Machine Learning Models for Yield Predictions of Biginelli Reaction

Iryna Boiko[1], Dragos Horvath[1], Serhiy Ryabukhin[2], Dmytro Volochnyuk[2], and Alexandre Varnek[1]

[1] Laboratory of Chemoinformatics, UMR 7140, CNRS University of Strasbourg, 4 rue Blaise Pascal, 67000 Strasbourg, France
[2] Enamine Ltd, 78 Winston Churchill Street, 02094 Kyiv, Ukraine

In recent years, the availability of large publicly accessible reaction datasets[1,2] along with the increasing interest in Machine Learning, has triggered progress towards computational modeling of chemical reactions.[3] Yield predictions are still less successful than other areas of reaction informatics.[3] Nevertheless, research in this field continues, as predicting a reaction's outcome can offer many advantages, reducing the costs invested into failed experiments.[4] However, in silico approaches are limited in predictive accuracy by the training data, raising critical issues with the notoriously noisy yields in literature or in-house datasets.[5–7]

Our work explores a dataset of 3869 unique Biginelli reaction products (figure 1), generated by Enamine's chemists, always following the same experimental protocol. Prediction models, both classification (feasible or not) and regression (of the estimated yield) were developed. Starting with simple product-based models, we moved to more complex representations, such as Condensed Graphs of Reactions. Different fragmentation schemes and coloring types for ISIDA descriptors, as well as other descriptor types, were systematically tested. However, we observed a relatively uniform performance of different ML methods, representations, and descriptor types: the R^2 in cross-validation was in the range of 0.34 - 0.40. The performance of regression models is limited by the inaccuracies in the reported yield values. The resynthesis of outlier molecules confirmed that the initial experimental yields were incorrect, demonstrating the ability of our approaches to be used for data quality control.

Still, the classification models are of satisfactory quality (Balanced Accuracy in cross-validation up to 0.8) and would be sufficient to improve decision-making and allow for the prioritization of compounds at Enamine. A combinatorial matrix of ~380 million potentially feasible Biginelli products was enumerated from all building blocks available at Enamine, and expected outcomes were predicted by a consensus model composed of three individual SVM classifiers. ~300 molecules, present in Enamine's REAL database, were selected for synthesis. The validation of the predictive ability of the models on new experimental data is pending.

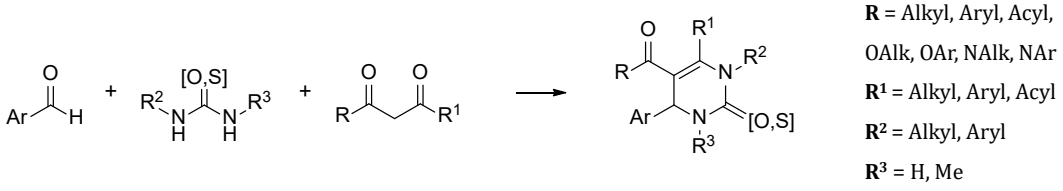

Figure 1. A schematic representation of the Biginelli reaction and a Markush structure, representing a Biginelli product

[1] D. Lowe, Extraction of Chemical Structures and Reactions from the Literature, **2012**.
[2] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science (80-.)*. **2018**, *360*, 186–190.
[3] T. I. Madzhidov, A. Rakhimbekova, V. A. Afonina, T. R. Gimadiev, R. N. Mukhametgaleev, R. I. Nugmanov, I. I. Baskin, A. Varnek, *Mendeleev Commun.* **2021**, *31*, 769–780.
[4] V. Voinarovska, M. Kabeshov, D. Dudenko, S. Genheden, I. V. Tetko, *J. Chem. Inf. Model.* **2024**, *64*, 42–56.
[5] P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016.
[6] 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.
[7] Z. Liu, Y. S. Moroz, O. Isayev, *Chem. Sci.* **2023**, *14*, 10835–10846.