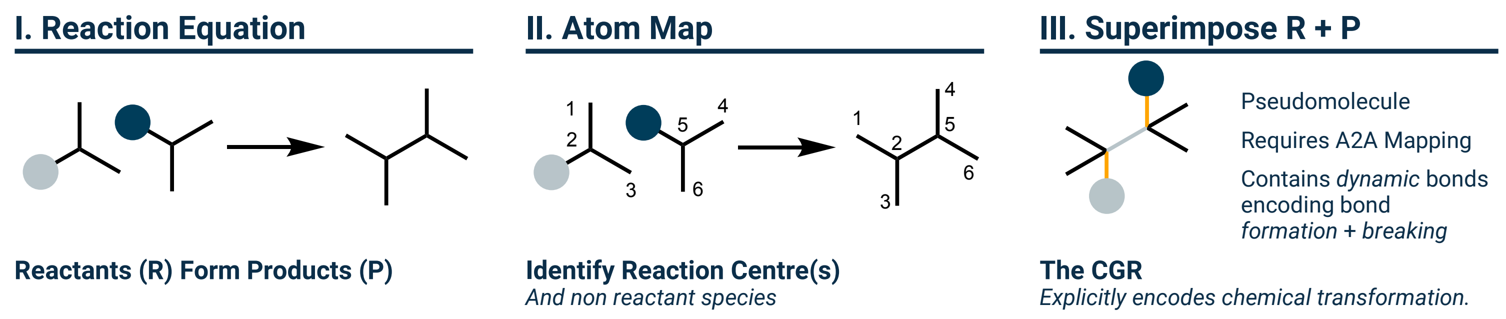
Predicting Reaction Conditions: A Data-Driven Perspective

Matt Ball[†‡], Dragos Horvath[†], Thierry Kogej[‡], Mikhail Kabeshov[‡] and Alexandre Varnek[†]

[†]Laboratory of Cheminformatics, University of Strasbourg, 67081 Strasbourg, France
[‡]Molecular AI, Discovery Sciences RD, AstraZeneca, 431 83 Gothenburg, Sweden

Introduction

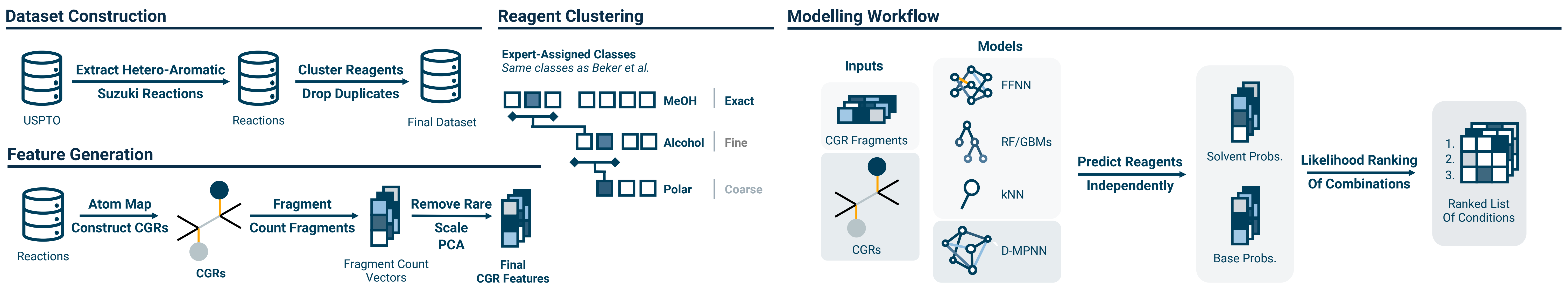
- Reaction conditions play a pivotal role in determining the outcomes of chemical syntheses, but despite this importance, there are still relatively few computational tools to predict optimal reaction conditions directly.
- Underlying problems with reaction data are well documented[1-3] but have underdiscussed implications for the **design** and **evaluation** of condition prediction models.
- These problems manifest themselves in **poor model performance**[4], where state-of-the-art approaches cannot significantly improve upon a literature popularity baseline.
- We suggest alternative approaches for the design and evaluation of condition prediction models and investigate the impact that **reaction representation** can have on existing model performance.


I. Reaction Equation
Reactants (R) Form Products (P)
II. Atom Map
1
2
3
4
5
6
Identify Reaction Centre(s)
And non reactant species
III. Superimpose R + P
Pseudomolecule
Requires A2A Mapping
Contains dynamic bonds encoding bond formation + breaking
The CGR
Explicitly encodes chemical transformation.

Constructing a CGR

The Impact Of Reaction Representation

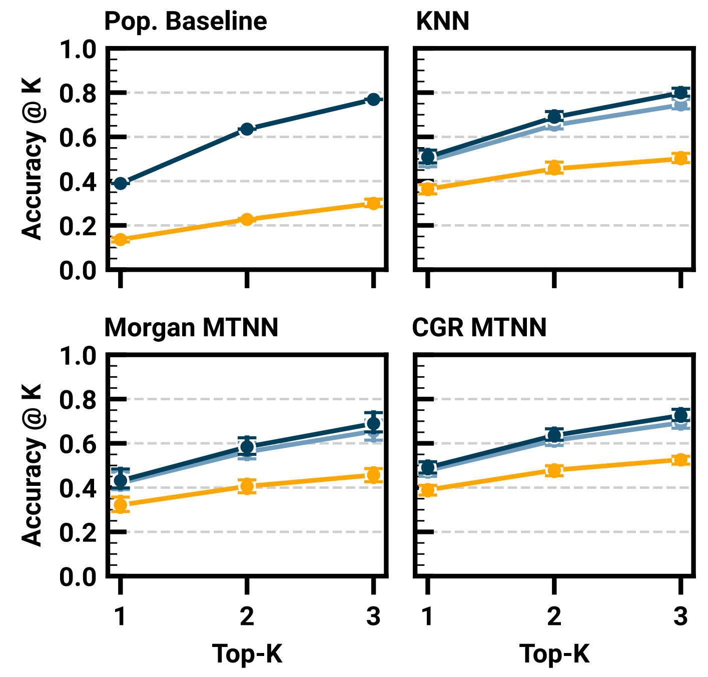
- It has previously been suggested that ML models **cannot significantly improve upon literature popularity baselines**, for a range of models and input representations[4].
- The author's best model, a Multi-Task neural network based on Morgan fingerprints, gave minor improvements compared to popularity when **predicting the expert-assigned class of solvent and base** for heteroaromatic Suzuki-Miyaura reactions.
- To investigate the impact that representation can have on model performance, we build and assess **Condensed Graph of Reaction**-based models on this dataset.



Dataset construction, feature generation and modelling workflow.

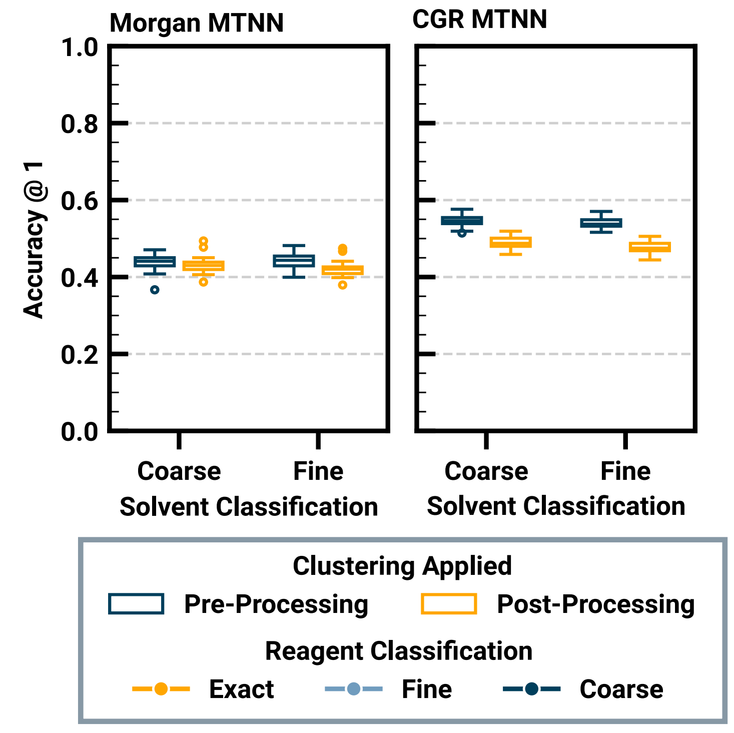
Clustering Impact on Top-K Accuracies

*'Exact' **Solvent** Predicted, Then Clustered.*



Clustering Ordering Matters

Solvent Top-1 Accuracies, Coloured By Clustering Application Time

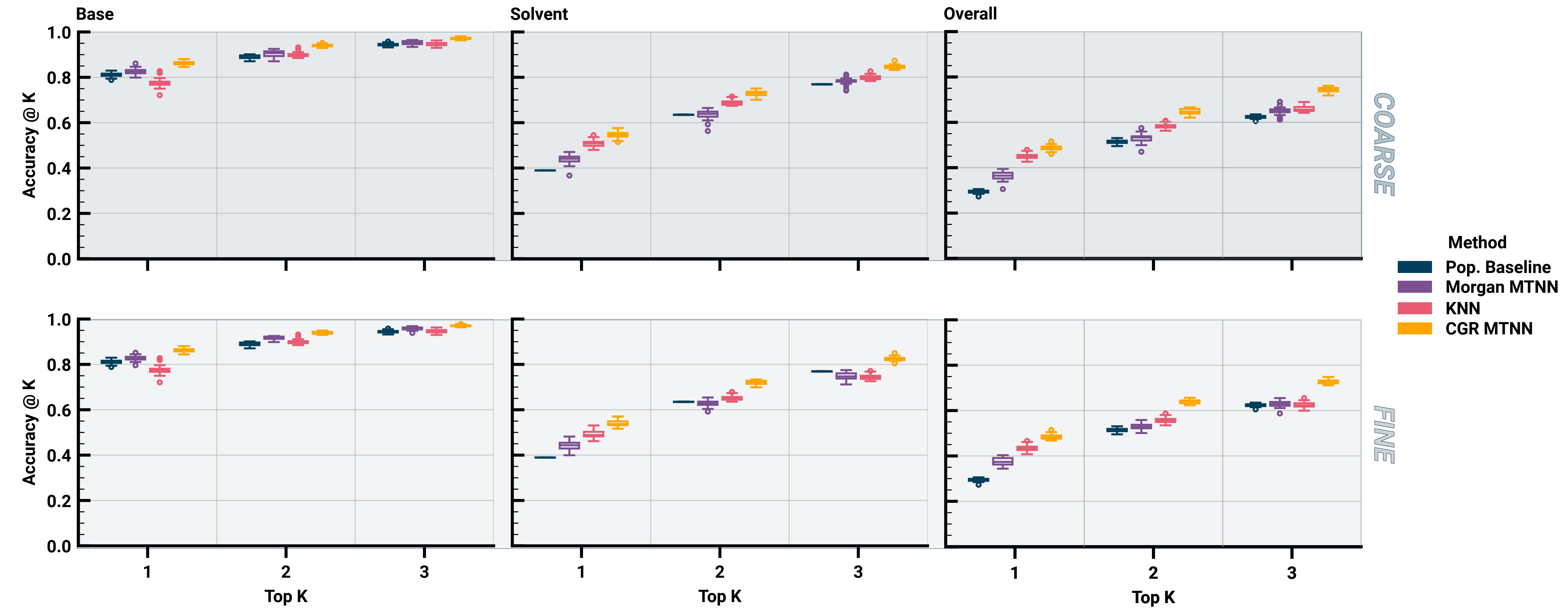


Chemically-Informed Condition Classes Improves Performance

- As expected, **a coarse-grained treatment of reaction conditions improves performance**, and represents a potential approach to **combat data sparsity** in large-scale condition prediction models.
- We see a noticeable difference in performance depending on *when* this categorisation is applied. With models **trained on the 'categorised' conditions performing better** than those trained on the 'exact' conditions and *then* applying categorisation to the outputs.

Top-K Accuracy Comparison For Selected Models

Solvent/Base = Independent Predictions; Overall = Likelihood Ranking of Combinations Applied



CGR-Based Representations Improve Upon A Challenging Literature Baseline

- **CGR-based** representations **improve performance significantly** above Morgan fingerprint and popularity baselines. This is particularly pronounced when considering the '**overall**' accuracy of predicting both solvent and base simultaneously.
- Even a simple similarity search based on these CGR-Fragments performs comparably to a more complex model based on Morgan fingerprints.

Conclusions

- Predictive models using literature data can surpass baseline performance with **appropriate reaction representations**.
- Further gains are possible through **improved input and output encoding** to address data biases and sparsity.
- Evaluation should go beyond **binary accuracy,** incorporating **expert knowledge** or **experimental validation.**
- Expert-defined reagent classes offer a promising strategy to mitigate sparsity, assuming **intra-class reactivity is consistent**.

Author Information

Matt Ball
1st Year PhD Fellow
MSCA-DN Project AiChemist
matthew.ball3@astrazeneca.com

Acknowledgements

This study was funded by the European Union's Horizon 2022 research and innovation program under the Marie Sklodowska-Curie Actions, grant agreement "Explainable AI for Molecules - AiChemist" No. 101120466. We would also like to thank Timur Madzhidov and Gilles Marcou for stimulating discussion.

References

1. Maloney, M. P. et al. Negative Data in Data Sets for Machine Learning Training. The Journal of Organic Chemistry 88, 5239–5241 (2023).
2. Strieth-Kalthoff, F. et al. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. Angewandte Chemie International Edition 61, (2022).
3. Varvara Voinarovska, Mikhail Kabeshov, Dmytro Dudenko, Genheden, S. & Tetko, I. V. When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges. Journal of Chemical Information and Modeling 64, 42–56 (2023).
4. Beker, W. et al. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. Journal of the American Chemical Society 144, 4819–4827 (2022).
5. Schilter, O. T., Baldassari, C., Laino, T. & Schwaller, P. Predicting solvents with the help of Artificial Intelligence. (2023) doi:https://doi.org/10.26434/chemrxiv-2023-hmml5.
6. Gao, H. et al. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. ACS Central Science 4, 1465–1476 (2018).
7. Wang, Z., Lin, K., Pei, J. & Lai, L. Reacon: a template- and cluster-based framework for reaction condition prediction. Chemical Science 16, 854–866 (2025).
8. Afonina, V. A. et al. Prediction of Optimal Conditions of Hydrogenation Reaction Using the Likelihood Ranking Approach. International Journal of Molecular Sciences 23, 248–248 (2021).