Predicting Reactivity and Reaction Yields in Parallel Synthesis: Meeting of Expectations and Reality

Iryna B. Boiko,[1,2] Anton V. Zhemera,[1,2] Dragos Horvath,[4] Dmytro G. Krotko,[1] Dmytro M. Volochnyuk,[1,2,3] Ihor V. Komarov,[1,2] Alexandre Varnek,[4] Serhiy V. Ryabukhin[1,2,3]

[1] Enamine Ltd, 78 Winston Churchill Street, 02094 Kyiv, Ukraine
[2] Taras Shevchenko National University Kyiv, 64 Volodymyrska Street, Kyiv 01601, Ukraine
[3] Enamine Scientific Research Institute, 78 Winston Churchill str., 02094 Kyiv, Ukraine
[4] Laboratory of Chemoinformatics, UMR 7140, CNRS University of Strasbourg, 4 rue Blaise Pascal, 67000 Strasbourg, France

Over the last decade, the availability of large, publicly accessible reaction datasets, combined with advances in machine learning, has driven growing interest in reaction informatics.[1] Among the topical tasks in this field is predicting reaction outcomes. However, predictive accuracy remains limited by data quality, as variability in purification procedures, side products, and experimental conditions introduce significant noise into reported yields.[1–3]

Here, we investigate reaction outcome prediction using a diverse dataset of ~4000 Biginelli products, generated under a standardized parallel synthesis protocol at Enamine Ltd,[4] comparing regression models for yield prediction with classification models predicting overall reaction feasibility. Different ML methods, representations, and descriptor types showed similar performance.

Consistent with previous studies,[3,5] regression models were strongly affected by yield outliers, likely arising from experimental inaccuracies (**figure 1**). Resynthesis of selected molecules confirmed data quality issues and demonstrated how collaboration with experimentalists enables iterative refinement of datasets and models. Our observations suggest that while models captured reactivity-related features, purification-related factors remained a major source of noise. In contrast, feasibility prediction proved more robust, achieving a balanced accuracy of ~0.70 on an external test set of over 400 newly synthesized products.

These findings highlight the role of machine learning not only in prediction, but also in data quality assessment and understanding of synthetic processes, emphasizing the critical impact of data quality on model performance. This study has established a foundation for future analyses of other reaction datasets, available at Enamine.

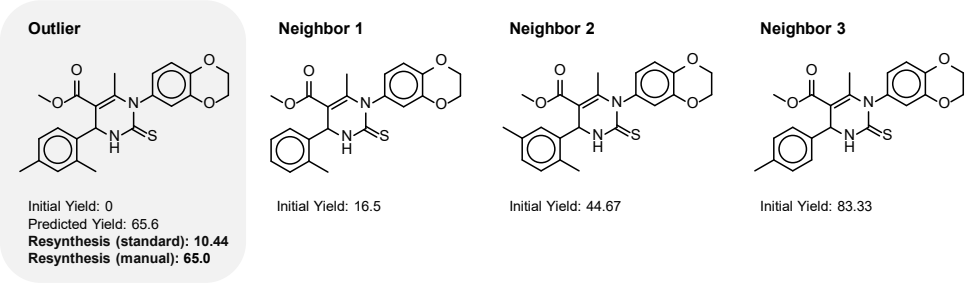

Figure 1. An example of a detected outlier and its 3 closest neighbors (by Tanimoto distance). The structural differences cannot rationalize the dramatically different experimental yields of these molecules; thus, the model cannot fit the outlier. The yields obtained in resynthesis, with both standard and manual purification, confirm that the product is, in fact, feasible.

Bibliography:

[1] T. I. Madzhidov, A. Rakhimbekova, V. A. Afonina, T. R. Gimadiev, R. N. Mukhametgaleev, R. I. Nugmanov, I. I. Baskin, A. Varnek, *Mendeleev Communications* **2021**, *31*, 769–780.
[2] R. G. Bergman, R. L. Danheiser, *Angew. Chem. Int. Ed.* **2016**, *55*, 12548–12549.
[3] M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle, N. V. Chawla, O. Wiest, *Chem. Sci.* **2023**, *14*, 4997–5005.
[4] E. N. Ostapchuk, A. S. Plaskon, O. O. Grygorenko, A. A. Tolmachev, S. V. Ryabukhin, *J. Heterocycl. Chem.* **2013**, *50*, 1299–1303.
[5] Z. Liu, Y. S. Moroz, O. Isayev, *Chem. Sci.* **2023**, *14*, 10835–10846.