SDU
University of Southern Denmark

TACsy

Training Alliance for Computational systems chemistry

UNIVERSITÄT LEIPZIG

SynCat: A light-weight model for reaction classification

Winterseminar-Bled

Presenter: Tieu-Long Phan

Date: 10.02.2025

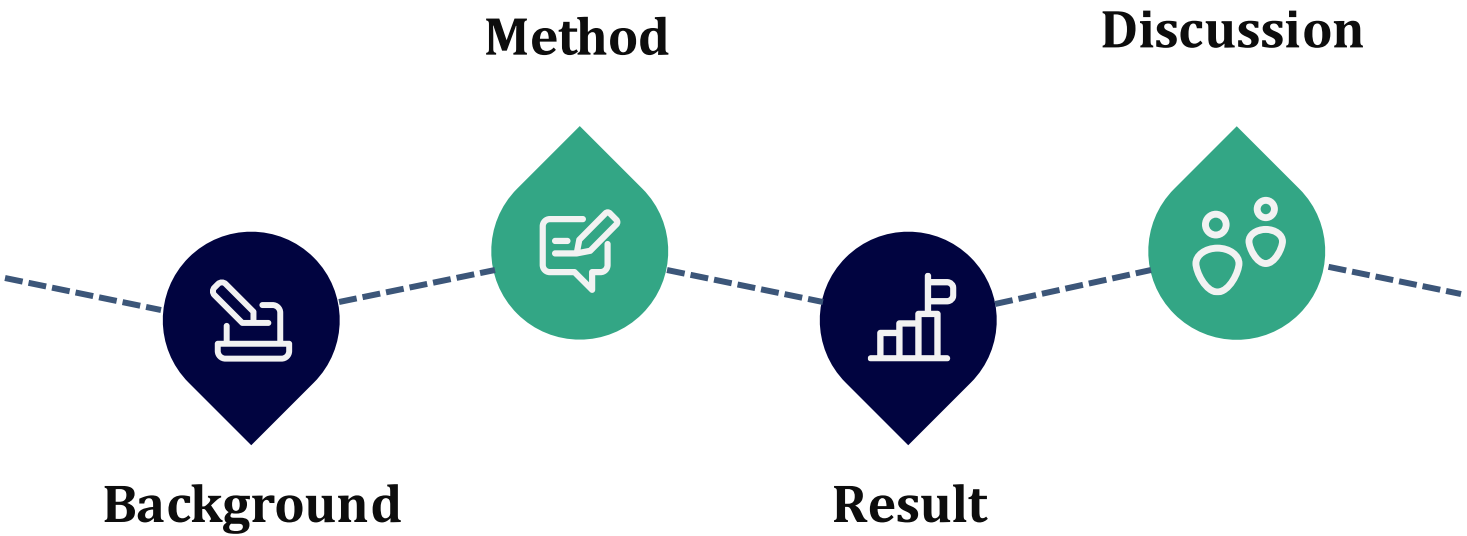
MARIE CURIE ACTIONS

Founded by the European Union

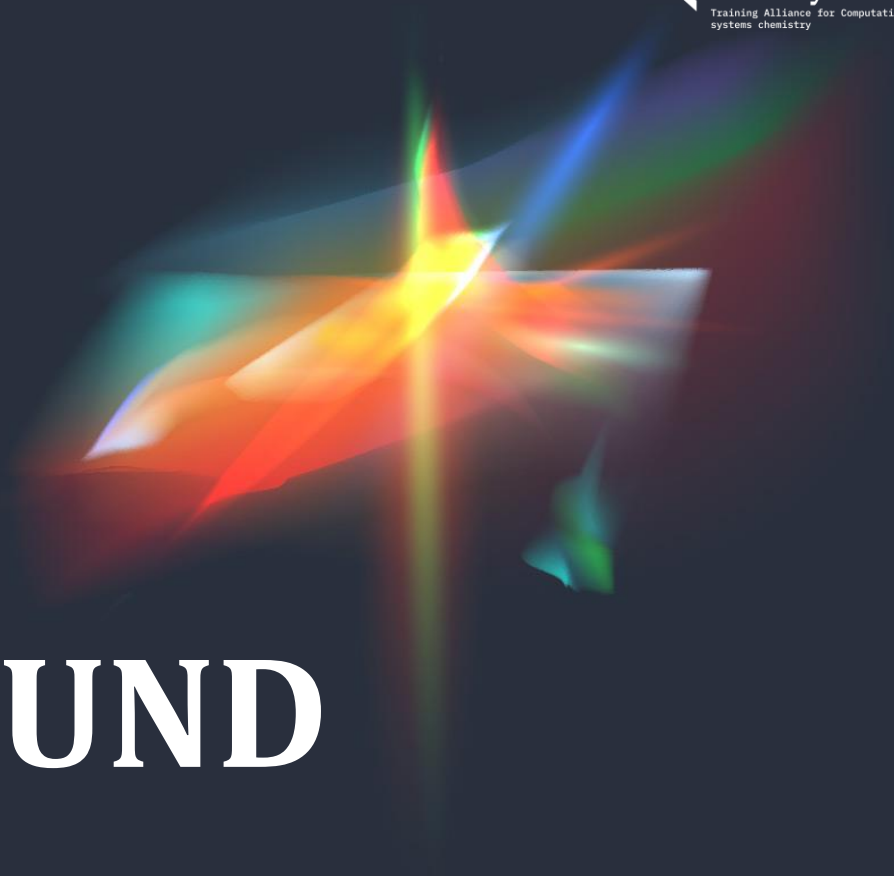
This project has received funding from the European Unions Horizon 2021 research and innovation programme under the Marie-Skłodowska-Curie grant agreement No 101072930

OUTLINE

TACsy
Training Alliance for Computational systems chemistry



TACsy
Training Alliance for Computational systems chemistry

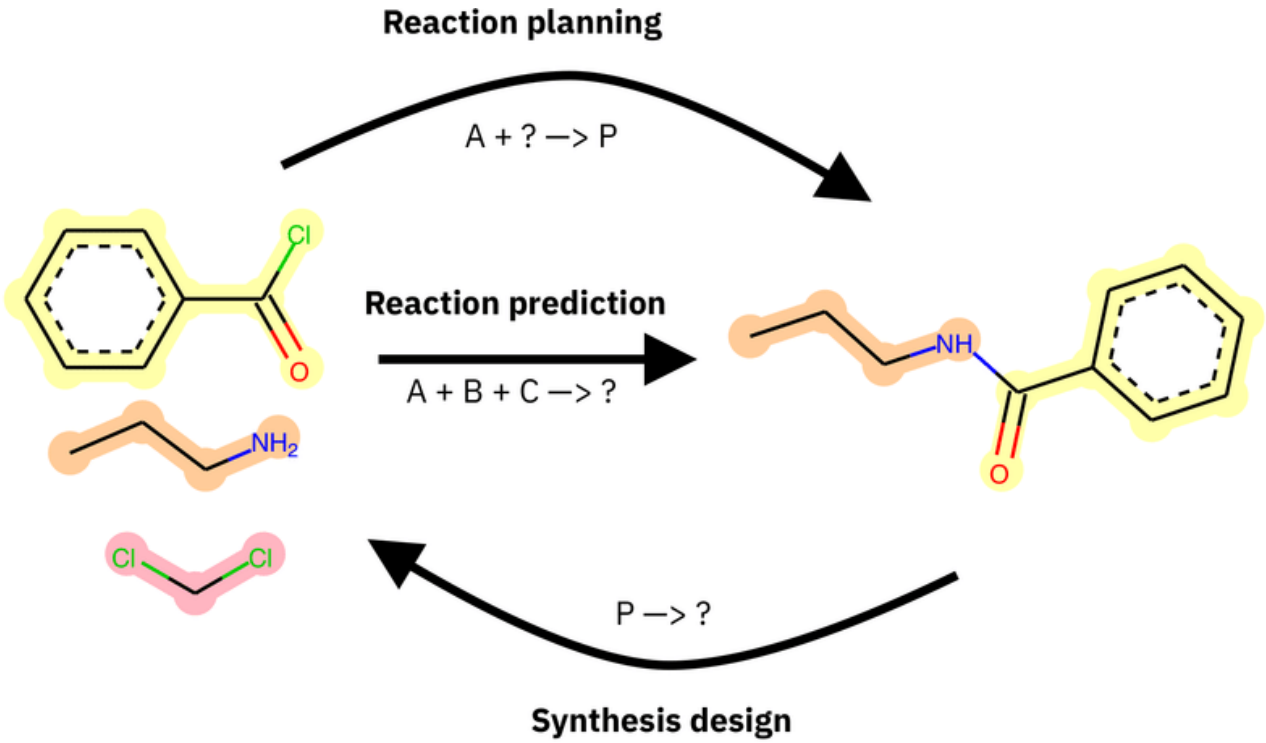


01

BACKGROUND

BACKGROUND

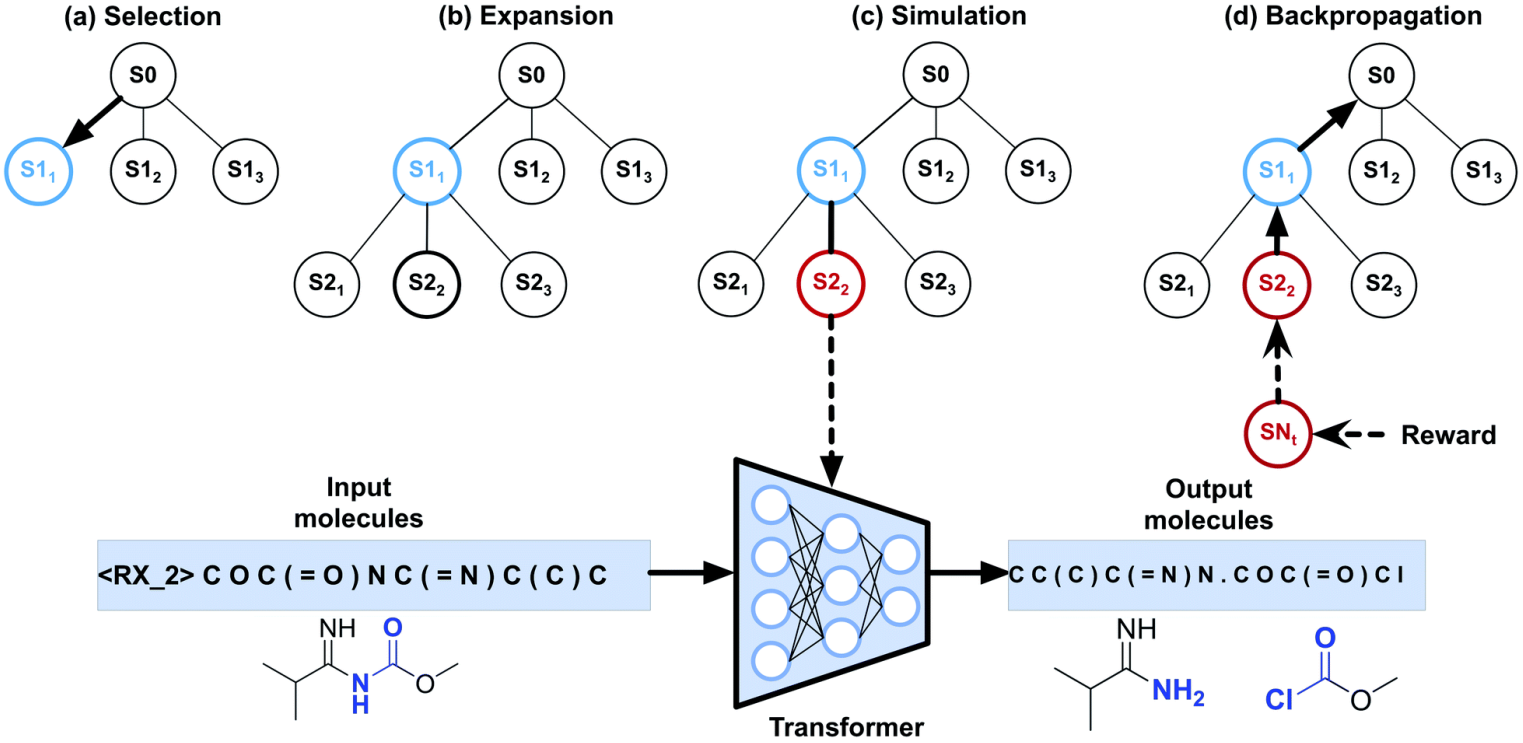
Synthesis Planning



Griffiths, R. R., et al. (2021). Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design. arXiv preprint arXiv:2105.02637.

BACKGROUND

Template-free synthesis planning

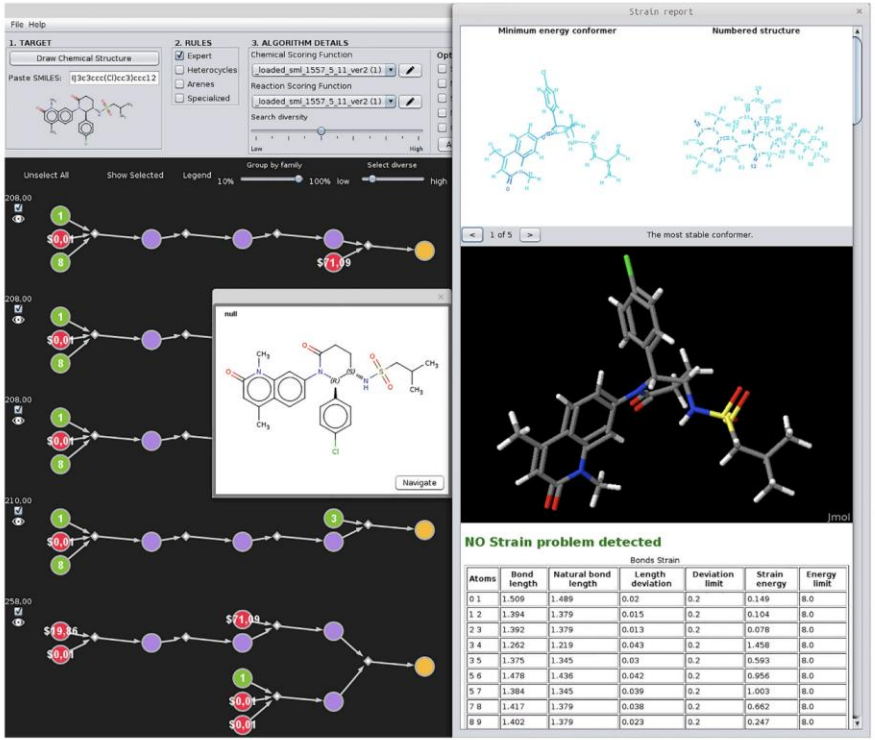


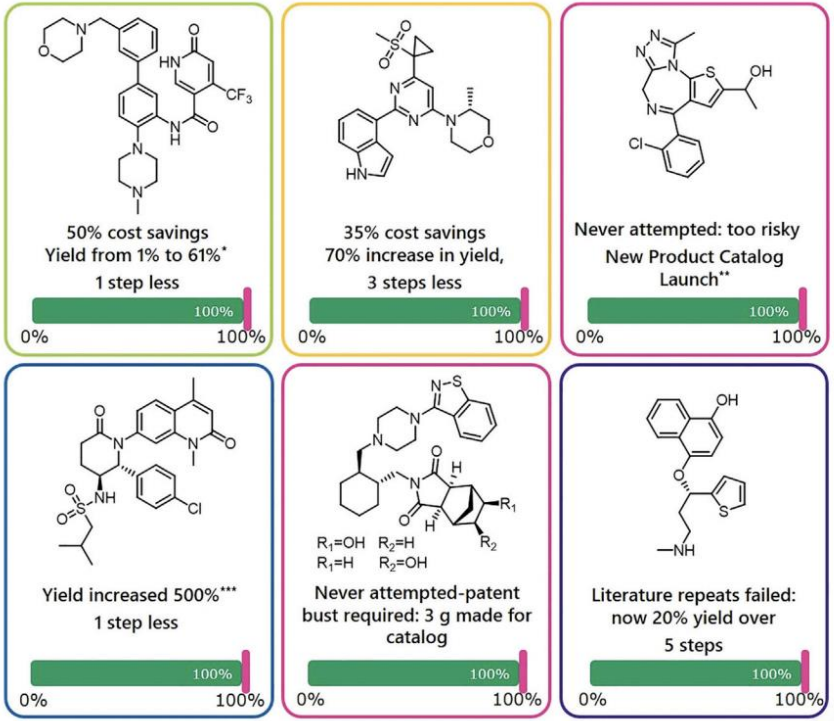
Lin, K., Xu, Y., Pei, J., & Lai, L. (2020). Automatic retrosynthetic route planning using template-free models. Chemical science, 11(12), 3355-3364.

BACKGROUND

TACsy
Training Alliance for Computational systems chemistry

Template-based synthesis planning

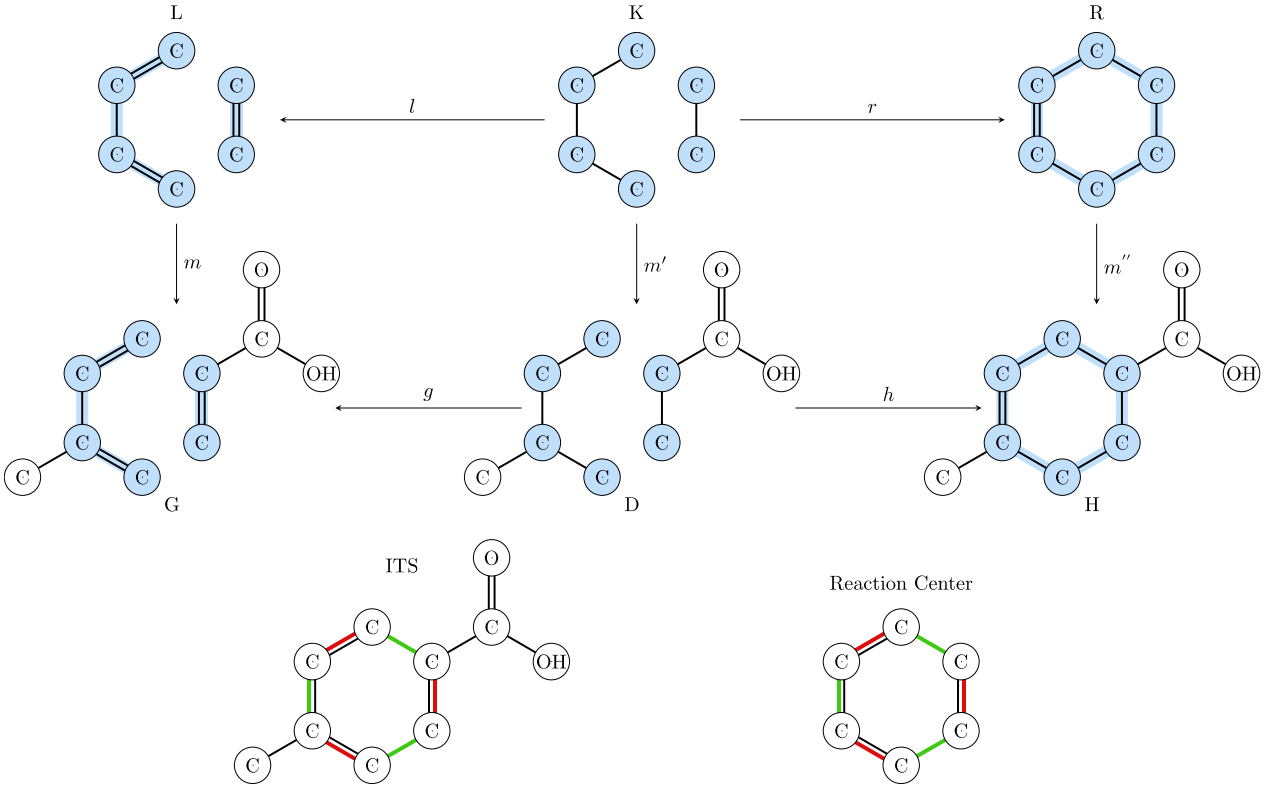




Grzybowski, B. A., Szymkuć, S., Gajewska, E. P., Molga, K., Dittwald, P., Wołos, A., & Klucznik, T. (2018). Chematica: a story of computer code that started to think like a chemist. Chem, 4(3), 390-398.

BACKGROUND

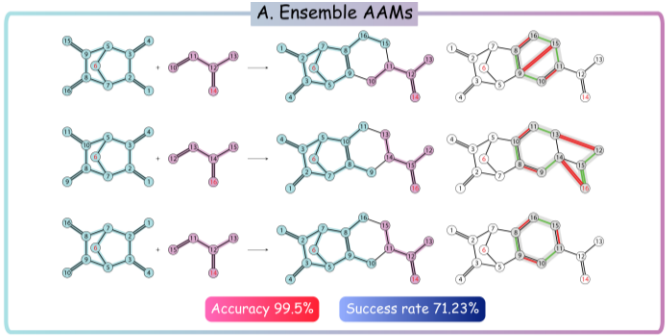
Double Pushout Rules

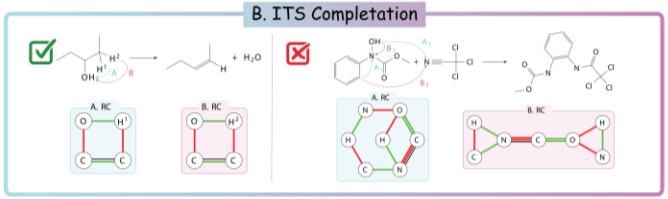


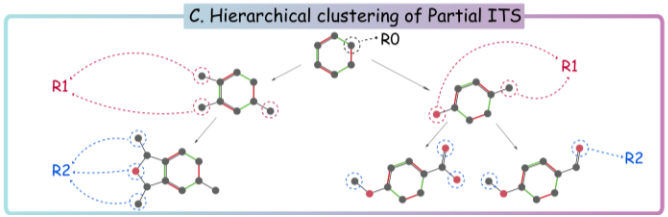
Phan, T. L., Weinbauer, K., Laffitte, M. E. G.,...,& Stadler, P. F. (2024). SynTemp: Efficient Extraction of Graph-Based Reaction Rules from Large-Scale Reaction Databases.

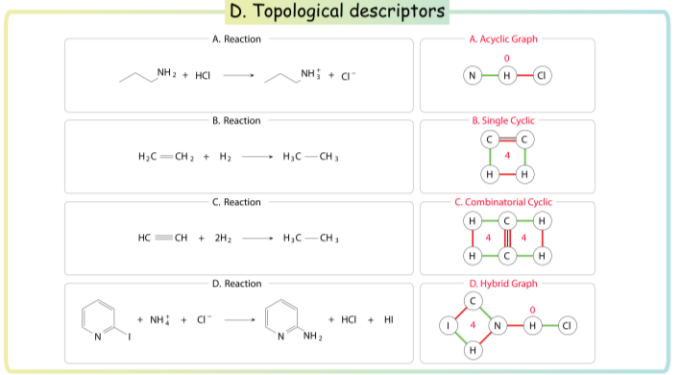
BACKGROUND

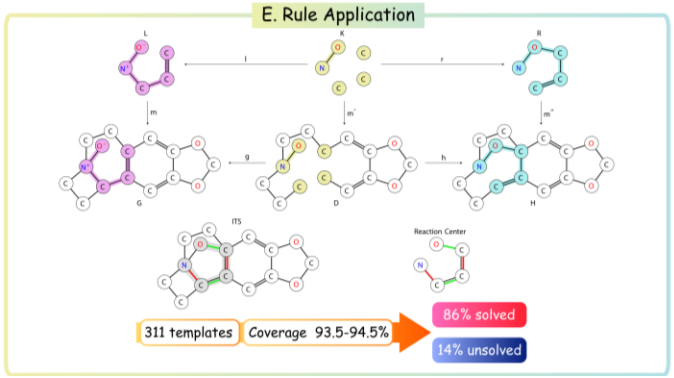
SynTemp









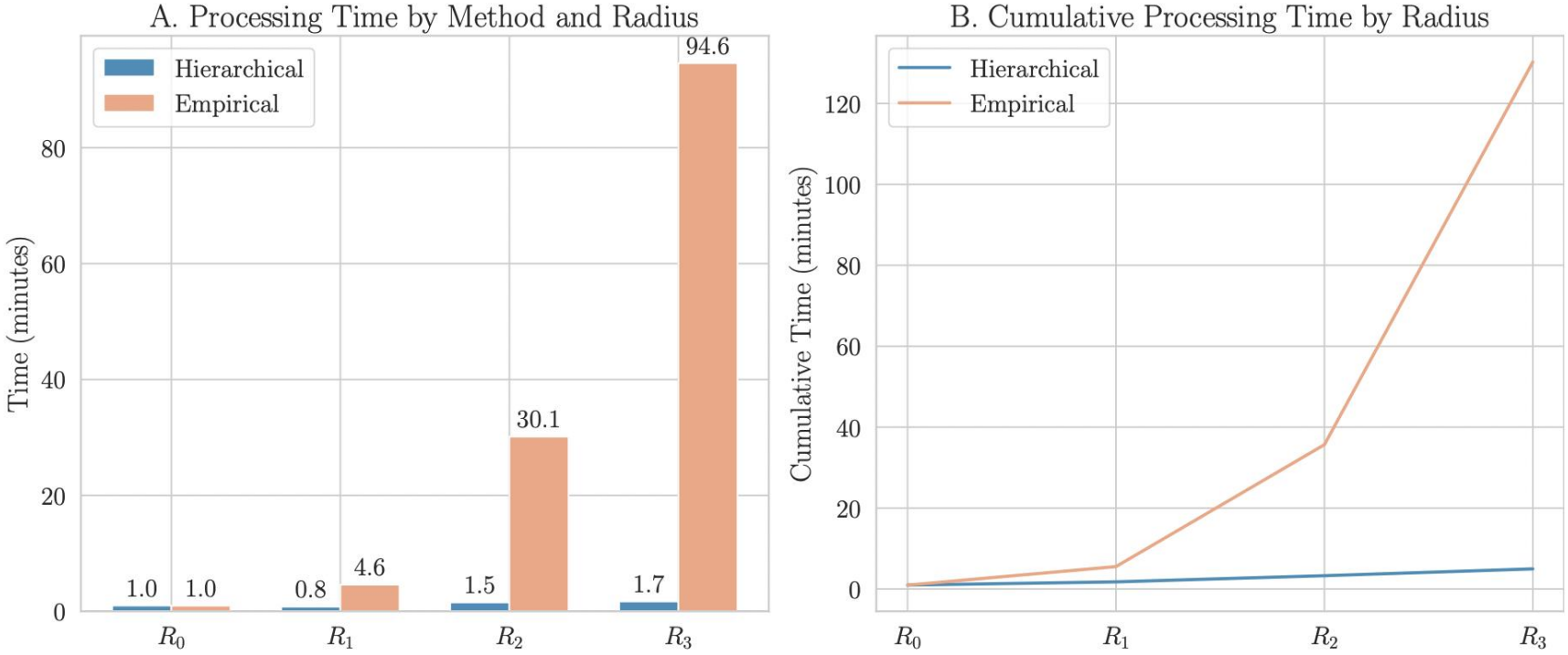


Phan, T. L., Weinbauer, K., Laffitte, M. E. G.,...,& Stadler, P. F. (2024). SynTemp: Efficient Extraction of Graph-Based Reaction Rules from Large-Scale Reaction Databases.

BACKGROUND

TACsy
Training Alliance for Computational systems chemistry

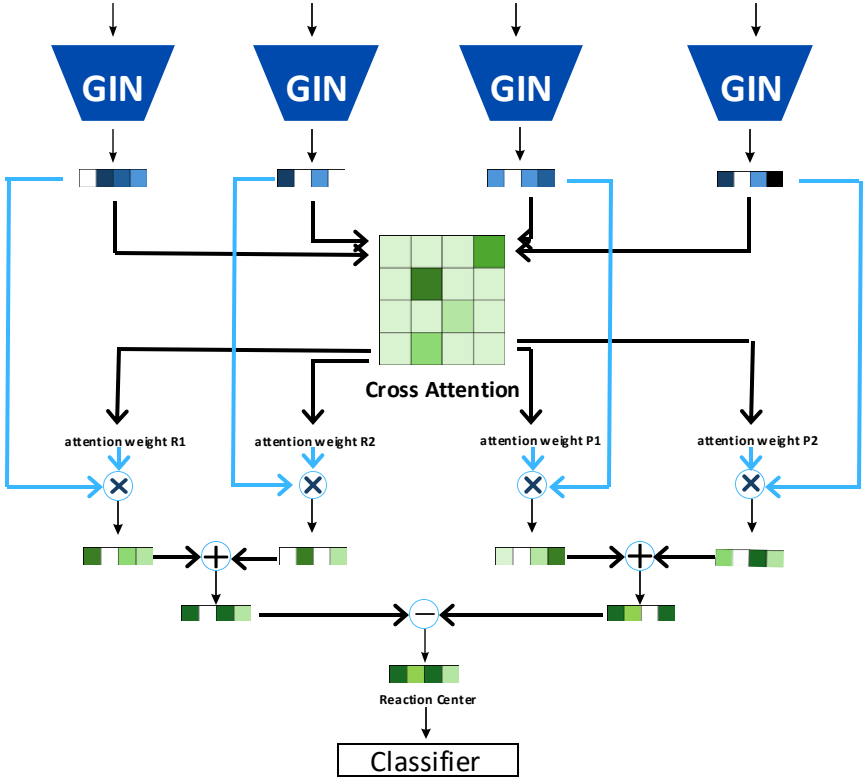
SynTemp - Challenge



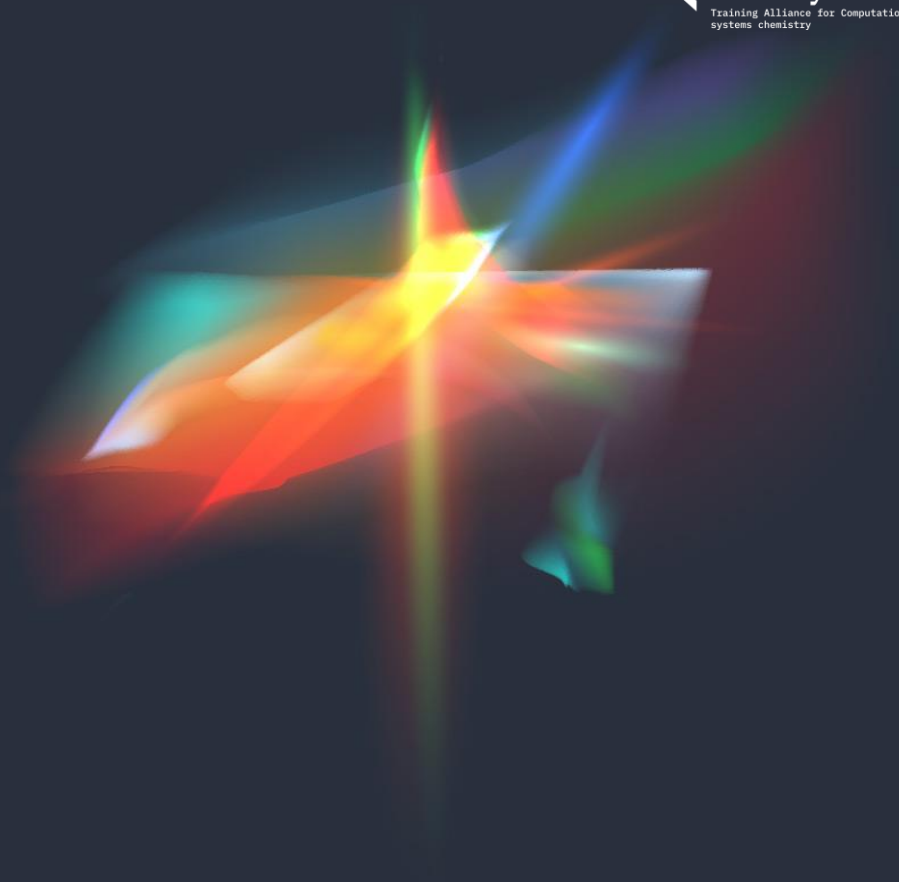
Phan, T. L., Weinbauer, K., Laffitte, M. E. G.,…,& Stadler, P. F. (2024). SynTemp: Efficient Extraction of Graph-Based Reaction Rules from Large-Scale Reaction Databases.

BACKGROUND

SynCat



TACsy
Training Alliance for Computational systems chemistry



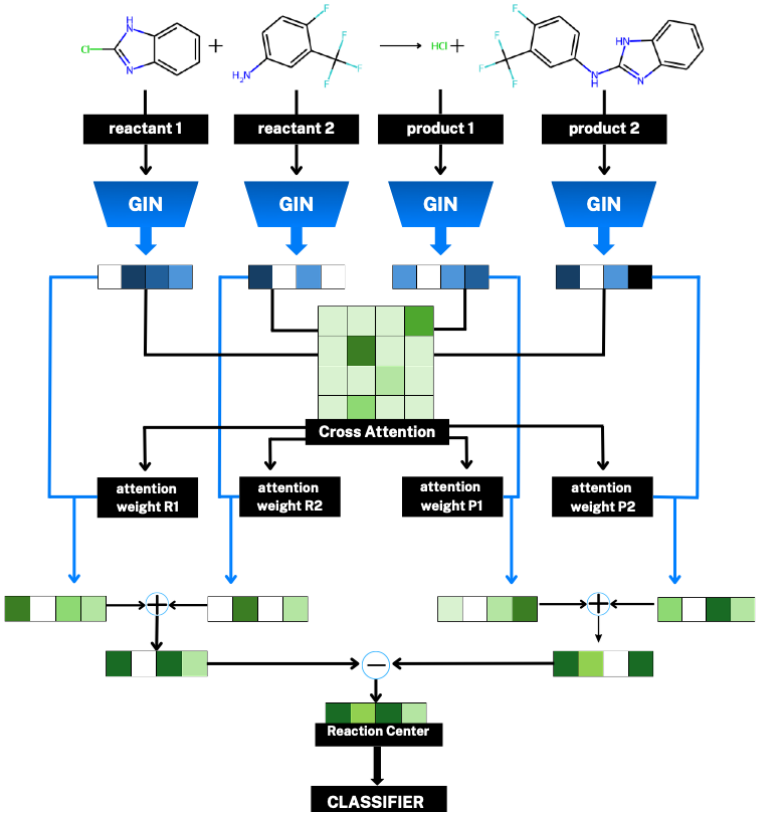
02

METHOD

METHOD

TACsy
Training Alliance for Computational systems chemistry

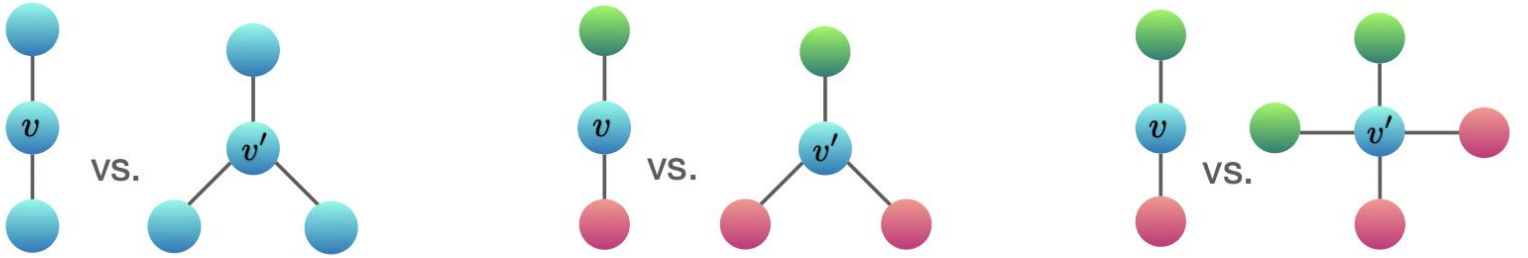
MODEL ARCHITECTURE



METHOD

TACsy
Training Alliance for Computational systems chemistry

Graph isomorphism network



(a) Mean and Max both fail (b) Max fails (c) Mean and Max both fail

We utilize the Graph Isomorphism Network (GIN) to enhance graph representation accuracy, employing learnable transformations:

$$\mathbf{h}_v^{(0)} = \phi_n(\mathbf{v}), \quad \mathbf{h}_e^{(k)} = \phi_e(\mathbf{e}^{(k)}), \tag{1}$$

METHOD

TACsy
Training Alliance for Computational systems chemistry

Graph isomorphism network

Vertex embeddings update through layers, integrating local and neighboring data:

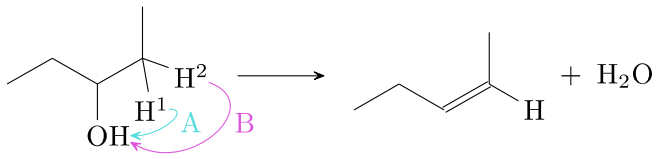
$$\mathbf{h}_v^{j,(l)} = \psi^{(l)} \left(\mathbf{h}_v^{j,(l-1)} + \sum_{k|\mathbf{e}^{j,k} \in E} \mathrm{ReLU}(\mathbf{h}_v^{j,(l-1)} + \mathbf{h}_e^{j,k}) \right), \qquad (2)$$

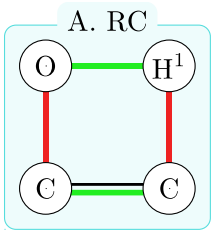
culminating in a comprehensive graph embedding:

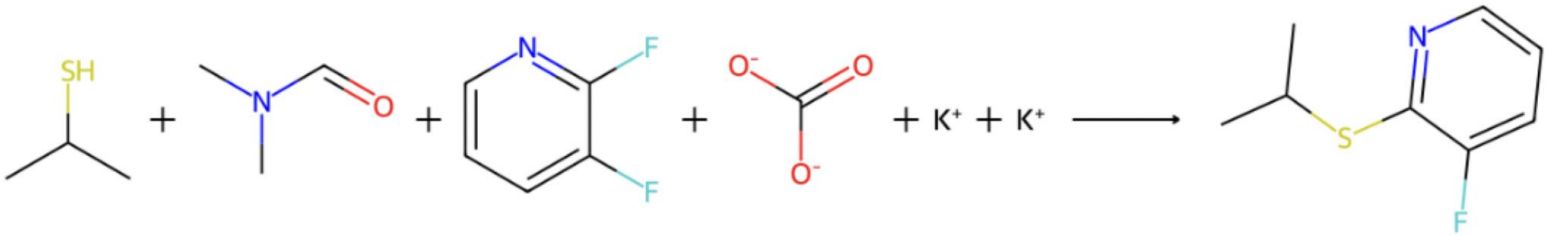
$$\mathbf{h}_g = \sum_{j|\mathbf{v}^j \in V} \mathbf{h}_v^{j,(L)}. \qquad (3)$$

METHOD

Cross-attention







METHOD

Cross-attention

TACsy
Training Alliance for Computational systems chemistry

The reaction center, Γ, is calculated by subtracting the aggregated product embeddings from the reactant embeddings:

$$\Gamma = \mathcal{R} - \mathcal{P}$$

where $\mathcal{R}$ and $\mathcal{P}$ denote the embeddings of reactants and products, respectively. This difference highlights the differential characteristics crucial for reaction classification.

METHOD

Cross-attention

TACsy
Training Alliance for Computational systems chemistry

To address noise from redundant substances, we implement cross-attention mechanisms that assign significance to compounds based on their contribution to the reaction center. The matrices $\mathcal{R}_{\text{org}}$ and $\mathcal{P}_{\text{org}}$ represent original reactant and product embeddings, from which we derive:

$$\mathcal{R}_{\text{updated}} = \begin{bmatrix} \mathcal{R}_{\text{org}} \\ S_{\mathcal{R}_{i,j}} \end{bmatrix}, \quad \mathcal{P}_{\text{updated}} = \begin{bmatrix} \mathcal{P}_{\text{org}} \\ S_{\mathcal{P}_{i,j}} \end{bmatrix}$$

Pairwise sums are integrated to enhance embeddings.

METHOD

Cross-attention

TACsy
Training Alliance for Computational systems chemistry

To address noise from redundant substances, we implement cross-attention mechanisms that assign significance to compounds based on their contribution to the reaction center. The matrices $\mathcal{R}_{\text{org}}$ and $\mathcal{P}_{\text{org}}$ represent original reactant and product embeddings, from which we derive:

$$\mathcal{R}_{\text{updated}} = \begin{bmatrix} \mathcal{R}_{\text{org}} \\ S_{\mathcal{R}_{i,j}} \end{bmatrix}, \quad \mathcal{P}_{\text{updated}} = \begin{bmatrix} \mathcal{P}_{\text{org}} \\ S_{\mathcal{P}_{i,j}} \end{bmatrix}$$

Pairwise sums are integrated to enhance embeddings.

METHOD

Cross-attention

TACsy
Training Alliance for Computational systems chemistry

Updated embeddings are processed to generate query and key vectors for attention calculations:

$$Q_{\mathcal{R}} = \mathcal{R}_{\text{updated}} \cdot W_q, \quad K_{\mathcal{R}} = \mathcal{R}_{\text{updated}} \cdot W_k$$

$$Q_{\mathcal{P}} = \mathcal{P}_{\text{updated}} \cdot W'_q, \quad K_{\mathcal{P}} = \mathcal{P}_{\text{updated}} \cdot W'_k$$

Attention weights are computed to emphasize the contributions of specific compounds.

METHOD

Cross-attention

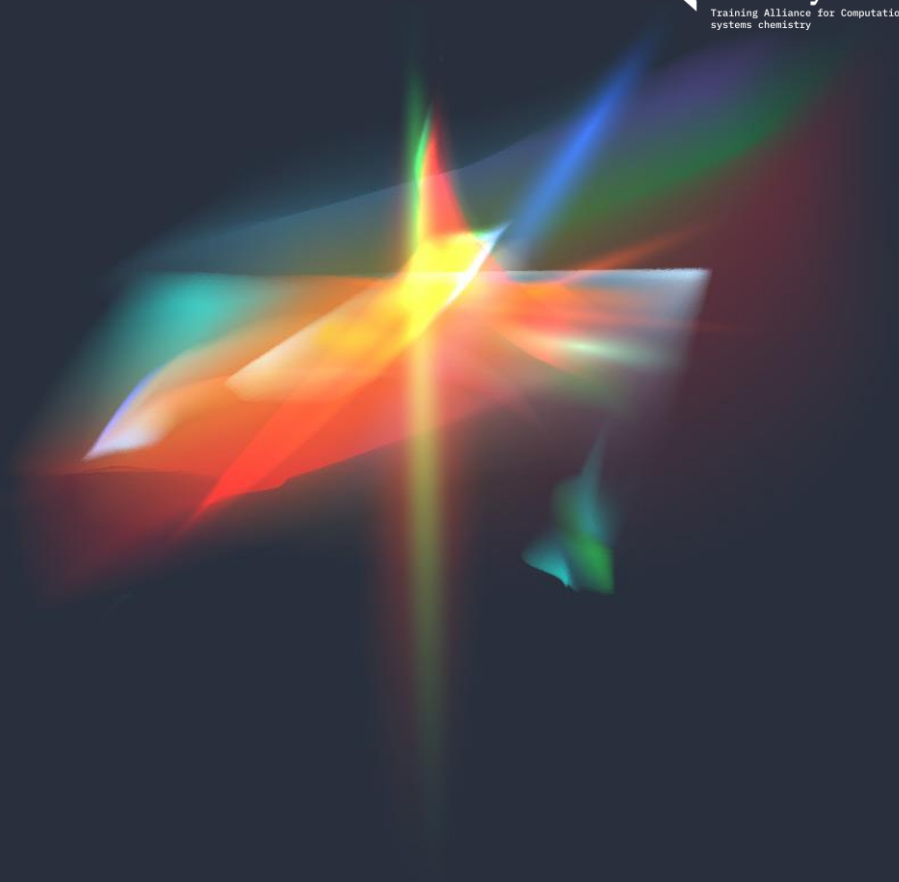
The final embeddings for reactants and products are derived by applying attention weights:

$$\mathcal{R} = \sum_{i=0}^{n_r-1} \mathcal{R}_{\text{updated}}[i,:] \cdot A_{\mathcal{R}_{\text{avg}}}[i], \quad \mathcal{P} = \sum_{i=0}^{n_p-1} \mathcal{P}_{\text{updated}}[i,:] \cdot A_{\mathcal{P}_{\text{avg}}}[i]$$

The reaction center Γ is then used in a neural network with softmax activation for classification, optimizing using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log(p_{i,c})$$

TACsy
Training Alliance for Computational systems chemistry



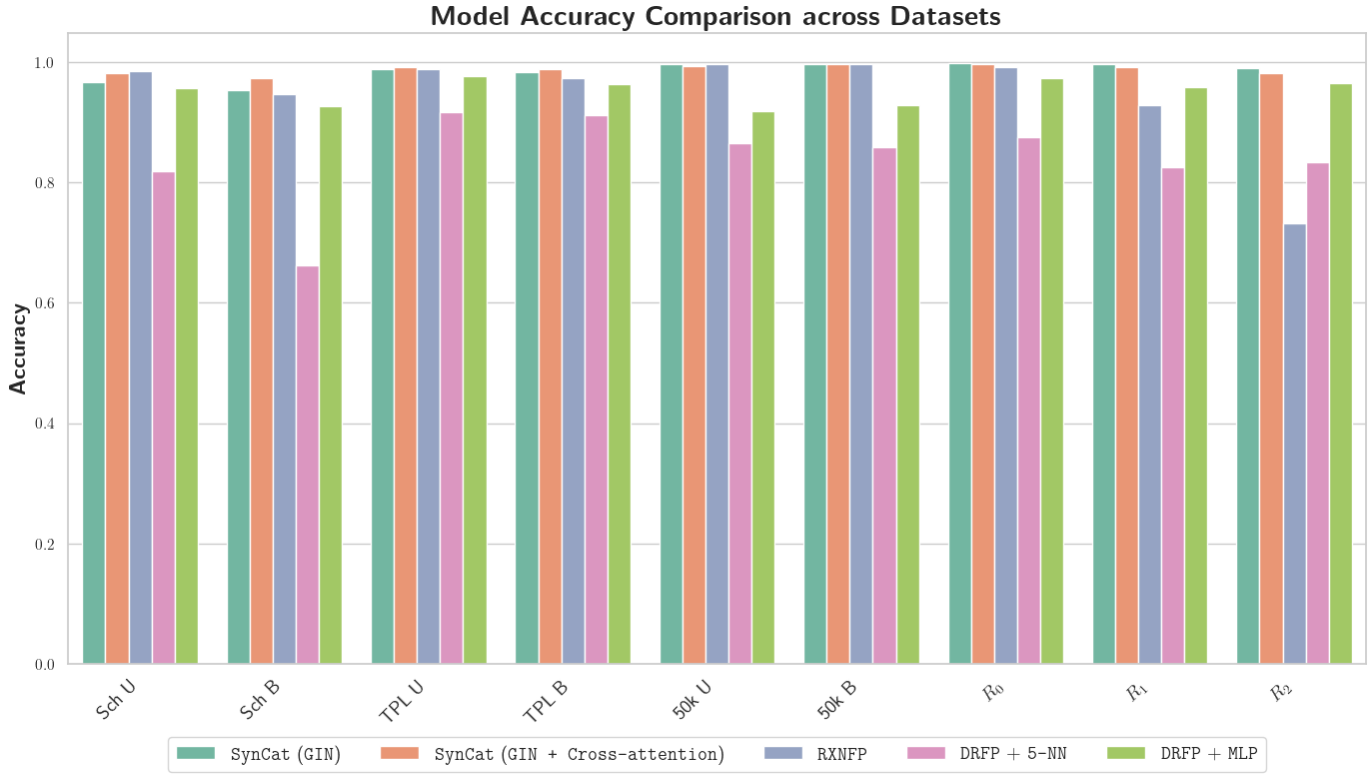
03

RESULT

RESULT

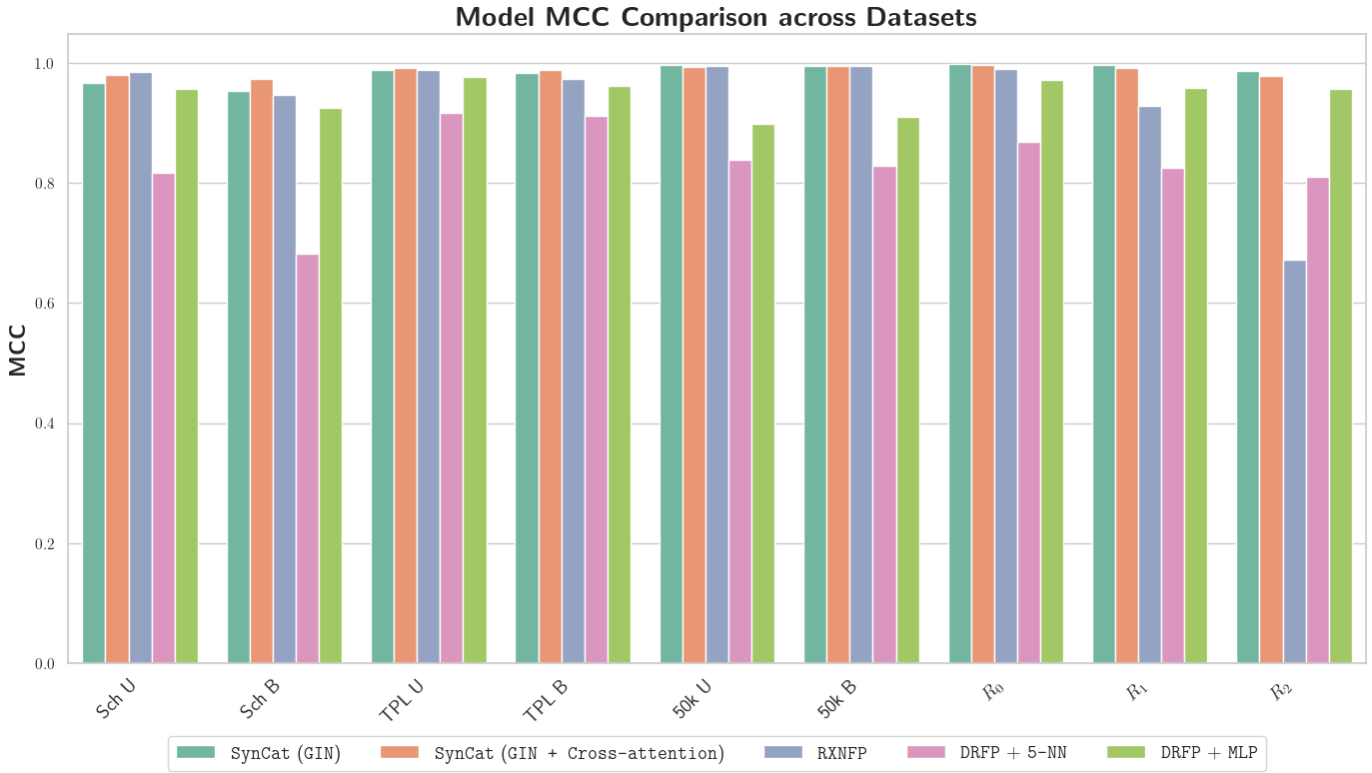
Benchmarking

TACsy
Training Alliance for Computational systems chemistry



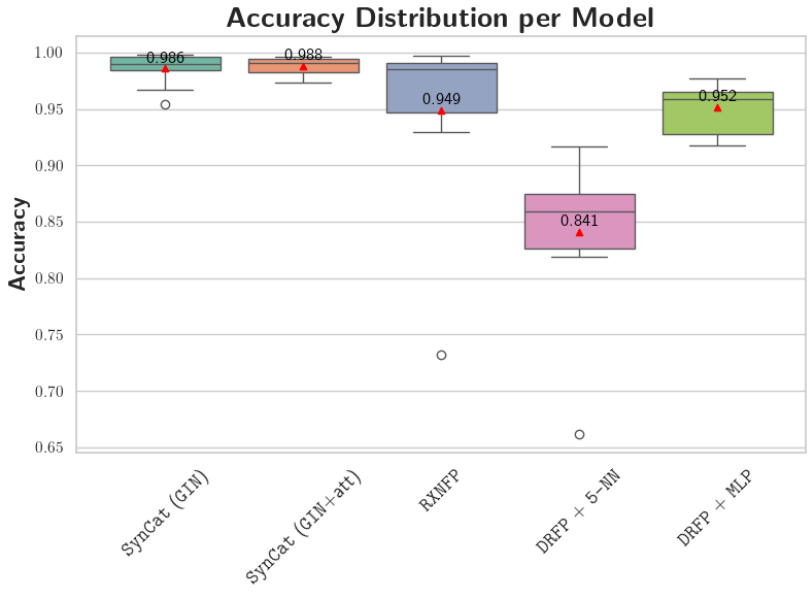
RESULT

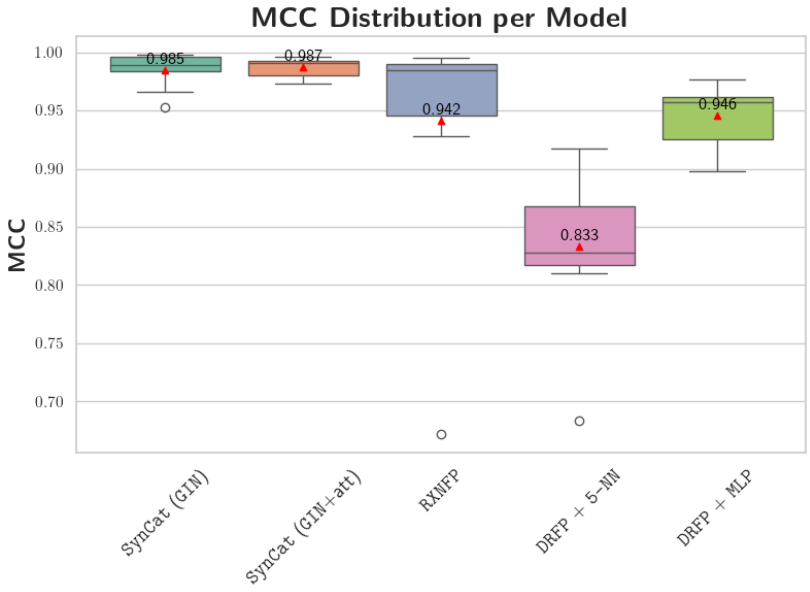
Benchmarking



RESULT

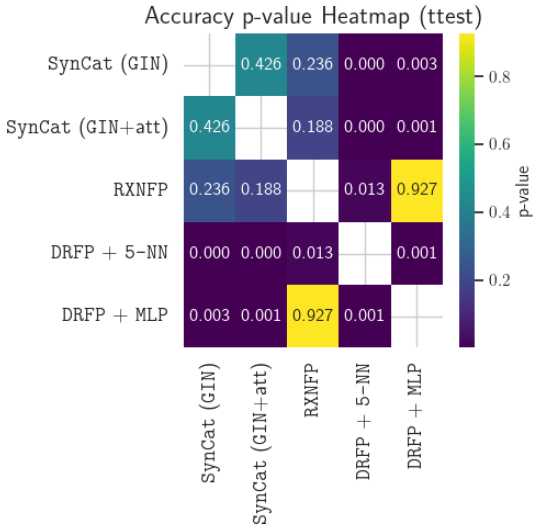
Benchmarking

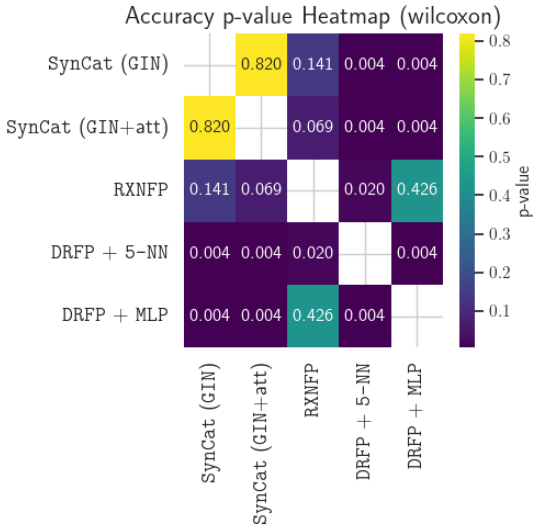


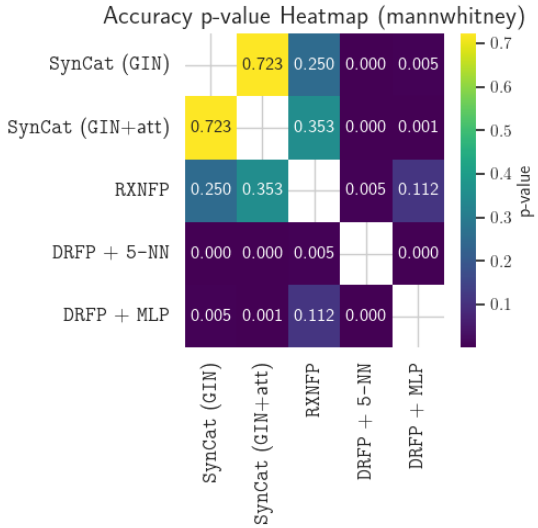


RESULT

Benchmarking

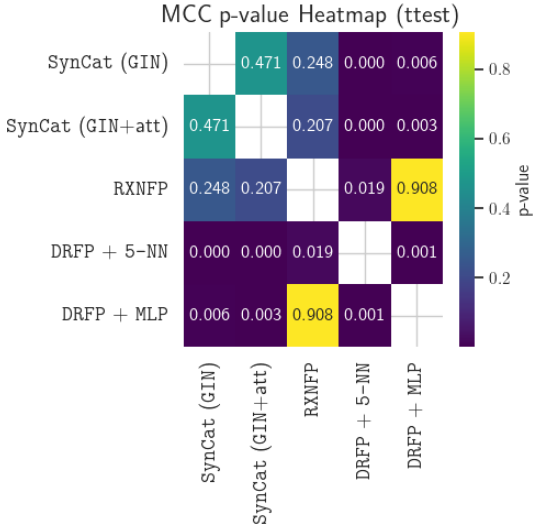


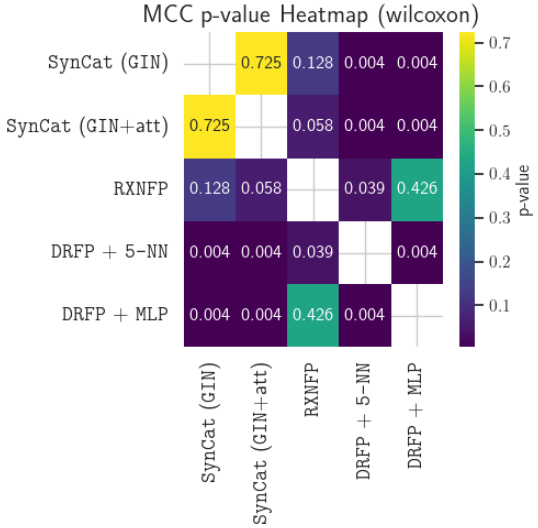


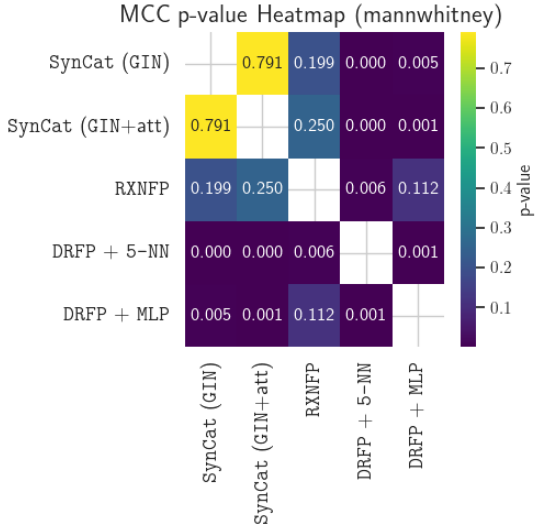


RESULT

Benchmarking

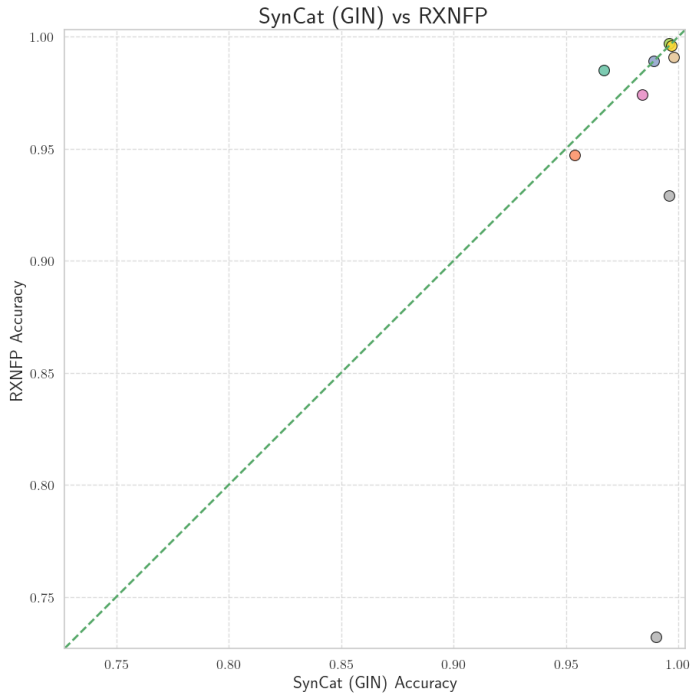


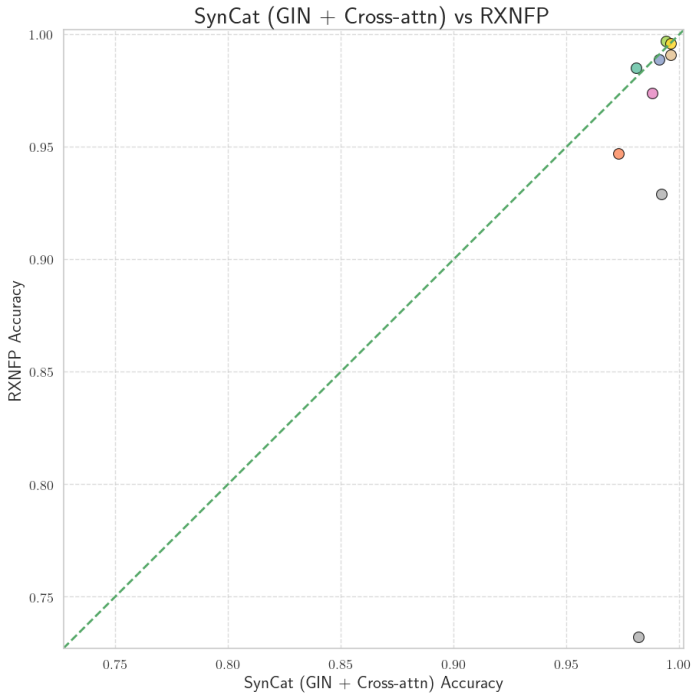




RESULT

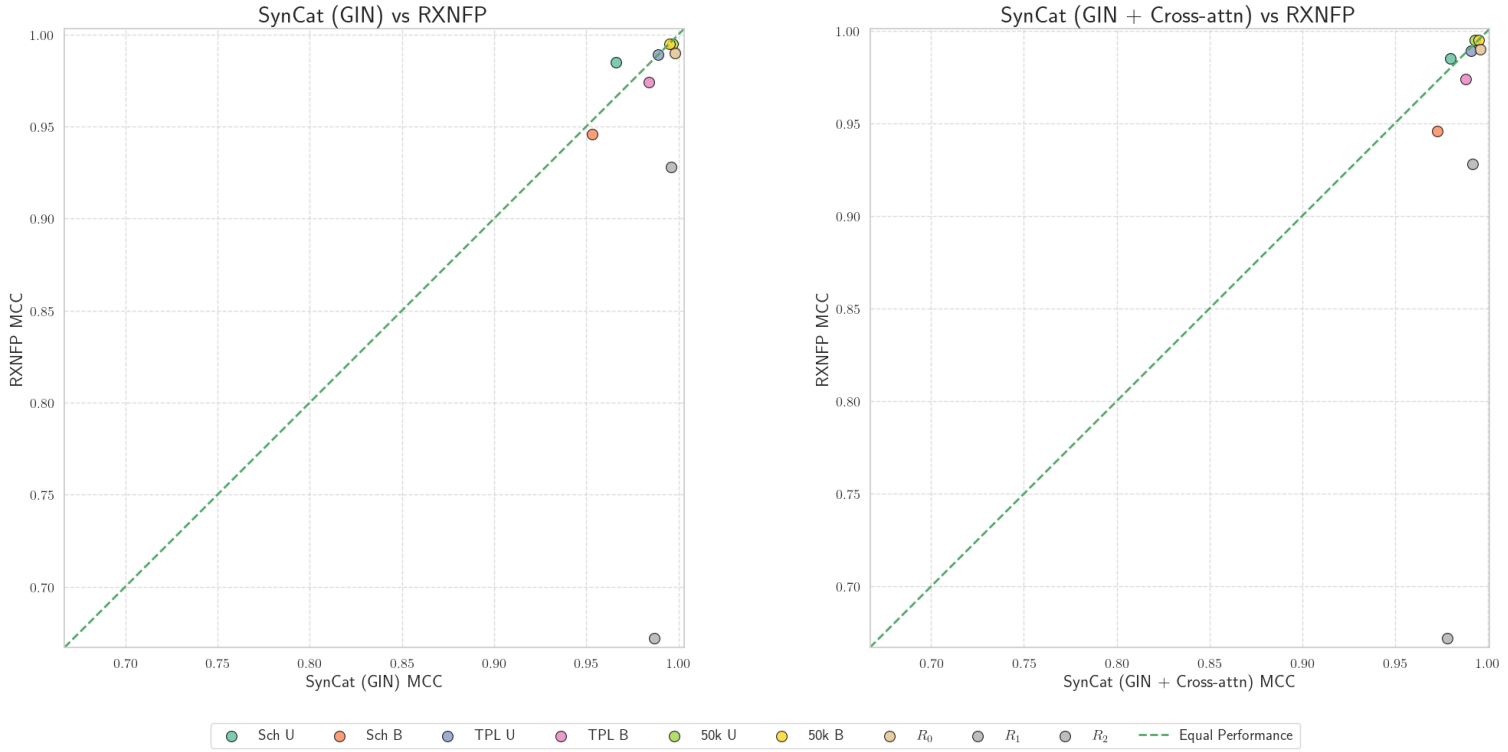
Benchmarking





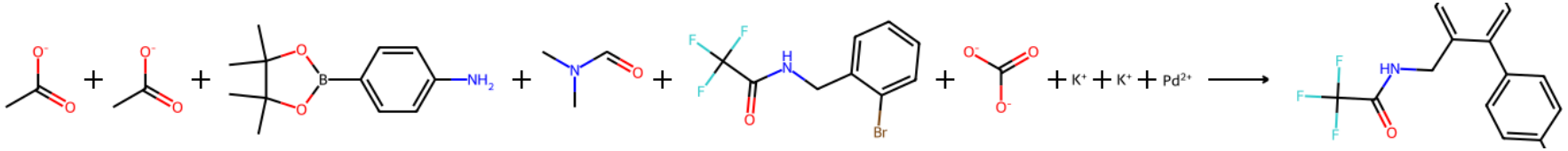
RESULT

Benchmarking

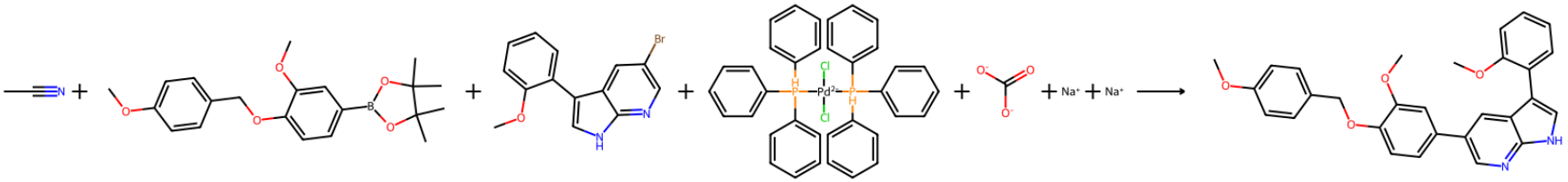


RESULT

Benchmarking



Dataset: Schneider
Label: Bromo Suzuki Coupling
Predict: Bromo Suzuki-type Coupling



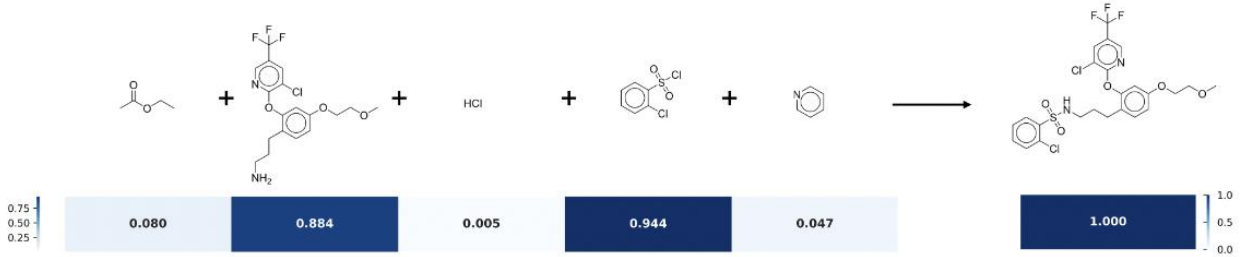
Dataset: Schneider
Label: Bromo Suzuki Coupling
Predict: Bromo Suzuki-type Coupling

RESULT

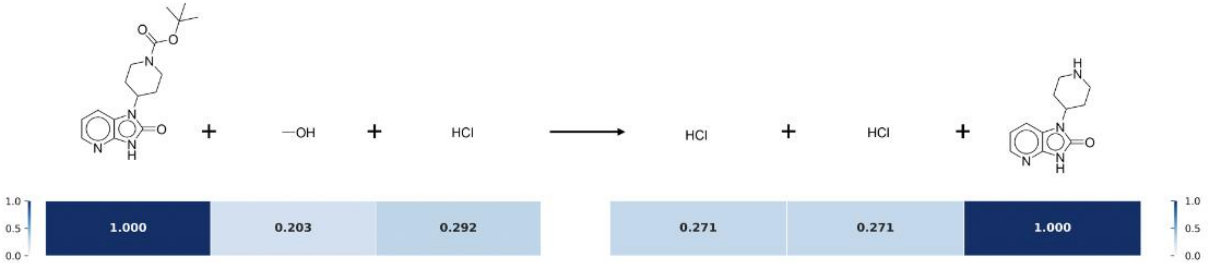
TACsy
Training Alliance for Computational systems chemistry

Reagent detection

Dataset: USPTO TPL
TPL label 43: [C:4][S:1](=[O:2])(=[O:3])Cl.[C:5][N:6]>>[C:5][N:6][S:1](=[O:2])(=[O:3])



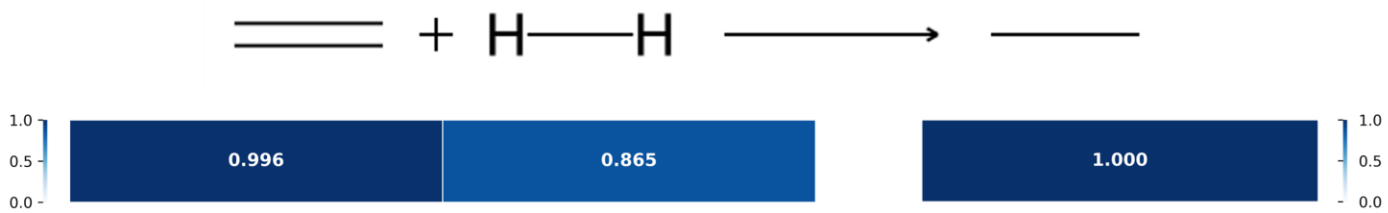
Dataset: Schneider
Label: N-Boc deprotection



RESULT

TACsy
Training Alliance for Computational systems chemistry

Reagent detection



TACsy
Training Alliance for Computational systems chemistry

04

DISCUSSION

DISCUSSION

TACsy
Training Alliance for Computational systems chemistry

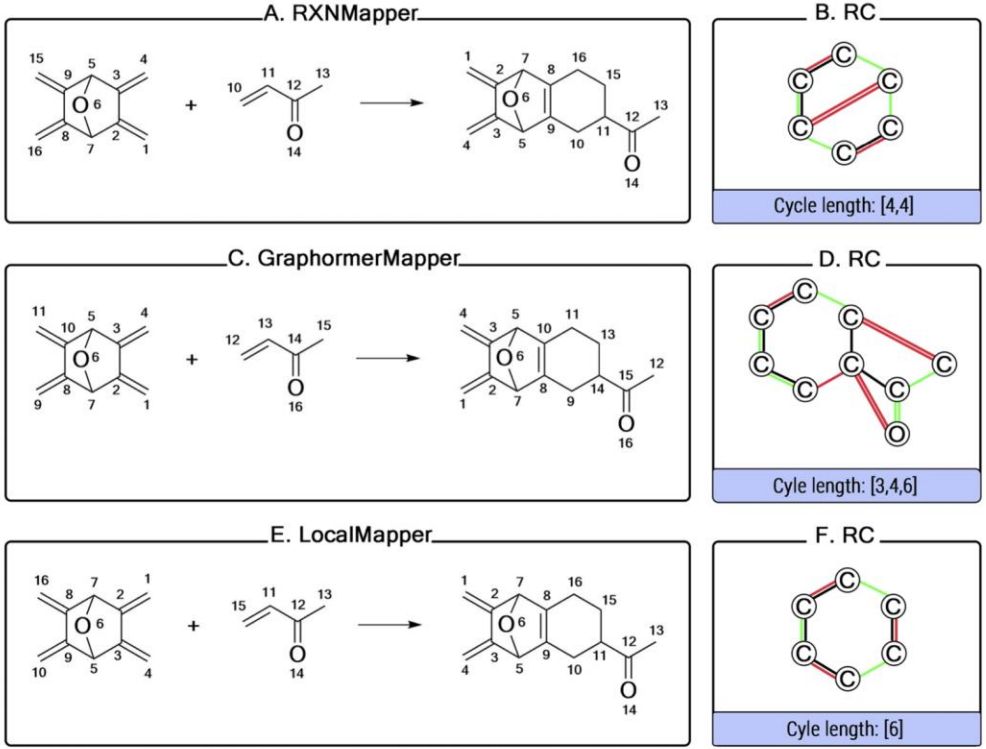
Database Organization

SynCat + SynTemp

DISCUSSION

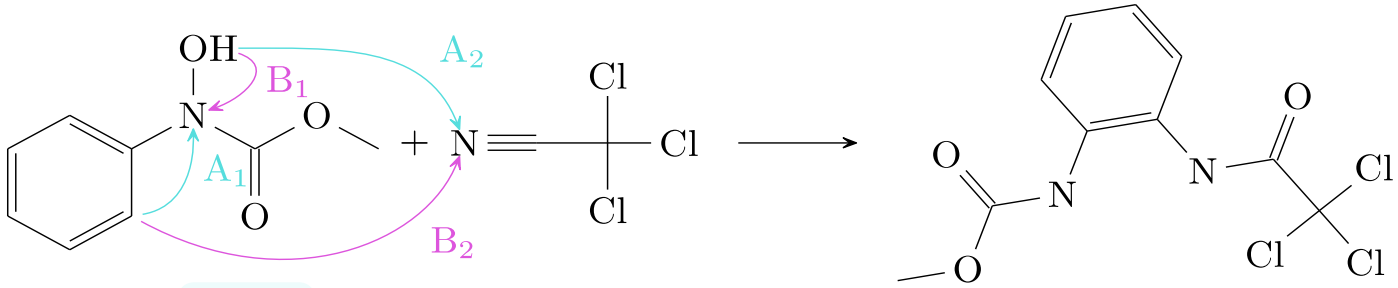
TACsy
Training Alliance for Computational systems chemistry

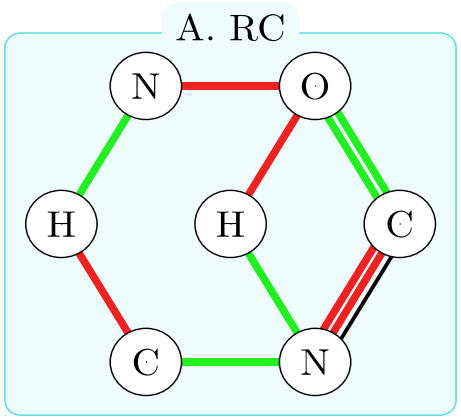
Inequivalent AAMs

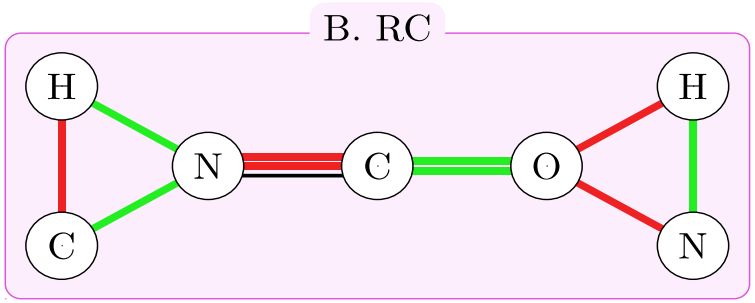


DISCUSSION

Ambiguous Hydrogen







Our team

TACsy
Training Alliance for Computational systems chemistry

Phuoc-Chung Van Nguyen
UMP

Ngoc-Vi Nguyen Tran
Uppsala University

Tieu-Long Phan
Leipzig University

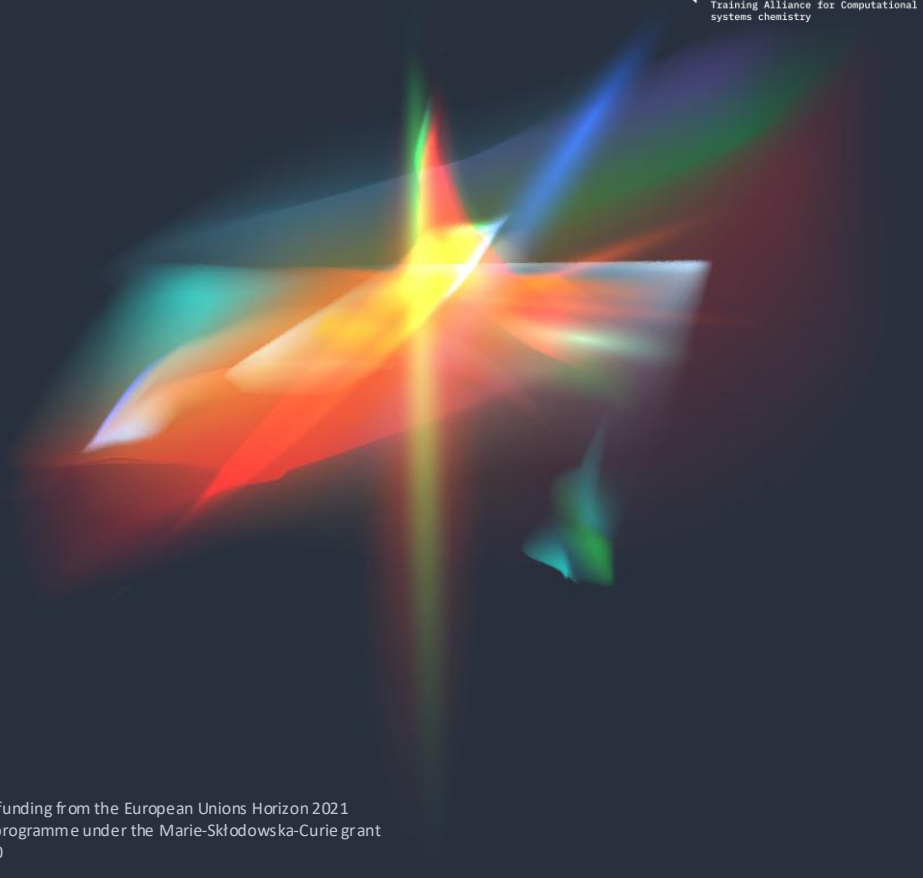
Peter F. Stadler
Leipzig University

SDU University of Southern Denmark

UNIVERSITÄT LEIPZIG

TACsy Training Alliance for Computational systems chemistry

Thank you for your attendance



MARIE CURIE ACTIONS

Founded by the European Union

This project has received funding from the European Unions Horizon 2021 research and innovation programme under the Marie-Skłodowska-Curie grant agreement No 101072930

TACsy
Training Alliance for Computational systems chemistry