Improving Route Development Using Convergent Retrosynthesis Planning

Paula Torren-Peraire[1,2], Jonas Verhoeven[1], Dorota Herman[1], Hugo Ceulemans[1], Igor Tetko[2], Jörg K Wegner[3]

1. In-Silico Discovery, Janssen Research & Development, Janssen Pharmaceutica N.V, Beerse, Belgium;

2. Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich, Neuherberg, Germany;

3. In-Silico Discovery, Janssen Research & Development, Janssen Research & Development LLC, Cambridge, US

Abstract

Computer-aided synthesis planning approaches have allowed a greater exploration of potential synthesis routes. However, these methods are generally developed to produce linear routes from a singular product to a set of proposed building blocks and are not designed to leverage potential shared paths between targets. These convergent routes allow the simultaneous synthesis of compounds, reducing the time and cost of synthesis across compound libraries. We introduce a novel planning approach to develop convergent synthesis routes, which can search multiple products and intermediates simultaneously, enhancing the overall efficiency and practical applicability of retrosynthetic planning. We evaluate the multistep synthesis planning approach using extracted convergent routes from Johnson & Johnson Electronic

Laboratory Notebooks (J&J ELN) and publicly available datasets and observe that solvability is generally very high across those routes, being able to identify a convergent route for over 90% of the test routes and showing an individual compound solvability of over 98%.

Methods

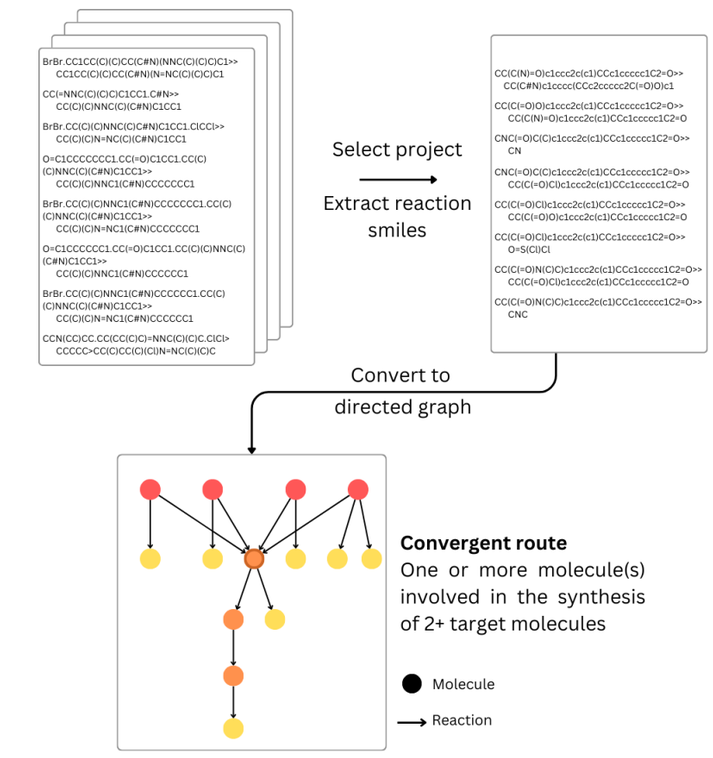

Figure 1. Convergent routes dataset. Extraction process of synthesis routes comprised of multiple target molecules resulting from common intermediates. Each project is processed individually, and the convergent routes collected per dataset.

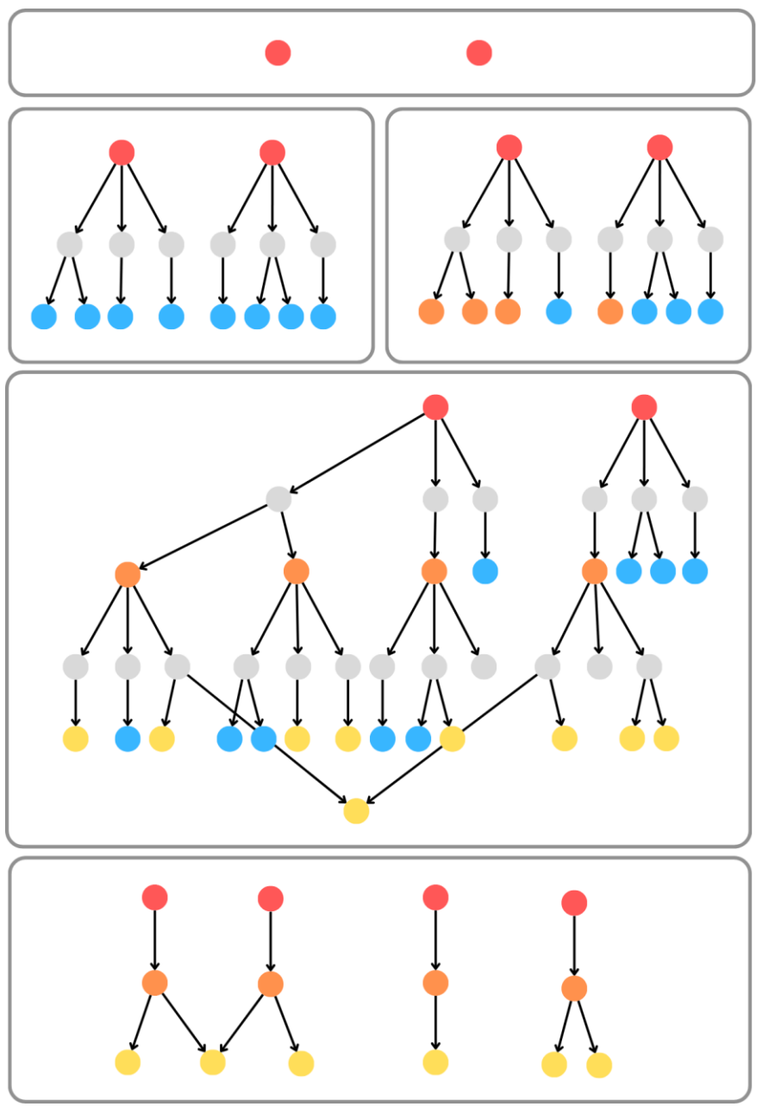

Figure 2. Multistep search process. The hypothetical example shown consists of two target molecules, with 3 sets of reactants proposed for each molecule node (*N*) and 4 molecule nodes followed up (*K*) with a maximum of two iterations. Only a sample of the potential extracted routes is shown.

Results & Discussion

Using J&J ELN and USPTO data separately we create a convergent route dataset of each reaction dataset

Table 1. Summary of raw data and extracted convergent routes from J&J ELN and USPTO data.

	J&J ELN	USPTO
Raw Data		
Total Reactions	-	3,746,981
Number of Projects	-	226,746
Convergent Routes		
Reactions Recovered	87%	70%
Projects Used	85%	36%
Number convergent routes		**94,521**

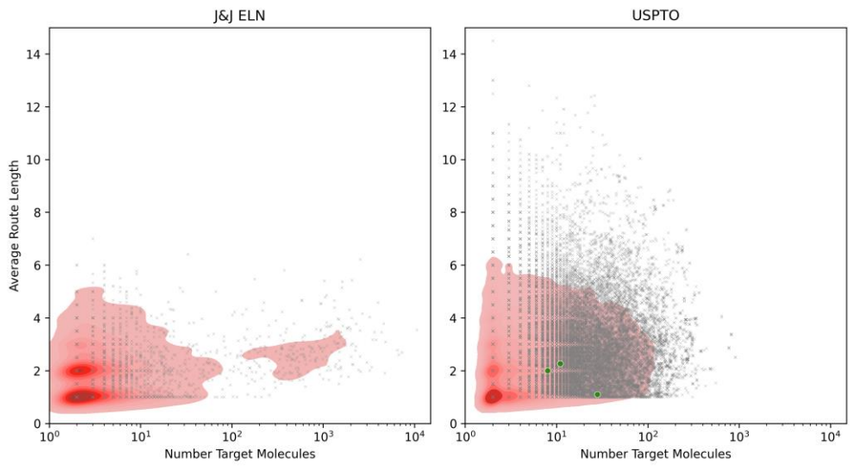

Figure 3. Distribution of the number of target molecules and average route length for convergent routes, per convergent route, from J&J ELN and USPTO.

We develop a new multi-step synthesis planning framework which instantiates multiple target molecules simultaneously, with the aim of convergent route development. Using the convergent route datasets developed for J&J ELN and USPTO we can search convergent routes for real compound libraries to assess the utility of the approach. We randomly select 500 and 1000 convergent routes from J&J ELN and USPTO respectively as the multi-step test set. We train a single-step retrosynthesis model based on the remaining data, fine-tuning the pre-trained Chemformer on each dataset, to guide the multi-step approach..

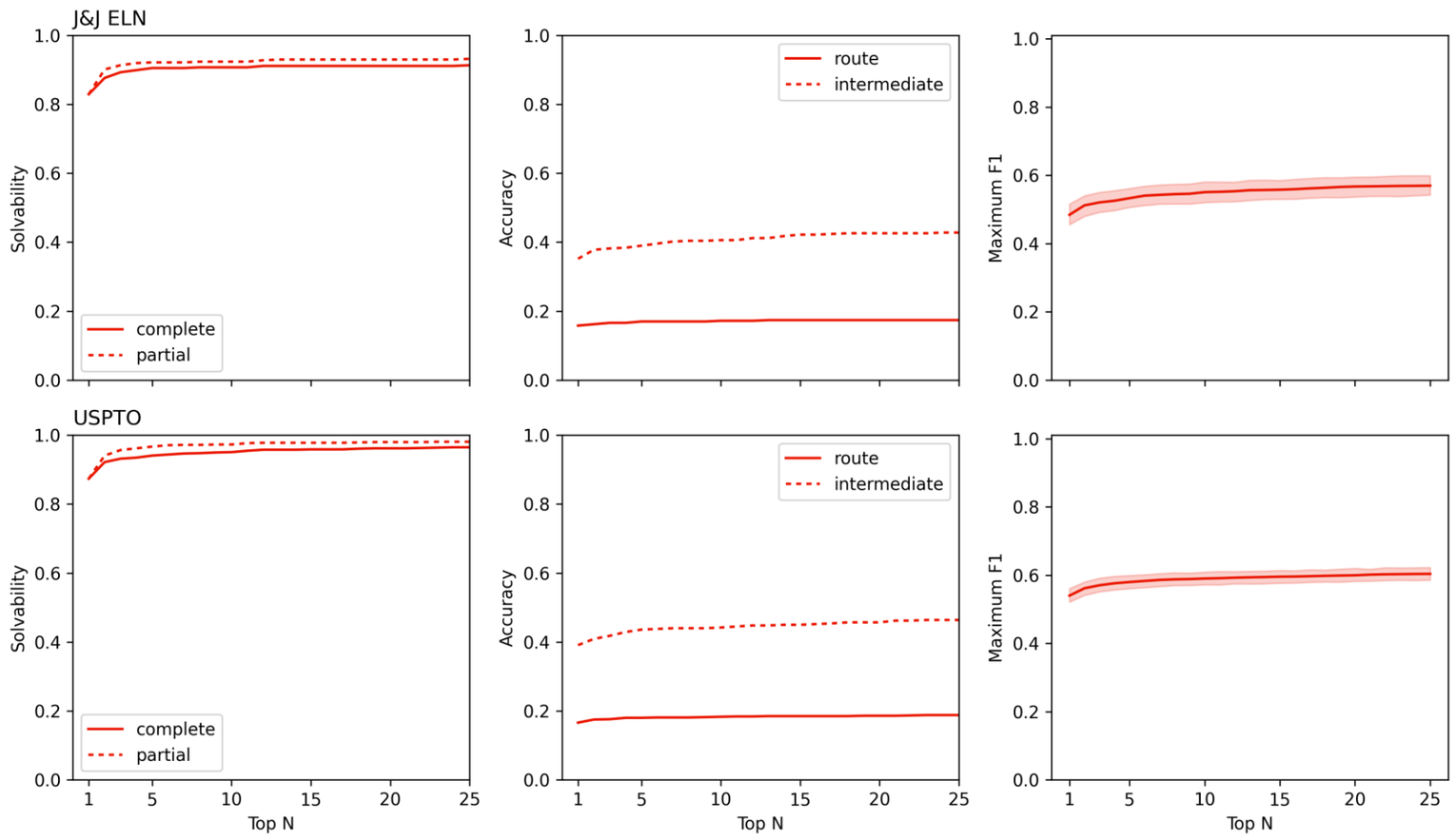

Figure 4. Solvability, accuracy and combined F1 score of proposed retrosynthetic routes using J&J ELN (top) and USPTO (bottom) compound library test sets. Accuracy and combined F1 score are calculated compared to the extracted experimentally validated retrosynthetic routes.

Table 2. Average statistics of the highest-ranked proposed retrosynthetic route for J&J ELN and USPTO compound library test sets.

	J&J ELN	USPTO
Fraction solved molecules	88.6%	89.7%
Common intermediates	3.1	4.2
Building blocks	6.3	9.3
Molecules	27.9	38.1
Reactions	19.8	26.1
Reactants per reaction	1.4	1.5
Target molecules per intermediate	3.3	3.6

Conclusion

- Convergent routes, producing the synthesis of multiple target molecules from a shared synthetic path, are a central and common part of medicinal chemistry
- We introduce a multi-step synthesis planning approach to develop convergent synthesis routes, which can search multiple products and intermediates simultaneously, enhancing the overall efficiency and practical applicability of retrosynthetic planning
- We evaluate the multi-step synthesis planning approach using a novel dataset of extracted convergent routes from industry-relevant and publicly available datasets,
- With the convergent route approach, we identify a convergent route for over 90% of the test routes and producing a synthesis route for over 98% of compounds found within the compound libraries.
- The proposed routes are similar to the experimentally validated routes in over a third of the compound libraries.

References

1. D. M. Lowe, “Extraction of chemical structures and reactions from the literature,” Thesis, 2012
2. R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, “Chemformer: a pre-trained transformer for computational chemistry,” Machine Learning: Science and Technology, vol. 3, no. 1, p. 015022, 2022

Acknowledgments

This study was partially funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement “Advanced machine learning for Innovative Drug Discovery (AIDD)” No. 956832.

Paper

LinkedIn

Johnson&Johnson

HELMHOLTZ MUNICH