UNIVERSITÄT LEIPZIG

TACsy

Training Alliance for Computational systems chemistry

Synthesis Rebalancing Framework

Tieu-Long Phan & Klaus Weinbauer

Date: 13.02.2024

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

AGENDA

TACsy
Training Alliance for Computational systems chemistry

01
Introduction

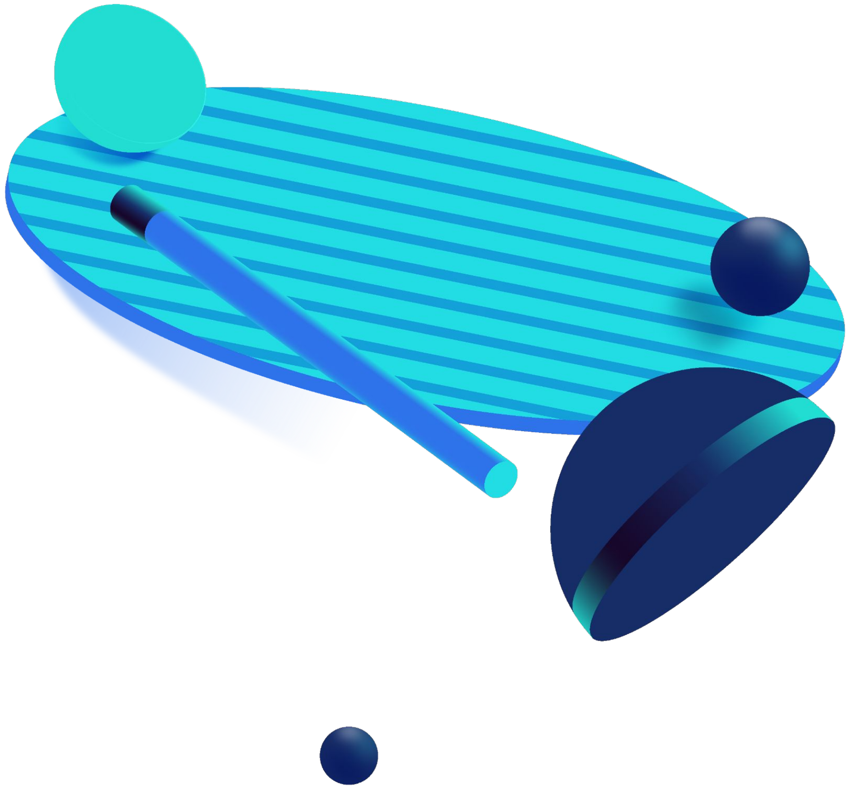
02
Proposed Method

03
Result Discussion

04
Conclusion

INTRODUCTION

01

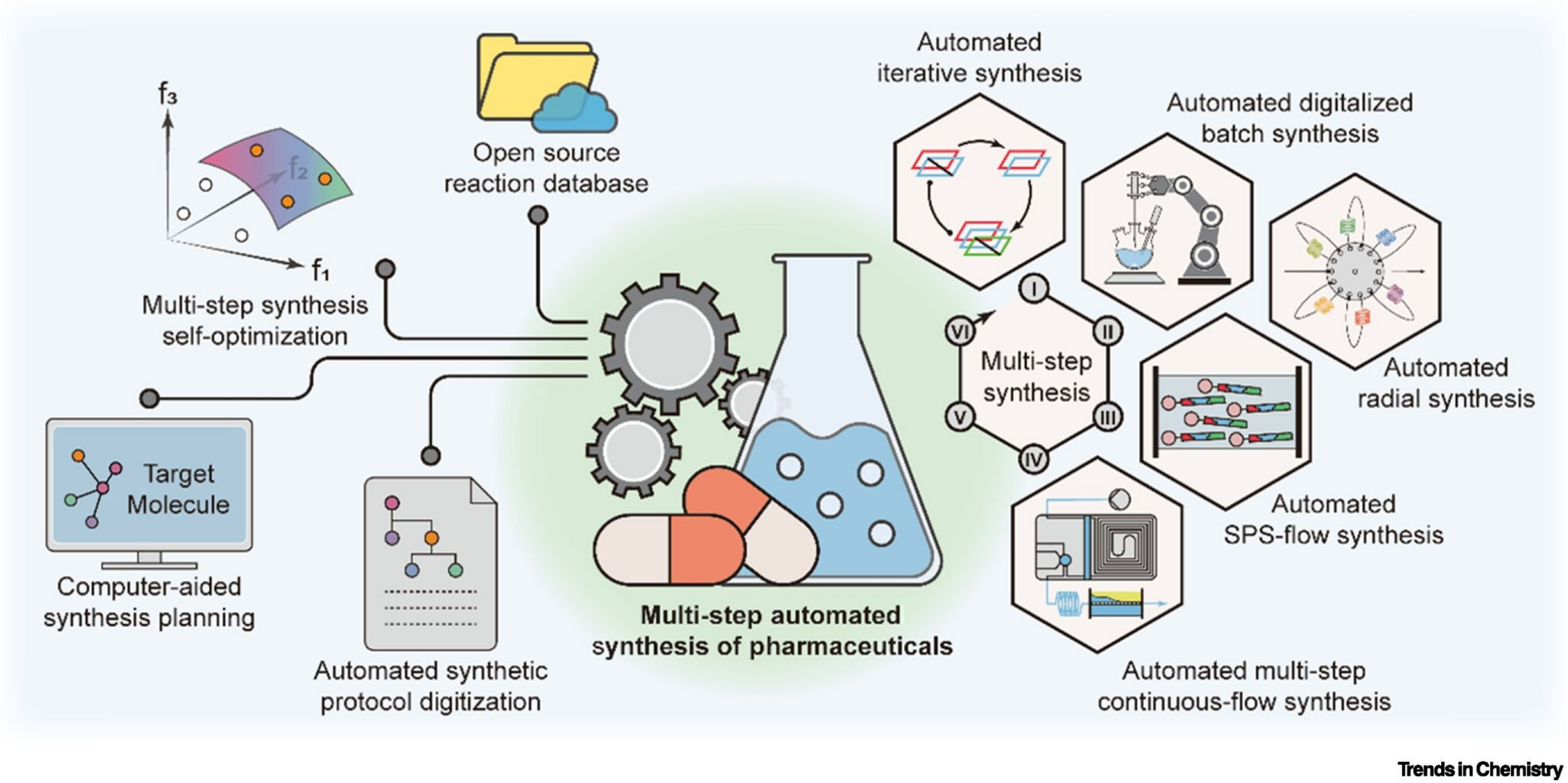


Introduction

INTRODUCTION

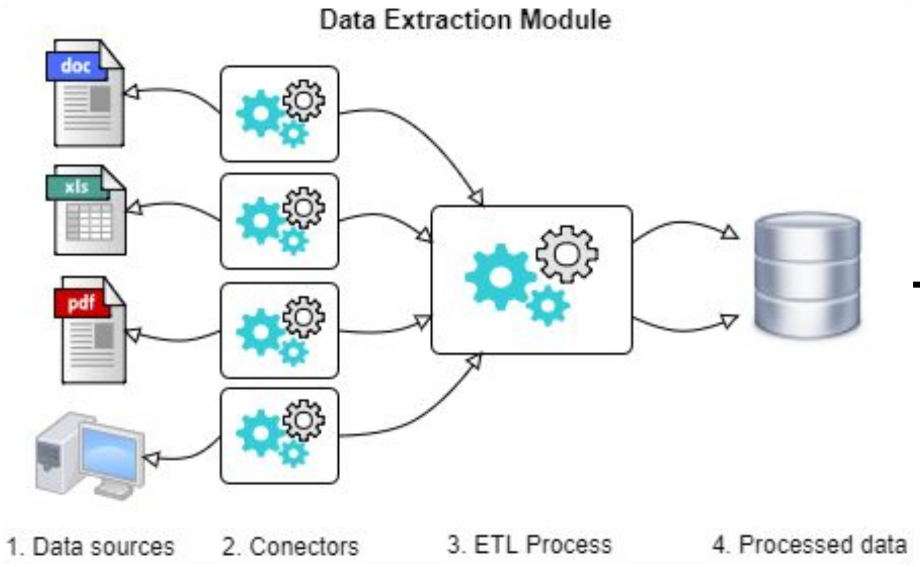
TACsy
Training Alliance for Computational systems chemistry

INTRODUCTION | METHOD | RESULT-DISCUSSION | CONCLUSION



Wang, G., Ang, H. T., Dubbaka, S. R., O'Neill, P., & Wu, J. (2023). Multistep automated synthesis of pharmaceuticals. Trends in Chemistry.

INTRODUCTION



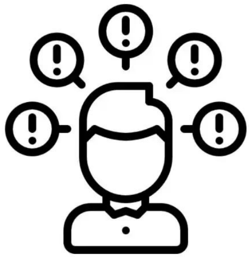
Insufficient data

Bad data = Bad model

(bad data can mess up how companies decide things)

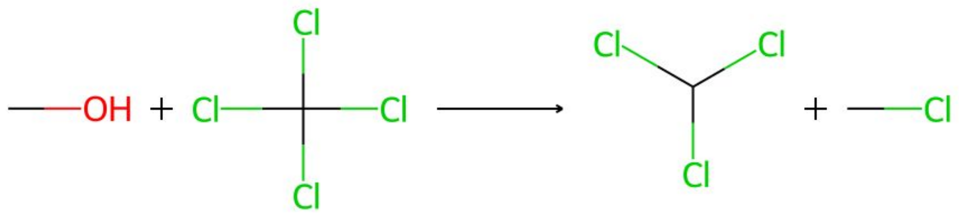
Incorrect algorithm selection
Incorrect hyperparameter tuning
Incorrect model deployment
Wrong evaluation metrics
Poorly collected requirements
...

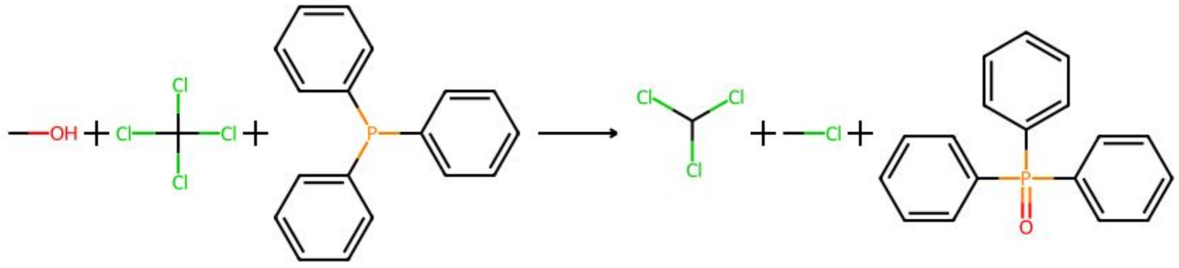
The effect of (bad) Data Quality on Model Accuracy in Supervised Machine Learning



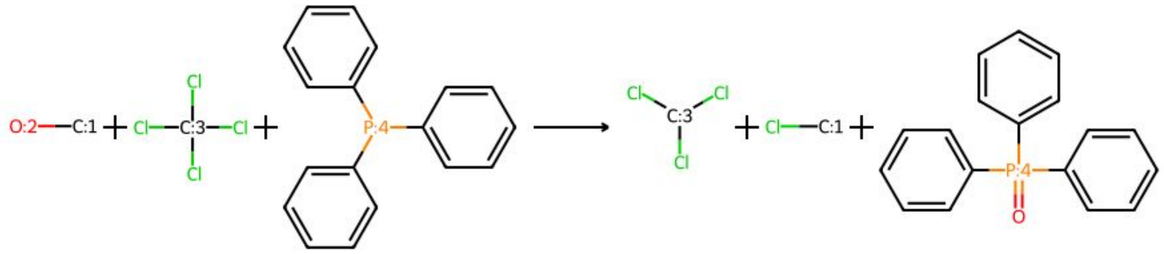
Saquicela, V., Baculima, F., Orellana, G., Piedra, N., Orellana, M., & Espinoza, M. (2018, March). Similarity Detection among Academic Contents through Semantic Technologies and Text Mining. In IWSW (pp. 1-12).

INTRODUCTION



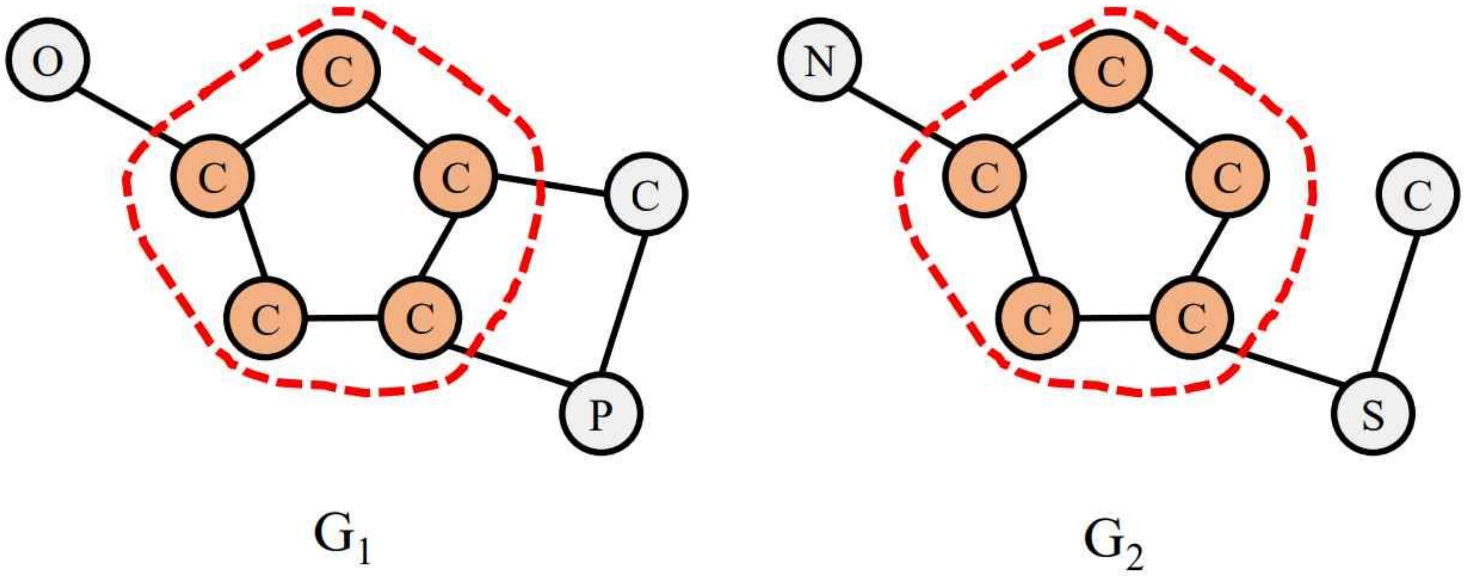


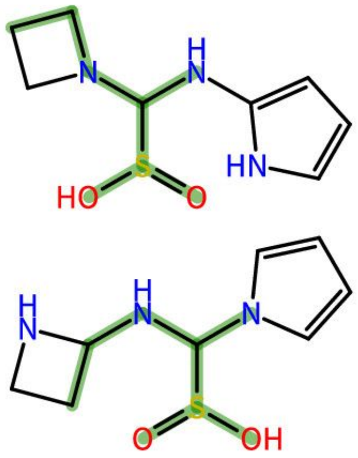
Rebalancing



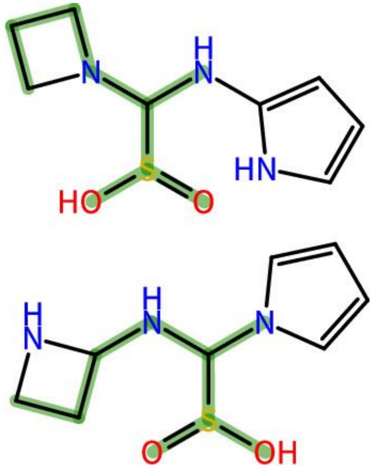
Atom mapping

Maximum-common-subgraph





(a) An example of the connected induced MCS (cMCIS).

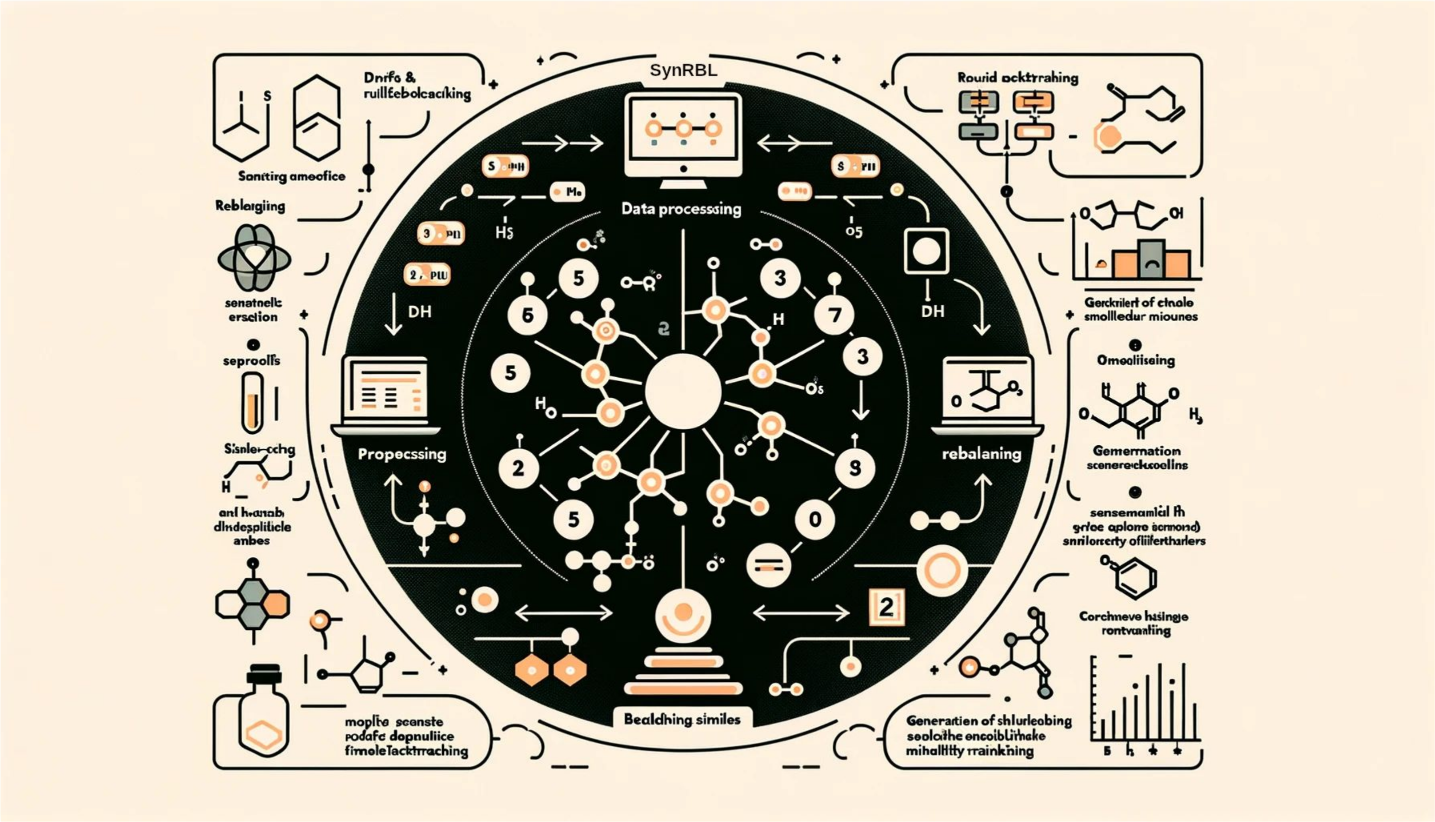


(b) An example of the connected noninduced MCS (cMCES).

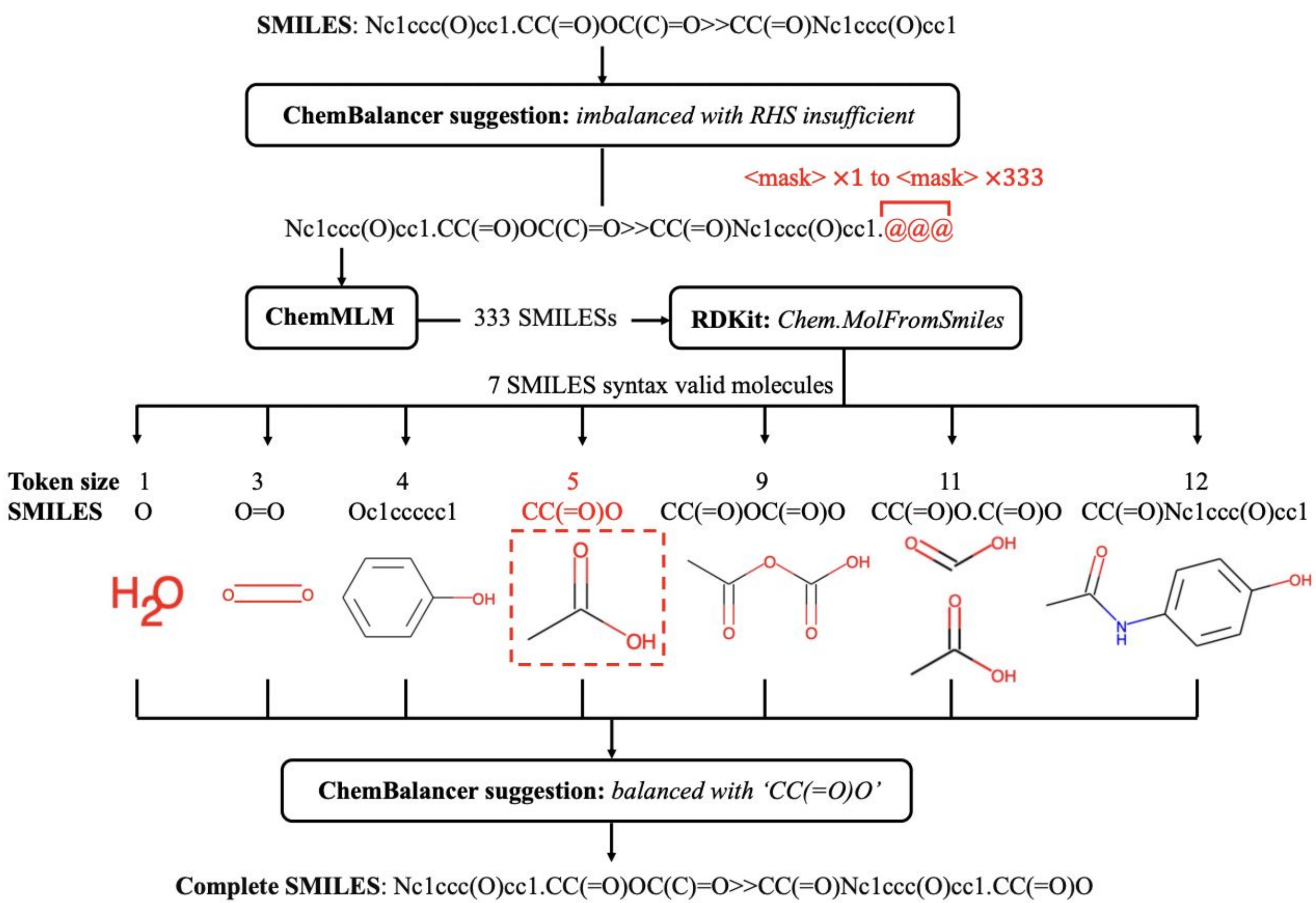
1. Bai, Y., Xu, D., Sun, Y., & Wang, W. (2021, July). Glsearch: Maximum common subgraph detection via learning to search. In International Conference on Machine Learning (pp. 588-598). PMLR.
2. Robert Schmidt, Florian Krull, Anna Lina Heinzke, and Matthias Rarey. Journal of Chemical Information and Modeling 2021 61 (1), 167-178 DOI: 10.1021/acs.jcim.0c00741

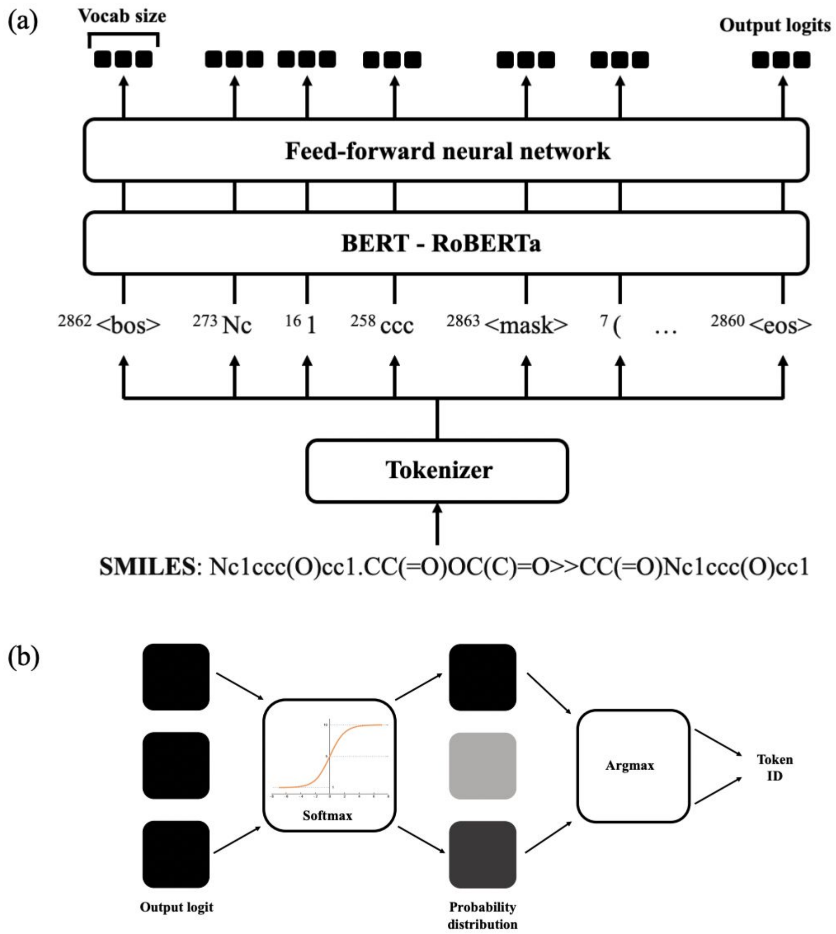
SynRBL: Synthesis Rebalancing Framework

SynRBL (Synthesis Rebalancing Framework) is a specialized toolkit designed for computational chemistry. Its primary focus is on rebalancing incomplete chemical reactions and providing rule-based methodologies for data standardization and analysis.



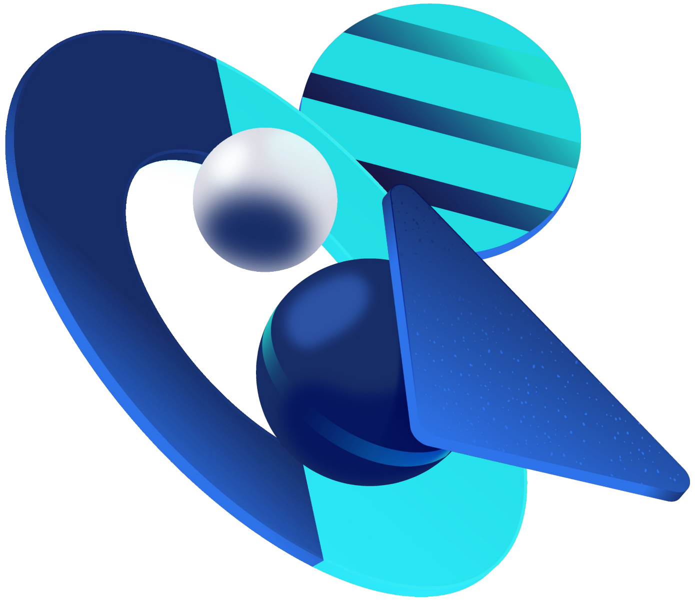
INTRODUCTION





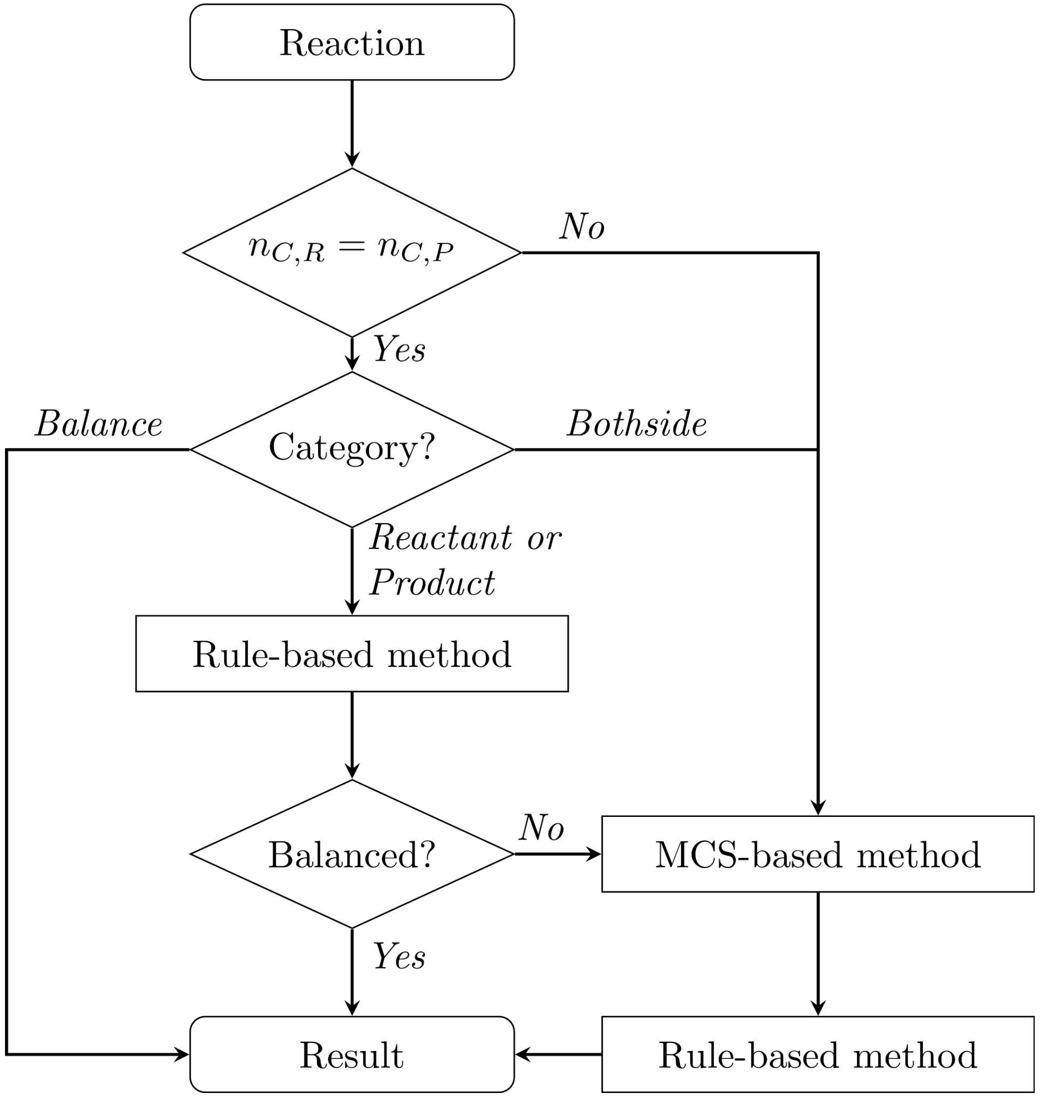
Zhang, C., Arun, A., & Lapkin, A. (2023). Completing and balancing database excerpted chemical reactions with a hybrid mechanistic-machine learning approach.

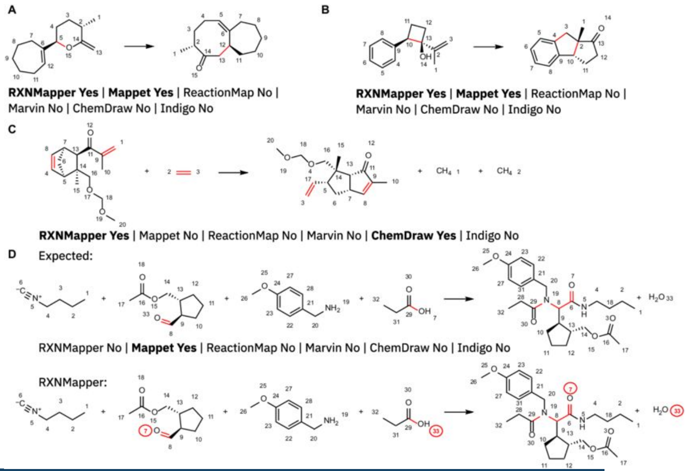
02



METHOD

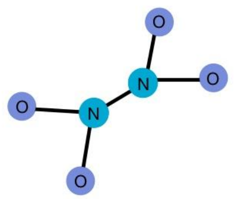
METHOD

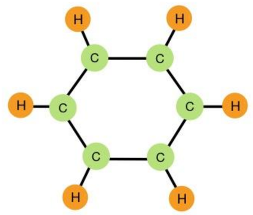


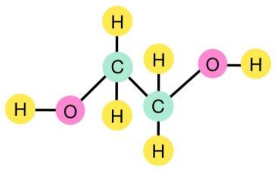


UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

Rule-based approach



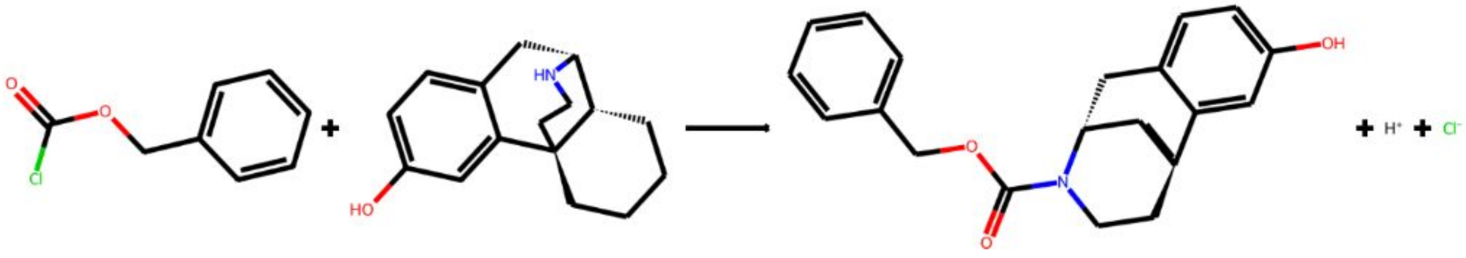




Molecular	Empirical
N_2O_4	NO_2
C_6H_6	CH
$C_2H_6O_2$	CH_3O

Molecular Representation
Eg: CH_3CHOOH
{C:2, H:4, O : 2, Q : 0}.

Rule-based approach

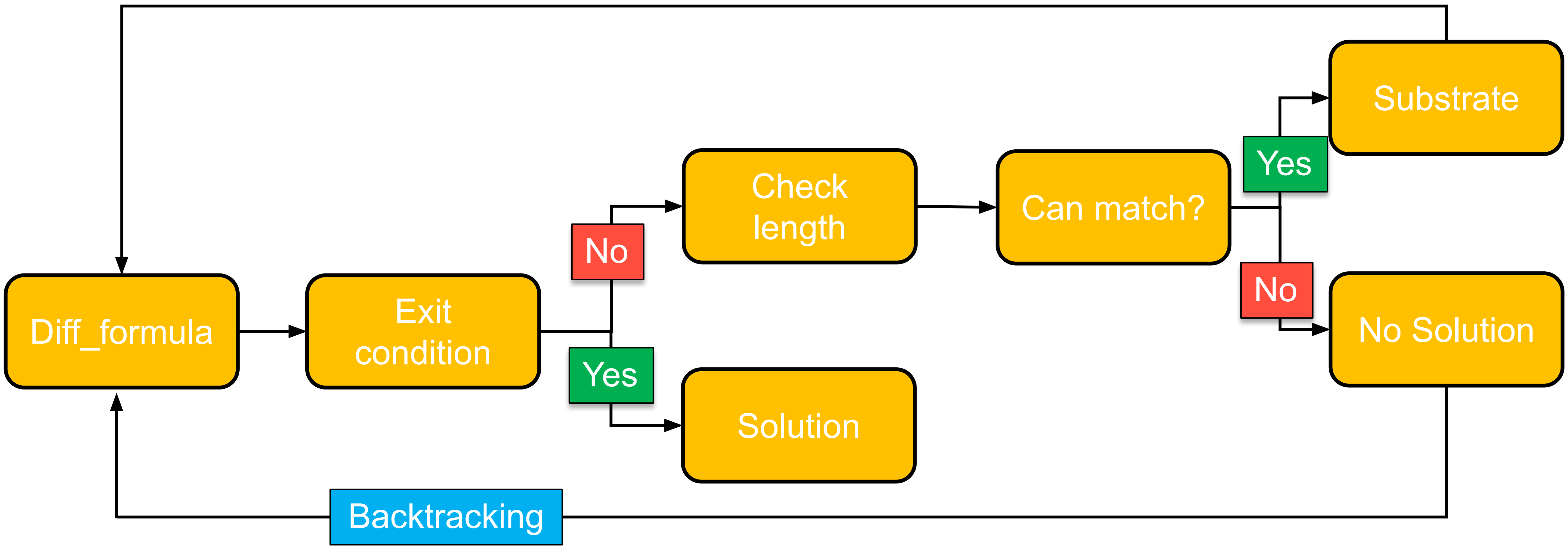


What is the current scale and comprehensiveness of the template library within this context?

METHOD

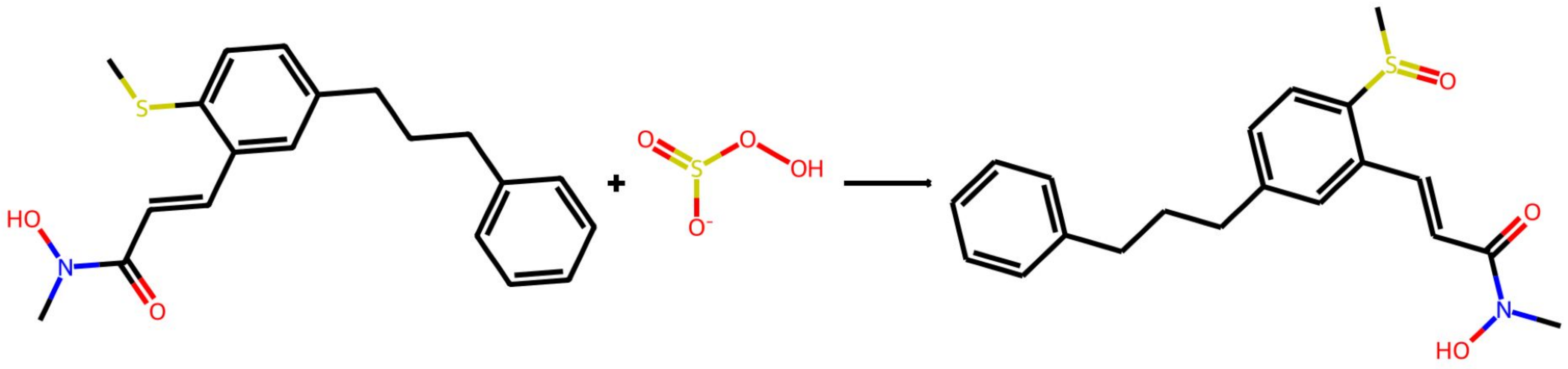
TACsy
Training Alliance for Computational systems chemistry

Rule-based approach



DFS search

Case Study

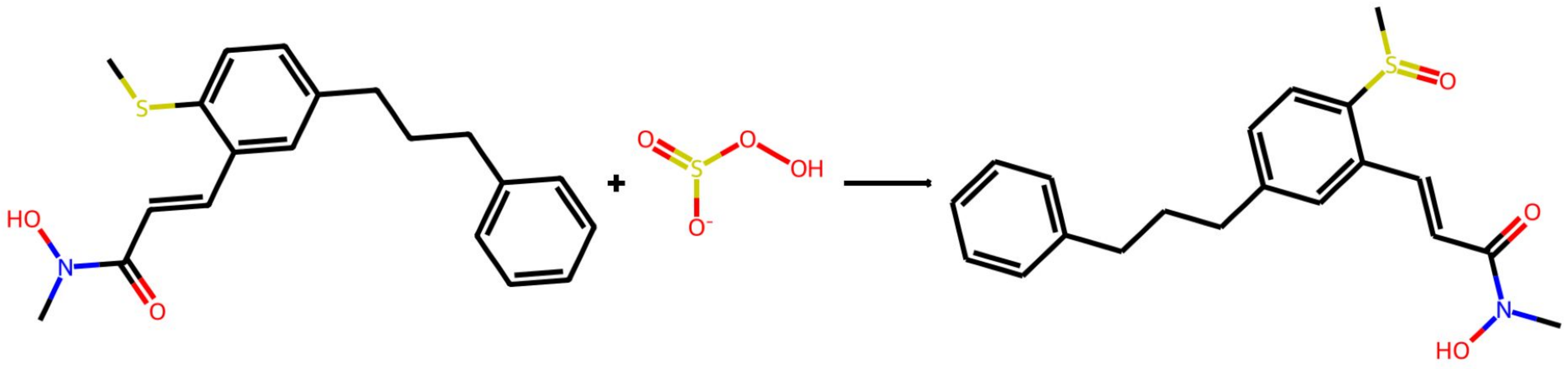


'Unbalance': 'Products'

'Diff_formula': {'S': 1, 'O': 3, 'H': 1, 'Q': -1},

METHOD

Case Study



Check length: 4

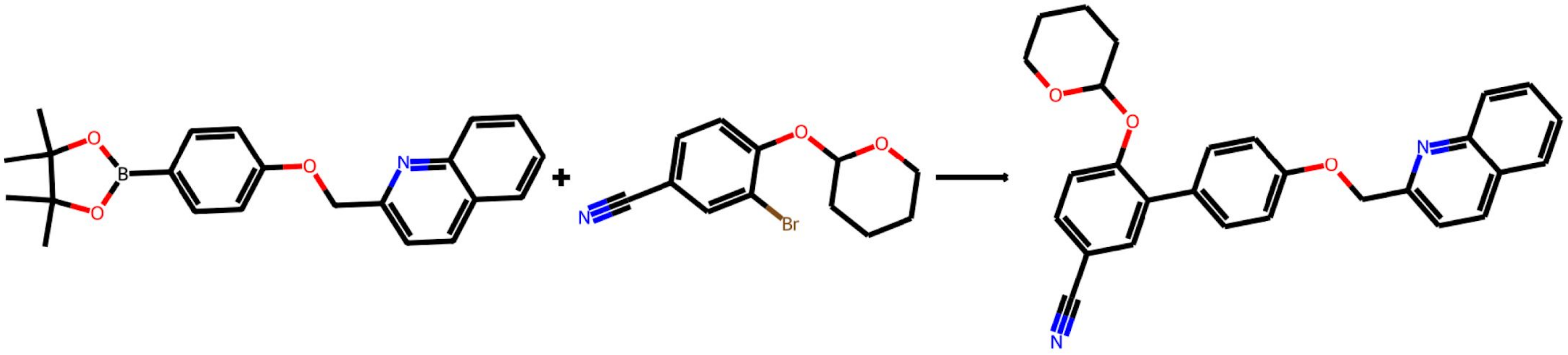
Search rules from length 4: SO_3^{2-} {'S': 1, 'O': 3, 'Q': -2}

Substrate: {‘H’:1, ‘Q’:+1}

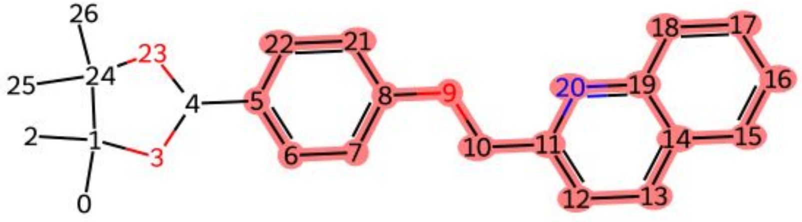
Check length: 2

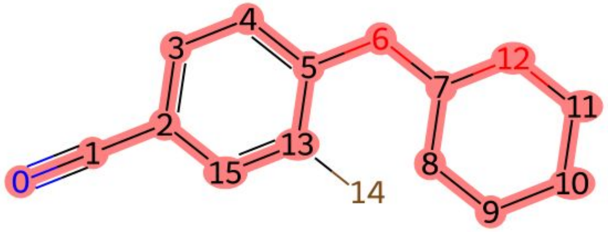
Search rules with length 2: H^+ {‘H’:1, ‘Q’:1}

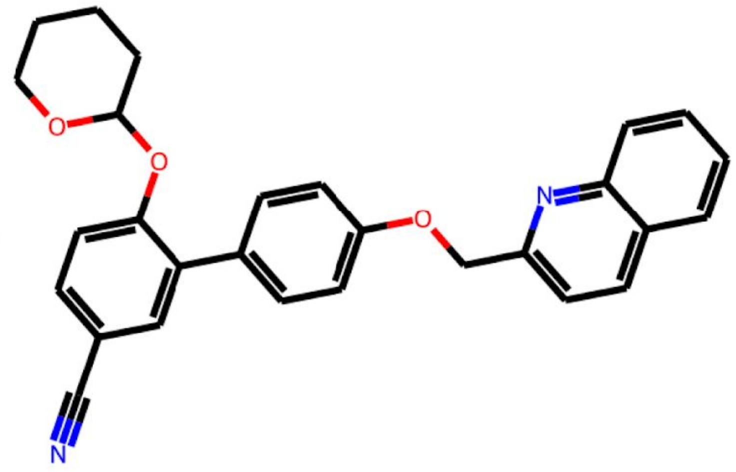
MCS-based approach



MCS-based approach

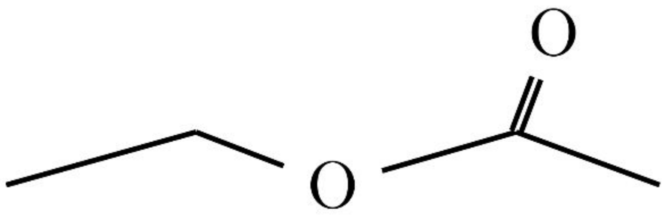






METHOD

Reactants

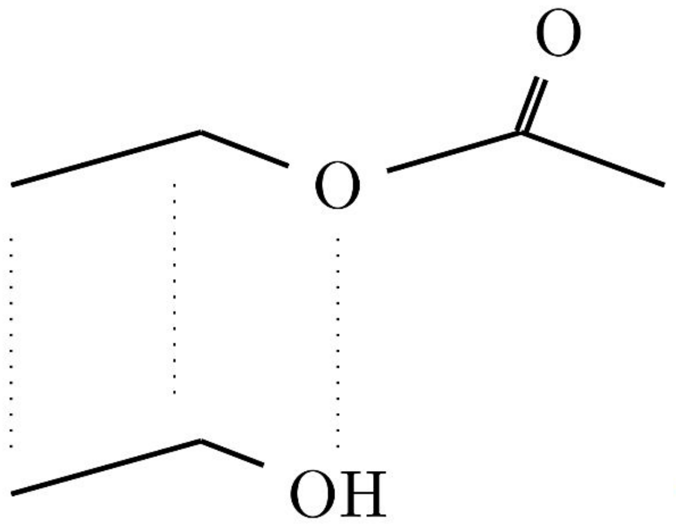


Products

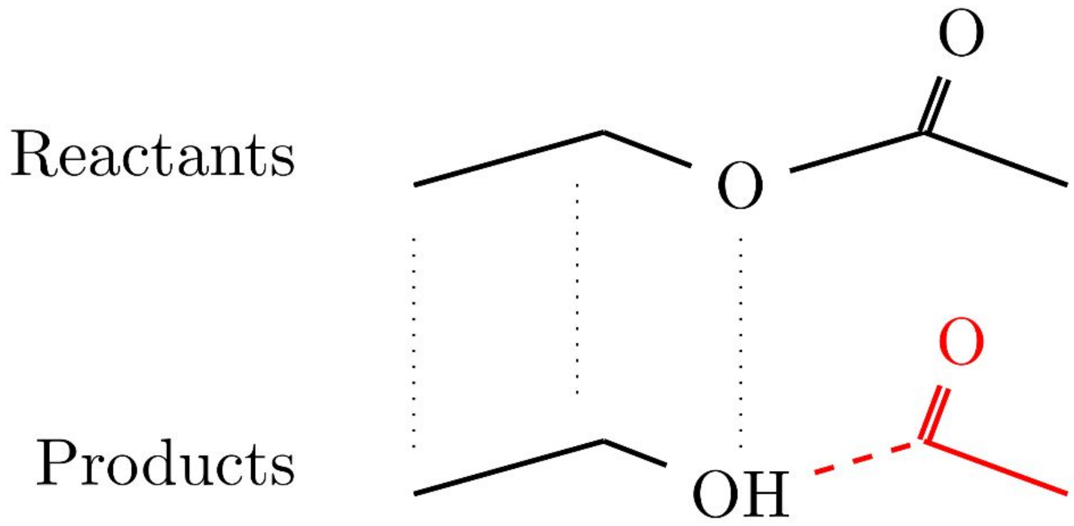
METHOD

Reactants

Products

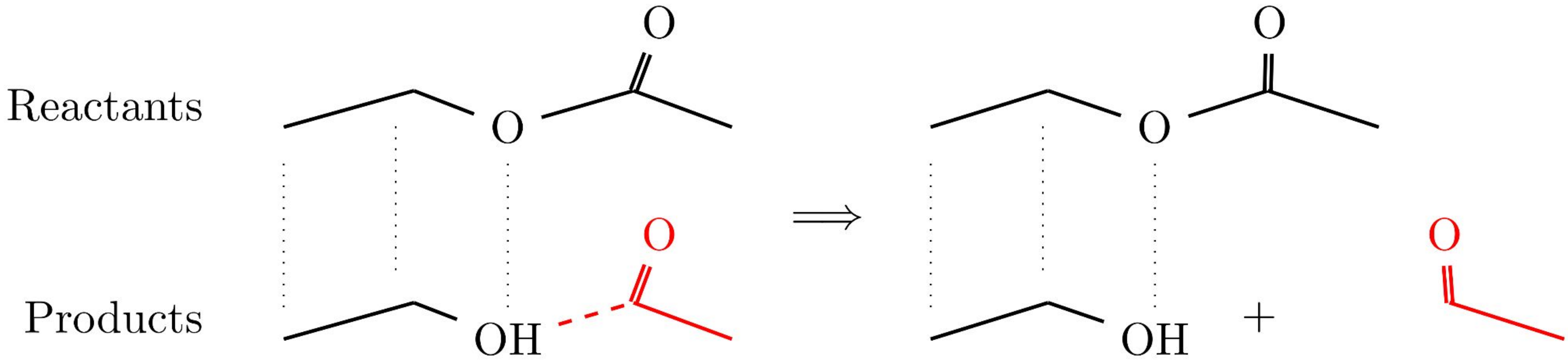


METHOD



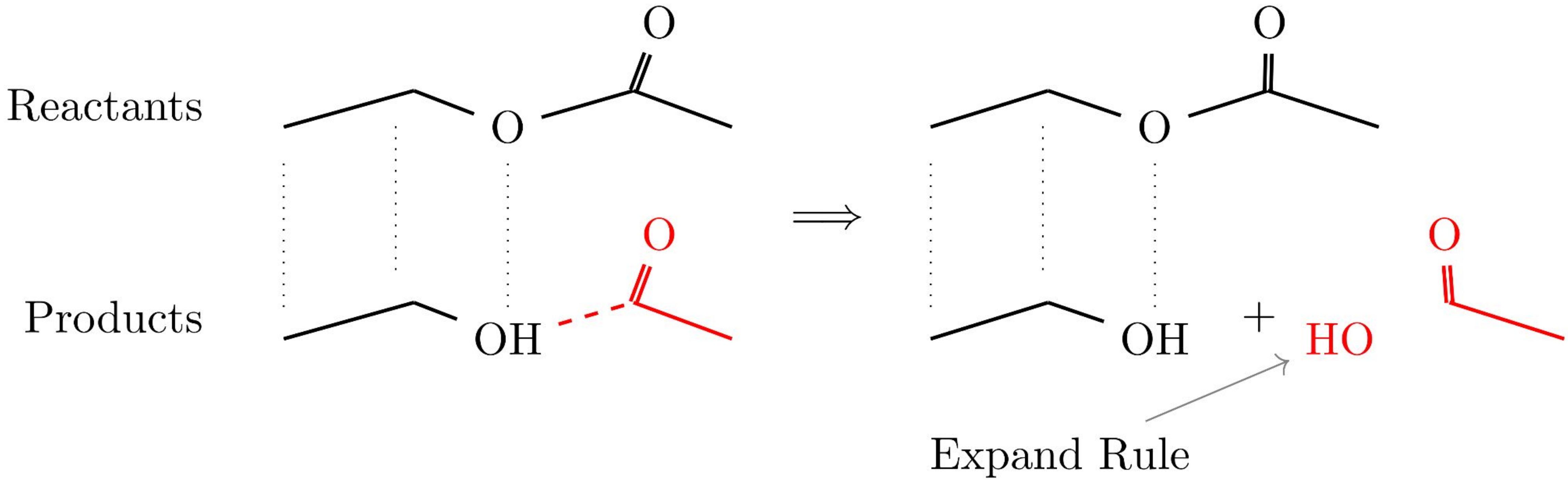
METHOD

Training Alliance for Computational systems chemistry

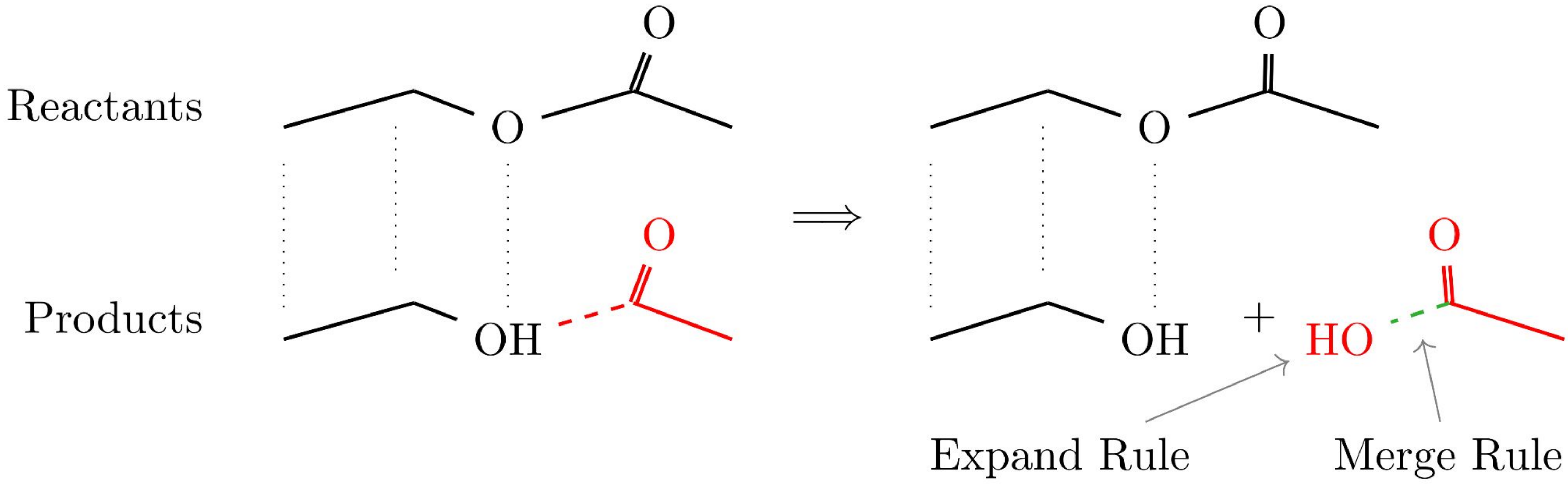


METHOD

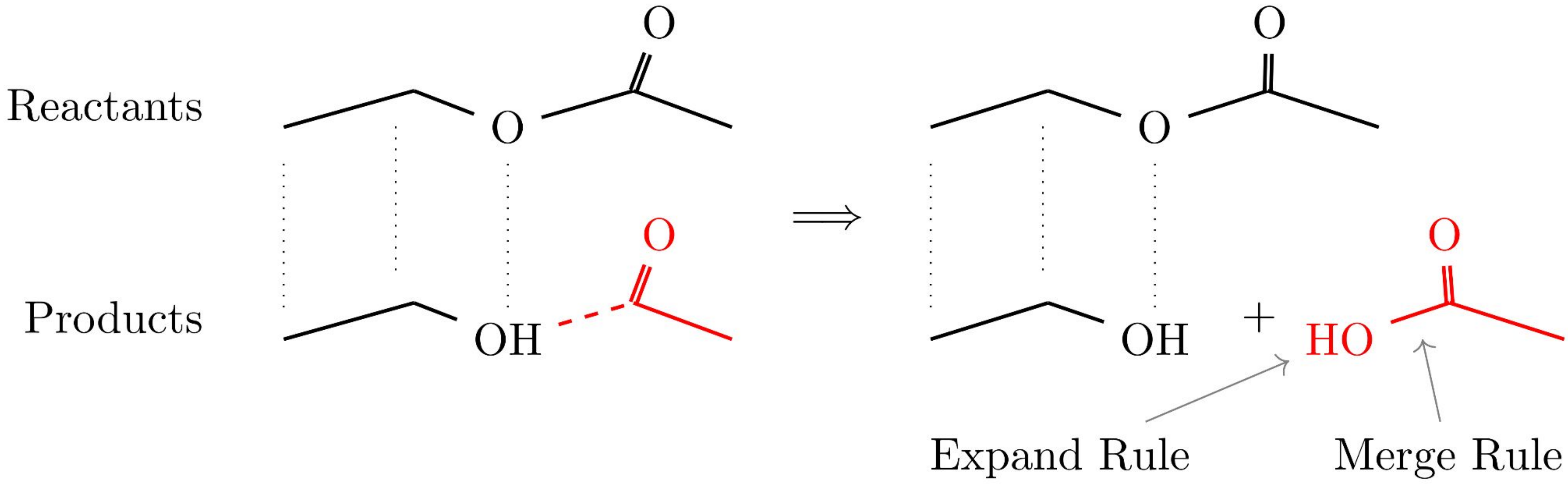
TACsy
Training Alliance for Computational systems chemistry



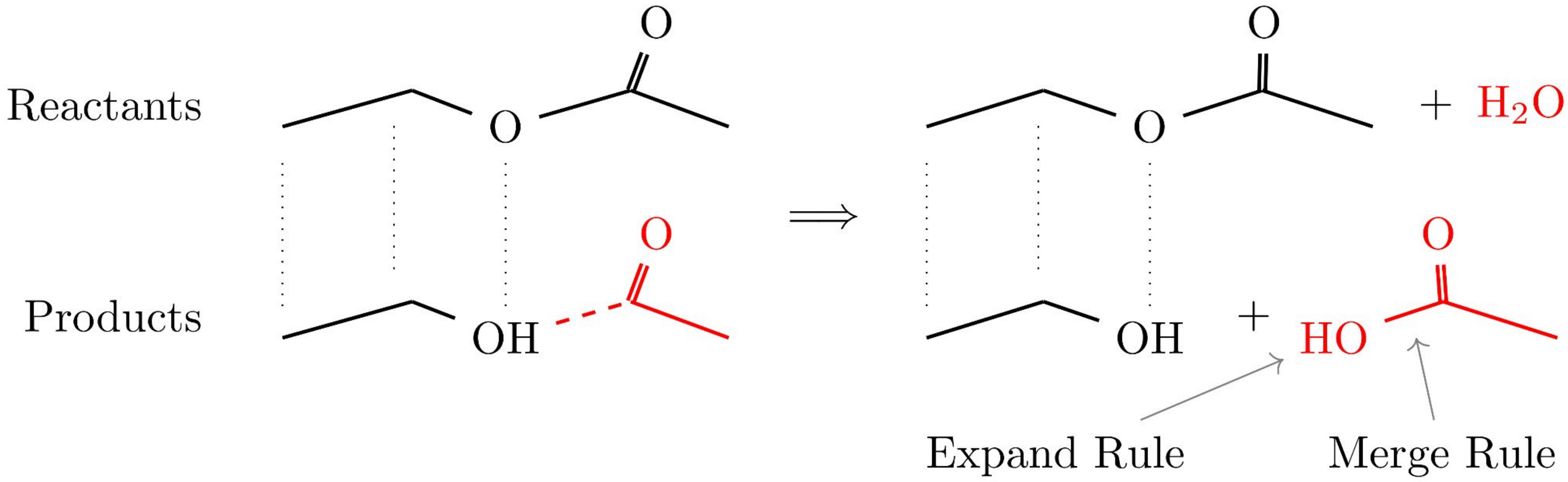
METHOD



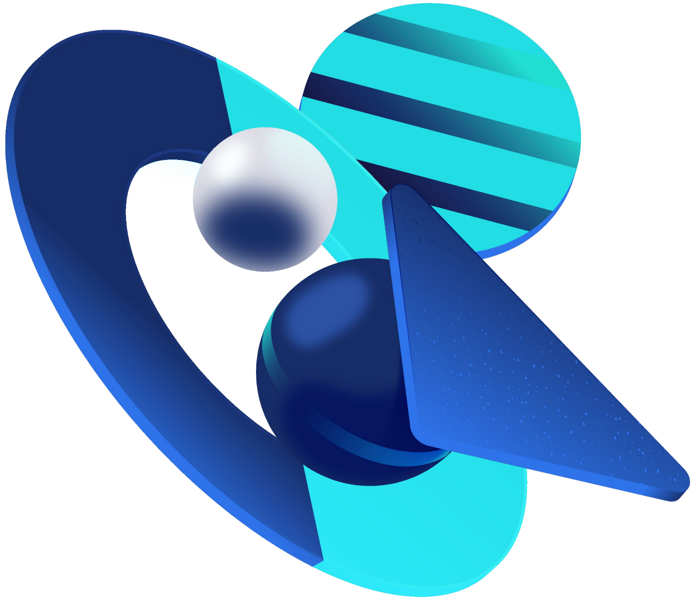
METHOD


Reactants
Products
O
O
OH
O
$\Longrightarrow$
O
O
OH
+
HO
O
Expand Rule
Merge Rule

METHOD

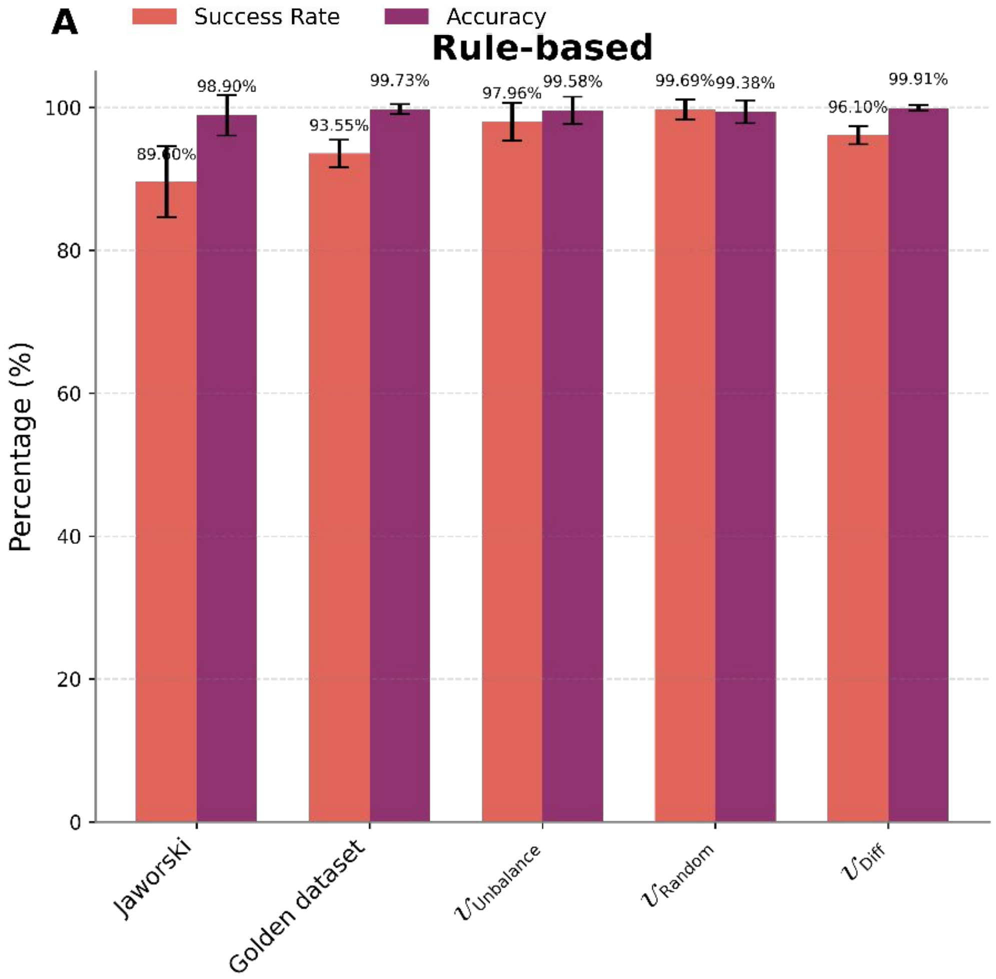

Reactants
Products
O
O
OH
O
O
O
OH
+ H_2O
+ HO
O
Expand Rule
Merge Rule

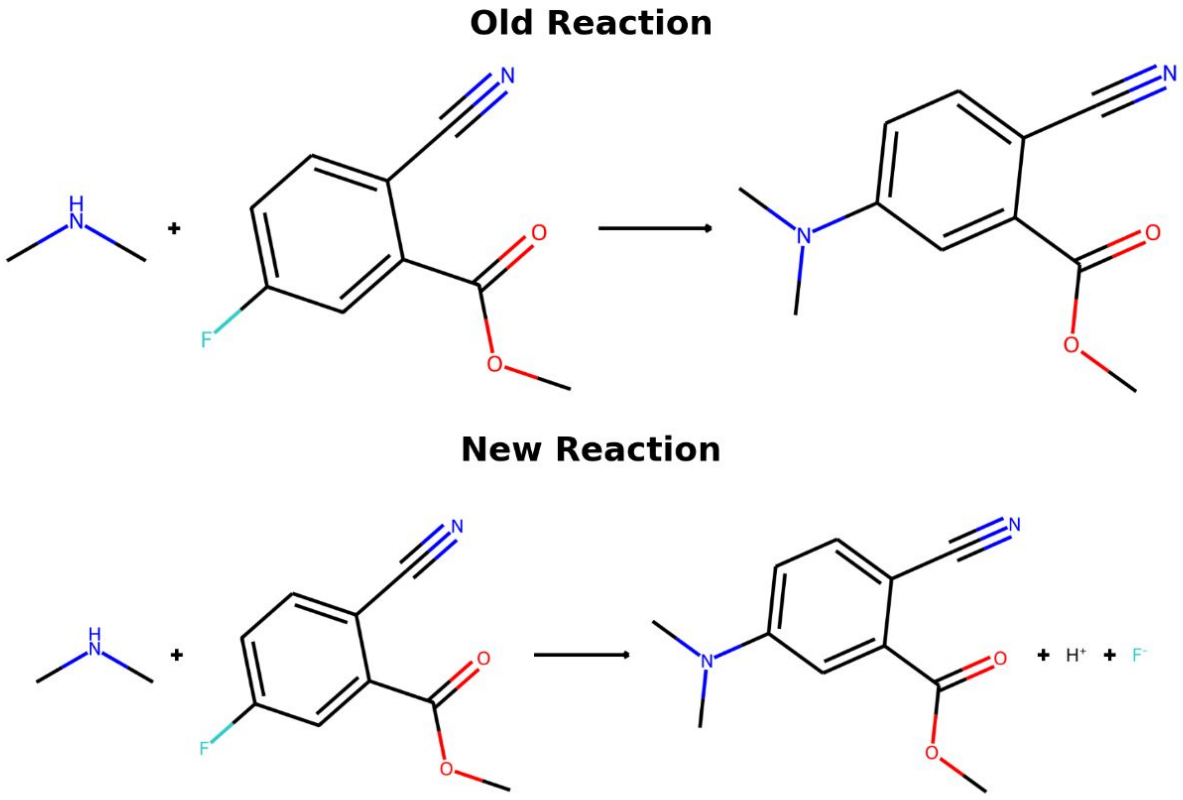
03



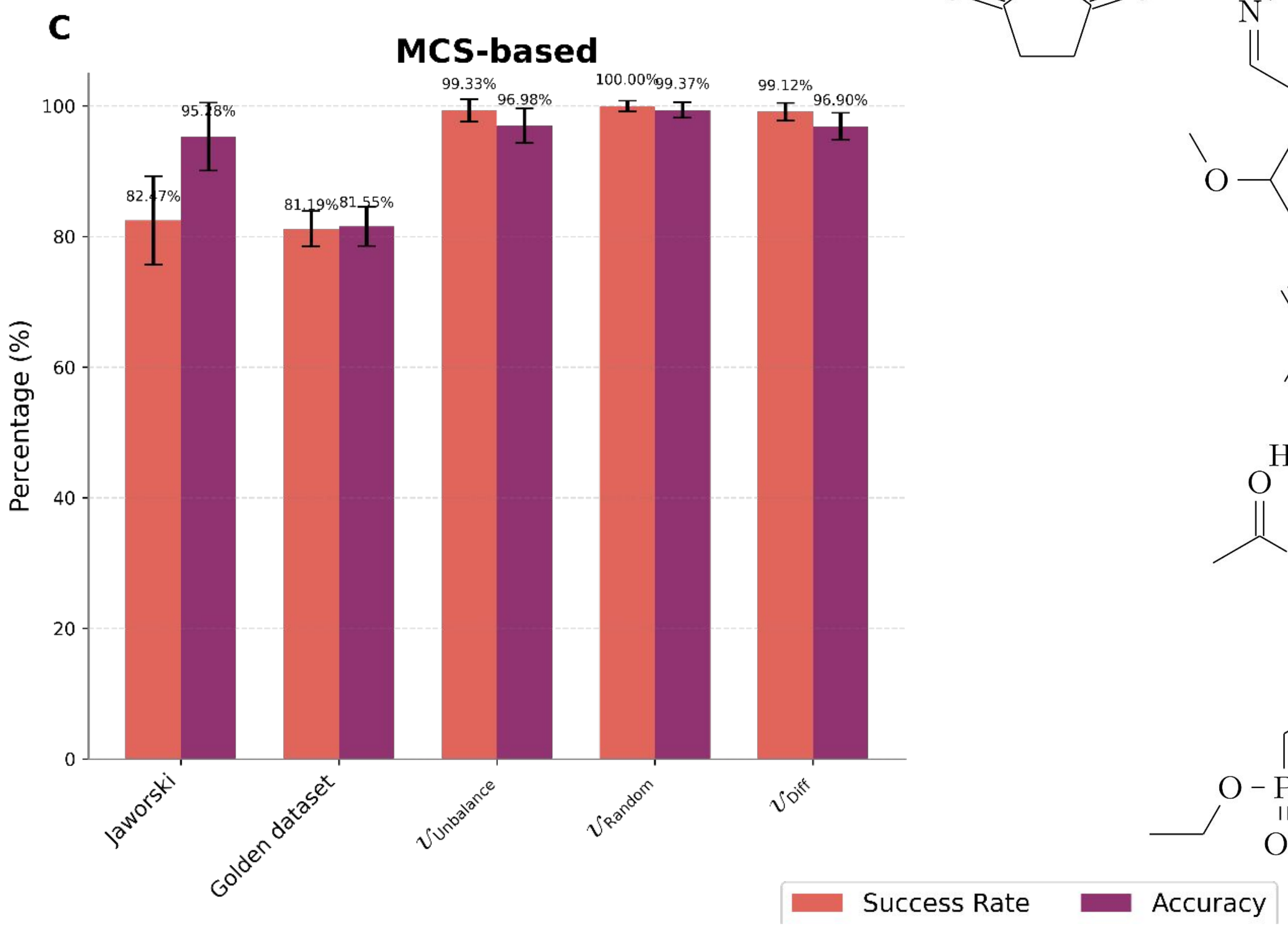
RESULT - DISCUSSION

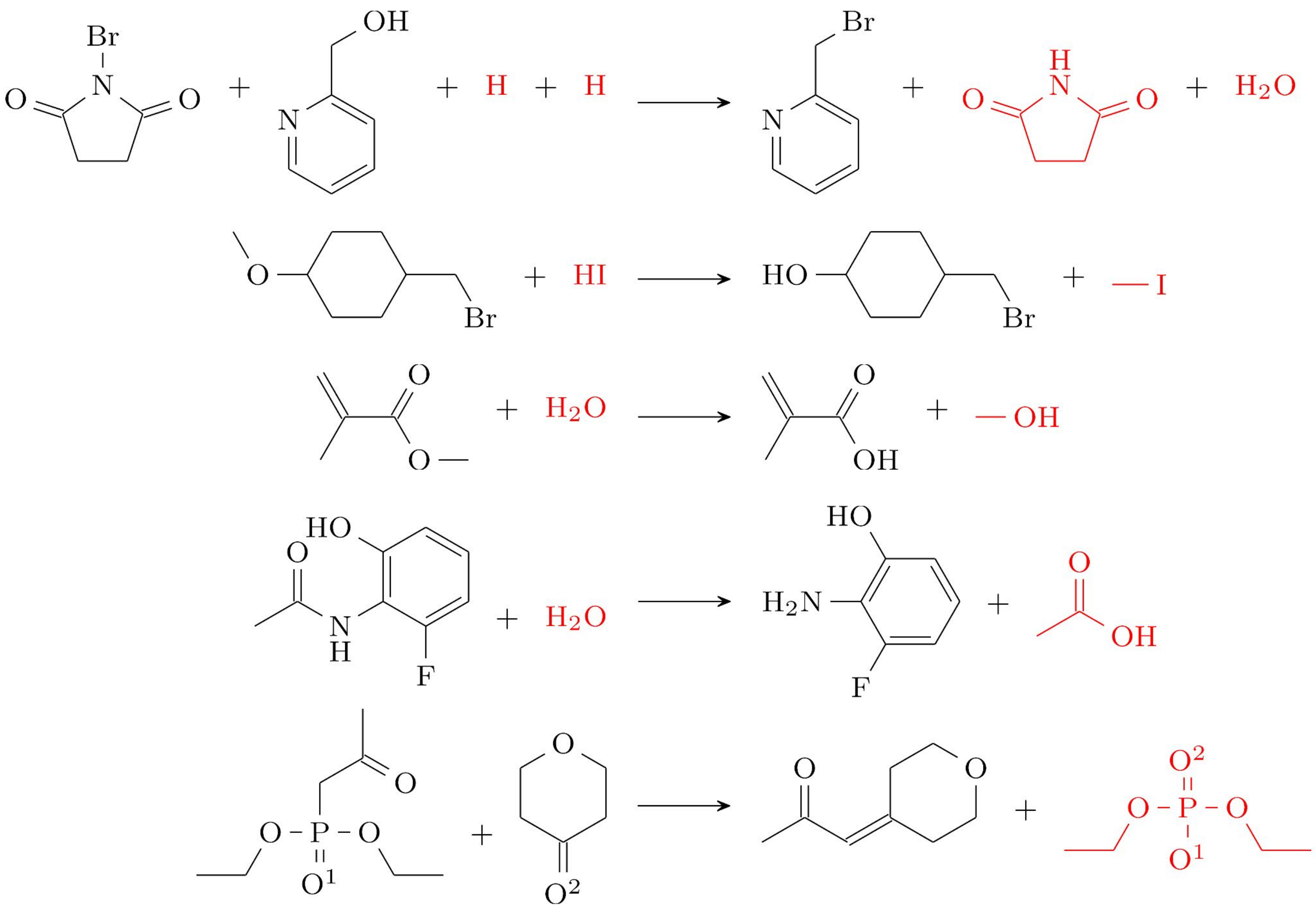
Rule-based approach





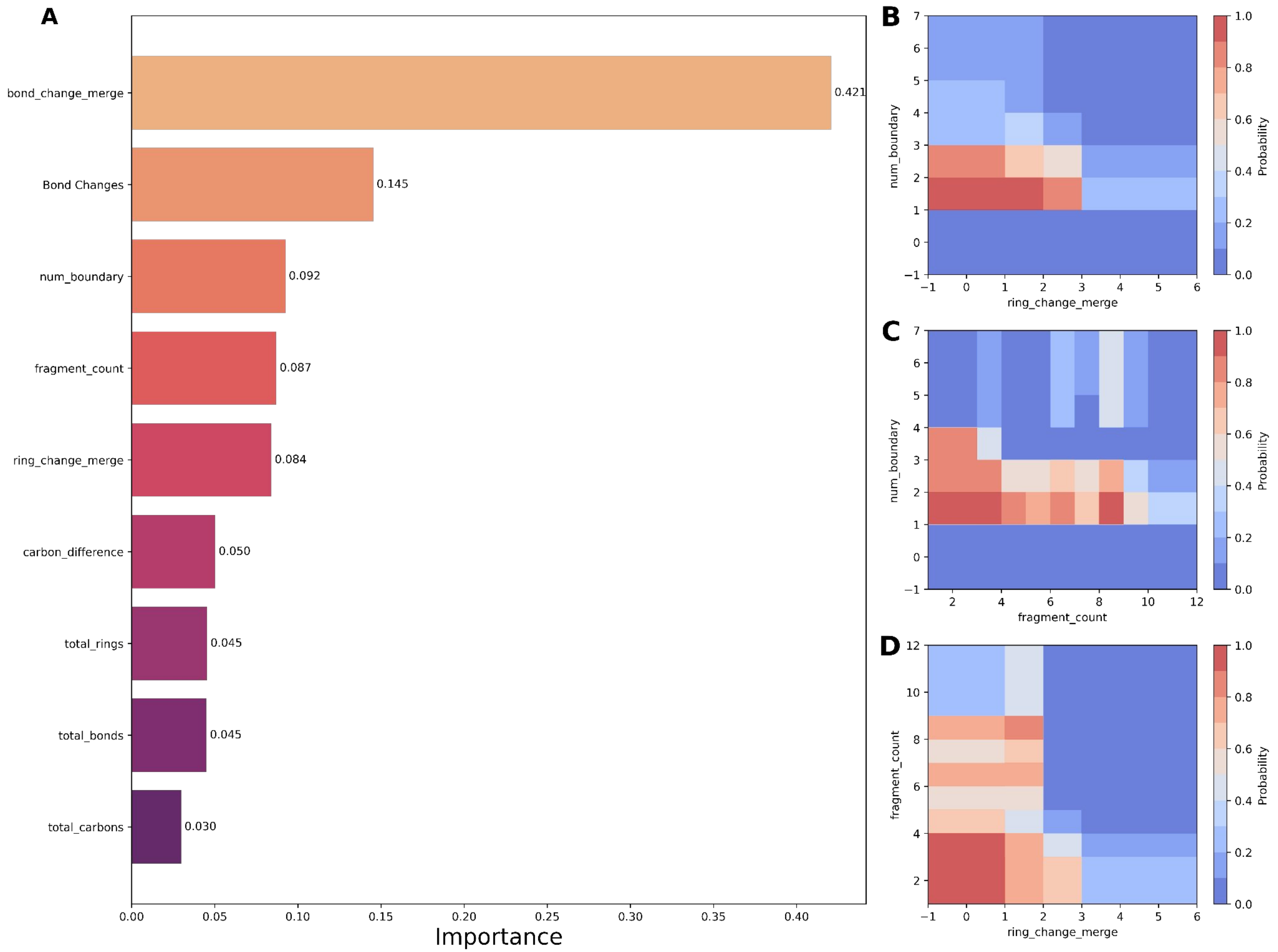
MCS-based approach





RESULT

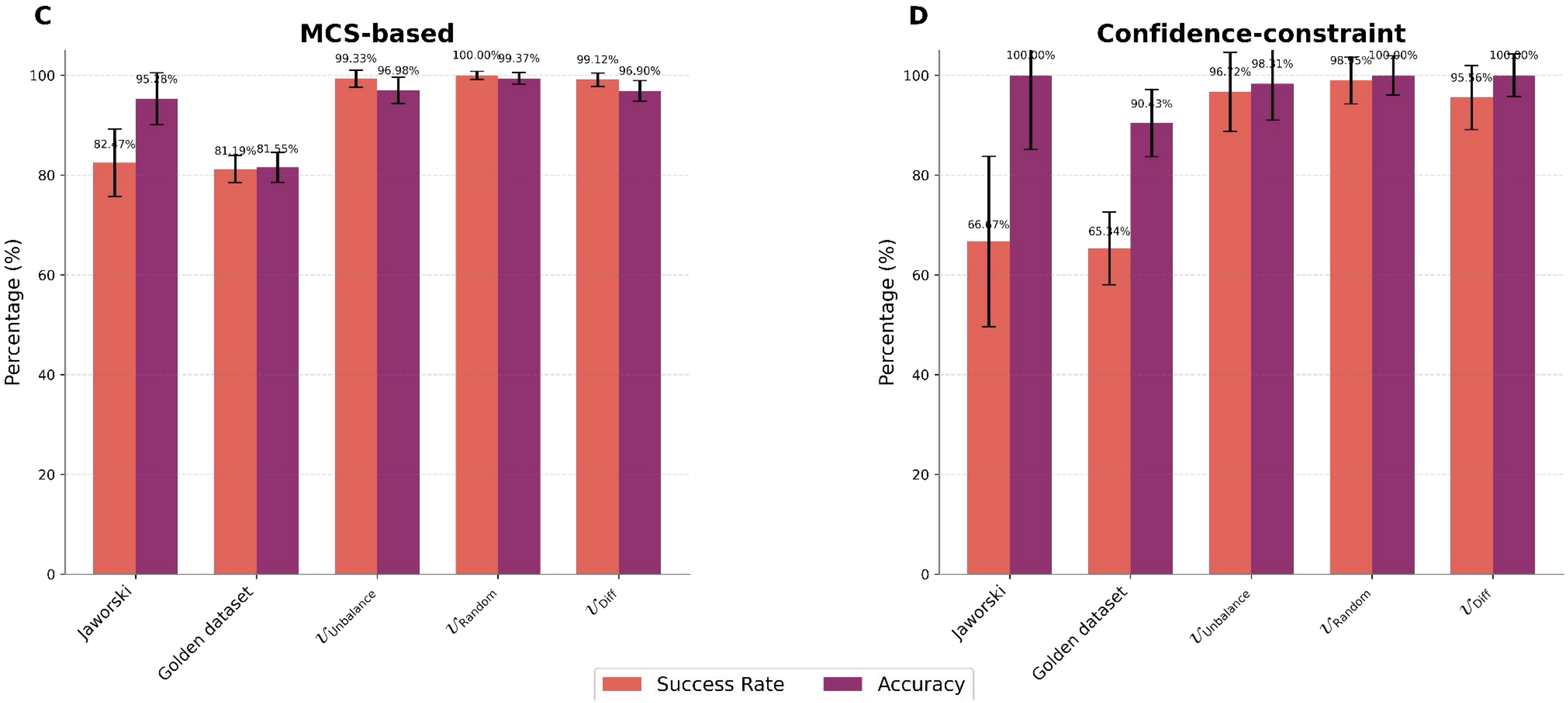
TACsy
Training Alliance for Computational systems chemistry



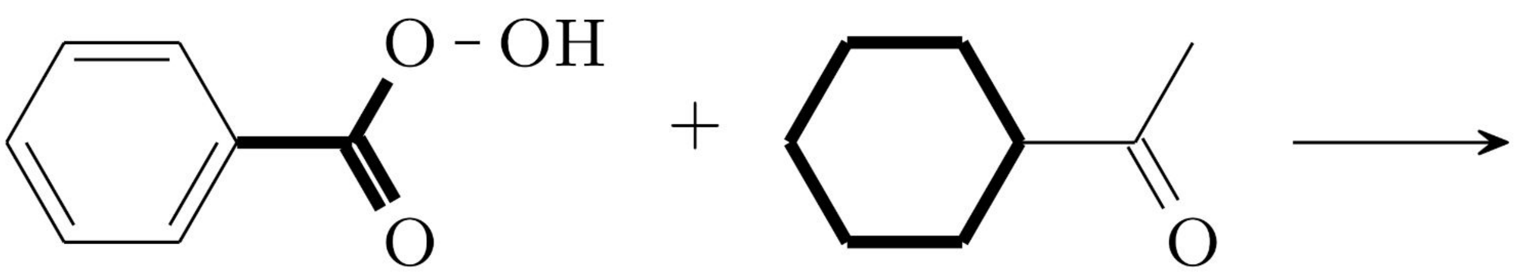
RESULT

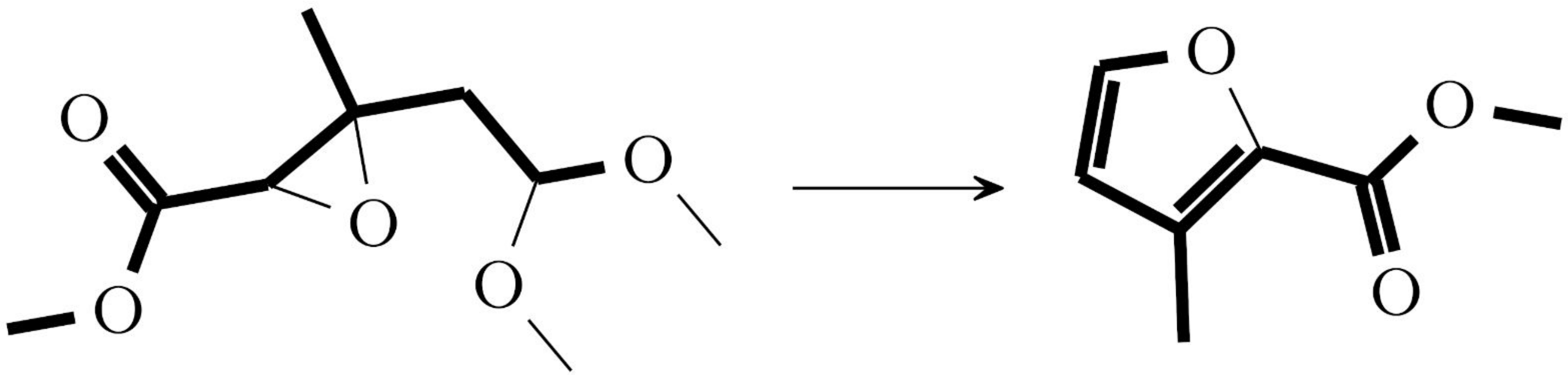
TACsy
Training Alliance for Computational systems chemistry

INTRODUCTION | METHOD | RESULT-DISCUSSION | CONCLUSION

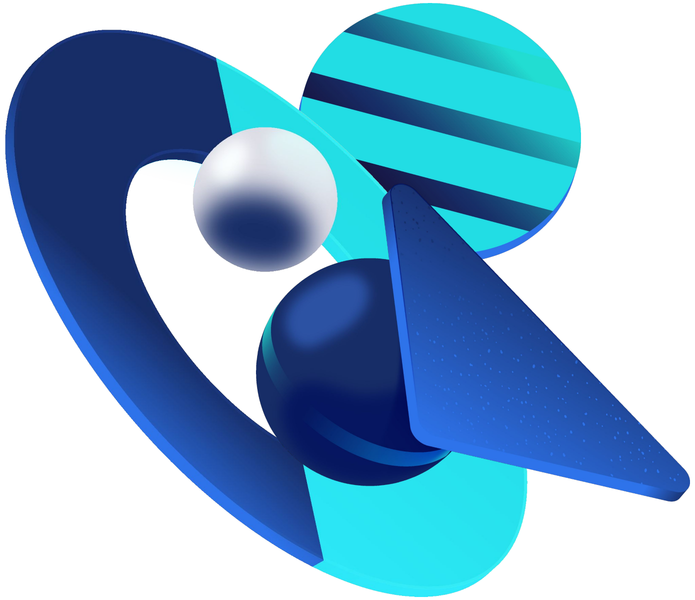


Limitations





04



CONCLUSION

TACsy
Training Alliance for Computational systems chemistry

Thank you for your attendance

MARIE CURIE ACTIONS

Founded by the European Union

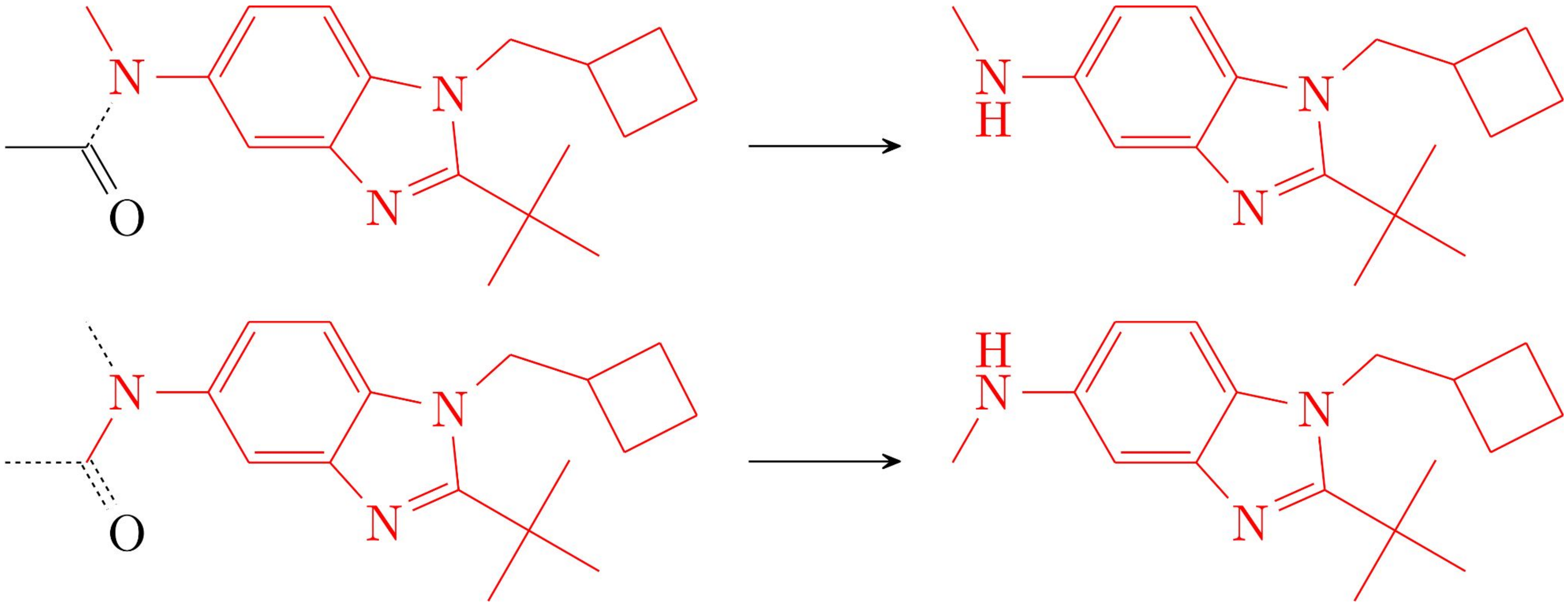
This project has received funding from the European Unions Horizon 2021 research and innovation programme under the Marie-Skłodowska-Curie grant agreement No 101072930

UNIVERSITÄT LEIPZIG

Appendix

RESULT

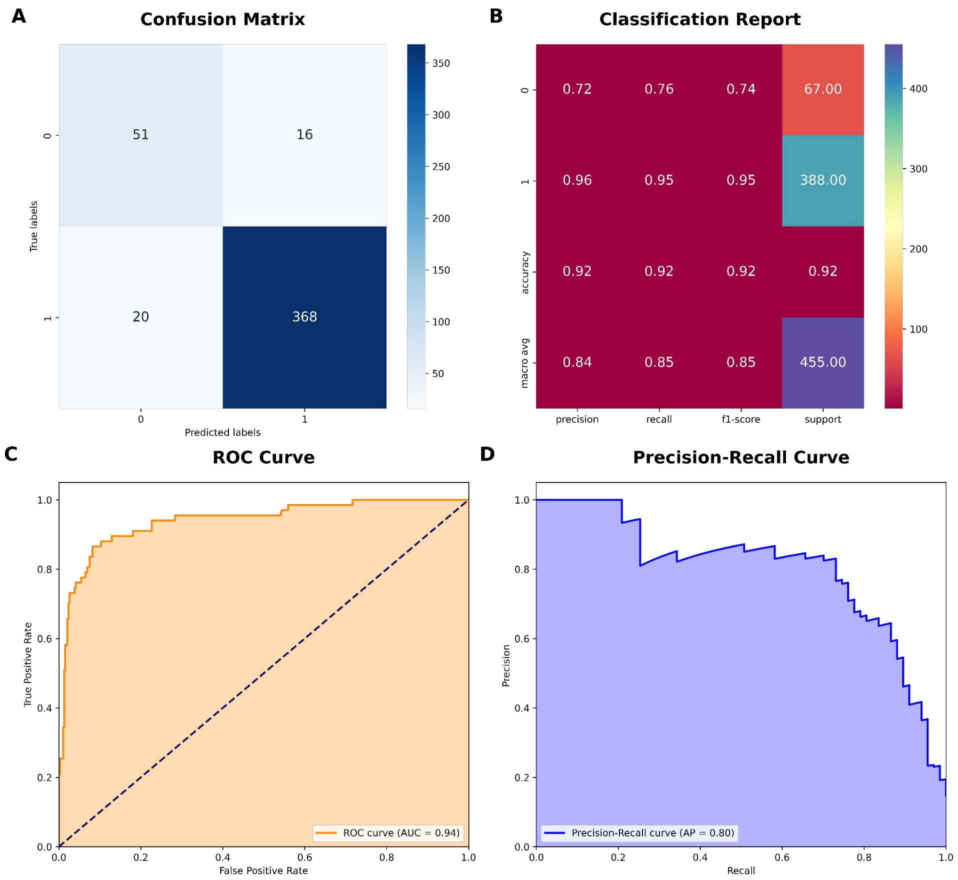
Equivariant Isomorphism

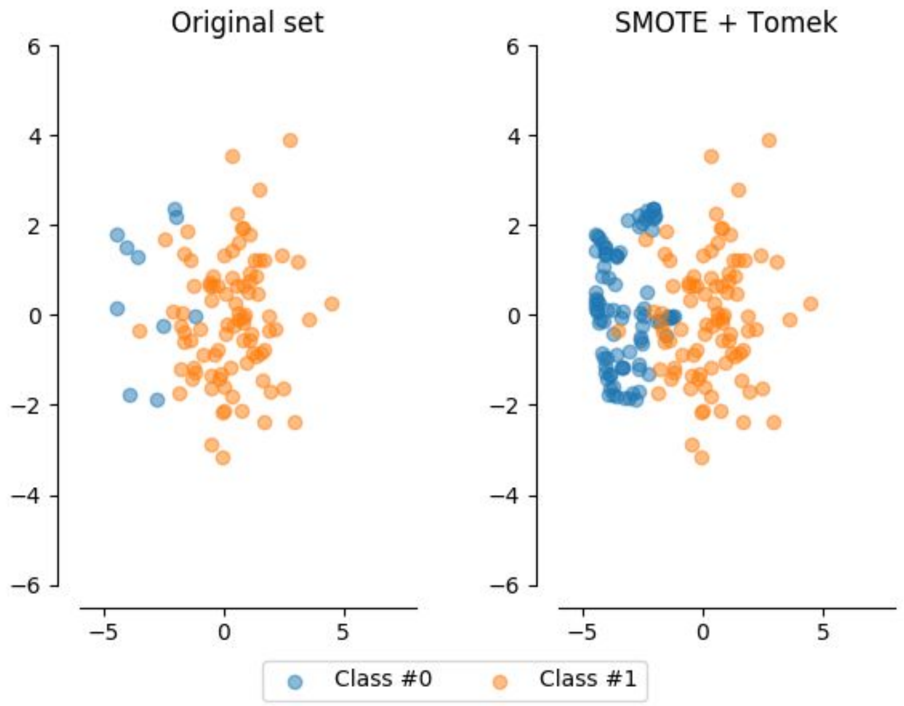


RESULT

TACsy
Training Alliance for Computational systems chemistry

Applicability Domain

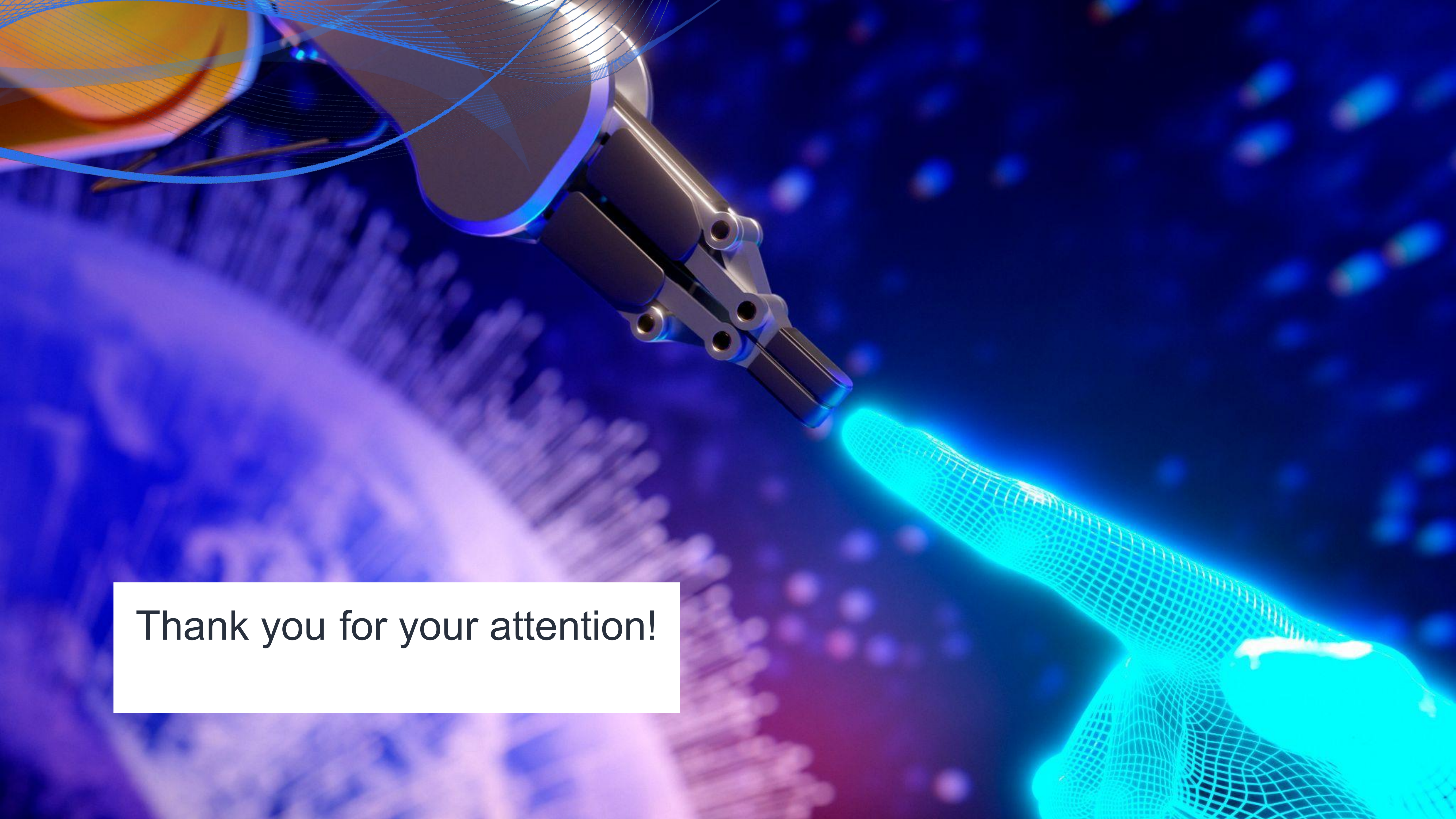




Imbalanced learn

dmlc XGBoost

1. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.
2. LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research, 18(17), 1-5.



Thank you for your attention!

INTRODUCTION

INTRODUCTION | **METHOD** | RESULT-DISCUSSION | CONCLUSION

INTRODUCTION | METHOD | **RESULT-DISCUSSION** | CONCLUSION

INTRODUCTION | METHOD | RESULT-DISCUSSION | **CONCLUSION**

Zhang, C., Arun, A., & Lapkin, A. (2023). Completing and balancing database excerpted chemical reactions with a hybrid mechanistic-machine learning approach.