Bioinformatics, 2026, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

A cross-attentive multi-task graph learning framework for chemical reaction modeling

Maryam Astero,[1,*] Anchen Li,[1] Elena Casiraghi[1,2] and Juho Rousu[1,3,*]

[1]Department of Computer Science, Aalto University, Finland, [2]Department of Computer Science, Università degli Studi di Milano, Italy and [3]Department of Computer Science, University of Helsinki, Finland
[*]Corresponding authors. maryam.astero@aalto.fi, juho.rousu@aalto.fi

Abstract

Motivation: Understanding chemical reactions requires bridging fine-grained molecular edits with broader semantic context. Reaction mechanisms are determined not only by local atom–bond transformations but also by the global reaction class. However, most existing approaches treat these tasks separately or rely on external atom-mapping tools, introducing noise and limiting end-to-end learnability. We introduce MARCC (Mapping-Assisted Reaction Center and Classification), a multi-task graph neural network that jointly predicts atom mappings, reaction centers, and reaction classes within a unified architecture.

Results: MARCC integrates three key innovations: (i) a mapping-guided cross-attention mechanism that aligns reactants and products for local edit detection, (ii) a dual-graph design that explicitly reasons about bond-level transformations, and (iii) pooled product embeddings for global reaction classification. On the USPTO-50K benchmark, MARCC achieves state-of-the-art results when trained with both reactants and products, including 98.2% atom mapping accuracy, 99.1% Top-1 edit localization accuracy, and 97.2% reaction classification accuracy. Even under the products-only setting, MARCC delivers competitive performance comparable to specialized baselines. Ablation studies confirm the value of mapping-guided attention and multi-task supervision, which enhance both predictive accuracy and interpretability. By unifying atom-level alignment, local reactivity, and global classification, MARCC provides a structured and interpretable framework for reaction understanding. Beyond benchmarks, MARCC has the potential to support applications in reaction annotation, template discovery, and mechanism inference; with additional domain-specific modeling and data, it could be extended to biochemical domains such as enzyme-catalyzed transformations and metabolic pathway modeling.

Availability and Implementation: The source code and implementation details are available at `https://github.com/maryamastero/MARCC` and archived at `https://doi.org/10.5281/zenodo.18500230`.

Contact: maryam.astero@aalto.fi, juho.rousu@aalto.fi

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Introduction

Understanding chemical reactions requires reasoning across multiple structural and semantic levels. Three core tasks underpin this process: *atom mapping*, which establishes atom-level correspondences between reactants and products; *reaction center identification*, which pinpoints the atoms and bonds altered during the transformation; and *reaction classification*, which captures the overall transformation type. Together, these tasks provide complementary perspectives on molecular reactivity and are essential for applications where both reaction sides are available, including mechanism inference, reaction template discovery, role assignment, and reaction annotation, as well as supervised training settings for retrosynthesis.

Atom mapping, traditionally solved by graph-theoretic search (Raymond and Willett, 2002; Heinonen et al., 2011), is now increasingly addressed by attention- or graph-based methods (Schwaller et al., 2021; Nugmanov et al., 2022; Astero and Rousu, 2024, 2025). Reaction center prediction initially relied on rule-based or template-driven systems (Coley et al., 2017; Dai et al., 2019), later reformulated as graph-labeling or subgraph selection tasks (Yan et al., 2020; Somnath et al., 2021; Lan et al., 2024). Reaction classification has similarly evolved from fingerprint-based methods (Gelernter et al., 1990; Schneider et al., 2015; Probst et al., 2022) to graph neural network models (Li et al., 2024). Despite these advances, most approaches treat these tasks in isolation, overlooking their interdependence: atom mappings provide a structural scaffold for localizing reactive centers, and both local edits and mappings jointly constrain global reaction classes.

Prior research on atom mapping, such as *SAMMNet* (Astero and Rousu, 2025), demonstrated that multitask learning can partially mitigate the inherent incompleteness of reaction data by coupling atom mapping with auxiliary supervision. Specifically, SAMMNet focuses primarily on correspondence prediction and incorporates a self-supervised auxiliary task of atom-type prediction to improve generalization and robustness in situations with incomplete reaction

information. However, it did not explicitly unify local reaction mechanisms with global reaction semantics, nor did it model bond-level transformations or reaction-level abstractions within a single framework.

We introduce MARCC (Mapping-Assisted Reaction Center and Classification), a cross-attentive multi-task framework that explicitly unifies atom mapping, reaction center identification, and reaction classification within a single learning architecture. MARCC leverages mapping-guided cross-attention to align reactant and product graphs, enabling joint reasoning over local structural changes and global reaction semantics. Furthermore, a dual-graph representation treats bonds as explicit nodes, allowing the model to directly capture bond-breaking and bond-forming events that define reaction mechanisms.

This unified design enables MARCC to jointly reason over atom-mapping prediction, mechanistic edits, and reaction semantics, rather than treating these components as independent prediction problems.

Evaluated on the USPTO-50K benchmark, MARCC delivers state-of-the-art performance in reaction center identification and classification, while achieving mapping accuracy on par with specialized systems. These results establish MARCC as a powerful and interpretable framework for chemical reaction understanding. Beyond chemical informatics, the approach has natural extensions to biological settings: many biochemical processes, such as enzyme-catalyzed transformations, metabolic pathway progression, and drug metabolism, can be represented as graph-structured reactions with localized centers and global functional classes. By jointly learning atom-level correspondences, reactive sites, and overall reaction types, MARCC offers a structured foundation for applications in systems biology, enzymology, and pharmacokinetics.

Our key contributions are as follows:

- We propose a unified multi-task GNN architecture that simultaneously performs reaction classification, reaction center prediction, and atom mapping within a single framework.
- We introduce a cross-graph attention mechanism that explicitly aligns reactant and product atoms, enabling joint inference across molecular structures.
- We design a dual-graph representation in which bonds are treated as explicit nodes, allowing the model to capture bond-level transformations more effectively.
- We develop an imbalance-aware optimization scheme that combines focal and dice losses to improve robustness on underrepresented reaction classes and rare transformation patterns.

To ensure fair comparison with prior reaction-center baselines, which typically operate in a retrosynthesis setting using only product graphs, we additionally evaluate a *products-only* variant of MARCC. In this setting, the cross-attention module is disabled for reactivity prediction and used solely for classification, aligning MARCC with the input constraints of existing baselines.

Problem Formulation

We formulate chemical reaction modeling as a multi-task learning problem over molecular graph transformations. Each reaction instance is represented by a pair of molecular graphs: the major product G_P and its corresponding reactants G_R, where atoms are nodes and bonds are edges.

We define a molecular graph as $G = (V, A, X, E)$, where V is the set of atoms, $A \in \{0,1\}^{|V|\times|V|}$ is the adjacency matrix, $X \in \mathbb{R}^{|V|\times d_x}$ encodes atom features (e.g., element, charge, hybridization), and $E \in \mathbb{R}^{|E|\times d_e}$ encodes bond features (e.g., bond type, conjugation, ring membership). Notably, G is not required to be a single connected component; it may represent a set of multiple molecules (e.g., reactants and reagents), allowing the model to represent complex multi-component reaction systems without structural modification.

Given a reaction (G_P, G_R), our goal is to jointly model two structurally coupled tasks—*reaction center identification* and *reaction classification*—while leveraging *atom mapping* as an auxiliary alignment signal.

Atom Mapping

Atom mapping establishes a one-to-one correspondence between atoms in the reactants and those in the product, capturing transformation mechanics at atomic resolution. This information directly informs both center prediction and reaction class classification.

In practice, patent-derived data often include side products or incomplete reactant specification, making direct mappings ambiguous. To mitigate this, we reverse the mapping direction: each product atom is softly aligned to candidate reactant counterparts, analogous to retrosynthetic reasoning.

We represent this alignment as a differentiable correspondence matrix: $M \in [0,1]^{|V_P|\times|V_R|}$, $M_{i,i'} = Pr(v_i^P \leftrightarrow v_{i'}^R)$, where $M_{i,i'}$ denotes the probability that product atom v_i^P aligns with reactant atom $v_{i'}^R$. The matrix M is constrained to be doubly stochastic, i.e., each row and column sums to one, ensuring a soft one-to-one correspondence.

Atom mappings are optimized by maximizing structural consistency:

$$M^* = \arg\max_M \sum_{i,j}\sum_{i',j'} A_P(i,j)\cdot A_R(i',j')\cdot M_{i,i'}\cdot M_{j,j'}.$$

Reaction Center Identification

The reaction center is the subset of atoms and bonds in G_P that undergo structural changes relative to their aligned counterparts in G_R. These changes include bond formation/cleavage, bond order modifications, or alterations in atomic valence, hydrogen count, or charge.

We model center detection as binary classification on the product graph. Each atom $v_i^P \in V_P$ and bond $e_{ij}^P \in E_P$ is assigned binary labels: $y_i^{\text{atom}} \in \{0,1\}$, $z_{ij}^{\text{bond}} \in \{0,1\}$, where a label of 1 indicates structural reactivity. Atom mappings provide the alignment needed for pairwise comparison between G_P and G_R.

Reaction Classification

At a global level, each reaction is assigned to one of K classes (e.g., substitutions, eliminations, additions) based on transformation semantics. We model this as multi-class prediction: $c^* = \arg\max_{k\in\{1,\ldots,K\}} \Pr(c = k \mid G_P, G_R)$, where c denotes the predicted reaction class.

Mapping-Assisted Reaction Center and Classification

We propose MARCC (Mapping-Assisted Reaction Center and Classification), a multi-task graph neural architecture for joint atom mapping, reaction center prediction, and reaction classification. The tasks are coupled through a shared embedding space and mapping-guided attention (Fig. 1). A detailed algorithm is given in Supplementary Material S1.

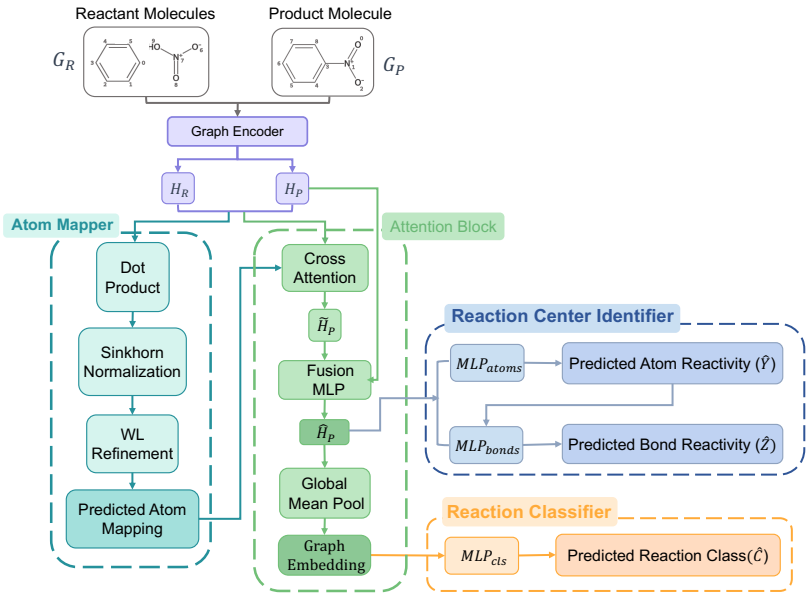

Fig. 1. Overview of the MARCC architecture. Reactant (G_R) and product (G_P) graphs are encoded by a shared GINE encoder, producing node embeddings H_R and H_P. The **Atom Mapper** computes soft alignments via dot-product similarity and Sinkhorn normalization, refined by Weisfeiler–Lehman (WL) symmetry resolution. These correspondences guide a cross-attention mechanism, yielding enriched product embeddings $\hat{H}_P$. Task-specific heads then perform **Reaction Center Identification** (atom/bond reactivity) and **Reaction Classification** (global reaction type).

Graph Encoder

Reactant and product graphs are independently encoded by a shared GINE encoder (Fey and Lenssen, 2019) with residual connections: $H_R = \text{GINE}(A_R, X_R, E_R)$, $H_P = \text{GINE}(A_P, X_P, E_P)$, where H_R and H_P denote node embeddings.

Atom Mapping

Atom correspondences are estimated by soft graph matching. A raw similarity matrix is computed as: $\hat{M} = H_P H_R^\top$, then normalized by Sinkhorn scaling (Sinkhorn and Knopp, 1967) into a doubly stochastic matrix M. Final mappings are obtained by row-wise softmax, aligning each product atom to a reactant candidate.

Symmetry Refinement.

To improve the quality of atom mapping predictions, especially in cases involving structural symmetries (e.g., aromatic rings or repetitive substructures), we apply a refinement procedure based on the Weisfeiler–Lehman (WL) test (Weisfeiler and Leman, 1968). The WL algorithm iteratively updates atomic labels via neighborhood aggregation. Atoms with identical labels across all iterations and matching atomic types are grouped into equivalence classes. These classes are used to disambiguate alignments during both training and inference, following (Astero and Rousu, 2025).

Mapping-Guided Cross-Attention

To connect mappings with downstream tasks, each product atom attends only to its most probable reactant counterpart:

$$q_i = W_Q H_P^{(i)}, \quad k_i = W_K H_R^{(M_i)}, \quad v_i = W_V H_R^{(M_i)},$$

$$\tilde{H}_P^{(i)} = \text{softmax}\left(\frac{q_i^\top k_i}{\sqrt{d_k}}\right) v_i,$$

with trainable weights W_Q, W_K, W_V. Final embeddings are obtained by concatenation: $\hat{H}_P^{(i)} = [H_P^{(i)} || \tilde{H}_P^{(i)}]$, where $||$ denotes feature-wise concatenation. Ground-truth mappings supervise training, while predicted mappings are used at inference.

Reaction Center Identification

Atom and bond reactivity are modeled as binary classification tasks on the product graph. Atom logits are: $\hat{y}_i = \sigma\left(\text{MLP}_{\text{atom}}(\hat{H}_P^{(i)})\right)$.

Bond predictions combine atom embeddings, bond features, dual-graph features, and reactivity logits:

$$\hat{z}_{ij} = \sigma\left(\text{MLP}_{\text{bond}}\left([\hat{H}_P^{(i)} || \hat{H}_P^{(j)} || E_P^{(ij)} || H_D^{ij} || \hat{y}_i || \hat{y}_j]\right)\right).$$

Dual Graph.

To capture bond-centric context, each bond is treated as a node in a dual graph, with edges connecting bonds that share an atom. Bond features form node inputs, while connectivity is inherited from shared atoms (see Supplementary Material S2).

Reaction Classification

Global semantics are captured by mean-pooling product embeddings: $\bar{H}_P = \frac{1}{|V_P|} \sum_{i \in V_P} \hat{H}_P^{(i)}$, $\hat{c} = \text{softmax}(\text{MLP}_{\text{class}}(\bar{H}_P))$.

Multi-Task Objective

The overall loss is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{map}} \mathcal{L}_{\text{map}} + \lambda_{\text{react}} (\mathcal{L}_{\text{atom}} + \mathcal{L}_{\text{bond}}) + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}.$$

Reactivity Loss.

Because reactive atoms/bonds are sparse, we combine focal loss (Lin et al., 2017) (to emphasize hard cases) with Dice loss (Milletari et al., 2016) (to optimize overlap):

$$\mathcal{L}_{\text{atom}} = \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}}^{\text{atom}} + (1 - \lambda_{\text{dice}}) \mathcal{L}_{\text{Focal}}^{\text{atom}},$$

$$\mathcal{L}_{\text{bond}} = \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}}^{\text{bond}} + (1 - \lambda_{\text{dice}}) \mathcal{L}_{\text{Focal}}^{\text{bond}}.$$

The full mathematical definitions of the Dice and focal loss terms are given in Supplementary Material S3.

Experiments

Experimental Setup

Dataset

We evaluate MARCC on the USPTO-50K benchmark dataset (Lowe, 2012; Schneider et al., 2016), using the standardized, atom-mapped release from Somnath et al. (Somnath et al., 2021). The dataset consists of 50,000 reactions annotated with atom mappings and categorized into 10 reaction classes, making it a widely adopted benchmark for reaction understanding.

To mitigate spurious correlations arising from atom index ordering, we adopt the canonicalization and remapping procedure of (Somnath et al., 2021; Astero and Rousu, 2024), which uses canonical product SMILES and regenerated alignments. This ensures that learned signals reflect true structural reasoning rather than positional artifacts.

Atom and bond reactivity labels are derived by comparing reactants and products: atoms are labeled reactive if their hydrogen count or formal charge changes; bonds are labeled reactive if broken or altered in type. As only 0.5% of bonds are newly formed, we focus on transformations over existing bonds, consistent with prior work (Somnath et al., 2021). Following the standard protocol, we adopt an 8:1:1 training, validation, and test split.

We prioritize the USPTO-50K benchmark over larger, uncurated corpora (e.g., USPTO-full) because its high-fidelity ground truth is essential for isolating the architectural contributions of our multi-task framework. This ensures that performance metrics reflect the model's structural reasoning rather than the noise or erroneous mappings prevalent in larger datasets.

Graph Construction and Features

Molecular graphs are constructed from SMILES representations using RDKit (Landrum et al., 2006). Atom features include atomic number, formal charge, chirality, hybridization, aromaticity, degree, hydrogen counts, and ring membership. Bond features capture bond type, conjugation, stereochemistry, and ring membership. All features are one-hot encoded or categorical and concatenated into atom and bond embeddings. A full specification is provided in Material S4.

Evaluation Metrics

We assess performance on reaction center prediction, reaction classification, and atom mapping.

- **Reaction Center Identification**: evaluated using Top-n Edit Accuracy ($n \in \{1, 3, 5\}$), which measures the proportion of true edits recovered among the top-ranked predictions.
- **Reaction Classification**: evaluated using Top-1 Accuracy, reflecting the proportion of reactions assigned to the correct class.
- **Atom Mapping**: evaluated using symmetry-aware accuracy (Astero and Rousu, 2025), which accounts for equivalence among symmetrically valid mappings.

Hyperparameter Search

We optimized hyperparameters using Optuna's Tree-structured Parzen Estimator (TPE) sampler (Bergstra et al., 2011). The search space covered architectural choices (embedding dimension, encoder depth, number of attention heads), optimizer settings, task weights, and loss coefficients. Continuous parameters (e.g., learning rate) were sampled on a logarithmic scale, while discrete and categorical options were sampled directly.

Hyperparameter selection was performed exclusively on the validation set. After identifying the best configuration, we retrained the model from scratch on the combined training and validation sets and reported results on the held-out test set.

Each trial was scored using a scalarized validation objective that jointly balances atom-level F1, bond-level F1, and reaction classification F1. To improve efficiency, underperforming trials were pruned early using Optuna's pruning mechanisms—primarily the Asynchronous Successive Halving (ASHA) strategy, with the Median Pruner used as a fallback (Akiba et al., 2019).

For every model–mode combination, we conducted 30 trials per study. The best-performing configuration was reused across all reported experiments to ensure comparability. Details of the search space, trial budget, and pruning strategies are provided in Supplementary Material S5.

Task Weighting

We adopt uncertainty-based weighting (Kendall et al., 2018) to balance the three tasks. Each task is assigned a learnable log-variance parameter, leading to the joint loss: $\mathcal{L}_{\text{total}} = \sum_i \left(\frac{1}{2\sigma_i^2} \mathcal{L}_{\text{task}_i} + \log \sigma_i \right)$, where σ_i denotes the homoscedastic uncertainty of task i. A larger σ_i downweights the corresponding loss, while a smaller σ_i increases its influence. This scheme adaptively rescales gradients, giving more weight to reliable tasks and reducing the impact of noisier ones. The uncertainty parameters σ_i are learned jointly with the model weights during training and remain fixed at inference.

Training Setup

All experiments were conducted on the Aalto University Triton HPC cluster using a single NVIDIA V100 GPU (32 GB memory) with PyTorch Geometric. A typical MARCC training run on the USPTO-50K dataset, with batch size 32, 5 GINEConv layers (512-dimensional embeddings), 8 attention heads, and the AdamW optimizer (learning rate 1×10^{-4}, weight decay 5×10^{-5}), required approximately 17 hours of wall-clock time. Early stopping with a patience of 10 epochs was applied, and models typically converged within 45–50 epochs.

The shared graph encoder consists of 5 GINEConv layers with dynamically weighted skip connections. The guided cross-attention module employs 8 heads and integrates atom mapping alignments. Reactivity prediction heads were optimized with a hybrid Dice–Focal loss ($\lambda_{\text{dice}} = 0.4$), with positive samples upweighted to address class imbalance. Thresholds for atom and bond reactivity were selected to maximize validation F1. Reaction classification was trained with standard cross-entropy loss across the 10 USPTO classes.

Overall Results and Baseline Comparison

We first evaluate MARCC on reaction center identification and reaction classification, comparing against strong baselines designed for retrosynthesis and reaction understanding. These include center-prediction methods (GraphRetro (Somnath et al., 2021), RetroXpert (Yan et al., 2020), RCSearcher (Lan et al., 2024)) and classification baselines (MolBERT (Fabian et al., 2020), RXGL (Li et al., 2024)). Results are summarized in Table 1.

A key consideration is that most reaction-center baselines are designed for retrosynthesis and operate solely on the product graph, without leveraging reactant information. To provide a fair comparison, we also report a *products-only* variant of MARCC, in which the model is restricted to the major product structure. In this configuration, cross-attention is disabled for reactivity prediction and used only for classification. As shown in Table 1,

Table 1. Comparison with prior methods on USPTO-50K. "–" indicates the method does not perform that task. Best results in bold.

Model	Reaction Center: Edit (%)			Reaction Classification (%)
	Top-1	Top-3	Top-5	Accuracy
GraphRetro	70.8	89.5	92.7	–
RetroXpert	50.4	61.1	62.3	–
RCSearcher	69.3	79.3	85.7	–
MolBERT	–	–	–	70.3
RXGL	–	–	–	93.2
MARCC (Products Only)	67.9	86.5	93.2	**97.6**
MARCC (Full)	**99.1**	**99.6**	**99.7**	**97.2**

MARCC in the full-reaction setting achieves a Top-1 edit accuracy of 99.1%, far exceeding product-only baselines such as GraphRetro (70.8%). Since GraphRetro and related approaches operate solely on product graphs, the improvement reflects both MARCC's richer input setting—leveraging explicit reactant–product alignment—and its integrated multi-task design. These gains are driven by three innovations: (i) multi-task supervision that couples local edits with global semantics; (ii) differentiable atom–atom alignment that resolves symmetries; and (iii) mapping-guided cross-attention that contextualizes node embeddings.

In the *products-only* setting, MARCC achieves a Top-1 edit accuracy of 67.9%. While this is comparable to product-only baselines, it highlights the intrinsic need for reactant–product alignment to reliably identify reactive atoms. Notably, classification accuracy remains very high (97.6%), surpassing specialized classifiers such as RXGL (93.2%) and MolBERT (70.3%). This suggests that while structural alignment is essential for fine-grained mechanistic inference, global reaction semantics can often be inferred directly from product structures alone.

Beyond reaction center prediction and classification, MARCC also learns atom-to-atom correspondences as part of its multitask training. In the following section, we evaluate its mapping accuracy and show that, despite not being the primary training objective, MARCC remains competitive with state-of-the-art atom mapping methods.

Atom Mapping Evaluation

Atom-to-atom alignment has been extensively studied, with several methods designed specifically for this task. To contextualize the performance of MARCC's auxiliary mapping head, we evaluate its accuracy on USPTO-50K against representative baselines: RXNMapper (Schwaller et al., 2021), GraphormerMapper (Nugmanov et al., 2022), and SAMMNet (Astero and Rousu, 2025). Table 2 summarizes accuracies across these systems.

Table 2. Atom mapping accuracy on USPTO-50K.

Model	Accuracy (%)
RXNMapper	98.8
SAMMNet	97.4
GraphormerMapper	94.5
MARCC	98.2

In the full multitask setting, MARCC achieves 98.2% symmetry-aware accuracy, closely approaching RXNMapper (98.8%) while surpassing other graph-based models. Importantly, this performance is obtained without dedicating the architecture solely to atom mapping: the mapping head serves as a structural regularizer within the joint training scheme. This demonstrates that accurate atom alignments can be learned as a byproduct of multitask supervision, while simultaneously enhancing reaction center and classification performance (see Section 4.4 for further analysis).

Ablation: Impact of Multitask Learning

To assess the contribution of each supervision signal, we ablate the MARCC training objective across seven task configurations (Table 3). Each configuration activates a subset of the available losses: reaction classification (Classification), reaction center prediction (Reactivity), and atom mapping (Mapping).

Single-task training yields narrow competence: reaction classification alone (Classification) generalizes poorly (73.3%), while reaction-center (Reactivity) and Mapping models achieve decent accuracy but lack broader contextualization. Pairwise combinations provide complementary benefits—e.g., (Classification + Reactivity) substantially improves both classification (96.3%) and edit localization (88.1%), indicating strong synergy between global semantics and local edits.

Notably, mapping-only training achieves near-perfect alignment (99.0%) yet offers little transfer, reflecting overfitting to structural correspondences. By contrast, integrating mapping with center prediction (Mapping + Reactivity) or classification (Mapping + Classification) yields clear mutual reinforcement, raising edit and classification accuracy by 5–20 points.

The full multitask model achieves the best balance across objectives: 99.1% edit accuracy, 97.2% classification, and 98.2% mapping. This suggests that atom mapping acts as a structural regularizer, center prediction sharpens local reactivity, and classification enforces global semantic consistency—together yielding robust, generalizable representations.

Furthermore, we investigate the specific impact of the dual-graph (bond-node) representation. As detailed in Supplementary Material S9, removing the dual-graph module while retaining guided cross-attention results in a significant decline in edit accuracy, dropping from 99.1% to 85.4%. This confirms that modeling bonds as first-class nodes is essential for precisely localizing chemical transformations, a task that standard atom-centric message passing fails to capture with the same fidelity.

Per-Class Performance Analysis

To assess how structural and distributional factors affect predictive performance, we analyze per-class statistics including class frequency, reactivity profiles, and MARCC's accuracy. Figure 2 summarizes MARCC's performance across reaction classes. The dataset is highly imbalanced, with Classes 1 and 2 dominating while Classes 8–10 are underrepresented, limiting generalization. Reactivity profiles vary: Class 4 shows the most complex edits, whereas Classes 6 and 9 exhibit minimal transformations. Despite this variability, MARCC achieves consistently high edit localization accuracy across all classes, even for complex or sparse categories. In

Table 3. Effect of supervision configurations on MARCC performance. Metrics include Top-1 Edit Accuracy, Reaction Classification Accuracy, and Mapping Accuracy. "–" indicates the task is not applicable. Best values in **bold**.

Configuration	Edit Acc. (%)	Class Acc. (%)	Map Acc. (%)
Classification	–	73.3	–
Reactivity	87.7	–	–
Mapping	–	–	**99.0**
Classification + Reactivity	88.1	96.3	–
Mapping + Reactivity	93.0	–	90.5
Mapping + Classification	–	91.8	97.6
Classification + Reactivity + Mapping (Full)	**99.1**	**97.2**	98.2

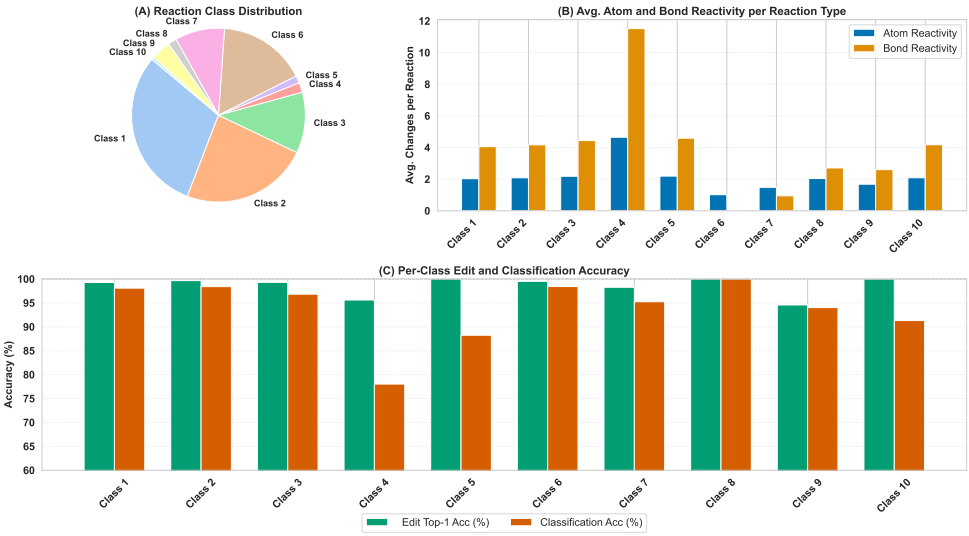

Fig. 2. Dataset Composition and Per-Class Performance of MARCC on USPTO-50K. (A) Distribution of reaction classes in the test set, showing strong imbalance across the 10 categories. (B) Average number of reactive atoms and bonds per reaction type. (C) Per-class Top-1 edit localization accuracy and reaction classification accuracy achieved by MARCC.

contrast, reaction classification accuracy is more affected by class imbalance, with weaker performance on rare classes (4, 5, 10). Importantly, prediction consistency analysis confirms that correct global classifications are strongly supported by accurate local edit predictions.

We further assess the internal consistency of MARCC's predictions by measuring the proportion of reactions where both the reaction classification and Top-1 edit prediction are correct. As shown in Supplementary Material S6, this alignment remains consistently high across classes, suggesting that global predictions are supported by accurate local mechanistic edits.

Case Study: Attention-Guided Interpretation

To further assess the interpretability of MARCC and understand the interaction between atom mapping and reaction center prediction, we present qualitative visualizations of two representative examples. Figure 3 illustrates how MARCC's cross-attention mechanism supports interpretability. In successful cases, attention weights align strongly with true reactive atoms, confirming that atom mapping and reaction center identification are correctly recovered. Even in challenging scenarios with partial symmetry, the model remains robust: despite imperfect mappings, attention focuses on chemically relevant atoms (e.g., oxygen), enabling correct reaction center prediction. These results demonstrate that cross-attention provides a reliable inductive bias, enhancing both accuracy and interpretability.

Case Study: Qualitative Analysis

Figure 4 illustrates three representative reactions from the USPTO-50K test set, highlighting MARCC's behavior across diverse prediction scenarios.

In Figure 4a, MARCC correctly predicts both atom mapping and the reaction center, including reactive bonds, reflecting the model's ability to handle non-trivial rearrangements. In Figure 4b, all atoms are mapped correctly, but the model falsely labels atom 9 (an oxygen) as reactive—likely due to overgeneralization from learned chemical priors. In Figure 4c, the reaction center is correctly predicted, but one product atom is misaligned during atom mapping. Specifically, carbon atom 17 is incorrectly mapped to reactant atom 20 instead of the ground-truth atom 24. This discrepancy likely stems from limited structural context in a small molecule, where peripheral atoms may lack distinctive features.

To quantitatively validate our qualitative interpretability claims, we performed a structural analysis of the attention mechanism (Supplementary Sections S7–S8). Specifically, Section S7 provides a Multi-Head Attention analysis, while Section S8 offers fine-grained visualizations of reactant–product alignments. To ensure these findings generalize across the entire dataset, we conducted a systematic evaluation of prediction consistency (Section S6) and analyzed the statistical correlation between structural mapping and reaction center localization (Section S10).

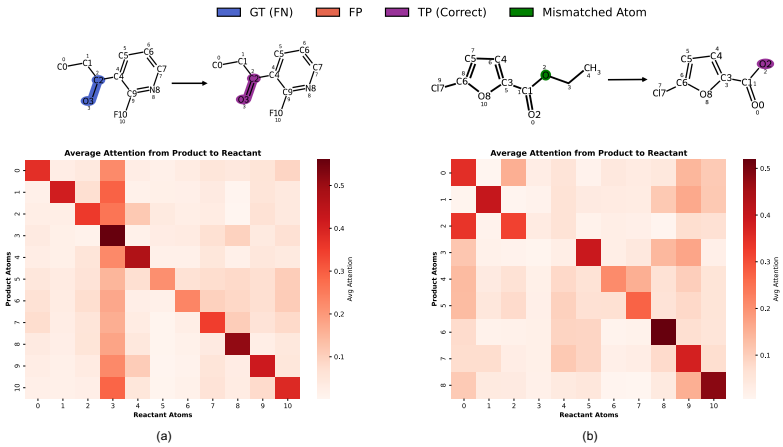

Fig. 3. Visualization of attention-based alignment for interpretability and error analysis. Each panel shows one reaction example: the top row depicts the reaction with atom-level annotations, and the bottom row shows the corresponding attention heatmap from product atoms (y-axis) to reactant atoms (x-axis). (a) Correct prediction where both atom mapping and reaction center identification are accurate; attention weights concentrate on the true reactive atoms. (b) Partially misaligned case where atom mapping is imperfect due to symmetry in the product. Despite this, the model correctly localizes the reaction center, with attention focused on transformation-relevant atoms (e.g., oxygen), illustrating robustness to alignment noise. *Color coding:* Blue = ground-truth reactive atoms (FN), Orange = predicted reactive atoms not in ground truth (FP), Purple = correctly predicted reactive atoms (TP), Green = mismatched mapped atoms. *Heatmap:* Dark red = high attention, white = low or no attention.

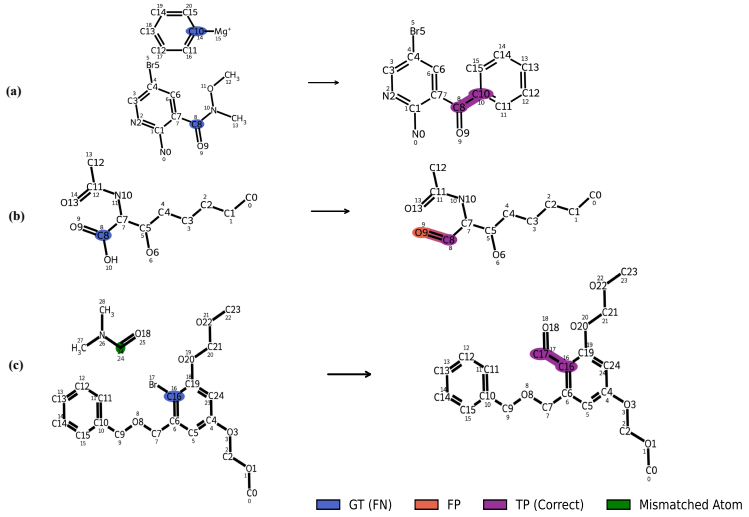

Fig. 4. Representative MARCC predictions on USPTO-50K test set reactions. Each row shows a different reaction: (a) MARCC accurately predicts both atom mapping and reactivity. (b) MARCC correctly maps atoms but falsely classifies atom 9 as reactive (orange). (c) MARCC identifies the correct reaction center but misaligns atom 17 in the mapping (green).

Conclusion and Future Work

We presented MARCC, a multi-task graph learning framework that unifies reaction center identification, reaction classification, and atom mapping within a single architecture. Its three main innovations—(i) mapping-guided cross-attention for reactant–product alignment, (ii) a dual-graph representation for bond-level reasoning, and (iii) imbalance-aware optimization—enable coherent reasoning across local and global aspects of molecular reactivity.

On the USPTO-50K benchmark, MARCC achieves state-of-the-art performance in reaction center localization and classification, while remaining competitive with dedicated atom-mapping systems. Ablation studies confirm that its multitask design enhances both accuracy and interpretability, making it well-suited for applications such as synthesis planning, reaction annotation, and mechanistic elucidation.

Beyond cheminformatics, MARCC offers a general design principle for reaction modeling: leveraging atom-to-atom alignment as a structural prior to unify local edits and global semantics. This paradigm holds promise for biochemical domains including enzymatic reactions, metabolic pathways, and drug metabolism. Future extensions will target retrosynthesis and multi-step synthesis planning, aiming to improve pathway inference and accelerate automated molecular discovery. In particular, we aim to enhance the model's robustness to data scarcity and noise.

While the current work validates MARCC on a high-fidelity benchmark, the architecture is designed with inherent flexibility to scale toward more complex chemical spaces. Our framework naturally supports multi-product reactions by treating the product

set as a disconnected graph, while the soft-attention mechanism in our cross-graph alignment remains robust to missing reagents by prioritizing existing structural correspondences. We intend to leverage these dataset-agnostic inductive biases to extend MARCC to larger, noisier corpora such as the Open Reaction Database (ORD) and USPTO-full. Future efforts will focus on utilizing the structural diversity of these datasets—combined with class-aware sampling and targeted data augmentation—to further enhance performance on rare transformations and complex multi-component systems. To further contextualize these advancements, future work will benchmark MARCC against architectures utilizing local environments and human-in-the-loop refinement, such as LocalMapper (Chen et al., 2024), exploring the interplay between fully automated global reasoning and expert-guided local mapping.

Availability of data and material

Code and data are available at `https://github.com/maryamastero/MARCC`; an archived version is available via Zenodo (`https://doi.org/10.5281/zenodo.18500230`).

Authors' contributions

M.A. contributed to the conception of the project and was responsible for designing the methodology, developing the model, conducting experimentation, analyzing the data, and writing the manuscript. A.L. assisted with benchmarking baselines for reaction classification. E.C. provided ongoing guidance during the project and contributed to the refinement of the manuscript by advising on its structure, clarity, and presentation. J.R. provided supervision, strategic guidance, and critical feedback throughout all stages of the work. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Wihuri Foundation; the Jane and Aatos Erkko Foundation through the BIODESIGN project; and the Helsinki Institute for Information Technology (HIIT). Additional support was provided by the Research Council of Finland under grants 339421 and 345802, and through the Flagship Programme (FCAI). E.C. acknowledges support from the FAIR project, funded by the NextGenerationEU program.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Maryam Astero and Juho Rousu. Learning symmetry-aware atom mapping in chemical reactions through deep graph matching. *Journal of Cheminformatics*, 16(1):46, 2024.

Maryam Astero and Juho Rousu. Enhancing atom mapping with multitask learning and symmetry-aware deep graph matching. *Journal of Cheminformatics*, 17(87), 2025.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Shuan Chen, Sunggi An, Ramil Babazade, and Yousung Jung. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nature Communications*, 15(1): 2250, 2024.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Herbert Gelernter, J Royce Rose, and Chyouhwa Chen. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *Journal of chemical information and computer sciences*, 30(4): 492–504, 1990.

Markus Heinonen, Sampsa Lappalainen, Taneli Mielikäinen, and Juho Rousu. Computing atom mappings for biochemical reactions without subgraph isomorphism. *Journal of Computational Biology*, 18(1):43–58, 2011.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Zixun Lan, Zuo Zeng, Binjie Hong, Zhenfu Liu, and Fei Ma. Rcsearcher: Reaction center identification in retrosynthesis via deep q-learning. *Pattern Recognition*, 150:110318, 2024.

Greg Landrum et al. Rdkit: Open-source cheminformatics, 2006.

Anchen Li, Elena Casiraghi, and Juho Rousu. Chemical reaction enhanced graph learning for molecule representation. *Bioinformatics*, 40(10):btae558, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

Ramil Nugmanov, Natalia Dyubankova, Andrey Gedich, and Joerg Kurt Wegner. Bidirectional graphormer for reactivity understanding: neural network trained to reaction atom-to-atom mapping task. *Journal of Chemical Information and Modeling*, 62(14):3307–3315, 2022.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital discovery*, 1(2):91–97, 2022.

John W Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of computer-aided molecular design*, 16:521–533, 2002.

Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification

and similarity. *Journal of chemical information and modeling*, 55 (1):39–53, 2015.

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

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34:9405–9415, 2021.

Boris Weisfeiler and Andrei Leman. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series*, 2(9):12–16, 1968.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems*, 33:11248–11258, 2020.