Learning Reaction-Condition Plausibility for Evaluation and Self-Curation Under Noisy and Non-Unique Supervision

Hyukjun Lim [1] **Yousung Jung** [1]

Abstract

Many scientific benchmarks are built on reference labels that are noisy, incomplete, or non-unique. For reaction-condition prediction, exact match can therefore mark a workable protocol as wrong when it differs from the archived record. We present UniCon, a framework for reaction-condition evaluation and self-curation that learns the plausibility of reaction-condition pairs instead of trying to reproduce a single archived label. UniCon aligns 2D graph representations of reaction transformations with fingerprint-based embeddings of chemical conditions in a shared latent space. It scores a candidate protocol against the archived one using a Bradley-Terry preference, which we call UniConScore. The same score can curate training data by comparing archived records with predictions from an independent condition prediction model and filtering chemically anomalous entries. Across expert preference studies, zero-shot high-throughput experimentation benchmarks, and self-curation analyses, UniConScore gives a ranking signal that tracks chemical viability better than exact match under noisy supervision.

1. Introduction

Translating retrosynthetic plans into executable laboratory protocols remains a major bottleneck in automated organic synthesis (Gao et al., 2018; Ball et al., 2025). Methods for target-oriented synthesis have improved substantially at predicting reactant sets (Sacha et al., 2021; Tu & Coley, 2022; Chen & Jung, 2021; Chen et al., 2020; Kim et al., 2021; Zhao et al., 2024; Wang & Montana, 2025), but a route can only be tested if its conditions are compatible with the intended transformation (Gao et al., 2018; Maser et al., 2021; Wang et al., 2023; 2025; Yan et al., 2025). Catalysts, ligands, solvents, bases, additives, and temperature jointly determine whether a proposed route can be converted into an experimentally meaningful protocol (Gao et al., 2018; Ball et al., 2025; Wang et al., 2025).

Scientific datasets often contain labels that are incomplete, noisy, or non-unique, which makes exact-match evaluation a weak proxy for validity (Thakkar et al., 2020; Northcutt et al., 2021; Wigh et al., 2024; Lee et al., 2024). Reaction-condition records are a clear example, since patent-derived datasets such as USPTO (Lowe, 2017) and curated commercial repositories such as Reaxys (Lawson et al., 2014) are assembled from text-mined or semi-structured records. As a result, reaction-condition annotations may omit essential catalysts, misassign solvent roles, or treat work-up agents such as water and brine as reaction-driving conditions (Schneider et al., 2016; Thakkar et al., 2020; Andronov et al., 2023; Wigh et al., 2024). Models trained directly on these annotations may learn purification artifacts rather than chemically decisive factors. They can reproduce incomplete historical records without identifying the reagents required for the transformation (Thakkar et al., 2020; Lee et al., 2024; Yan et al., 2025).

A second challenge is that reaction conditions are inherently non-unique (Ball et al., 2025; Wang et al., 2025; Shim et al., 2025). Many transformations admit multiple viable protocols that differ in catalyst family, solvent system, base, or additive choice. Exact-match evaluation collapses this diversity into a binary outcome: a prediction is correct only when it reproduces the archived record exactly (Wang et al., 2025; Yan et al., 2025). This penalizes chemically plausible alternatives because they differ from one stored example, so benchmark performance becomes partly a measure of database incompleteness rather than chemical validity.

These considerations motivate a different objective. Instead of asking only whether a model can recover an archived label, we ask whether it can score the plausibility of a reaction-condition pair, compare competing protocols, flag suspect database entries, and rank candidates against external experimental outcomes. We introduce the **Uni**fied Framework for Reaction-**Con**dition Evaluation and Self-Curation (UniCon), which centers on a learned reaction-condition plausibility

[1]Department of Chemical and Biological Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea. Correspondence to: Yousung Jung <yousung.jung@snu.ac.kr>.

Accepted at ICML 2026 AI for Science Workshop.

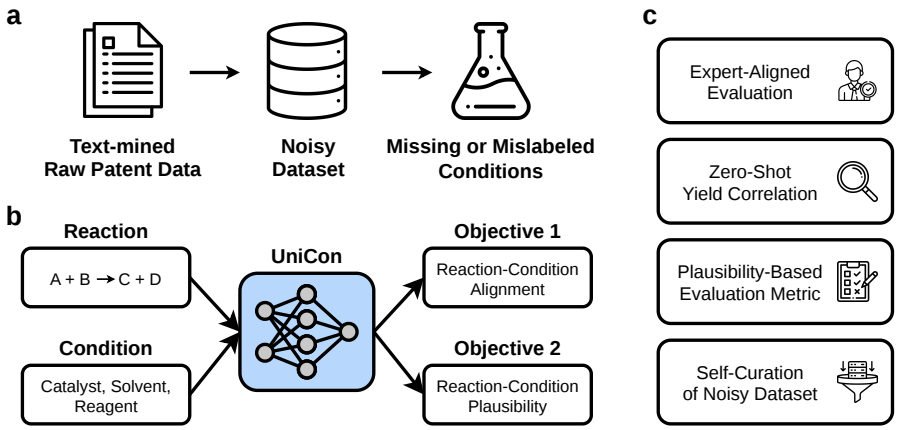


Figure 1. **Overview of the proposed framework.** (a) Identification of data extraction challenges, specifically noisy and incomplete condition labels derived from raw patent text. (b) The proposed multimodal transformer architecture, integrating reaction and condition inputs. The model is optimized for global alignment and reaction-condition plausibility scoring. (c) Downstream applications of the derived plausibility score, including expert-aligned evaluation, dataset self-curation, and zero-shot yield correlation.

score. UniCon supports evaluation and self-curation when the archived labels are noisy or incomplete (see Figure 1).

The framework combines structural reaction representations, role-ordered condition inputs, and structure-aware reagent embeddings within a shared multimodal transformer. Global alignment organizes reactions and conditions in a common latent space, and plausibility scoring provides a continuous signal for ranking and comparison. When alternative protocols are available, the same plausibility score can compare them and guide dataset refinement. In experiments, UniCon-Score aligns with expert judgments, supports self-curation of noisy datasets, and transfers to zero-shot prioritization on external high-throughput experimentation benchmarks.

2. Related work

2.1. Chemical Reaction-Condition Prediction

Reaction-condition prediction is a core problem in computer-aided synthesis planning because a retrosynthetic route is useful only when it can be translated into an experimentally executable protocol. Early data-driven systems formulated this task as the prediction of suitable catalysts, solvents, reagents, and temperatures from reaction structure (Gao et al., 2018). Subsequent work has explored multilabel classification, molecular translation, graph-based representations, template-aware prediction, and robust recommendation strategies (Maser et al., 2021; Andronov et al., 2023; Wang et al., 2023; 2025; Yan et al., 2025). Despite these architectural differences, most methods are trained and evaluated against archived condition labels as if each reaction had a single canonical protocol. That assumption breaks down when several protocols are viable or when the recorded label is incomplete. UniCon shifts the target from label reconstruction to plausibility assessment of reaction-condition pairs.

2.2. Noisy Reaction-Condition Datasets

The reliability of reaction-condition models depends heavily on the datasets used for training and evaluation. Large reaction corpora derived from patents, publications, or semi-structured databases contain useful chemical information, but they also contain extraction errors, missing reagents, inconsistent role assignments, and work-up species that are not reaction-driving conditions (Schneider et al., 2016; Thakkar et al., 2020; Toniato et al., 2021; Andronov et al., 2023). Recent efforts have improved reaction data quality through preprocessing pipelines, role assignment, benchmark construction, noise analysis, and automated re-extraction or correction (Wigh et al., 2024; Lee et al., 2024; Yuan et al., 2025). These studies show that data quality can determine what a model learns. UniCon complements this line of work by using a learned plausibility scorer to compare archived labels with candidate alternatives, allowing suspect condition records to be flagged or refined within a self-curation loop.

2.3. Evaluation Beyond Exact Match

Reaction-condition prediction is commonly evaluated with exact match, partial match, or role-specific top-k accuracy against archived labels. Such metrics are convenient but can be chemically misleading when multiple protocols are viable for the same transformation (Ball et al., 2025; Wang et al., 2025). A prediction may fail exact match while still being chemically reasonable, and an exact match may reproduce an incomplete or noisy historical record. Recent studies

have therefore called for evaluation criteria that account for diversity and noisy supervision (Ball et al., 2025; Wang et al., 2025; Yan et al., 2025). UniCon builds on this approach but makes the evaluation target explicit. Instead of testing whether a prediction reproduces a stored condition set, it assesses whether the proposed conditions are more plausible than the archived labels.

2.4. Preference-Based Scoring and Learned Evaluation

When a single reference label is insufficient, relative scoring provides a natural way to compare candidate outputs. Pairwise ranking objectives have long been used to learn scoring functions from relative comparisons, as in Bradley-Terry models and neural learning-to-rank methods such as RankNet (Bradley & Terry, 1952; Burges et al., 2005). Preference and reward models have also been widely used in generative modeling to evaluate or optimize outputs from pairwise or ranked comparisons (Christiano et al., 2017; Ouyang et al., 2022; Xu et al., 2023). In chemical reaction modeling, recent work has explored ranking-based formulations for prioritizing reaction conditions rather than predicting a single condition label (Shim et al., 2025). UniCon adapts this perspective to reaction-condition assessment. Given a reaction, the model assigns plausibility scores to candidate condition sets and converts score differences into a relative preference probability. This learned relative score supports plausibility-aware evaluation and dataset self-curation under noisy, non-unique supervision.

3. Method

3.1. Problem Formulation

We formulate reaction-condition evaluation as plausibility scoring. Let $\mathcal{R}$ be a reaction graph containing the reactants and products, and let $\mathcal{C} = (c_1, c_2, \ldots, c_K)$ be a condition set. Instead of asking the model to recover one fixed label, we learn a scoring function $s(\mathcal{R}, \mathcal{C})$ that estimates how plausible a condition set is for a reaction. The score compares alternative condition sets for the same reaction and flags archived labels that look chemically questionable. An auxiliary condition-prediction model can provide the alternative condition sets used in evaluation and self-curation.

3.2. UniConScore as an Evaluation Metric

UniCon scores each reaction-condition pair $(\mathcal{R}, \mathcal{C})$ with a single scalar. A higher score means the model finds the condition set more plausible for that reaction. For evaluation, we compare a predicted condition set with the recorded protocol through a Bradley-Terry preference model:

$$\text{UniConScore} := P(\text{Pred} \succ \text{GT}) = \sigma(s_{\text{Pred}} - s_{\text{GT}}), \quad (1)$$

where s_{Pred} and s_{GT} are the scores for the predicted and recorded condition sets. Scores above 0.5 favor the prediction over the archived label, while scores near 0.5 indicate little preference between the two. The metric is therefore a test of chemical compatibility, not exact string overlap.

3.3. Candidate Proposal and Self-Curation

For self-curation, UniCon needs candidate condition sets to compare with archived labels. An external proposal routine proposes candidates for each training reaction. The scorer then evaluates the candidate and archived condition sets side by side. If a candidate receives a much higher score, the archived label can be reviewed or replaced. The revised data can be used to retrain the scorer in another curation round.

3.4. UniCon Architecture

UniCon uses a shared multimodal transformer that combines reaction structure with condition representations to score candidate condition sets (see Figure 2). The model has a reaction encoder, a condition encoder, and a shared transformer backbone for alignment and plausibility scoring.

Reaction Encoder. To encode chemical transformations, we use a Directed Message Passing Neural Network (D-MPNN) (Heid & Green, 2021). The reaction graph $\mathcal{R}$ is converted into molecular graphs, and message passing aggregates local information. The encoder returns atom-level embeddings for cross-attentive scoring and a pooled reaction context for global alignment. The backbone therefore receives both local and pooled views of the reaction.

Condition Encoder. We map training examples into a shared vocabulary over catalyst, solvent, and reagent roles, then pack them into the fixed role order described above. In the default configuration, each indexed condition token retrieves a fixed Morgan fingerprint (Rogers & Hahn, 2010), which an MLP projects into the model space. This keeps the training label space consistent and still exposes molecular structure to the scorer. At evaluation time, out-of-vocabulary reagents can be scored by supplying fingerprint tensors directly.

Multimodal Transformer Backbone. The backbone processes condition embeddings with task-specific tokens for Reaction-Condition Alignment (RCA) and Reaction-Condition Plausibility (RCP) scoring. The design follows the architecture used in ALBEF (Li et al., 2021), but uses reaction-condition pairs instead of image-text pairs. For global alignment, the condition stream is processed without cross-attention to the reaction. For plausibility scoring, condition embeddings attend to atom-level reaction features. Thus the same backbone handles global alignment and pairwise plausibility scoring.

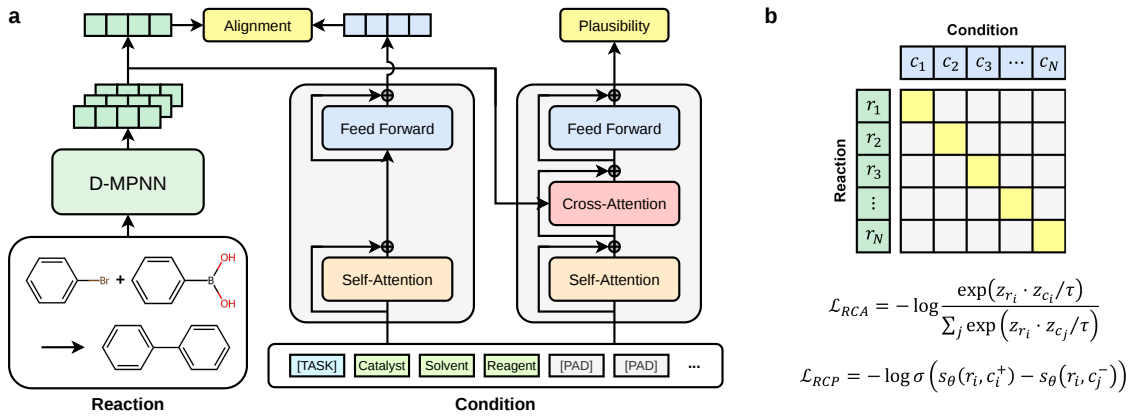


Figure 2. **Architecture of the UniCon model.** (a) The reaction graph is encoded via a directed message passing neural network. Concurrently, role-ordered condition tokens are processed through a transformer backbone, splitting into distinct alignment and plausibility pathways. The alignment pathway establishes a shared latent space via self-attention, whereas the plausibility pathway utilizes cross-attention over reaction memory to evaluate reaction-condition pairs. (b) The reaction-condition alignment employs a contrastive objective, assigning matched pairs as positive instances. Plausibility scoring is optimized via a Bradley-Terry preference objective, prioritizing recorded conditions over role-preserving hard negatives.

3.5. Training Objectives

Reaction-Condition Alignment (RCA). RCA learns a shared space for reactions and condition sets. Let $\mathbf{z}_{r,i}$ and $\mathbf{z}_{c,i}$ be the L2-normalized reaction and condition embeddings from the RCA path. We compute the batch similarity matrix as $A_{ij} = \frac{\mathbf{z}_{r_i}^\top \mathbf{z}_{c_j}}{\tau}$, where τ is a learnable temperature parameter. Let $\mathcal{V}(i) = \{j \in \{1, \ldots, B\} \mid j = i \text{ or } \mathcal{C}_j \neq \mathcal{C}_i\}$ denote the valid RCA denominator for reaction i. The symmetric contrastive loss for RCA is:

$$\begin{aligned} \mathcal{L}_{r\to c} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(A_{ii})}{\sum_{j\in\mathcal{V}(i)} \exp(A_{ij})}, \\ \mathcal{L}_{c\to r} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(A_{ii})}{\sum_{j\in\mathcal{V}(i)} \exp(A_{ji})}, \\ \mathcal{L}_{\text{RCA}} &= \frac{1}{2}\left(\mathcal{L}_{r\to c} + \mathcal{L}_{c\to r}\right). \end{aligned} \tag{2}$$

This pulls each recorded reaction-condition pair together while pushing valid in-batch mismatches apart. It helps align the embedding space beyond individual pairwise matches.

Reaction-Condition Plausibility Scoring (RCP). RCP learns the plausibility score. It assigns each reaction-condition pair a scalar score $S_{ij} = s_\theta(\mathcal{R}_i, \mathcal{C}_j)$ using cross-attention. The training set reports one successful condition set per reaction, without failed trials or explicit preference labels. For each matched pair, we create synthetic hard negatives by replacing one condition token. The replacement comes from the same role pool and is sampled in proportion to its training frequency. For reaction i, let $\mathcal{C}_i^+$ be the reported condition set and $\mathcal{C}_i^-$ be its synthetic hard negative. The scorer assigns $s_i^+ = s_\theta(\mathcal{R}_i, \mathcal{C}_i^+)$ and $s_i^- = s_\theta(\mathcal{R}_i, \mathcal{C}_i^-)$. We treat the reported condition set as preferred over the synthetic alternative. With the pairwise margin $\Delta_i = s_i^+ - s_i^-$, this defines a logistic ranking loss (Bradley & Terry, 1952; Burges et al., 2005):

$$\mathcal{L}_{\text{RCP}} = -\frac{1}{B}\sum_{i=1}^{B} \log \sigma(\Delta_i). \tag{3}$$

This objective trains the scorer to rank the recorded condition set above close role-preserving alternatives.

Momentum Distillation (MoD). Following ALBEF (Li et al., 2021), UniCon applies MoD to the RCA objective to reduce sensitivity to noisy archived records. We keep a momentum teacher with parameters $\bar{\theta}$, updated as an exponential moving average of the student parameters θ. We replace one-hot targets with a convex combination of the hard label and the teacher's soft pseudo-label. Let $q_{ij}^{r\to c} = (1-\alpha)\mathbb{1}[i=j] + \alpha \bar{p}_{ij}^{r\to c}$ and $q_{ji}^{c\to r} = (1-\alpha)\mathbb{1}[i=j] + \alpha \bar{p}_{ji}^{c\to r}$ be the smoothed targets, where $\bar{p}$ is the teacher softmax distribution and p is the corresponding student distribution. The distilled RCA objective is:

$$\begin{aligned} \mathcal{L}_{r\to c}^{\text{distill}} &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{V}(i)} q_{ij}^{r\to c} \log p_{ij}^{r\to c}, \\ \mathcal{L}_{c\to r}^{\text{distill}} &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{V}(i)} q_{ji}^{c\to r} \log p_{ji}^{c\to r}, \\ \mathcal{L}_{\text{RCA}}^{\text{distill}} &= \frac{1}{2}\left(\mathcal{L}_{r\to c}^{\text{distill}} + \mathcal{L}_{c\to r}^{\text{distill}}\right). \end{aligned} \tag{4}$$

We train RCA and RCP together with the summed loss, $\mathcal{L} = \mathcal{L}_{\text{RCA}}^{\text{distill}} + \mathcal{L}_{\text{RCP}}$.

4. Experiments

4.1. Dataset and Experimental Setup

For comparability with prior condition prediction work, we train UniCon on the USPTO split released with Reacon (Wang et al., 2025). The preprocessing removes patent records with unparseable SMILES or rare reaction templates. It keeps reactions with at most one catalyst, two solvents, and three reagents. The final dataset contains 690,872 reaction-condition pairs, with 439 catalysts, 542 solvents, and 2,746 reagents. We use the same random split, with 80% for training, 10% for validation, and 10% for testing.

We train with AdamW, a batch size of 16, an initial learning rate of 1×10^{-4}, and a peak learning rate of 1×10^{-3}. The learning rate warms up linearly for the first 3% of training and then decays to zero with a cosine schedule. UniCon jointly optimizes the RCA and RCP losses for 1 epoch. The temperature parameter τ for the RCA loss is initialized at 0.07 and updated during training. For MoD in the RCA objective, the weight α increases linearly from 0 to 0.4 during warm-up, and the momentum coefficient is 0.995. All models are trained on a single NVIDIA RTX 5070 GPU.

4.2. Expert-Aligned Plausibility Evaluation

Exact-match metrics treat every deviation from the historical record as wrong, even when the alternative is chemically workable. We test whether the learned plausibility score agrees with expert judgment. We ran a blind A/B study with senior synthetic chemists on 50 reactions sampled from the test set. For each reaction, experts were shown the reaction graph and two condition sets, one from the archived label and one proposed by a Reacon condition predictor (Wang et al., 2025). They then selected condition set A, condition set B, both, or neither as plausible for the reaction.

Table 1 reports the results. The archived condition record was often not the only reasonable answer. Among the 50 reviewed reactions, experts preferred the independent model's protocol in 21 cases and the archived ground-truth protocol in 13 cases. They judged both protocols acceptable in 12 cases and rejected both in 4 cases. Restricting to the 34 cases with a strict preference, the alternative protocol was preferred in 61.8% of cases. Most disagreements came from routine annotation issues, not unusual chemistry. Some archived labels omitted essential bases such as tert-butyllithium, misidentified work-up reagents like water as reaction conditions, or logged leaving-group byproducts. These annotations illustrate why exact match alone is a poor proxy for condition plausibility.

The two baselines provide useful context for this agreement rate. Because each strict-preference case requires choosing between two condition sets, uniform random selection has an expected accuracy of 50.0%. The popularity baseline, which chooses more frequently observed conditions, improves only modestly to 58.8%, suggesting that frequency capture some broad dataset priors but do not reliably resolve reaction-specific plausibility. Thus, expert alignment cannot be explained simply by choosing common protocols. It requires sensitivity to the compatibility between the reaction and the proposed condition set.

Table 1. Blind expert preferences comparing archived records with proposed condition sets, and UniCon agreement on cases with a strict preference.

Expert Judgment	**Count**
Alternative preferred	21
Archived preferred	13
Both protocols acceptable	12
Neither protocol acceptable	4
Model	**Accuracy (%)**
UniCon	**94.1**
Popularity	58.8
Random	50.0

Notably, in 94.1% of the cases with a strict expert preference, the preference implied by UniConScore agreed with expert judgment. These cases match the intended use of UniConScore, where the archived label may be incomplete and a different protocol may still be chemically reasonable. Among the highest-scoring examples, UniCon often preferred protocols that were more complete than the archived label. For instance, in a silylation reaction of an alcohol using a silyl chloride, the archived label specifies tetrahydrofuran and an amine base. The candidate pipeline instead predicted dimethylformamide as the solvent and imidazole as the catalyst. Although this prediction fails exact-match evaluation, Corey & Venkateswarlu (1972) reported that silylation under conventional conditions in tetrahydrofuran proceeds slowly and gives unsatisfactory yields, whereas imidazole in dimethylformamide is highly effective. UniConScore prefers the predicted condition over the archived label with a probability of 0.9899. Here, the score follows chemical viability rather than exact agreement with the dataset record.

4.3. External validation on high-throughput experimentation data

We evaluate UniConScore outside the patent-derived labels used for training. The Buchwald-Hartwig high-throughput experimentation (HTE) dataset reports measured yields for a dense grid of condition combinations (Ahneman et al., 2018). It studied a palladium-catalyzed Buchwald–Hartwig amination of aryl and heteroaryl halides with 4-

Table 2. Zero-shot UniCon performance and reagent coverage on the Buchwald-Hartwig HTE dataset. Most reagents in this benchmark are absent from the USPTO training split.

Dataset Statistics	**Value**
Unique Reagents (Unknown / Total)	28 / 31 (90.3%)
Occurrences (Unknown / Total)	13,513 / 19,775 (68.3%)
Datapoints w/ Unknown Reagents	3,955 / 3,955 (100.0%)
Metric	**Performance**
Spearman Rank Correlation (ρ)	0.3851
AUROC (Yield $\geq$ 10%)	0.6864
Enrichment Factor:	
Top 1%	1.61×
Top 5%	1.47×
Bottom 1%	0.22×
Bottom 5%	0.61×

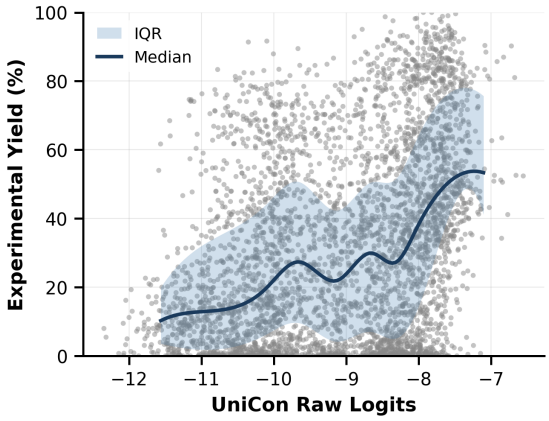


Figure 3. Zero-shot evaluation on the Buchwald-Hartwig high-throughput experimentation dataset, comparing measured yield with UniCon raw score.

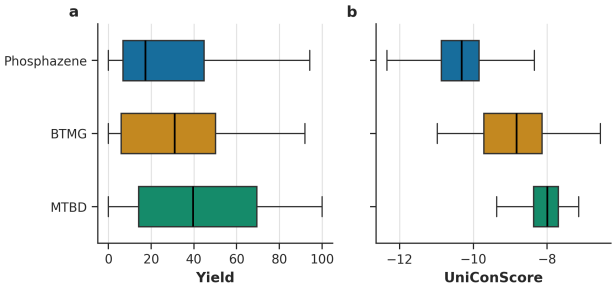


Figure 4. Base-dependent distributions of measured yield and UniCon raw score in the HTE dataset.

methylaniline under a combinatorial condition matrix. The experimental design varied four components, including 15 aryl or heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives, and produced 4,608 reactions including controls. Following the commonly used preprocessed version of this dataset, we exclude control reactions, reactions involving additive 7, and entries with missing yield values, resulting in 3,955 datapoints. In our evaluation, the aryl halide is treated as the coupling substrate, while the ligand, base, additive, and fixed reaction components are represented as condition agents.

This benchmark is largely open-set and therefore tests whether UniCon can rank successful condition combinations above unsuccessful ones outside its training distribution. As shown in Table 2, 28 of the 31 unique reagents in the HTE set are absent from the training vocabulary, and every datapoint contains at least one unseen reagent. Closed-vocabulary condition models are therefore limited in this regime. UniCon can still score each reaction-condition pair because the condition encoder receives molecular fingerprints rather than only learned reagent IDs. The raw plausibility score s_{raw} has a Spearman correlation of 0.3851 with experimental yield, averaged on aryl halide groups. For a screening threshold, s_{raw} achieves an AUROC of 0.6864 for identifying reactions with yield $\geq 10\%$.

The trends in Figure 3 show that UniCon remains informative in this near zero-shot regime. Since every HTE datapoint contains at least one reagent unseen during training and most unique reagents are outside the training vocabulary, the model cannot rely on memorized condition identities. Instead, the score appears to transfer through the molecular fingerprint representation of the condition agents. Low and high s_{raw} values concentrate among low- and high-yield reactions, so the scorer can help separate weak condition combinations from stronger ones before experimental screening. The enrichment analysis supports the same conclusion. Conditions in the top 1% by s_{raw} achieve a median yield of 46.3%, which is 1.61-fold higher than the random-selection median of 28.76%. In contrast, conditions in the bottom 1% have a median yield of only 6.3%, corresponding to 0.22-fold of the random baseline. UniCon is not trained as a yield predictor, but these results show that its plausibility signal transfers to external HTE data and can help prioritize conditions for screening.

As shown in Figure 4, the score distribution recovers a base-dependent split in the HTE measurements. Although the distributions overlap, the central tendencies for both yield and s_{raw} exhibit clear shifts dictated by the chemical identity of the base. MTBD, a bicyclic guanidine base, has the highest median yield at 39.6% and a corresponding median s_{raw} of -7.996. BTMG and a phosphazene derivative yield medians of 31.0% and 17.3%, with s_{raw} of -8.827 and -10.318, respectively.

This result is consistent with generalization through molecular representations rather than memorization of reagent identifiers. MTBD and the phosphazene base are absent from the training data, yet the model ranks their relative performance in the same order as the measured yields. UniCon assigns the highest score to MTBD, which is less sterically

Table 3. Comparison of Reacon and the popularity baseline under UniConScore, exact match, and partial match.

Model	UniConScore	Exact Match	Partial Match
Reacon	0.3009	9.11%	37.44%
Popularity	0.2085	1.90%	13.01%

hindered than the other bases in this comparison. That ranking is chemically plausible because lower steric demand can make transition-state formation more favorable in this reaction family.

4.4. Plausibility-Aware Reassessment of Baseline Models

We use UniConScore to reassess reaction-condition prediction models beyond exact-match metrics. We sample 1,000 reactions from the test set and evaluate top-1 predictions from Reacon (Wang et al., 2025) and a popularity baseline. The popularity baseline directly uses the extracted reaction template. When a test reaction matches a template observed in the training set, it predicts the most frequently observed condition set associated with that template. When no matching template is found, it abstains from prediction. Table 3 reports exact match, partial match, and UniConScore.

The model can also identify cases where the archived label is chemically incomplete. One example is the chemoselective reduction of a functionalized β-lactam, where a specific ethyl ester is reduced to a primary alcohol. The archived ground-truth label specifies only ethanol, a solvent mechanistically incapable of acting as a reducing agent. The candidate protocol correctly proposes the combination of ethanol and sodium borohydride. While standard exact-match and partial match evaluation penalizes this prediction for deviating from the database entry, the inclusion of a hydride donor is an absolute requirement for the transformation. UniCon registers this mechanistic necessity, assigning a UniConScore of 0.9999 to the candidate protocol. This example shows how the score can favor a chemically necessary condition set over an incomplete historical label.

4.5. Self-Curation of Noisy Reaction-Condition Data

We evaluate whether the plausibility model can improve training data through self-curation. An external candidate-proposal routine first proposes condition sets for training reactions. Here, Reacon supplies the candidates (Wang et al., 2025), but the curation rule itself is not tied to the predictor. The plausibility scorer then compares each candidate with the archived label. If the candidate set is preferred with probability at least 0.75, the archived label is flagged as implausible in the curated dataset.

The filtering process flagged 29.0% of the original entries. In practice, the most informative cases are those where the archived label appears incomplete, chemically inconsistent, or dominated by work-up terms. For example, one archived label specifies only toluene, which is a solvent and cannot drive the depicted transformation by itself. The candidate protocol adds active reagents, including p-toluenesulfonic acid and butyllithium, alongside the solvent. The model assigns a preference score of 0.9366 to the candidate, reflecting the missing reactivity in the archived record. These results indicate that UniCon can help curate reaction-condition corpora assembled from noisy historical sources. This procedure can be iterated to retrain the scorer on a cleaner dataset and further refine the training set.

Table 4. Ablation study evaluating the contribution of the RCA loss.

Metric	UniCon	w/o RCA
Expert Agreement	94.1%	85.3%
Spearman	0.3851	0.2134
EF@Top1%	1.61×	0.90×
EF@Bottom1%	0.22×	1.08×

4.6. Ablation Study

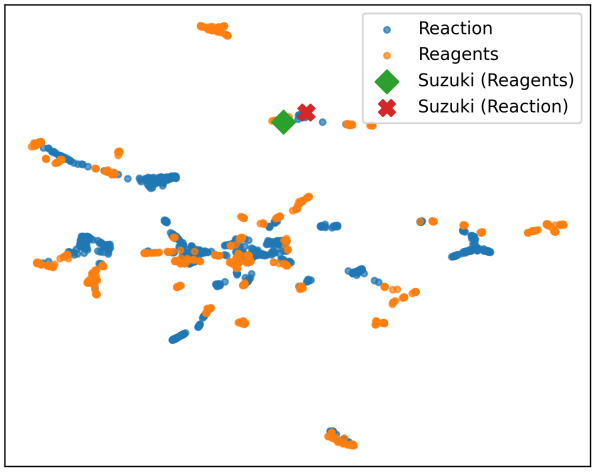
We conduct an ablation study to evaluate the contribution of the RCA objective. This comparison changes only the objective terms, while keeping the data and candidate proposal procedure fixed. Table 4 reports agreement with strict expert preferences, Spearman correlation, and enrichment factors on the HTE benchmark.

Removing the RCA loss shows a drop in all metrics. This is consistent with prior work on multimodal alignment (Li et al., 2021), which demonstrates that globally aligning unimodal representations prior to cross-modal pairwise scoring improves representation learning.

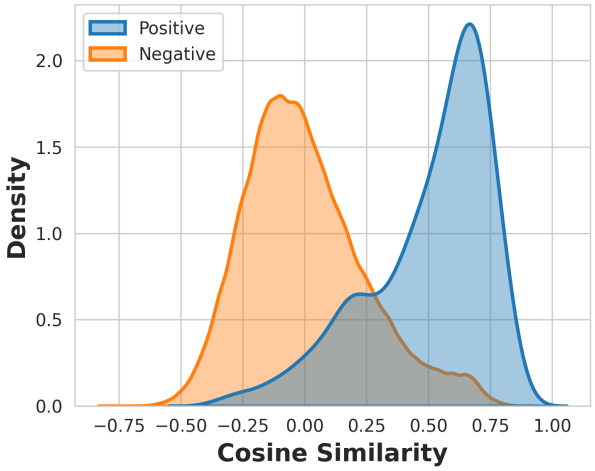
4.7. Analysis of the Latent Compatibility Space

We also examine the geometry of the learned compatibility space. The embeddings provide a qualitative check on whether compatible and incompatible reaction-condition pairs are separated.

As shown in Figure 5a, when projected into two dimensions using UMAP (McInnes et al., 2018), the latent space groups reaction-condition pairs into chemically coherent neighborhoods, even without explicit reaction-class supervision. For example, the highlighted Suzuki coupling reaction and reagent embeddings map closely to one another in the learned space. The cosine similarity distributions in Figure 5b further show that positive and negative reaction-condition pairs are separated. The visualizations support the narrower claim that UniCon learns a compatibility representation useful for reaction-condition assessment, prioritiza-



(a) Reaction-condition compatibility space.



(b) Cosine similarity distribution of pairs.

Figure 5. Visualizations of the learned compatibility space. (a) The UMAP projection groups reaction-condition pairs by compatibility patterns, including the highlighted Suzuki coupling example. (b) The density plot compares cosine similarity for compatible and incompatible reaction-condition pairs.

tion, and curation.

5. Discussion

UniConScore captures relative reaction-condition plausibility under noisy supervision. In practice, it can compare a predicted protocol with an archived label, keep plausible candidates during self-curation, and rank condition sets for external screening.

The score is still learned from historical reaction-condition data. It therefore inherits missing reaction context, reporting bias, and gaps in the training distribution. For some reactions, multiple protocols may be similarly plausible yet differ in yield, robustness, cost, or operational convenience. UniCon is useful for protocol assessment, curation, and prioritization, but it does not replace experimental validation.

The expert preference study and HTE benchmark test this claim from different angles. Agreement with chemist judgments indicates that the score captures protocol-level plausibility in cases where exact match is chemically misleading. The HTE correlation shows that the score can also rank unseen condition combinations in an external dataset. Together, these results make a case for benchmarking and self-curation methods that use learned plausibility rather than exact label recovery alone.

Self-curation can propagate bias if the training set systematically underrepresents a chemistry family or omits context such as concentration, order of addition, or atmosphere. Rare transformations are also hard to score reliably when few related examples appear during training. Future work should test UniCon across more reaction classes and include richer experimental metadata, so the scorer can use more of the actual protocol context.

6. Conclusion

We presented UniCon as a chemistry-grounded framework for evaluating, curating, and prioritizing reaction conditions under non-unique and noisy supervision. Instead of treating the archived label as the only correct answer, UniCon learns a plausibility score that compares alternative protocols, filters noisy labels during self-curation, and ranks candidates in zero-shot HTE screening. The results point to a practical direction for reaction-condition models that account for plausible alternatives and dataset noise, rather than only rewarding strict reproduction of historical labels.

References

Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D., and Doyle, A. G. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360 (6385):186–190, 2018.

Andronov, M., Voinarovska, V., Andronova, N., Wand, M., Clevert, D.-A., and Schmidhuber, J. Reagent prediction with a molecular transformer improves reaction data quality. *Chemical Science*, 14(12):3235–3246, 2023.

Ball, M., Horvath, D., Kogej, T., Kabeshov, M., and Varnek, A. Predicting reaction conditions: a data-driven perspective. *Chemical Science*, 16(38):17523–17541, 2025.

Bradley, R. A. and Terry, M. E. Rank analysis of incom-

plete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Chen, B., Li, C., Dai, H., and Song, L. Retro*: learning retrosynthetic planning with neural guided a* search. In *International conference on machine learning*, pp. 1608–1616. PMLR, 2020.

Chen, S. and Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Corey, E. and Venkateswarlu, A. Protection of hydroxyl groups as tert-butyldimethylsilyl derivatives. *Journal of the American Chemical Society*, 94(17):6190–6191, 1972.

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., and Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476, 2018.

Heid, E. and Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, 2021.

Kim, J., Ahn, S., Lee, H., and Shin, J. Self-improved retrosynthetic planning. In *International Conference on Machine Learning*, pp. 5486–5495. PMLR, 2021.

Lawson, A. J., Swienty-Busch, J., Géoui, T., and Evans, D. The making of reaxys—towards unobstructed access to relevant chemistry information. In *The Future of the History of Chemical Information*, pp. 127–148. ACS Publications, 2014.

Lee, C., Chen, S., Ong, K. T.-i., Yeo, J., and Jung, Y. Noise analysis and data refinement for chemical reactions from us patents via large language models. 2024.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Lowe, D. Chemical reactions from us patents (1976-sep2016). 2017.

Maser, M. R., Cui, A. Y., Ryou, S., DeLano, T. J., Yue, Y., and Reisman, S. E. Multilabel classification models for the prediction of cross-coupling reaction conditions. *Journal of Chemical Information and Modeling*, 61(1): 156–166, 2021.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Sacha, M., Błaz, M., Byrski, P., Dabrowski-Tumanski, P., Chrominski, M., Loska, R., Włodarczyk-Pruszynski, P., and Jastrzebski, S. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284, 2021.

Schneider, N., Stiefl, N., and Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12): 2336–2346, 2016.

Shim, E., Tewari, A., Cernak, T., and Zimmerman, P. M. Recommending reaction conditions with label ranking. *Chemical Science*, 16(9):4109–4118, 2025.

Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O., and Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical science*, 11(1): 154–168, 2020.

Toniato, A., Schwaller, P., Cardinale, A., Geluykens, J., and Laino, T. Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence*, 3(6):485–494, 2021.

Tu, Z. and Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling*, 62(15):3503–3513, 2022.

Wang, M. and Montana, G. Retrosynthesis planning via worst-path policy optimisation in tree-structured MDPs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=m7uj1vIZ62.

Wang, X., Hsieh, C.-Y., Yin, X., Wang, J., Li, Y., Deng, Y., Jiang, D., Wu, Z., Du, H., Chen, H., et al. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research*, 6:0231, 2023.

Wang, Z., Lin, K., Pei, J., and Lai, L. Reacon: a template- and cluster-based framework for reaction condition prediction. *Chemical Science*, 16(2):854–866, 2025.

Wigh, D. S., Arrowsmith, J., Pomberger, A., Felton, K. C., and Lapkin, A. A. Orderly: data sets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling*, 64(9):3790–3798, 2024.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935, 2023.

Yan, X., Zhong, H., and Wang, X. Robust chemical reaction condition recommendations via label mix strategy. *Journal of Chemical Information and Modeling*, 65(23): 12775–12785, 2025.

Yuan, S., Gong, S., and Xu, H. Uspto-llm: A large language model-assisted information-enriched chemical reaction dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 817–820, 2025.

Zhao, D., Tu, S., and Xu, L. Efficient retrosynthetic planning with mcts exploration enhanced a* search. *Communications Chemistry*, 7(1):52, 2024.