Supplementary material of

A General-Purpose Framework for Chemical Reaction Representation with Atomic Correspondence and Flexible Condition Adaptation

Kaipeng Zeng[1, 2†], Xianbin Liu[1†], Yu Zhang[1], Xiaokang Yang[1], Yaohui Jin[1*], Yanyan Xu[1*]

[1*]MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, Shanghai, China.
[2]Shanghai Innovation Institute, Shanghai, 200231, Shanghai, China.

*Corresponding author(s). E-mail(s): jinyh@sjtu.edu.cn; yanyanxu@sjtu.edu.cn;
†These authors contributed equally to this work.

1 Data Preparation

1.1 Descriptions and Statistical Information of Datasets

The detailed descriptions of all datasets we used are presented as follows.

- **USPTO-Condition dataset** [1]. USPTO-Condition dataset is a dataset processed by Wang et al. [1]. This dataset comprises 680,741 reactions, with each reaction condition consistently consisting of up to one catalyst, two reagents, and two solvents.
- **USPTO-500MT dataset** [2]. USPTO-500MT dataset is a dataset collected by Lu et al. [2]. It can be used to train and test five different types of reaction tasks. In this paper, we only use it for the reaction condition combination generation tasks.

- **Buchwald-Hartwig cross-coupling dataset** [3]. This dataset provides 10 random splits and four ligand-based out-of-sample splits for evaluation. The test sets under the four out-of-sample data splits contain reaction additives that are not included in the training sets. We use the raw data and the data split provided by Probst el al. [4] and add the atom mapping for reactions according to the reaction template.
- **chiral C-H functionalization dataset** [5]. The dataset contains 6,115 reactions that functionalize C-H bonds of heterocycles (via free radical reactions), with regioselectivity due to differences in reaction sites. The dataset were randomly divided into a training set and a test set with a 7: 3 ratio for 10 times.
- **chiral phosphoric acid-catalyzed thiol addition dataset** [6, 7]. The dataset contains the reactions regarding chiral phosphoric acid–catalyzed thiol addition to N-acylimines. It includes 43 catalysts and combinations of 5 $\times$ 5 reactants, which form 1075 reactions. The dataset were randomly divided into a training set and a test set with a 7: 3 ratio for 10 times.

We summarize the statistical information of all the datasets used in this work in Supplementary Table 1.

Supplementary Table 1: The statistical information of the datasets

Dataset	Split-type	Train	Val	Test
USPTO-Condition	random split	544,591	68,075	68,075
USPTO-500MT	random split	116,360	12,937	14,238
Buchwald-Hartwig cross-coupling	random split	2,491	277	1187
	Test1	2,751	306	898
	Test2	2,749	306	900
	Test3	2,752	306	897
	Test4	2,749	306	900
chiral phosphoric acid-catalyzed thiol addition	random split	677	75	323
radical C-H functionalization	random split	3,851	428	1,835

1.2 Adding Atom Mapping

The Buchwald-Hartwig cross-coupling dataset features a consistent reaction template across all its reaction data, enabling the derivation of atom mapping through rule-based approaches. For all the other datasets except the Buchwald-Hartwig cross-coupling dataset, we have employed RXNMapper [8] to obtain atom mapping. It is noteworthy that we have re-annotated the labels for the USPTO-500MT dataset, treating both the originally provided reactants and reagents as input reactants during the atom-mapping annotation for this dataset. For further details on the processing of the USPTO-500MT dataset, please refer to Supplementary Sec. 1.3.

1.3 Generation of Reagents in USPTO-500MT dataset

We restructured the USPTO-500MT dataset. Following the annotations provided by the dataset, we divided the reactants involved in the reactions into a series of charge-balanced ions or molecular clusters. Here, we treat a cluster as a single molecule rather than segmenting molecules based on the delimiter '.' as in prior studies. Utilizing RXNMapper [8], we appended atom mapping to the reactions, where molecules in the original reactants that have atoms present in the products are considered reactants, while the remaining portions are classified as reagents. We categorized the reagents according to the following hierarchy:

- If a molecule is a free metal, contains a cyclic structure along with a metal or phosphorus atom, or is a metal halide, it is designated as Type I.
- If a molecule is an organic compound, it is designated as Type II.
- The remaining molecules are categorized as Type III.

Our data labels are constructed in the order of Type I, Type II, and Type III. When reagents contain multiple molecules of the same type, these molecules are sorted in ascending order of their SMILES string lengths within the same category. The tokenization of the SMILES representations of the reagents is aligned with the work [9].

2 The Full Experiment Results

2.1 Reaction Condition Recommendation

Experimental results for the USPTO-Condition and USPTO-500MT datasets are tabulated in Supplementary Tables 2, Supplementary Table 3, and Supplementary Table 4. For USPTO-Condition, we additionally report the prediction accuracies for catalysts, solvents, and reagents. Each entry may contain up to two solvents and two reagents; crucially, the prediction order is not considered in calculating the solvents'and reagent's prediction accuracy.

2.2 Reaction Yield Prediction

We evaluated our model across ten random data splits and four out-of-distribution (OOD) data splits on the Buchwald-Hartwig cross-coupling dataset. We use MAE, RMSE, and R^2 as evaluation metrics. The results are summarized in Supplementary Table 5, Supplementary Table 6, and Supplementary Table 7. For the random data splits, we report the mean and standard deviation across the ten splits. For the four OOD data splits, we conduct experiments using five random seeds (2023 to 2027) and present the mean and standard deviation of the results.

2.3 Reaction Selectivity Prediction

We evaluated our model's performance on both the thiol addition and C–H functionalization datasets. Each dataset was partitioned into ten distinct random splits. Using MAE, RMSE, and R^2 as evaluation metrics, we report the mean and standard deviation of these metrics across all ten splits for each dataset. These results are

Supplementary Table 2: **Overall accuracy of reaction condition prediction on USPTO-Condition dataset.** The best performance is in **bold**. The second-best performance is underlined. * indicates that the model has been pretrained. IFL stands for information fusion layer.

Model	Top-k accuracy (%)			
	1	3	5	10
T5chem*	<u>28.31</u>	<u>42.32</u>	<u>48.15</u>	<u>55.32</u>
Parrot*	27.42	41.86	46.98	50.95
GCNN	12.81	21.95	26.40	32.17
FPRCR	16.90	26.38	31.16	36.96
Reagent Transformer	22.74	34.09	39.36	46.01
T5Chem	16.42	27.75	33.53	40.79
Parrot[1]	23.97	36.72	41.59	45.96
Ours	**33.30**	**47.11**	**52.43**	**59.19**

[1] The checkpoint of non-pretrained version is not provided, the performance is copied from the original paper [1].

summarized in Supplementary Table 8 and Supplementary Table 9. We also report the performance of different models on three out-of-distribution test sets of the thiol addition dataset. Five random seeds (2025 to 2029) were used for training, and the mean and standard deviation of all metrics are reported in Supplementary Table 10, Supplementary Table 11, and Supplementary Table 12.

2.4 Qualitative Analysis of Attention Distributions

2.4.1 Scope, Limitations, and Caveats of this Qualitative Illustration

We provide the following explicit caveats regarding the interpretation of the attention weights presented in this section. **It is crucial to emphasize that these visualizations are intended for illustrative purposes only and do not constitute causal explanations of the model's decision-making process.** The chemical narratives provided below are post-hoc interpretations that, in these specific instances, appear to align with chemical intuition. **These observations are intended to offer a qualitative glimpse into how the model organizes input information and should not be interpreted as systematic evidence of intrinsic chemical reasoning capabilities, nor as a substitute for rigorous quantitative interpretability metrics**. The following cases were selected to illustrate the range of attention patterns observed and do not represent a quantitative analysis of model behavior.

Supplementary Table 3: **Accuracy of each component on USPTO-Condition dataset for reaction condition prediction.** The detailed prediction accuracy on each type of components is displayed. The best performance is in **bold** and the second-best is underlined. *indicates the the model is pretrained

Method	Top-k accuracy (%)											
	catalyst				solvents				reagents			
	1	3	5	10	1	3	5	10	1	3	5	10
Parrot*	<u>92.12</u>	<u>94.91</u>	<u>95.97</u>	<u>97.28</u>	<u>44.20</u>	<u>59.23</u>	<u>63.71</u>	<u>66.16</u>	<u>46.47</u>	<u>62.04</u>	<u>68.19</u>	<u>73.93</u>
T5Chem*	91.64	94.76	95.88	97.07	44.17	59.02	65.14	72.95	46.23	62.37	68.55	75.66
Reagent Transformer	89.80	93.30	94.52	95.88	37.97	51.29	57.47	65.44	39.11	55.20	61.83	69.80
FPRCR	91.22	92.82	93.57	94.74	38.51	49.72	54.64	61.56	37.54	47.95	52.69	58.48
GCNN	90.59	91.80	92.40	93.39	32.39	46.02	52.51	61.06	35.84	46.37	50.61	55.99
T5Chem	88.60	92.38	93.68	95.07	30.30	46.66	54.63	64.57	32.71	47.98	54.86	63.25
Parrot[1]	91.96	91.96	91.96	91.96	-	-	-	-	-	-	-	-
Ours	**92.65**	**95.56**	**96.46**	**97.47**	**48.60**	**63.09**	**68.81**	**75.76**	**49.68**	**66.34**	**72.28**	**79.02**

[1] The checkpoint of non-pretrained version is not provided, the prediction accuracy on solvents and reagents can not be calculated. The prediction accuracy on catalyst is copied from the original paper [1].

Supplementary Table 4: **Overall accuracy of reaction condition generation on USPTO-500MT dataset.** The best performance is in **bold**. The second-best performance is underlined. * indicates that the model has been pretrained.

Model	Top-k accuracy (%)			
	1	3	5	10
T5Chem*	<u>25.36</u>	<u>38.12</u>	<u>43.42</u>	<u>49.54</u>
Reagent Transformer	17.88	25.47	28.78	33.27
T5Chem	17.32	28.21	32.98	38.76
Ours	**26.62**	**39.46**	**44.69**	**50.94**

Supplementary Table 5: **The results on Buchwald-Hartwig cross-coupling dataset under ten random splits.** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

Model	**Type**	MAE ↓	RMSE ↓	R^2 ↑
MFF	ML	6.4411 ± 0.8157	9.6382 ± 1.0206	0.8739 ± 0.0292
DRFP	ML	4.0995 ± 0.1191	6.2424 ± 0.2636	0.9474 ± 0.0050
Chemprop	DL	4.6430 ± 0.1405	6.4306 ± 0.1938	0.9441 ± 0.0042
YieldBert	DL	10.9327 ± 0.6514	15.7460 ± 0.6475	0.6655 ± 0.0311
T5Chem	DL	11.2666 ± 0.6432	15.8840 ± 0.7096	0.6601 ± 0.0253
ReaMVP	DL	**3.3921 ± 0.1453**	**5.1034 ± 0.2717**	**0.9648 ± 0.0039**
Ours	DL	<u>3.3972 ± 0.1530</u>	<u>5.1582 ± 0.2162</u>	<u>0.9642 ± 0.0026</u>
YieldBert*	DL	3.5532 ± 0.1600	5.4480 ± 0.3240	0.9599 ± 0.0050
T5Chem*	DL	<u>3.5059 ± 0.1562</u>	<u>5.3181 ± 0.2482</u>	<u>0.9607 ± 0.0047</u>
ReaMVP*	DL	3.7213 ± 0.0413	5.6183 ± 0.2772	0.9574 ± 0.1637
Ours†	DL	**3.2763 ± 0.1313**	**5.0662 ± 0.1609**	**0.9654 ± 0.0021**

Supplementary Table 6: **The results on Buchwald-Hartwig cross-coupling dataset under out-of-distribution data split Test1 and Test2.** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

Model	**Type**	**Test1**			**Test2**		
		MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MFF	ML	10.9730 ± 1.1066	17.4750 ± 2.9942	0.5795 ± 0.1364	9.6391 ± 0.1465	14.0512 ± 0.2180	0.7309 ± 0.0083
DRFP	ML	**8.0137 ± 0.0702**	**11.3992 ± 0.0845**	**0.8252 ± 0.0026**	<u>9.0642 ± 0.0725</u>	<u>13.5332 ± 0.1913</u>	<u>0.7504 ± 0.0070</u>
Chemprop	DL	9.7982 ± 0.9586	<u>14.0620 ± 1.3639</u>	<u>0.7320 ± 0.0523</u>	10.0890 ± 0.7180	13.6659 ± 0.8532	0.7447 ± 0.0308
YieldBert	DL	13.2461 ± 0.2341	18.8416 ± 0.4287	0.5222 ± 0.0216	13.0776 ± 0.9276	18.9203 ± 1.4600	0.5101 ± 0.0751
T5Chem	DL	12.3271 ± 0.5404	17.1757 ± 0.7090	0.6263 ± 0.0302	12.7461 ± 0.6414	17.8660 ± 0.8463	0.5896 ± 0.0333
ReaMVP	DL	13.2885 ± 1.3030	18.6288 ± 2.0344	0.5275 ± 0.0963	9.9564 ± 0.7027	15.5496 ± 0.6603	0.6700 ± 0.0281
Ours	DL	<u>8.5583 ± 1.6501</u>	14.3919 ± 2.7852	0.7130 ± 0.1183	**8.0124 ± 1.9118**	**12.2923 ± 3.2324**	**0.7827 ± 0.1025**
YieldBert*	DL	7.7793 ± 0.4532	<u>11.5510 ± 0.6199</u>	0.8201 ± 0.0197	7.0805 ± 0.8213	10.4167 ± 1.3162	0.8503 ± 0.0371
T5Chem*	DL	<u>7.3660 ± 0.2220</u>	12.0129 ± 0.2372	<u>0.8058 ± 0.0077</u>	<u>6.1430 ± 0.2137</u>	<u>9.4270 ± 0.3318</u>	<u>0.8788 ± 0.0086</u>
ReaMVP*	DL	9.7030 ± 0.9769	14.3139 ± 1.1189	0.7227 ± 0.0449	7.9062 ± 0.2206	12.2573 ± 0.6379	0.7947 ± 0.0217
Ours†	DL	**6.6637 ± 0.3786**	**10.8275 ± 0.8035**	**0.8414 ± 0.0236**	**5.5398 ± 0.1932**	**8.3805 ± 0.2555**	**0.9042 ± 0.0059**

Supplementary Table 7: **The results on Buchwald-Hartwig cross-coupling dataset under out-of-distribution data split Test3 and Test4.** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

Model	Type	Test3			Test4		
		MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MFF	ML	**9.6022 ± 0.0996**	**14.8230 ± 0.2124**	**0.7216 ± 0.0080**	15.9437 ± 2.0114	20.9527 ± 1.7867	0.3688 ± 0.1086
DRFP	ML	10.1413 ± 0.0914	15.9365 ± 0.0954	0.6782 ± 0.0039	**12.7281 ± 0.0332**	**19.0366 ± 0.0644**	**0.4819 ± 0.0035**
Chemprop	DL	10.2583 ± 0.4631	15.2133 ± 0.5174	0.7065 ± 0.0197	14.5552 ± 0.7769	20.3402 ± 0.9106	0.4076 ± 0.0522
YieldBert	DL	15.4959 ± 0.5699	22.1421 ± 0.8237	0.3783 ± 0.0462	17.7884 ± 0.9201	23.2340 ± 0.9895	0.2273 ± 0.0656
T5Chem	DL	11.3490 ± 0.7840	16.7147 ± 0.4914	0.6566 ± 0.0185	20.6032 ± 0.7013	26.8503 ± 1.0382	0.2284 ± 0.0426
ReaMVP	DL	13.0350 ± 1.8263	19.7192 ± 1.9116	0.5028 ± 0.0962	18.2115 ± 2.6947	24.7563 ± 3.4819	0.1065 ± 0.2614
Ours	DL	10.8734 ± 0.4776	17.7394 ± 0.8088	0.6007 ± 0.0367	15.7305 ± 0.5098	23.5543 ± 0.7971	0.2062 ± 0.0535
YieldBert*	DL	9.6526 ± 0.1736	15.4455 ± 0.7289	0.6972 ± 0.0286	14.3879 ± 1.1374	19.6979 ± 1.5110	0.4427 ± 0.0834
T5Chem*	DL	9.1328 ± 0.4950	14.4337 ± 1.3158	0.7339 ± 0.0485	13.5883 ± 0.5530	19.7144 ± 0.6983	0.4437 ± 0.0393
ReaMVP*	DL	9.7782 ± 0.2469	14.5569 ± 0.2282	0.7315 ± 0.0084	**12.6095 ± 1.0512**	**17.8684 ± 0.7946**	**0.5420 ± 0.0531**
Ours†	DL	**8.9069 ± 0.4436**	**13.8243 ± 0.8078**	**0.7572 ± 0.0282**	14.0090 ± 0.2941	20.3306 ± 0.5318	0.4087 ± 0.0311

Supplementary Table 8: **The results on radical C-H functionalization dataset under random splits.** The best performance is in **bold**. The second-best performance is underlined. * indicates that the model has been pretrained.

Model	Type	MAE ↓	RMSE ↓	R^2 ↑
RXNFP*	DL	**0.3744 ± 0.0094**	**0.5378 ± 0.0204**	**0.9862 ± 0.0010**
T5Chem*	DL	0.5930 ± 0.0148	0.7912 ± 0.0219	0.9702 ± 0.0016
ReaMVP*	DL	0.5908 ± 0.0099	0.8237 ± 0.015	0.9677 ± 0.0012
MFF	ML	2.2923 ± 0.0400	2.9388 ± 0.0393	0.5891 ± 0.0095
DRFP	ML	0.9435 ± 0.0293	1.2943 ± 0.0387	0.9203 ± 0.0042
Chemprop	DL	0.3593 ± 0.0099	0.5396 ± 0.0299	0.9861 ± 0.0015
RXNFP	DL	1.0034 ± 0.0393	1.4389 ± 0.0638	0.9013 ± 0.0089
T5Chem	DL	2.5398 ± 0.0549	3.3156 ± 0.0925	0.4765 ± 0.0322
ReaMVP	DL	0.6390 ± 0.0216	0.8653 ± 0.0307	0.9643 ± 0.0026
Ours	DL	**0.3362 ± 0.0158**	**0.5375 ± 0.0379**	**0.9862 ± 0.0018**

Supplementary Table 9: **The results on chiral phosphoric acid-catalyzed thiol addition dataset under random splits.** The best performance is in **bold**. The second-best performance is underlined.

Model	Type	MAE ↓	RMSE ↓	R^2 ↑
MFF	ML	**0.1421 ± 0.0093**	0.2122 ± 0.0169	**0.9055 ± 0.0119**
DRFP	ML	0.1460 ± 0.0086	**0.2109 ± 0.0141**	0.9066 ± 0.0115
Chemprop	DL	0.1626 ± 0.0118	0.2283 ± 0.0174	0.8907 ± 0.0130
RXNFP	DL	0.2717 ± 0.0150	0.3935 ± 0.0223	0.6747 ± 0.0368
T5Chem	DL	0.3941 ± 0.0287	0.5177 ± 0.0459	0.4293 ± 0.1344
ReaMVP	DL	0.1920 ± 0.0889	0.2721 ± 0.1219	0.8127 ± 0.2325
Ours	DL	0.1580 ± 0.0079	0.2249 ± 0.0117	0.8937 ± 0.0115
RXNFP*	DL	0.1650 ± 0.0111	0.2313 ± 0.0197	0.8872 ± 0.0193
T5Chem*	DL	0.1812 ± 0.0128	0.2614 ± 0.0213	0.8568 ± 0.0171
ReaMVP*	DL	0.1647 ± 0.0073	0.2373 ± 0.0146	0.8817 ± 0.0142
Ours†	DL	**0.1531 ± 0.0094**	**0.2215 ± 0.0147**	**0.8970 ± 0.0117**

Supplementary Table 10: **The results on chiral phosphoric acid-catalyzed thiol addition dataset(Cat test set).** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

Model	**Type**	MAE ↓	RMSE ↓	R^2 ↑
DRFP	ML	0.2741 ± 0.0018	**0.4027 ± 0.0026**	**0.6701 ± 0.0043**
MFF	ML	**0.2665 ± 0.0060**	0.4095 ± 0.0228	0.6580 ± 0.0368
Chemprop	DL	0.3053 ± 0.0136	0.4491 ± 0.0147	0.5893 ± 0.0268
ReaMVP	DL	0.4307 ± 0.0104	0.6718 ± 0.0184	0.0813 ± 0.0512
T5Chem	DL	0.5026 ± 0.0256	0.6441 ± 0.0175	0.1557 ± 0.0464
RXNFP	DL	0.4276 ± 0.0137	0.6450 ± 0.0135	0.1534 ± 0.0356
Ours	DL	0.3316 ± 0.0095	0.4477 ± 0.0117	0.5921 ± 0.0217
RXNFP*	DL	0.3924 ± 0.0574	0.5372 ± 0.0562	0.4078 ± 0.1268
T5Chem*	DL	0.3274 ± 0.0178	0.4734 ± 0.0334	0.5424 ± 0.0650
ReaMVP*	DL	0.3948 ± 0.0186	0.5039 ± 0.0226	0.4826 ± 0.0470
ours†	DL	**0.2699 ± 0.0092**	**0.3803 ± 0.0132**	**0.7055 ± 0.0203**

Supplementary Table 11: **The results on chiral phosphoric acid-catalyzed thiol addition dataset(Sub test set).** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

Model	Type	MAE ↓	RMSE ↓	R^2 ↑
DRFP	ML	0.1873 ± 0.0066	0.2631 ± 0.0082	0.8551 ± 0.0091
MFF	ML	0.2741 ± 0.0061	0.3843 ± 0.0107	0.6910 ± 0.0169
Chemprop	DL	0.1487 ± 0.0098	0.2036 ± 0.0105	0.9132 ± 0.0091
ReaMVP	DL	0.2223 ± 0.0875	0.2923 ± 0.1043	0.8032 ± 0.1517
T5Chem	DL	0.4610 ± 0.0187	0.5885 ± 0.0211	0.2749 ± 0.0515
RXNFP	DL	0.3783 ± 0.0229	0.4610 ± 0.0228	0.5547 ± 0.0446
Ours	DL	**0.1469 ± 0.0045**	**0.1971 ± 0.0039**	**0.9187 ± 0.0032**
RXNFP*	DL	0.1849 ± 0.0407	0.2381 ± 0.0458	0.8779 ± 0.0503
T5Chem*	DL	0.2173 ± 0.0262	0.2935 ± 0.0198	0.8192 ± 0.0247
ReaMVP*	DL	0.3135 ± 0.0227	0.4002 ± 0.0313	0.6634 ± 0.0541
Ours†	DL	**0.1485 ± 0.0024**	**0.2009 ± 0.0033**	**0.9156 ± 0.0028**

2.4.2 Illustrative Case Studies: Observed Attention Patterns

To qualitatively observe how the proposed Reaction-Center-Aware (RC-aware) cross-attention mechanism influences the model's focus, we visually compare the attention weight distributions of the RC-aware heads, the global heads within our full model, and the attention heads of a non-RC-aware baseline variant. This baseline is identical to our full model but employs a standard Transformer decoder with multi-head cross-attention over all encoder nodes, leaving the Atom-Aligned Encoder and all other

Supplementary Table 12: **The results on chiral phosphoric acid-catalyzed thiol addition dataset(Sub-Cat test set).** The best performance of pretrained and not pretrained model is in **bold** and the second-best performance of pretrained and not pretrained models is underlined. * indicates that the model has been pretrained, and † indicates that our model uses a pretrained condition encoder.

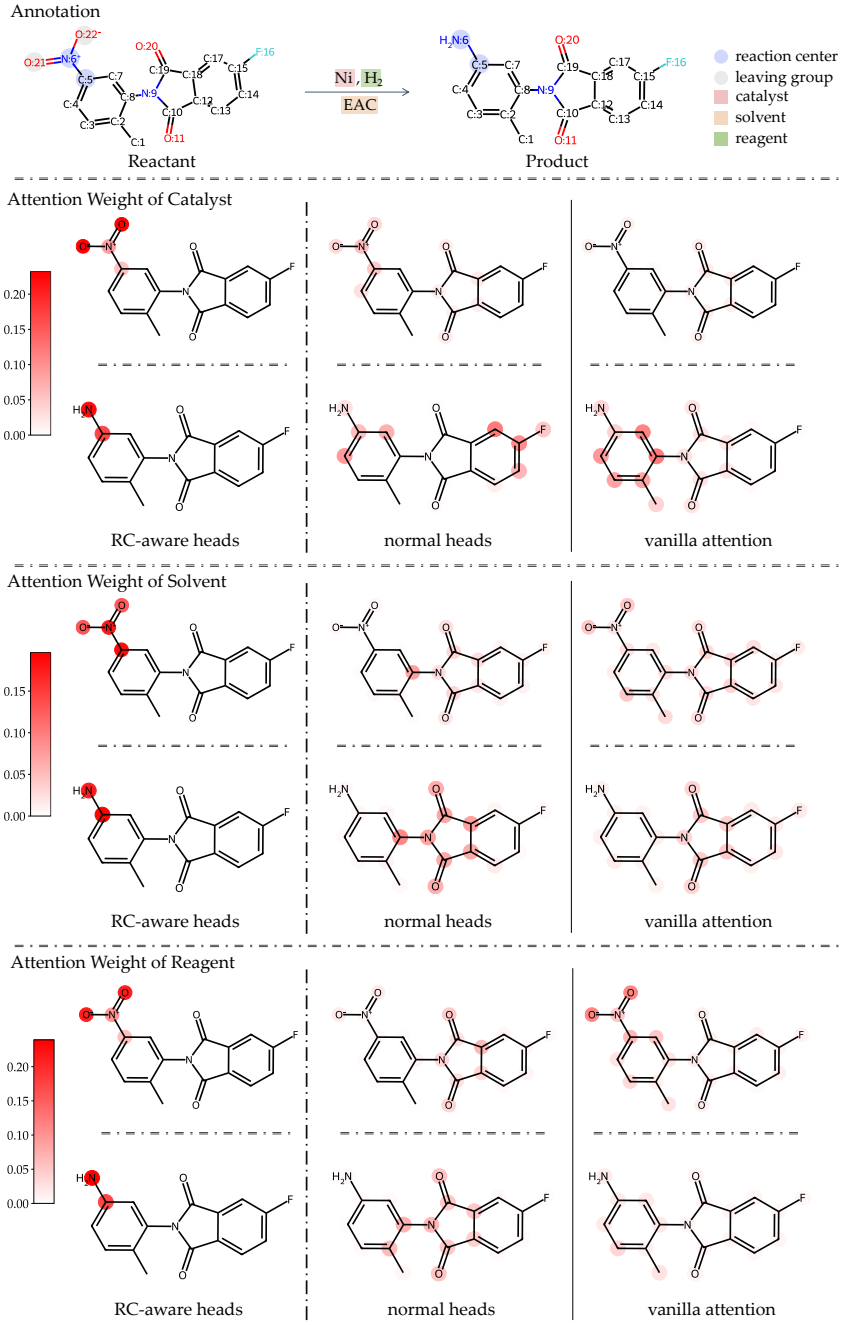
Model	Type	MAE ↓	RMSE ↓	R^2 ↑
DRFP	ML	**0.3213 ± 0.0080**	0.4846 ± 0.0103	0.5643 ± 0.0186
MFF	ML	0.3943 ± 0.0201	0.5616 ± 0.0300	0.4136 ± 0.0604
Chemprop	DL	0.3345 ± 0.0178	**0.4747 ± 0.0191**	**0.5815 ± 0.0339**
ReaMVP	DL	0.5602 ± 0.1106	0.7824 ± 0.0958	-0.1492 ± 0.2900
T5Chem	DL	0.5521 ± 0.0125	0.7496 ± 0.0213	-0.0429 ± 0.0592
RXNFP	DL	0.4975 ± 0.0129	0.7109 ± 0.0179	0.0620 ± 0.0465
Ours	DL	0.3581 ± 0.0126	0.4912 ± 0.0161	0.5521 ± 0.0297
RXNFP*	DL	0.3697 ± 0.0793	0.6053 ± 0.0641	0.3142 ± 0.1471
T5Chem*	DL	0.3920 ± 0.0359	0.5573 ± 0.0492	0.4203 ± 0.1009
ReaMVP*	DL	0.5403 ± 0.0367	0.7090 ± 0.0460	0.0644 ± 0.1225
Ours*	DL	**0.2977 ± 0.0090**	**0.4215 ± 0.0151**	**0.6702 ± 0.0234**

components unchanged. All models are trained on the USPTO-Condition dataset. We present four representative cases below. The attention weights of each node in the reactants and products, while predicting the reaction condition combination, are visualized.

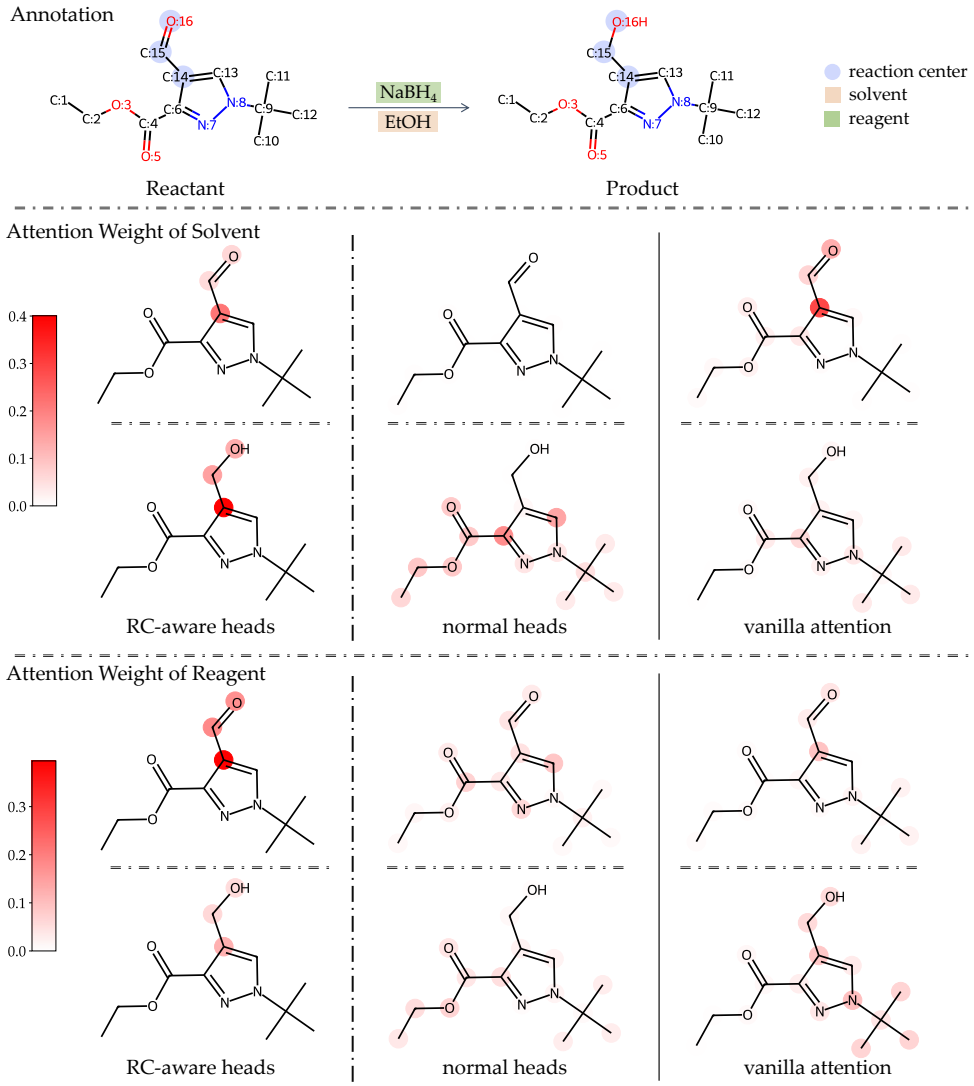
- **Case 1: Reduction of an Aromatic Nitro Group**
  Fig. 1 presents a representative case of the reduction of an aromatic nitro group, with attention weights averaged across heads of each type. This transformation is typically achieved via metal catalyzed hydrogenation, a common industrial method valued for its relative eco friendliness and efficiency [10]. However, this approach can become unsuitable when the aromatic ring bears competing groups, such as halogens [11, 12].
  In the non-RC-Aware baseline, the model consistently focuses on atoms within topologically complex structures (e.g., rings) or the reaction centers across all prediction steps, with attention weights distributed relatively evenly among all atoms of the molecule. This undifferentiated pattern offers limited insight into which specific structural features inform the model's decisions for different reaction components.
  In contrast, our RC-Aware decoder exhibits a clear division of labor by design. The RC-constrained heads are architecturally restricted to attend exclusively to the reaction center. This fixed focus on the transformation site anchors the model's understanding of the reaction type. Consequently, the remaining global heads are liberated from this responsibility and can flexibly attend to peripheral substituents with task-specific patterns. During catalyst prediction, the global heads exhibit distinct attention on the presence of halogen atoms on the aromatic ring, a structural feature that can influence catalyst selection. During solvent prediction, they shift focus to regions outside the reaction center, aligning with the structural features relevant to solvent selection. In reagent prediction, while the RC-constrained heads

attend to the leaving group, which ultimately combines with hydrogen to complete the reduction, the global heads simultaneously monitor other relevant functional groups.

- **Case 2: Reduction of an Aldehyde**
  In the second case presented in Supplementary Fig. 2, which involves the reduction of an aldehyde, the prediction of the reagent sodium borohydride draws the attention of the non-RC-Aware baseline to the reaction center, as well as to the topologically complex tert-butyl group and aromatic ring. However, it overlooks the ester group on the opposite side of the molecule, which represents a potential competing site for reduction and can be reduced by the more reactive reagent lithium aluminum hydride, included in the candidate reagent library of the USPTO-Condition dataset. In contrast, while the RC-Aware heads focus on the reaction center, the normal heads of our model concurrently attend to this potential competing site, reflecting a more comprehensive consideration of the molecular context.
- **Case 3: Synthesis of a Ketone from a Weinreb Amide**
  In the reaction shown in Supplementary Fig. 3, a ketone is synthesized from a Weinreb amide and 4-methylpyridine using a lithium reagent. Lithium diisopropylamide (LDA) is selected as the base for this transformation because the reaction requires a metal ion, such as lithium, to form a stable five-membered chelated intermediate with the Weinreb amide. This chelation is essential for facilitating subsequent mild nucleophilic addition [13, 14]. Additionally, the strong electron-withdrawing nature of the pyridine ring enhances the acidity of the 4-methyl protons in 4-methylpyridine, enabling a strong base to deprotonate this site and generate the carbon nucleophile. The ideal reagent must, therefore, fulfill three roles: it must provide a metal ion for chelation catalysis, it must be sufficiently basic to deprotonate 4-methylpyridine, and it must exhibit low nucleophilicity to avoid direct attack on the Weinreb amide. LDA, as a sterically hindered lithium amide base, satisfies all three requirements. Within this chemical context, the normal attention heads of our model exhibit a focus pattern that aligns closely with the underlying logic of the transformation. Significant attention is allocated to atoms `N:18` and `O:17` of the Weinreb amide, precisely the atoms involved in forming the essential five-membered chelation ring. Concurrently, the model maintains substantial attention on the entire pyridine ring of 4-methylpyridine, consistent with the electron-withdrawing character that renders its 4-methyl group a potential reactive site. This attention distribution suggests that the model integrates multiple factors influencing the reaction outcome.
- **Case 4: Selective Reduction of an Ester to an Aldehyde**
  Supplementary Fig. 4 shows that when predicting the reagent, the normal heads compensate for the attention of the RC-Aware heads to the aldehyde group at the reaction center in the product, which helps the model predict the correct reducing agent, diisobutylaluminum hydride, rather than a more powerful reductant, such as lithium aluminum hydride.



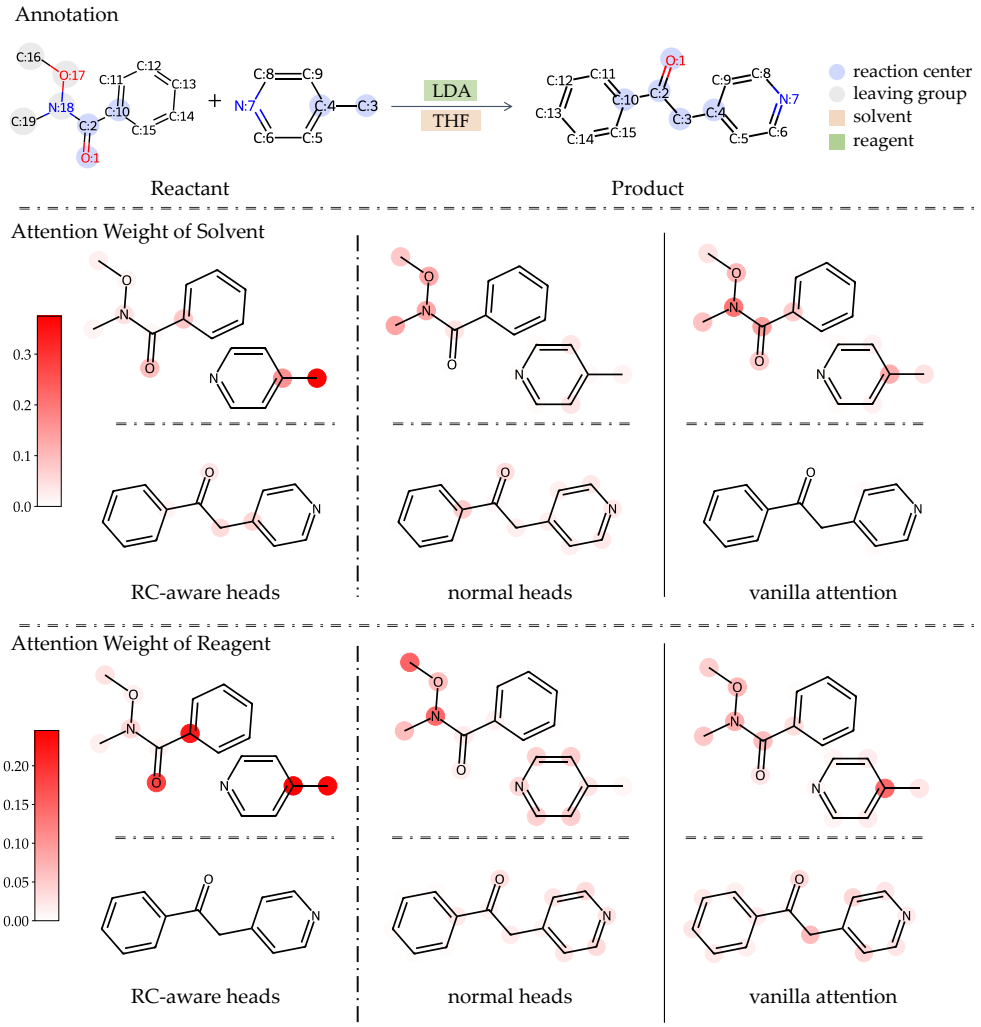
Supplementary Fig. 1: **The visualization the attention weights of each node in reactants and products while predicting the reaction condition combination for the reduction of nitro groups.**



Supplementary Fig. 2: Supplementary case one on the RC-Aware cross attention mechanism. The attention weights of each node in reactants and products while predicting the reaction condition combination of the given reaction is visualized.

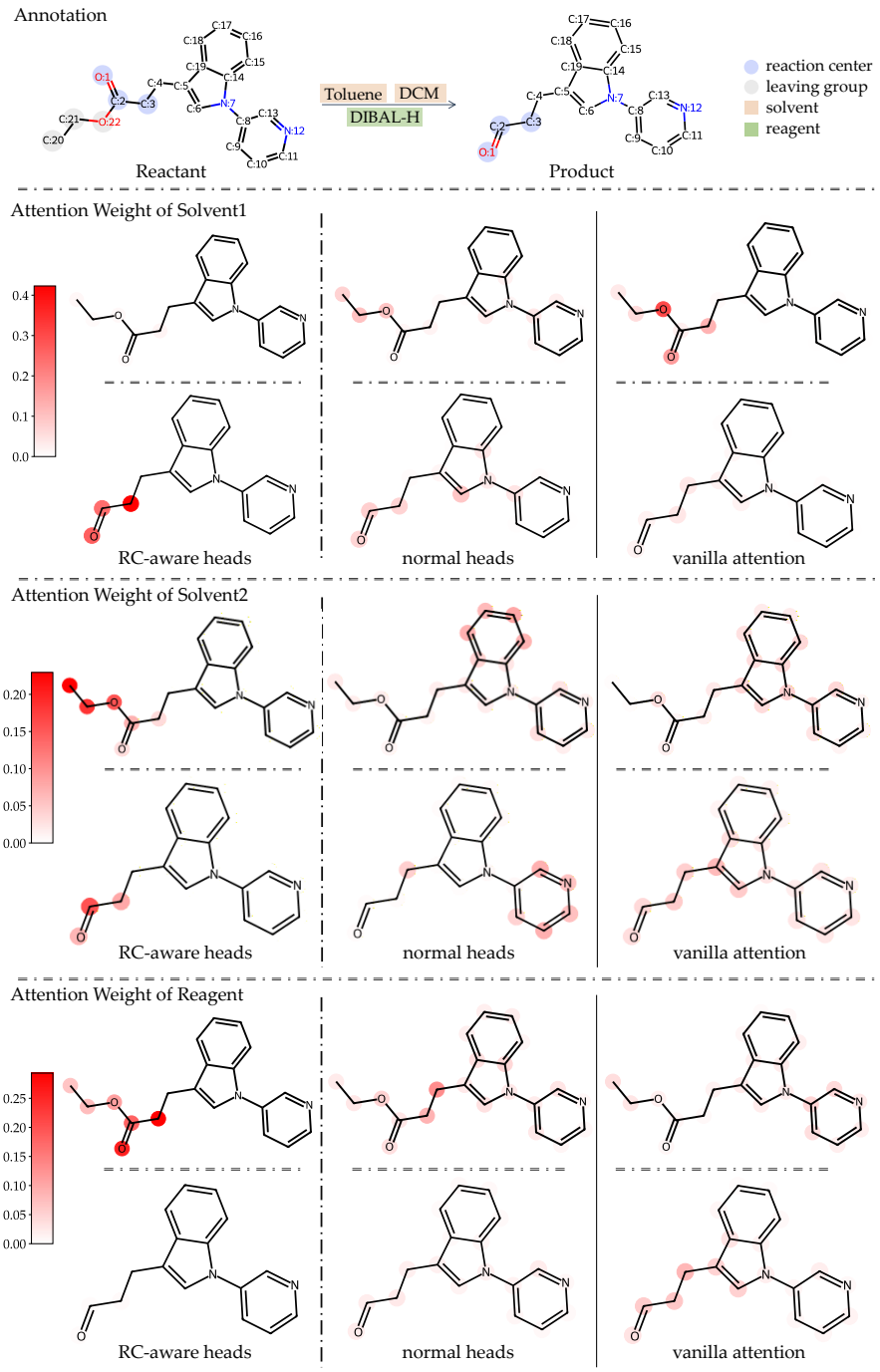
2.4.3 Summary of Qualitative Observations

Synthesizing the qualitative patterns observed across these four distinct reactions reveals several key characteristics of the RC-Aware framework's information aggregation. While conventional attention mechanisms (the non-RC-aware baseline) often concentrate focus on regions with high topological complexity or localized reaction



Supplementary Fig. 3: Supplementary case two on the RC-Aware cross attention mechanism. The attention weights of each node in reactants and products while predicting the reaction condition combination of the given reaction is visualized.

centers, our framework facilitates a more structured partitioning of focus. Specifically, the results suggest a dual-stream attention strategy. The RC-aware heads, as architecturally intended, consistently focus on the predefined reaction center atoms. Meanwhile, the behavior of the global heads adapts to the chemical context: in complex molecules, they prioritize potential interfering functional groups or distal motifs relevant to condition selection. Notably, as illustrated in Cases 3 and 4, the global



Supplementary Fig. 4: Supplementary case three on the RC-Aware cross attention mechanism. The attention weights of each node in reactants and products while predicting the reaction condition combination of the given reaction is visualized.

heads operate in synergy with the RC-aware heads, providing complementary reinforcement that stabilizes the representation of the reaction environment when distal functional groups are absent.

These observations do not constitute proof of the model's causal reasoning or its understanding of chemistry. However, the visual alignment between the observed attention patterns and chemically intuitive concepts in these specific instances suggests that the proposed architectural bias can lead to attention patterns that are more structured and context-dependent than the baseline. This observation presents a potential avenue for future hypothesis-driven research, which would require rigorous quantitative validation and perturbation-based analysis to establish any causal links.

3 Implementation details

3.1 Initial node/edge feature extraction

We use the atom and bond encoder provided by Open Graph Benchmark [15]. These encoders transform atoms and chemical bonds into integers based on their characteristics and molecular structures. Nine atom descriptors are provided, including the atom type, formal charge, and other properties that can be calculated by RDKit [16]. For chemical bonds, three descriptors are used: bond type, bond stereochemistry, and whether the bond is conjugated. Each descriptor corresponds to a learnable embedding table. The initial node and edge features are derived by aggregating the embeddings corresponding to each descriptor.

3.2 MPNN Layer in the Atom-Aligned Encoder

In contrast to many graphs where edges convey limited information, the chemical bonds within molecules play a pivotal role in determining their properties. Consequently, we employ a variant of the Graph Attention Network(GAT) [17], akin to those utilized in the works of Yan et al. [18] and Sacha et al. [19], to integrate information from chemical bonds into the node features during the message-passing process. The MPNN layer we implement can be mathematically formulated as follows:

$$
\begin{aligned}
\tilde{e}_{u,v} &= \mathrm{FFN}_e(e_{u,v}), \\
\tilde{h}_u &= \mathrm{FFN}_n(h_u), \\
c_{u,v} &= \mathbf{a}^T[\tilde{h}_u \| \tilde{h}_v \| \tilde{e}_{u,v}], \\
\alpha_{u,v} &= \frac{\exp(\mathrm{LeakyReLU}(c_{u,v}))}{\sum_{v' \in \mathcal{N}(u)\cup\{u\}} \exp(\mathrm{LeakyReLU}(c_{u,v'}))}, \\
h'_u &= \sum_{v \in \mathcal{N}(u)\cup\{u\}} \alpha_{u,v} \left(\tilde{h}_u + \tilde{e}^{(k)}_{u,v} \right),
\end{aligned}
\tag{1}
$$

where h_u represents the input node feature of node u, $e_{u,v}$ represents the input edge feature of edge (u, v), and h'_u represents the output node feature of u in the message passing layer.

3.3 Condition Encoders for Buchwald–Hartwig Dataset and thiol addition Selectivity Dataset

The Buchwald–Hartwig and thiol addition selectivity datasets employ chemical reagents as reaction conditions, with condition features generated via graph neural networks. Given the limited diversity of reagents in these datasets, and considering that several baselines were pretrained on large-scale reaction corpora, we adopted Hu et al.'s pretrained molecular representation model [20] as our condition encoder for a fair comparison. This lightweight model underwent pretraining exclusively through random atom masking. The output representation of its k-th layer is formalized as:

$$\begin{aligned} X &= \sum_{u \in \mathcal{N}(v) \cup \{v\}} h_u^{(k-1)} + \sum_{e=(u,v)} \sum_{u \in \mathcal{N}(v) \cup \{v\}} h_e^{(k-1)}, \\ h_v^{(k)} &= \mathrm{ReLU}\left(\mathrm{MLP}^{(k)}\left(X\right)\right), \end{aligned} \tag{2}$$

where $h_u^{(k)}$ is the node feature of node u in the k-th layer, and $h_e^{(k)}$ is the edge feature of edge e in the k-th layer. To conduct pure architectural comparisons, we also implemented a model using a simple GAT as the condition encoder, trained from scratch, and benchmarked it against the non-pretrained versions of all baselines. In these experiments, the simple GAT implementation follows Eq. 1.

3.4 Model Implementation details

We implement our model based on `PyTorch 1.13` [21] and `torch_geometric 2.2.0` [22]. All models are trained using the Adam optimizer [23]. We summarize the training and model architecture hyper-parameters in Supplementary Table 13, Supplementary Table 14, and Supplementary Table 15. For the reaction condition recommendation task, we use the following strategy to adjust the learning rate during training: we gradually increase our learning rate to the maximum during the first few epochs and then decrease it using exponential decay. An early stopping strategy is also used, which halts the training when the evaluation metrics on the validation set do not increase for 20 epochs.

Supplementary Table 13: The training and model architecture hyper-parameters for reaction condition recommendation.

dataset	USPTO-Condition	USPTO-500MT
hidden size	384	256
learning rate	1.5e-4	1.5e-4
dropout ratio	0.1	0.1
number of encoder layers	6	6
number of decoder layers	6	6
number of attention heads	6	8
number of training epochs	200	200

Supplementary Table 14: The training and model architecture hyper-parameters for the regression tasks (with pretrained condition encoder).

dataset	Buchwald-Hartwig	radical C-H functionalization	thiol addition
hidden size of main model	128	128	128
number of layers of main model	3	5	3
hidden size of condition encoder	300	-	300
number of layers of condition encoder	5	-	5
learning rate	5e-4	5e-4	5e-5
dropout ratio	0.1	0	0
number of attention heads	8	8	8
number of training epochs	200	200	250

The training is conducted on Ubuntu 22.04.5, with two NVIDIA RTX 6000 GPUs. The training on the USPTO-Condition dataset uses two GPUs and takes around 60 hours. The training of other datasets is conducted on a single GPU. The training on the USPTO-500MT dataset takes around 24 hours with early stopping. The training for yield prediction on a single split with a single random seed takes around 2 hours. The training for selectivity prediction on the radical C-H functionalization dataset takes approximately 2 hours with a single split and a single random seed. The training on the chiral phosphoric acid-catalyzed thiol addition dataset with a single split and a single random seed takes around 1 hour.

4 Baselines

4.1 Chemical Reaction Condition Prediction

We chose T5Chem [2], Parrot [1], GCNN [24], Reagent Transformer [25], and FPRCR [26] as our baselines for chemical reaction condition recommendation; the details are as follows.

- Parrot is a transformer-based model that accepts reaction SMILES as input to predict reaction condition combinations. It specifically targets condition sets with a fixed set of components. The model was pretrained on curated SMILES from USPTO 1976-2016sep [27] using standard BERT-style masked language modeling [28].
- Reagent Transformer is a transformer-based model applied to both the USPTO-Condition and USPTO-500MT datasets. For a fair comparison on the USPTO-Condition dataset, we replaced its SMILES-generation vocabulary with a predefined molecular library and reformulated condition prediction as a fixed-length sequence generation task, adapting the original decoder design that produces SMILES strings.
- T5Chem is a language model pretrained on the PubMed dataset [29] using self-supervised learning and on the USPTO-500MT dataset [2] with five supervised tasks, including reaction condition generation. We evaluated this baseline on both the USPTO-Condition and USPTO-500MT datasets. Analogous to our adaptation of Reagent Transformer for USPTO-Condition, we replaced T5Chem's

Supplementary Table 15: The training and model architecture hyper-parameters for the regression tasks (without pretrained condition encoder).

dataset	Buchwald-Hartwig		radical C-H functionalization	thiol addition	USPTO-Yield
datasplit	random	OOD	-	all	-
hidden size of main model	128	64	128	128	384
number of layers of main model	3	3	5	3	6
hidden size of condition encoder	128	64	-	128	384
number of layers of condition encoder	3	3	-	3	6
learning rate	5e-4	5e-4	5e-4	5e-5	1e-4
dropout ratio	0.1	0.1	0	0	0.1
number of attention heads	8	8	8	8	8
number of training epochs	200	200	200	250	200

SMILES-generation vocabulary with a predefined molecular library and reformulated condition prediction as a fixed-length sequence generation task producing five components.

- FPRCR is a fingerprint-based method specifically designed for reaction condition combinations with a fixed number of components. The model comprises a set of MLPs that correspond to the number of components in each reaction condition combination. All MLPs are invoked in the order of prediction, using the fingerprints of reactants, products, and the already predicted components of the reaction condition combination as input to forecast the next molecule that should be used as a condition.
- GCNN is a graph-based method that takes the molecular graphs of reactants and products as input to generate chemical reaction representations for predicting reaction condition combinations. The reaction condition combinations in the USPTO-Condition dataset consist of up to one catalyst, two solvents, and two reagents. We have extended the model's prediction head from a single multilabel classification head to five separate single-label classification heads.

4.2 Chemical Reaction Yield/Selectivity Prediction

The tasks of predicting chemical reaction yield and selectivity are highly analogous regression tasks; hence, their baselines are identical. **It is noteworthy that the emphasis of this paper is on introducing a novel architecture for chemical reaction representation learning, rather than being limited to the specific task of chemical reaction yield/selectivity prediction.** Therefore, we have not included baselines with complex pretraining tasks or intricate training loss and inference strategies. Pretraining and specialized loss functions tailored for reaction yield prediction are orthogonal to our research. The details of the chosen baselines are as follows.

- DRFP [4] is a type of chemical reaction fingerprint for yield prediction that calculates the fingerprint of the symmetric difference between the n-grams of reactants and products.
- MFF [30] utilizes multiple fingerprint features from all molecules in one reaction as input for the regressor.
- Chemprop [31] is a type of Message Passing Network that encodes reactions. It converts the reactant and product SMILES pairs into a single pseudo-molecule known as the condensed graph representation (CGR) of the reaction and processes it through a directed messaging passing network (D-MPNN) to extract reaction features.
- RXNFP [32] is a transformer model that has been pre-trained on the Pistachio dataset [33] using self-supervised learning tasks. Building upon this work, the original research team has also proposed a finetuned version for predicting chemical reaction yields, termed YieldBert [34].
- T5Chem [2] is a language model pretrained on the PubMed dataset [29] with a self-supervised task and the USPTO-500MT dataset [2] with five different supervised tasks, including reaction yield prediction.

- ReaMVP [35] is a multi-modal model that takes both SMILES and molecular graphs as input. It is pretrained on USPTO 1976-2016 sep [27] and CJHIF [36] using contrastive learning.

5 Integrating Diverse Reaction Conditions: A Case Study

To explore the extensibility of our proposed Condition Adapter and its potential to integrate non-molecular reaction conditions, we conducted a preliminary case study using the open-source USPTO 1976-2016sep dataset [27] for yield prediction.

5.1 Data Extraction

We converted the original XML files to JSON and extracted the following information:

- Reactant list and their corresponding amounts.
- Reagent list and their corresponding amounts.
- Product SMILES.
- Temperatures associated with all steps in the procedural sequence.
- Product yield.
- Mapped reaction SMILES.

5.2 Data Curation Pipeline

The extracted data underwent the following cleaning and standardization steps:

1. **Atom Mapping Assignment:** Based on the mapped reaction SMILES, we assigned atom mapping numbers to each reactant and reagent. Components were matched by canonical SMILES, with priority given to reactants. Reactions with duplicate atom-mapping numbers were removed.
2. **Re-refining Reactants & Reagents:** Components containing atom mapping numbers present in the products were re-classified as reactants; all others were treated as reagents.
3. **Amount Normalization:** We processed only three units: mole, mass, and volume. Other units were discarded (marked as `None`). Mass amounts were converted to moles using molecular weight. All amounts were normalized to base units: moles (mol) for quantity and liters (L) for volume. The smallest non-`None` molar amount among reactants was defined as the theoretical maximum yield (1.0 equivalent). All reagent amounts were normalized by this equivalent, and the reaction yield was calculated accordingly.
4. **Temperature Assignment:** The temperature value that is furthest from room temperature among the recorded procedural steps was selected as the reaction temperature.
5. **Data De-duplication:** Reactions were considered duplicates if they shared identical combinations of (a) reactants and their amounts, (b) products, (c) reagents

and their amounts, and (d) reaction temperature. All duplicate entries were aggregated, and their yields were averaged to produce a single representative value for each unique reaction condition.

Following this pipeline, we obtained 756,544 reactions, which were randomly split into training, validation, and test sets in an 8:1:1 ratio.

5.3 Model Implementation and Experimental Results

Reagents (molecular), their normalized amounts, and the reaction temperature were used as conditions. We trained two model variants:

- Only molecular reagents were used as conditions.
- Molecular reagents, *along with* their amounts and the reaction temperature, were used as conditions. To prevent label leakage, the amounts of the products were excluded.

Implementation Details:

- Molecular reagents were encoded using a Simple GAT, whose message-passing layer follows Eq. 1.
- Numerical conditions (amounts, temperature) were projected into a $\mathbb{R}^{1\times d}$ feature vector. We adopted a cluster-based embedding method [37], implemented as Linear(Softmax(Linear(x))), where x is the input numerical value.
- To integrate amount information into reagent embeddings, the adapter structure defined in Section 4.3 of the main text was used to fuse the outputs of the Simple GAT and the numerical condition embeddings.

Supplementary Table 16: Performance Comparison With and Without Non-Molecular Reaction Conditions

Amount	Temperature	R^2 ↑	MAE ↓	RMSE ↓
✓	✓	0.2495	18.2864	22.7346
✓	×	0.2527	18.1232	22.6864
×	✓	0.2291	18.5609	23.0410
×	×	0.2373	18.2872	22.9180

The results are summarized in Supplementary Table 16, which details the impact of incorporating two specific types of non-molecular conditions on the yield regression task. Interestingly, we observe that the inclusion of temperature as a condition leads to a decline in model performance across both model variants, regardless of whether the reagent amount is also included. We attribute this counterintuitive result to the inherent noise in the temperature data, which is parsed from the procedural sequence using a rule-based heuristic: the temperature value furthest from room temperature is selected as the reaction temperature. This heuristic introduces considerable uncertainty and may not accurately reflect the true reaction conditions. In

contrast, when the reagent amount is introduced as an additional condition, a noticeable improvement in predictive accuracy is observed. This result suggests that despite the challenges posed by the noisy temperature feature, the integration of the reagent amount provides a meaningful and complementary signal that enhances model performance. Nevertheless, this experiment serves as a proof-of-concept demonstrating that our adapter architecture can technically accommodate diverse non-molecular condition modalities. While further refinement and higher-quality datasets are required to fully unlock its predictive benefits, it lays the groundwork for future developments in more comprehensive reaction representations.

References

[1] Wang, X. *et al.* Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research* **6** (2023).

[2] Lu, J. & Zhang, Y. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling* **62**, 1376–1387 (2022).

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

[4] Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discovery* (2022).

[5] Li, X., Zhang, S.-Q., Xu, L.-C. & Hong, X. Predicting regioselectivity in radical c- h functionalization of heterocycles through machine learning. *Angewandte Chemie International Edition* **59**, 13253–13259 (2020).

[6] Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).

[7] Li, S.-W., Xu, L.-C., Zhang, C., Zhang, S.-Q. & Hong, X. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nature Communications* **14**, 3569 (2023).

[8] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **7**, eabe4166 (2021).

[9] Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572–1583 (2019).

[10] Orlandi, M., Brenna, D., Harms, R., Jost, S. & Benaglia, M. Recent developments in the reduction of aromatic and aliphatic nitro compounds to amines. *Organic Process Research & Development* acs.oprd.6b00205 (2016).

[11] Shi, Q. *Catalytic Transfer Hydrogenation of Nitronarenes under Mild Conditions*. Ph.D. thesis, Dalian University of Technology (2007).

[12] Yao, M. & Zhang, R. The progress of the reduction reaction of aromatic nitro compounds. *Chemical Industry and Engineering Progress* 19–25 (1984).

[13] Zhao Wei, L. W. Progresses of weinreb amides in organic synthesis. *Chinese Journal of Organic Chemistry* **35**, 55 (2015). URL https://sioc-journal.cn/Jwk_yjhx/EN/abstract/article_344621.shtml.

[14] Nahm, S. & Weinreb, S. M. N-methoxy-n-methylamides as effective acylating agents. *Tetrahedron Letters* **22**, 3815–3818 (1981).

[15] Hu, W. *et al.* Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems* **33**, 22118–22133 (2020).

[16] Landrum, G. *et al.* Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8**, 31 (2013).

[17] Veličković, P. *et al.* Graph attention networks (2018). International Conference on Learning Representations.

[18] Yan, C. *et al.* Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems* **33**, 11248–11258 (2020).

[19] Sacha, M. *et al.* Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling* **61**, 3273–3284 (2021).

[20] Hu, W. *et al.* Strategies for pre-training graph neural networks (2020). URL https://openreview.net/forum?id=HJlWWJSFDH. International Conference on Learning Representations.

[21] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[22] Fey, M. *et al.* PyG 2.0: Scalable learning on real world graphs (2025). Temporal Graph Learning Workshop @ KDD.

[23] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Maser, M. R. *et al.* Multilabel classification models for the prediction of cross-coupling reaction conditions. *Journal of Chemical Information and Modeling* **61**,

156–166 (2021).

[25] Andronov, M. *et al.* Reagent prediction with a molecular transformer improves reaction data quality. *Chemical Science* **14**, 3235–3246 (2023).

[26] Gao, H. *et al.* Using machine learning to predict suitable conditions for organic reactions. *ACS Central Science* **4**, 1465–1476 (2018).

[27] Lowe, D. Chemical reactions from US patents (1976-Sep2016) (2017). Figshare https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

[28] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding (2019). URL https://doi.org/10.18653/v1/n19-1423. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers).

[29] Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**, D1388–D1395 (2020). URL https://doi.org/10.1093/nar/gkaa971.

[30] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020). URL https://www.sciencedirect.com/science/article/pii/S2451929420300851.

[31] Heid, E. *et al.* Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling* **64**, 9–17 (2024). URL https://doi.org/10.1021/acs.jcim.3c01250. PMID: 38147829.

[32] Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **3**, 144–152 (2021).

[33] Mayfield, J., Lowe, D. & Sayle, R. Pistachio: Search and faceting of large reaction databases (2017). ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY.

[34] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2**, 015016 (2021).

[35] Shi, R., Yu, G., Huo, X. & Yang, Y. Prediction of chemical reaction yields with large-scale multi-view pre-training. *Journal of Cheminformatics* **16**, 22 (2024).

[36] Jiang, S. *et al.* When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access* **9**, 85071–85083 (2021).

[37] Liu, G., Xu, J., Luo, T. & Jiang, M. Graph diffusion transformers for multi-conditional molecular generation. *Advances in Neural Information Processing Systems* **37**, 8065–8092 (2024).