# MIRA: PREDICTION-LEVEL MULTISCALE DECOMPOSITION FOR MOLECULAR MIXTURE PROPERTIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The property of a molecular mixture depends on its constituents, their proportions, and how they behave together. However, existing methods suffer from two key limitations: (1) interaction learning approaches typically fuse all signals into a single latent embedding before prediction, limiting interpretability and obscuring the separation between component-level and mixed-level contributions; (2) physically motivated decompositions offer clearer structure but often depend on property-specific equations or assumptions, restricting their transferability. To address these limitations, we introduce **Mixture Interaction and Residual Aggregation (MIRA)**, a unified property-agnostic framework for molecular mixture property prediction. MIRA performs prediction-level multi-scale decomposition by assigning explicit predictive roles to composition-weighted constituents, molecular interactions, and residual whole-mixture effects. This design separates component-wise, pairwise, and formulation-level contributions while preserving a common architecture across diverse properties. Across 11 CheMixHub tasks, MIRA achieves the lowest or tied-lowest mean test MAE at the reported precision, reducing five-fold MAE by up to 61.0%. It further generalizes to external mixture-property datasets, including surface tension and electrolyte-property prediction, demonstrating the broad applicability of its architectural design.

## 1 INTRODUCTION

Molecular mixtures are ubiquitous in electrolytes (Yang et al., 2026b), solvents (Leenhouts et al., 2025), and pharmaceutical formulations (Amiri & Khaleseh, 2025). Their properties depend on the identities and fractions of the constituent molecules as well as measurement conditions. While a composition-weighted reference of pure constituents captures part of the property variation, intermolecular interactions cause actual behavior to deviate from this baseline, with the magnitude of deviation depending on molecular pairings and mixing ratios. Such effects matter for mixing enthalpy, viscosity, ionic conductivity, and solubility. Predicting these properties therefore requires models that retain constituent information while learning how the response changes across compositions and formulations.

Existing methods introduce mixture structure through aggregation, explicit interactions, or physical constraints. Set-based models use composition-weighted or permutation-invariant pooling (Zhang et al., 2024; Leenhouts et al., 2025; Rajaonson et al., 2025), but typically predict from a single mixture representation. Interaction-based methods enrich that representation with interaction graphs, geometry, or cross-molecule communication (Qin et al., 2023; Li et al., 2025; Shi et al., 2026; Yang et al., 2026a). Physics-guided models expose reference and correction terms more explicitly, but often rely on property-specific equations or constraints (Rittig et al., 2023; Rittig & Mitsos, 2024; Zhu et al., 2024; Bai et al., 2026).

Together, these approaches point to a gap. Learned mixture representations can model interactions, but usually leave the predictive roles of constituents and interactions implicit; physically structured decompositions expose terms through forms that are often specific to a property. We ask whether one architecture can instead assign separate predictions to constituents, pairs, and the whole mixture, then learn their roles for each target without changing the form of the readout.

We introduce **MIRA** (**M**ixture **I**nteraction and **R**esidual **A**ggregation) to organize these scopes at the prediction level. A shared molecular encoder produces composition-neutral and composition-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

aware views that feed three predictive pathways. The *component pathway* produces a composition-weighted reference from individual constituents. The *pairwise pathway* scores unordered molecular pairs, with contributions gated by their co-presence. The *whole-mixture pathway* uses the complete formulation and available measurement conditions to refine the prediction. Each pathway returns a scalar, and the three outputs are combined only at the final readout. This *prediction-level decomposition* preserves permutation invariance without prescribing a different physical mixing law for each property. It also makes the trained model's allocation across molecular scopes directly observable.

Across all eleven CheMixHub tasks, MIRA achieves the lowest or tied-lowest five-fold mean test MAE at the reported precision and is further evaluated on five external datasets. Encoder controls, pathway analyses, and system-disjoint tests examine the source, behavior, and generalization of these gains.

Our main contributions are summarized as follows:

- **Prediction-level multiscale decomposition.** We propose a prediction-level decomposition in which component, pairwise, and whole-mixture pathways make separate scalar predictions that are combined only at the final readout.
- **Unified architecture with high predictive accuracy.** MIRA uses the same core structure across properties without property-specific mixing equations. It achieves the lowest or tied-lowest mean test MAE on all eleven CheMixHub tasks at the reported precision and lower primary error than the listed references on five external datasets.
- **Readout-level gains and inspectable pathway behavior.** Encoder-matched comparisons show that prediction-level separation preserves predictive performance while exposing component-, pairwise-, and whole-mixture behavior. System- and pair-disjoint tests further show improved generalization to new combinations of observed molecules.

## 2 Related Work

**Learned representations for molecular mixtures.** Mixture models commonly encode constituent molecules and combine their representations before prediction. Existing approaches use invariant or composition-aware pooling (Zaheer et al., 2017; Zhang et al., 2024; Leenhouts et al., 2025), explicit intermolecular interactions (Qin et al., 2023; Li et al., 2025; Shi et al., 2026; Yang et al., 2026a), or pretrained molecular representations adapted to multicomponent inputs (Wadell et al., 2026). These methods enrich the mixture representation, but contributions from different molecular scopes are not retained as separate prediction terms. MIRA instead keeps component, pairwise, and whole-mixture outputs separate until the final prediction.

**Physics-guided and structured mixture prediction.** Physical structure can also enter through a training objective or a prescribed prediction function. Thermodynamic models regularize activity coefficients using the Gibbs–Duhem relation (Rittig et al., 2023), derive properties from learned excess Gibbs free energy (Rittig & Mitsos, 2024), or impose pure-component boundary behavior (Specht et al., 2024). DiffMix predicts parameters of Redlich–Kister and Vogel–Fulcher–Tammann relations (Zhu et al., 2024), while electrolyte models incorporate domain-specific dependencies (Yang et al., 2026b). Mixture-PINN offers a different combination of interaction learning and physical guidance: it aggregates constituent and pair features into a mixture representation, then regularizes composition weights, pair symmetry, and temperature response where applicable (Bai et al., 2026). These approaches encode prior knowledge in response equations or training constraints, some of which must be chosen for the target property. MIRA instead builds its structural bias into the readout: composition weighting defines the constituent reference, a symmetric pair operator and abundance gate govern pair corrections, and the whole-mixture pathway captures formulation-level context. The same prediction structure applies across properties without specifying a new response equation or property-specific constraint for each one.

## 3 Method

Given a mixture's constituent molecules, composition fractions, and measurement conditions, MIRA decomposes the prediction into a composition-weighted component reference, a pairwise

correction, and a whole-mixture correction (Figure 1). Their scalar outputs are learned jointly and summed only at the final prediction.

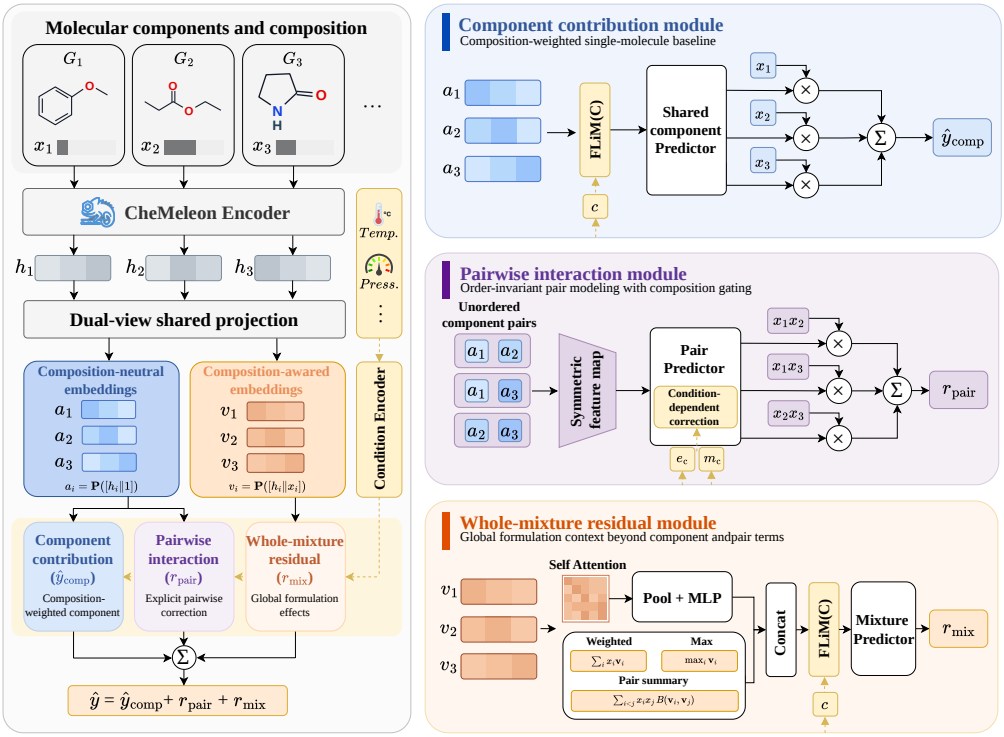


Figure 1: **Overview of MIRA.** Composition-neutral and composition-aware molecular views feed component, pairwise, and whole-mixture pathways. Measurement conditions modulate the predictive pathways, whose scalar outputs are summed only at the final prediction.

### 3.1 Shared Molecular Encoding and Composition Views

We represent a formulation as a set of molecular graphs and fractions, together with any available measurement conditions:

$$\mathcal{M} = \{(G_i, x_i)\}_{i=1}^{n}, \qquad x_i \geq 0, \qquad \sum_{i=1}^{n} x_i = 1, \tag{1}$$

$$\hat{y} = F_\Theta(\mathcal{M}, c) \in \mathbb{R}, \tag{2}$$

where $G_i$ and $x_i$ denote a molecular graph and its fraction, and $c$ collects conditions such as temperature. $F_\Theta$ is invariant to constituent ordering. A shared pretrained CheMeleon encoder (Burns et al., 2026) then maps each molecular graph to an initial molecular representation,

$$h_i = E_\theta(G_i). \tag{3}$$

The prediction pathways use two views of this representation. The component and pairwise routes start from molecular features independent of the observed fraction, whereas the whole-mixture route requires composition information before combining constituents. Both views are obtained from one shared affine projector $P_\phi$ by supplying either a fixed composition input of 1 or the observed fraction $x_i$:

$$a_i = P_\phi([h_i; 1]), \qquad v_i = P_\phi([h_i; x_i]). \tag{4}$$

Thus, $a_i$ stays fixed as $x_i$ changes and feeds the component and pairwise pathways, whereas $v_i$ varies with composition and feeds the whole-mixture pathway. Measurement conditions are introduced

only in the downstream predictive pathways and do not enter $E_\theta$ or $P_\phi$; implementation details appear in Appendix A.2.

### 3.2 Prediction-Level Multiscale Readout

The two input views feed three property-valued pathways. Each pathway first produces its own scalar prediction, and the three outputs are combined only at the final readout:

$$\hat{y} = \hat{y}_{\text{comp}} + r_{\text{pair}} + r_{\text{mix}}. \tag{5}$$

The component reference $\hat{y}_{\text{comp}}$, pairwise correction $r_{\text{pair}}$, and whole-mixture correction $r_{\text{mix}}$ can therefore be examined separately across compositions, although only their sum is supervised by the property label.

**Component prediction.** The component pathway predicts a contribution for each molecule without seeing its partners and takes the fraction-weighted sum. To account for conditions such as temperature, we apply feature-wise linear modulation (FiLM) (Perez et al., 2018) before prediction; the pairwise pathway uses the same conditioned features, while tasks without conditions bypass FiLM:

$$\widetilde{a}_i = \gamma_{\text{comp}}(c) \odot a_i + \beta_{\text{comp}}(c), \qquad \hat{y}_{\text{comp}} = \sum_{i=1}^{n} x_i f_{\text{comp}}(\widetilde{a}_i). \tag{6}$$

At fixed conditions, each contribution stays constant along a binary composition sweep, so the weighted component prediction is affine in composition. Interior peaks or troughs must therefore arise from the correction pathways.

**Pairwise correction.** The component reference cannot adjust a molecule's contribution for its partner. The pairwise pathway introduces this dependence while preserving exchange symmetry and contributing zero when either molecule is absent. Symmetric features handle ordering, whereas abundance weighting handles co-presence. For vectors $p, q$ and scalar $\delta$, we define

$$B(p, q, \delta) = [p + q;\ |p - q|;\ p \odot q;\ \delta]\,, \tag{7}$$

$$b_{ij} = B\big(\widetilde{a}_i, \widetilde{a}_j, (x_i - x_j)^2\big)\,, \tag{8}$$

where semicolons denote concatenation and the last entry records composition contrast. The squared difference preserves exchange symmetry. A shared predictor scores $b_{ij}$, with an additional condition-dependent residual when conditions are available. Let $e_c$ denote the encoded condition features and let $m_c \in \{0, 1\}$ indicate whether any valid condition is available:

$$s_{ij} = f_{\text{pair}}(b_{ij}) + m_c f_{\text{pair},c}([b_{ij}; e_c]), \qquad r_{\text{pair}} = \sum_{i<j} 4 x_i x_j s_{ij}. \tag{9}$$

The factor $4x_i x_j$ makes the correction zero when either constituent is absent and equal to one for an equimolar binary pair. Without conditions, FiLM is bypassed and the condition-dependent residual vanishes. In a binary mixture, the pairwise correction is symmetric about equal composition, whereas the whole-mixture pathway can additionally represent asymmetric responses.

**Whole-mixture correction.** A pair score sees only two constituents, whereas the whole-mixture pathway reads all composition-aware views $\{v_i\}_{i=1}^n$ to capture global formulation context and asymmetric binary responses. It combines two complementary summaries: self-attention exchanges information among constituents before pooling, while permutation-invariant statistics retain a weighted mean, feature-wise maxima, and an aggregated summary of pair relations:

$$z_{\text{attn}} = \rho_{\text{attn}}(\text{MeanPool}[\text{SelfAttn}(\{v_i\}_{i=1}^n)])\,, \tag{10}$$

$$z_{\text{inv}} = \rho\left[\sum_i x_i v_i;\ \max_i v_i;\ \sum_{i<j} x_i x_j\, \phi\big(B(v_i, v_j, (x_i - x_j)^2)\big)\right]. \tag{11}$$

Here $\rho_{\text{attn}}$, $\rho$, and $\phi$ are learned transformations. The final term in $z_{\text{inv}}$ is a vector-valued pair summary, distinct from the scalar pair correction in Eq. (9).

Measurement conditions may also affect the formulation-level response. We concatenate the two summaries, modulate them with FiLM, and predict a scalar correction:

$$z_{\text{mix}} = \gamma_{\text{mix}}(c) \odot [z_{\text{attn}}; z_{\text{inv}}] + \beta_{\text{mix}}(c), \qquad r_{\text{mix}} = f_{\text{mix}}(z_{\text{mix}}). \tag{12}$$

The final layer of $f_{\text{mix}}$ is initialized to zero, so this correction starts at zero. Unlike the pairwise correction, the whole-mixture output remains active at pure-component endpoints and can therefore refine the component prediction throughout the composition range.

All three outputs are permutation invariant through weighted constituent sums, symmetric unordered-pair sums, and pooled set representations, respectively. Equation (5) describes one compositional stream; for tasks with role-separated inputs, such as solute and solvent, we apply the same readout to each stream and combine them with a symmetric comparison adapter (Appendix A.2). Only the final prediction is supervised: individual pathways have no separate labels or auxiliary losses, and all parameters are trained jointly with Huber loss (threshold $0.5$) in standardized target space. We analyze the resulting pathway outputs in Section 4.4.

# 4 Experiments

We evaluate MIRA across 11 CheMixHub tasks, five external datasets, and unseen mixture systems. Encoder-matched comparisons separate the contribution of the molecular representation from that of the readout architecture. Pathway ablations, composition sweeps, and frozen-model interventions are then used to examine how the component, pairwise, and whole-mixture outputs contribute to the final prediction.

## 4.1 Experimental Setup

Our primary benchmark is CheMixHub (Rajaonson et al., 2025), with eleven tasks and official five-fold splits. External evaluation uses CALiSol and DiffMix (Li et al., 2025), the Bamboo-Mixer dataset (Yang et al., 2026b), MixtureSolDB mix-exp (Malikov et al., 2026; Leenhouts et al., 2026), and an independently collected surface-tension dataset. Table 1 includes published baselines, encoder-matched CheMeleon controls, and D-MPNN+MIRA (Yang et al., 2019); MolSets (Zhang et al., 2024) and MolPool (Leenhouts et al., 2025) are re-evaluated on the same CheMixHub folds. CheMixHub results report five-fold test MAE as mean $\pm$ sample standard deviation. Models are trained for up to 1,000 epochs, with a 50-epoch warmup followed by cosine decay. Training and dataset-specific protocols are given in Appendix A.

## 4.2 Accuracy Across Heterogeneous Properties

Table 1 compares MIRA with published baselines and encoder-matched controls on all eleven CheMixHub tasks. The same component-pair-whole-mixture readout is used for every property. Across all eleven CheMixHub tasks, MIRA achieves the lowest or tied-lowest five-fold mean test MAE at the reported precision among the models in Table 1. Relative to Bamboo-Mixer, MAE decreases from 0.054 to 0.030 for IL viscosity, from 0.068 to 0.043 for IL conductivity, and from 0.050 to 0.038 for vaporization enthalpy. The same architecture is used across all properties without introducing a task-specific mixing formula.

The encoder-matched comparison is clearest on IL viscosity: MAE falls from 0.106 with CheMeleon+DeepSets to 0.030 with MIRA. Replacing CheMeleon with D-MPNN retains much of this performance: D-MPNN+MIRA matches MIRA on Density and IL viscosity at the reported precision and remains close on the other tasks. This indicates that the benefit of the multiscale readout is not tied to a single molecular encoder.

Table 1: **Performance on the eleven CheMixHub tasks.** Five-fold test MAE (↓), reported as mean with sample standard deviation in parentheses; rounded zero standard deviations are shown as $< 0.001$. Values are shown to three decimals; complete results and reporting details are given in Appendix B. Abbreviations: DS = DeepSets, XGB = XGBoost, CheM = CheMeleon.

**(a) Density, thermodynamic, solubility, polymer, and conductivity tasks**

| Model | Density | $\Delta H_{\mathrm{mix}}$ | $\Delta H_{\mathrm{vap}}$ | Drug Sol. | SPE | IL Cond. |
|---|---|---|---|---|---|---|
| GNN+DS | 0.003 (<0.001) | 0.159 (0.002) | 0.406 (0.668) | 0.065 (0.005) | 0.315 (0.067) | 0.226 (0.017) |
| MolT5+XGB | 0.009 (<0.001) | 0.269 (0.004) | 0.306 (0.003) | 0.028 (0.001) | 0.222 (0.007) | 0.071 (0.001) |
| RDKit+XGB | 0.009 (<0.001) | 0.225 (0.005) | 0.295 (0.002) | 0.028 (0.001) | 0.223 (0.008) | 0.073 (0.002) |
| Bamboo | 0.003 (<0.001) | 0.165 (0.002) | 0.050 (0.018) | 0.012 (0.001) | – | 0.068 (0.017) |
| MIST | 0.003 (<0.001) | 0.163 (0.002) | 0.067 (0.001) | – | – | 0.644 (0.014) |
| CheM+DS | 0.005 (0.001) | 0.156 (0.003) | 0.055 (0.004) | 0.033 (0.003) | 0.275 (0.008) | 0.317 (0.043) |
| CheM+Pair | 0.003 (<0.001) | 0.198 (0.003) | 0.075 (0.001) | 0.025 (0.001) | 0.225 (0.009) | 0.465 (0.015) |
| MolSets(DMPNN) | 0.003 (<0.001) | 0.173 (0.004) | 0.053 (0.003) | 0.210 (0.304) | 0.314 (0.005) | 0.467 (0.022) |
| MolPool(MPNN) | 0.003 (<0.001) | 0.157 (0.003) | 0.043 (0.001) | 0.034 (0.003) | 0.185 (0.003) | 0.126 (0.018) |
| MIRA(DMPNN) | 0.003 (<0.001) | 0.160 (0.002) | 0.041 (0.001) | 0.014 (<0.001) | 0.171 (0.008) | 0.047 (0.004) |
| **MIRA(CheM)** | **0.003** (<0.001 | **0.154** (0.003) | **0.038** (0.001) | **0.011** (0.001) | **0.165** (0.007) | **0.043** (0.003) |

**(b) Viscosity, fuel, and perceptual tasks**

| Model | IL Visc. | NIST-full | NIST-small | MON | Olfactory |
|---|---|---|---|---|---|
| GNN+DS | 0.206 (0.020) | 0.131 (0.010) | 0.091 (0.005) | 5.990 (1.382) | 0.146 (0.010) |
| MolT5+XGB | 0.078 (0.002) | 0.148 (0.001) | 0.059 (0.001) | 5.002 (0.538) | 0.128 (0.006) |
| RDKit+XGB | 0.076 (0.002) | 0.055 (<0.001) | 0.048 (0.002) | 4.570 (0.348) | 0.125 (0.006) |
| Bamboo | 0.054 (0.003) | 0.029 (<0.001) | 0.016 (0.001) | 4.460 (0.556) | 0.125 (0.009) |
| MIST | 0.433 (0.005) | 0.309 (0.002) | 0.036 (0.001) | 4.028 (0.631) | 0.112 (0.008) |
| CheM+DS | 0.106 (0.035) | 0.088 (0.032) | 0.031 (0.010) | 4.382 (0.345) | 0.098 (0.005) |
| CheM+Pair | 0.104 (0.005) | 0.068 (0.001) | 0.023 (<0.001) | 4.010 (0.183) | 0.093 (0.009) |
| MolSets(DMPNN) | 0.120 (0.003) | 0.098 (0.002) | 0.037 (0.003) | 5.334 (1.638) | 0.120 (0.005) |
| MolPool(MPNN) | 0.076 (0.006) | 0.035 (0.001) | 0.017 (0.001) | 4.213 (1.272) | 0.106 (0.002) |
| MIRA(DMPNN) | 0.030 (0.001) | 0.028 (<0.001) | 0.015 (<0.001) | 3.659 (0.749) | 0.095 (0.005) |
| **MIRA(CheM)** | **0.030** (0.001) | **0.027** (0.001) | **0.013** (<0.001) | **3.558** (0.260) | **0.092** (0.002) |

### 4.3 Evaluation Beyond CheMixHub

We evaluate the same MIRA readout on five external datasets (Table 2). Their targets, data splits, and primary metrics differ, so each result is compared with the reference reported under the corresponding evaluation protocol.

Table 2: **Evaluation beyond CheMixHub.** Single-model results on each dataset's primary error metric. Gain is the relative error reduction from the listed reference, computed before rounding.

| | | | Primary error ↓ | | |
|---|---|---|---|---|---|
| Dataset | Metric | Reference model | Reference | MIRA | Gain ↑ |
| CALiSol | MSE ↓ | GeoMix-TFN | 0.432 | 0.243 | 43.8% |
| DiffMix | MSE ↓ | GeoMix-TFN | 0.035 | $1.40 \times 10^{-4}$ | 99.6% |
| Bamboo | log-MAE ↓ | Bamboo-Mixer | 0.058 | 0.054 | 6.7% |
| MixtureSolDB | RMSE ↓ | D-MPNN + FFN | 0.066 | 0.047 | 29.3% |
| Surface tension | MAE ↓ | DeepSets | 1.826 | 1.212 | 33.7% |

MIRA yields a lower primary error than the listed reference in all five settings. The surface-tension comparison holds the CheMeleon encoder and grouped split fixed; MAE falls from 1.826 to 1.212. The other datasets follow their own benchmark protocols. Notably, the MixtureSolDB reference trains with available pure-compound observations, whereas MIRA uses mixture rows only. DiffMix uses simulated formulations and a random split; its very small error should therefore be interpreted as interpolation within the simulated composition-property surface. Appendix A reports split definitions, fold-level uncertainty, and ensemble results.

### 4.4 Pathway Complementarity and Analysis

**Pathway complementarity.** Table 3 reports models retrained with individual pathways removed or isolated. These ablations measure the predictive value of each pathway while allowing the remaining architecture to adapt. Removing the whole-mixture pathway causes the largest degradation on several tasks. On IL conductivity, MAE rises from 0.043 to 0.458 after the remaining pathways are retrained, and IL viscosity shows a similar dependence. IL conductivity also degrades substantially without the component or pairwise pathway, indicating that the three pathways provide complementary predictive information. Drug Solubility behaves differently: the pairwise-only model matches Full MIRA at the displayed precision, whereas the unstructured control performs poorly.

Table 3: **Ablation of the three MIRA predictive pathways.** Test MAE (↓), reported as mean with sample standard deviation in smaller parentheses under the matched ablation protocol; rounded zero standard deviations are shown as $< 0.001$. Full MIRA is retrained under the matched ablation protocol, so its values may differ slightly from those in Table 1.

| Variant | IL Visc. | NIST-small | IL Cond. | $\Delta H_{\mathrm{mix}}$ | Drug Sol. |
|---|---|---|---|---|---|
| **Full MIRA** | 0.030 (0.001) | **0.013** (**<0.001**) | **0.043** (**0.003**) | **0.154** (**0.003**) | 0.012 (0.001) |
| *Leave-one-pathway-out* | | | | | |
| w/o Component | 0.030 (0.001) | 0.014 (<0.001) | 0.169 (0.009) | 0.157 (0.003) | 0.015 (0.001) |
| w/o Pairwise | 0.032 (0.001) | 0.015 (0.001) | 0.210 (0.033) | 0.159 (0.003) | 0.014 (0.001) |
| w/o Whole | 0.102 (0.002) | 0.022 (0.001) | 0.458 (0.015) | 0.198 (0.003) | 0.012 (0.001) |
| *Single-pathway and unstructured controls* | | | | | |
| Component only | 0.293 (0.004) | 0.054 (0.001) | 0.906 (0.021) | 0.446 (0.005) | 0.014 (0.001) |
| Pairwise only | 1.317 (0.031) | 0.137 (0.003) | 1.094 (0.013) | 0.204 (0.004) | 0.012 (0.001) |
| Whole only | 0.033 (0.001) | 0.015 (<0.001) | 0.194 (0.028) | 0.158 (0.003) | 0.014 (<0.001) |
| None | 0.086 (0.023) | 0.024 (0.004) | 0.324 (0.017) | 0.160 (0.004) | 0.239 (0.011) |

**Property-dependent pathway allocation.** The ablations compare separately trained models. To examine how the pathways are used within Full MIRA, we freeze each model and sweep 51 compositions for 20 binary systems from mixing enthalpy, NIST-small viscosity, and surface tension, keeping measurement conditions fixed within each system. Figure 2(a) summarizes the relative output magnitudes.

For each binary system, let $\mathcal{X}$ be its 51-point composition grid. We measure the relative magnitudes of the three outputs in Eq. (5) directly:

$$(C_{\mathrm{comp}}, C_{\mathrm{pair}}, C_{\mathrm{mix}}) = \frac{\left(\sum_{x\in\mathcal{X}} |\hat{y}_{\mathrm{comp}}(x)|,\ \sum_{x\in\mathcal{X}} |r_{\mathrm{pair}}(x)|,\ \sum_{x\in\mathcal{X}} |r_{\mathrm{mix}}(x)|\right)}{\sum_{x\in\mathcal{X}} \left(|\hat{y}_{\mathrm{comp}}(x)| + |r_{\mathrm{pair}}(x)| + |r_{\mathrm{mix}}(x)|\right)}. \tag{13}$$

The whole-mixture output has the largest median relative magnitude for mixing enthalpy ($C_{\mathrm{mix}} = 0.595$) and surface tension, whereas the component output is largest for NIST-small ($C_{\mathrm{comp}} = 0.524$). The pairwise share is also larger for mixing enthalpy than for surface tension. These differences arise under the same architecture, showing that the relative use of the three pathways depends on the target property.

**Departures from additive mixing.** Relative magnitude alone does not show whether the correction pathways capture the shape of the composition response, and aggregate error can hide errors in composition-dependent departures from additive mixing (Leenhouts et al., 2026). For a binary system, let $x$ denote the fraction of its second constituent. We compare the endpoint-aligned correction with the observed deviation from an additive reference:

$$\begin{aligned} r_{\mathrm{corr}}(x) &= r_{\mathrm{pair}}(x) + r_{\mathrm{mix}}(x) - (1-x)r_{\mathrm{mix}}(0) - x r_{\mathrm{mix}}(1), \\ \delta(x) &= y(x) - y_{\mathrm{add}}(x). \end{aligned} \tag{14}$$

At fixed conditions, $\hat{y}_{\mathrm{comp}}(x)$ in Eq. (6) is affine in $x$, and Eq. (9) gives $r_{\mathrm{pair}}(0) = r_{\mathrm{pair}}(1) = 0$. Thus, $r_{\mathrm{corr}}$ equals the departure of the full MIRA prediction from the linear interpolation of its endpoints. For mixing enthalpy, $y_{\mathrm{add}} = 0$ because the target is already an excess property; for NIST-small and surface tension, it interpolates the observed values at the extreme available compositions.

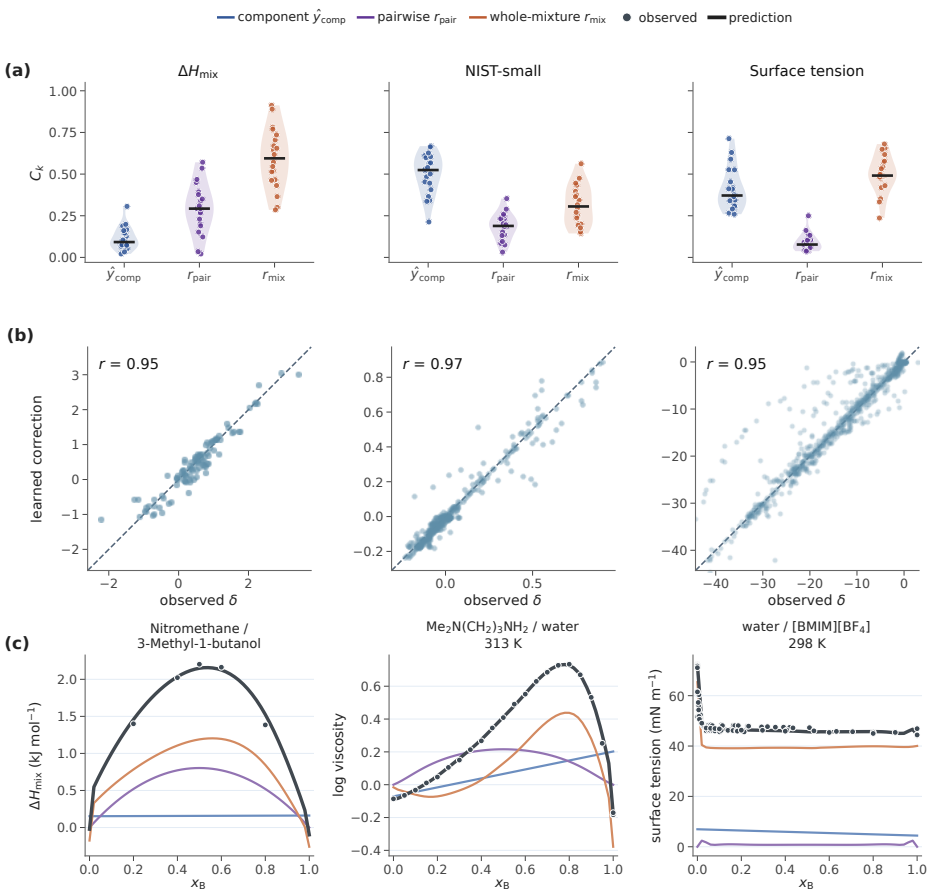


Figure 2: **Property-dependent behavior of the learned MIRA pathways.** **(a)** Relative pathway magnitudes across binary mixture systems for mixing enthalpy, NIST-small viscosity, and surface tension. **(b)** Endpoint-aligned learned corrections versus observed departures from an additive reference; dashed lines denote $y = x$. **(c)** Representative composition trajectories of the three pathway outputs, final prediction, and observations.

The learned correction closely follows the observed departure from the additive reference in all three property families, with Pearson correlations of $0.954$, $0.972$, and $0.951$ for mixing enthalpy, NIST-small viscosity, and surface tension, respectively (Figure 2(b)). This comparison attributes the composition-dependent departure to the combined correction $r_{\text{pair}} + r_{\text{mix}}$; it does not uniquely assign it to either pathway.

**Representative composition trajectories.** Figure 2(c) illustrates three distinct behaviors. Pairwise and whole-mixture corrections form the interior response for mixing enthalpy; all three pathways shape the nonlinear NIST-small curve; and surface tension is dominated by the whole-mixture correction with a small pairwise term. Thus, the same architecture realizes qualitatively different composition responses across properties.

**Frozen-model interventions.** Unlike retrained ablations, frozen-model interventions probe how the fitted Full MIRA prediction depends on each pathway without allowing the other pathways to adapt. We set $\hat{y}_{\text{comp}}$, $r_{\text{pair}}$, and $r_{\text{mix}}$ to zero one at a time and measure the increase in test MAE. On IL conductivity, zeroing the whole-mixture output increases MAE by 1.926, compared with 0.119 for the pairwise output; NIST-small is most sensitive to the component output.

For mixing enthalpy, the effect of removing the whole-mixture output increases toward more asymmetric compositions. Because zeroing also removes any constant pathway offset, we repeat the intervention by replacing each output with its training-set mean. The error still increases, showing that the pathways carry sample-dependent information rather than only constant offsets. On IL viscosity, mean replacement increases MAE by 0.895 for the component output and 0.745 for the whole-mixture output, compared with a Full MIRA MAE of 0.030. Complete mean-replacement results and fold/seed stability are reported in Appendix C.

### 4.5 Generalization to Unseen Mixture Systems

We define a mixture system as the unordered set of constituent molecules, independent of composition and measurement conditions. The official folds hold out observations, so other compositions of a test mixture system may remain in training. We instead assign every observation from an unordered molecular system to the same fold. Each test constituent still occurs in another training system. We report system-macro MAE, which gives each test system equal weight regardless of how many compositions or conditions were measured for it.

Table 4: **Generalization to unseen mixture systems and molecular pairs.** Five-fold system-macro MAE (↓), reported as mean ± sample standard deviation. DeepSets and MIRA use the same CheMeleon encoder. All test molecules occur in the training data. Surface tension is reported in mN/m.

| Property | Split | DeepSets | MIRA |
|---|---|---|---|
| NIST-small | System | $0.0457 \pm 0.0035$ | $0.0377 \pm 0.0048$ |
| | Pair | $0.0506 \pm 0.0048$ | $0.0387 \pm 0.0042$ |
| IL viscosity | System | $0.346 \pm 0.031$ | $0.285 \pm 0.048$ |
| | Pair | $0.332 \pm 0.052$ | $0.261 \pm 0.032$ |
| Surface tension | System | $2.409 \pm 0.338$ | $2.014 \pm 0.228$ |
| | Pair | $2.065 \pm 0.286$ | $1.584 \pm 0.291$ |
| $\Delta H_{\mathrm{mix}}$ | System | $0.1464 \pm 0.0005$ | $0.1442 \pm 0.0022$ |
| | Pair | $0.328 \pm 0.038$ | $0.306 \pm 0.005$ |

MIRA has lower system-macro MAE than the encoder-matched DeepSets readout on all four properties under both splits. The reductions are 16–18% for NIST-small, IL viscosity, and surface tension under system-disjoint evaluation, and 21–24% under the stricter pair-disjoint split; mixing enthalpy shows smaller reductions of 1.5% and 6.7%, respectively. Thus, MIRA gives lower error than simple set aggregation when known molecules are recombined into unseen systems or pairs. Complete results are in Appendix D.

## 5 Conclusion

In this work, we introduced **MIRA**, a unified framework for molecular mixture property prediction that separates component, pairwise, and whole-mixture contributions at the prediction level and combines them only at the final readout. Across eleven CheMixHub tasks and five external datasets, MIRA achieves strong predictive performance while retaining the same core structure across properties. Composition sweeps, endpoint-aligned analyses, and frozen-model interventions further show different and complementary pathway behavior across properties. For the three properties examined by composition sweeps, the combined correction closely follows observed departures from additive mixing, with Pearson correlations above 0.95.

**Limitations and outlook.** MIRA's pathways are optimized jointly without separate physical supervision, so their outputs should be interpreted as predictive components rather than uniquely identifiable physical contributions. The whole-mixture pathway remains active at pure-component endpoints and may capture effects also represented by the other pathways. Our system- and pair-disjoint tests form new mixtures from molecules seen during training rather than testing unseen molecular scaffolds. Future work will examine clearer pathway separation and scaffold-disjoint generalization.

## AI USE STATEMENT

We used ChatGPT to assist with language polishing and manuscript editing. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript.

## REFERENCES

Masoud Amiri and Farnaz Khaleseh. Predicting drug solubility in binary solvent mixtures using graph convolutional networks: A comprehensive deep learning approach. *Scientific Reports*, 15:45711, 2025. doi: 10.1038/s41598-025-28272-3. URL https://doi.org/10.1038/s41598-025-28272-3.

Tianyou Bai, Huan Wang, Mingchen Gao, Fangyue Lin, Pinze Ren, Zhenlin Zhao, and Siming Dong. Chem World: A large-scale benchmark and physics-informed framework for trustworthy chemical property prediction, 2026. URL https://arxiv.org/abs/2607.28079.

Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H. Green. Deep Learning Foundation Models from Classical Molecular Descriptors, 2026. URL https://arxiv.org/abs/2506.15792v2. Version 2.

Judith Kleinheins. Data for publication "a unified surface tension model for multi-component salt, organic and surfactant solutions". Zenodo, version 1, 2024.

Roel J. Leenhouts, Nathan Morgan, Emad Al Ibrahim, William H. Green, and Florence H. Vermeire. Pooling solvent mixtures for solvation free energy predictions. *Chemical Engineering Journal*, 513:162232, 2025. doi: 10.1016/j.cej.2025.162232.

Roel J. Leenhouts, Nathan K. Morgan, William Green, Jan G. Rittig, and Florence H. Vermeire. A Systematic Evaluation of Molecular Mixture Behavior Prediction, 2026. URL https://arxiv.org/abs/2605.29698v1.

Anyi Li, Jiacheng Cen, Songyou Li, Mingze Li, Yang Yu, and Wenbing Huang. Geometric mixture models for electrolyte conductivity prediction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 13236–13264, 2025. doi: 10.52202/085713-0448. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/13edb7a3b525022fd754a87d2a96710d-Abstract-Conference.html. Main Conference Track.

Dmitry Malikov, Lev Krasnov, Marina Kiseleva, Elizaveta Meshcheriakova, Fedor Kuznetsov, Vladimir Elistratov, Matvei Vasiyarov, Sergei Tatarin, and Stanislav Bezzubov. Dataset of solubility values for organic compounds in binary mixtures of solvents at various temperatures. *Scientific Data*, 13:727, 2026. doi: 10.1038/s41597-026-07047-z. URL https://doi.org/10.1038/s41597-026-07047-z.

National Institute of Standards and Technology. Experimental data sets for viscosity and surface tension of binary mixtures at the temperatures (293.15 to 323.15) k and the pressures (99.325 to 103.325) kpa. NIST Materials Data Repository, 2023.

Kamil Paduszynski. Extensive databases and group contribution QSPRs of ionic liquid properties. 3: Surface tension. *Industrial & Engineering Chemistry Research*, 60(15):5705–5720, 2021. doi: 10.1021/acs.iecr.1c00783.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/AAAI.V32I1.11671.

Shiyi Qin, Shengli Jiang, Jianping Li, Prasanna Balaprakash, Reid C. Van Lehn, and Victor M. Zavala. Capturing molecular interactions in graph neural networks: A case study in multi-component phase equilibrium. *Digital Discovery*, 2:138–151, 2023. doi: 10.1039/D2DD00045H.

Dihao Qing, Zhengyang Lu, Linxi Hou, Wei Dai, Jianquan Hong, and Xin Ge. Predicting the surface tension of mixed surfactant systems through theory and machine learning. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 727:138427, 2025. doi: 10.1016/j.colsurfa.2025.138427.

Ella Miray Rajaonson, Mahyar Rajabi Kochi, Luis Martin Mejia Mendoza, Seyed Mohamad Moosavi, and Benjamin Sanchez-Lengeling. Chemixhub: Datasets and benchmarks for chemical mixture property prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1107. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2f803abdcad9de35b45d5a656dade45c-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Demian Riccardi, Zachary Trautt, Ala Bazyleva, Eugene Paulechka, Vladimir Diky, Joe W. Magee, Andrei F. Kazakov, Scott Townsend, and Chris Muzny. Towards improved FAIRness of the ThermoML archive. *Journal of Computational Chemistry*, 2022. doi: 10.1002/jcc.26842.

Jan G. Rittig and Alexander Mitsos. Thermodynamics-consistent graph neural networks. *Chemical Science*, 15:18504–18512, 2024. doi: 10.1039/D4SC04554H.

Jan G. Rittig, Kobi C. Felton, Alexei A. Lapkin, and Alexander Mitsos. Gibbs–duhem-informed neural networks for binary activity coefficient prediction. *Digital Discovery*, 2:1752–1767, 2023. doi: 10.1039/D3DD00103B.

Zhiwei Shi, Miao Ma, Hanyang Ning, Bo Yang, and Liping Ding. A multi-level cross-molecular interaction neural network for mixture property prediction. *Physica A: Statistical Mechanics and its Applications*, 686:131340, 2026. doi: 10.1016/j.physa.2026.131340.

Thomas Specht, Mayank Nagda, Sophie Fellenz, Stephan Mandt, Hans Hasse, and Fabian Jirasek. HANNA: Hard-constraint neural network for consistent activity coefficient prediction. *Chemical Science*, 15(47):19777–19786, 2024. doi: 10.1039/D4SC05115G.

Vishwesh Venkatraman, Sigvart Evjen, and Kallidanthiyil Chellappan Lethesh. Ionic liquid properties. Zenodo data set, 2019.

Alexius Wadell, Anoushka Bhutani, Victor Azumah, Austin R. Ellis-Mohr, Andrew J. Stier, Kareem Hegazy, Alexander Brace, Hancheng Zhao, Celia Kelly, Anuj K. Nayak, Yuhan Chen, Dimitrios Simatos, Hongyi Lin, Murali Emani, Venkatram Vishwanath, Kevin Gering, Melisa Alkan, Tom Gibbs, Jack Wells, Wesley W. Qian, Richard C. Gerkin, Benjamin Amorelli, Alexander B. Wiltschko, Lav R. Varshney, Bharath Ramsundar, Karthik Duraisamy, Michael W. Mahoney, Arvind Ramanathan, and Venkatasubramanian Viswanathan. Foundation Models for Discovery and Exploration in Chemical Space, 2026. URL https://arxiv.org/abs/2510.18900v2. Version 2.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, July 2019. ISSN 1549-960X. doi: 10.1021/acs.jcim.9b00237. URL http://dx.doi.org/10.1021/acs.jcim.9b00237.

Xiao Yang, Baojiang He, Lingbo Ji, Xiankuan Huo, Jun Hu, and Zhan Qu. Cross-molecule attention unifies diverse mixture property predictions, 2026a. ChemRxiv preprint, version 3.

Zhenze Yang, Yifan Wu, Xu Han, Ziqing Zhang, Haoen Lai, Zhenliang Mu, Tianze Zheng, Siyuan Liu, Zhichen Pu, Zhi Wang, Zhiao Yu, Sheng Gong, and Wen Yan. A unified predictive and generative solution for liquid electrolyte formulation. *Nature Machine Intelligence*, 8(2):186–196, January 2026b. ISSN 2522-5839. doi: 10.1038/s42256-025-01173-w. URL http://dx.doi.org/10.1038/s42256-025-01173-w.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Hengrui Zhang, Tianxing Lai, Jie Chen, Arumugam Manthiram, James M. Rondinelli, and Wei Chen. Learning molecular mixture property using chemistry-aware graph neural network. *PRX Energy*, 3(2):023006, 2024. doi: 10.1103/PRXEnergy.3.023006.

Shang Zhu, Bharath Ramsundar, Emil Annevelink, Hongyi Lin, Adarsh Dave, Pin-Wen Guan, Kevin Gering, and Venkatasubramanian Viswanathan. Differentiable modeling and optimization of non-aqueous Li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15:8649, 2024. doi: 10.1038/s41467-024-51653-7.

# A Implementation and Evaluation Protocols

## A.1 Training and Model Selection

CheMixHub runs use a trainable pretrained CheMeleon encoder and one MIRA architecture across tasks. Training uses Huber loss, batch size 256, at most 1,000 epochs, and patience 100. A 50-epoch warmup starts at 0.1 of the learning rate, followed by cosine decay to 0.02 of it. The lowest-validation-MAE checkpoint is evaluated on the test fold. Bamboo-electrolyte and surface-tension runs use validation-based selection with patience 100 and budgets of 2,000 and 500 epochs, respectively.

## A.2 MIRA Architecture and Conditioning

Table 5 specifies the shared architecture. All scalar heads use two-hidden-layer MLPs. The base pair head has standard initialization; the condition-pair residual and whole-mixture output are zero-initialized, so joint training begins with the component and base-pair routes active.

Table 5: **MIRA architecture used in the main experiments.**

| Block | Configuration |
|---|---|
| Molecular encoder | Trainable CheMeleon; 128-dimensional output |
| Shared projector | Affine projection to $d = 128$ |
| Attention summary | 2 layers, 8 heads, mean pooling, 3-layer 128-dimensional MLP |
| Invariant summary | Mean, maximum, and pair sum; pair MLP $385 \to 256 \to 128$; fusion $384 \to 128$ |
| Scalar heads | 2 GELU hidden layers of width 100 and a scalar output |
| Condition modules | Component FiLM: 128; mixture FiLM: 256; pair-condition encoder: $64 \to 32$ |
| Regularization | Zero dropout in the main configuration |

Component and whole-mixture FiLM use affine maps of $c$ followed by sigmoid activations to produce scale and shift vectors; tasks without conditions bypass FiLM. The pair-condition encoder appends availability masks to transformed fields, distinguishing missing values from zero. When available, temperature is the first field and contributes centered inverse-temperature features. A 64-unit GELU layer maps the result to $e_c \in \mathbb{R}^{32}$. With no valid condition, the encoder returns zero and masks the conditioned pair residual.

The attention and pair-summary branches both scale as $O(n^2)$ in the number of constituents. Pair aggregation retains the composition weights and is not divided by the number of pairs.

For a record with $S \geq 2$ semantic streams, let $\hat{y}_s$ and $z_s$ denote the single-stream prediction and mixture embedding, and let $\ell_s$ be a learned 16-dimensional role embedding. We form $q_s = [z_s; \hat{y}_s; \ell_s]$ and use the symmetric stream feature $D(q_s, q_t) = [q_s + q_t; |q_s - q_t|; q_s \odot q_t]$. The multistream prediction is

$$\hat{y}_{\text{multi}} = \frac{1}{S}\sum_{s=1}^{S} \hat{y}_s + g_{\text{stream}}\left(\frac{1}{\binom{S}{2}}\sum_{s<t} D(q_s, q_t)\right), \tag{15}$$

where $g_{\text{stream}}$ is a zero-initialized MLP with hidden width 100. For $S = 1$, the adapter returns $\hat{y}_1$ exactly. The Drug Solubility and MixtureSolDB configurations use this adapter for their role-separated inputs.

## A.3 Architectural Properties

**Permutation invariance.** The molecular encoder and projector share parameters across constituents. The pair map is symmetric and summed over unordered pairs; self-attention is permutation equivariant in the absence of positional encodings; and the weighted mean, maximum, and pooled attention statistics are invariant. Consequently, jointly permuting the constituent identities and compositions does not change the final prediction.

**Endpoint behavior and scaling.** At a pure-component endpoint, all products $x_i x_j$ with $i \neq j$ vanish, so the explicitly pair-weighted correction is zero. The component and whole-mixture pathways can remain active. With $n$ constituents, explicit pair enumeration uses $n(n-1)/2$ pairs and component self-attention has a quadratic attention matrix. MIRA does not apply attention over the pair tokens.

### A.4 MIXTURESOLDB MIX-EXP PROTOCOL

**Data representation and split.** The Mixture Property Benchmark standardization of MixtureSolDB mix-exp has 125,932 mixture rows and 782 compounds (Malikov et al., 2026; Leenhouts et al., 2026). Its target is standardized $\log_{10}$ mole-fraction solubility. The two solvents form a mixture with $x_1 + x_2 = 1$; the solute has a separate role stream with fraction one, and temperature enters through FiLM. We train only on mixture rows. The official random-rowwise folds contain about 88,152/12,593/25,187 train/validation/test rows (25,184 test rows in fold 4). They differ from the CheMixHub Drug folds and do not test molecular extrapolation. The published D-MPNN+FFN reference also trains on available pure-compound observations (Leenhouts et al., 2026).

**Model and optimization.** The model uses trainable CheMeleon, self-attention IMEX, exchange IMEX, and the Drug-task multistream adapter. We train with Huber loss and Adam for at most 1,000 epochs (batch size 256; eight workers; seed 42; patience 100). The 50-epoch warmup and cosine schedule use factors 0.1 and 0.02; parameter-group learning rates are $10^{-4}$ and $5 \times 10^{-4}$. Validation RMSE selects checkpoints. Absolute test RMSE is primary, with MAE reported secondarily.

Table 6: **MixtureSolDB mix-exp five-fold results.** Test metrics for the official random-rowwise folds. The final row reports the mean $\pm$ sample standard deviation across folds; RMSE is the primary metric.

| Fold | MAE | RMSE |
|---|---|---|
| 0 | 0.01966 | 0.04535 |
| 1 | 0.01856 | 0.04837 |
| 2 | 0.02101 | 0.04462 |
| 3 | 0.01838 | 0.04986 |
| 4 | 0.01835 | 0.04513 |
| Mean $\pm$ std. | $0.01919 \pm 0.00115$ | $0.04667 \pm 0.00231$ |

### A.5 METRICS AND COMPARISON PROTOCOLS

**Metric aggregation.** We compute MAE, RMSE, $R^2$, MSE, and Pearson correlation within each test fold, then report their arithmetic mean and sample standard deviation. Target transforms are reversed unless the benchmark defines a transformed target (e.g., log viscosity). Ensemble metrics are computed after averaging aligned predictions.

**Comparison scope.** The CheMixHub baseline block follows Rajaonson et al. (2025); Bamboo-Mixer and MIST values follow their publications (Yang et al., 2026b; Wadell et al., 2026). MIST uses pooled attention for Olfactory. CheMeleon+DeepSets, CheMeleon+Raw Pair, MIRA (D-MPNN), and MIRA (CheMeleon) use the official five folds. CALiSol and DiffMix retain GeoMix's 70/20/10 splits, original-space MSE, and Pearson correlation. Surface tension uses a grouped hold-out and reports MAE in mN/m. MixtureSolDB uses separate random-rowwise folds and absolute RMSE, so it is distinct from CheMixHub Drug; its D-MPNN+FFN reference includes pure-compound training observations.

## B COMPLETE CHEMIXHUB BASELINE MATRIX

Tables 7 and 8 report the per-model results summarized in Table 1. Published values retain the precision reported by their sources; evaluations in this work are five-fold means ± sample standard deviations.

Table 7: **Complete CheMixHub comparison for density, thermodynamic, solubility, polymer, and conductivity tasks.** Test MAE (↓).

| Model | Density | $\Delta H_{\mathrm{mix}}$ | $\Delta H_{\mathrm{vap}}$ | Drug Sol. | SPE / Polymer | IL Cond. |
|---|---|---|---|---|---|---|
| *Published results* | | | | | | |
| GNN + Attention | 0.018 (0.020) | 0.158 (0.002) | 0.098 (0.006) | 0.087 (0.006) | 0.312 (0.043) | 0.276 (0.044) |
| GNN + DeepSets | 0.003 (<0.001) | 0.159 (0.002) | 0.406 (0.668) | 0.065 (0.005) | 0.315 (0.067) | 0.226 (0.017) |
| MolT5 + XGBoost | 0.009 (<0.001) | 0.269 (0.004) | 0.306 (0.003) | 0.028 (0.001) | 0.222 (0.007) | 0.071 (0.001) |
| MolT5 + Attention | 0.005 (0.001) | 0.157 (0.002) | 0.125 (0.077) | 0.082 (0.027) | 0.279 (0.006) | 0.244 (0.011) |
| MolT5 + DeepSets | 0.008 (0.005) | 0.157 (0.003) | 0.071 (0.002) | 0.130 (0.013) | 0.328 (0.010) | 0.196 (0.013) |
| RDKit + XGBoost | 0.009 (<0.001) | 0.225 (0.005) | 0.295 (0.002) | 0.028 (0.001) | 0.223 (0.008) | 0.073 (0.002) |
| RDKit + Attention | 0.006 (0.001) | 0.167 (0.002) | 0.199 (0.030) | 0.070 (0.006) | 0.394 (0.028) | 0.407 (0.019) |
| RDKit + DeepSets | 0.005 (<0.001) | 0.207 (0.008) | 0.079 (0.005) | 0.179 (0.011) | 0.344 (0.016) | 0.290 (0.059) |
| Bamboo-Mixer | 0.003 (<0.001) | 0.165 (0.002) | 0.050 (0.018) | 0.012 (0.001) | – | 0.068 (0.017) |
| MIST Linear Combination | 0.003 (<0.001) | 0.163 (0.002) | 0.067 (0.001) | – | – | 0.644 (0.014) |
| *Evaluations in this work* | | | | | | |
| CheMeleon + DeepSets | 0.005 (0.001) | 0.156 (0.003) | 0.055 (0.004) | 0.033 (0.003) | 0.275 (0.008) | 0.317 (0.043) |
| CheMeleon + Raw Pair | 0.003 (<0.001) | 0.198 (0.003) | 0.075 (0.001) | 0.025 (0.001) | 0.225 (0.009) | 0.465 (0.015) |
| MolSets-DMPNN | 0.003 (<0.001) | 0.173 (0.004) | 0.053 (0.003) | 0.210 (0.304) | 0.314 (0.005) | 0.467 (0.022) |
| MolPool-MPN | 0.003 (<0.001) | 0.157 (0.003) | 0.043 (0.001) | 0.034 (0.003) | 0.185 (0.003) | 0.126 (0.018) |
| MIRA (D-MPNN) | 0.003 (<0.001) | 0.160 (0.002) | 0.041 (0.001) | 0.014 (<0.001) | 0.171 (0.008) | 0.047 (0.004) |
| **MIRA (CheMeleon)** | **0.003** (<0.001) | **0.154** (0.003) | **0.038** (0.001) | **0.011** (0.001) | **0.165** (0.007) | **0.043** (0.003) |

Table 8: **Complete CheMixHub comparison for viscosity, fuel, and perceptual tasks.** Test MAE (↓).

| Model | IL Visc. | NIST-full | NIST-small | MON | Olfactory |
|---|---|---|---|---|---|
| *Published results* | | | | | |
| GNN + Attention | 0.154 (0.084) | 0.136 (0.010) | 0.035 (0.004) | 10.240 (1.658) | 0.129 (0.005) |
| GNN + DeepSets | 0.206 (0.020) | 0.131 (0.010) | 0.091 (0.005) | 5.990 (1.382) | 0.146 (0.010) |
| MolT5 + XGBoost | 0.078 (0.002) | 0.148 (0.001) | 0.059 (0.001) | 5.002 (0.538) | 0.128 (0.006) |
| MolT5 + Attention | 0.083 (0.035) | 0.076 (0.004) | 0.030 (0.001) | 4.660 (0.603) | 0.123 (0.005) |
| MolT5 + DeepSets | 0.132 (0.003) | 0.162 (0.009) | 0.056 (0.004) | 5.296 (0.585) | 0.121 (0.006) |
| RDKit + XGBoost | 0.076 (0.002) | 0.055 (<0.001) | 0.048 (0.002) | 4.570 (0.348) | 0.125 (0.006) |
| RDKit + Attention | 0.100 (0.003) | 0.069 (0.006) | 0.056 (0.004) | 11.297 (2.110) | 0.148 (0.010) |
| RDKit + DeepSets | 0.107 (0.007) | 0.137 (0.005) | 0.047 (0.003) | 7.625 (1.874) | 0.150 (0.008) |
| Bamboo-Mixer | 0.054 (0.003) | 0.029 (<0.001) | 0.016 (0.001) | 4.460 (0.556) | 0.125 (0.009) |
| MIST Linear Combination | 0.433 (0.005) | 0.309 (0.002) | 0.036 (0.001) | 4.028 (0.631) | 0.112 (0.008) |
| *Evaluations in this work* | | | | | |
| CheMeleon + DeepSets | 0.106 (0.035) | 0.088 (0.032) | 0.031 (0.010) | 4.382 (0.345) | 0.098 (0.005) |
| CheMeleon + Raw Pair | 0.104 (0.005) | 0.068 (0.001) | 0.023 (<0.001) | 4.010 (0.183) | 0.093 (0.009) |
| MolSets-DMPNN | 0.120 (0.003) | 0.098 (0.002) | 0.037 (0.003) | 5.334 (1.638) | 0.120 (0.005) |
| MolPool-MPN | 0.076 (0.006) | 0.035 (0.001) | 0.017 (0.001) | 4.213 (1.272) | 0.106 (0.002) |
| MIRA (D-MPNN) | 0.030 (0.001) | 0.028 (<0.001) | 0.015 (<0.001) | 3.659 (0.749) | 0.095 (0.005) |
| **MIRA (CheMeleon)** | **0.030** (0.001) | **0.027** (0.001) | **0.013** (<0.001) | **3.558** (0.260) | **0.092** (0.002) |

# C Pathway Diagnostics

## C.1 Stability Across Folds and Seeds

We repeat the composition analysis on the same 20 systems using all five models trained on the official folds for mixing enthalpy and NIST-small, and three independently trained surface-tension models. Table 9 reports variation across models. The dominant property-level pathway is unchanged: the whole-mixture pathway dominates mixing enthalpy and surface tension, whereas the component pathway dominates NIST-small. The endpoint-aligned correction correlation also varies little across models. System rankings are most stable for the component and whole-mixture pathways; the weaker pairwise pathway shows greater ranking variation. In every model, the three outputs reconstruct the prediction to a maximum absolute error below $2.4 \times 10^{-7}$.

Table 9: **Stability of the learned pathway organization.** Pathway magnitudes are 20-system means within each model; these and the correction correlations are reported as mean $\pm$ sample standard deviation across trained models evaluated on the same 20 composition series (five folds for mixing enthalpy and NIST-small; three seeds for surface tension). System rank reports the mean pairwise Spearman correlation of the pathway-wise system rankings across models. Consensus counts systems whose dominant pathway is identical across all models.

| Property | $C_{\mathrm{comp}}$ | $C_{\mathrm{pair}}$ | $C_{\mathrm{mix}}$ | $\rho_P(r_{\mathrm{corr}}, \delta)$ | Rank $\rho_S$ comp/pair/mix | Cons. |
|---|---|---|---|---|---|---|
| $\Delta H_{\mathrm{mix}}$ | $0.120 \pm 0.022$ | $0.301 \pm 0.018$ | $0.579 \pm 0.027$ | $0.942 \pm 0.011$ | $0.910/0.840/0.866$ | $15/20$ |
| NIST-small | $0.583 \pm 0.045$ | $0.131 \pm 0.036$ | $0.286 \pm 0.022$ | $0.974 \pm 0.004$ | $0.879/0.769/0.885$ | $16/20$ |
| Surface tension | $0.376 \pm 0.030$ | $0.106 \pm 0.015$ | $0.518 \pm 0.016$ | $0.951 \pm 0.001$ | $0.900/0.681/0.730$ | $17/20$ |

On fixed fold 0, we evaluate seeds 42–46 on the same 20 systems and 51-point grids. Table 10 reports each seed's 20-system median as mean $\pm$ sample standard deviation across seeds. The combined correction share $C_{\mathrm{corr}}$ adds $C_{\mathrm{pair}}$ and $C_{\mathrm{mix}}$ per system before taking the median, so the displayed medians need not add. System-rank $\rho_S$ averages Spearman correlations over the ten seed pairs for system-level shares, not pointwise trajectories.

Table 10: **Pathway allocation across five seeds on fixed fold 0.** Each seed is evaluated on the same 20 binary systems and 51-point composition grids. Pathway shares are 20-system medians, reported as the mean $\pm$ sample standard deviation over seeds 42–46. The final column gives the mean seed-pair Spearman correlation of system rankings for the component, pairwise, and whole-mixture shares, respectively.

| Property | $C_{\mathrm{comp}}$ | $C_{\mathrm{pair}}$ | $C_{\mathrm{mix}}$ | $C_{\mathrm{corr}}$ | Rank $\rho_S$ comp/pair/mix |
|---|---|---|---|---|---|
| $\Delta H_{\mathrm{mix}}$ | $0.118 \pm 0.040$ | $0.272 \pm 0.026$ | $0.589 \pm 0.051$ | $0.882 \pm 0.040$ | $0.60/0.61/0.73$ |
| NIST-small | $0.546 \pm 0.030$ | $0.146 \pm 0.029$ | $0.298 \pm 0.013$ | $0.454 \pm 0.030$ | $0.67/0.88/0.73$ |
| Surface tension | $0.348 \pm 0.015$ | $0.092 \pm 0.012$ | $0.514 \pm 0.033$ | $0.652 \pm 0.015$ | $0.77/0.63/0.72$ |

The dominant property-level pathway is unchanged across the five seeds: whole-mixture for mixing enthalpy and surface tension, and component for NIST-small. Pairwise mass remains smallest for surface tension. The system-rank correlations are moderate for some pathways, so the stable property-level ordering does not imply identical system-level allocations. The three outputs reconstruct each prediction with a maximum absolute error of $2 \times 10^{-7}$ in this seed analysis.

## C.2 Mean-Replacement Interventions

Zeroing a pathway removes both its sample-dependent signal and any constant offset. We therefore repeat the frozen-model intervention by replacing $\hat{y}_{\mathrm{comp}}$, $r_{\mathrm{pair}}$, or $r_{\mathrm{mix}}$ with that output's mean on the corresponding training partition, one at a time. This preserves the pathway's training-set mean while removing its sample-specific variation. Table 11 reports the resulting increase in test MAE for each pathway; every increase is positive. For mixing enthalpy, NIST-small, and IL conductivity, the dominant mean-replacement effects remain close to their zero-out counterparts. On IL viscosity, centering reduces the component and whole-mixture effects, but the remaining increases of 0.895

and 0.745 remain large relative to the full-model MAE of 0.030. The intervention therefore captures sample-dependent predictive information in addition to pathway offsets. As with zeroing, it measures functional dependence of the fitted predictor rather than unique attribution.

Table 11: **Zero-out and mean-replacement pathway interventions.** Test MAE is mean ± sample standard deviation. Intervention columns report the mean increase in test MAE across five folds after either setting a pathway to zero or replacing it with its training-partition mean. All models remain frozen.

| | | Zero-out | | | Mean replacement | | |
|---|---|---|---|---|---|---|---|
| **Task** | **Full MAE** | $\Delta_{\text{comp}}$ | $\Delta_{\text{pair}}$ | $\Delta_{\text{mix}}$ | $\Delta_{\text{comp}}^{\text{mean}}$ | $\Delta_{\text{pair}}^{\text{mean}}$ | $\Delta_{\text{mix}}^{\text{mean}}$ |
| IL Cond. | $0.043 \pm 0.003$ | 0.924 | 0.119 | 1.926 | 0.218 | 0.125 | 1.901 |
| IL Visc. | $0.030 \pm 0.001$ | 2.342 | 0.209 | 2.399 | 0.895 | 0.214 | 0.745 |
| NIST-small | $0.013 \pm 0.000$ | 0.201 | 0.050 | 0.094 | 0.201 | 0.026 | 0.076 |
| $\Delta H_{\text{mix}}$ | $0.154 \pm 0.003$ | 0.008 | 0.185 | 0.304 | 0.007 | 0.128 | 0.293 |

## D Generalization to Unseen Mixture Systems

**Split construction.** All observations from an unordered molecular system are assigned to the same fold. A test system is absent from training, while every constituent molecule appears in at least one different training system. The pair-disjoint stress test is stricter: every unordered molecular pair in the test fold is absent from training, although the individual molecules remain represented. These splits therefore evaluate recombination of known constituents rather than generalization to unseen molecular identities.

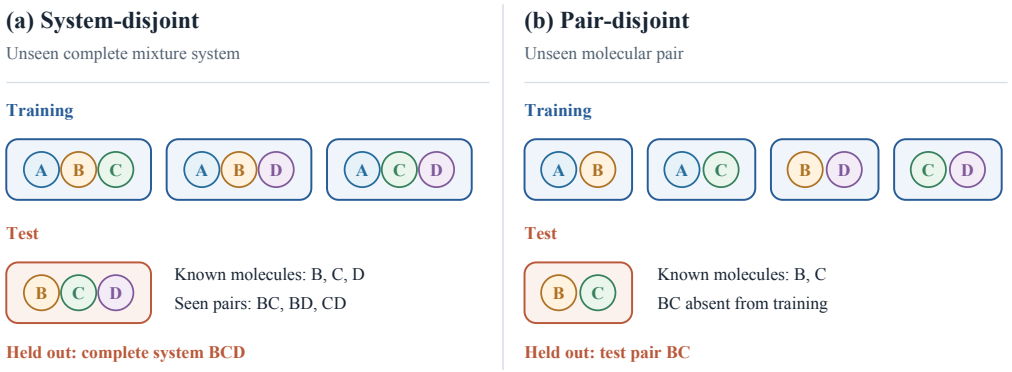


Figure 3: **Schematic comparison of generalization splits.** **(a)** In the system-disjoint setting, the complete test mixture system is absent from training while its constituent molecules remain observed. Constituent pairs may also have appeared in other training systems, as in the example shown. **(b)** In the pair-disjoint setting, every test molecular pair is absent from training although the individual molecules remain observed.

**Molecule-disjoint evaluation.** We additionally evaluate MIRA under the molecule split of Leenhouts et al. (2026), where all molecular identities in the test mixtures are absent from training. This setting is substantially stricter than our system- and pair-disjoint splits, which recombine molecules observed during training. MIRA shows mixed performance in this regime. It obtains the lowest mean RMSE for flash point (11.8) among the evaluated methods and remains competitive for viscosity, but does not consistently outperform the reference models on vaporization enthalpy, solvation free energy, or solubility. These results indicate that extrapolation to entirely unseen molecular identities remains challenging.

Table 12: **Generalization to unseen mixture systems and molecular pairs.** Five-fold system-macro MAE (↓), reported as mean ± sample standard deviation. DeepSets and MIRA use the same CheMeleon encoder. All test molecules occur in the training data. Surface tension is reported in mN/m.

| Property | Split | DeepSets | MIRA |
|---|---|---|---|
| NIST-small | System | $0.0457 \pm 0.0035$ | $0.0377 \pm 0.0048$ |
| | Pair | $0.0506 \pm 0.0048$ | $0.0387 \pm 0.0042$ |
| IL viscosity | System | $0.346 \pm 0.031$ | $0.285 \pm 0.048$ |
| | Pair | $0.332 \pm 0.052$ | $0.261 \pm 0.032$ |
| Surface tension | System | $2.409 \pm 0.338$ | $2.014 \pm 0.228$ |
| | Pair | $2.065 \pm 0.286$ | $1.584 \pm 0.291$ |
| $\Delta H_{\mathrm{mix}}$ | System | $0.1464 \pm 0.0005$ | $0.1442 \pm 0.0022$ |
| | Pair | $0.328 \pm 0.038$ | $0.306 \pm 0.005$ |

**Aggregation levels.** Let $\mathcal{S}$ denote the test systems, $I_s$ the observations belonging to system $s$, and $N_s = |I_s|$. Table 12 reports system-macro MAE, which averages system-level errors equally:

$$\mathrm{MAE}_{\mathrm{sys}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left( \frac{1}{N_s} \sum_{i \in I_s} |y_i - \hat{y}_i| \right). \tag{16}$$

Unlike observation-level MAE, this prevents systems with dense composition or condition sweeps from dominating the summary.

## E Surface-Tension Task Data and Provenance

**Scope and composition.** We assembled the external surface-tension task in Table 2 from heterogeneous liquid–gas surface-tension records, including measured data and compiled pure-component values. The frozen task contains 76,980 observations, 9,636 component identifiers, and 10,930 unordered molecular systems. Figure 4 summarizes its record composition, and Table 13 reports the corresponding system counts and data splits.

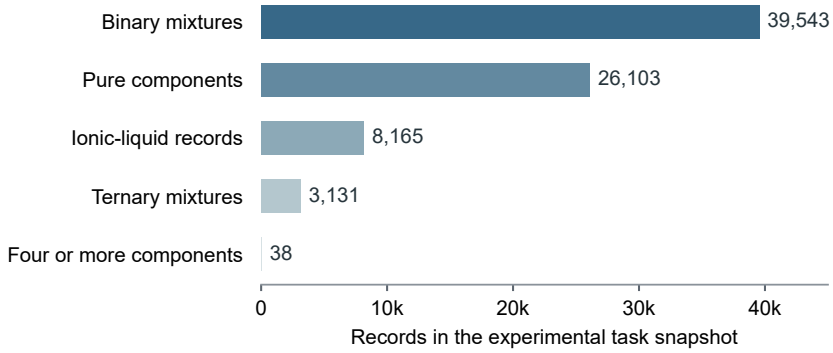


Figure 4: **Composition of the frozen surface-tension task.** Number of records in each data class of the 76,980-record task snapshot.

Table 13: **Coverage of the frozen surface-tension task.** Systems are distinct unordered sets of component identifiers within each class. The total deduplicates systems shared between classes, so system counts by class need not sum to the total. Rows are observations, not independent experiments.

| Record class | Rows | Systems | Train | Val. | Test |
|---|---|---|---|---|---|
| Pure components | 26,103 | 8,693 | 18,194 | 2,654 | 5,255 |
| Binary mixtures | 39,543 | 1,516 | 27,860 | 3,917 | 7,766 |
| Ionic-liquid records | 8,165 | 861 | 5,956 | 733 | 1,476 |
| Ternary mixtures | 3,131 | 80 | 2,153 | 241 | 737 |
| Four or more components | 38 | 11 | 29 | 2 | 7 |
| **Total** | **76,980** | **10,930** | **54,192** | **7,547** | **15,241** |

**Split and value checks.** The 29,332 formulation groups are keyed by component identifiers and mole fractions, keeping measurements at different temperatures of the same formulation in one partition. No formulation group crosses the 54,192/7,547/15,241 train/validation/test split. Molecular systems can recur across partitions; the system-disjoint test in Appendix D assesses a different generalization setting. All 76,980 records have unique table identifiers, matched component and fraction vector lengths, mole fractions in $[0, 1]$ summing to one within $10^{-5}$, and positive temperature and target values. Temperature spans 183.00–743.71 K (median 303.15 K; interquartile range 298.15–318.15 K), and surface tension spans 0.026–99.08 mN/m (median 30.39; interquartile range 24.30–41.00). Source labels are present for all rows, although 985 rows retain placeholder literature-reference fields.

**Sources in the training data.** Table 14 counts only rows in the frozen 76,980-record task, including its training, validation, and test partitions. The reference field has different levels of detail across collections: ThermoML rows retain article DOIs, NIST binary rows retain source-package reference IDs, and several other collections retain only a dataset or paper label (National Institute of Standards and Technology, 2023; Riccardi et al., 2022).

Table 14: **Source labels in the frozen surface-tension task.** Counts refer to model-input records across all three partitions, not to separate source-module releases. The last column describes the recorded reference field, which does not always identify an original experiment.

| Class and source-label group | Rows | Reference retained |
|---|---|---|
| Pure: ML-Sigma | 11,910 | Dataset label |
| Pure: Yaws Chapter 19 compilation | 7,227 | Compound identifier or placeholder |
| Pure: NIST ThermoML | 6,902 | Article DOI |
| Pure: book and individual study | 64 | Book or study citation |
| Binary: NIST binary-mixture package | 34,910 | 628 NIST reference IDs |
| Binary: NIST ThermoML | 2,775 | Article DOI |
| Binary: remaining labels | 1,858 | NIST mirror and mixed literature labels |
| Ionic-liquid collections | 8,165 | Collection citations; some original-paper details |
| Ternary collections | 3,131 | Article DOIs or study/collection labels |
| $\geq$ 4-component study labels | 38 | Study labels |
| **Total** | **76,980** | |

**Pure-component provenance.** The 26,103 pure rows comprise 18,194 training, 2,654 validation, and 5,255 test records. ML-Sigma is a published dataset label (**?**); the frozen rows do not carry its underlying experimental citations. All 6,902 pure ThermoML rows retain article DOIs. The 7,227 locally compiled rows were drawn from Yaws Chapter 19, as were 63 further rows carrying an explicit book label (**?**). In the frozen table, however, 985 of the 7,227 rows have placeholder references and most of the remainder retain compound identifiers rather than the book citation. We therefore attribute these rows to the Yaws compilation without treating every value as a separately traced original measurement.

Among the remaining records, the binary subset includes 1,146 rows with a generic LGDB NIST-mirror label. The ionic-liquid subset combines 5,309 Paduszynski-labeled rows and 2,856 Zenodo-labeled rows (Paduszynski, 2021; Venkatraman et al., 2019). The ternary subset includes 1,098 ThermoML rows, 895 rows under a Kleinheins collection label, and 356 surfactant-labeled rows (Kleinheins, 2024; Qing et al., 2025). The Kleinheins-labeled rows cite eight different underlying studies, so their collection label alone does not establish each record's original source.

**Record overlap.** In the frozen model table, 3,840 rows repeat an exact formulation–temperature–value key when source labels are ignored, whereas only two rows repeat that key when the source label is included. These source-overlapping records remain distinguishable, and the formulation split keeps every repeated key within one partition.