# Reaction Condition Prediction: A Data-Driven Perspective: Supplementary Material

Matt Balla, Dragos Horvath, Thierry Kogej, Mikhail Kabeshov and Alexandre Varnek

e-mail:varnek@unistra.fr

Additional explanations of the Solvent/Base clusters. Additional plots and metrics for the trained models.

## Solvent/Base Classes

Following the lead of Beker et al., the 'fine-grained' solvent classification consisted of 13 distinct classes:

- Alcohols
- Alcohols/Aromatics
- Alcohols/Aromatics/Water
- Alcohols/Water
- Amides
- Amides/Water
- Aromatics
- Aromatics/Water
- Ethers

- Ether/Water
- Low Boiling Point/Water (e.g. MeCN and Water)
- Water
- Other

While the 'coarse-grained' solvent classification contained 6 distinct classes:

- Alcohols/Aromatics (consisting of the Aromatics/Water, Alcohols/Aromatics and Alcohols/Aromatics/Water classes from the 'fine-grained' solvent classification)
- Aromatics
- Ethers
- Ether/Water
- Polar (consisting of the Alcohols, Low Boiling Point/Water, Alcohols/Water, Amides/Water, Water and Amide classes from the 'fine-grained' solvent classification)
- Other

The bases were split into 7 different classes:

- Acetates
- Amines
- Carbonates
- Fluorides
- Hydroxides
- Phosphates
- Other (e.g. Alkoxides)

# Modelling Overview

Additional diagrams describing the modelling workflow, explaining CGRs and explaining Likelihood Ranking.

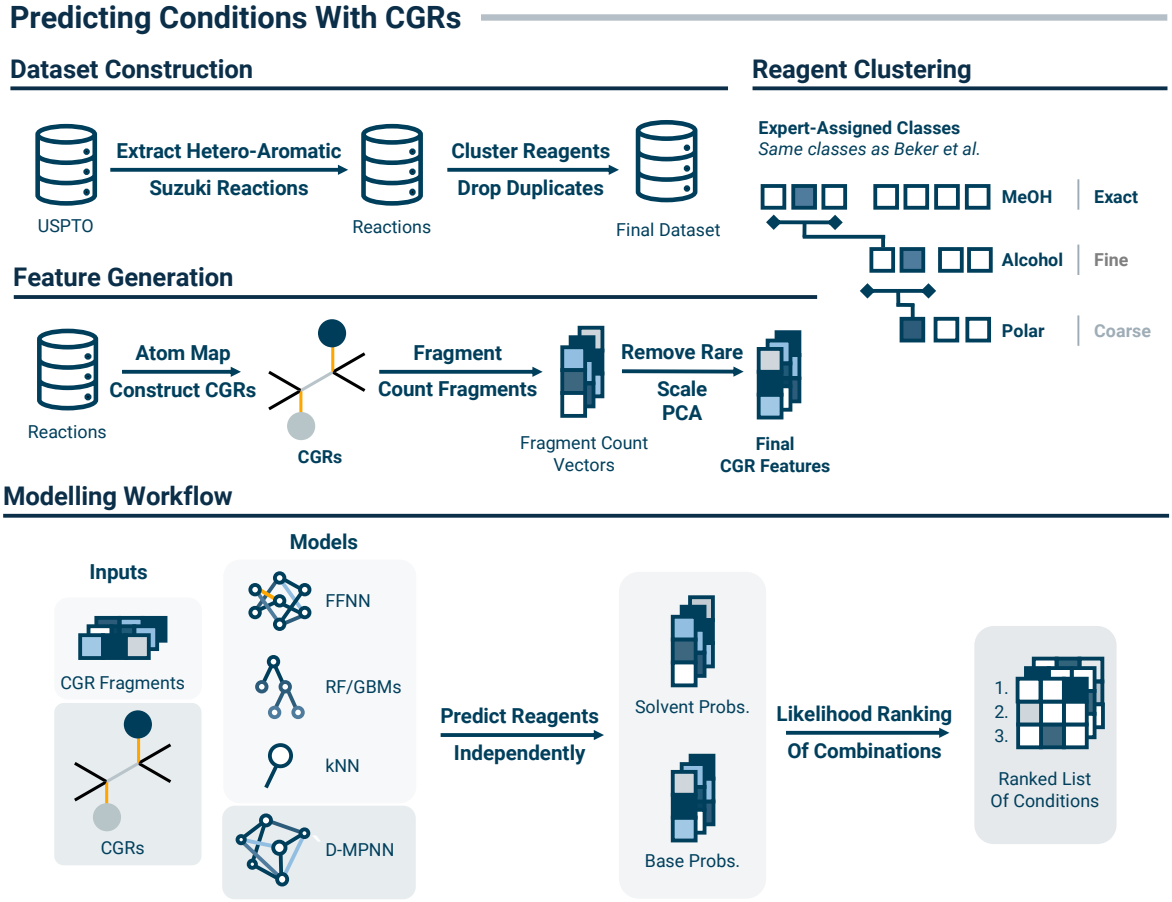


Figure S1: The workflow for the CGR condition prediction case study and an explanation of a CGR. Our workflow differs from Beker et al. in the choice of representation, where we use a different representation, the CGR, which that work did not examine.[1]

# Additional Modelling Figures and Metrics

Additional figures from the models trained in the Case Study.

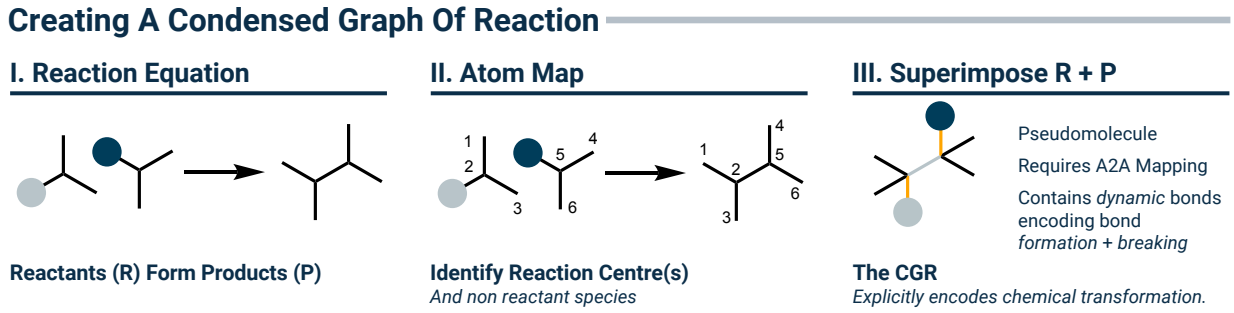


Figure S2: The creation of a CGR. The key advantage of a CGR, over other reaction representations such as Morgan fingerprints, is that it explicitly encodes the chemical transformation occurring in the reaction, contained within the 'dynamic bonds'.

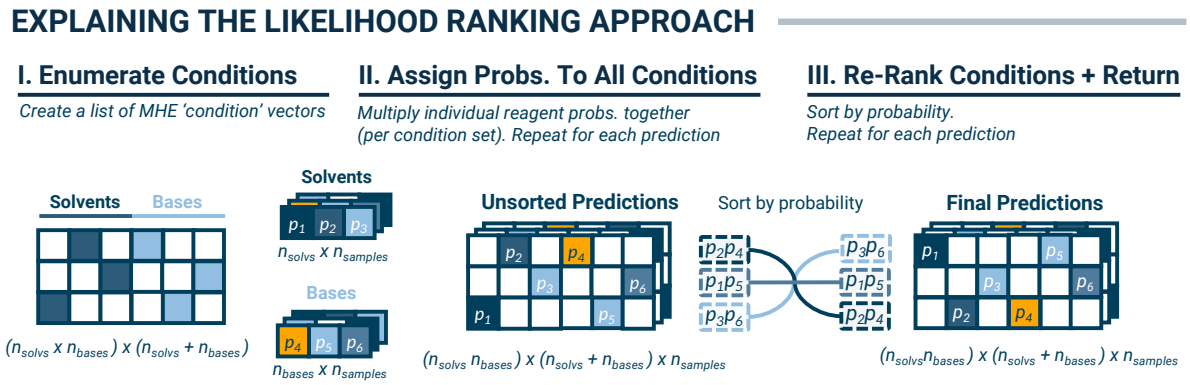


Figure S3: Explaining likelihood ranking. In the case where all tasks are multi-class classification (which is the case in our case study), there are $n_i$ possible options per reagent type, where $i$ is the number of classes for that reagent type. This means that there are $n_{solvs} \times n_{bases}$ total combinations (therefore there are $6 \times 7 = 42$ for 'coarse-grained' solvent classification, and $13 \times 7 = 91$ for 'fine-grained' solvent classification).

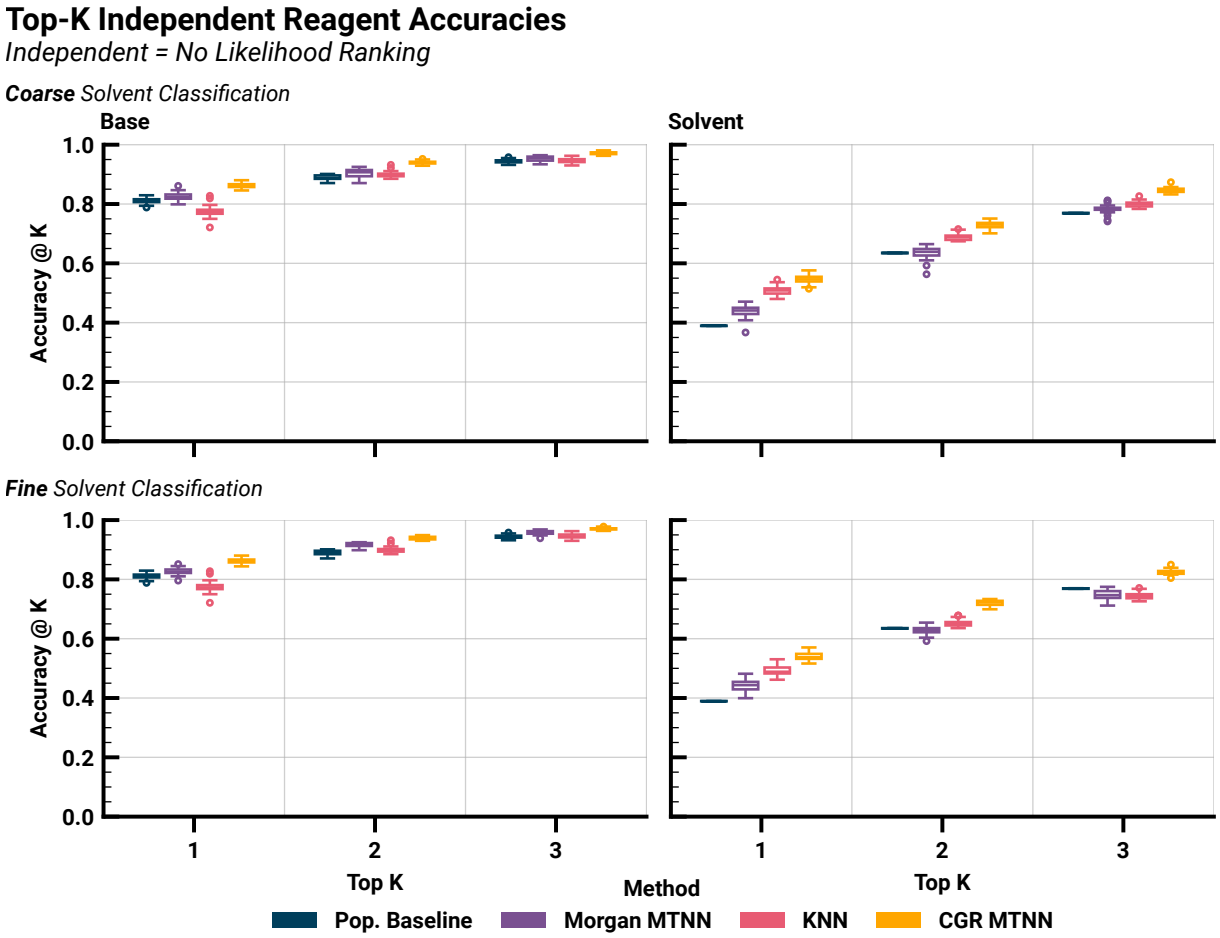


Figure S4: Box plots of the distribution of top-k accuracies reported for the predictions of the solvent and base *independently*, i.e. no likelihood ranking was performed. The CGR multi-task network is highlighted in yellow, and outperforms the literature baseline, as well as similarity and best existing ML method (tested by Beker et al.).[1]

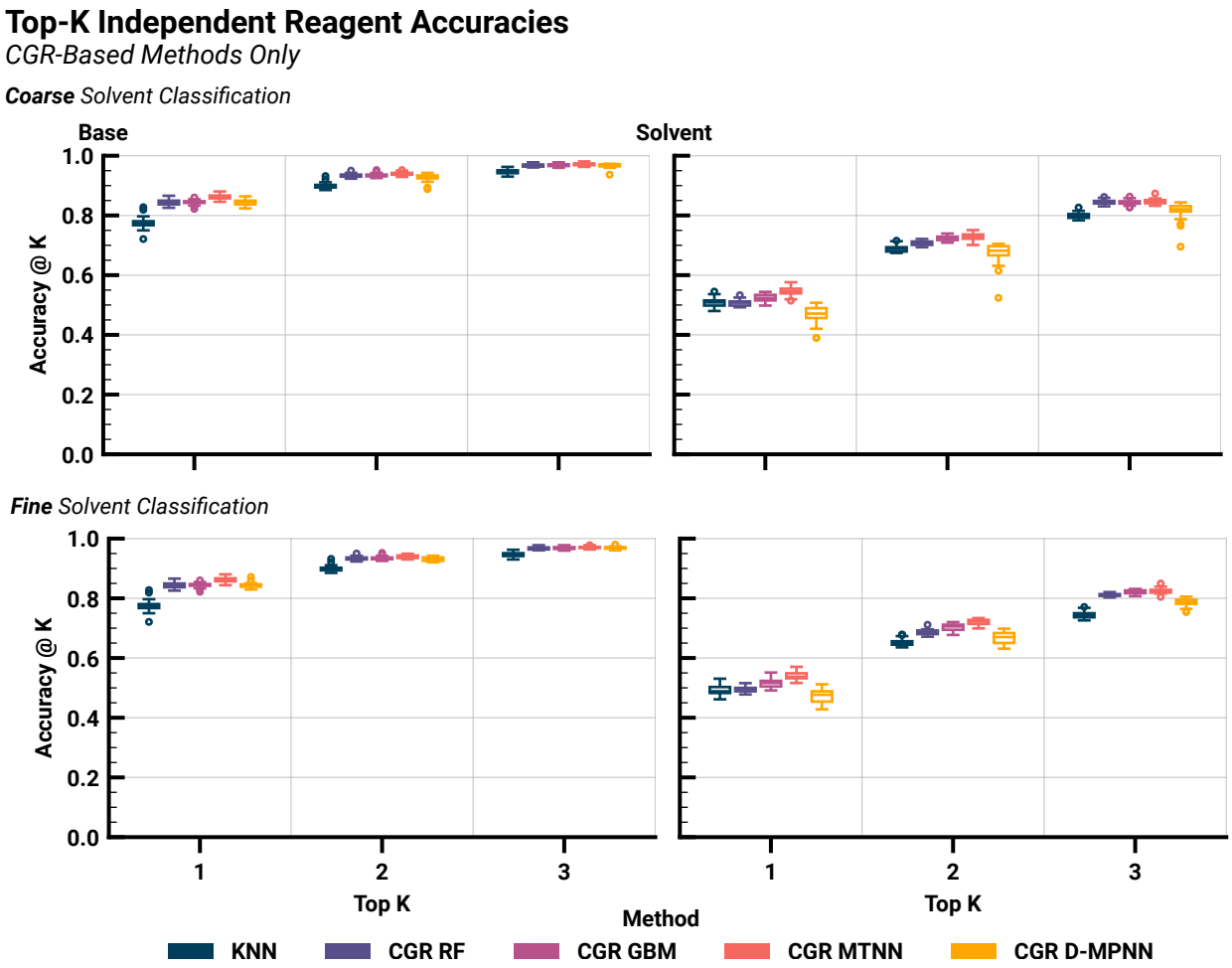


Figure S5: Independent accuracies, using only the CGR-based methods. Although the traditional ML models (CGR GBM in particular) showed comparable performance, the CGR MTNN was selected due to its strong performance on other metrics (see Fig. S8), and fast inference times.

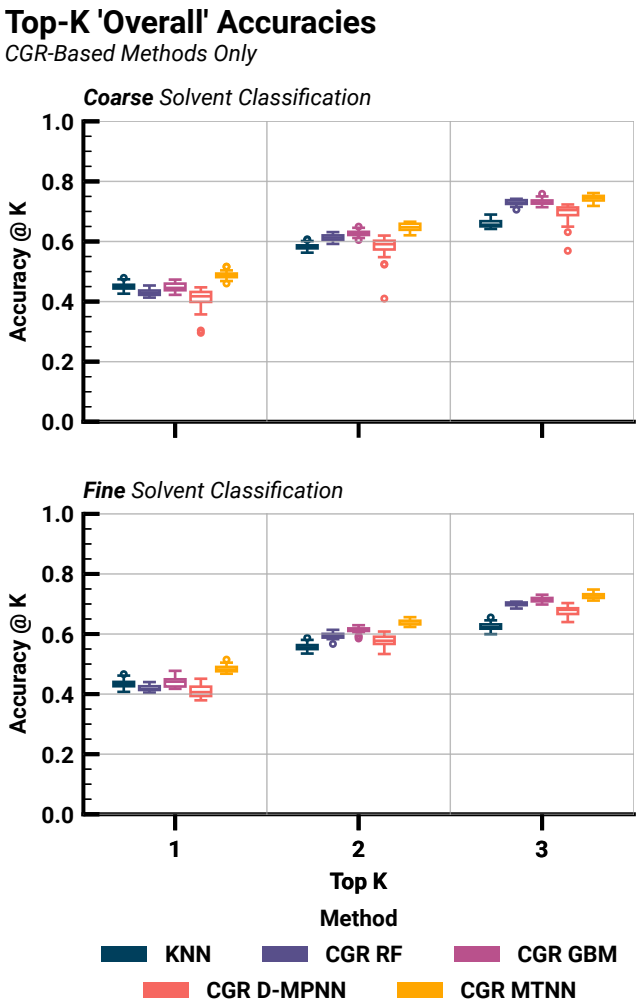


Figure S6: Overall accuracies, considering only CGR-based modelling methods.

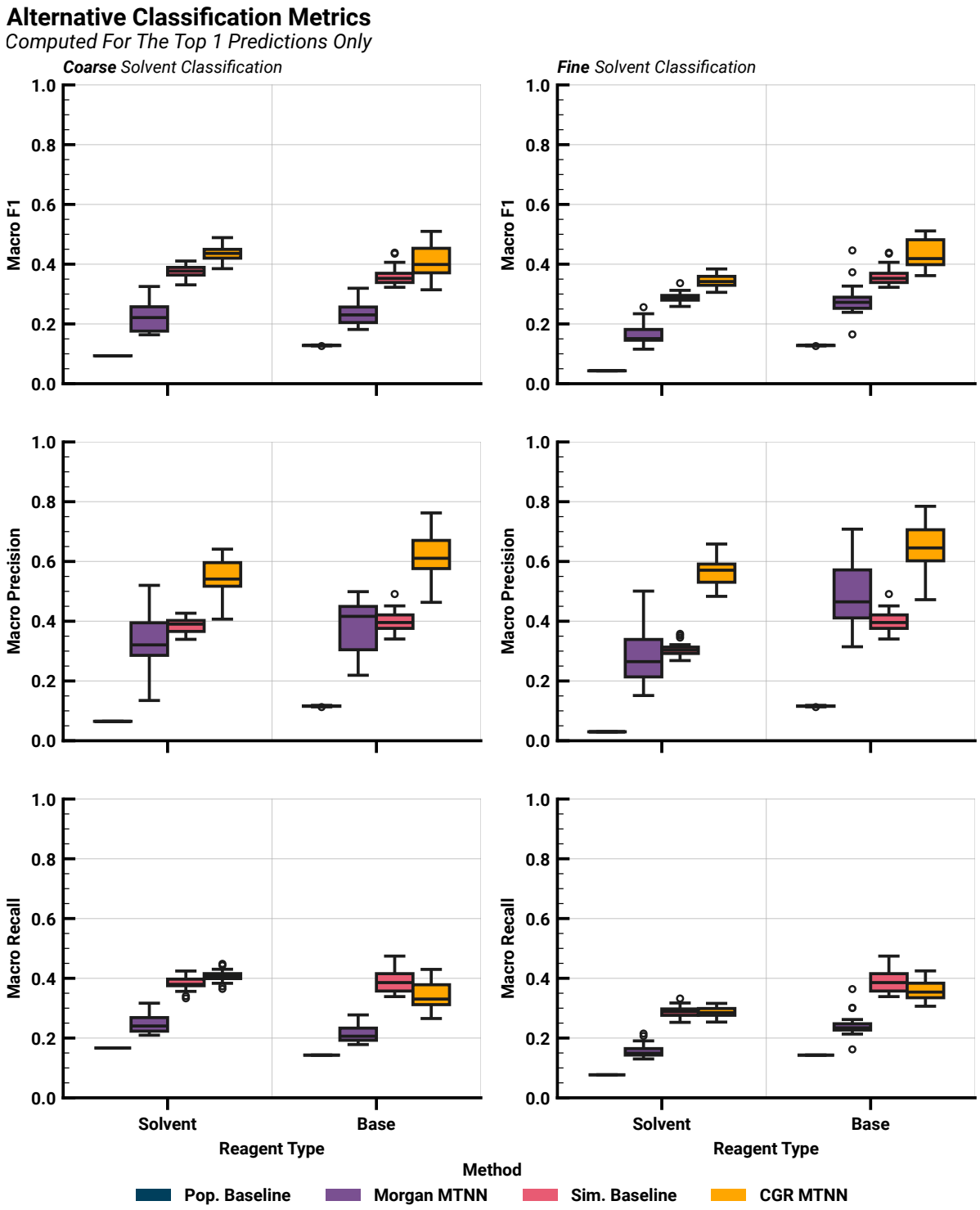


Figure S7: Alternative classification metrics for the models. Here, 'macro' averaging refers to the calculation of the metric for each individual class in multi-class classification, before taking the mean. We use equal weighting for all classes, since all classes (e.g. solvent classes are equally valid). Despite the strong accuracy scores (see Fig. S4), here scores on other classification metrics beyond accuracy are poorer. This highlights how further work is required to improve these models to predict correct conditions across all reagent clusters.

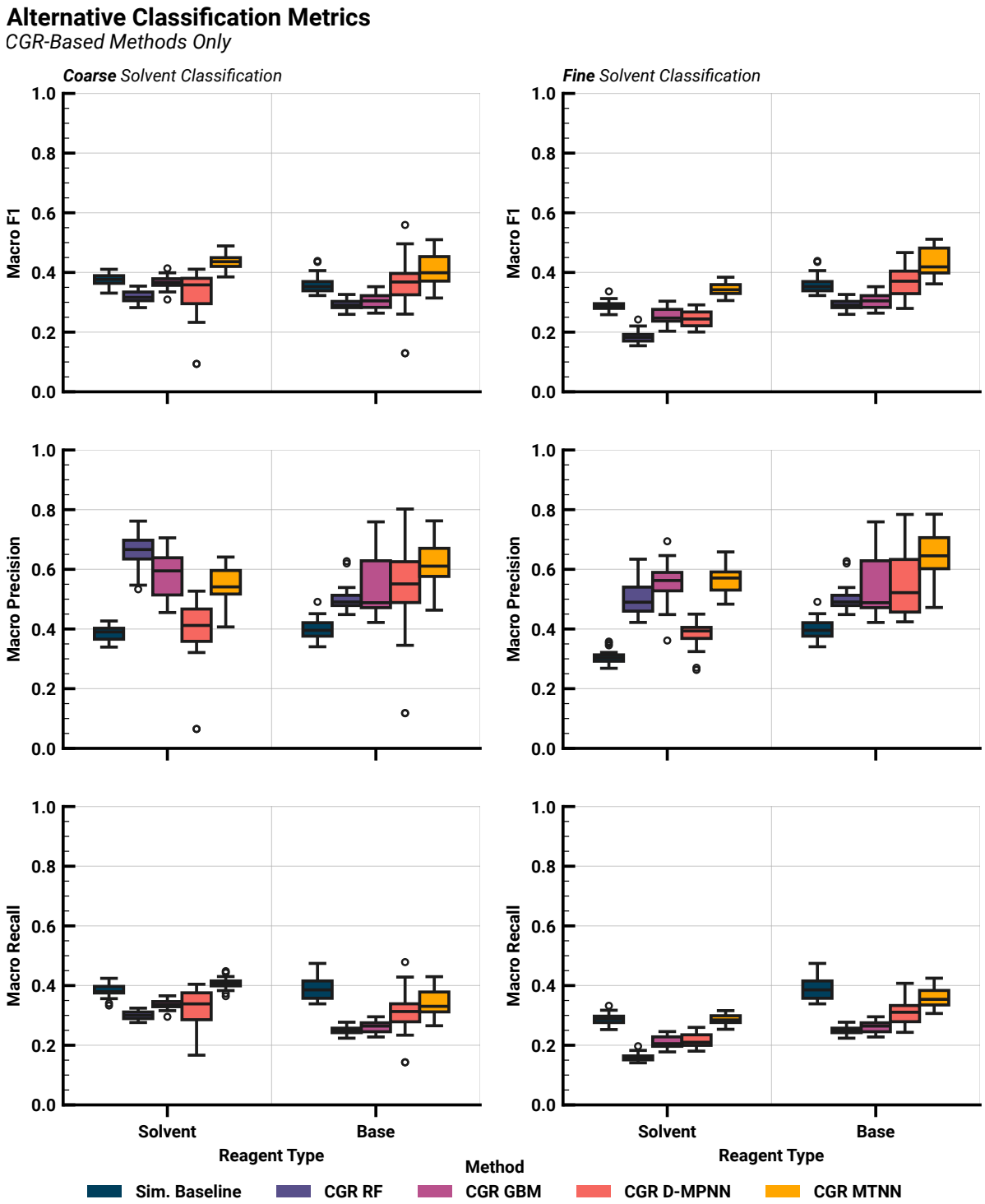


Figure S8: Alternative classification metrics for the models, showing only CGR-based methods.

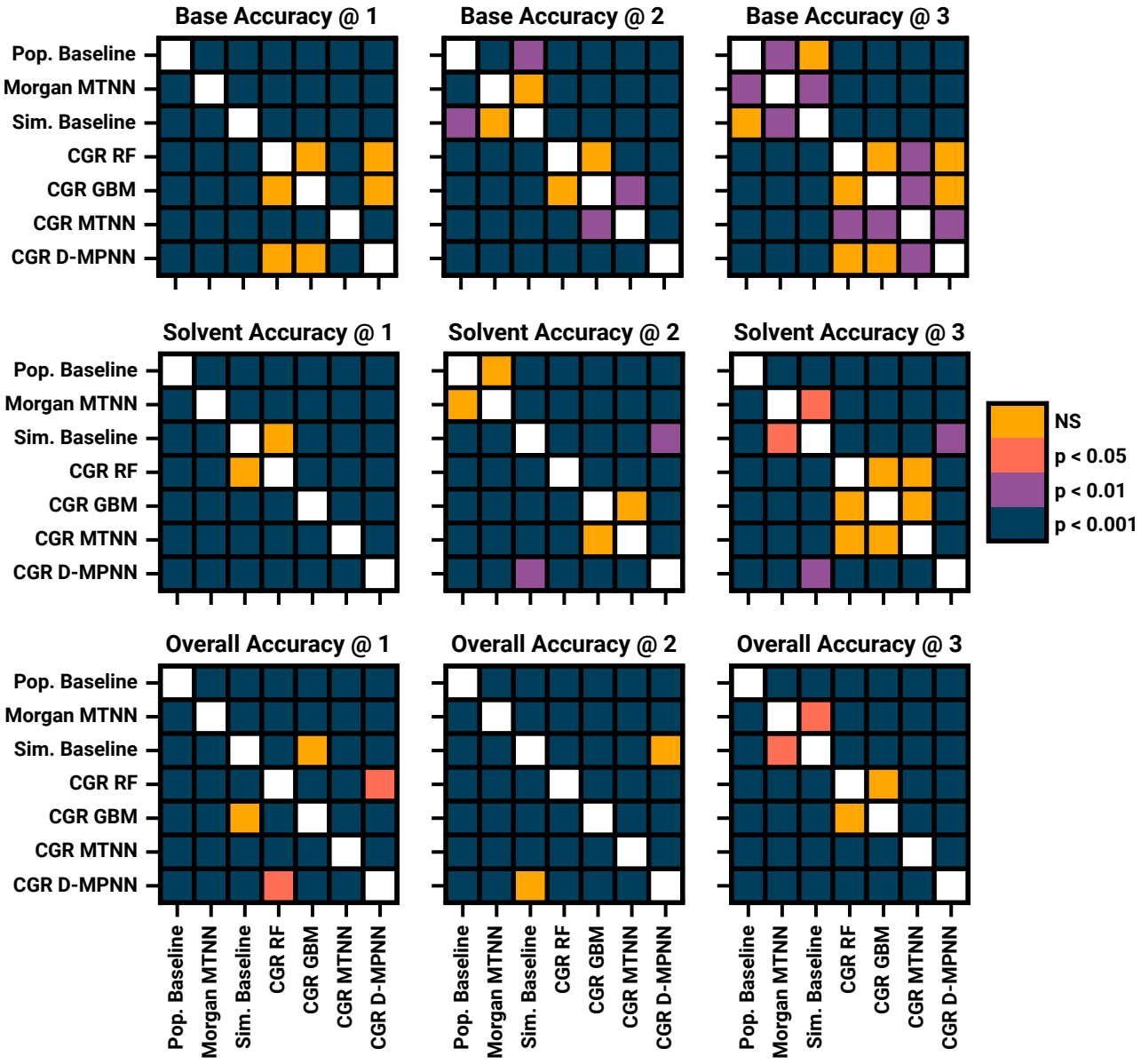


Figure S9: p-Values from statistical testing following the workflow set out by Ash et al..[2] A Holm-Bonferroni method is used, implemented using SciPy. Here, the accuracies for Base and Solvent are taken from the independent predictions (no likelihood re-ranking is performed), since this is what Beker et al. use in their work. Yellow indicates a non-significant difference between the distributions of values from the method corresponding to the row, and the method corresponding to the column. We can see that for the Top-1 accuracies, the CGR-MTNN results are statistically significant ($p \leq 0.001$) with all other methods. Crucially, the improvement from the CGR-MTNN to the popularity baseline and Morgan fingerprint baseline is statistically significant ($p \leq 0.001$) for all accuracy metrics.

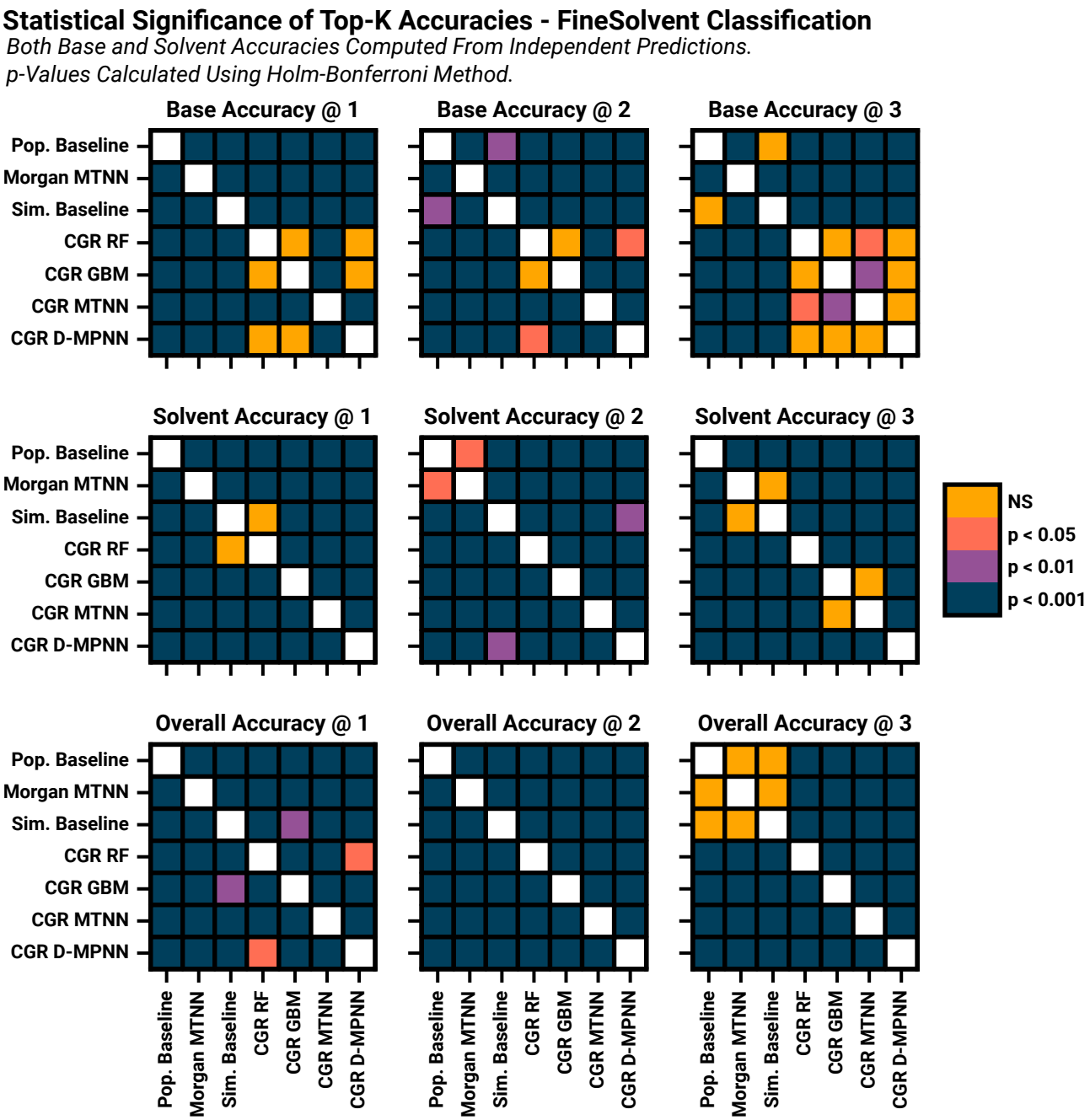


Figure S10: p-Values from statistical testing following the workflow set out by Ash et al..[2] Same as Fig. S9, but using the 'fine' solvent classification.

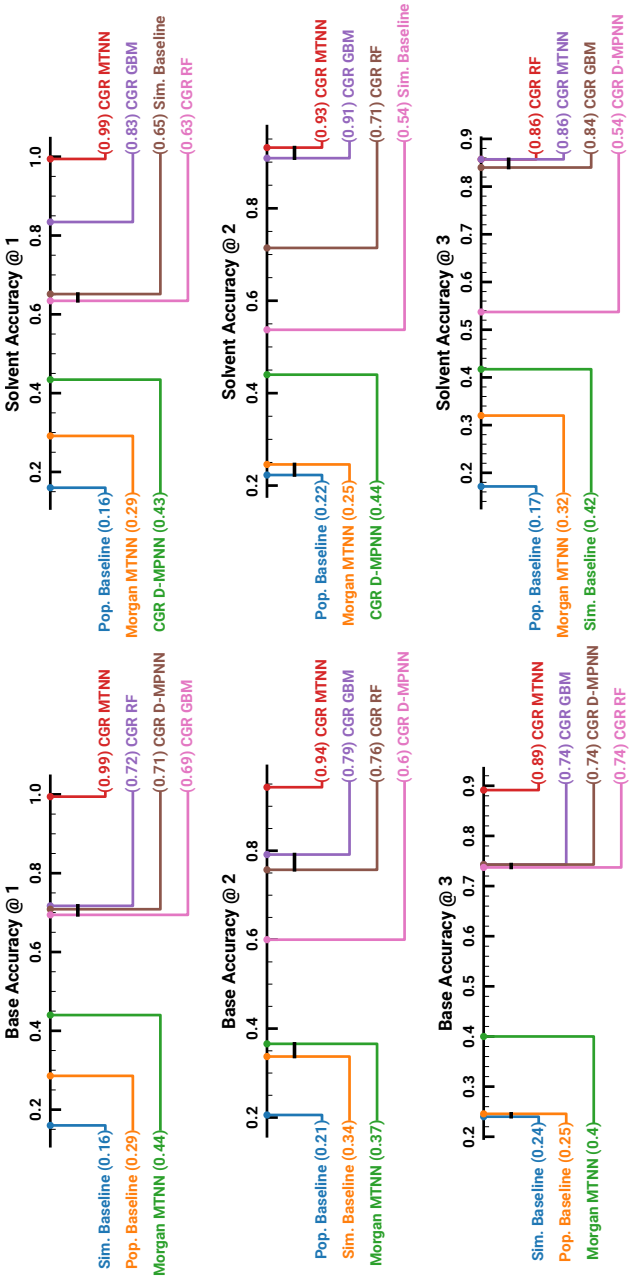


Figure S11: Critical difference diagrams for the different models across different metrics, using 'coarse' solvent classification. Shows the mean rank of each model across the 5x5 CV. A line between methods indicates a non-significant difference. We can clearly see that the CGR MTNN is the best model for most of these metrics, outperforming the literature baseline and Morgan fingerprint model.

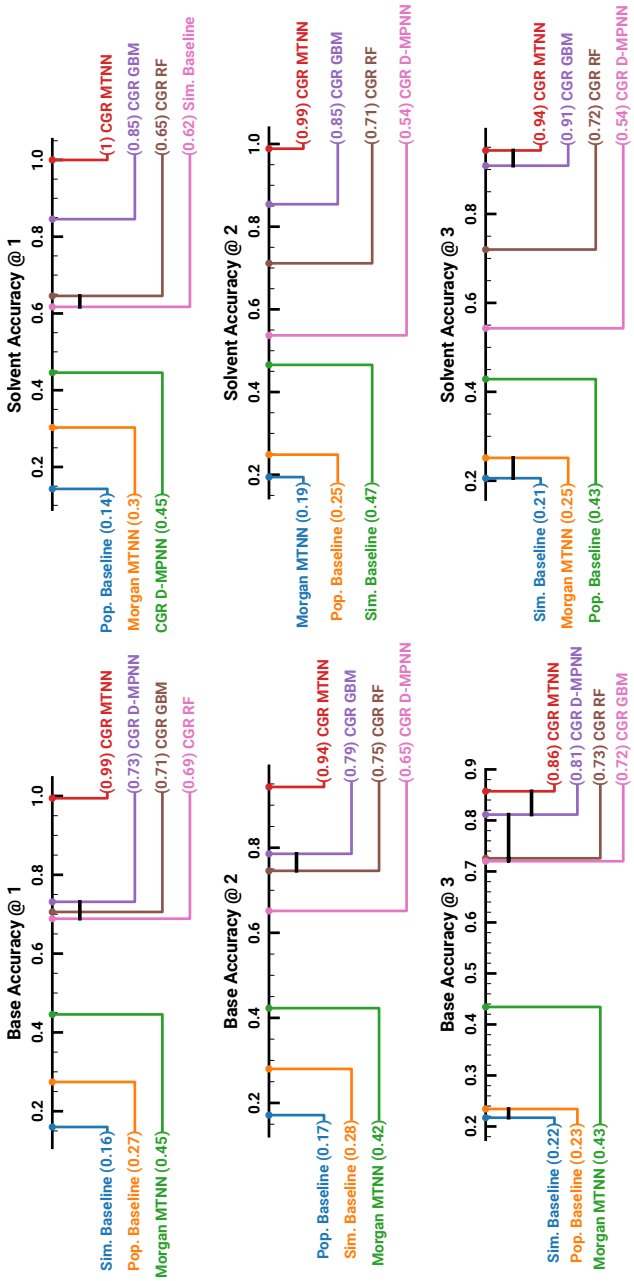


Figure S12: Critical difference diagrams for the different models across different metrics, using 'fine' solvent classification. Like Fig. S11 the CGR-MTNN outperforms literature baseline and Morgan fingerprint model across most folds in a statistically significant way.

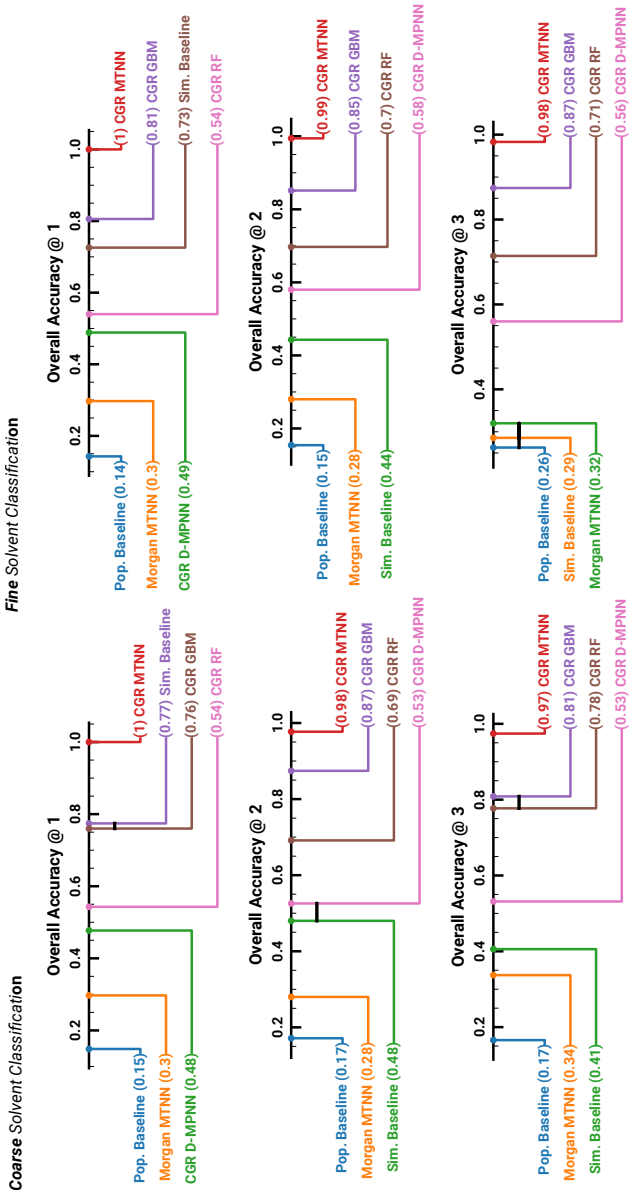


Figure S13: Critical difference diagrams for the different models across the 'overall' accuracies (getting both the base and solvent correct).

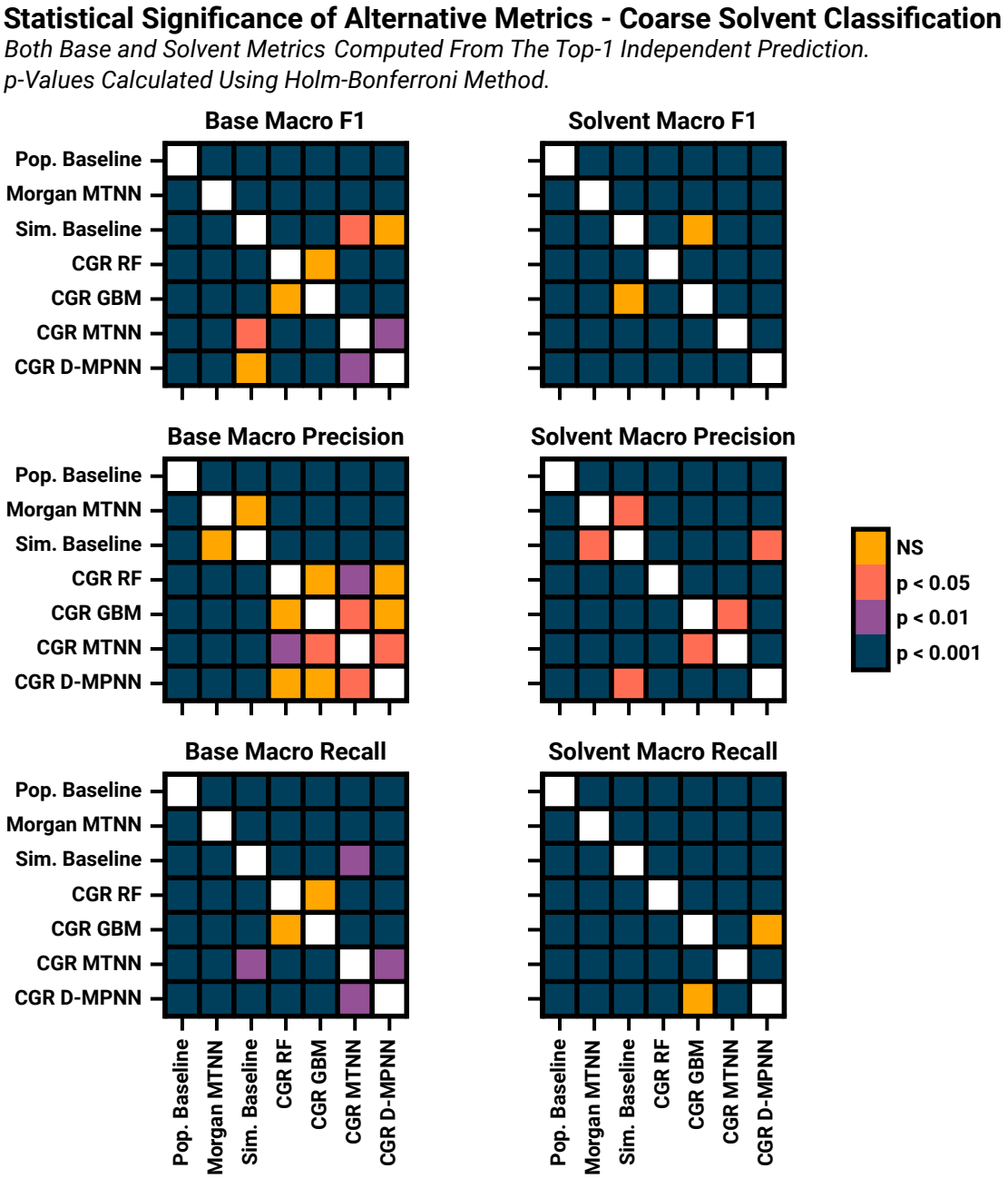


Figure S14: p-Values from statistical testing following the workflow set out by Ash et al..[2] Comparing differences in the distributions of the alternative classification metrics, for the 'coarse' solvent classification.

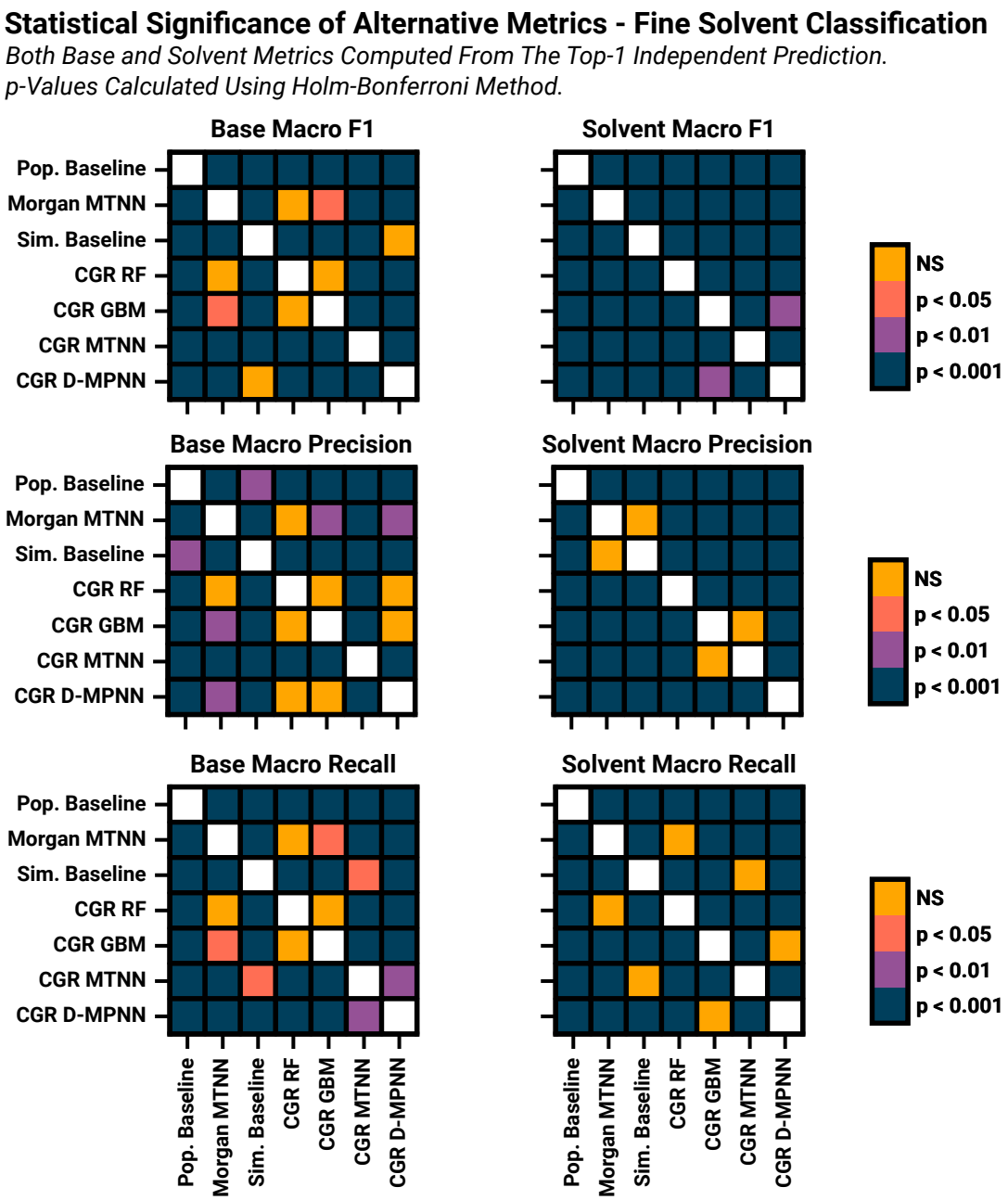


Figure S15: p-Values from statistical testing following the workflow set out by Ash et al..[2] Same as Fig. S14, but using the 'fine' solvent classification.

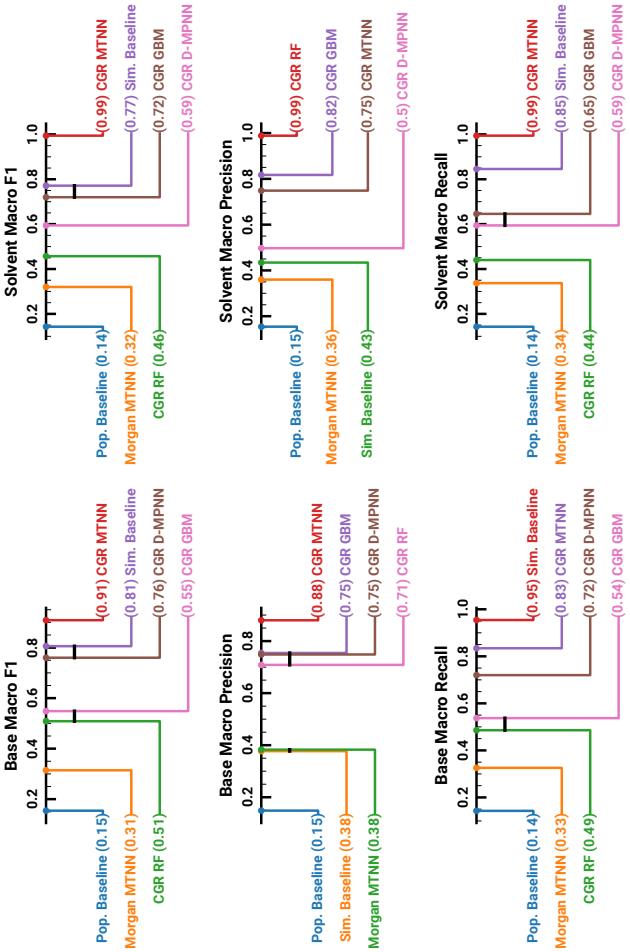


Figure S16: Critical difference diagrams for the different models across the alternative classification metrics, using 'coarse' solvent classification. Shows the mean rank of each model across the 5x5 CV. A line between methods indicates a non-significant difference. We can clearly see that the CGR MTNN is among the best models for all of these metrics, outperforming the literature baseline and Morgan fingerprint model.

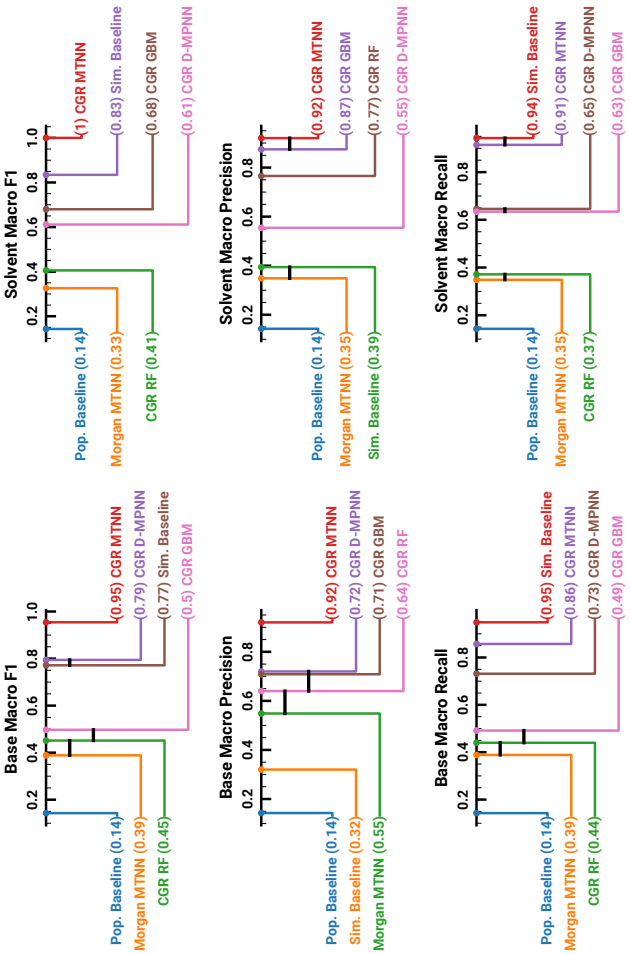


Figure S17: Critical difference diagrams for the different models across the alternative classification metrics, using 'fine' solvent classification. Like Fig. S17 the CGR-MTNN outperforms literature baseline and Morgan fingerprint model across most folds in a statistically significant way.

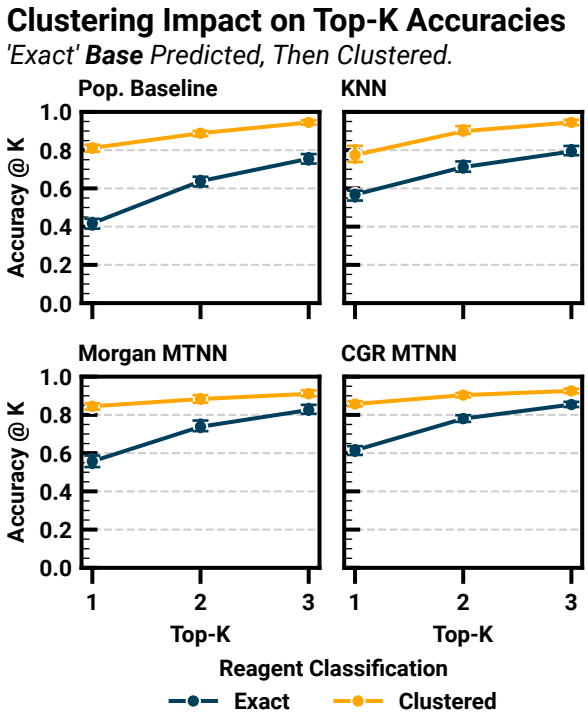


Figure S18: The impact of clustering on base Top-K accuracies. Like Fig. 7, clustering causes a large increase in model performance, suggesting that they are successfully predicting the correct solvent class, even if the counter-ion is incorrect.

# References

(1) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *Journal of the American Chemical Society* **2022**, *144*, 4819–4827.

(2) Ash, J. R.; Wognum, C.; Rodríguez-Pérez, R.; Aldeghi, M.; Cheng, A. C.; Clevert, D.-A.; Engkvist, O.; Fang, C.; Price, D. J.; Hughes-Oliver, J. M.; Walters, W. P. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. 2024; `https://chemrxiv.org/engage/chemrxiv/article-details/672a91bd7be152b1d01a926b`.