Unsupervised Multiple Sequence Alignment of Reaction SMILES

Babak Mahjour
Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
bmahjour@mit.edu

Connor Coley
Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
ccoley@mit.edu

Abstract

Multiple sequence alignment (MSA) establishes consistent positional correspondences across biological sequences, enabling comparative analysis and model interpretability. Drawing inspiration from this paradigm, we introduce an unsupervised framework for multiple alignment of chemical reactions represented as atom-mapped SMILES. Each reaction is decomposed into a set of atom-centered SMARTS fragments, which serve as local sequence analogues describing atomic environments and transformations. These SMARTS-derived features are embedded into a global token space whose parameters are optimized through a fast expectation–maximization (EM) procedure. During training, a ranking-based objective learns both the global embedding and positional multipliers that weight atomic contexts, while periodic recanonicalization of the SMARTS templates reorders reactants and token structures according to the evolving embedding. This feedback loop yields a self-consistent embedding that captures reaction role semantics and structural correspondences across the dataset. The final learned embedding is then applied to recanonicalize full reaction SMILES, producing standardized, embedding-guided representations that function analogously to consensus sequences in bioinformatics. The resulting canonicalization enables unsupervised role discovery, reaction alignment, and mechanistically interpretable representations for downstream modeling and generation. A case study illustrates the framework's ability to recover consistent reactant-role orderings and reaction family structure across diverse transformation classes.

1 Introduction

Chemical reactions can be encoded as one-dimensional text strings using line notations such as SMILES[1]. While canonicalization procedures for individual molecular SMILES are well established and ensure a unique representation for any given structure, no comparable framework exists for reaction SMILES. In particular, aligning reactions such that reactant roles and token sequences follow a consistent, chemically meaningful order remains an unsolved problem. Equivalent reactions may differ purely by the permutation of reactants or by variations in the ordering of atomic tokens within mapped molecules, introducing inconsistency that hinders data integration, model training, and mechanistic interpretation.

In molecular biology, multiple sequence alignment (MSA) addresses a comparable problem: establishing consistent positional correspondences across related protein, DNA, or RNA sequences to enable comparative analysis and feature extraction.[2] The core idea of MSA is that local sequence similarity can be used to infer a global alignment that captures conserved structure and function. By analogy, reaction canonicalization can be viewed as an alignment problem in which chemical tokens

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SimBioChem 2025.

(atoms, bonds, and functional features) must be consistently ordered to reveal underlying structural and mechanistic roles across reactions.

Recently, we introduced RDCanon, a canonicalization algorithm for SMARTS[3]. RDCanon accepts an embedding of SMARTS primitives and sorts the order of tokens and roles within a reaction SMARTS according to their learned embedding values. This allows embeddings to serve as alignment functions, aligning tokens across reaction sets in a data-driven manner. Traditional reaction SMILES standardization involves canonicalization of reactants, agents, and products—often followed by alphanumerical ordering of components—while more sophisticated methods employ graph neural networks to learn embeddings from atomic environments.[4] However, with template extraction algorithms,[5] chemically rich reaction SMARTS templates can be generated at controlled radii around the reaction core, identified by atoms that change state across the transformation, offering a structured substrate for embedding-based alignment.

Here, we report an unsupervised multiple sequence alignment framework for reaction SMILES. Each reaction is decomposed into token-level feature vectors describing local atomic environments derived from extracted SMARTS templates. A global embedding is learned via a fast expectation–maximization procedure that optimizes both token weights and positional multipliers to reproduce consistent reactant ordering across the dataset. Periodic recanonicalization of the SMARTS under the evolving embedding plays a role analogous to iterative realignment in MSA,[6,7] progressively refining token and reactant order to minimize global inconsistency. The learned embedding is subsequently applied to the full reaction SMILES, yielding an embedding-guided canonicalization that unifies reaction representation across datasets and enables unsupervised role discovery, mechanistic alignment, and improved interpretability for downstream modeling.

2 Methods

2.1 Data Representation and Template Extraction

Each reaction is represented as an atom-mapped SMILES string of the form `reactants > agents > products`. Reactants are parsed into molecular graphs, and the set of atoms that undergo a change in formal charge, hybridization, or bonding pattern defines the **reaction core**. Surrounding atomic environments are expanded to a specified radius ($r = 6$ in this study) and extracted using RDKit. This yields environment-rich SMARTS atom-typed templates that describe the local topology and token composition of each reactant and product. For benchmarking purposes, all reactions are canonicalized with the initial embedding to calculate the initial alignment score. Instances of reactions that do not match the consensus number of components for that reaction class are filtered out (for instance, intramolecular reactions where the number of reactants is fewer than its intermolecular counterpart).

For each extracted template, atom-mapped environments are encoded as vectors of token occurrence counts describing chemical primitives such as element type, ring membership, hydrogen count, charge, and hybridization (e.g., `#6`, `R1`, `H1`, `+1`). These per-atom vectors are assembled into a structured **feature tensor** $\mathbf{X} \in \mathbb{R}^{N\times K\times P\times T}$, where N is the number of reactions, K the number of reactants per reaction, P the number of positional slots (equal to the number of learned positional multipliers), and T the vocabulary size. Each entry $\mathbf{X}_{n,k,p,t}$ represents the weighted count of token t at positional index p of the k-th reactant in reaction n. Before use in training, all features are z-score normalized along the dataset dimension to ensure comparable scaling across positions and token types.

2.2 Token Embedding and Feature Construction

Each token t in the vocabulary is assigned a scalar embedding weight w_t, forming an **embedding vector** $\mathbf{E} = w_t$. For a given reaction, per-atom feature vectors are constructed from local token counts collected at each atom-map index, weighted by their corresponding embedding values. These vectors describe the local atomic environments in terms of chemical primitives such as element type, charge, ring membership, and hybridization.

A set of **positional multipliers** $\mathbf{m} = [m_1, m_2, \ldots, m_P]$ modulates the contribution of each atomic position within a reactant, reflecting decreasing influence with distance from the reaction center. Unlike static heuristic weights, these multipliers are *learned jointly* with the embedding parameters, adapting dynamically to capture positional patterns that improve alignment consistency across reactions.

For each reaction r, a **feature tensor** $\mathbf{X} \in \mathbb{R}^{N\times K\times P\times T}$ is constructed, where N is the number of reactions, K the number of reactants, P the number of positional slots, and T the vocabulary size. Each slice $\mathbf{X}_{n,k,:,:}$ contains the weighted token vectors for the k-th reactant in reaction n. To capture transformation information, the token basis is augmented with "delta" features d_t representing differences between product and reactant environments. Before training, all features are z-score normalized to ensure consistent scale across tokens and positions.

2.3 Role Assignment as Global Ranking Alignment

Each reaction is represented by a feature tensor $\mathbf{X} \in \mathbb{R}^{N\times K\times P\times T}$, where N is the number of reactions, K the number of reactants per reaction, P the number of positional slots, and T the vocabulary size. A global embedding vector $\mathbf{W} \in \mathbb{R}^T$ assigns a weight to each token type, while a set of positional multipliers $\mathbf{m} = [m_1, m_2, \ldots, m_P]$ modulates the contribution of each positional index to the overall reactant score.

For each reaction n and reactant k, a scalar score is computed as

$$s_{n,k} = \sum_{p=1}^{P} m_p \, \langle \mathbf{X}_{n,k,p,:}, \mathbf{W} \rangle, \tag{1}$$

which quantifies the alignment between that reactant's local token composition and the global embedding. Within each reaction, reactants are ranked by $s_{n,k}$, producing an ordered tuple of roles that is consistent across the dataset. This ranking formulation inherently enforces a "one-of-each" constraint, as each reaction yields a unique ordered configuration in which every reactant occupies a distinct role position.

2.4 Expectation–Maximization with Embedding-Guided Recanonicalization

Training follows an unsupervised expectation–maximization (EM) process in which both the embedding weights $\mathbf{W}$ and positional multipliers $\mathbf{m}$ are iteratively optimized to maximize alignment consistency across reactions. Alignment is evaluated by comparing the predicted reactant ordering against the canonical role indices within each reaction SMILES. For a dataset of N reactions, each containing K reactants, a reaction is considered *aligned* if the predicted rank order $\hat{\pi}_n = \mathrm{argsort}(s_{n,1}, \ldots, s_{n,K})$ matches the canonical order $\pi_n = [0, 1, \ldots, K-1]$. Cluster purity is then defined as the fraction of reactions whose predicted ordering is correct:

$$\text{Purity} = 1 - \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[\hat{\pi}_n \neq \pi_n], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. This metric measures the proportion of reactions for which all reactants are assigned to their correct relative roles, providing a global measure of alignment consistency across the dataset.

1. **E-step (Scoring and Role Ordering):** Given the current embedding $\mathbf{W}$ and multipliers $\mathbf{m}$, compute scores $s_{n,k}$ for all reactants and reorder reactants within each reaction by ascending score. The proportion of reactions whose ordering remains unchanged from the previous iteration defines the *alignment purity*, serving as a global consistency metric.
2. **M-step (Ranking-Based Optimization):** The parameters $\mathbf{W}$ and $\mathbf{m}$ are updated to improve ordering consistency by minimizing a ranking loss over all pairwise reactant score differences:

$$\mathcal{L}_{\text{rank}} = \sum_{n=1}^{N} \sum_{i<j} \sigma\left(\frac{s_{n,i} - s_{n,j} - \delta}{\tau} \right), \tag{3}$$

   where σ is the sigmoid function, δ is a ranking margin, and τ is a temperature parameter controlling smoothness. Gradients are computed for both $\mathbf{W}$ and $\mathbf{m}$, with gradient clipping applied to maintain numerical stability. The multipliers $\mathbf{m}$ are constrained to remain positive, ensuring consistent positional weighting.

3. **Recanonicalization Step:** After a fixed number of optimization steps, the learned embedding $\mathbf{W}$ is used to re-canonicalize all reaction SMARTS templates. This reordering of reactants and atom tokens under the updated embedding acts analogously to iterative realignment in multiple sequence alignment, progressively refining the global ordering of reactions and improving alignment purity across the dataset.

Typical training parameters include a learning rate of 0.5 for $\mathbf{W}$ and 5×10^{-3} for $\mathbf{m}$, a ranking margin $\delta = 0.25$, temperature $\tau = 2.5$, and gradient clipping threshold of 5.0. At convergence, the algorithm yields a global embedding and set of positional multipliers that define a self-consistent canonical ordering across all reactions, providing a reproducible alignment of reactant roles and token sequences suitable for downstream modeling.

2.5 Case Study

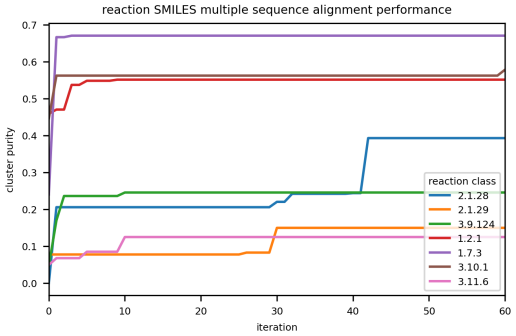

Figure 1: Performance of unsupervised reaction SMILES alignment across seven representative organic reaction classes from the Pistachio dataset. The cluster purity (fraction of reactions whose reactant ordering matches the majority role assignment for that index) is plotted over epochs of expectation–maximization (EM) iterations. Each curve corresponds to a distinct named reaction, illustrating how alignment quality evolves as the algorithm refines both role-specific embeddings and reactant token ordering.

This embedding serves as a role-agnostic alignment function that generalizes across reaction classes, enabling unsupervised canonicalization of new reactions, reaction clustering by role, and transfer learning for mechanistic or reactivity prediction tasks. We assessed the performance of the unsupervised alignment framework on seven representative reaction classes drawn from the Pistachio dataset: the Ugi reaction (2.1.28), Passerini reaction (2.1.29), aldehyde–alkyne–amine coupling (3.9.124), aldehyde reductive amination (1.2.1), ethyl esterification (1.7.3), Friedel–Crafts acylation (3.10.1), and the Mannich reaction (3.11.6). Each dataset consisted of atom-mapped reaction SMILES that were independently trained to convergence under the same expectation–maximization protocol.

Table 1: Performance of unsupervised reaction SMILES alignment across representative reaction classes from the Pistachio dataset.

Reaction	**Code**	**Batch Size**	**Initial Purity**	**Final Purity**
Ugi	2.1.28	550	0.000	0.393
Passerini	2.1.29	194	0.077	0.150
Aldehyde–alkyne–amine	3.9.124	106	0.047	0.250
Aldehyde reductive amination	1.2.1	987	0.459	0.551
Ethyl esterification	1.7.3	983	0.240	0.670
Friedel–Crafts acylation	3.10.1	998	0.446	0.578
Mannich	3.11.6	979	0.049	0.125

Across representative reaction classes, alignment (as assessed by cluster purity) improved consistently following iterative optimization and recanonicalization. Most two-component transformations, such as esterifications and reductive aminations, displayed monotonic improvement in purity, typically

stabilizing within two to three recanonicalization cycles. In contrast, multicomponent reactions such as the Ugi and Mannich transformations exhibited lower initial purity and slower convergence, reflecting greater combinatorial diversity and role ambiguity among reactive participants. Despite these differences, all reaction classes demonstrated measurable and consistent gains in alignment, with final purities exceeding their initial baselines. These results indicate that the embedding-guided canonicalization procedure effectively discovers stable, role-consistent orderings of reactants and atom tokens without the use of supervision or rule-based constraints, generalizing across both bimolecular and multicomponent reaction types.

3 Conclusion

We have introduced an unsupervised multiple sequence alignment framework for reaction SMILES that learns to canonicalize and align chemical transformations without supervision, templates, or predefined atom ordering rules. By representing reactions as token sequences and iteratively optimizing role-specific embeddings through expectation–maximization, the method establishes consistent positional correspondences across diverse reaction classes. The integration of recanonicalization as a realignment step enables the model to progressively refine both its internal embedding space and the ordering of reactant tokens. Applied across several canonical reaction types in the Pistachio dataset, the framework demonstrated rapid convergence toward stable role assignments, highlighting that unsupervised alignment of reaction SMILES is both feasible and robust. More broadly, this approach offers a pathway toward learning chemically meaningful token embeddings from structure–reactivity data, bridging discrete reaction representations with continuous latent spaces that may serve as a foundation for downstream tasks such as reaction clustering, mechanistic classification, and generative modeling of synthetic pathways.

References

[1] D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**,.

[2] J.-F. Taly, C. Magis, G. Bussotti, J.-M. Chang, P. Di Tommaso, I. Erb, J. Espinosa-Carrasco, C. Kemena and C. Notredame, *Nature Protocols*, 2011, **6**, 1669–1682.

[3] B. A. Mahjour and C. W. Coley, *Journal of Chemical Information and Modeling*, 2024, **64**, 2948–2954.

[4] K. Zeng, B. Yang, X. Zhao, Y. Zhang, F. Nie, X. Yang, Y. Jin and Y. Xu, *Journal of Cheminformatics*, 2024, **16**,.

[5] C. Coley, W. H. Green and K. F. Jensen, *Journal of Chemical Information and Modeling*, 2004, **59**,.

[6] F. Sievers and D. G. Higgins, *Protein science*, 2018, **27**, 135–145.

[7] R. C. Edgar, *BMC Bioinformatics*, 2004, **5**,.