Are the historically available experimental data valid for the AI prediction of the chemical reactivity?

Serhiy V. Ryabukhin[1,2,3], Dmytro M. Volochnyuk[1,2,3]

[1] *Enamine Ltd, 78 Winston Churchill str., 02094 Kyiv, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, 60 Volodymyrska str., 01601 Kyiv, Ukraine*
[3] *Institute of Organic Chemistry, National Academy of Sciences of Ukraine, 5 Akademik Kuhar str., 02660 Kyiv, Ukraine*
[4] *Enamine Scientific Research Institute, 78 Winston Churchill str., 02094 Kyiv, Ukraine*

Nowadays, machine learning has achieved promising results in predicting chemical reactions from high-throughput experimentation (HTE) data. However, most available data that come from historical preparative synthesis are more heterogeneous but operationally realistic. A critical question remains: what can ML models learn from standardized but imperfect real-world synthesis data? What is the difference between the data from open literature, parallel synthesis ELN, and HTE?

Using over 20 years of statistics from a model reaction of Enamine's combinatorial chemistry department, we examine several thousand historical records collected under a standardized parallel-synthesis protocol. Unlike literature datasets (biased toward successful cases) or narrow HTE campaigns, this dataset combines broad substrate diversity, a common workflow, and, importantly, numerous unsuccessful attempts. Multiple molecular representations and ML algorithms were evaluated for yield regression and feasibility classification.

Isolated yield proved to be a contradictory characteristic for machine learning. Models showed similar performance regardless of architecture or descriptors, suggesting the main limitation was data quality itself. Isolated yield conflates product formation with workup efficiency - making it an unreliable measure of intrinsic reactivity. Outlier analysis revealed that most prediction errors were due to technical issues or workflow failures in the synthetic process rather than to model shortcomings. In contrast, feasibility classification (whether standardized workflow delivers target material) proved robust and practically meaningful.

Also, ML serves as a valuable quality-control tool, identifying suspicious records warranting experimental re-examination. Historical parallel-synthesis data contain valuable information about real workflows and reagent scope, but should not be treated as clean ground truth for quantitative reactivity modeling. Successful AI in synthetic chemistry requires not only better algorithms but better-defined reaction records that distinguish product formation from isolation, and chemical failure from procedural failure. [1]

Bibliography:

[1] Boiko, I. B.; Zhemera, A. V.; Horvath, D. ; Krotko, D. G. ; Volochnyuk, D. M.; Komarov, I. V.; Varnek, A; Ryabukhin. S. V. ChemRxiv 2026 DOI: 10.26434/chemrxiv.15000783/v1