[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.

A Appendix

A.1 Random seeds for training

NeuralSym and ChemBERTa classification models were trained using the set of seeds shown in Table A2, with a constant set of hyperparameters but varying the loss function.

7137799	129388	7971049	813804	6215678
9672708	131184	9718656	3685980	839341
7687853	3472862	3928806	3347752	8066535

Table A2: All the trained NeuralSym and ChemBERTa models were trained on this set of seeds.

A.2 Other DTk distributions.

In the context of differentiable top-k loss functions, we define a *pure k training approach* as that where a single k is given maximum importance, i.e. $P_K(k) = 1$. We experimented with such strategies for k between 1 and 5 with the NeuralSym architecture, and report the obtained top-k accuracies in the template task for k in [1, 2, 3, 4, 5, 10, 20] . The results of this are show in Table A3.

Loss Function	Top-k template accuracy						
	1	2	3	4	5	10	20
Cross entropy	**40.12**	50.94	56.51	59.70	61.79	67.21	71.35
P_k = {0,1,0,0,0}	39.06	**51.98**	**58.19**	**61.72**	**64.02**	69.16	72.52
P_k = {0,0,1,0,0}	36.31	50.59	57.66	61.43	63.91	69.26	72.63
P_k = {0,0,0,1,0}	31.17	43.83	50.25	54.65	56.97	65.27	70.39
P_k = {0,0,0,0,1}	33.58	47.85	55.41	60.00	63.05	**69.39**	**72.93**

Table A3: Top-k accuracies on the template prediction task, from training with a pure k approach, for k between 1 and 5. The case with k=1 is just the exact cross entropy.

The results for k>1 show in general a lower performance in top 1 accuracy, they do however tend to perform better for top-2 to top-5 accuracies, especially the pure approach with k=2. Top-10 and top-20 accuracies are also generally improved, and the best performance is achieved by the approach with k=5. The results show that it is in principle possible to improve in each one of these metrics, only by targeting the cost function with appropriate values of k.

A.3 Model predictions

This section explores several cases in which top-k accuracy evaluation is inadequate for assessing the performance of models in the one-step retrosynthesis task. The attention is centered in two cases: (1) when the ground truth is not the top-1 prediction, but is found within the first top-10 predictions, and (2) when ground truth is not found within the top-10 predictions, but the model still predicts applicable templates. In our test set, we find that the first case occurs for 27.35% of the test products, while the second is the case for 15.88% of them.

A.3.1 Case 1: Ground truth predicted in top-(k>1).

Normal cross entropy punishes the model for not classifying the ground truth as the top-1 prediction, however in the cases in which the ground truth label is found within the top-k predictions, the model also predicts reasonable and more diverse disconnections with a higher rank (Figure A2). Normal cross entropy Loss thus hinders learning, as it prevents the model from proposing paths different than the typically incomplete ground truth.

A.3.2 Case 2: Ground truth is not within top-k.

In these cases, we find that the models generally lack understanding of the chemical environment, and thus the top-k predicted templates either are not applicable, or are at best risky options, as other reactions could undergo due to the presence of interfering functional groups (Figure A3).

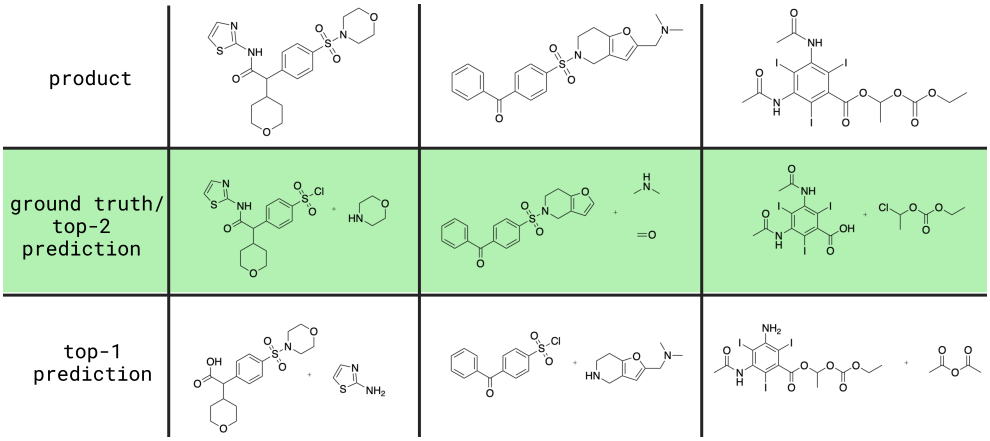

Figure A2: Example of model predictions where ground truth is not classified as the best template, but it is found within the top-4 predictions. The figure illustrates how in these cases, most of the other predicted top-4 sets of precursors corresponds to equally valid disconnections.

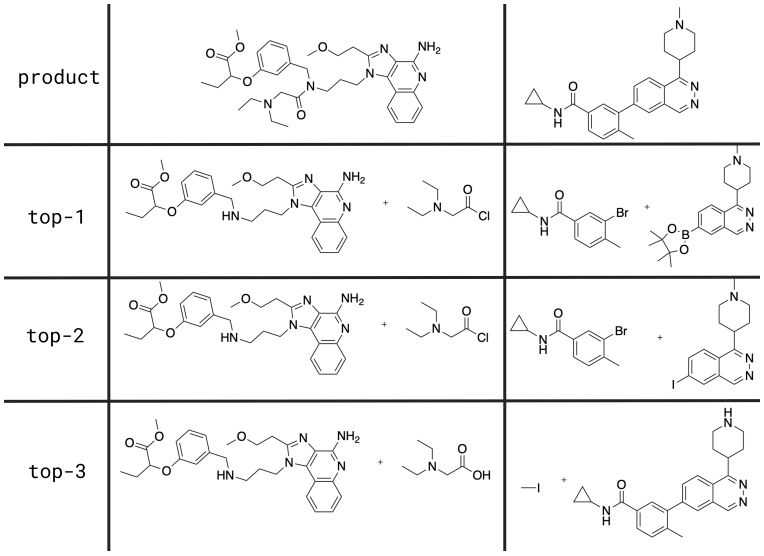

Figure A3: Model results where ground truth is not found within top-10 predictions. These cases tend to be more complex and require careful consideration of the chemical environment.