Differential top-k learning for template-based single-step retrosynthesis

Andres M Bran
Laboratory of Artificial Chemical Intelligence (LIAC) &
National Centre of Competence in Research (NCCR) Catalysis,
Institute of Chemical Sciences and Engineering
Ecole Polytechnique Fédérale de Lausanne (EPFL),
andres.marulandabran@epfl.ch

Philippe Schwaller
Laboratory of Artificial Chemical Intelligence (LIAC) &
National Centre of Competence in Research (NCCR) Catalysis,
Institute of Chemical Sciences and Engineering
Ecole Polytechnique Fédérale de Lausanne (EPFL),
philippe.schwaller@epfl.ch

Abstract

Retrosynthesis is one of the core tasks in the organic molecule design cycle, yet it is still a computational challenge to produce suitable sets of precursors for a desired product. Commonly used template-based approaches reduce the problem to a multi-class classification task for single steps. However, reactions in available datasets are noisy and incomplete, making usual training methods problematic. In this work, considering that multiple disconnections are possible for a product, we propose training models using differential top-k losses. We show that using these loss functions yields improvements in every top-N metric, with little overhead relative to cross-entropy. The use of more powerful models, more diverse and complete datasets, and other methodologies, is expected to yield significant improvements on this task when combined with the training approach presented here.

1 Introduction

The development of novel materials and molecules is at the foundation of societal development towards a sustainable future. However, synthesis –i.e. the reaction steps for how to make a given molecule– remains one of the crucial bottlenecks during the design cycles of such substances [1], heavily slowing down the discovery and production of such novel materials and molecules. Human experts attempt to solve this problem by applying their chemical knowledge in retrosynthetic planning, where the goal is to find a synthetic path for a target molecule, such that it can be resolved to commercially available or easily synthetically accessible substances. The search is performed backwards by inspection of the target molecule, while suggesting sets of precursors that, upon reaction under adequate conditions, would lead to the desired product.

Current computational approaches for the multi-step problem [2–6] pair single-step retrosynthetic prediction models with search algorithms, that iteratively compose single-step predictions until a suitable synthetic plan is obtained from commercially available starting materials. In turn, multiple strategies have been published towards solving the single-step problem [7], such as template-based [8–12], graph edit-based [13,14] and sequence-based approaches [4,15–18]. Since the pioneering work of Corey [19,20], researchers have attempted to encode chemical transformations into expert-curated [21,22] and auto-

36th Conference on Neural Information Processing Systems, AI4Mat workshop (NeurIPS 2022).

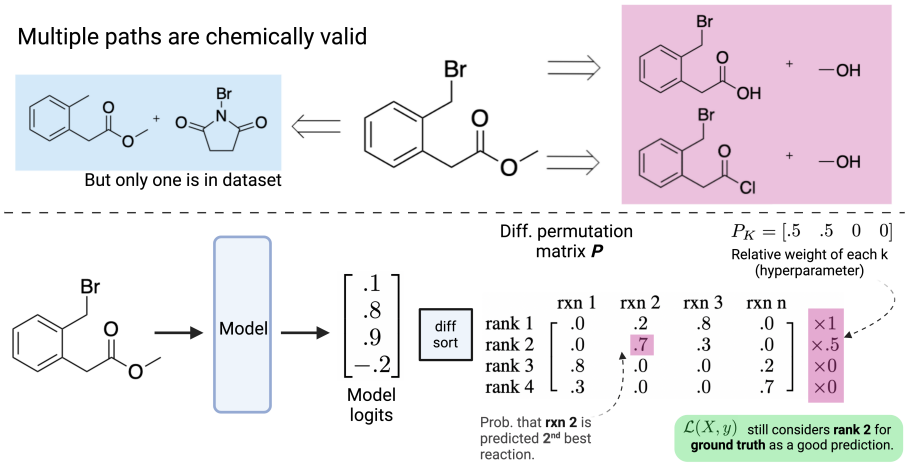

Figure 1: **Top.** Multiple plausible synthetic paths exist for a given product, however only a small subset has been reported. Using cross entropy loss, models get punished for proposing synthetic paths alternative to the ground truth. **Bottom.** Differentiable top-k based loss functions (below) help this by also considering higher ranks (k) as valid, their importance given by a distribution P_K.

matically extracted reaction rules[2,8], that describe types of reactions found in data –very much like "name reactions" in organic chemistry. When a template library is built, the single-step retrosynthetic problem can be reframed as a multi-class classification problem[2], where the task is to find the most suitable reaction leading to the desired product. Although such approaches are limited by the size of the template library, they are popular in synthesis planning tools thanks to their speed, as the problem is reduced to a much simpler classification task. In this work, we focus solely on template-based and single-step retrosynthesis approaches, and how mapping the training task closer to reality leads to overall prediction improvements.

The challenge with single-step retrosynthesis task is that multiple correct possibilities exist to synthesize the same target molecule[15] (Figure 1 top). However, typically only one of the correct possibilities is recorded in the current reaction benchmark datasets and hence the ground truth is generally incomplete. As reactions other than the ground truth ones could equally be correct, the cross entropy loss on which previous template-based approaches were trained is ill-posed[8–10,12].

Inspired by recent improvements in image classification by Petersen et al.[23], we introduce differential top-k learning (DTk) for the single-step retrosynthetic task (Figure 1 bottom). We show that a straightforward change in the loss function, which relaxes the distribution over the target templates, leads better top-N accuracies compared to the same models trained on top-1 cross entropy. Most strikingly, even the top-1 prediction accuracy typically increases as a result. We conclude that the use of this type of cost functions is better suited for this task, both from the chemical and the computational perspectives. We expect DTk learning to become valuable for other applications in machine learning for chemical sciences, where the ground truth is incomplete and contains only a fraction of the correct classes.

2 Methods

2.1 Template-based single-step retrosynthesis models

Several template-based models and molecular representations have been proposed in the recent years to tackle the single-step retrosynthesis problem. In the seminal work of Segler et al.[2], the authors used a highway neural network architecture[24] (NeuralSym) that takes as input an ECFP4 fingerprint of the desired molecule[25], to predict the most suitable template from a set of automatically extracted transformations. More recent approaches are based on conditional graph logic networks (GLN)[9], modern Hopfield networks[12], and re-ranking of predictions using energy-based models[26], which all improve over the simple highway neural network baseline. Though not ideal, all studies use top-k

accuracy, whether the ground truth is found in the k most likely prediction or not, as a proxy for evaluating and comparing single-step retrosynthesis models.

We investigated NeuralSym[2], GLN[9] and a Transformer classifier (ChemBERTa)[27], whose input is the molecular SMILES of the desired product. NeuralSym and ChemBERTa were trained and tested using a set of 15 random seeds each time, using the same set of hyperparameters for each loss function (see Appendix A.1). GLN was only run once for each loss function, due to its high computational cost during training and inference.

2.2 Differential top-k learning

Conventionally used loss functions for classification tasks are built to maximize top-k, for a given positive integer k[28,29], cross entropy loss being the special case for $k = 1$. Building on recent advances in differentiable sorting[30,31], Petersen et al.[23] further relaxed the need to choose a value of k, and instead use a distribution over this parameter, giving a relative importance to each top-k objective. They thus propose a family of loss functions, each being specified by such distribution. The authors thoroughly test the method in a range of benchmarks for image classification, and show that mixed top-k strategies not only improve top-5 accuracies but also top-1 accuracy, achieving a new state-of-the-art on ImageNet[32].

Differentiable sorting works by producing a differentiable permutation matrix $\mathbf{P}$[30] from a model's raw logits $f_\Theta(X)$. Each entry $\mathbf{P}_{k,j}$ of this matrix is interpreted as the predicted probability that class j is the k-th best prediction. Finally, the loss function is fully specified by the distribution over ks P_K, so that $P_K(k)$ is the relative importance assigned to the pure top-k objective.

The model is then trained by minimizing the following loss function

$$\mathcal{L}(X, y) = -\log\left(\sum_{k=1}^{n} P_K(k)\left(\sum_{m=1}^{k} \mathbf{P}_{m,y}(f_\Theta(X))\right)\right). \tag{1}$$

To highlight the importance of top-k predictions other than top-1, and thus the advantage of DTk learning, examples are shown in the Appendix A.3, where models fail to predict the ground truth precursors. In the cases where the ground truth is highly ranked, usually the other highly ranked predictions also correspond to equally valid disconnections and possible transformations. However the cross entropy loss still punishes the model for these "wrong" predictions. DTk learning prevents this, thus giving more flexibility to the models which in turn allow them to predict more diverse disconnections and reaction types.

2.3 Data and representation

Schneider et al.[33] initially published the USPTO 50k dataset, and Liu et al.[15] later curated it and created the retrosynthesis benchmark dataset. It contains 50k reactions from 10 reaction superclasses. We use the same train/valid/test split as used by Lin et al.[26], and the same template library for NeuralSym and ChemBERTa. The templates are extracted from atom-mapped reactions[34], and their quality might be limited by the quality of the mapping[35]. For the NeuralSym models, the target molecules are represented with extended-connectivity fingerprints (ECFPs)[25]. The ChemBERTa model directly takes tokenized SMILES as input. For GLN, the data is featurized as in the original paper[9].

3 Results & Discussion

We investigate three types of template-based single-step retrosynthesis prediction models on the USPTO 50k dataset: NeuralSym[2] based on the highway neural network architecture[24], Graph Logic Network (GLN)[9], and a SMILES Transformer Classifier based on the pretrained ChemBERTa architecture[27]. Each model is trained using the standard cross entropy loss function as a baseline, and a set of selected P_k distributions for DTk learning. As the original GLN model used a different optimization objective, the architecture was slightly modified to train with an explicit cross entropy loss, allowing in turn training with DTk losses. Thus for this architecture we additionally report results of training with the original loss as well as with cross entropy and DTk losses.

As shown in Table 1, trainig with DTk losses typically provide accuracy improvements of as much as 3% in some top-k accuracies relative to the cross entropy baseline on the same model. Note that such gains in accuracy are achieved by simply changing the loss function while using the same hyperparameters as for cross entropy. This method thus provides better results with little extra overhead. In general we find that the best results are achieved typically with a combination of strong top-1 ($9/10$ in table) with small additions of other top-k objectives, which tend to yield better top-1 test accuracies as well as improvements in the other metrics. Other combinations work good as well in some cases, such as flat distributions over the first n values of k, e.g. $\{1/3, 1/3, 1/3\}$ or $\{1/4, 1/4, 1/4, 1/4\}$ and so on. However, training of the transformer model appears to be unstable under some DTk losses, yielding far from optimal results in these particular cases. Further results are shown in the Appendix A.2.

Notably, using DTk loss on the very simple NeuralSym architecture improves top-3 and top-5 accuracies, relative to the optimal results with the much more complex GLN model[9]. This shows the potential of this training technique for this, as well as other classification tasks in chemistry[36]. Further experiments are however required to determine training strategies that can work better in general, across models and tasks. This is however left as future work.

Model type	Loss function	Top-k accuracy				
		1	3	5	10	20
NeuralSym	Cross entropy	45.43	66.85	74.09	81.34	85.92
	P_k = {1/5,1/5,1/5,1/5,1/5}	46.24	69.12	76.79	83.12	86.55
	P_k = {1/4,1/4,1/4,1/4, 0}	46.21	**69.26**	**76.93**	83.29	86.66
	P_k = {1/3,1/3,1/3, 0, 0}	46.24	**69.25**	**76.90**	83.27	**86.74**
	P_k = {1/10,0,0,0, 9/10}	45.56	68.90	76.87	**83.37**	86.68
	P_k = {9/10,0,0,0, 1/10}	46.43	68.21	75.22	81.71	85.49
	P_k = {9/10,0,0, 1/10,0}	**46.57**	68.14	75.43	81.59	85.44
ChemBERTa	Cross entropy	44.08	66.61	74.17	81.47	85.50
	P_k = {1/5,1/5,1/5,1/5,1/5}	43.68	**67.30**	**74.90**	82.09	86.16
	P_k = {1/4,1/4,1/4,1/4, 0}	22.40	40.88	50.28	61.29	66.64
	P_k = {1/3,1/3,1/3, 0, 0}	37.95	59.96	68.76	77.30	82.03
	P_k = {1/10,0,0,0, 9/10}	42.50	66.52	74.68	82.16	86.15
	P_k = {9/10,0,0,0, 1/10}	44.75	67.03	74.72	82.05	86.09
	P_k = {9/10,0,0,1/10, 0}	**45.10**	**67.21**	**74.87**	**82.33**	**86.34**
GLN	Original	**52.27**	66.56	74.05	82.28	88.09
	Cross entropy	51.27	66.32	73.07	81.82	88.37
	P_k = {1/5,1/5,1/5,1/5,1/5}	51.21	66.76	74.05	82.50	88.95
	P_k = {1/4,1/4,1/4,1/4, 0}	50.89	66.25	73.50	82.54	88.24
	P_k = {1/3,1/3,1/3, 0, 0}	51.55	**67.56**	**75.19**	**83.46**	88.71
	P_k = {1/10,0,0,0, 9/10}	50.43	66.70	74.33	83.02	**89.15**
	P_k = {9/10,0,0,0, 1/10}	**52.27**	66.10	73.69	81.74	88.45
	P_k = {9/10,0,0, 1/10,0}	51.95	66.63	73.28	81.81	87.88

Table 1: Top-k accuracies on test set, for each combination of model type and loss function tested. DTk stands for Differential Top-k loss with the P_k distribution shown in front. Numbers in bold correspond to the maximum top-k accuracy achieved for the model type specified on the left column. The numbers for NeuralSym and ChemBERTa models are average accuracies for 15 models trained with the same parameters, but with the different random seeds shown in Appendix A.1. For GLN, only one model was trained due to its long training and inference time. The additional "Original" loss function is reported for GLN, as the original implementation[9] had to be modified for training with cross entropy loss.

4 Conclusion

Cross entropy is the most commonly used loss function for classification tasks in machine learning. Here we argue that this function is not adequate for the task of single-step retrosynthesis planning, as datasets for this task are incomplete and different reactions are usually as valid. Differentiable top-k learning is proposed to alleviate this limitation, and it is shown that using this strategy for training systematically leads to improvements on top-k accuracies. Notably, we show that the adoption of a

DTk strategy can make a simple model like NeuralSym[2] overperform much more complex models, like GLN[9], in certain top-k accuracies. This shows that the use of such loss functions can boost the performance of retrosynthesis models when coupled with powerful models. We additionally expect this approach to be useful and become a standard for other multi-label classification tasks in chemistry[36,37], beyond template based retrosynthesis. In the future, our work will be extended to consider more diverse datasets, such as USPTO-full[9], and more powerful models[12]. The code and data to reproduce the results will be made available upon publication.

Acknowledgments

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References

[1] Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry* **2018**, *10*, 383–394.

[2] Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

[3] Coley, C. W.; Thomas III, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, eaax1566.

[4] Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science* **2020**, *11*, 3316–3325.

[5] Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical science* **2020**, *11*, 154–168.

[6] Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics* **2020**, *12*, 1–9.

[7] Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, e1604.

[8] Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **2017**, *3*, 1237–1245.

[9] Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems* **2019**, *32*.

[10] Bjerrum, E. J.; Thakkar, A.; Engkvist, O. Artificial applicability labels for improving policies in retrosynthesis prediction. *Machine Learning: Science and Technology* **2020**, *2*, 017001.

[11] Chen, S.; Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **2021**, *1*, 1612–1620.

[12] Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G. Improving Few-and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *Journal of chemical information and modeling* **2022**, *62*, 2111–2120.

[13] Sacha, M.; Błaz, M.; Byrski, P.; Dabrowski-Tumanski, P.; Chrominski, M.; Loska, R.; Włodarczyk-Pruszynski, P.; Jastrzebski, S. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling* **2021**, *61*, 3273–3284.

[14] Somnath, V. R.; Bunne, C.; Coley, C.; Krause, A.; Barzilay, R. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems* **2021**, *34*, 9405–9415.

[15] Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **2017**, *3*, 1103–1113.

[16] Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling* **2019**, *60*, 47–55.

[17] Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature communications* **2020**, *11*, 1–11.

[18] Wang, X.; Li, Y.; Qiu, J.; Chen, G.; Liu, H.; Liao, B.; Hsieh, C.-Y.; Yao, X. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal* **2021**, *420*, 129845.

[19] Corey, E. J. General methods for the construction of complex molecules. *Pure and Applied chemistry* **1967**, *14*, 19–38.

[20] Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **1985**, *228*, 408–418.

[21] Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-assisted synthetic planning: the end of the beginning. *Angewandte Chemie International Edition* **2016**, *55*, 5904–5937.

[22] Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P., et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **2018**, *4*, 522–532.

[23] Petersen, F.; Kuehne, H.; Borgelt, C.; Deussen, O. Differentiable Top-k Classification Learning. International Conference on Machine Learning (ICML). 2022.

[24] Srivastava, R. K.; Greff, K.; Schmidhuber, J. Highway networks. *arXiv preprint arXiv:1505.00387* **2015**,

[25] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

[26] Lin, M. H.; Tu, Z.; Coley, C. W. Improving the performance of models for one-step retrosynthesis through re-ranking. *Journal of cheminformatics* **2022**, *14*, 1–13.

[27] Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. **2020**, arXiv:2010.09885 [physics, q-bio].

[28] Lapin, M.; Hein, M.; Schiele, B. Loss functions for top-k error: Analysis and insights. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 1468–1477.

[29] Berrada, L.; Zisserman, A.; Kumar, M. P. Smooth loss functions for deep top-k classification. *arXiv preprint arXiv:1802.07595* **2018**,

[30] Petersen, F.; Borgelt, C.; Kuehne, H.; Deussen, O. Differentiable Sorting Networks for Scalable Sorting and Ranking Supervision. International Conference on Machine Learning (ICML). 2021.

[31] Petersen, F.; Borgelt, C.; Kuehne, H.; Deussen, O. Monotonic Differentiable Sorting Networks. International Conference on Learning Representations (ICLR). 2022.

[32] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. CVPR09. 2009.

[33] Schneider, N.; Stiefl, N.; Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling* **2016**, *56*, 2336–2346.

[34] Thakkar, A.; Reymond, J.-L. Automatic Extraction of Reaction Templates for Synthesis Prediction. *CHIMIA* **2022**, *76*, 294–294.

[35] Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, eabe4166.

[36] Saini, K.; Ramanathan, V. Predicting odor from molecular structure: a multi-label classification approach. *Scientific reports* **2022**, *12*, 1–11.

[37] Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel classification models for the prediction of cross-coupling reaction conditions. *Journal of Chemical Information and Modeling* **2021**, *61*, 156–166.