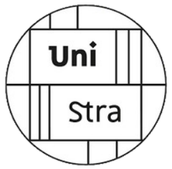


Predicting Organic Reaction Conditions: A Data-Driven Perspective

AiChemist Workshop

Matt Ball - 25th April 2025

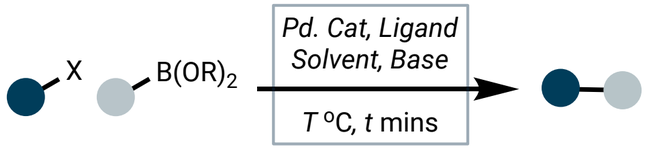
Introduction

What Are Reaction Conditions?

Key Points

- Chemical species or parameters that facilitate a chemical reaction occurring.
- From chemical species (reagents), to physical parameters.
- At its most fine-grained level, conditions encapsulates **all non-reactant variables** in a reaction.
- Optimal conditions can be either '**general**' (best over a range of reactants) or '**substrate-specific**' (best for a specific reactant pair).

What Are Reaction Conditions?



The components 'above the arrow'
which facilitate a chemical reaction

What Are Reaction Conditions Composed Of?

REAGENTS
CHEMICAL VARIABLES

Reacting species which **do not contribute a heavy atom** to the product.

Here:
- *Pd Cat.*
- *Ligand*
- *Solvent*
- *Base*

Categorical

PHYSICAL PARAMETERS
NON-CHEMICAL VARIABLES

Other non-chemical variables that influence a reaction.

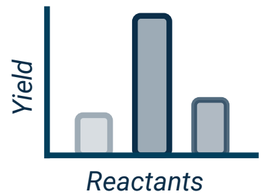
Here:
- *Temp.*
- *Time*

+ many more…

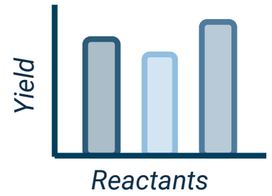
Continuous

What Are 'Optimal' Conditions?

Substrate-Specific:
Best for a single reactant pair



General:
Best across a range of reactants



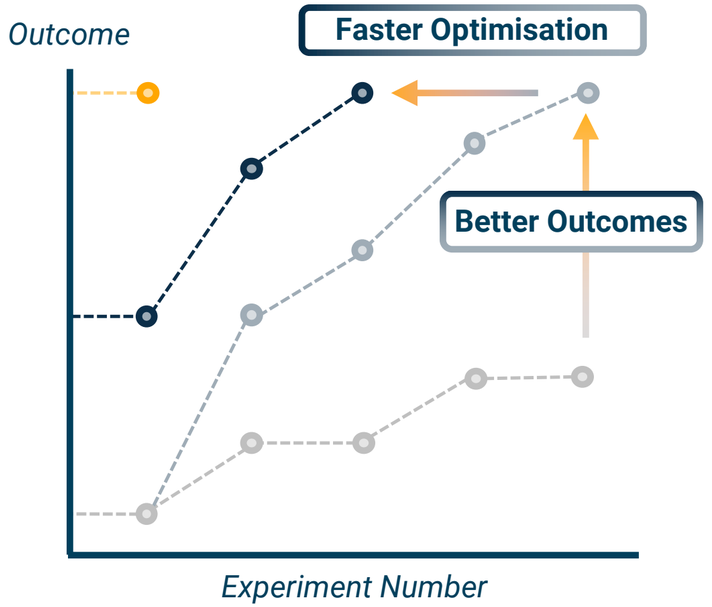
Introduction

Why Do We Care?

Key Points

- Reaction conditions have a **large impact** on the **success** of chemical transformations.
- Even **small** changes in conditions can lead to completely **different reactivity**.
- Predicting which conditions will lead to ‘successful’ reactions is therefore critically important in any chemical synthesis.

What Role Can ML Play?



- **Ideal** ML Model Condition Prediction
  Predict the best conditions, immediately
- ML-Guided **Initial** Condition Prediction + **BO**
  Improve starting points, fewer experiments required
- ML-Assisted Experiment Planning e.g. **BO**
  More informed experiment design
- No Computational Help e.g. DoE, OFAT
  Inefficient, can’t capture complex relationships

Modelling: Theory

How Can We Predict Optimal Reaction Conditions?

Key Points

- Modelling reaction outcomes typically requires both the reaction equation and conditions as input.
- To predict the **best** conditions, we can either:
  - Enumerate all condition combinations, **predict the outcome under each set of conditions**, and pick the conditions leading to the desired outcome.
  - **Directly predict the conditions**, using the reaction equation alone.

Mathematical Formulation

TRADITIONAL REACTION MODELLING

$\boldsymbol{r}$: Reaction
$\boldsymbol{c}$: Conditions
$\boldsymbol{y}$: Outcome

- Feasibility
- Yield
- k

$$\hat{\boldsymbol{y}} = f(\boldsymbol{r}, \boldsymbol{c})$$

CONDITION PREDICTION

VIRTUAL CONDITION SCREENING

Predict outcome for *all* conditions

Select the ***best*** performing conditions

$$\boldsymbol{c}_{\mathbf{opt}} = \underset{\boldsymbol{c} \in C}{\operatorname{argmax}}\, f(\boldsymbol{r}, \boldsymbol{c})$$

DIRECT PREDICTION

Directly predict conditions that give ***desired*** outcome

$$\hat{\boldsymbol{c}} = f^{-1}(\boldsymbol{r}, \boldsymbol{y})$$

Modelling: Challenges

Literature Data Isn't Great

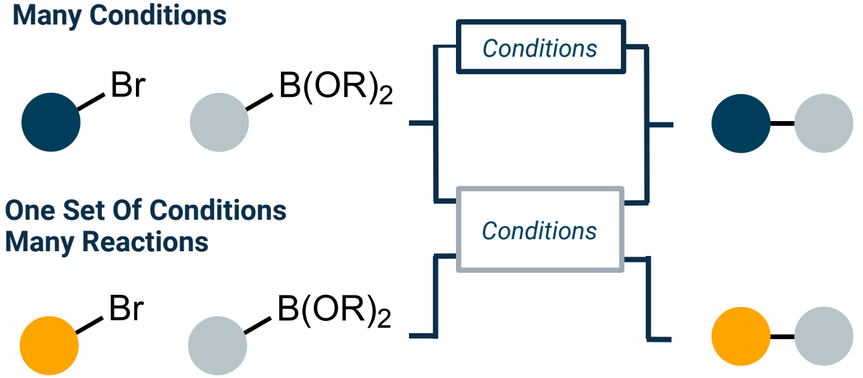
Key Points

- The *many-to-many* nature of condition prediction makes reaction-condition space **combinatorially large**
- Literature data suffers from a number of problems:
  - **Reporting Bias:** tendency to only report successful reactions
  - **Selection Bias:** tendency to rely on established and available routines
  - **Experimental Noise:** variance in reaction outcomes for the same reaction protocol
- The first two points lead to a lack of negative data, and **data sparsity**

Reactions can proceed under many conditions, but only a small number are reported.

Many-To-Many Correspondence

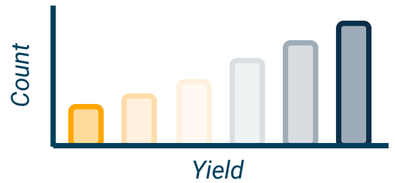
One Reaction
Many Conditions



One Set Of Conditions
Many Reactions

This Complicates Model Design + Evaluation
And causes *data sparsity.*

Lack Of Negative Data



Biases in reaction data favour successful reactions.

Leads to **imbalanced** datasets

1. F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, Angewandte Chemie International Edition, 2022, 61, e202204647.

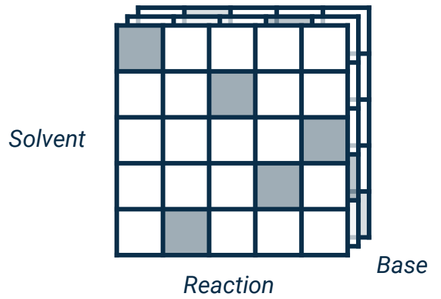
Modelling: Challenges

Data Sparsity

Key Points

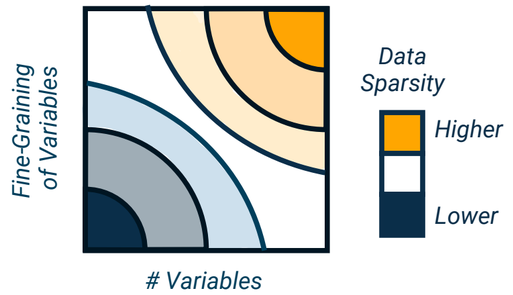
- The extent of data sparsity is dependent on the number of variables that we want to model.
- More variables → condition space is larger → data sparsity gets worse
- Data sparsity will also get worse when considering higher fidelity variables.
- Other confounding variables too, like personal preference and availability of reagents in the lab complicate prediction of conditions.

Data Sparsity



Most Reactions Only Appear Under A Single Set Of Conditions

Making it difficult for models to learn trends in both reactant and condition reactivity.



Sparsity Becomes Worse When Modelling More Condition Variables

Therefore, models must balance the scope and granularity of their predictions.

1. 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).

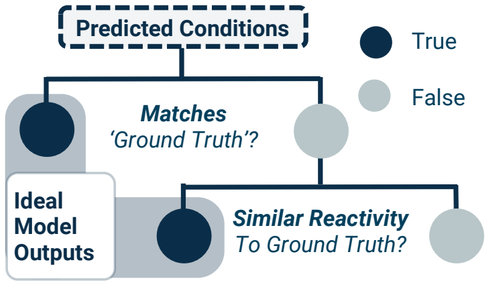
Modelling: Challenges

How Should We Evaluate Predictions?

Key Points

- Models are typically evaluated using **top-k accuracy,** but this doesn't tell the full story.
- Gold standard: **experimental validation**
- *In Silico?*
  - Expert-assigned reagent classes
    Requires selection of reagent classes.
  - Condition similarity score
    Requires a meaningful encoding for the reaction conditions.

Example Evaluation Workflow



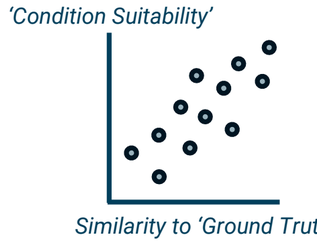
Assessing Prediction Suitability

QUALITATIVELY

*Would a **synthetic chemist** be **willing to try** these conditions?*

QUANTITATIVELY

*What is the **similarity** between the **predicted** condition and the **'ground truth'** condition?*



Assessing Condition Similarity

EXPERT-ASSIGNED REAGENT CLASSES

Does the predicted reagent fall into the same 'class' as the 'ground truth' reagent?

Reagent Classes Follow Known Reactivity Trends

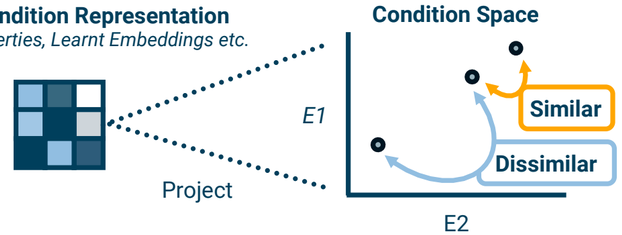
I.e. all reagents within the same class should lead to similar outcomes

'FEATURISING' CONDITIONS

How similar are the representations of the predicted reagent/condition compared to the 'ground truth'?

Numeric Condition Representation

Physical Properties, Learnt Embeddings etc.



1. 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).
2. Gao, H. et al. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. ACS Central Science 4, 1465–1476 (2018).
3. 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).

The Impact Of Data Problems

Models Can’t Outperform Literature Popularity

Key Points

- Beker et al. have previously suggested that **models can’t significantly outperform popularity baselines,** using a case study on heteroaromatic Suzuki-Miyaura couplings.
- Tested a range of representations and model types
- These models couldn’t improve upon simply choosing the most popular conditions from the literature

Journal of the American Chemical Society > Vol 144/Issue 11 > Article

Open Access

Cite Share Jump to Expand

ARTICLE | March 8, 2022

Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling

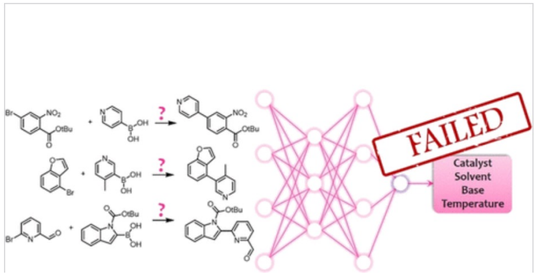
Click to copy article link

Wiktor Beker, Rafał Roszak, Agnieszka Wołos, Nicholas H. Angello, Vandana Rathore, Martin D. Burke*, and Bartosz A. Grzybowski*

Open PDF | Supporting Information (1) | open URL

Abstract

Applications of machine learning (ML) to synthetic chemistry rely on the assumption that large numbers of literature-reported examples should enable construction of accurate and predictive models of chemical reactivity. This paper demonstrates that abundance of carefully curated literature data may be insufficient for this purpose. Using an example of Suzuki–Miyaura coupling with heterocyclic building blocks—and a carefully selected database of >10,000



1. *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).*

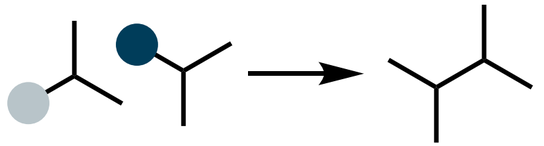
Case Study: Representation

Can We Improve On Literature Popularity?

Key Points

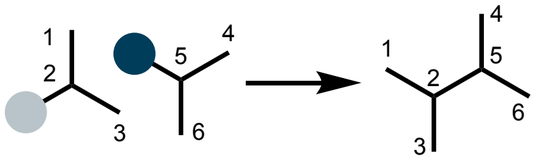
- We want to investigate if a Condensed Graph of Reaction representations can improve model performance, despite underlying data problems.
- Specifically, can alternative reaction representations improve model performance?
- CGR-Based methods have shown strong performance in the prediction of other reaction properties:
  - Activation Energies
  - Reaction Rates

I. Reaction Equation



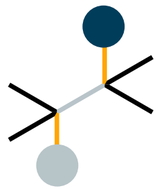
Reactants (R) Form Products (P)

II. Atom Map



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.

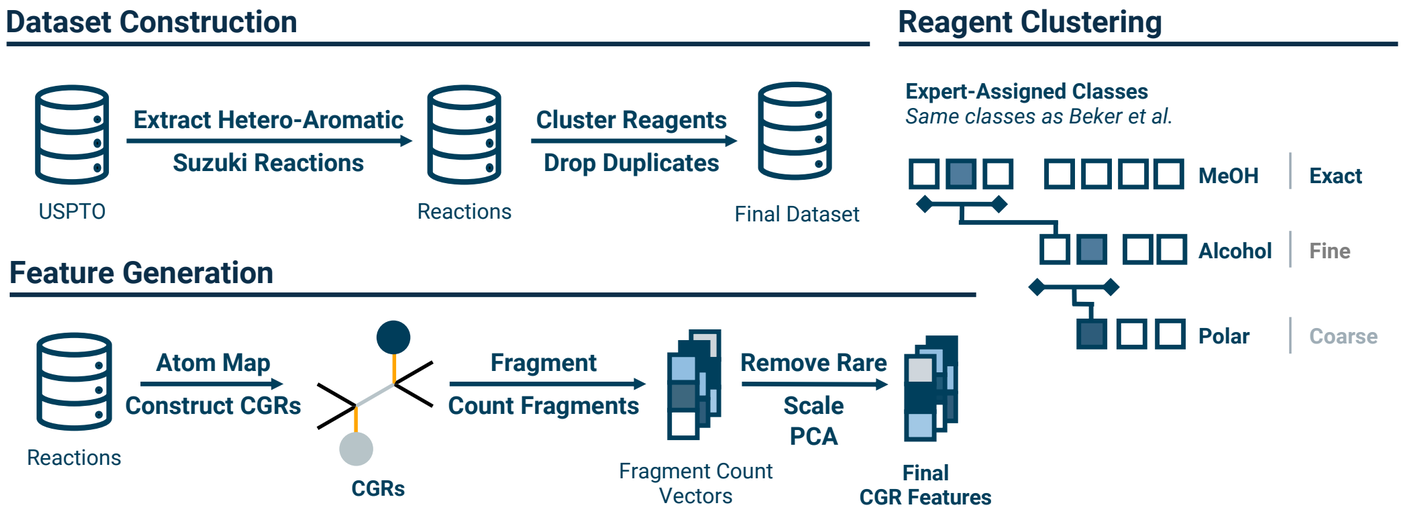
1. *Varnek, A., Fourches, D., Hoonakker, F. & Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 19, 693–703 (2005).*

Case Study: Representation

Dataset Construction

Key Points

- Extract of reactions from USPTO, followed by categorization of solvent and bases into their classes.
- 2x Multiclass-classification tasks:
  - Base: 7 Classes
  - Solvent: 13 Classes (‘Fine’-Grained) or 6 Classes (‘Coarse’-Grained)
- Create **CGR fragment features** (or just **CGRs** themselves for use with ChemProp).



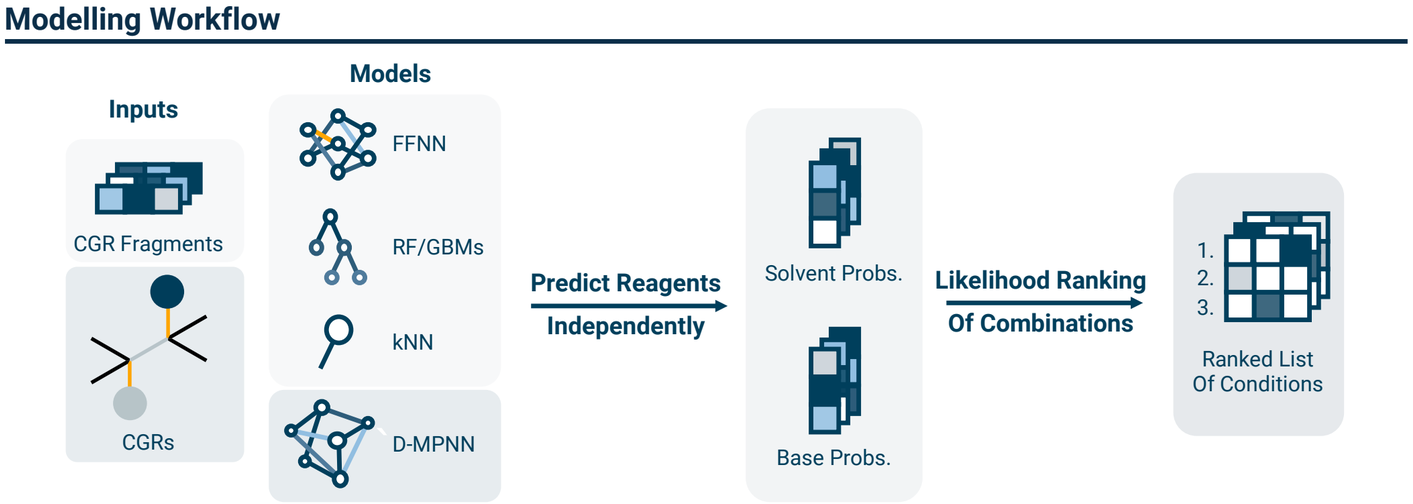
1. *Beker, W. et al. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. Journal of the American Chemical Society 144, 4819–4827 (2022).*
2. *Heid, E. et al. Chemprop: A Machine Learning Package for Chemical Property Prediction. Journal of Chemical Information and Modeling 64, 9–17 (2023).*

Case Study: Representation

Modelling Workflow

Key Points

- CGRs or CGR Fragments as input
- Predict Solvent and Base **independently**
- Use the '**Likelihood Ranking**' approach to combine independent predictions into a combined prediction of both base and solvent.



1. *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).*

Case Study

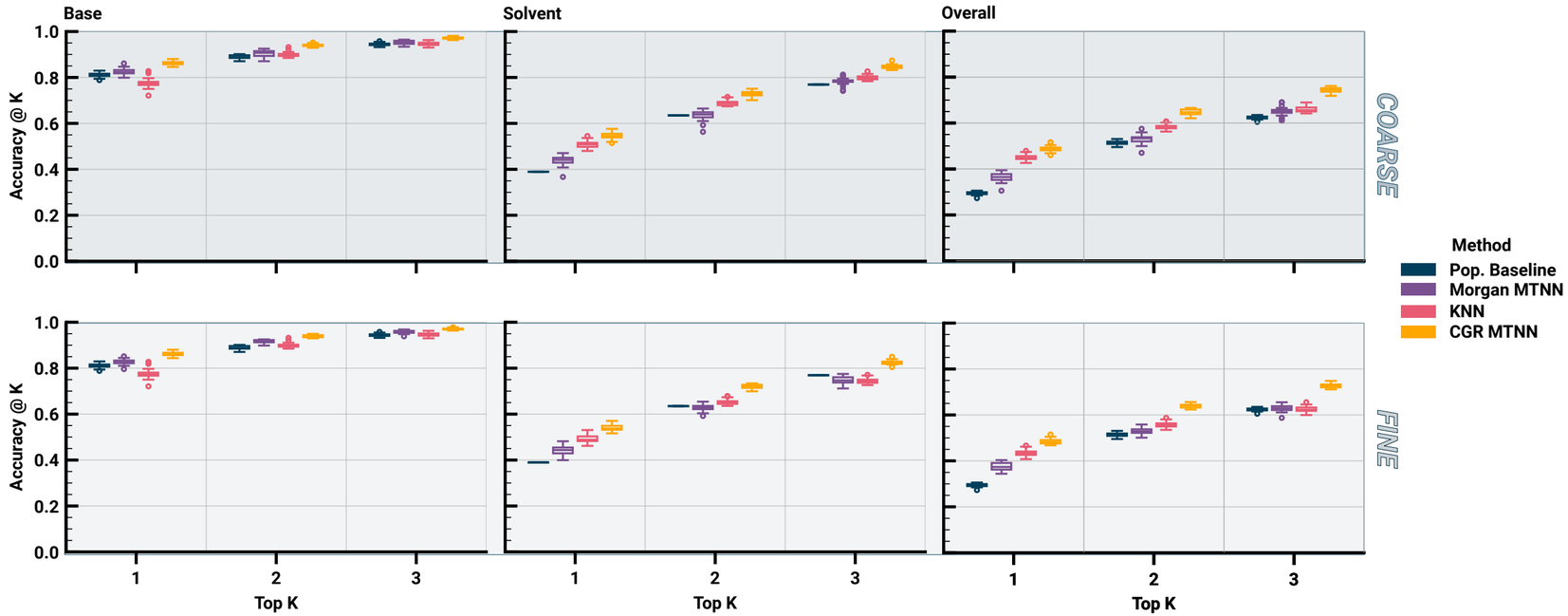
Results: The Impact Of Representation

Key Points

- CGRs **can improve** upon the performance of Morgan fingerprint-based models.
- Even **kNN** with CGRs performs comparably, or even better than, the MorganFP-based model.
- Whilst they do outperform literature baselines, there is **still some way to go** in terms of performance – particularly for solvent.

Top-K Accuracy Comparison For Selected Models

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



Case Study

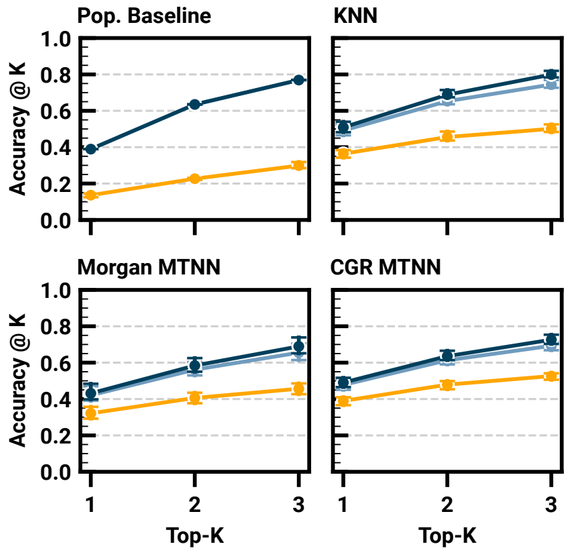
Results: The Impact Of Condition Classes

Key Points

- We can also demonstrate the impact of ‘clustering’ conditions into classes.
- As one might expect, reducing the number of classes dramatically improves model performance.
- By performing this clustering in pre-processing we can improve results over performing this in post-processing.

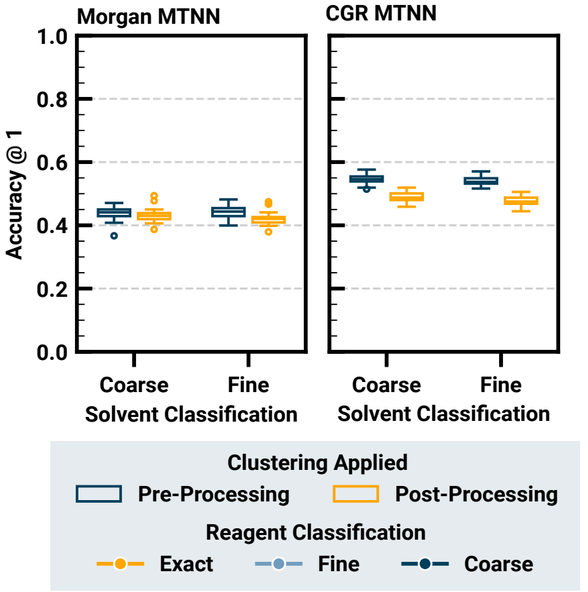
Clustering Impact on Top-K Accuracies

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



Clustering Ordering Matters

Solvent Top-1 Accuracies, Coloured By Clustering Application Time



Finishing Up

Conclusions And Takeaways

Conclusions

- Understanding the role that models trained on literature data can have on synthesis.
  - Large-scale ‘global’ models: can’t **directly** predict *exact optimal* conditions but can **suggest useful starting points** for BO and HTE.
  - Small-scale ‘local’ models: underlying data quality higher – can predict *optimal* conditions.
- Issues with literature data necessitate countermeasures, like creation of **condition classes** or **featurization of conditions**, to mitigate data sparsity.
  - This is particularly important if we want to model **more variables**.
- We suggest that models **can** outperform literature popularity baselines, provided they use the **correct representation**.
- The **clustering of similar conditions** can also improve performance, and represents an important way to mitigate **data sparsity**, provided the reactivity of conditions within a cluster is consistent.

Future Work

- Develop a method to automatically assign conditions to clusters based on their reactivity
- Use this representation (or derivatives of this) as a target for new models

Thank you for listening

Any Questions?