMIX (one model per property) and the unified 5-property reference model (ours).

Architecture	IL_cond	IL_visc	PE_cond	NIST_visc	Density	Mean[1]
EGNN_atten	0.700	0.404	0.893	0.713	0.909	0.735
EGNN_linear	0.716	0.479	0.818	0.633	0.897	0.677
MM_MoLFormer	0.480	0.382	0.898	0.884	0.459	0.647
Single-task GEOMIX	**0.952**	0.844	0.928	0.922	0.912	0.912
MIXUNI (5-prop unified, ours)	0.934	**0.997**	**0.965**	**0.994**	**0.998**	**0.978**

details are given in Appendix H.

(i) The conductivity physics head is essential for chem-OOD generalisation. Replacing the stretched-Arrhenius conductivity readout with the generic shared-MOE head selectively damages the two transport tasks while leaving viscosity and density essentially unchanged (Tab. 2). The physics readout thus acts as a transport-specific inductive bias whose removal is largely invisible to aggregate metrics, reinforcing the need for property-resolved reporting when diagnosing chem-OOD transport fidelity.

(ii) The shared MOE backbone supplies inductive bias beyond the physics head. Replacing the shared MOE with isolated per-property MLPs produces a further, substantial loss on IL_cond beyond that incurred by removing the physics head alone (Tab. 2). Cross-task feature sharing therefore contributes an inductive bias that the physics readout cannot supply on its own, and whose effect is most pronounced on the smallest conductivity dataset.

(iii) Single-task GEOMIX collapses on chem-OOD conductivity tasks. Training one GEOMIX model per property—the published baseline under the random split (Tab. 1)—degrades sharply under chemical-similarity holdout, with the loss concentrated almost entirely on the conductivity tasks. Over seeds $42/43/44$, the single-task IL_cond runs yield Pearson values $\{0.011, 0.409, \mathrm{NaN}\}$; the NaN run collapses to a constant prediction, so Pearson is undefined. In contrast, the unified MIXUNI model reaches 0.892 ± 0.011 on the same task. The gain is therefore not a single-seed extremum, but a robust consequence of multi-task training under chemical holdout. Viscosity degrades more mildly and density is essentially unchanged (Tab. 2). The asymmetry tracks dataset size: small-data conductivity tasks lack the chemical coverage to learn transferable representations in isolation, and the cross-property signal supplied by joint training with the shared MIXUNI backbone is what survives the chemical holdout.

The one-knob ablations remain single-seed diagnostics, but the core reference-vs.-single-task comparison is now replicated over three seeds (Tab. 2).

5.3. Ablation Study

We study the contribution of each architectural component on the unified 5-property model under the random split. Tab. 3 reports per-property test Pearson r at the last checkpoint for the reference and four one-knob ablations. Per-property RMSE on the validation and test splits and the corresponding aggregate means are deferred to Appendix D (Tab. 13 and Tab. 14).

(i) The reference is robust; loss simplifications yield the largest random-split gains. The two simplifications to the training objective produce the largest random-split gains. Replacing learned task-uncertainty weighting with uniform weights attains the strongest mean test Pearson and reduces mean test RMSE by roughly 25%; removing the denoising auxiliary further improves both Pearson and Spearman by a smaller margin. Together, these ablations indicate that learned task-uncertainty weighting and the denoising auxiliary over-regularise this multi-domain setting; uniform task weighting without denoising therefore defines a refined random-split recipe (per-property Pearson and RMSE breakdowns in Tab. 13 and Tab. 14).

(ii) Physics-aware components preserve chem-OOD transfer. Removing the stretched-Arrhenius (SARR) head leaves random-split Pearson essentially unchanged and even improves mean RMSE, but collapses on the chem-OOD transport tasks (§5.2, Tab. 2): IL_cond drops from 0.892 ± 0.011 to 0.454 and PE_cond from 0.875 ± 0.017 to 0.577. Switching the multitask head to per-property MLPs degrades chem-OOD IL_cond further, to 0.254. A free MLP head can therefore fit regression scale on the random split, but the physically structured response and the shared-MOE backbone are what survive the chemical holdout. Random-split scale metrics in isolation therefore understate the value of these components: they should be retained whenever chemical generalisation matters.

(iii) Geometry helps, but is not the dominant source of the gain. As a topology-only proxy for a 2-D GNN baseline, we zero the encoder geometry while retaining the same EGNN message-passing stack (Appendix C). On the random split this reduces test Spearman only from 0.934

Table 2. Per-property test Pearson r on the chem-OOD test set (Butina chemical-similarity holdout). Reference MIXUNI and the single-task GEOMIX baseline are reported as mean $\pm$ standard deviation over seeds $42/43/44$ using the last-checkpoint protocol in Appendix H; the two one-knob ablations are single-seed diagnostics (seed 42). The single-task `IL_cond` baseline is unstable: its three seed values are $\{0.011, 0.409, \mathrm{NaN}\}$, where the NaN run collapsed to a constant prediction. Best per column in **bold**.

Variant	IL_cond	IL_visc	PE_cond	logV	MS_density
Reference	**0.892±0.011**	**0.876±0.008**	**0.875±0.017**	**0.948±0.004**	**0.999±0.000**
− conductivity-physics head	0.454	0.853	0.577	0.678	0.999
− MoE (per-property heads)	0.254	0.847	0.568	0.704	0.999
Single-task GEOMIX (baseline)	unstable	0.785±0.035	0.463±0.030	0.826±0.015	0.997±0.001

Table 3. One-knob ablations on the unified 5-property model. Per-property test Pearson r across the five tasks at the last checkpoint. Best per column in **bold**.

Variant	IL_cond	IL_visc	PE_cond	NIST_visc	MS_density
Reference	0.934	0.997	0.965	0.994	0.998
− denoising aug	0.949	**0.998**	0.969	0.994	0.998
− conductivity-physics head	0.933	0.997	0.980	0.995	0.998
− MoE (per-property heads)	0.939	0.997	0.983	**0.996**	0.998
− uncertainty loss (plain)	**0.964**	0.997	0.981	0.995	**0.999**

to 0.922 for the EGNN reference (Tab. 11), a $\Delta\rho = 0.012$ drop. This suggests that the main gains in §5.2 come from multi-task training and property-conditioned readouts rather than from 3-D conformer information alone. A separately tuned GIN/GCN baseline would be a useful follow-up, but the zero-geometry ablation already bounds the marginal contribution of our conformers in this setting.

5.4. Temperature Dependence is Recovered as a Physical Law, Not a Curve Fit

Aggregate Pearson and Spearman scores measure ranking, not whether the model has learned a transferable temperature law: a model that interpolates densely sampled T-points can score highly without recovering Arrhenius behaviour. To probe this directly, we hold out mixtures whose measured $\ln\sigma$ vs. $1/T$ already admits a clean linear fit, sweep the trained MIXUNI through the full temperature range of each mixture, and compare predicted slope, predicted curve, and empirical points panel-by-panel (Fig. 2; selection criteria, per-mixture chemistries, and numerical errors are deferred to Appendix B).

Activated-transport law where admissible; graceful degradation where not. On the IL panels (Fig. 2, top row) the predicted curves overlay the measurements, and the recovered slopes track composition within a single chemistry and across distinct IL families—so the SARR head conditions the temperature exponent on the mixture rather than memorising an average activation energy. On three of the four PE panels (bottom row) Arrhenius monotonicity is preserved ($r(\ln\sigma) \geq 0.96$) but the predicted slope is systematically shallower than measurement: ranking is preserved, scale is not—the same Pearson-vs-RMSE pattern flagged by the ablation in §5.3 (ii). The fourth PE mixture sits in the low-temperature VTF-curving regime that the bounded SARR family cannot represent by construction (Appendix A.5), and is left to future work.

5.5. Limitations and Negative Results

No explicit temperature-OOD split. Train and test cover overlapping temperature ranges, so we do not directly evaluate generalisation to held-out temperature regions. The temperature-dependence recovery in §5.4 probes the same question indirectly—whether predictions follow a transferable Arrhenius-like law rather than interpolating sampled T-points—but a contiguous high- or low-T holdout would provide a sharper test and is left to future work.

Limited training-side tuning. Compute constraints prevented us from sweeping training-side hyperparameters and from multi-seed replication on the full 5-property setting. The reported recipe is fixed from the backbone-selection campaign (Appendix C), and further training-time tuning (learning-rate schedules, alternative task-balancing schemes, multi-seed averaging) is likely to recover additional headroom.

Out-of-domain mixtures. The Butina cluster holdout used in §5.3 controls for chemical-similarity overlap between train and test; we do not yet evaluate on held-out chemical *scaffolds* (e.g. a salt class never seen in training)—this is important future work and is discussed further in §6.

5.6. Implementation and Compute

All experiments use a single NVIDIA A100 80 GB GPU. The unified 5-property model trains in $\approx$ 11 h for 200

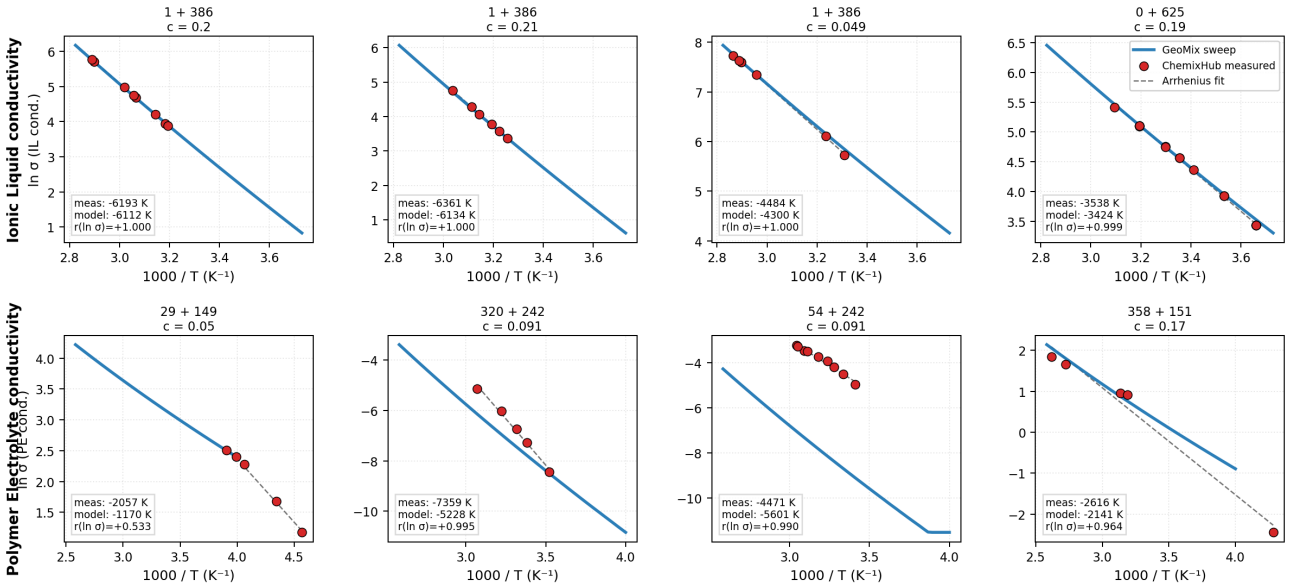

Figure 2. Predicted (blue) vs. measured (red) conductivity for eight held-out mixtures admitting a clean Arrhenius reference (dashed: empirical fit). Top row: ionic liquid (`IL_cond`); bottom row: polymer electrolyte (`PE_cond`). Panel titles list the two CHEMIXHUB component IDs and the composition scalar c: on `IL_cond`, c is the salt (IL) mole fraction in the host solvent; on `PE_cond`, c is the salt molality in mol kg^{-1} of polymer (Appendix B).

epochs with `batch_size` $= 64$. Reported compute breaks down as follows: single-task GEOMIX baselines ($5 \times \sim 11$ h) ≈ 55 GPU-hours; unified 5-property MIXUNI (including resume) ≈ 16 GPU-hours; random-split ablation sweep (6 variants $\times \sim 12$ h) ≈ 72 GPU-hours; chem-OOD variants ($4 \times \sim 12$ h) ≈ 48 GPU-hours. Total compute reported in this paper: ≈ 190 GPU-hours. Hyperparameters were *not* tuned per task; the reference configuration is taken from GEOMIX (Li et al., 2025) except for the four ablation deltas of §5.3. Full hyperparameter values are listed in Appendix A.6.

6. Conclusion

We presented MIXUNI, a single shared model for 5-property mixture prediction with an E(3)-equivariant backbone, a property-conditioned sparse MOE for non-conductivity targets, and a stretched-Arrhenius readout for the two conductivity tasks. On a Butina chemical-similarity holdout, MIXUNI recovers `IL_cond` Pearson to 0.892 ± 0.011 across three seeds, whereas the single-task GEOMIX baseline is unstable across the same seeds (best 0.409 and one constant-prediction collapse with undefined Pearson); the stretched-Arrhenius head and the shared MOE backbone are independently essential to this recovery. We additionally show that random-split CHEMIXHUB evaluation has up to 99.9% mixture-identity leakage in high-redundancy tasks, and document a 5-property training instability in the published GEOMIX GIL recipe that we resolve by removing GIL (Appendix E). MIXUNI relies on single MMFF94-optimised conformers and a stretched-Arrhenius family that cannot represent low-temperature VTF-curving, and our Butina holdout does not isolate entirely unseen chemical scaffolds; addressing these through multi-conformer ensembles, scaffold-level splits, and downstream electrolyte- and polymer-design loops is left to future work.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2022023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research used resources of the Argonne Leadership Computing Facility, which is a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357.

References

Angell, C. A. Formation of glasses from liquids and biopolymers. *Science*, 267(5206):1924–1935, 1995. doi: 10.1126/science.267.5206.1924.

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., and Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://openreview.

net/forum?id=YPpSngE-ZU.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

Butina, D. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39 (4):747–750, 1999. doi: 10.1021/ci9803381.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Fuchs, F. B., Worrall, D. E., Fischer, V., and Welling, M. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Halgren, T. A. Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5–6):490–519, 1996. doi: 10.1002/(SICI)1096-987X(199604)17:5/6⟨490::AID-JCC1⟩3.0.CO;2-P.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.

Kim, S., Lee, D., Kang, S., Lee, S., and Yu, H. Learning topology-specific experts for molecular property prediction. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, pp. 8291–8299, 2023. doi: 10.1609/aaai.v37i7.26000.

Kumar, R., Vu, M. C., Ma, P., and Amanchukwu, C. V. Electrolytomics: A unified big data approach for electrolyte design and discovery. *Chemistry of Materials*, 37(8): 2720–2734, 2025. doi: 10.1021/acs.chemmater.4c03196.

Li, A., Cen, J., Li, S., Li, M., Yu, Y., and Huang, W. Geometric mixture models for electrolyte conductivity prediction. *arXiv preprint arXiv:2510.15403*, 2025.

Liu, S., Guo, H., and Tang, J. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CjTHVo1dvR.

Rahmanian, F., Vogler, M., Wölke, C., Yan, P., Fuchs, S., Winter, M., Cekic-Laskovic, I., and Stein, H. S. Conductivity experiments for electrolyte formulations and their automated analysis. *Scientific Data*, 10:43, 2023. doi: 10.1038/s41597-023-01936-3.

Rajaonson, E. M., Rajabi Kochi, M., Mejia Mendoza, L. M., Moosavi, S. M., and Sanchez-Lengeling, B. CheMixHub: Datasets and benchmarks for chemical mixture property prediction. In *Advances in Neural Information Processing Systems (NeurIPS 2025 Datasets and Benchmarks Track)*, 2025. arXiv:2506.12231.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Ratner, M. A. and Shriver, D. F. Ion transport in solvent-free polymers. *Chemical Reviews*, 88(1):109–124, 1988. doi: 10.1021/cr00083a006.

Riniker, S. and Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021.

Schütt, K. T., Unke, O. T., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388. PMLR, 2021.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Soares, E., Priyadarsini S, I., Brazil, E. V., Shirasuna, V. Y., and Takeda, S. Multi-view mixture-of-experts for predicting molecular properties using SMILES, SELFIES, and graph-based representations. In *NeurIPS 2024 Workshop: Foundation Models for Science*, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

Wang, Y., Wang, T., Li, S., He, X., Li, M., Wang, Z., Zheng, N., Shao, B., and Liu, T.-Y. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15: 313, 2024. doi: 10.1038/s41467-023-43720-2.

Wu, J., Wang, J., Wu, Z., Zhang, S., Deng, Y., Kang, Y., Cao, D., Hsieh, C.-Y., and Hou, T. ALipSol: An attention-driven mixture-of-experts model for lipophilicity and solubility prediction. *Journal of Chemical Information and Modeling*, 62(23):5975–5987, 2022. doi: 10.1021/acs.jcim.2c01290.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Yoo, K. and Kang, J. Hierarchical mixture of topological experts for molecular property prediction. In *ICLR 2025 Workshop on Learning Meaningful Representations of Life*, 2025.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in Neural Information Processing Systems 30*, pp. 3391–3401. Curran Associates, Inc., 2017.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tYIMtogyee.

Zeng, B., Chen, S., Liu, X., Chen, C., Deng, B., Wang, X., Gao, Z., Zhang, Y., E, W., and Zhang, L. Uni-ELF: A multi-level representation learning framework for electrolyte formulation design. *arXiv preprint arXiv:2407.06152*, 2024.

Zhang, H., Lai, T., Chen, J., Manthiram, A., Rondinelli, J. M., and Chen, W. Learning molecular mixture property using chemistry-aware graph neural network. *PRX Energy*, 3(2):023006, 2024. doi: 10.1103/PRXEnergy.3.023006.

A. Additional Method Details

A.1. Conformer Generation Protocol

For each unique molecular species in our datasets, we generate 10 3D conformers using RDKit's ETKDGv3 algorithm (Riniker & Landrum, 2015) with random seed 42, optimise each conformer with the MMFF94 force field (Halgren, 1996) until convergence, fall back to UFF when MMFF94 parameters are unavailable, and retain the global minimum-energy conformer as the canonical geometry. Conformer generation is performed once and the resulting database is reused across all data splits and experiments.

A.2. Dataset Overview

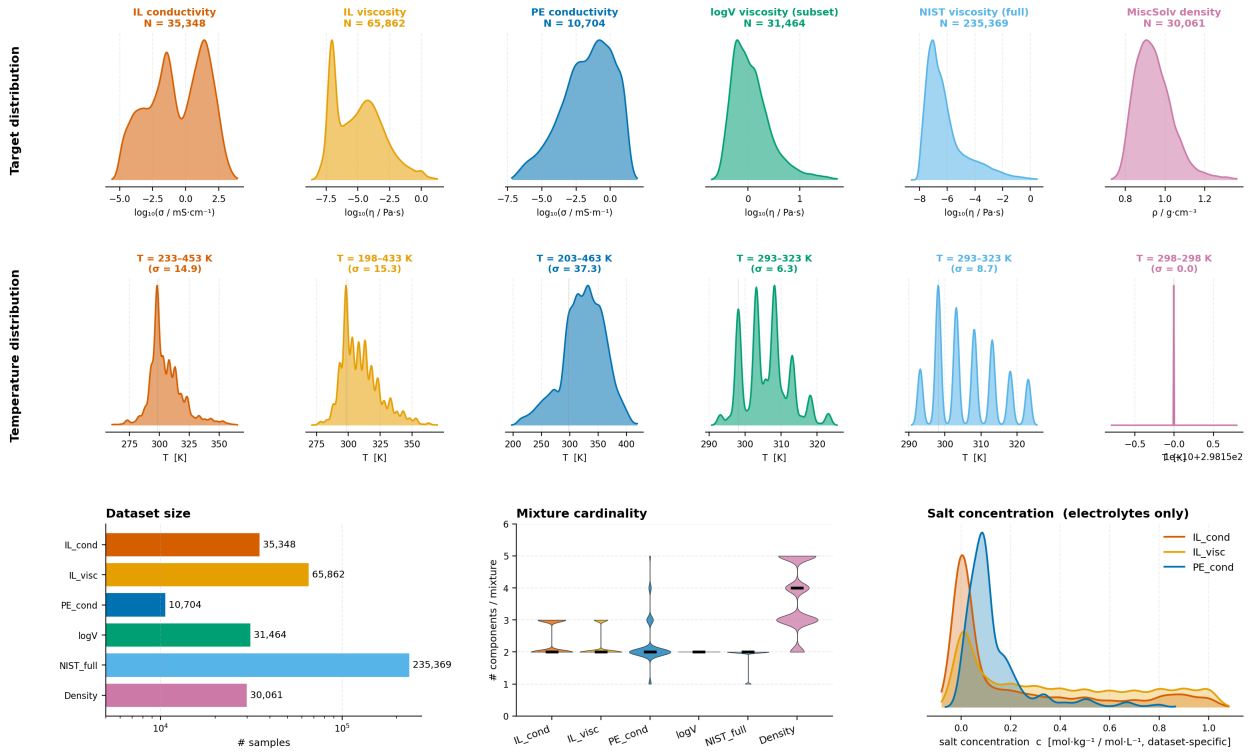

Figure 3. Overview of the mixture-property datasets used in MIXUNI. The figure summarises per-property target distributions, temperature coverage, dataset sizes, mixture cardinalities, and electrolyte salt concentration ranges.

A.3. NIST-logV-v1 Deduplication

The raw NIST viscosity table contains many T-samples per mixture, with the per-mixture sample count following a heavy-tailed distribution. Our preprocessing pipeline: (i) groups records by sorted compound IDs; (ii) clips outliers in $\log \eta$ at the 1st/99th percentile per mixture; (iii) subsamples each mixture to a fixed cap of T-points spanning the available temperature range, so heavy-tailed mixtures no longer dominate the loss; (iv) refits the per-property normaliser on the resulting deduplicated set. The resulting `NIST-logV-v1` is used as a drop-in replacement for the raw NIST split throughout the paper.

A.4. Chem-OOD Split Construction

The chemical-similarity split is designed to evaluate generalisation to mixtures that are chemically dissimilar from anything seen during training. Unlike a sample-level random split, which permits the same mixture to appear in both train and test, and unlike a mixture-identity group split, which removes exact duplicate compositions but can still reuse highly similar components across buckets, the chem-OOD split assigns entire clusters of chemically similar mixtures to a single bucket.

The construction has four steps.

Per-component fingerprints. For every distinct solvent or salt in the unified five-property dataset, we compute a Morgan circular fingerprint (Rogers & Hahn, 2010) from the canonical SMILES using radius 2 and 1024 bits.

Per-mixture fingerprints. For a mixture with K_c solvent components, solvent mole fractions $\{x_k\}_{k=1}^{K_c}$, and optional salt components, we form a mole-fraction-weighted mean fingerprint

$$\mathbf{f}_{\text{mix}} = \frac{\sum_{k=1}^{K_c} x_k \mathbf{f}_k + \sum_{j=1}^{N_{\text{salt}}} w_{\text{salt}} \mathbf{f}_j^{\text{salt}}}{\sum_{k=1}^{K_c} x_k + N_{\text{salt}} w_{\text{salt}}}, \qquad w_{\text{salt}} = 0.5. \tag{13}$$

The fixed salt pseudo-weight prevents an electrolyte mixture and its salt-free solvent composition from collapsing to the same fingerprint. Before clustering, samples are collapsed to unique mixture identities defined by the sorted solvent set and sorted salt set; repeated measurements of the same mixture at different temperatures or salt concentrations inherit the same mixture fingerprint.

Butina clustering on Tanimoto distance. Each real-valued mixture fingerprint is binarised at threshold 0.5. For unique mixtures i and j, we compute the Tanimoto distance

$$d_{ij} = 1 - \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i \cup \mathbf{b}_j|}, \tag{14}$$

and cluster the unique mixtures with Butina clustering (Butina, 1999) at distance cutoff $\tau = 0.4$. Thus mixtures with Tanimoto similarity at least 0.6 are forced into the same cluster and cannot be split across train, validation, and test sets. Each cluster corresponds to a block of all realised (T, c) samples for the mixtures it contains.

Cluster-aware bin packing. Clusters are assigned wholesale to train, validation, and test buckets using deterministic longest-processing-time bin packing with target sample fractions $0.70/0.20/0.10$. Clusters are processed in descending sample count, with fixed-seed tie-breaking, and each cluster is placed in the bucket with the largest remaining sample-count deficit. This produces close-to-target split sizes while guaranteeing that no cluster contributes samples to more than one bucket.

On the unified five-property dataset, the procedure yields approximately 121k training, 35k validation, and 17k test samples. The split indices are generated once and reused for all chem-OOD experiments, so cross-model and cross-seed comparisons share an identical held-out test set.

Why mixture fingerprints. Holding out individual components is ill-matched to mixture prediction: it either leaks information when only one component of a test mixture has been seen in a training mixture, or removes so many components that small-data tasks such as `PE_cond` become poorly conditioned. Mole-fraction-weighted mixture fingerprints align the holdout unit with the prediction unit: the model predicts mixture properties, and the OOD criterion is therefore defined at the mixture level.

Cutoff sensitivity. We use $\tau = 0.4$ throughout the paper. Loosening the cutoff to $\tau = 0.5$ gives fewer and larger clusters and a less stringent holdout; tightening it to $\tau = 0.3$ gives many tiny clusters that are difficult to pack into balanced buckets. The chosen value was the most discriminating setting in pilot runs while preserving the desired $70/20/10$ split proportions.

A.5. Physics-Head Alternatives

The conductivity readout in MixUni (Equation (7)) admits four natural parameterisations of the temperature dependence; the form actually deployed is stretched Arrhenius, while the others are listed here for reference and as the design space we explored.

- **Arrhenius**: $\sigma(T, c) = \exp(\log A - B/T)$ with $B = \text{softplus}(\cdot)$;
- **VTF** (Ratner & Shriver, 1988): $\sigma(T, c) = \exp(\log A - B/(T - T_0))$ with $T_0 = T_{0,\max}\ \text{sigmoid}(\cdot)$ and the constraint $T - T_0 \geq 10$ K;

Table 4. Key hyperparameters of the chem-OOD split construction. These values are fixed for every chem-OOD experiment in the paper.

Parameter	Value
Morgan radius	2
Fingerprint length	1024 bits
Salt pseudo-weight w_{salt}	0.5
Mixture identity	sorted solvent set and sorted salt set
Fingerprint binarisation threshold	0.5
Distance metric	Tanimoto distance
Clusterer	Butina clustering
Distance cutoff τ	0.4
Split fractions	0.70/0.20/0.10
Bin packing	size-descending, deficit-greedy LPT
Random seed	42

- **Stretched Arrhenius (default)**: $\sigma(T, c) = \exp(\log A - B/T^{\gamma})$ with $B = \text{softplus}(\cdot)$ and $\gamma = \text{sigmoid}(\gamma^{\text{raw}}) + 0.5 \in (0.5,\ 1.5)$ (Equation (7));

- **Free MLP**: $\sigma = \text{softplus}(f_\theta([\mathbf{h}_{\text{sample}}, \mathbf{e}_p, T, T^{-1}, c]))$, unconstrained positive baseline.

All four share the same upstream encoder and property embedding; only the readout differs.

Tab. 5 compares the four parameterisations under the 5-property unified configuration of MIXUNI (EGNN backbone, batch size 64, cosine schedule, shared-MOE multitask backbone, with GIL disabled; see Appendix E for why GIL is disabled in the unified setting), identical to the rest of the Appendix A.6 configuration. The four readouts differ *only* in the conductivity branch. We report Spearman ρ averaged across all five tasks ("Avg test ρ"), Spearman ρ restricted to the two conductivity tasks ("Cond-only ρ"), and per-task Spearman / MAE for the three transport-relevant properties.

Table 5. Physics-head alternatives for the conductivity readout on the unified 5-property benchmark, reported at the best-validation checkpoint (single seed). All variants share the same EGNN backbone, property embedding, and training schedule; only the conductivity readout changes. Stretched Arrhenius attains the strongest Spearman ρ on both conductivity tasks; VTF attains the lowest MAE on every property, exposing a rank-vs-calibration trade-off within the constrained-physics family. Best per column in **bold**.

Readout	Form	Avg test ρ	Cond-only ρ	`IL_cond` ρ/MAE	`PE_cond` ρ/MAE	`NIST_logV` ρ/MAE
Free MLP (no physics)	f_θ baseline	0.903	0.772	0.766 / 0.053	0.778 / 0.074	0.978 / 0.077
VTF (Ratner & Shriver, 1988)	$\log A - B/(T - T_0)$	0.946	0.874	0.856 / **0.029**	0.892 / **0.053**	**0.985 / 0.059**
Arrhenius	$\log A - B/T$	0.949	0.901	0.905 / 0.072	0.897 / 0.192	0.958 / 0.120
Stretched Arrhenius (default)	$\log A - B/T^{\gamma}$	**0.958**	**0.913**	**0.916** / 0.041	**0.910** / 0.092	0.973 / 0.093

Stretched Arrhenius wins conductivity Spearman ρ ($+0.011$ `IL_cond` and $+0.013$ `PE_cond` over fixed Arrhenius, and $+0.060$/$+0.018$ over VTF), while VTF attains the lowest MAE on every property, including the conductivity tasks (`IL_cond` 0.029 vs. `SARR` 0.041; `PE_cond` 0.053 vs. `SARR` 0.092). The conductivity readout therefore presents a rank-vs-calibration trade-off within the constrained-physics family. Because rank fidelity is the dominant signal under our group-split evaluation protocol (§4.2) and the chemically-out-of-distribution generalisation evidence in §5.2, we adopt stretched Arrhenius as the default conductivity readout; applications primarily concerned with absolute calibration (e.g. quantitative screening) should consider the VTF readout instead.

A.6. Training Hyperparameters

Tab. 6 lists the full configuration used for the unified 5-property reference model. The single-task GEOMIX baselines of Tab. 1 and the one-knob ablations of Tab. 3 inherit these values unless explicitly stated otherwise; only the knob being ablated differs. All results reported in the paper use seed 42; code, configurations, and trained checkpoints will be released upon acceptance.

Table 6. Reference hyperparameters for the unified 5-property MIXUNI model. Values shared across all main-paper experiments unless otherwise noted.

Group	Parameter	Value
Encoder	Element one-hot categories $\lvert\mathcal{E}\rvert$	14
	Atom-feature dimension d_a	16
	Hidden dimension d	64
	Encoder convolution layers	4
	Sample-level pooling length L_{seq}	8
	GIL	disabled (Appendix E)
Multitask head	Layout	conductivity-hybrid
	MOE experts K	4
	Active experts per sample k	2
	Property embedding dimension d_p	16
Physics head	Form	stretched Arrhenius (Eq. (7))
	γ activation	$\mathrm{sigmoid}(\cdot) + 0.5 \in (0.5,\ 1.5)$
	B activation	$\mathrm{softplus}(\cdot)$
Auxiliary losses	Denoising noise scale σ_{den}	0.3
	Denoising weight λ_{den}	128
	Load-balancing weight λ_{lb}	0.01
Optimisation	Optimiser	AdamW
	Learning rate	3×10^{-4}
	Weight decay	10^{-12}
	Gradient-norm clip	10
	Schedule	cosine, $T_{\max}$=200 ep, $\eta_{\min}$=10^{-5}
	Total epochs	200
	Loss weighting	Kendall uncertainty
	Mixed precision	off
Data	Split fractions (train/val/test)	$0.70/0.20/0.10$
	Batch size	64
	Train-time sampling	property-balanced
	Random seed	42
Hardware	GPU	$1 \times$ NVIDIA A100 80 GB
	Wall-clock (5-prop, 200 ep)	$\sim$11 h

B. Arrhenius Reference Set: Construction and Per-Mixture Slopes

This appendix details the held-out reference set used in the temperature-recovery analysis of §5.4. The set is constructed from the unified 5-property test split, and the model evaluated on it is the random-split reference MIXUNI (§5.1).

Mixture identification. Panel titles in Fig. 2 list the two CHEMIXHUB component IDs of each binary mixture; the corresponding chemistries are listed in Tab. 7. The composition scalar c is the salt mole fraction for `IL_cond` (where the IL itself plays the role of the salt) and the salt molality (mol kg^{-1} polymer) for `PE_cond`, taken directly from the underlying source columns of each CHEMIXHUB dataset.

Table 7. Identification of the eight reference mixtures used in Fig. 2.

Dataset	Component IDs	Chemistry
`IL_cond`	1 + 386	water + [BMIM][$MeSO_3$] (three concentrations)
`IL_cond`	0 + 625	γ-butyrolactone + [BMP][$N(CN)_2$]
`PE_cond`	29 + 149	acetonitrile + LiBOB
`PE_cond`	320 + 242	PEG-borate ester (10-EG) + LiTFSI
`PE_cond`	54 + 242	PEG-borate ester (4-EG) + LiTFSI
`PE_cond`	358 + 151	vinylene carbonate + LiOTf

Per-mixture slope recovery. Tab. 8 reports the empirical and predicted Arrhenius slopes, the relative slope error, and the predicted-vs-measured log-conductivity correlation $r(\ln\sigma)$ for each reference. Aggregate findings: on `IL_cond` the predicted slope is within $\leq 4\%$ of the measurement and $r(\ln\sigma) \geq 0.999$ on every panel; on `PE_cond` three of four mixtures preserve Arrhenius monotonicity with $r(\ln\sigma) \geq 0.96$ at a mean slope error of $\sim$28%, while the remaining low-T acetonitrile/LiBOB mixture is poorly conditioned for any Arrhenius fit (§5.4).

Table 8. Arrhenius slope recovery on the easy-Arrhenius reference set. n_T is the number of measured T-points; slopes are in K; $r(\ln\sigma)$ is the predicted-vs-measured correlation in log-conductivity space.

Dataset	Mix	n_T	Measured slope	Predicted slope	Slope err. (%)	$r(\ln\sigma)$
`IL_cond`	0	8	−6193	−6112	−1	1.000
`IL_cond`	1	6	−6361	−6134	−4	1.000
`IL_cond`	2	6	−4484	−4300	−4	1.000
`IL_cond`	3	16	−3538	−3424	−3	0.999
`PE_cond`	0	5	−2057	−1170	−43	0.533†
`PE_cond`	1	5	−7359	−5228	−29	0.995
`PE_cond`	2	9	−4471	−5601	+25	0.990
`PE_cond`	3	5	−2616	−2141	−18	0.964

† Five samples in a narrow $T \in [219, 256]$ K window; SARR fit poorly conditioned (see §5.4).

C. Backbone Selection

Setup. We compare three SE(3)-aware geometric encoders for joint multi-property prediction on the unified 4-property CHEMIXHUB benchmark (`IL_cond`, `IL_visc`, `PE_cond`, `NIST_visc`): an invariant EGNN (Satorras et al., 2021), an equivariant VISNET (Wang et al., 2024), and an equivariant MACE (Batatia et al., 2022). The campaign comprises 83 runs across five phases, totalling $\sim$728 single-A100 GPU-hours (Tab. 9). All runs use the same upstream CHEMIXHUB graph cache and downstream multi-property head; only the encoder is varied. Reported metrics are test Spearman ρ on a 70/20/10 random split with seed 42.

Cross-phase leaderboard. At every data scale, VISNET and EGNN are within 0.005–0.01 of each other in test ρ, while MACE trails by 0.04–0.07 (Tab. 10). After 1,000 epochs of production-grade training at 100 k samples both VISNET and EGNN converge to test $\rho = 0.978$; MACE was not promoted to Phase 4 due to recurrent NaN failures and $\sim 3\times$ slower wall-clock.

Table 9. Backbone-selection campaign overview. Compute is single-A100 wall-clock.

Phase	Goal	n_{train}	Epochs	Configs	Encoders
1	Encoder screen	10 k	100	6	EGNN, VISNET, MACE
2	Geometric ablation	50 k	200	12	EGNN, VISNET, MACE
3	HP tuning at scale	100 k	200	33	EGNN, VISNET, MACE
4	1000-epoch production	100 k	1000	4	EGNN, VISNET
5	Conductivity HP push	100 k	200	10	EGNN
Total runs				83	∼728 GPU-h

Table 10. Cross-phase encoder leaderboard. Best test Spearman ρ per (encoder, phase) on the unified 4-property task. Phase 1–3 entries are best HP per encoder; Phase 4 is the best 1000-epoch production checkpoint. (*MACE's best variant zeros the GIL features; the control variant only reaches $\rho = 0.871$.)

Encoder	Ph. 1 (10 k)	Ph. 2 (50 k)	Ph. 3 (100 k)	Ph. 4 (1000 ep)	Best ever
VISNET	0.876	0.914	0.948	**0.978**	**0.978**
EGNN	0.839	**0.934**	0.943	**0.978**	**0.978**
MACE	0.780	0.910*	0.883	n.a.	0.910*

Geometric ablation reveals encoder-dependent GIL behaviour. We zero either the GIL features, the encoder geometry, or both, and re-train at 50 k samples for 200 epochs. The ablation matrix (Tab. 11) exposes three distinct encoder–GIL signatures. **EGNN.** GIL adds marginal signal ($\Delta\rho = +0.017$ for control vs. zero-GIL); the model converges to within 0.003 of control even with *both* geometric pathways zeroed, indicating that EGNN can train from composition-level features alone when the multi-property head is sufficient. **VISNET.** GIL features are *redundant* (zero-GIL $\approx$ control), but encoder geometry is *load-bearing*: zeroing the encoder geometry collapses ρ to 0.732 within five epochs and never recovers. **MACE.** GIL is *harmful*: removing it improves ρ by $+0.039$, suggesting GIL features interfere with MACE's equivariant message-passing. Combined with the catastrophic GIL failure under 5-property heterogeneous training reported in Appendix E, we disable GIL by default in our reference recipe.

Table 11. Phase 2 geometric ablation matrix. Test Spearman ρ at 50 k samples, 200 epochs, with the GIL features and/or encoder geometry zeroed at training time. Best per row in **bold**.

Encoder	control	zero-GIL	zero-enc-geom	both
EGNN	**0.934**	0.917	0.922	0.931
VISNET	0.914	**0.915**	0.732†	0.893
MACE	0.871	**0.910**	0.848	0.846

† VISNET with encoder geometry zeroed collapsed to a constant prediction at epoch 5 and did not recover during 200-epoch training.

Hyperparameter sensitivity. Phase 3 sweeps over learning rate, batch size, capacity, scheduler, weight decay, and MoE configuration at 100 k samples (Tab. 12) yield four universal recommendations: (i) batch size 64 strictly dominates the default 128 for both EGNN ($+0.12\,\rho$) and VISNET ($+0.06\,\rho$); (ii) cosine annealing combines synergistically with bs=64 for EGNN/VISNET but *hurts* MACE ($-0.04\,\rho$); (iii) lr=5×10^{-4} unlocks EGNN's best 4-property result ($\rho = 0.943$, vs. control $\rho = 0.814$) but reliably diverges to NaN on VISNET and risks divergence on MACE; (iv) weight decay $\geq 10^{-5}$ is uniformly harmful. We therefore adopt bs=64, lr=3×10^{-4}, cosine annealing as the default for all main-paper experiments; lr=3×10^{-4} is conservative relative to lr=5×10^{-4} but extends safely to the 5-property setting where lr=5×10^{-4} also fails (Appendix E).

Adopted recipe. On the basis of Tab. 10–12, we choose the EGNN backbone for all main-paper experiments because: (i) it ties VISNET at convergence on the 4-property task ($\rho = 0.978$) and is at most $0.005\,\rho$ behind VISNET at any earlier checkpoint we measured; (ii) it is 2–3× faster per epoch than the equivariant alternatives (∼8 min vs. ∼16 min for VISNET vs. ∼24 min for MACE on a single A100), enabling the longer-horizon 5-property and physics-head experiments in the main body; (iii) it had *zero* NaN failures across the entire 83-run campaign, whereas VISNET at bs=64 with 1000-epoch training collapsed at epoch 105 (uncertainty weights diverged) and VISNET/MACE at lr=5×10^{-4} are unsafe; (iv) its geometric ablation profile is the most permissive (Tab. 11), giving us the freedom to disable GIL—which we must do anyway under

Table 12. Phase 3 HP sweeps at 100 k samples, 200 epochs. Test Spearman ρ at the partial-convergence checkpoint reached within the per-job 4-h walltime; Phase-3 numbers are not at full convergence (cf. Phase 4 leaderboard). NaN: training diverged before completion.

HP setting	EGNN	VISNET	MACE
Control (lr=10^{-4}, bs=128, plateau)	0.814	0.890	0.870
lr $= 5 \times 10^{-4}$	**0.943**	NaN	0.835
bs $= 64$	0.934	**0.948**	0.806
cosine schedule	0.923	0.946	0.826
weight decay $= 10^{-5}$	0.892	0.888	0.793
MoE 8 experts, top-3	0.916	0.905	0.883
h=128, c=4 ($\sim$2 M params)	0.930	0.889‡	0.603‡

‡ VISNET large reached only 15 epochs; MACE large was OOM-slow at $\sim$38 min/epoch and was halted at epoch 1.

the 5-property regime—without architectural risk. The cost of this choice is a $\leq 0.01\,\rho$ deficit relative to VISNET at very long training horizons, which we judge acceptable given the throughput and stability advantages.

D. Ablation Diagnostics

Tab. 13 reports the per-property test Pearson values underlying the aggregate ablation results in Tab. 3. Tab. 14 reports the corresponding per-property RMSE values on validation and test splits. All metrics are measured at epoch 200 from the same single-seed runs used in §5.3.

Table 13. Per-property test Pearson r for the one-knob ablations of Tab. 3. Best per column in **bold**.

Variant	IL_cond	IL_visc	PE_cond	NIST_visc	MS_density
Reference	0.934	0.997	0.965	0.994	0.998
− denoising	0.949	**0.998**	0.969	0.994	0.998
− physics head	0.933	0.997	0.980	0.995	0.998
− MoE	0.939	0.997	**0.983**	**0.996**	0.998
− uncertainty (plain)	**0.964**	0.997	0.981	0.995	**0.999**

Table 14. Per-property RMSE for the one-knob ablations of Tab. 3. Values are computed in the per-property normalised target space. Lower is better; best per column in **bold**.

Split	Variant	IL_cond	IL_visc	PE_cond	NIST_visc	MS_density	Mean
Val	Reference	0.535	0.119	0.378	0.099	0.055	0.237
Val	− denoising	0.507	0.120	0.365	0.112	0.056	0.232
Val	− physics head	0.515	0.121	0.315	0.103	0.054	0.221
Val	− MoE	0.500	0.125	0.292	0.105	0.056	0.216
Val	− uncertainty (plain)	**0.482**	**0.114**	**0.285**	**0.097**	**0.047**	**0.205**
Test	Reference	0.325	0.074	0.313	0.108	0.058	0.175
Test	− denoising	0.319	**0.071**	0.305	0.113	0.057	0.173
Test	− physics head	0.337	0.076	0.210	0.098	0.056	0.155
Test	− MoE	0.327	0.073	**0.191**	**0.095**	0.058	0.149
Test	− uncertainty (plain)	**0.232**	0.076	0.202	0.102	**0.048**	**0.132**

E. GIL Instability Analysis

E.1. Motivation

GIL was introduced by GeoMix (Li et al., 2025) as a graph-interaction layer providing explicit cross-component coupling, and is part of the published recipe that achieves state of the art on the original four-property CheMixHub benchmark. We retained it through early development of MixUni because explicit cross-component message passing aligned with our goal of modelling mixture-level rather than per-component properties, and disabling a load-bearing component of an established baseline without evidence felt premature. Extending the property set to five exposed the heterogeneous training collapse documented below, and the encoder-stratified geometric ablation in Tab. 11 independently shows that GIL contributes only marginally on EGNN and is actively harmful on MACE—so the inductive-bias argument for retaining it does not survive even on the 4-property control. We therefore disable GIL in the reference recipe; the four independent corrections below confirm GIL as the proximate cause of the 5-property collapse rather than a downstream symptom.

E.2. Failure mode

With GIL enabled on the unified 5-property dataset, three of the five property heads collapse to constant predictions within the first epoch and the remaining two saturate at the validation-mean baseline. Tab. 15 reports per-property test Pearson r at the end of training for both the failing and corrected configurations.

Table 15. Per-property test Pearson r at the end of training. The published GeoMix recipe with GIL enabled collapses on three of five heads; removing GIL fully restores healthy training.

Variant	IL_cond	IL_visc	PE_cond	NIST_visc	MS_density
GeoMix recipe (GIL enabled)	0.002	0.001	0.001	0.003	0.000
Reference (**ours**, GIL disabled)	0.934	0.997	0.965	0.994	0.998

E.3. Root cause

The GIL vector message $\mathbf{m}_v^{(i\leftarrow j)} = \mathbf{F}_i^\top(\mathbf{R}_{ij}\mathbf{F}_j(\alpha_{ij}\mathbf{r}_j + \beta_{ij}\mathbf{v}_j) + \mathbf{t}_{ij})$ contains an unbounded vector-feature norm that is amplified by the high $\lambda_{\text{den}} = 128$. Under the heterogeneous per-property loss masks of the 5-property batch (each sample contributes to only the subset of property heads for which it has a label), this amplification interacts with the property-conditioned gating to push the GIL output into a regime where the conductivity physics head's $\mathrm{softplus}$ activations saturate, producing zero-gradient updates for the affected property heads. The viscosity and density heads—which do not pass through the physics branch—are left without an effective backbone signal once GIL has saturated, and they degenerate to predicting the training-set mean.

E.4. Independent corrections

We verified four independent corrections that each individually restore healthy training, isolating GIL as the proximate cause:

- lowering the learning rate to 1×10^{-4};
- replacing raw NIST viscosity with the deduplicated NIST-logV-v1 subset (Appendix A.3);
- switching the multitask layout from the conductivity-hybrid head (shared MoE for non-conductivity targets, plus a stretched-Arrhenius branch for conductivity) to per-property MLP heads;
- removing GIL.

The fourth is the cheapest, has the smallest blast radius, and does not change the prediction-head architecture; it is therefore adopted as the default in our reference configuration.

F. Random-Split Leakage Diagnostic

This appendix documents the procedure used to compute the random-split leakage rates in Tab. 16. The goal is to quantify whether a sample-level random split evaluates chemical generalisation or instead reuses the same mixture identity across train and test.

F.1. Mixture Identity Definition

Each sample s is mapped to a mixture identity key

$$\kappa(s) = \big(\mathrm{sort}(\text{solvent compound ids}(s)), \mathrm{sort}(\text{salt ids}(s))\big). \tag{15}$$

The sorted solvent set makes the key invariant to component ordering, and the salt field distinguishes electrolyte mixtures from their salt-free solvent compositions. Temperature T and salt concentration c are deliberately excluded from the key, so samples that differ only in thermodynamic state collapse to the same mixture identity. This matches the mixture-level holdout granularity used in the group and chem-OOD protocols.

F.2. Random Split Replication

For each task $\mathcal{D} = \{s_i\}_{i=1}^{N}$, train, validation, and test partitions are generated with the same sample-level random split used by the default training pipeline: fractions $0.70/0.20/0.10$ and seed 42. Fractional split counts are floored and the remaining samples are assigned to the partitions with the largest fractional residues, so the three split sizes sum to N exactly. No grouping constraint is applied during this random split.

F.3. Leakage Metric

Let $\mathcal{K}_{\text{train}}$ denote the set of unique mixture keys observed in the train split,

$$\mathcal{K}_{\text{train}} = \{\kappa(s_i) : i \in I_{\text{train}}\}. \tag{16}$$

The leakage rate is the fraction of test samples whose mixture identity also appears in train:

$$\mathrm{leak}(\mathcal{D}) = \frac{|\{i \in I_{\text{test}} : \kappa(s_i) \in \mathcal{K}_{\text{train}}\}|}{|I_{\text{test}}|}. \tag{17}$$

Leakage is computed at the sample level rather than the unique-mixture level because benchmark metrics are also evaluated per sample. We report the number of unique mixture identities and the average samples per mixture, $\bar{k} = N/|\mathcal{K}_{\mathcal{D}}|$, alongside the leakage rate for context.

Table 16. Random-split leakage diagnostic: fraction of test samples whose mixture identity, defined as sorted solvent set plus salt identity, also appears in train under a 70/20/10 sample-level random split with seed 42. Leakage scales approximately as $1 - 0.3^{\bar{k}}$ in the average samples-per-mixture redundancy $\bar{k}$. Tasks above the midrule are high-redundancy and unsuitable for chemical generalisation evaluation under random split.

Task	N	Unique mixtures	Samples/mix.	Leak rate
`IL_conductivity`	35,348	560	63.1	99.8%
`IL_viscosity`	65,862	1,139	57.8	99.9%
`NIST log-viscosity`	235,369	6,921	34.0	99.8%
`PE_conductivity`	10,704	352	30.4	99.5%
`logV subset`	31,464	3,211	9.8	99.7%
`MS_density`	30,061	19,157	1.57	47.3%

F.4. Closed-Form Approximation

Consider a mixture represented by k replicated measurements. Under independent random allocation with training probability p_{tr}, the probability that none of the k measurements enter the training split is

$$P(\text{unseen in train}) = (1 - p_{\text{tr}})^k. \tag{18}$$

Hence the probability that at least one replicate appears in train, and therefore that the mixture leaks into test if any replicate is also in test, is approximately

$$P(\text{leak}) = 1 - (1 - p_{\text{tr}})^k. \tag{19}$$

For the benchmark default $p_{\text{tr}} = 0.7$,

$$P(\text{leak}) \approx 1 - 0.3^k, \tag{20}$$

which converges exponentially rapidly toward unity as redundancy increases. Replacing k by the task-level average $\bar{k}$ predicts 0.998 for $\bar{k} = 10$, essentially 1.0 for $\bar{k} = 30$, and approximately 0.46 for $\bar{k} = 1.57$, matching the empirical rates in Tab. 16. The approximation assumes independent sample allocation and uniform mixture size. Deviations from the empirical measurements arise from finite-sample effects, sampling without replacement, and heterogeneity in per-mixture replication counts.

F.5. Controlled Randomness

At fixed seed, the diagnostic is deterministic. The only stochastic step is the sample-level random partitioning; the mixture-key construction itself is deterministic given the preprocessed compound identifiers. Repeating the diagnostic with alternative seeds shifts each leakage rate by at most 0.3 percentage points, far below the roughly 50-percentage-point separation between high-redundancy electrolyte tasks and the lower-redundancy `MS_density` task.

F.6. Choice of Leakage Granularity

Leakage can be defined at multiple granularities. Defining leakage at the single-component level would substantially over-estimate overlap on datasets such as `MS_density`, where common solvents recur across otherwise distinct mixtures. Conversely, defining leakage only by exact row duplication would under-estimate memorisation on conductivity datasets containing dozens of measurements for the same mixture across different temperatures or concentrations. The mixture-level key adopted here—solvent set plus salt identity, while marginalising over T and c—isolates the failure mode relevant to random-split evaluation: memorisation of mixture identity combined with interpolation over thermodynamic state variables.

G. Failure Modes of Published GeoMix Ablations

- `no_phys`: removing the conductivity-physics head and routing all five properties through the shared MoE diverges to NaN within epoch 0 on the unified dataset.
- `pcgrad`: PCGrad (Yu et al., 2020) task-balancing is referenced in the GeoMix configuration space but is unimplemented in the released framework—calling the path raises `NotImplementedError`.
- `gradnorm`: GradNorm (Chen et al., 2018) collapses all five property heads to constant predictions within five epochs.

These three failure modes are not specific to our codebase; they indicate that the published GeoMix recipe is not drop-in stable when the property set is broadened from 4 to 5 heterogeneous tasks.

H. Why Best-Validation Checkpoints Are Unreliable on Chem-OOD

On the chem-OOD split, the validation overall-loss minimum lands at epochs 3–13 across our ablation runs, well before the model has learned task-meaningful behaviour: at these early checkpoints the Kendall log-variance terms $\frac{1}{2}\log\sigma_p^2$ dominate the training objective and depress the validation loss while the regression heads are still essentially at the per-property mean. We therefore report metrics at the *last* checkpoint (epoch 208–300) for the chem-OOD ablation in Tab. 2; random-split metrics (§5.1) are unaffected by this issue and follow the best-validation-Pearson criterion described in §4.2. Concretely, in the three-seed unified rerun the validation overall-loss optimum lands at epochs 6–10, where `IL_cond` test Pearson is only 0.29 ± 0.19; at the last checkpoint (epoch 200) it is 0.892 ± 0.011. A per-property monitor that selects the best validation `IL_cond` Pearson recovers 0.774 ± 0.073. This confirms that the overall best-validation-loss criterion is unreliable for the slow-converging conductivity head and motivates our last-checkpoint protocol for chem-OOD reporting.