SynCat: Molecule-Level Attention Graph Neural Network for Precise Reaction Classification

Phuoc-Chung Van Nguyen [a], Van-Thinh To [a], Ngoc-Vi Nguyen Tran [b], Tieu-Long Phan [*cd], Tuyen Ngoc Truong [a], Thomas Gärtner [e], Daniel Merkle [fd], and Peter F. Stadler [bcghij]

[a] *School of Pharmacy, University of Medicine and Pharmacy at Ho Chi Minh City, Dinh Tien Hoang, Ho Chi Minh City, Vietnam.*
[b] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Ridebanevej 9, DK-1870, Frederiksberg, Denmark.*
[c] *Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics & School for Embedded and Composite Artificial Intelligence (SECAI), Leipzig University, Härtelstraße 16–18, Leipzig, Germany.*
[d] *Department of Mathematics and Computer Science, University of Southern Denmark, Odense M DK-5230, Denmark.*
[e] *Machine Learning Research Unit, TU Wien Informatics, Wien M A-1040, Austria.*
[f] *Algorithmic Cheminformatics Group, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany.*
[g] *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103, Leipzig, Germany.*
[h] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090, Vienna, Austria.*
[i] *Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia.*
[j] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, 87501, USA.*
[*] *Corresponding author*

Supplementary information

1 Methods details

1.1 Benchmarking metrics

We report Accuracy and Matthews correlation coefficient (MCC).

$$\mathrm{Acc} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{S1}$$

$$\mathrm{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{S2}$$

where TP, TN, FP, FN are counts (for multiclass use classwise counts TP_k, FP_k, FN_k).

1.2 Contamination

Finally, to quantify data leakage risk in the RXNFP pre-training corpus, here refer to the *contamination* term, we extracted the 2.8 million reactions used during pre-training and measured the overlap with each test split. We define the *contamination rate* as

$$\mathrm{Contamination} = \frac{\left|\mathrm{SMILES}_{\mathrm{test}} \cap \mathrm{SMILES}_{\mathrm{pretrain}}\right|}{\left|\mathrm{SMILES}_{\mathrm{test}}\right|} \times 100\%. \tag{S3}$$

2 Additional figures

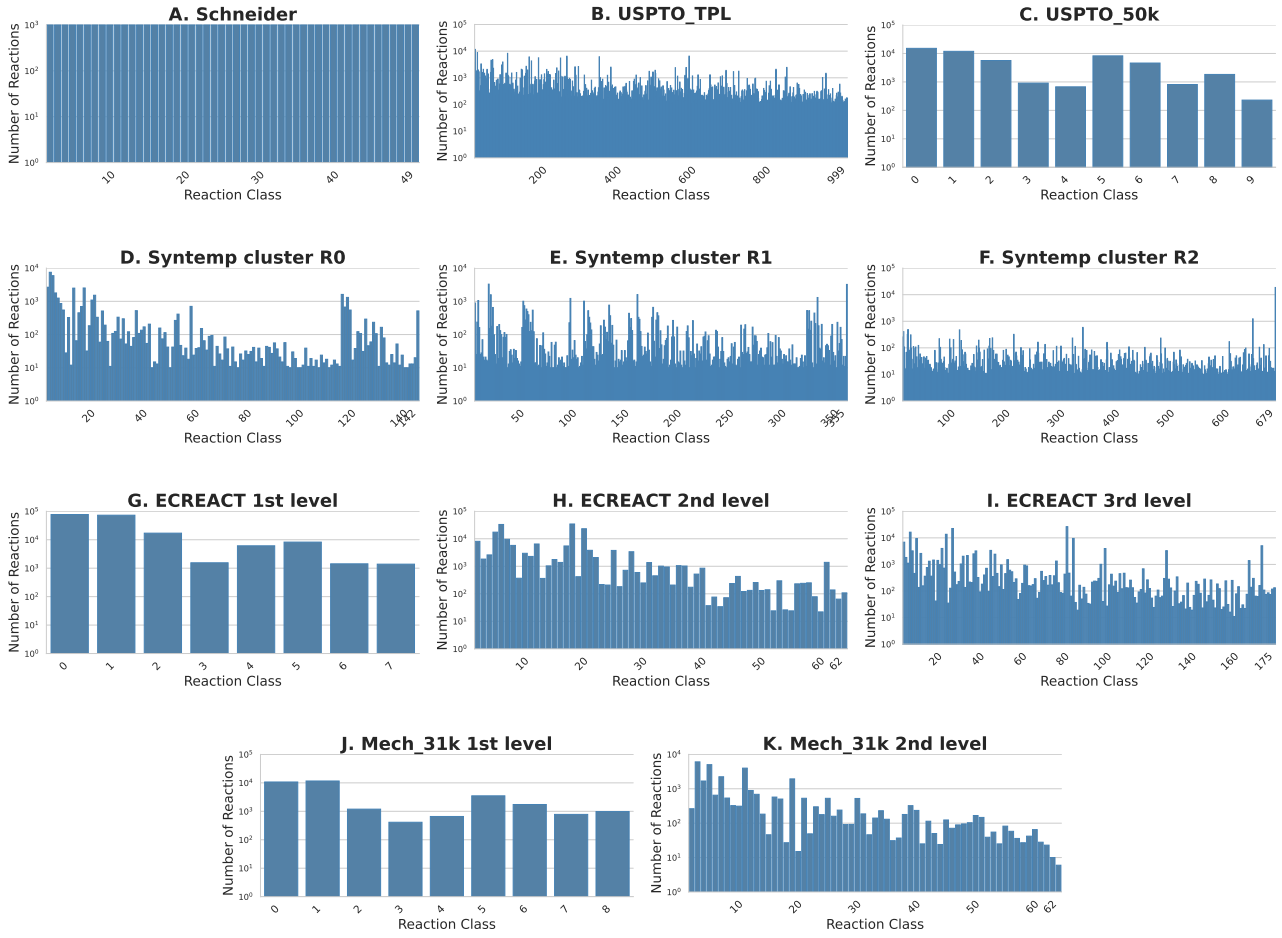

Fig. S1 Class distribution across datasets. Bars show the number of reactions per class; the y-axis is in base-10 logarithmic scale and the x-axis indicates class index (number of classes in each dataset). Only the `Schneider` dataset exhibits a *balanced class ratio*.

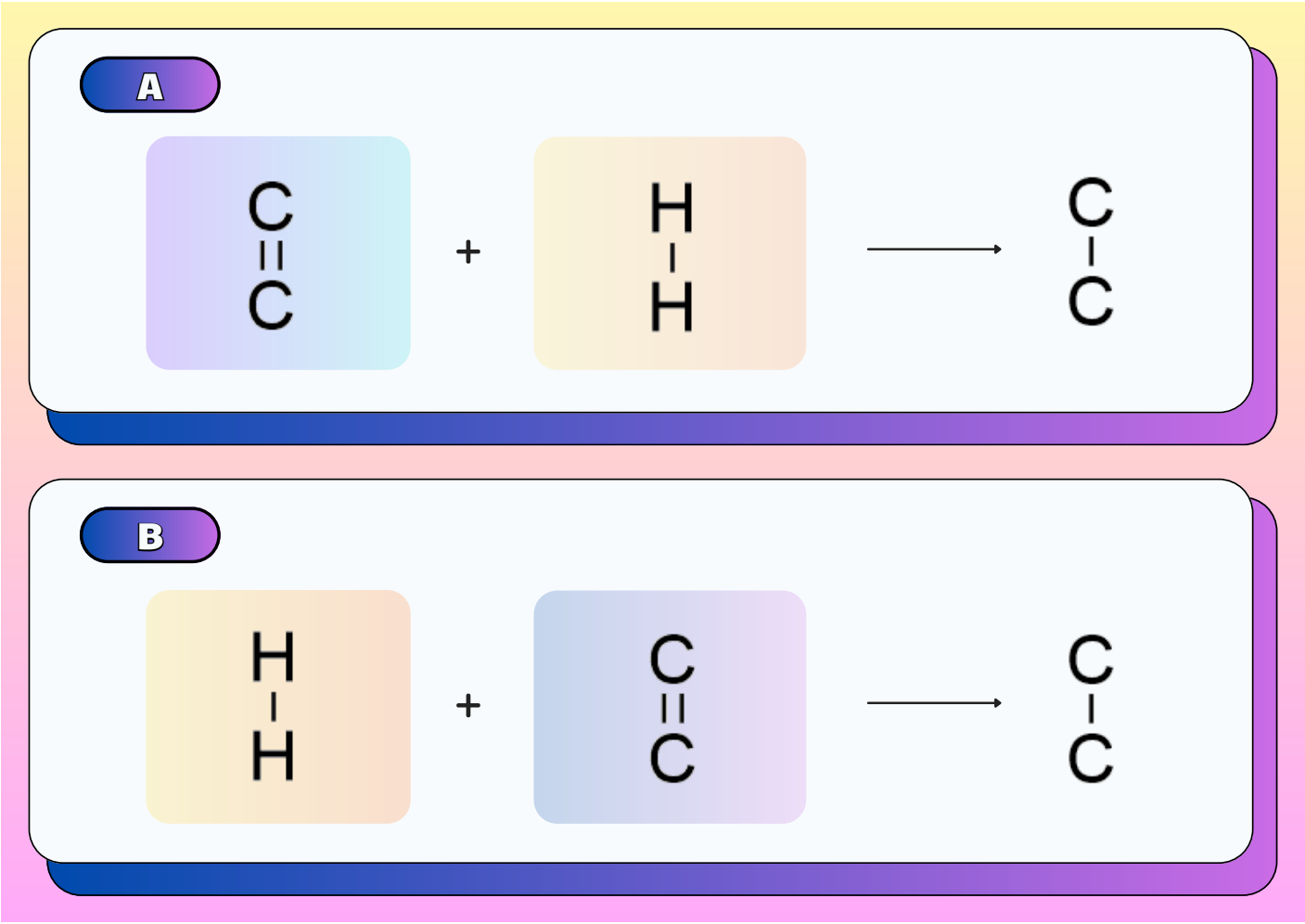

Fig. S2 Permutation-invariance test for reactant-side hydrogen positions.

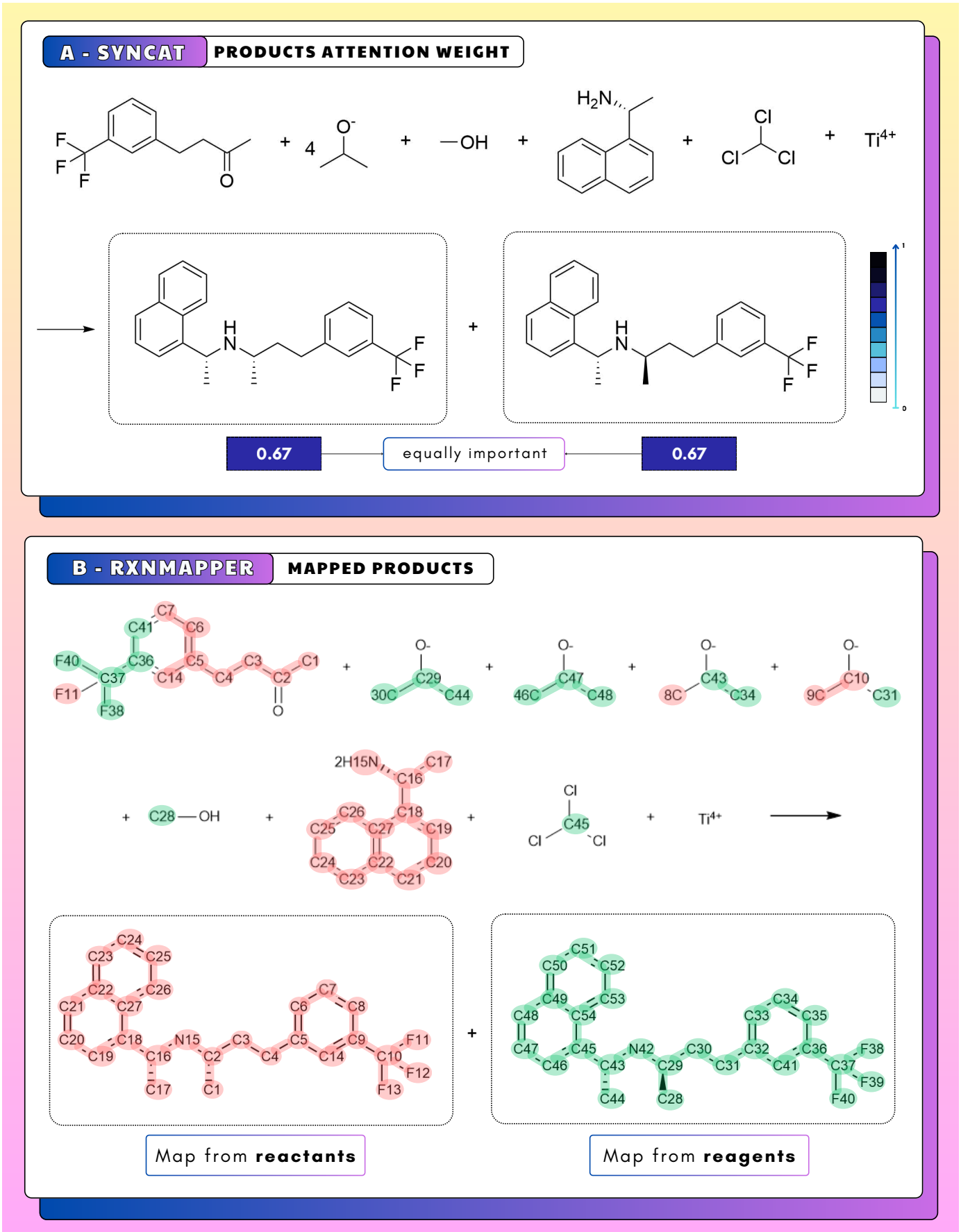

Fig. S3 `SynCat` **vs** `RXNMapper` **on a racemic reaction.** (A) `SynCat` assigns comparable attention to enantiomeric products in a racemic mixture, highlighting its potential to capture stereochemical outcomes as richer stereochemical encodings are incorporated. (B) `RXNMapper` produces erroneous atom mappings for the same reaction.

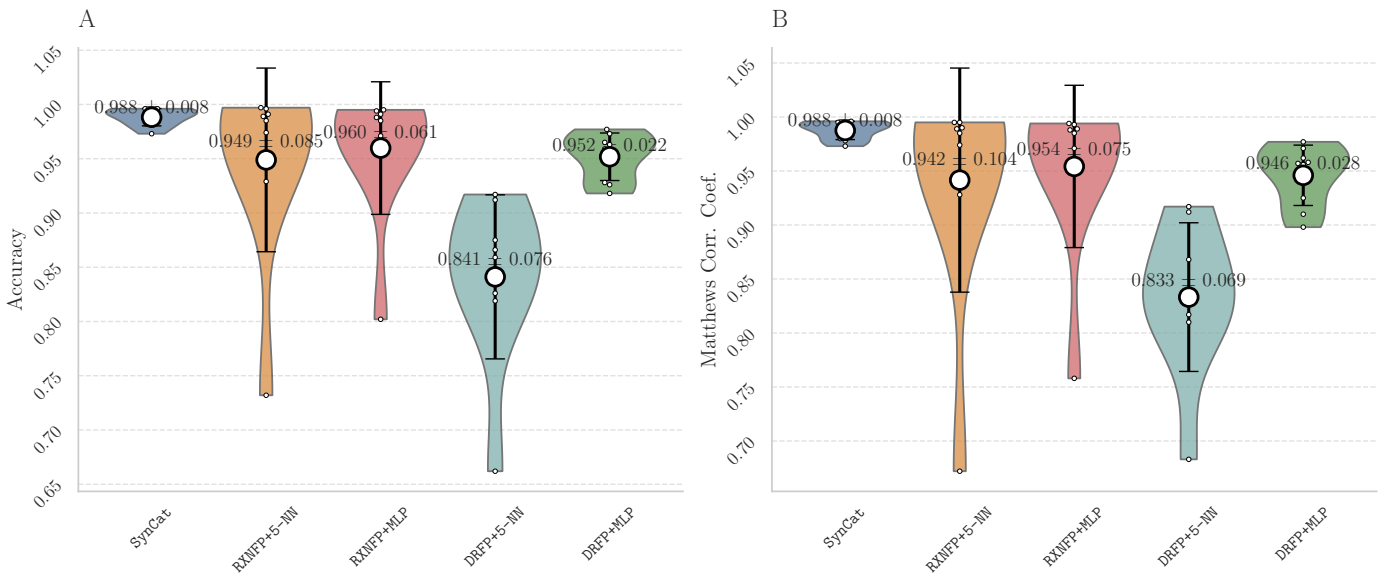

Fig. S4 The boxplot, segmented into panels (A) Accuracy and (B) Matthew's Correlation Coefficient (MCC), illustrates the performance of `SynCat` compared to benchmark methods across nine datasets.

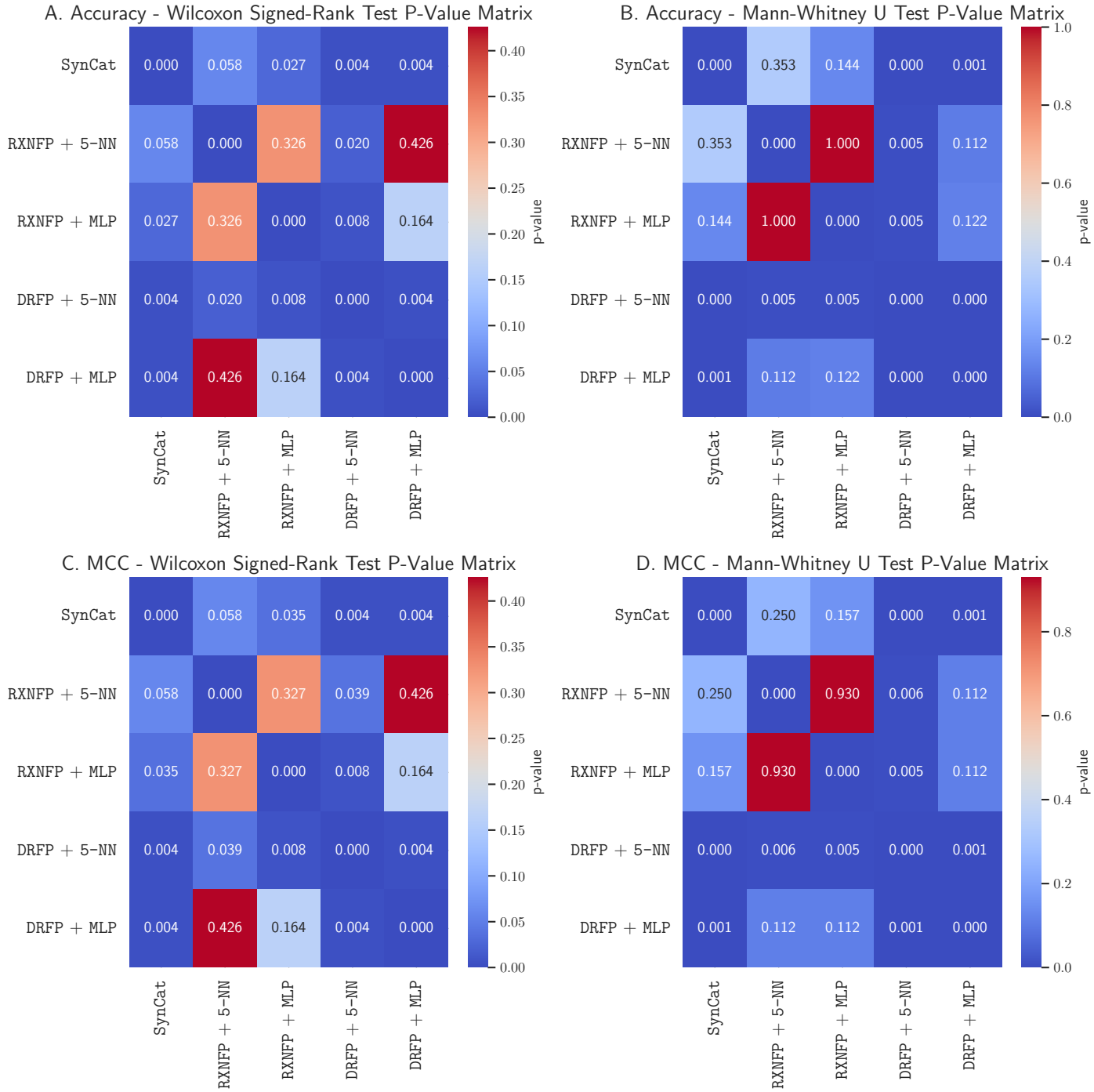

Fig. S5 Heatmaps summarizing pairwise statistical comparisons across methods. Panels (A) and (C) display the results of Wilcoxon signed-rank tests for Accuracy and MCC, respectively; Panels (B) and (D) show the results of Mann–Whitney U tests for Accuracy and MCC, respectively.

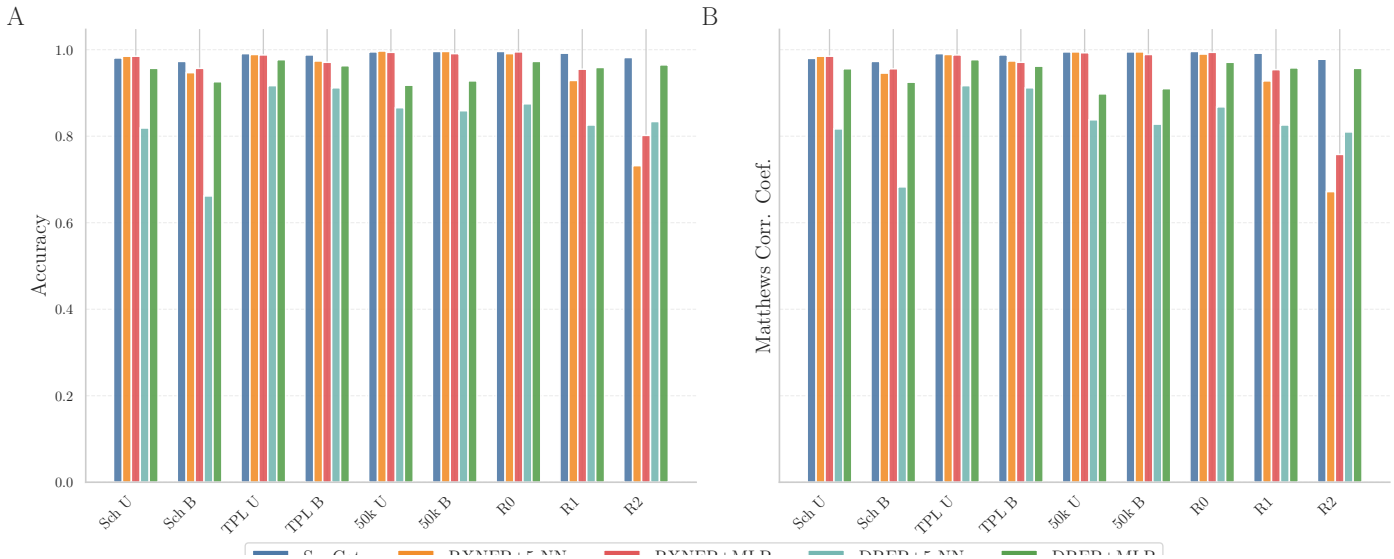

Fig. S6 Comparison of classification methods across nine datasets. (A) Mean *Accuracy*. (B) Mean *Matthews correlation coefficient* (MCC). Bars indicate method-wise performance aggregated across datasets

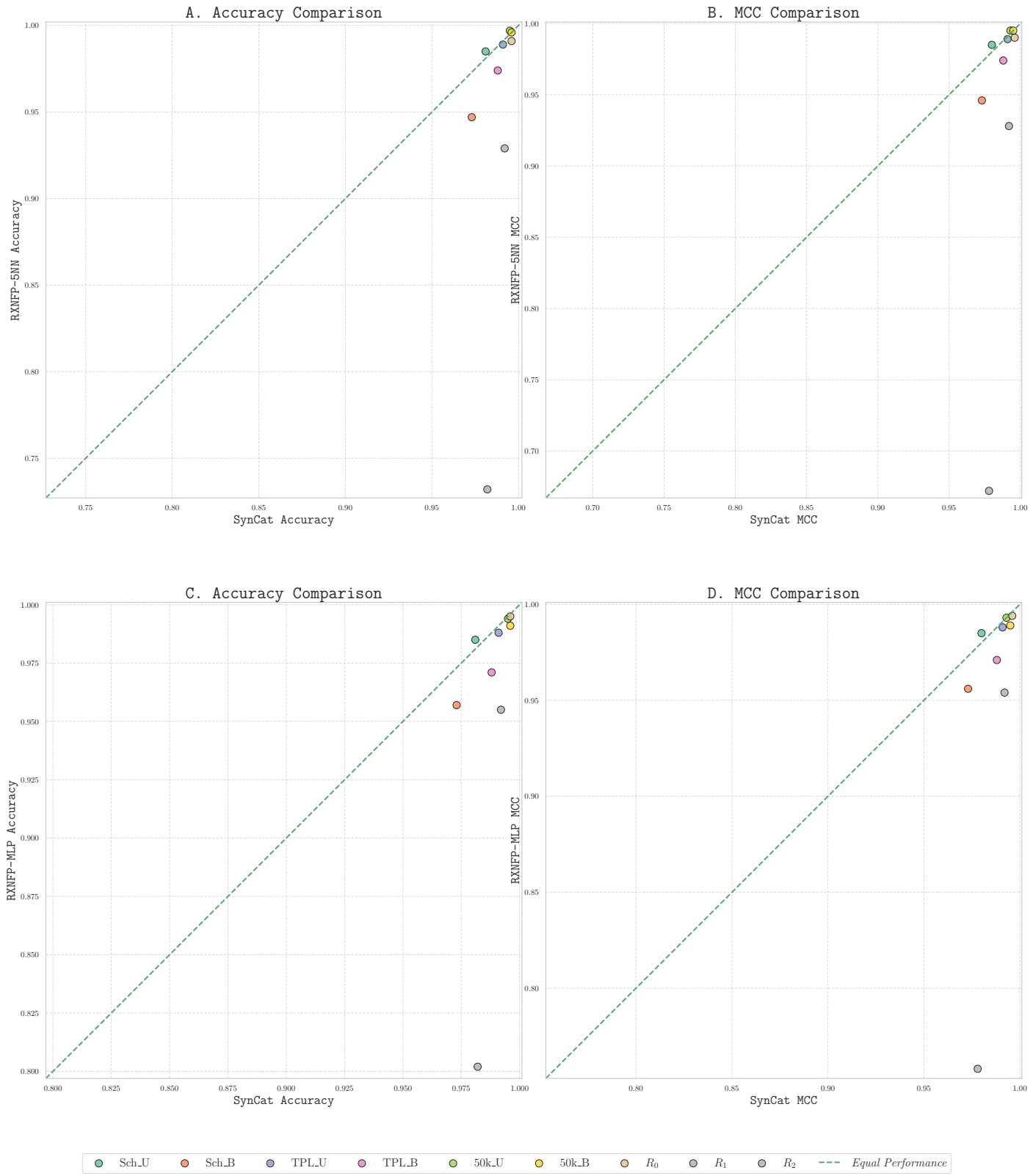

Fig. S7 Comparison of RXNFP performance against SynCat. Panels (A) and (B) show Accuracy and MCC, respectively, for the RXNFP+5-NN; Panels (C) and (D) show Accuracy and MCC for the RXNFP+MLP.

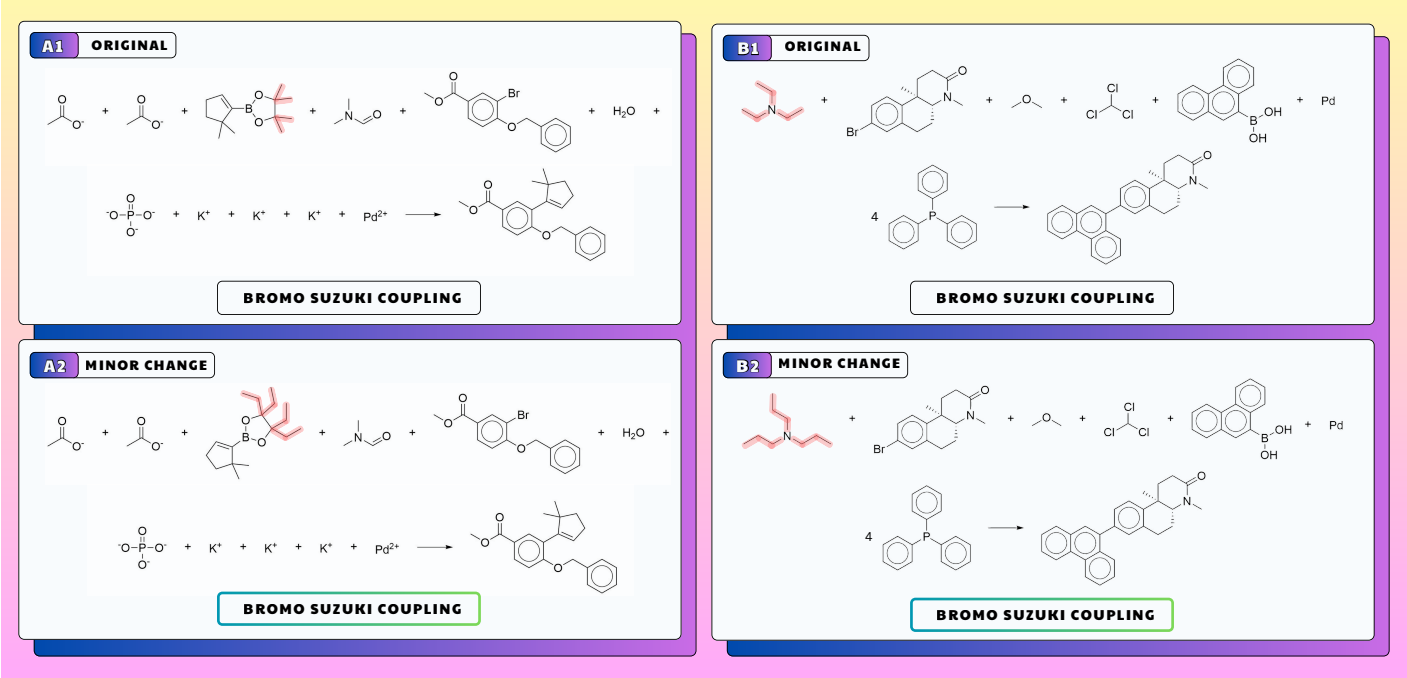

Fig. S8 (A1) Original Suzuki substrate containing Bpin–Me. **(A2)** Modified Suzuki substrate with Bpin–Et. **(B1)** Original reagent triethylamine (TEA). **(B2)** Substituted reagent tri-*n*-propylamine (TNPA)

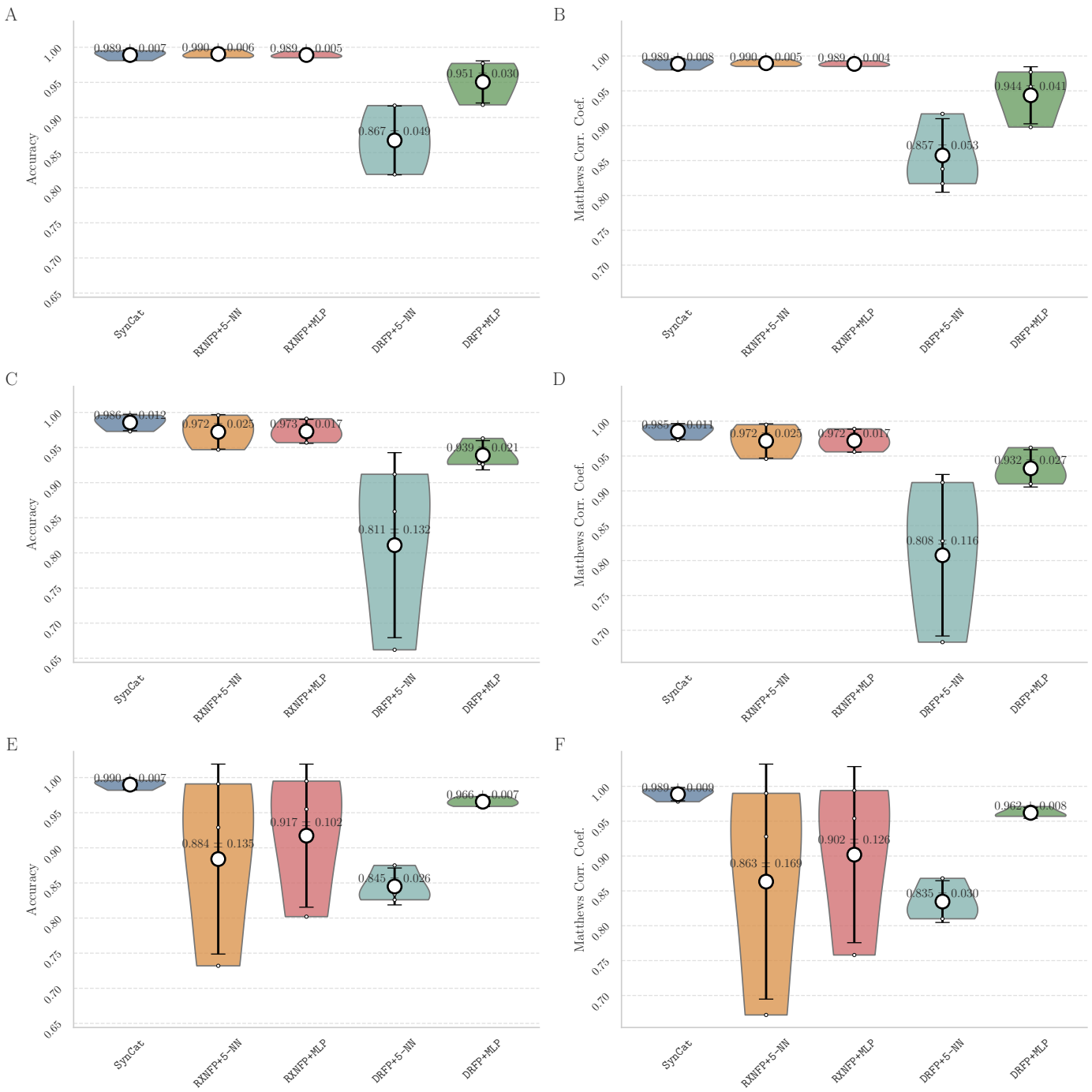

Fig. S9 Subgroup comparison of reaction classification performance. Panels (A) and (B) report Accuracy and MCC, respectively, on the unbalanced reaction dataset; Panels (C) and (D) report Accuracy and MCC on the balanced reaction dataset; and Panels (E) and (F) report Accuracy and MCC on the SynTemp-generated reaction templates.

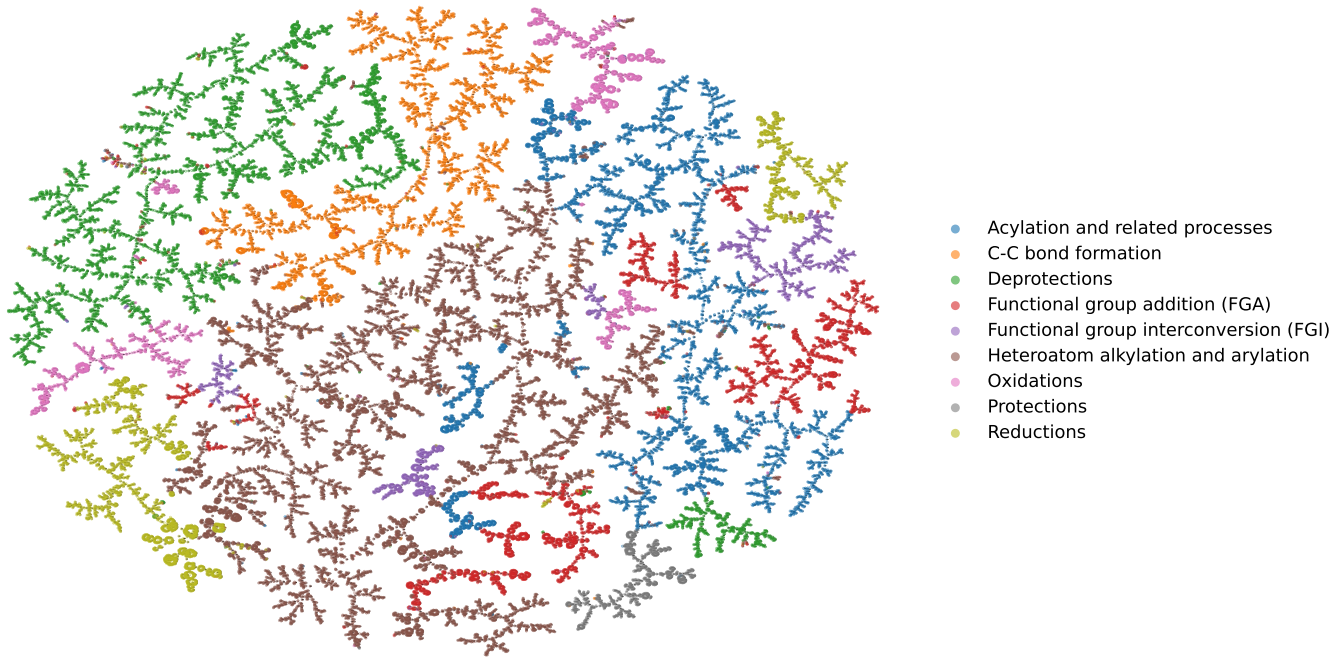

Fig. S10 TMAP based on `SynCat`'s embeddings for chemical reactions in `Schneider50k` dataset

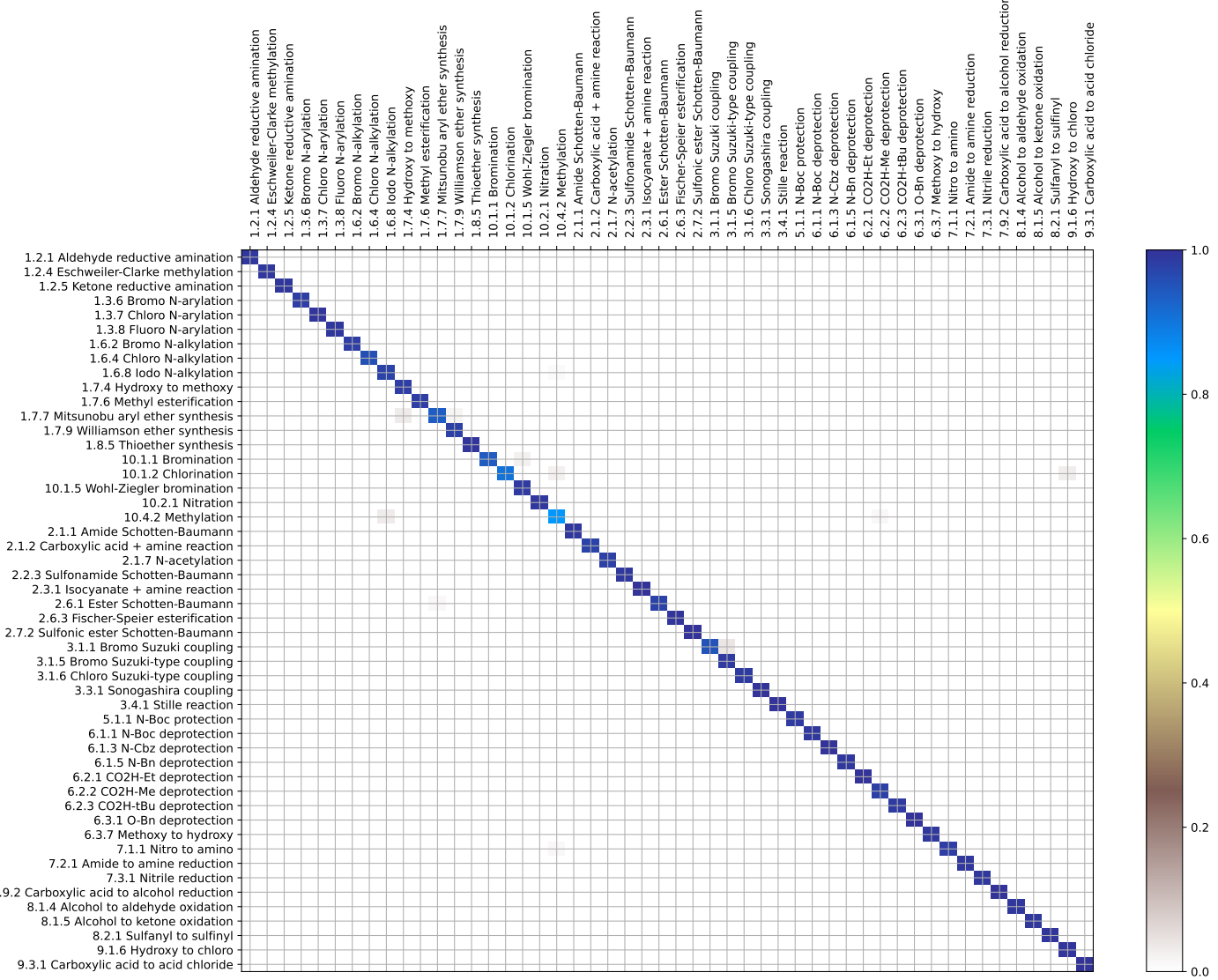

Fig. S11 The Confusion Matrix for `SynCat` on the `Schneider50k` dataset demonstrates exceptional performance across a set of chemical reaction classes. The strong diagonal dominance, where nearly all predicted labels align with true labels.

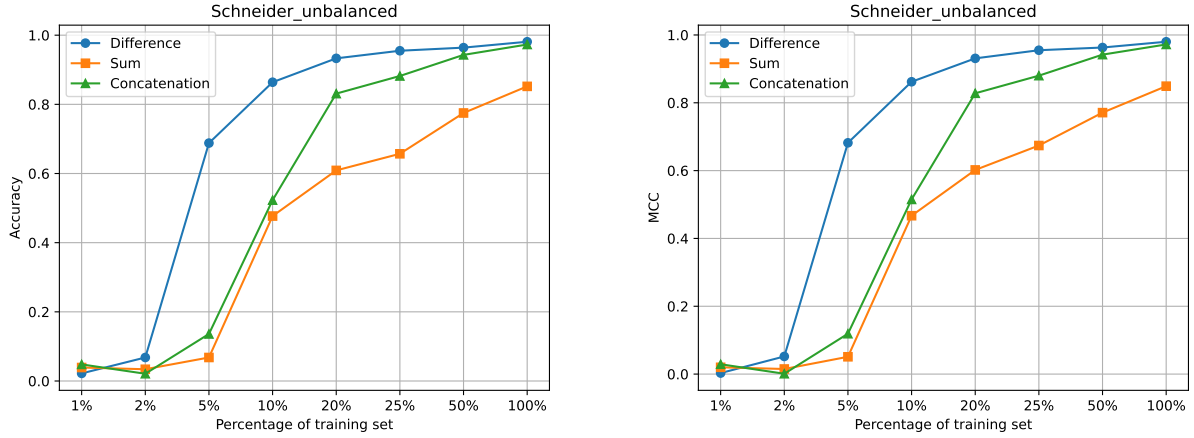

Fig. S12 Performance on the `Schneider`-Unbalanced dataset comparing three fusion operators including difference (Δ), sum (Σ), and concatenation (`Cat`), across increasing training-set fractions.

3 Additional tables

Table S1 Dataset partition summary

Dataset	Total size	Split ratio (Train/Val/Test)	Classes	Stratification keys	Random seed
Schneider-Unbalance	50 000	9:1:40	50	reaction class	42
Schneider-Balance	50 000	9:1:40	50	reaction class	42
USPTO_TPL-Unbalance	445 115	8:1:1	1000	reaction class	42
USPTO_TPL-Balance	445 115	8:1:1	1000	reaction class	42
USPTO_50k-Unbalance	50 016	8:1:1	10	reaction class	42
USPTO_50k-Balance	50 016	8:1:1	10	reaction class	42
SynTemp_R0	43 441	8:1:1	143	reaction class	42
SynTemp_R1	43 441	8:1:1	356	reaction class	42
SynTemp_R2	43 441	8:1:1	680	reaction class	42
ECREACT_1st	185 734	8:1:1	7	reaction class	42
ECREACT_2nd	185 734	8:1:1	63	reaction class	42
ECREACT_3rd	185 734	8:1:1	175	reaction class	42
Mech_31k_1st	31 673	8:1:1	9	reaction class	42
Mech_31k_2nd	31 673	8:1:1	63	reaction class	42

Table S2 Detailed descriptions of atomic features.

Feature	Possible Chemical Values	# of Features
Atom type τ	C, O, N, S, P, F, Cl, Br, I, H, ...	118
Formal charge q	$\ldots,-2,-1,0$ (neutral), $+1,+2,\ldots$	4
Degree d	$0,1,2,3,4,5,6,\ldots$	6
Hybridization hyb	$\mathrm{sp}, \mathrm{sp}^2, \mathrm{sp}^3, \mathrm{sp}^3\mathrm{d}, \mathrm{sp}^3\mathrm{d}^2$	5
H-count n_H	$0,1,2,3,4,\ldots$	4
Valence val	$1,2,3,4,5,6,\ldots$	6
Donor/acceptor DA	Donor, Acceptor	2
Chirality χ	R, S	2
Ring size r_s	$3,4,5,6,7,8$	6
Aromaticity aro	0 (non-aromatic), 1 (aromatic)	2

Table S3 Detailed descriptions of bond features.

Feature	Possible Chemical Values	# of Features
Bond type b_t	single, double, triple, aromatic, selfloop	5
Stereochemistry st	cis, trans	2
is_conjugated_ring icr	is_ring, is_conjugated	2

Table S4 Configuration of SynCat

Dataset	Number of GINE layers	Dimension of embedding
Schneider-Unbalance	2	256
Schneider-Balanced	2	256
USPTO_TPL-Unbalanced	3	384
USPTO_TPL-Balanced	3	384
USPTO_50k-Unbalance	2	256
USPTO_50k-Balanced	2	256
SynTemp_R0	2	256
SynTemp_R3	2	256
SynTemp_R2	2	256
ECREACT_1st	2	256
ECREACT_2nd	2	256
ECREACT_3rd	2	256
Mech_31k_1st	2	256
Mech_31k_2nd	2	256

Table S5 Contamination analysis of RXNFP pre-training model

Dataset	Contamination %
Schneider-Unbalance	72.58%
Schneider-Balance	6.23%
USPTO_TPL-Unbalance	99.97%
USPTO_TPL-Balance	14.70%
USPTO_50k-Unbalance	12.95%
USPTO_50k-Balance	0.32%
SynTemp_R0	0.39%
SynTemp_R1	0.39%
SynTemp_R2	0.44%

Table S6 Performance across training-set fractions, comparing three fusion operators (Δ, Σ, and Cat) on the Schneider-Unbalance dataset.

				Accuracy ↑				
Percentage	**1%**	**2%**	**5%**	**10%**	**20%**	**25%**	**50%**	**100%**
difference (Δ)	0.022	**0.068**	**0.688**	**0.864**	**0.933**	**0.955**	**0.964**	**0.981**
sum (Σ)	0.039	0.034	0.068	0.477	0.609	0.680	0.775	0.852
concatenation (Cat)	**0.048**	0.021	0.136	0.523	0.831	0.882	0.943	0.973
				MCC ↑				
Percentage	**1%**	**2%**	**5%**	**10%**	**20%**	**25%**	**50%**	**100%**
difference (Δ)	0.003	**0.052**	**0.682**	**0.862**	**0.931**	**0.955**	**0.963**	**0.980**
sum (Σ)	0.020	0.015	0.051	0.467	0.602	0.674	0.771	0.849
concatenation (Cat)	**0.029**	0.001	0.119	0.515	0.828	0.880	0.942	0.972