Evaluation Metrics for Single-Step Retrosynthetic Models

Philippe Schwaller
IBM Research – Zurich, Switzerland
DCB, University of Bern, Switzerland
phs@zurich.ibm.com

Vishnu H. Nair
IBM Research – Zurich, Switzerland
har@zurich.ibm.com

Riccardo Petraglia
IBM Research – Zurich, Switzerland
rpe@zurich.ibm.com

Teodoro Laino
IBM Research – Zurich, Switzerland
teo@zurich.ibm.com

Abstract

Data-driven chemical synthesis is expected to accelerate the discovery and development of new molecules and materials. One of the main strategies to do chemical synthesis design is retrosynthesis. A retrosynthetic step involves the break down of the target molecule into available starting materials by imaginary disconnection of bonds or by functional group interconversion, which leads, upon iteration, to a tree of possible synthesis pathways. Multiple data-driven retrosynthetic models have been proposed in the last years to help chemists construct optimal routes. However, their performance is typically evaluated with a top-N accuracy metric, which is the probability of finding the ground truth output within the first N recommendation of the predictive model. In this work, we analyze the drawback of using a top-N accuracy and propose an analysis over three of the four evaluation metrics introduced in a recent publication of ours: round-trip accuracy, coverage and diversity. We show that it is possible to train a transformer-based retrosynthetic model, reaching a round-trip accuracy of 82.4%, while covering 96.4% of the reactions.

1 Introduction

The synthesis of novel materials and molecules, for example, new medicinal drugs, agrochemicals and polymers, has a tremendous impact on modern society. Recently, the search for strategies to accelerate the discovery of new molecules led to innovative algorithms for molecular design [1, 2]. However, apart from few exceptions [3, 4, 5], most de-novo molecules were not experimentally verified due to obstacles in their synthetic routes. Current molecule generation strategies do not account for the synthesizability of a target molecule. Therefore, computer-based algorithms that efficiently evaluates the synthesis of a candidate structure and provides a way to penalize the design of hard-to-synthesize molecules will soon be a key component to bring molecular design closer to experimental validation.

The dream of automating the design of synthesis has been around for many decades and was first formulated by Corey[6], who pioneered the concept of retrosynthetic analysis. In retrosynthesis, a synthetic route is designed starting from the desired product and following a backward analysis. In every retrosynthetic step, candidate precursors are suggested, which by reacting together would form the molecules from the preceding step, until (commercially) available precursors are found [7, 8]. For a comprehensive review of computer-aided synthesis, we refer the reader to recent publications [9, 10].

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

A crucial part of a retrosynthetic tool is the algorithm that suggests candidate precursors for each step, generating a tree of possible reactions, where the ending nodes are (commercially) available starting materials. Whenever a reacting molecule, which is not commercially available, is suggested, the tree is further expanded at that molecule. However, one of the intrinsic difficulties is evaluating and comparing the single-step prediction models. In fact, various single-step retrosynthetic models have been proposed in the last two years by prioritizing automatically extracted reaction rules [11, 7], traditional seq2seq [12], molecular similarity [13], transformer-based models [14, 15, 16, 17, 18, 19] and a graph-2-graph and graph-2-seq two-step method [20]. To simplify the prediction task, authors removed reagents (molecules not contributing atoms to the products) from the reactants using rule-based reaction-reagent role assignment [21, 22] and evaluated their models using a top-N accuracy. Top-N accuracy means that the exact precursors (matching the same entry as the one reported in the data set) are found within the N most likely predictions of the model.

Unfortunately, chemical reactivity is not a one-to-one function but rather as a many-to-one, in which multiple sets of precursors can react to produce a given product. What makes one of them more preferable than the other are things such as yields, ease of execution and price. An example is shown in Figure 1. This means that any top-N accuracy reported in previous literature is simply an evaluation of how good the model is in retrieving precisely the same information that was stored in the evaluation data set. To make things worse, the reactions reported in this set very likely are not the optimal ones. Consequently, evaluating single-step retrosynthetic models by comparing the top-N accuracy is only a metric of how good the model is performing on information retrieval tasks and not of the quality of its predictions.

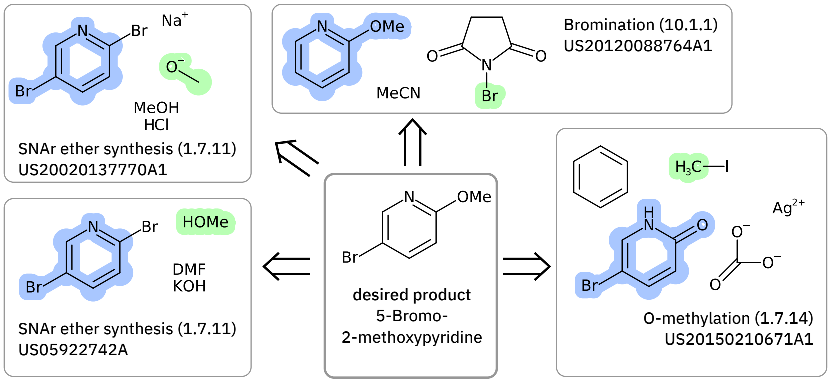

Figure 1: Highlighting few of the ground truth precursors and reactions to form 5-Bromo-2-methoxypyridine.

Therefore, we should focus on evaluating if the model predictions are valid combinations that could react to give the specified product without pretending to retrieve the exact entry of the validation set. Recently we disclosed a novel end-to-end retrosynthetic framework [23], where we introduce four different metrics to optimize single-step retrosynthetic models. Here, we discuss more in detail three of the four metrics based on a model-score, namely, round-trip accuracy, coverage and diversity. We dispute the previous use of top-N accuracy and initiate a discussion about new ways of evaluating and improving retrosynthetic models. A more elaborate analysis on the complete set of metrics for single-step retrosynthetic models within multi-step pathway predictions, is available in a recent work of ours [23].

2 Evaluation Metrics for Single-Step Retrosynthesis Models

An incontrovertible evaluation of a model prediction involves an assessment by human experts followed by validation with wet-lab experiments. Such an evaluation is unfortunately not scalable, and while it can be done in a few cases, it demands effort and investments. The closest approach to the use of human experts is the use of a forward chemical reaction prediction model to validate the suggestions of the retrosynthetic models [24, 7]. In fact, given a set of reactants/reagents, a forward prediction model predicts the corresponding product and by-products (possibly due to selectivity issues). Model scores, as an alternative to human annotators, have been already used to evaluate generative adversarial networks [25].

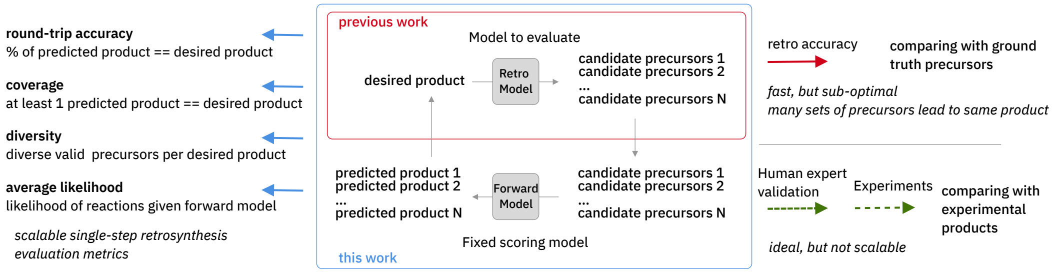

Figure 2: Overview of single-step retrosynthesis evaluation metrics.

Here we analyze three of four metrics recently introduced [23] to evaluate retrosynthetic models based on the use of a chemical reaction prediction model (forward model) and report an overview in Figure 2. First, the **round-trip accuracy** quantifies what percentage of the retrosynthetic suggestions are valid. It is desirable to have as many valid suggestions as possible. Letting a retrosynthetic model make more suggestions, for example, using a beam search, might lead to a smaller percentage of valid suggestion, as the suggestion quality decreases in the higher beams.

Second, the **coverage** quantifies for how many of the products at least one valid suggestion of the set of reactants could be found. A model could make many valid suggestions for some reactions, but none for the rest. This would result in small coverage. A retrosynthetic model should be able to produce valid suggestions for various products.

Third, the **diversity** counts the number of diverse valid precursors after removing the available (buyable) molecules. It is desirable for a single-step retrosynthesis model to predict structurally different precursors and not just reactant sets with, for example, just a different solvent. Similar to a skilled chemist, a model that is predicting precursors with a high diversity will have a better chance to circumvent failing reactions in a synthesis route. Another way to assess the diversity of the suggested reactions is to use a reaction classifier and count the number of reaction classes [23].

Beyond the three metrics reported here, it is important to measure the bias the model introduces in suggesting different classes as a consequence of imbalanced training data sets. We cover elsewhere [23] this important aspect, complementing the diversity metric with a statistical analysis of the probability distributions across different suggested classes, using a novel reaction class prediction model [26].

Additionally, it is also important that the model produces syntactically valid molecules. We check this using the open-source chemoinformatics software RDKit [27].

3 Results & Discussion

We used the metrics suggested above can be applied to any retrosynthesis model. Here, we apply the metrics to a retrosynthesis model based on the Molecular Transformer [28, 29, 30], trained using the same method and hyperparameters as described in the Molecular Transformer paper. The main difference is that compared to the training of the reaction prediction Molecular Transformer, we trained the retrosynthesis by exchanging source and targets. Similar to previous work [31, 12, 32], the molecules used as source and target are strings represented with the simplified molecular-input line-entry system (SMILES) [33], a line notation to describe molecular graphs. e.g. the SMILES for a benzene ring would be "c1ccccc1". Recent work [14, 15, 16, 17, 18] have also used the transformer architecture for retrosynthesis but limited themselves to the predictions of reactants and a top-N evaluation.

Table1 shows the development of the three evaluation metrics during the training of a retrosynthesis model. For this experiment, we fixed the beam size at 10. A beam search with beam size X allows obtaining the X most likely precursor sets for a given product. As expected, all three metrics increase continuously before converging at around 100k time steps, while the percentage of invalid SMILES simultaneously decreases.

Table 1: Development of the metrics during training.

Model	Beam	Total rxns	Round-trip accuracy	Coverage	Diversity per rxn	Invalid SMILES
stereo 10k	10	100k	56.9%	87.4%	1.87	4.03 %
stereo 20k	10	100k	73.8%	93.8%	2.15	1.72 %
stereo 50k	10	100k	78.7%	95.0%	2.29	0.81 %
stereo 100k	10	100k	81.6%	95.8%	2.28	0.65 %
stereo 150k	10	100k	81.3%	95.8%	2.29	0.62 %
stereo 200k	10	100k	81.0%	95.8%	2.32	0.59 %
stereo 250k	10	100k	81.5%	95.9%	2.23	0.58 %

In Table 2, we compare the evaluation metrics for three different beam sizes and models trained on different amounts of data. The training data sets, differ as follows: *stereo* contains 1M reactions from the USPTO dataset [34], *stereo&text* contains additionally 900K textbook reactions generated by Nam & Kim [31]. Adding the additional data seems to be beneficial for all the metrics and to decrease the percentage of invalid SMILES in the suggested precursors. As seen in Table 2, increasing the beam size leads to a slight decrease in the round-trip accuracy but the number of diverse reactions and the coverage increases.

Table 2: Evaluation of retrosynthesis models with different training data, evaluated on the same validation set with different beam sizes.

Model	Beam	Total	Round-trip accuracy	Coverage	Diversity per rxn	Invalid SMILES
stereo	5	50k	82.4%	93.5%	1.6	0.57 %
stereo&text	5	50k	83.6%	94.2%	1.6	0.52 %
stereo	10	100k	81.5%	95.9%	2.2	0.59 %
stereo&text	10	100k	82.4%	96.4%	2.3	0.49 %
stereo	20	200k	79.8%	97.1%	3.1	0.65 %
stereo&text	20	200k	80.8%	97.5%	3.2	0.87 %

For every reaction in the USPTO data set [34], we checked that all the molecule names were correctly recognized and could be converted to an InChI [35, 36]. As available molecules for the diversity score, we have taken a data set from emolecules [37]. The validation set, on which the experiment was performed, contained 10k unseen reactions from the USPTO dataset. The experiment were run using the code found on [29].

4 Conclusion

The evaluation of single-step retrosynthetic models is an overlooked research topic. The top-N accuracy, which is usually reported to rank the predictive quality of such models is far from being ideal, as it assesses the information retrieval capabilities rather than the quality of the predictions. Ideally, the optimal assessment would involve human experts. This approach is unfortunately not feasible if not in a few cases. Here we present the use of a chemical prediction model as a surrogate of human expertise to discuss three of four recently introduced metrics based on a forward prediction model: round-trip accuracy, coverage and diversity. The extended analysis, introducing the fourth metric on the statistical significance of the diversity through the Jensen-Shannon divergence can be found in [23].

We proposed metrics to evaluate any retrosynthetic single-step model based on a forward prediction transformer architecture [28]. Compared to recent work [14, 15, 16, 17, 18] using a similar transformer-based retrosynthetic model, we predict not only reactants but also reagents. Despite the increased difficulty of the task, we showed that it is possible to reach a round-trip accuracy of

82.4%. This means that for 82.4% of the precursors suggested by the retrosynthesis model, the forward reaction prediction model predicted the correct product. We hope that this work will initiate a discussion on how to best develop more robust single-step retrosynthesis models while improving their prediction rate and help researchers in the field of machine learning and chemical synthesis design.

References

[1] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[2] Daniel C Elton, Zois Boukouvalas, Mark D Fuge, and Peter W Chung. Deep learning for molecular design-a review of the state of the art. *Molecular Systems Design & Engineering*, 2019.

[3] Daniel Merk, Lukas Friedrich, Francesca Grisoni, and Gisbert Schneider. De novo design of bioactive small molecules by artificial intelligence. *Molecular informatics*, 37(1-2):1700153, 2018.

[4] Francesca Grisoni, Claudia S Neuhaus, Miyabi Hishinuma, Gisela Gabernet, Jan A Hiss, Masaaki Kotera, and Gisbert Schneider. De novo design of anticancer peptides by ensemble artificial neural networks. *Journal of Molecular Modeling*, 25(5):112, 2019.

[5] Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019.

[6] Elias James Corey. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (nobel lecture). *Angewandte Chemie International Edition in English*, 30(5):455–465, 1991.

[7] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604, 2018.

[8] John S Schreck, Connor W Coley, and Kyle JM Bishop. Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 2019.

[9] Connor W Coley, William H Green, and Klavs F Jensen. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5):1281–1289, May 2018.

[10] A Filipa de Almeida, Rui Moreira, and Tiago Rodrigues. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry*, 1:1–16, August 2019.

[11] Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

[12] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.

[13] Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

[14] Pavel Karpov, Guillaume Godin, and Igor Tetko. A transformer model for retrosynthesis. may 2019.

[15] Kangjie Lin, Youjun Xu, Jianfeng Pei, and Luhua Lai. Automatic retrosynthetic pathway planning using template-free models. *arXiv preprint arXiv:1906.02308*, may 2019.

[16] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reaction using self-corrected transformer neural networks. *arXiv preprint arXiv:1907.01356*, July 2019.

[17] Hongliang Duan, Ling Wang, Chengyun Zhang, and Jianjun Li. Retrosynthesis with attention-based nmt model and chemical analysis of the" wrong" predictions. *arXiv preprint arXiv:1908.00727*, August 2019.

[18] Qingyi Yang, Vishnu Sresht, Peter Bolgar, Xinjun Hou, Jacquelyn Klug-McLeod, Christopher Butler, et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 2019.

[19] Alpha A. Lee, Qingyi Yang, Vishnu Sresht, Peter Bolgar, Xinjun Hou, Jacquelyn L. Klug-McLeod, and Christopher R. Butler. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.*, pages –, 2019.

[20] XiangGen Liu, Pengyong Li, and Sen Song. Decomposing retrosynthesis into reactive center prediction and molecule generation. *bioRxiv*, page 677849, 2019.

[21] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

[22] Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. 2018.

[23] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting Retrosynthetic Pathways Using a Combined Linguistic Model and Hyper-Graph Exploration Strategy. 10 2019.

[24] Hiroko Satoh and Kimito Funatsu. Sophia, a knowledge base-guided reaction prediction system-utilization of a knowledge base derived from a reaction database. *Journal of chemical information and computer sciences*, 35(1):34–44, 1995.

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[26] Philippe Schwaller, Alain C. Vaucher, Vishnu H Nair, and Teodoro Laino. Data-Driven Chemical Reaction Classification with Attention-Based Neural Networks. 9 2019.

[27] Greg Landrum, Paolo Tosco, Brian Kelley, sriniker, gedeck, NadineSchneider, Riccardo Vianello, Andrew Dalke, Ric, Brian Cole, AlexanderSavelyev, Samo Turk, Matt Swain, Alain Vaucher, Dan N, Maciej Wójcikowski, Axel Pahl, JP, Francois Berenger, strets123, JLVarjo, Noel O'Boyle, David Cosgrove, Patrick Fuller, Jan Holst Jensen, Gianluca Sforna, DoliathGavid, Karl Leswing, Susan Leung, and Jeff van Santen. rdkit/rdkit: 2019_03_4 (q1 2019) release, August 2019.

[28] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 0(0):null, 0.

[29] Molecular Transformer, https://github.com/pschwllr/moleculartransformer. (Accessed Aug 29, 2019).

[30] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[31] Juno Nam and Jurae Kim. Linking the neural machine translation and the prediction of organic chemistry reactions. *arXiv preprint arXiv:1612.09529*, 2016. (Accessed Aug 29, 2019).

[32] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.

[33] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[34] Daniel Lowe. Chemical reactions from US patents (1976-Sep2016). 6 2017.

[35] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5(1):7, 2013.

[36] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

[37] emolecules, https://www.emolecules.com. (Accessed Aug 29, 2019).