Supporting Information

Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions

An Su, ‡[a] Xinqiao Wang, ‡[b] Ling Wang, [b] Chengyun Zhang, [b] Yejian Wu, [b] Xinyi Wu, [b] Zhaoqingjie, [d] and Hongliang Duan*[bc]

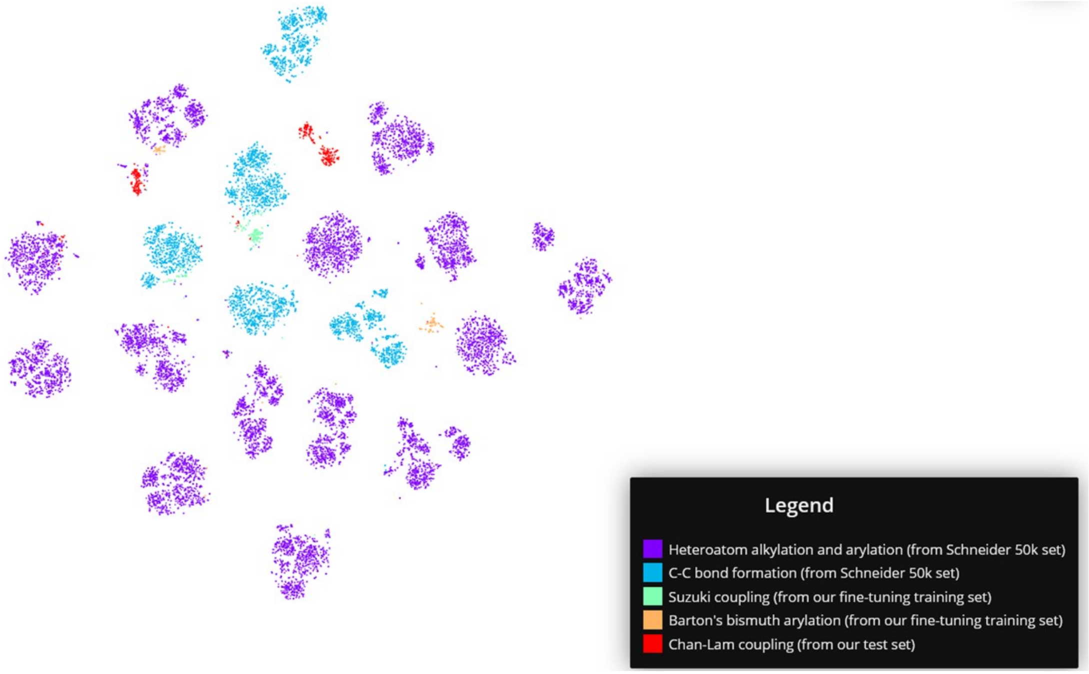
[a] College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, P. R. China

[b] Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, P. R. China.

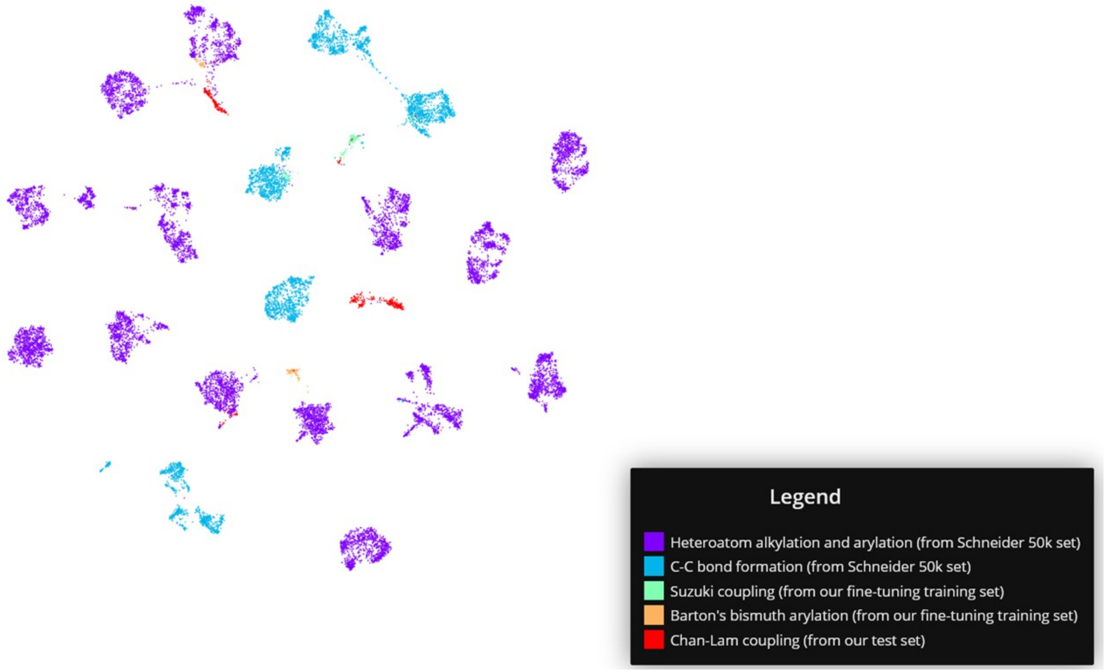
[c] State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica (SIMM), Chinese Academy of Sciences, Shanghai 201203, China.

*Email: hduan@zjut.edu.cn

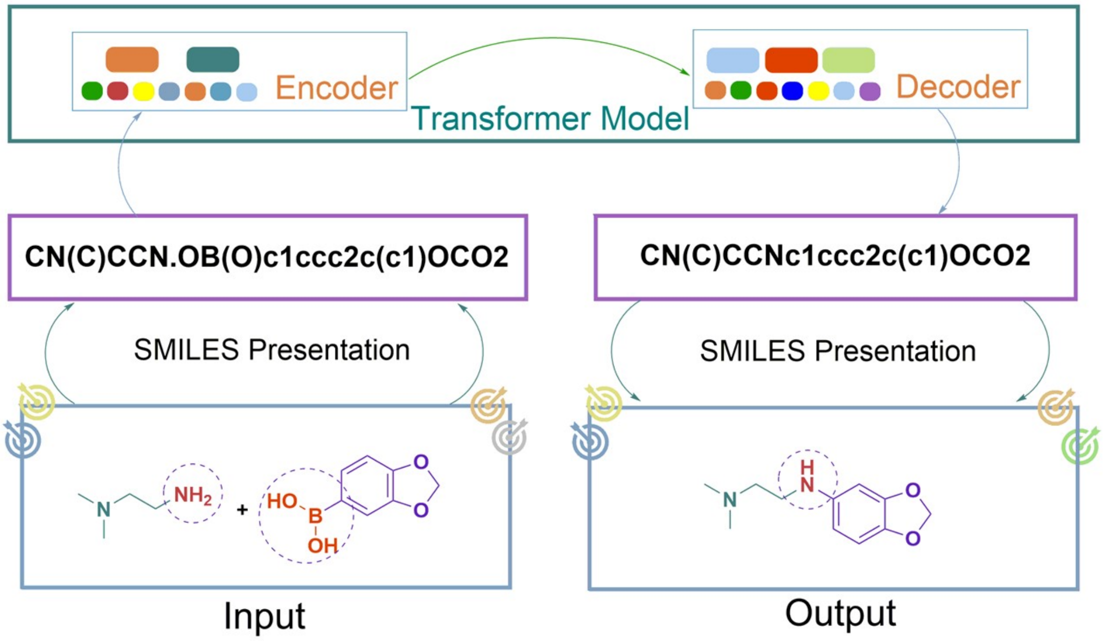
[d] Shanghai Institute of Material Medical, Chinese Academy of Sciences, Shanghai 201203, P. R. China.



Supplementary Fig. 1 Reactions in the fine-tuned dataset and test set of our study are classified, as well as two related types of reactions in the Schneider 50K.[1] The Schneider 50K set is Not used to train our model, but only for visualization purposes. The fingerprints are generated using *rxnfp*,[2] and the reactions are visualized using t-SNE algorithm.[3]

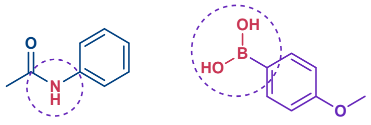
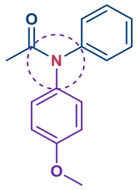
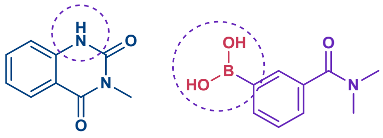
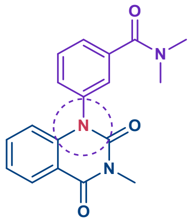
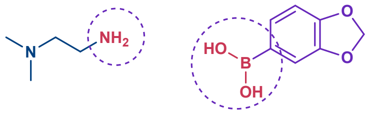
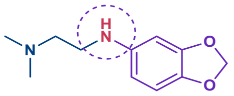
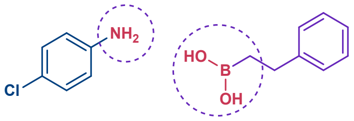
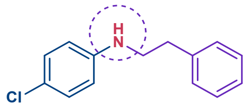
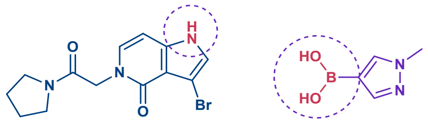
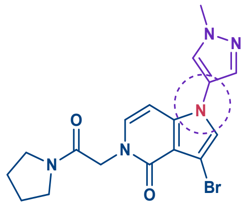


Supplementary Fig. 2 Reactions in the fine-tuned dataset and test set of our study are classified, as well as two related types of reactions in the Schneider 50K[1]. The Schneider 50K set is Not used to train our model, but only for visualization purposes. The fingerprints are generated using *rxnfp*[2], and the reactions are visualized using UMAP algorithm[4].



Supplementary Fig. 3 The process of reaction prediction by the Transformer. The Transformer model is composed of encoder and decoder.

Supplementary Table 1 Examples of C-N coupling reactions correctly predicted via ZSRP model.

Reactant type	Reactants	Zero-shot learning's prediction
Amide (linear)		
Amide (cyclic)		
Aliphatic amine		
Aromatic amine		
N-aromatic heterocyclic		

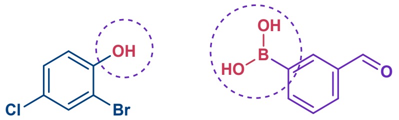
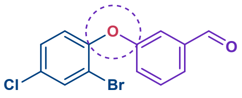
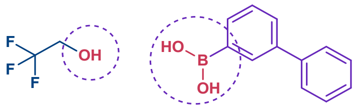
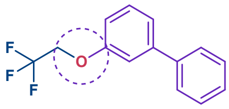
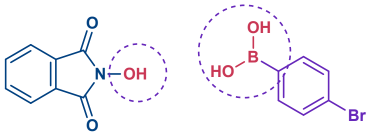
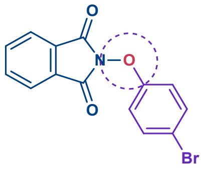
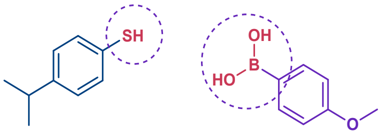
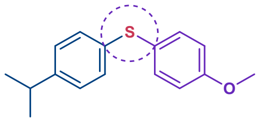
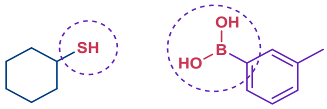
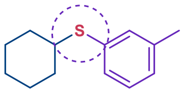
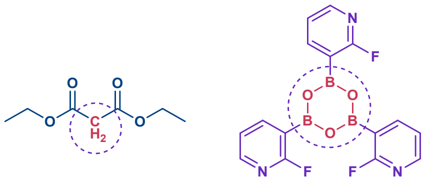
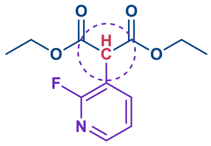
Supplementary Table 2 Performance of ZSRP model for C-O Chan-Lam coupling categorized by reactant type.

Reactant type	Test samples	Top-1 accuracy (%)
Aromatic alcohol	133	69.9
Aliphatic alcohol	23	34.8
Amide alcohol	4	75.0
Total	160	65.0

Supplementary Table 3 Performance of ZSRP model for C-S Chan-Lam coupling categorized by reactant type.

Reactant type	Test samples	Top-1 accuracy (%)
Thiophenol	20	85.0
Thiol	6	83.3
Total	26	84.6

Supplementary Table 4 Examples of C-O/C-S/C-C Chan-Lam coupling correctly predicted via ZSRP model.

Reactant type	Reactant	Zero-shot learning's prediction
Aromatic alcohol (C-O)		
Aliphatic alcohol (C-O)		
Amide alcohol (C-O)		
Thiophenol (C-S)		
Thiol (C-S)		
Methylene (C-C)		

Supplementary Table 5 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the first reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	83.1	86.8
100	83.6	86.8
150	84.0	86.8
200	84.0	86.6
250	83.8	87.1

300	83.6	87.1
350	83.8	87.1
400	83.8	87.1
450	83.6	87.1
500	83.8	86.4
550	83.8	86.8
600	83.8	86.8

Supplementary Table 6 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the second reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	76.4	81.5
100	77.5	82.0
150	78.0	82.4
200	78.2	82.6
250	78.2	82.6
300	78.9	82.8
350	78.9	83.5
400	78.9	83.5
450	79.1	83.5
500	78.9	83.7
550	78.9	83.9
600	79.1	83.9

Supplementary Table 7 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the third reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	77.3	79.4
100	77.3	80.9
150	77.5	81.5
200	77.5	81.5
250	772	81.3
300	77.2	81.3
350	77.5	81.3
400	77.5	81.3
450	77.8	81.3
500	77.5	81.3
550	76.9	81.3
600	77.2	81.3

Supplementary Table 8 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the fourth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	60.1	59.9
100	59.6	61.0
150	60.1	61.2
200	59.9	61.6
250	60.1	62.3
300	60.1	62.7
350	59.6	62.7
400	59.6	62.3
450	59.8	62.7
500	59.8	62.9
550	60.1	63.1
600	60.5	63.1

Supplementary Table 9 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the fifth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	61.6	67.4
100	63.0	67.8
150	64.1	68.2
200	65.0	68.8
250	65.2	68.8
300	65.0	69.1
350	65.2	69.3
400	65.2	69.1
450	65.0	69.3
500	65.0	69.9
550	65.2	69.9
600	66.1	70.8

Supplementary Table 10 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the sixth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7

50	78.4	79.2
100	78.4	79.4
150	79.3	79.4
200	80.0	79.6
250	80.7	79.4
300	81.4	79.4
350	81.6	79.4
400	81.6	79.6
450	82.0	79.9
500	81.8	79.9
550	81.6	79.9
600	81.6	79.9

Supplementary Table 11 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the seventh reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	68.1	69.5
100	68.1	69.3
150	68.1	69.1
200	67.9	68.8
250	67.7	69.5
300	67.7	69.3
350	67.4	69.3
400	67.2	69.5
450	67.2	69.5
500	67.2	69.5
550	67.2	69.5
600	67.2	69.5

Supplementary Table 12 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the eighth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	66.6	71.6
100	67.0	71.8
150	67.5	72.4
200	67.7	72.5
250	68.1	73.1
300	68.1	73.7
350	68.4	73.9

400	68.6	74.4
450	68.6	74.4
500	68.6	74.6
550	69.3	75.0
600	69.3	75.2

Supplementary Table 13 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the ninth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	57.6	61.6
100	58.7	62.1
150	58.7	62.1
200	58.5	61.9
250	58.3	61.9
300	58.5	61.9
350	58.3	61.6
400	58.0	61.6
450	58.0	61.6
500	58.3	61.6
550	58.7	61.5
600	58.9	61.8

Supplementary Table 14 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the tenth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	64.8	65.0
100	63.9	65.0
150	64.1	65.2
200	63.4	65.5
250	63.0	65.0
300	62.5	65.2
350	62.5	65.0
400	62.3	64.6
450	62.3	64.0
500	62.3	63.8
550	62.3	63.8
600	62.3	63.8

Supplementary Table 15 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the eleventh reaction

in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	68.8	70.3
100	69.0	70.5
150	69.5	70.8
200	69.5	71.2
250	69.5	71.0
300	69.5	71.2
350	69.0	71.0
400	69.0	71.0
450	69.0	70.8
500	68.6	70.8
550	68.8	70.8
600	68.8	70.8

Supplementary Table 16 OSRP performance of the training Chan-Lam coupling reaction varies with the training steps, the training Chan-Lam coupling reaction sample is the twelfth reaction in the Table 4.

Steps	Top-1 accuracy (%)	
	Validation	Test
0	55.1	55.7
50	61.9	64.2
100	61.9	63.7
150	61.4	63.5
200	61.0	63.5
250	60.8	64.0
300	61.0	64.0
350	61.0	63.8
400	61.0	63.8
450	60.8	63.8
500	60.8	64.0
550	60.6	64.0
600	60.6	63.6

Supplementary Table 17 The corresponding performance of ZSRP and OSRP model trained with reactions which contained reagents.

model			Top-1 accuracy (%)
ZSRP			51.4
OSRP	C-N	N-aromatic heterocyclic	76.9
	C-N	N-aromatic heterocyclic	83.8
	C-N	Aromatic amine	77.9
	C-N	Aliphatic amine	67.6
	C-N	Amide (cyclic)	75.1

	C-N	Amide (linear)	81.1
	C-O	Aromatic alcohol	75.0
	C-O	Aliphatic alcohol	71.2
	C-O	Amide alcohol	69.5
	C-O	Aromatic alcohol	71.2
	C-S	Thiophenol	68.9
	C-C	Methylene	67.9

Supplementary Table 18 Prediction performance with different fine-tuning strategies.

Training Set	Fine-tuning	Top-1 Accuracy (%)
USPTO + Suzuki + Barton	1 sample (one-shot)	87.1 (highest), 61.6(lowest), 72.3 (mean)
USPTO + Suzuki + Barton	12 samples (few-shot)	92.2
USPTO + Suzuki + Barton	101 samples	94.3

Supplementary Table 19 Distribution of the transformer model with OSRP categorized by coupling type.

Coupling type	Count of examples[a]
C-N	106
C-O	45
C-S	2
Total	153

[a]Examples of reactions where the transformer model with ZSRP predicts wrong but the model with OSRP predicts correct.

Supplementary Table 20 Test performance categorized by coupling type.

Coupling type	Test Samples	Top-1 accuracy (%)	
		One-shot (1 sample, highest)	**Few-shot (12 samples)**
C-N	283	83.4	89.0
C-O	160	93.1	96.3
C-S	26	88.5	100.0
C-C	3	100.0	100.0
Total	472	87.1	92.2

Supplementary Table 21 Test performance for C-N Chan-Lam coupling categorized by reactant type.

Reactant type	Test Samples	Top-1 accuracy (%)	
		One-shot	**Few-shot**

		(1 sample, highest)	**(12 samples)**
Amide (linear)	19	89.5	78.9
Amide (cyclic)	62	79.0	83.9
Aliphatic amine	49	73.5	85.7
Aromatic amine	77	91.0	93.5
N-aromatic heterocyclic	76	84.2	93.4
Total	283	83.4	89.0

Supplementary Table 22 Test performance for C-O Chan-Lam coupling categorized by reactant type.

Reactant type	**Test Samples**	**Top-1 accuracy (%)**	
		One-shot	**Few-shot**
		(1 sample, highest)	**(12 samples)**
Aromatic alcohol	133	94.7	98.5
Aliphatic alcohol	23	86.9	87.0
Amide alcohol	4	75.0	75.0
Total	160	93.1	96.3

Supplementary Table 23 Test performance for C-S Chan-Lam coupling categorized by reactant type.

Reactant type	**Test Samples**	**Top-1 accuracy (%)**	
		One-shot	**Few-shot**
		(1 sample, highest)	**(12 samples)**
Thiophenol	20	90.0	100.0
Thiol	6	83.3	100.0
Total	26	88.5	100.0

Supplementary Table 24 Major reaction datasets used in this study.

Dataset	**Count**
USPTO dataset	367726
Suzuki reaction	200
Barton's bismuth arylation	148
Chan-Lam coupling reaction	1031

References

1. N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity, *J. Chem. Inf. Model.*, 2015, **55**, 39-53.

2. P. Schwaller *et al.,* Mapping the space of chemical reactions using attention-based neural networks,

Nat. Mach. Intell., 2021, **3**, 144-152.

3. L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**, 2579-2605.

4. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).