Neural Network Guided Tree-Search Policies for Synthesis Planning

Amol Thakkar^{1,2}(✉), Esben Jannik Bjerrum¹(✉), Ola Engkvist¹,
and Jean-Louis Reymond²

¹ Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca,
Gothenburg, Sweden
{amol.thakkar, esben.bjerrum}@astrazeneca.com,
av.mol.th@gmail.com

² Department of Chemistry and Biochemistry,
University of Bern, Bern, Switzerland

Abstract. Developments and accessibility of computational methods within machine learning and deep learning have led to the resurgence of methods for computer assisted synthesis planning (CASP). In this paper we introduce our viewpoints on the analysis of reaction data, model building and evaluation. We show how the models' performance is affected by the specificity of the extracted reaction rules (templates) and outline the direction of research within our group.

Keywords: Reaction informatics · Neural networks · Classification · Synthesis prediction

1 Introduction

With the increasing availability of reaction data, developments and accessibility of computational methods, and a drive to further automate design, make, test, analyze (DMTA) cycles within drug discovery [1], computer assisted synthesis planning (CASP) has seen renewed interest as of late [2]. This has been spurred by recent achievements in the application of neural networks combined with search algorithms [3, 4], learning from breakthroughs in their application to games such as chess and Go [5].

CASP or retrosynthetic analysis refers to the strategy used by chemists to deconstruct a compound into its simpler precursors. In likeness with games, both are fundamentally decision-making tasks, in intractable search spaces, and with complex optimal solutions, where evaluation of the position and available moves at each step is difficult (Table 1). A knowledge base of reactions a chemist has learned throughout their career, coupled with extensive literature searching, form the basis of an initial pattern recognition step, from which applicable reactions can be identified and prioritized.

Recent studies have shown that neural network policies framed as multi-class classification problems can identify likely reactions through the noisy knowledge base [3, 4]. However, we have found they are heavily weighted towards frequently occurring reactions, owing to imbalanced datasets. Thus, miss out on less frequent yet feasible alternatives. In the present study, we explore and tune neural network architectures

Table 1. Comparison of search spaces in games vs retrosynthetic analysis.

	Chess	Go	Retrosynthesis
Search breadth	35	250	>500,000
Search depth	80	150	ca. 10

with the aim of maximizing the number of synthetically feasible options at each step. This is supplemented by curation and analysis of the underlying knowledge base, extracted from available reaction datasets. The number of which is limited when publicly available data is considered.

2 Methods

The US patent office extracts are a set of text mined reactions from the patent literature [6]. Given the reaction SMILES, an extension of the SMILES notation used to represent molecular structures [7], we used a modified version of Coley and coworkers algorithm to extract reaction templates [8]. That is the transformation required to convert the reactants into the products. These form the core of our knowledge base from which we can train a policy to enumerate retrosynthetic pathways in the form of a tree. To evaluate performance on a state-of-the-art model, we have opted to reimplement a variant of the policy used by Segler and Waller [4].

3 Results

We found that the applicability of the predicted templates from our trained policies varied with the template specificity (Fig. 1), the size of the template library, and the network architecture. Additionally, we developed a method for determining the selectivity of our templates during extraction and validation, the effect of which we are in the process of investigating on the trained model.

Preliminary evaluation of the models was performed on a random selection of 10,000 compounds from each ChEMBL [9] and FDB17 [10]. This enabled assessment of both the model’s predictive ability, and the validity of templates across a range of druglike and novel scaffolds. Using this assessment criteria, we aim to maximize the number of options available to our policy at each step in the subsequent tree search. Thereby, enabling the later prediction of full synthetic pathways, which is a necessity in accelerating automated DMTA cycles.

Whilst our viewpoint on the performance of the models has shed new light on the way in which a model may be evaluated, there is still much detail to investigate. This paper introduces preliminary results for a template-based synthesis planning methodology, and highlights that a more rigorous study is currently underway. This will encompass larger datasets, template design, data curation, the network architecture, and the implementation of appropriate metrics. These results will follow in a more rigorous study of the problems faced in computer assisted synthesis planning.

20,000 Randomly Sampled FDB17 and ChEMBL -
Successful Templates of Top N predictions for a Given Template Radius

Fig. 1. Assessment and comparison of the model's ability to predict templates that can be successfully applied for a subset of 10,000 randomly sampled compounds from each ChEMBL and FDB17. The radius refers to the number of bonds from the reaction center that are considered.

Funding. Amol Thakkar is supported financially by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 676434, "Big Data in Chemistry" ("BIGCHEM," <http://bigchem.eu>).

References

1. Nicolaou, C.A., et al.: Idea2Data: toward a new paradigm for drug discovery. *ACS Med. Chem. Lett.* (2019). <https://doi.org/10.1021/acsmmedchemlett.8b00488>
2. Boström, J., Brown, D.G., Young, R.J., Keserü, G.M.: Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709 (2018). <https://doi.org/10.1038/nrd.2018.116>
3. Segler, M.H.S., Preuss, M., Waller, M.P.: Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 (2018). <https://doi.org/10.1038/nature25978>
4. Segler, M.H.S., Waller, M.P.: Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**(25), 5966–5971 (2017). <https://doi.org/10.1002/chem.201605499>
5. Silver, D., et al.: Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484 (2016). <https://doi.org/10.1038/nature16961>
6. Daniel, L.: Extraction of chemical structures and reactions from the literature, Doctoral thesis, University of Cambridge (2012)
7. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**(1), 31–36 (1988). <https://doi.org/10.1021/ci00057a005>

8. Coley, C.W., Barzilay, R., Jaakkola, T.S., Green, W.H., Jensen, K.F.: Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**(5), 434–443 (2017). <https://doi.org/10.1021/acscentsci.7b00064>
9. Gaulton, A., et al.: The ChEMBL database in 2017. *Nucleic Acids Res.* **45**(D1), D945–D954 (2017). <https://doi.org/10.1093/nar/gkw1074>
10. Visini, R., Awale, M., Reymond, J.-L.: Fragment database FDB-17. *J. Chem. Inf. Model.* **57**(4), 700–709 (2017). <https://doi.org/10.1021/acs.jcim.7b00020>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.