Investigating the Interpretability and Reliability of Machine Learning Frameworks for Chemical Retrosynthesis

Friedrich Hastedt[1], Klaus Hellgardt[1], Sophia N. Yaliraki[1], Ehecatl Antonio del Rio Chanona[1]*, Dongda Zhang[2]*

1 Imperial College London, SW7 2BX, UK; 2 University of Manchester, Oxford Rd, UK

**Corresponding author: a.del-rio-chanona@imperial.ac.uk, dongda.zhang@manchester.ac.uk*

Highlights

- Novel interpretability study to understand the "black-box" nature of deep-learning retrosynthesis models
- Open-source benchmarking pipeline for fair model comparison within the community
- Superiority of traditional reaction-rule model over prominent deep-learning models

1. Introduction

The discovery of synthesis routes towards novel (target) molecules is a central challenge for reaction engineers. This challenge can be solved through retrosynthesis in a backward fashion. By iteratively breaking down the target molecule into simpler reactant molecules, plausible synthesis routes can be discovered. As retrosynthesis planning is highly time-intensive, researchers have developed machine-learning (ML) tools to assist professionals. These models either rely on reaction rules (templates) extracted from literature[1] or acquire knowledge directly from a reaction database[2,3]. Although both approaches have demonstrated remarkable success on paper, their practical application in real-life scenarios has faced slow adaptation due to a lack of model interpretability and reliability. To address these limitations, we propose a novel interpretability study to understand the reasoning provided by the model for retrosynthetic predictions. Furthermore, for the first time, we investigate the reliability of prominent retrosynthesis models through a comprehensive benchmark for chemical (prediction) feasibility – a critical performance measure overlooked by the conventional "top-k accuracy" metric.

2. Methods

Our interpretability study aims to reveal whether ML-based retrosynthesis models propose reaction predictions due to a profound chemical understanding or simply due to dataset memorization. In this context, chemical understanding refers to the model's ability to identify important functional groups (nodes) in the target molecule that lead to thermodynamic stabilization of the reaction product over the reactant molecules. Retrosynthesis models heavily rely on the Transformer or Graph Neural Network (GNN) deep-learning architectures. As such, we utilize powerful interpretable AI techniques, namely GNNExplainer[4] and Attention Maps, to understand the black-box nature of the GNN and Transformer models, respectively. Two Transformers (vanilla and masked) and two GNNs (EGAT and D-MPNN) are trained and subsequently tested on five industrially relevant case studies. We present one case study below. Moreover, we introduce an open-source and automated benchmarking pipeline within this work. The benchmarking pipeline, based on several different performance metrics, quantifies the model's ability to predict diverse, valid and chemically feasible reactions. Herein, chemical feasibility refers to the likelihood that a reaction would be successful when tested experimentally. We select 12 state-of-the-art retrosynthesis models and evaluate their performance on our pipeline.

3. Results and discussion

Our benchmarking pipeline reveals that models based on reaction templates (from literature precedent) propose the most feasible and diverse predictions. On the other hand, it is shown that "purely" data-driven models suffer from chemically infeasible and invalid predictions. This demonstrates that the incorporation of chemical reaction rules strongly benefits the machine-learning model and boosts its performance and reliability.

The interpretability case studies underscore the limitations of purely data-driven models. As an example, a simple substitution reaction is shown in Figure 1. This reaction primarily proceeds due to the stronger C-N bond compared to the C-Br bond. As such, the models should indicate the importance of the C-N bond in the product molecule. In Figure 2, the atom importance is visualized for the different model architectures. Both Transformer models (subplots a/b) fail to highlight this bond and instead prioritize the secondary alcohol functionalization. Moreover, both GNN models propose different bond disconnections compared to Figure 1. The EGAT model implies an inefficient attachment to the benzene ring (subplot e), while the D-MPNN indicates a C-C bond formation without offering comprehensive interpretability for its prediction (subplots d & f). This highlights the challenge faced by purely data-driven models proposing chemically feasible reaction predictions, possibly stemming from a deficiency in chemical awareness.

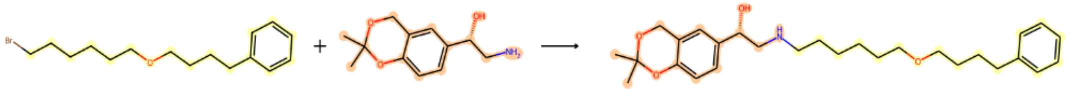
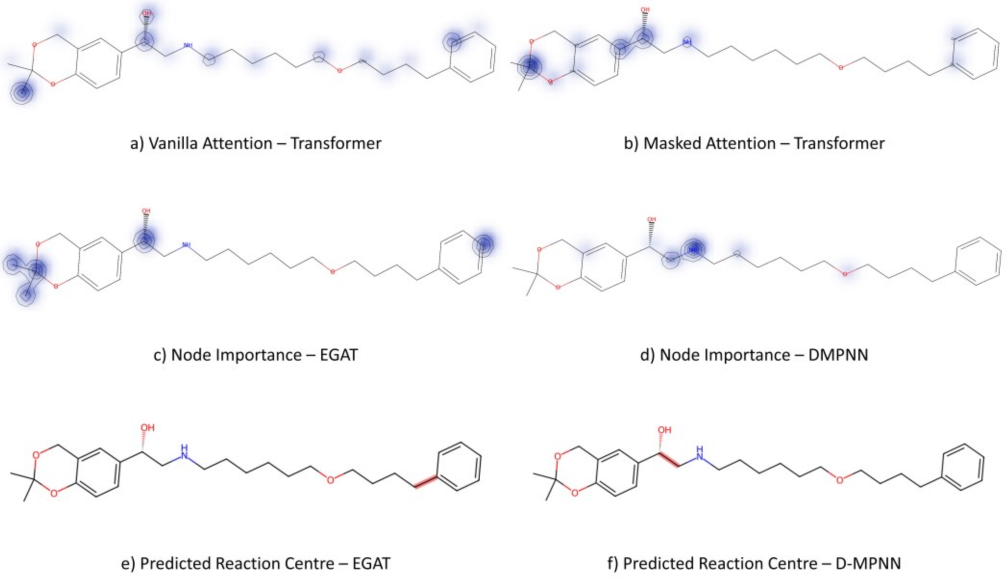


Figure 1: Example case study - substitution reaction



a) Vanilla Attention – Transformer

b) Masked Attention – Transformer

c) Node Importance – EGAT

d) Node Importance – DMPNN

e) Predicted Reaction Centre – EGAT

f) Predicted Reaction Centre – D-MPNN

Figure 2: Functional group (atom) importance by different models

4. Conclusions

In this work, we reveal that retrosynthesis models based on templates extracted from literature propose the most feasible and diverse predictions. Moreover, these models are easily interpretable thanks to the template's literature precedence. On the other hand, we observe that purely data-driven models, such as Transformers or GNNs, often lack interpretable predictions indicating a strong reliance on dataset patterns. Future models would therefore strongly benefit from chemically aware descriptors such as bond dissociation energy, weak interactions or electronegativity to instill a chemical prior to the model. Finally, this research provides guidance for future research directions to the community and the construction of more reliable and interpretable retrosynthesis models.

References

[1] Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2017). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, *4*(1), 120–131.

[2] Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, H., & Ning, X. (2023). G2Retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, *6*(1).

[3] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, *5*(9), 1572–1583

[4] Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J. (2019) 'GNNExplainer: Generating Explanations for Graph Neural Networks', arXiv [cs.LG]. Available at: http://arxiv.org/abs/1903.03894.

Keywords

Chemical Retrosynthesis, Machine Learning, Interpretable AI, Data-driven Reaction Engineering