AI for Interpretable Chemistry: Predicting Radical Mechanistic Pathways via Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

1 Appendix

In this appendix, we provide a comprehensive description of the experimental details and environments in which the experiments were conducted. Additionally, we present the detailed information and data pertaining to the pathway search. Furthermore, we offer an explanation of the various interfaces of the RMechRP software, which serves as the pioneering online radical reaction predictor. Each section in this appendix corresponds to the section with the same title in the main article. Finally, all the experiments are conducted using a single NVidia Titan X GPU.

1.1 Two Step Prediction

This method consists of two distinct steps, within each, we trained several neural networks. Here we explained the parameters used during the training of these networks.

1.1.1 Reactive Site Identification

For the Atom Fingerprint model, we constructed a fingerprint of length 800 for each atom. This fingerprint includes 700 graph topological features explained in [1] and 85 atomic features including a one-hot vector for atom type, and chemical features of the atoms such as valance and electronegativity. The graph topological features are extracted using a neighborhood if size three. The extracted fingerprints are fed into a fully connected model with an output layer for binary classification. For the GNN model, we used the atomic feature for the initial representations of atoms. The model consists of four GNN layers with an output layer for binary classification.

Combining both training sets presented in RMechDB [2], we extracted over 51000 atoms to train each of the models above. Both models are evaluated using a combination of two test sets in RMechDB and the topN accuracy of models are reported in Table 2 of the main article. Table 1 represents the parameters used for training the models.

Table 1: The parameters used for training the models for reactive site identification.

Model	Batch Size	Num Layers	Layers Dim	Act	Reg	Num Att Heads
Atom Fingerprint	32	3	512-256-1	GELU	L_2(5e-5)	-
GNN	32	4	64-64-64-1	ReLU	Dropout (0.3)	2

1.1.2 Plausibility Ranking

For the plausibility ranking experiments, we used the following four methods for representing a chemical reactions:

Feature Extraction: We use the same features explained in [1] which results in extracting a vector of length 3200 for each reaction.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

reaction*fp*: We use the RDKit [3] implementation of reaction*fp* [4]. For all three fingerprint types (Atom Pair, Morgan2, and Toplogical Torsions), we use a fingerprint of size 2048, with a bit ratio of 0.2. We considered non agent molecules with a weight of 0.4 and agent molecules with a weight of 1.0.

DRFP: We use the DRFP fingerprint [5] with a size of 2048 with a min and max radius of zero and four, while including the hydrogen atoms and rings.

Feature Extraction: We use the the default tokenizer and pretrained model for the *rxnfp* [6] which results in fingerprints of length 256.

For training, we use a combination of both training sets in RMechDB. For each sample of the training data (productive reaction), we generate (at most) 40 negative samples (unproductive reactions) by randomly sampling molecular orbitals other that the reactive MOs (m_1^*, m_2^*). This results in a data set of over 185000 pair of productive and unproductive reactions. To train the plausibility rankers for each method, we use the parameters explained in Table 2.

Table 2: The parameters used for training the models for the plausibility ranking.

Model	Batch Size	Num Layers	Layers Dim	Act	Reg
Feature Extraction	32	3	512-256-1	GELU	Dropout (0.5)
reaction*fp*	32	3	400-200-1	GELU	Dropout (0.5)
DRFP	32	3	400-200-1	GELU	Dropout (0.5)
rxnfp	64	2	128-1	GELU	Dropout (0.5)

1.2 Contrastive Learning

1.2.1 Atom Pairs and Atom Descriptor

For the contrastive learning method using the atom pairs and atom descriptor, we use the same atomic feature and graph topological features above to represent one single atom. Specifically, for the graph topological features, we use the neighborhood of size one. These features plus the atomic features results in a vector of length 140 for atom representation. Using these vectors, we train a contrastive model depicted in Figure 2 (left) of the main article. The objective function to train this contrastive model is as follows:

$$\mathcal{L} = 1 - \sigma([f(a_1^*) \times g(a_2^*)] - [f(a_1') \times g(a_2')]) \tag{1}$$

Where a_1^* and a_2^* are the atoms of the reactive MOs m_1^* and m_2^*, while a_i' are randomly chosen atoms. Both f and g functions are characterized by a fully connected neural network. The first reactive atoms in both productive an unproductive reactions, are fed through the same network f, and similarly the second reactive atoms are fed through the same network g. The outputs of both f and g are single real-valued numbers, which, when multiplied together, yield a score for the respective reaction. These scores are then utilized to construct the objective function, aiming to maximize the score of the productive reaction compared to the unproductive reactions using the same reactant set. Figure 1 represents a schematic depiction for this contrastive model.

We use a combination of both training sets in RMechDB to train f and g. For each productive reaction, we form unproductive reactions by considering at most 15 samples of (a_1', a_2^*), (a_1^*, a_2'), and (a_1', a_2'). This negative sampling results in a dataset of over 200000 pair of productive and unproductive atom pairs. We use this training dataset to minimize the objective function 1.

Both f and g has similar architectures that consists of three fully connected layers with GELU activation function and a dropout with a rate of 0.5 applied to all layers. The dimension of the layers are 128, 64, 1.

1.2.2 Rxn-Hypergraph

We use the Rxn-Hypergraph to replace form atom descriptors that are extracted automatically for minimizing the objective function 1. After processing the Rxn-hypergraph for N layers, the generated atom descriptors are used in the same setting above for the same minimization objective. Here in Table 3 we describe the parameters we use for training the Rxn-Hypergraph.

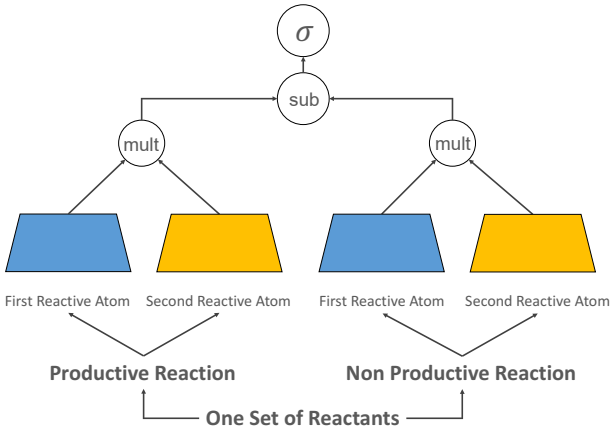

Figure 1: The architecture of the the constrastive learning approach.

1.3 Text Representation and Sequence to Sequence Models

In order to develop a text-based radical reaction predictor, we utilize the pretrained molecular transformer which was trained using the USPTO_MIT_mixed dataset. We also used the tokenizer developed by molecular transformer. This tokenizer yields 523 distinct tokens for the USPTO_MIT_mixed dataset. There are nine tokens from the RMechDB dataset that do not match the 573 tokens of the USPTO. Therefore, we used the *unknown token* to represent these nine tokens.

For the fine-tuning the pretrained model, we used the combination of both RMechDB training sets. We fine-tune the model using a simple data augmentation described in Section 4.5 for 10 epochs. Finally, for the evaluation of the text-based models, we considered all the generated *unknown token* as correct tokens.

1.4 Pathway Search

In the Pathway Search section, we conducted an experiment involving the execution of the pathway search for 100 specific reactants. Each of these reactants was associated with a desired target molecule, which was expected to be found within the mechanistic pathway tree. Additionally, a set of distinct parameters was assigned to each reactant to guide the pathway search process.

To provide detailed information and facilitate reproducibility, we have included supplementary materials accompanying the paper. Among these materials, you will find a file named *pathways.csv*. This file contains the reactants, corresponding targets, the provided context (if any), and the anticipated depth at which the target molecule is expected to appear within the mechanistic pathway tree.

Furthermore, we have included another file titled *pathway_results.txt* in the supplementary materials. This file comprises the identified pathways leading to the specified target molecules. It presents the discovered pathways that were found during the experiment.

It is worth noting that the 100 pathways and their results will be published alongside the paper, subject to acceptance. These materials serve to provide comprehensive insights into the pathway search process and its outcomes, enabling readers to reproduce and further explore the obtained results.

1.5 RMechRP Software

In addition to the methods and results presented in the main article, we have developed an online web server that enables users to utilize the trained models for predicting the outcomes of mechanistic

Table 3: The parameters used for training the Rxn-Hypergraph for the contrastive model.

Batch Size	Num Layers	Layers Dim	Act	Reg	Num Att Heads	Learning Rate
32	5	all 64	GELU	L_2(5e-5)	2	0.001

radical reactions with the highest levels of interpretability of the outcome. RMechRP (**R**adical **Mech**anistic **R**eaction **P**redictor) accessible via the anonymized link `http://128.195.8.137:8081/rrp/`. RMechRP offers two interfaces: Single-step prediction and Pathway search.

Single-Step Prediction predicts the outcome if a mechanistic reaction with a single transition state. Users have the option to either input the reactants in written form or draw them using a drawing tool provided on the web server. Additionally, users can specify the reaction conditions, with the current option being standard temperature and pressure. The number of reactive molecular orbitals (MOs) to be considered can also be specified by the user.

To ensure flexibility, users can choose to filter out reactions that violate specific chemical rules, such as Bredt's rule [7]. Once the input and conditions are set, the user can click the predict button. The system will then run the two-step prediction model, as described above, to generate and rank the potential products. These predicted products will be displayed, accompanied by additional information such as arrow codes, reactive MOs, and the mass of the products. The single step predictor is accessible via the anonymized link `http://128.195.8.137:8081/rrp/singlestep`. Figure 2 shows the single step interface and the displayed predictions for a simple reaction.

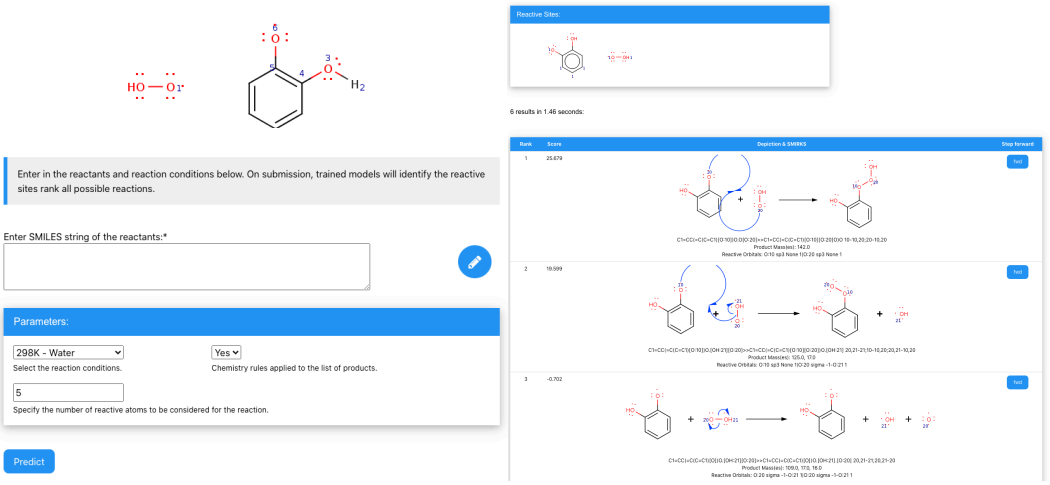

Figure 2: The single step interface with the predictions of a simple reaction. Left: the input panel. Right: the table displaying the ranked predictions.

Pathway Search forms the tree of the mechanistic pathways up to a given depth and breadth. Users have the option to either input the reactants in written form or draw them using a drawing tool provided on the web server. Users must also input a set of targets (either mass or chemical structure) to look for within the expanded tree of the mechanistic pathways. users have the ability to provide a context for the reactions. The context consists of a set of molecules along with their corresponding frequencies of appearance within the mechanistic pathway tree. When a molecule from the context is consumed in a reaction, the system can automatically reintroduce that molecule back into the pathway tree. The frequency of appearance indicates how many times a molecule can be added to the mechanistic pathway tree.

In addition to the context, there are several additional parameters that can be specified by the user. These parameters include:

Depth of Pathway Search: Users can define the depth of the pathway search, which determines how many reaction steps will be explored in the mechanistic pathway tree.

Breadth (Branching Factor) of Pathway Search: This parameter controls the branching factor of the pathway search, influencing the number of alternative reaction pathways that will be considered.

Application of Chemistry Rules: Users have the option to apply certain chemistry rules during the pathway search. These rules can be used to filter out reactions that violate specific chemical principles or constraints.

Score Threshold: Users can set a threshold value to consider only reactions with scores higher than the specified threshold. This helps narrow down the focus to more favorable or promising reactions.

These additional parameters allow users to customize their pathway search and refine the results based on their specific requirements and preferences. By leveraging these features, users can gain deeper insights into the mechanistic pathways and explore a wider range of possible reaction outcomes. The pathway search interface is accessible via the anonymized link `http://128.195.8.137:8081/rrp/pathway`. Figure 3 shows the pathway interface and the required parameters.

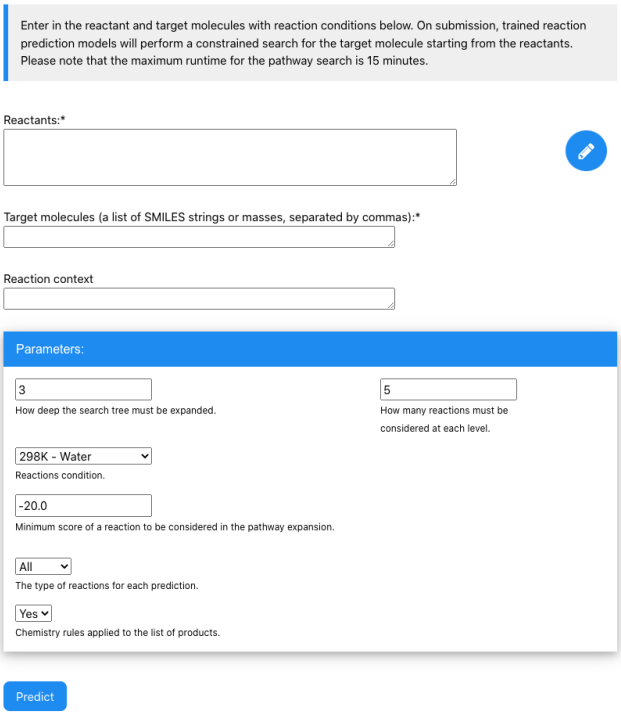

Figure 3: The pathway search interface.

References

[1] David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 2018.

[2] Mohammadamin Tavakoli, Yin Ting T Chiu, Pierre Baldi, Ann Marie Carlton, and David Van Vranken. Rmechdb: A public database of elementary radical reaction steps. *Journal of Chemical Information and Modeling*.

[3] RDKit: Open-source cheminformatics. `http://www.rdkit.org`. [Online; accessed 11-April-2013].

[4] 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.

[5] 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.

[6] Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence*, 3(2):144–152, 2021.

[7] J Bredt, Jos Houben, and Paul Levy. Ueber isomere dehydrocamphersäuren, lauronolsäuren und bihydrolauro-lactone. *Berichte der deutschen chemischen Gesellschaft*, 35(2):1286–1292, 1902.