Clustering of synthetic routes using tree edit distance

Samuel Genheden, Ola Engkvist, Esben Bjerrum

Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, Sweden

* Corresponding author: samuel.genheden@astrazeneca.com

Abstract

We present a novel algorithm to compute the distance between synthesis routes based on a tree edit distance calculation. Such distances can be used to cluster synthesis routes from a retrosynthesis prediction tool. We show that the clustering of routes from a retrosynthesis analysis is performed in less than ten seconds on average, and only constitutes seven percent of the total time (prediction + clustering). Furthermore, we are able to show that representative routes from each cluster can be used to reduce the set of predicted routes. Finally, we show with a number of examples that the algorithm gives intuitive clusters that can be easily rationalized. The algorithm is included in the latest version of the open-source AiZynthFinder software.

Keywords: computer-aided synthesis prediction, retrosynthesis analysis, tree edit distance, clustering

Introduction

Computer-aided synthesis prediction is an important tool in medicinal and process chemistry because it can provide suggestions on how to synthesize a compound (retrosynthesis analysis), how to optimize the reaction condition and evaluate the feasibility of a reaction. Most of such algorithms have benefited from the rise of deep learning methods and data-driven approaches in the last decade.[1,2,3,4] The synthesis of a target compound can be described as a route or reaction tree, showing what reactions need to be carried out in order to produce the target compound. The precursors of the target compound are typically smaller molecules (building blocks) available in a storage or can be synthesized from such molecules.

Usually, a retrosynthesis tool predicts more than one route and it is up to a chemist or a downstream software to select what routes to proceed with. A well-designed scoring function could aid with this task and in addition to simple ones such as the number of steps or the total price of precursors, more elaborate scoring functions have been suggested such as a recursive price estimator[5] and the aggregation of single-step likelihoods.[6,7] However, many routes are typically very similar and differ for instance only in the kind of halogen substituent on one of the precursors or two pairs of pre-cursors producing the same two products. In such situations, it is unclear if a scoring function is able to distinguish between competing routes. If this is the case, clustering of the predictions could help in deciding what routes to proceed with by reducing the set of predicted routes giving a better overview of the suggestions. Recently Mo et al. suggested an LSTM-based neural network architecture to encode routes[8] and used it to distinguish between predicted routes from ASKCOS[9] and human-designed routes. Relevant for our aim with route clustering, they also suggested using the latent space encoding of the routes as the basis for distance calculation and clustering.

In this text, we will take another approach and introduce a novel clustering algorithm that is based on a tree edit distance calculation.[10] The algorithm was implemented in the AiZynthFinder retrosynthesis tool,[11,12] and we will show that it produces intuitive clusters in a reasonable time.

Methods

Route predictions. We selected 5,000 random compounds from ChEMBL[13] and tautomeric state was selected with RDKit.[14] The SMILES strings of the compounds were used as input to the AiZynthFinder software to predict synthetic routes. The expansion and filter policies used in the search were derived from the USPTO dataset, as discussed previously.[11,15] In-house available and Enamine building blocks were used as termination criteria. The search was performed for 100 iterations, after which between 5 and 25 routes were extracted, depending on the scores[11] of the routes.

Distance calculations. A synthetic route or a reaction tree is a bipartite tree consisting of molecule and reaction nodes, with the target molecule as the root. The distance between two trees (T_1 and T_2) can be computed by a tree edit algorithm.[10] The algorithm consists of three possible operations: 1) insertion of a node, 2) deletion of a node, and 3) substitution of two nodes. For each of these operations, we define a cost. The tree edit distance (TED) is then defined as the minimum-cost sequence of such operations that transform T_1 into T_2. To compute TED we use the APTED (all path TED) algorithm,[16,17] which is available as a python package.[18] APTED only guarantees an optimal solution for an ordered tree, i.e., a tree where the children of a node have an inherent order. A reaction tree is however an unordered tree, but finding the solution to such a tree is NP-complete. Specialized algorithms to compute TED for unordered trees have been suggested, [10,19,20] but we found none of them appropriate for our task or a reference implementation was missing. Therefore, we decided to impose a number of heuristics on top of the APTED algorithm.

These heuristics are based on the observation that the branching factor of a reaction tree is small (a reaction node typically has one or two children, and at maximum five) and that the size of the reaction tree is small. Therefore we can in many instances enumerate all possible trees for a reaction tree by permuting the children (see Figure 1), and we define the number of possible trees as N_T. If the product of $N_T(T_1)$ and $N_T(T_2)$, i.e., the number of possible combinations of trees for T_1 and T_2, is low we can afford to do an exhaustive search and compute the minimum TED over all tree combinations. We set this limit to 20, based on some small tests. If the product $N_T(T_1)$ and $N_T(T_2)$ is larger than 20, but at least one of $N_T(T_1)$ and $N_T(T_2)$ is at most 20, we do a semi-exhaustive search: we compute the minimum TED over all enumerations of the tree with the smallest N_T and a single representation of the other reaction tree. If both $N_T(T_1)$ and $N_T(T_2)$ are larger than 20, then we generate 20 random enumerations of T_1 and T_2 and compute the minimum TED over all these enumerations.

The insertion and deletion cost is set to unity, and the substitution cost is set to the Jaccard distance[21] between the fingerprints of a node. The fingerprint of a molecule node was taken as 2048-bit fingerprint (ECFP4, computed by the Morgan algorithm in RDKit[14,22]). The fingerprint of a reaction node was taken as the difference between the fingerprints of the reactants and products.

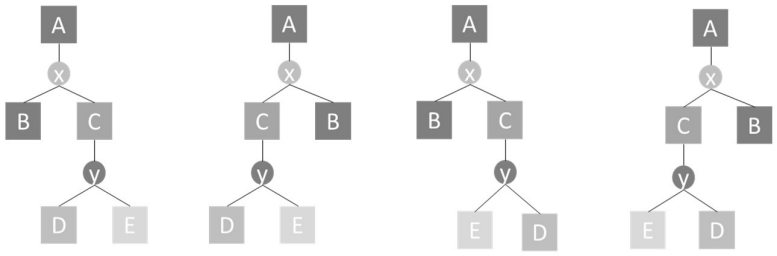


Figure 1 – Four possible enumerations of the same reaction tree. Molecule nodes are A, B, C, D and E, and reaction nodes are x and y.

Clustering. The clustering is based on the tree edit distance matrix. Because TED is not a distance in Cartesian space, we used hierarchical clustering with single linkage as implemented in SciKit-Learn.[23] The optimal cluster size was determined by the Silhouette method.[24] Because the number of analyzed routes (<=25) was small, the maximum number of clusters was set to 5. The representative of each cluster was taken as the route with the highest prediction score.[11]

LSTM-based clustering. We downloaded the LSTM (long short-term memory)-based neural network model of Mo et al. from Github. [8,25] The available model is trained from molecular fingerprints of size 2048, and an LSTM output-size of 256. The routes from the AiZynthFinder predictions were fed to the model producing the latent space encodings of the routes. The latent space encodings were then used to compute a distance matrix with a Euclidean metric. The distance matrix was finally used in clustering as described above for TED. For predictions only consisting of a single molecule (the target compound), the computation was skipped because it is not supported by the network architecture.

Results and discussion

We will first discuss some statistics of the clustering based on the predictions of all the 5,000 compounds selected from ChEMBL. We will then show some illustrative examples of the synthetic route clustering for a few selected compounds. Finally, we will compare the TED-based clustering approach to the deep learning method of Mo et al. [8]

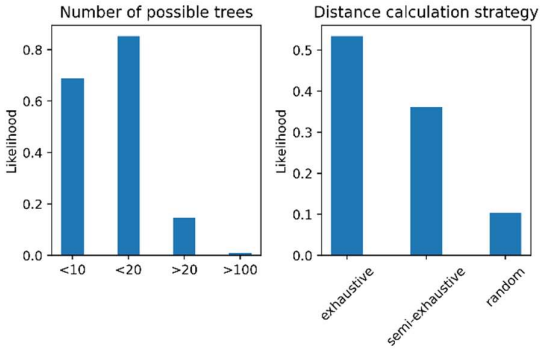


Figure 2 – Statistics from the distance calculations. Left) bar chart showing parts of the distribution of the number of possible trees (N_T), Right) the likelihood of using a particular strategy when computing the distance between two routes.

Predicted routes tend to be small and trees can easily be enumerated. The motivation of the heuristics imposed upon the APTED method was based on the observation that the predicted routes are generally small. For the 51,694 routes produced for the 5,000 compounds in this study, the average number of reactions and molecules is 3.8 and 7.5, respectively. As shown in Figure 2, number of possible trees (N_T) is less than 100 for a majority of those routes. At the cut-off we used for the distance

calculation, we capture 85% of the trees. For only a very small fraction of routes, we would have to enumerate more than 100 trees. In the distance calculation, the relatively small N_T leads to an exhaustive search in 53% of the computations (see Figure 2). For 90% of the distance calculations, either an exhaustive or semi-exhaustive strategy was used. As an alternative to these heuristic strategies, we also explored an option to form ordered tree by sorting the molecule nodes on their InChI keys. And although the TED computed in this way agrees with the heuristic TED in 43% of the comparisons made in this study, in 54% of the comparisons we could find a shorter distance with the heuristic approach.

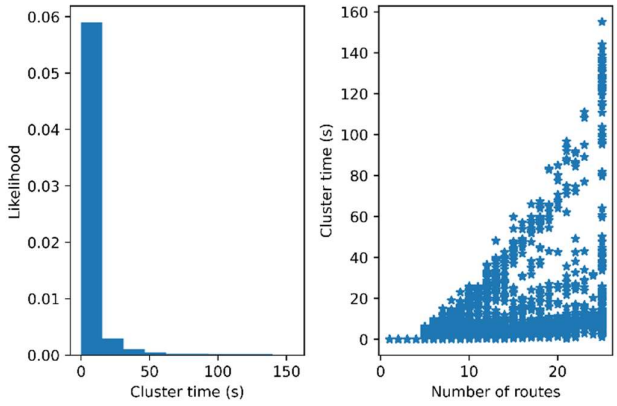


Figure 3 – Timings of the clustering: Left) The distribution of the clustering time over all 5,000 ChEMBL compounds, Right) the relationship between number of analyzed routes and the cluster time.

The clustering algorithm is on average fast. On average, the time to complete the clustering of the routes for a compound is 6 s. The average route prediction time is 78 s and compared to that the average clustering time is fast. On average the clustering only amounts to 6% of the total time (prediction + clustering). However, the distribution of clustering time is heavily skewed and has a long tail (see Figure 3), and although the median time is 2 s the worst time is 155 s. The clustering time is naturally correlated with the number of routes, but the correlation is not clear as seen in Figure 3 as the shape of the route is also an important factor. There are compounds for which 25 routes were clustered in less than 1 s as well as there are compounds for which 25 routes were clustered in 3 minutes. However, for 96% of the compounds, the clustering was done in less than 30s, which is acceptable considering the average route prediction time.

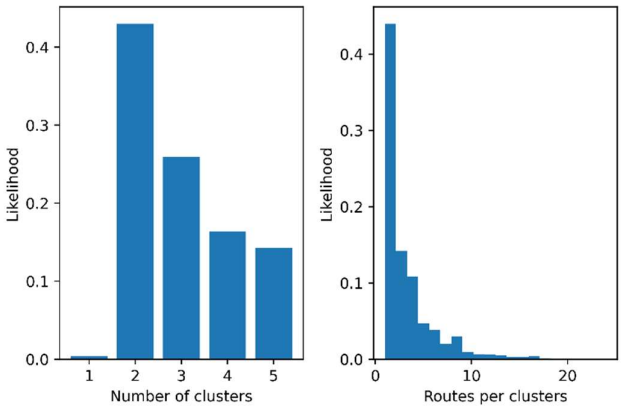


Figure 4 – Statistics of the cluster optimization: Left) the likelihood of forming a particular number of clusters, Right) The distribution of the number of routes per cluster

The cluster optimization produced few clusters with a small number of routes. Considering the small number of routes analyzed per compound (<=25), we set an upper limit when optimizing the number of clusters to five. The distribution of the number of formed clusters is shown in Figure 4, and for almost half of the compounds, the optimum number of clusters is two. If we then consider how many routes there are in each cluster, we obtain the distribution shown in Figure 4, showing that most of the clusters contain rather few routes. 85% of the clusters contain at most 5 routes, and 96% of the clusters contain at most 10 routes. If we would extract more routes from the retrosynthesis analysis, these statistics would of course change, but in our experience, there is rarely any point in inspecting more than the top-25 routes.

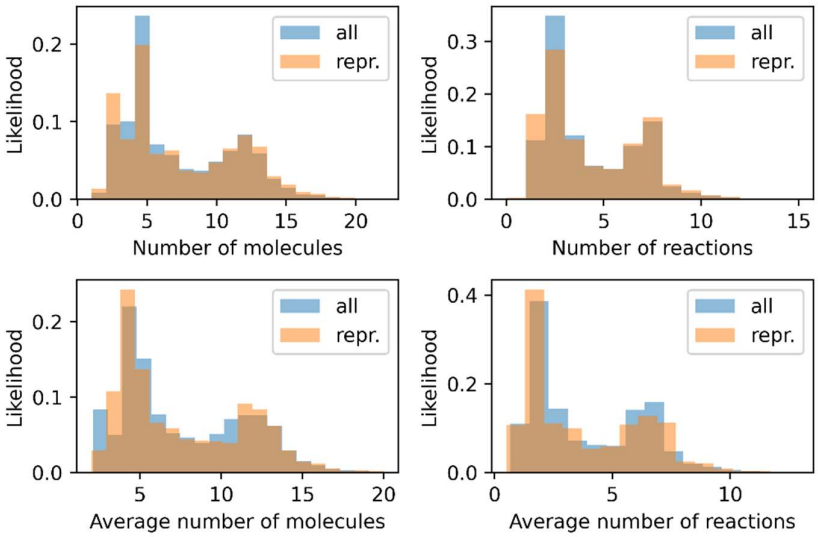


Figure 5 – Distribution of all routes or representative routes (repr.): top) the number of molecules and reactions in the routes, bottom) the average number of molecules and reactions for a compound

The clustering preserves the distribution of the shape of routes. If we select one route from each cluster it should represent all the other routes in that cluster, which implies that the distribution of all routes for a compound should be similar to the distribution of the representative routes. To analyze if the TED-based clustering algorithm has this property, we first looked at the distribution of the number of molecules and the number of reactions among all routes or only the representative routes (the highest scored route in each cluster). As seen in Figure 5, the distribution from all routes and all representatives are very similar. It seems that the clustering leads to a selection of slightly shorter routes, both in terms of number of molecules and number of reactions – but the difference is not great. Next, we computed the average number of molecules and reactions for a compound if we considered all routes or only representative routes. Also, this analysis shows that the clustering algorithm preserves the distribution of the routes (see Figure 5). Here, we see a larger difference in the distribution of the average number of reactions compared to the distribution of the average number molecules. However, the overall shape of the distributions is preserved by the clustering algorithm. Thus, we can conclude that representative routes from the clustering can be used to reduce the set of predicted routes.

Tree edit distance-based clustering produces qualitatively intuitive clusters. To show a few illustrative examples of routes and the clusters formed, we selected three compounds. In the main text, we will show the cluster representatives for the compounds (Figure 6-8), and in the supporting information, we will show all of the routes (Table S1-S3).

For the first compound, which can be synthesized in two simple steps, the routes differ mainly in what order the two reactions are taking place, i.e. whether the pyrrolidine is attached first or second (see Table S1). The other difference lies in the substitution on the methylthiophene molecule, whether it is a hydroxyl group, a keto group, or a bromine. The clustering algorithm produces two clusters and the representative routes are shown in Figure 6. The two clusters are made up of routes where the pyrrolidine is attached first and second, respectively – a natural and intuitive grouping.

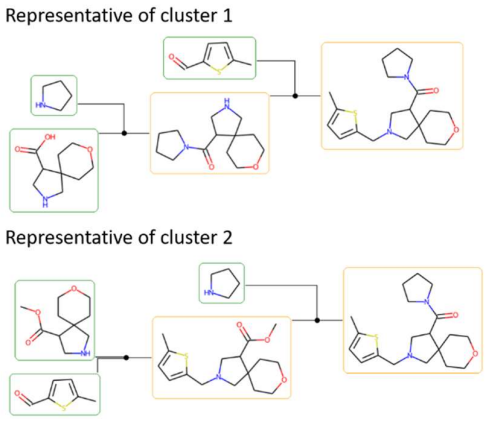


Figure 6 – Cluster representatives for the first example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

The second example is a slightly more complex compound to synthesize as it is made up of two aromatic cycles and two non-aromatic cycles. The representative clusters are shown in Figure 7. The main reactions are two arylation reactions that form bonds to the two non-aromatic cycles and one Suzuki coupling that form the bond between the two aromatic rings. In cluster 1, the order is arylation, followed by Suzuki coupling and finally another arylation. There is another route in this cluster that has an additional Suzuki coupling to attach a methyl group to one of the aromatic rings (see Table S2). In cluster 2, the routes start with a Suzuki coupling, followed by two arylations. Furthermore, some of the routes in this cluster contain some additional halogenations. Cluster 3 is formed from the most elaborate routes. Both routes in this cluster start with a protection step that enables the following tosyloxy alkylation reaction. The third reaction is the Suzuki coupling, followed by a deprotection step and finally an arylation. Again, it is clear that the clustering groups similar routes based on the order of reactions and complexity of the route.

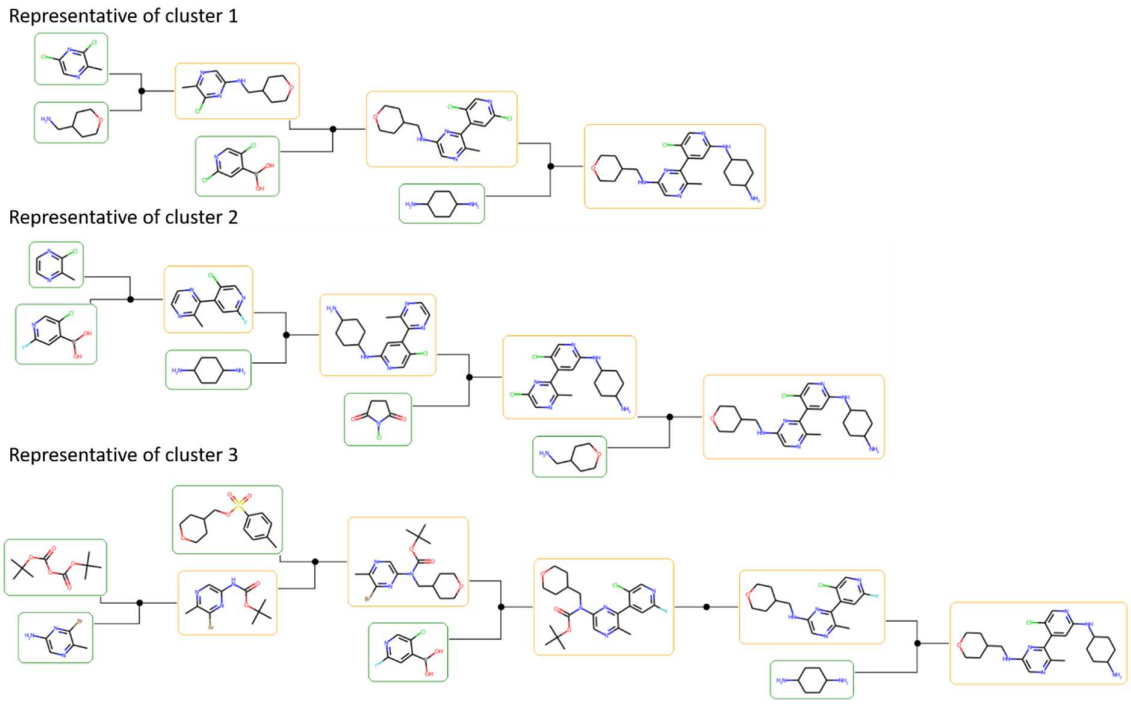


Figure 7 – Cluster representatives for the second example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

The third and final example highlights a convergent route where in the last step an ether bond is formed by a substitution reaction. What differs between the clusters lies in how the two molecules forming the ether bond are synthesized. The first molecule is formed by a Suzuki coupling followed by a reduction in clusters 1 and clusters 2, but in cluster 3 the reduction is not necessary. The second molecule forming the ether bond is synthesized by forming a substituted 2,5-pyrroledione. In cluster 1 and 2, the pyrroledione is formed from an anhydride and a substituted cyclopentane followed by a reduction, whereas in cluster 2 it is formed from a ring-forming reaction. The routes within the clusters differ mainly in what precursors are used in the Suzuki coupling and the substituent of the anhydride (see Table S3). As with the other example compounds, this is also a reasonable clustering of the routes.

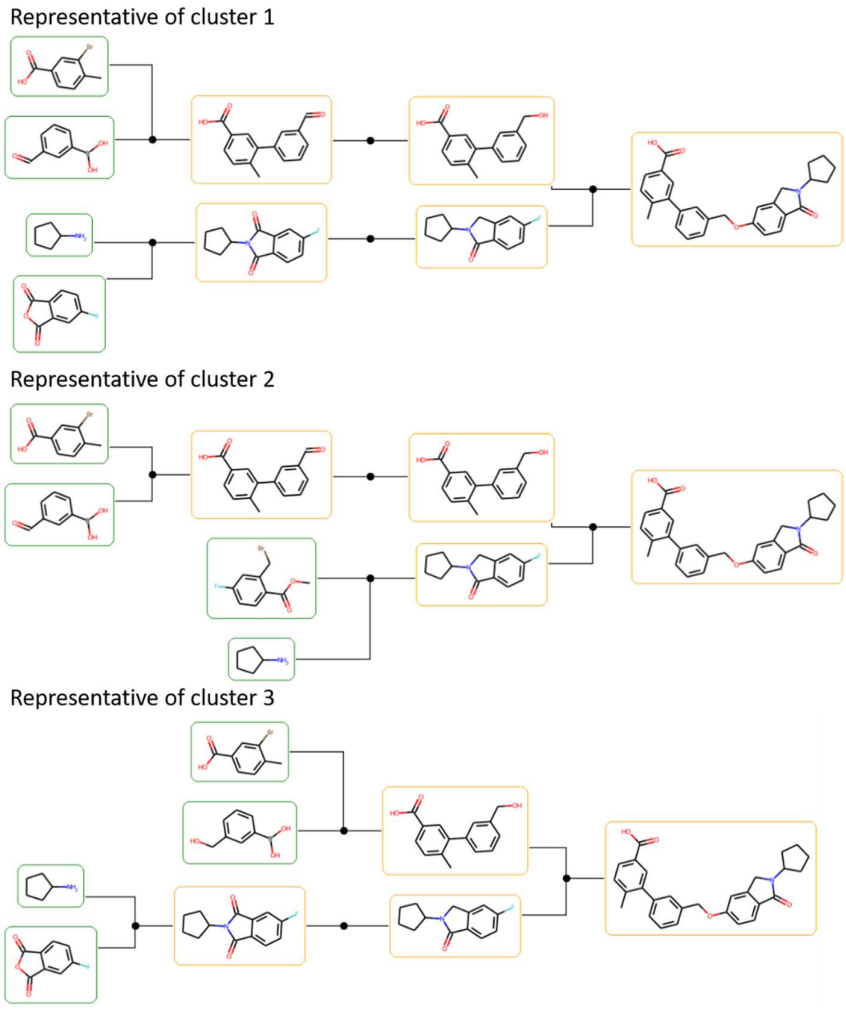


Figure 8 – Cluster representatives for the third example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

TED-based and LSTM-based clustering does not always produce the same labels. In Figure 9 we show the correlation between the TED and the Euclidean distance of the LSTM network-based latent space encoding for a sample of routes. There is a clear but weak correlation, and the correlation coefficient, *r*, is only 0.50. This naturally affects the hierarchical clustering. For each compound, we computed the cluster similarity as the average agreement of cluster labels over all pairs of routes. The distribution of the cluster similarity is shown in Figure 9 as well, and the average over all compounds is 0.70. This implies that on average only about 2/3 of the routes are in the same cluster when comparing TED-based and LSTM-based clustering. For only 23% of the compounds, the cluster similarity is more than 0.9 and the lowest cluster similarity is 0.21. Overall, the LSTM-based clustering leads to fewer clusters with more routes as seen in Figure S1. For the example compound 2 and 3, discussed above, the LSTM-based clustering produced four clusters for both compounds instead of three.

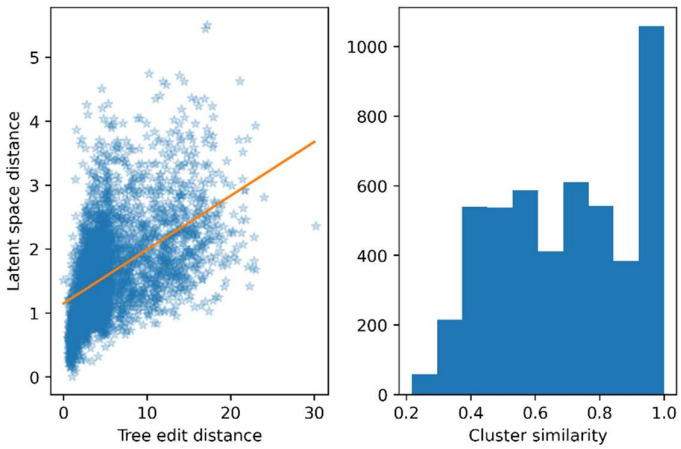


Figure 9 – Similarity of TED-based and LSTM-based clustering. Left) Correlation between the underlying distances. A sub-sample of all the routes is shown as scatter together with a linear regression line based on all routes. Right) the distribution of the cluster similarity for all compounds

Cluster assignments can be rationalized but optimal number of clusters can be subjective. We analyzed the difference between TED-based and LSTM-based clustering closer by selecting an example compound where the cluster similarity was low (0.23) and another one where the cluster similarity was close to the average (0.70). For the compound where the cluster similarity was low, we show the dendrograms formed from the two distance matrices in Figure 8a and 8b, and in Table S4, we show all of the routes as clustered by the TED matrix. TED-based clustering leads to five clusters, whereas LSTM-based clustering only leads to two clusters. The first TED-based cluster is formed from routes 0, 2, 3, 11, and 13 and this cluster is characterized by a first step adding a sulfonyl group to an aromatic ring followed by a sulfonylation step forming a N-S bond. The second TED-based cluster is formed from routes 1, 10 and 12 and this cluster differs from cluster 1 by the first step which is an acylation forming a N-C bond. The second step is the same as in cluster 1. The remaining three clusters contain two routes each. Cluster 3 formed of routes 4 and 6 is characterized by an initial deprotection step followed by the acylation step forming the N-C bond. Cluster 4 formed from routes 8 and 9 shares the first step with cluster 2 but the second step is an addition of an ethyl acetate group. Cluster 5 formed from routes 5 and 7 is characterized by the reverse order of the steps in cluster 2. All of these clusters can be rationalized quite easily, although it could be argued that some of the clusters (such as 2 and 4) could be merged. However, for the LSTM-based clustering, it is hard to rationalize that the first cluster consists of only route 0 and the second cluster is formed from all the other routes. If we should form only two clusters from the TED matrix, they would be formed from routes 4 to 7 and the rest of the routes, respectively, as is clear from the dendrogram in Figure 8. It should be pointed out the LSTM network was trained on discriminating between human-made and predicted routes, and not to produce intuitive clusters.[8] Therefore, it is likely that we can find examples where the LSTM-based clustering gives sub-optimal solutions.

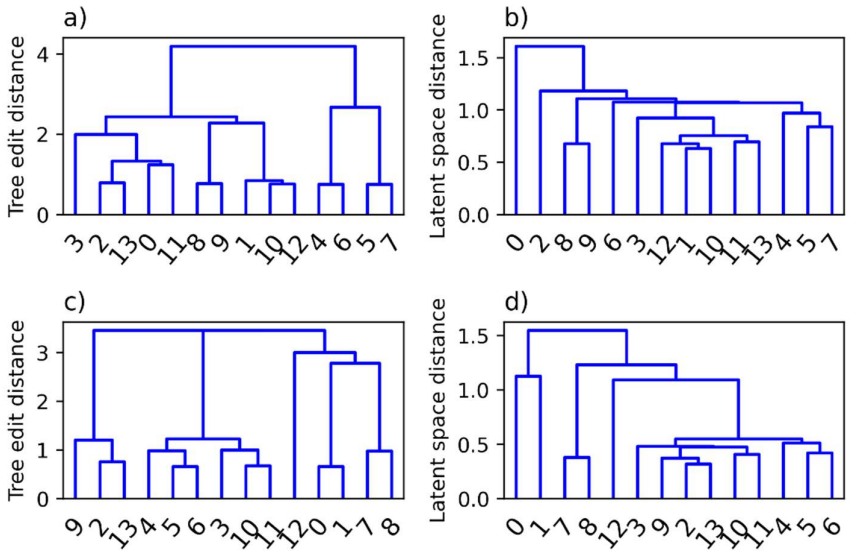


Figure 8 – Dendrogram from distance matrices giving very different clusters. a) and b) Dendrograms for an example compound for which cluster similarity is 0.23, c) and d) Dendrograms for an example compound for which cluster similarity is 0.70

For the example compound where the cluster similarity is average, we show the dendrograms in Figure 8c and 8d and all the routes in Table S5. For this compound TED-based clustering gives five clusters, whereas LSTM-based clustering gives three clusters. Cluster 1 is formed from two one-step routes (routes 0 and 1) and this cluster is given by both TED-based and LSTM-based clustering. Cluster 2 from TED-based clustering is formed from routes 3 to 6 and 10 to 11, and is characterized by a first step similar to the single step in cluster 1 followed by the addition of an amine group. The routes in this cluster also form a cluster based on the latent space distances, but is joined with clusters 3 and 4 from the TED-based clustering. Cluster 3 formed from routes 2, 9 and 13 and has a similar first step to the routes in cluster 2 but requires the addition of a bigger substituent to the non-aromatic ring in step 2. Cluster 4 in the TED-based clustering consists of only route 12 that starts with a sulfonylation step. In the dendrogram of the TED matrix, route 12 is closer to routes 0, 1, 7 and 8 (see Figure 8c), whereas in the dendrogram of the latent space distance matrix it is closer to for instance route 3 (see Figure 8d), explaining why they are clustered together. The final cluster, cluster 5 in the TED-based clustering is identical to the third LSTM-based cluster and is formed from routes 7 and 8. For this example compound, it is clear that we can rationalize both the TED-based and LSTM-based clustering. The perceived optimal cluster sizes would likely depend on the judgment of an expert[26,27] and because we are only discussing a few examples, we cannot conclude that one method is superior to the other. We simply conclude that the two clustering approaches are different and that it seems that TED-based clustering generally is more discriminative than LSTM-based clustering and leads to more clusters.

Conclusions

We have presented a novel algorithm to compute distances between synthesis routes and used it to cluster predictions from retrosynthesis analysis. The clustering algorithm is on average fast, and we have been able to show that it preserves the distribution of the routes. The clusters also appear to be intuitive as they can be easily rationalized. This implies that the clustering algorithm can reduce the number of predicted routes and thereby aid in the selection of routes for wet-lab experimentation. We have included the clustering algorithm in the latest release of the AiZynthFinder software and envisage it will be useful in future synthesis prediction tasks.

References

[1] Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

[2] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

[3] 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.* **2017**, *3* (5), 434–443. https://doi.org/10.1021/acscentsci.7b00064.

[4] Johansson, S.; Thakkar, A.; Kogej, T.; Bjerrum, E.; Genheden, S.; Bastys, T.; Kannas, C.; Schliep, A.; Chen, H.; Engkvist, O. AI-Assisted Synthesis Prediction. *Drug Discovery Today: Technologies*. Elsevier Ltd July 11, **2020**. https://doi.org/10.1016/j.ddtec.2020.06.002.

[5] Badowski, T.; Molga, K.; Grzybowski, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10* (17), 4640–4651. https://doi.org/10.1039/c8sc05611k.

[6] Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using a Combined Linguistic Model and Hyper-Graph Exploration Strategy. arXiv:1910.08036

[7] Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic Retrosynthetic Route Planning Using Template-Free Models. *Chem. Sci.* **2020**, *11* (12), 3355–3364. https://doi.org/10.1039/c9sc03666k.

[8] Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. Evaluating and Clustering Retrosynthesis Pathways with Learned Strategy. *Chem. Sci.* **2021**. https://doi.org/10.1039/D0SC05078D.

[9] Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; John Hart, A.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science (80-.).* **2019**, *365* (6453). https://doi.org/10.1126/science.aax1566.

[10] Bille, P. A Survey on Tree Edit Distance and Related Problems. *Theor. Comput. Sci.* **2005**, *337* (1–3), 217–239. https://doi.org/10.1016/j.tcs.2004.12.030.

[11] 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. *Chem. Sci.* **2020**, *11* (1), 154–168. https://doi.org/10.1039/c9sc04944d.

[12] Genheden, S.; Thakkar, A.; Chadimova, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. *J. Cheminf.* **2020**, *12* https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00472-1

[13] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G., Smit, I.; & Leach, A. R. The ChEMBL database in 2017. *Nucl. acids res.*, **2017**, *45(D1)*, D945–D954. https://doi.org/10.1093/nar/gkw1074

[14] RDKit: Open-source cheminformatics, http://www.rdkit.org.

[15] Genheden, S.; Engkvist, O.; Bjerrum, E. J. A Quick Policy to Filter Reactions Based on Feasibility in AI-Guided Retrosynthetic Planning. *ChemRxiv* Preprint. **2020**. https://doi.org/10.26434/CHEMRXIV.13280495.V1.

[16] Pawlik, M.; Augsten, N. Tree Edit Distance: Robust and Memory-Efficient. *Inf. Syst.* **2016**, *56*, 157–173. https://doi.org/10.1016/j.is.2015.08.004.

[17] Pawlik, M.; Augsten, N. Efficient Computation of the Tree Edit Distance. *ACM Trans. Database Syst.* **2015**, *40* (1), 1–40. https://doi.org/10.1145/2699485.

[18] https://github.com/JoaoFelipe/apted

[19] McVicar, M.; Sach, B.; Mesnage, C.; Lijffijt, J.; Spyropoulou, E.; De Bie, T. SuMoTED: An Intuitive Edit Distance between Rooted Unordered Uniquely-Labelled Trees. *Pattern Recognit. Lett.* **2016**. https://doi.org/10.1016/j.patrec.2016.04.012.

[20] Yoshino, T.; Higuchi, S.; Hirata, K. A Dynamic Programming A* Algorithm for Computing Unordered Tree Edit Distance. In Proceedings - 2nd IIAI International Conference on Advanced Applied Informatics, *IIAI-AAI* **2013**; 2013; pp 135–140. https://doi.org/10.1109/IIAI-AAI.2013.71.

[21] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983-996. https://doi.org/10.1021/ci9800211.

[22] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754

[23] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[24] Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20* (C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

[25] https://github.com/moyiming1/Retrosynthesis-pathway-ranking (accessed 2020-11-24)

[26] Estivill-Castro, V. Why so Many Clustering Algorithms. *ACM SIGKDD Explor. Newsl.* **2002**, *4* (1), 65–75. https://doi.org/10.1145/568574.568575.

[27] Budka, M. Clustering as an Example of Optimizing Arbitrarily Chosen Objective Functions. *Stud. Comput. Intell.* **2013**, *457*, 177–186. https://doi.org/10.1007/978-3-642-34300-1_17.

Supporting information

Clustering of synthetic route predictions using tree edit distance

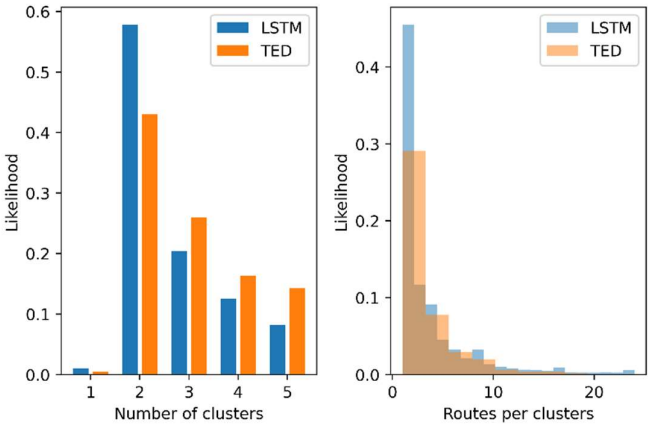


Figure S1 – Statistics of the cluster optimization: Left) the likelihood of forming a particular number of clusters, Right) The distribution of the number of routes per cluster. The data labeled with TED is the same data as in Figure 4, and is included here for easier comparison.

Table S1 – the routes for the first example compound

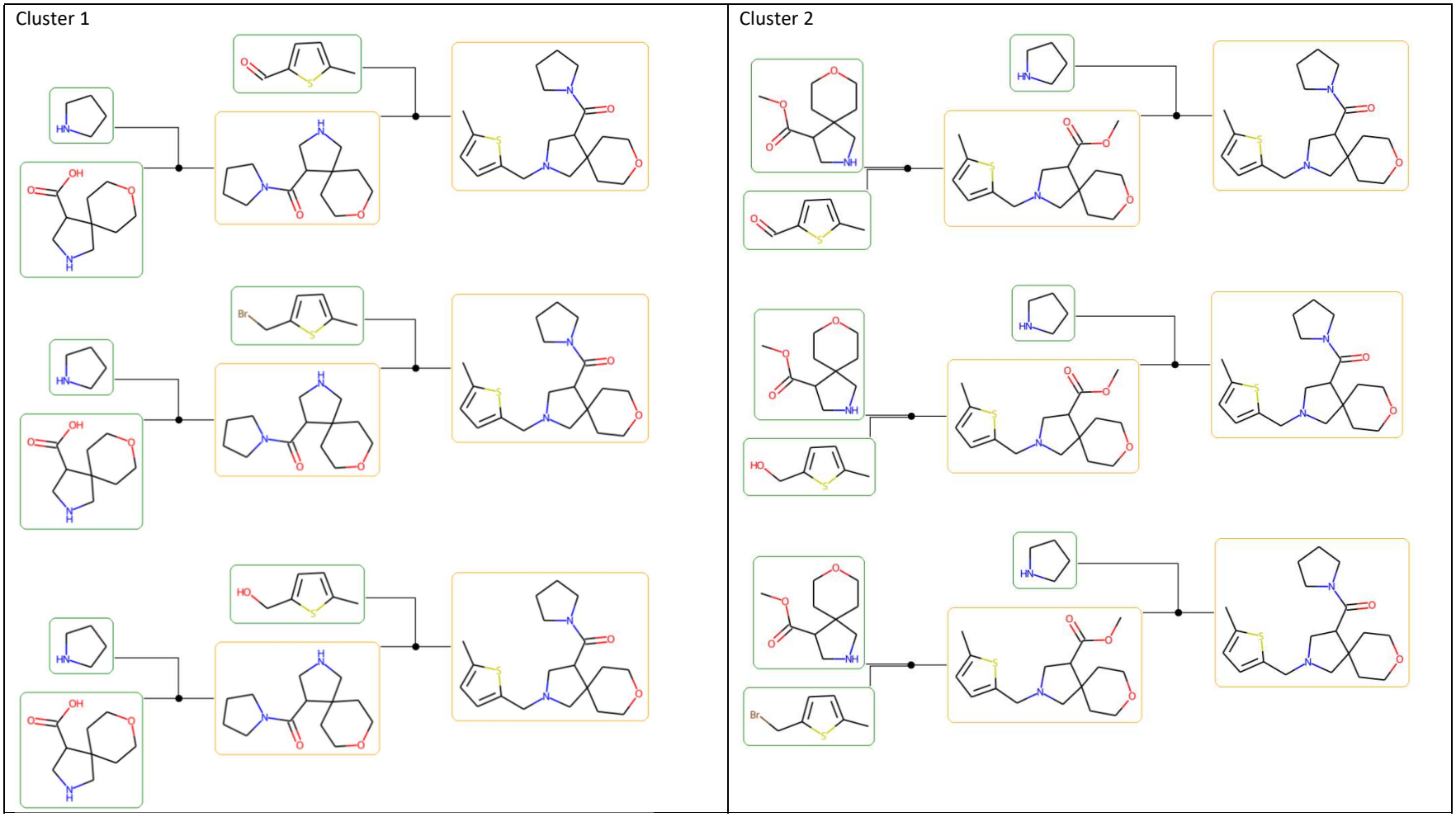
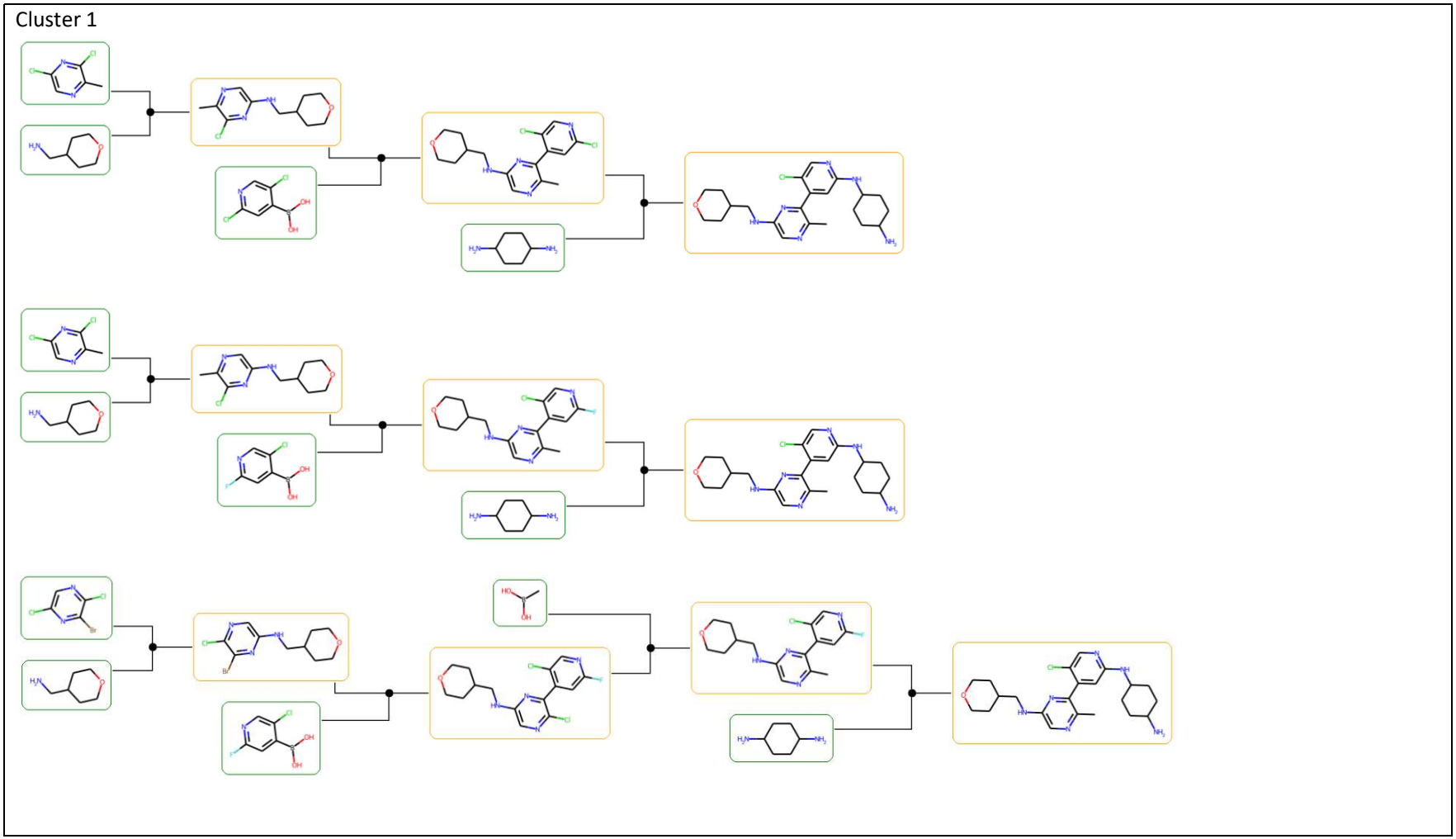
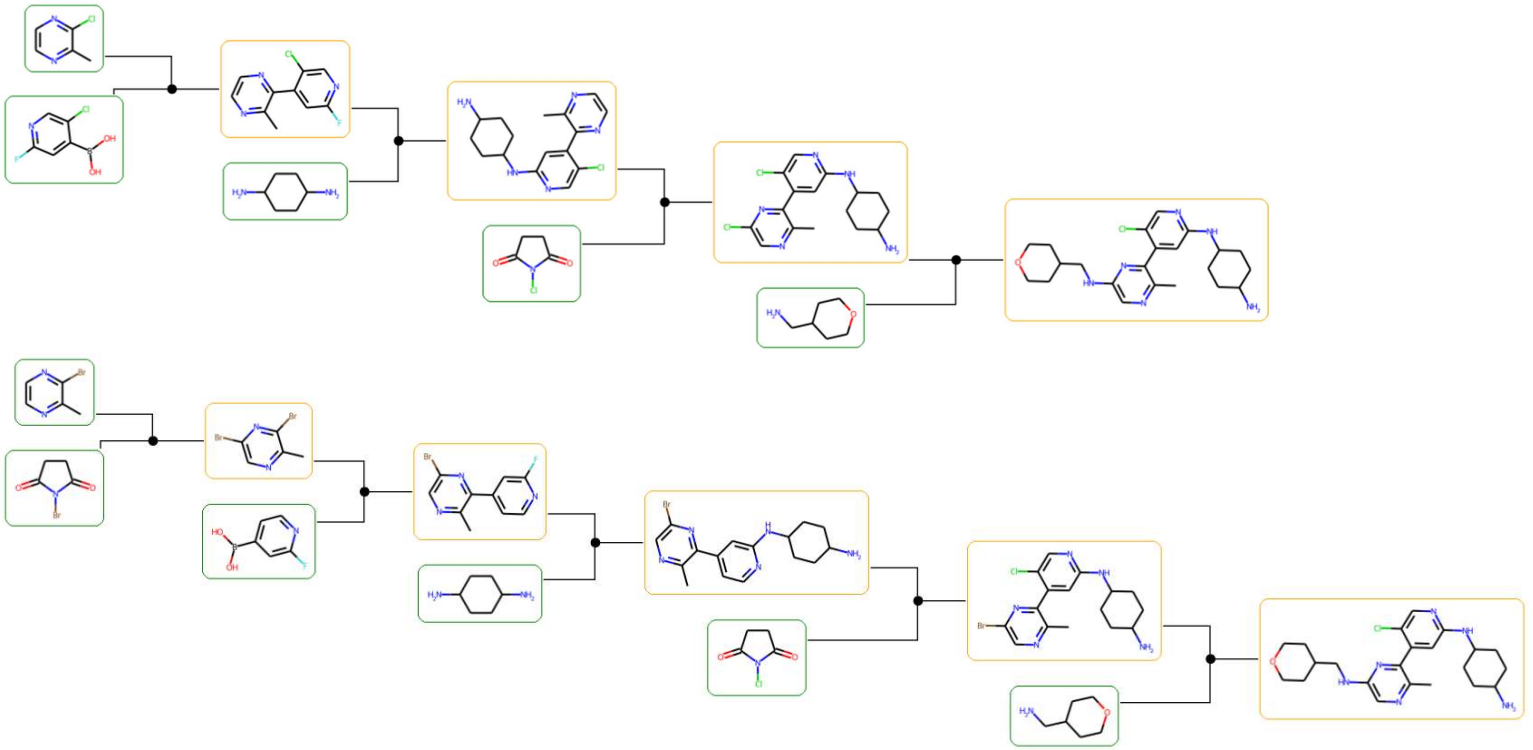


Table S2 – the routes for the second example compound



Cluster 2



Cluster 3

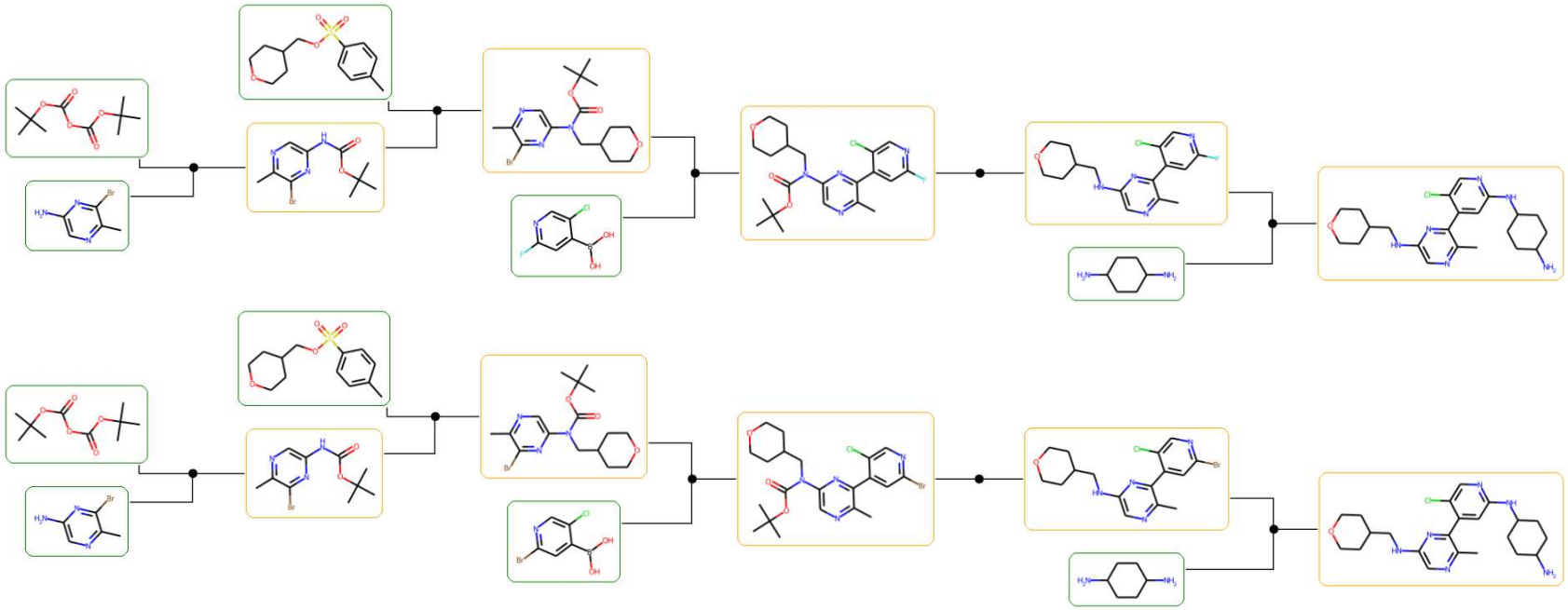
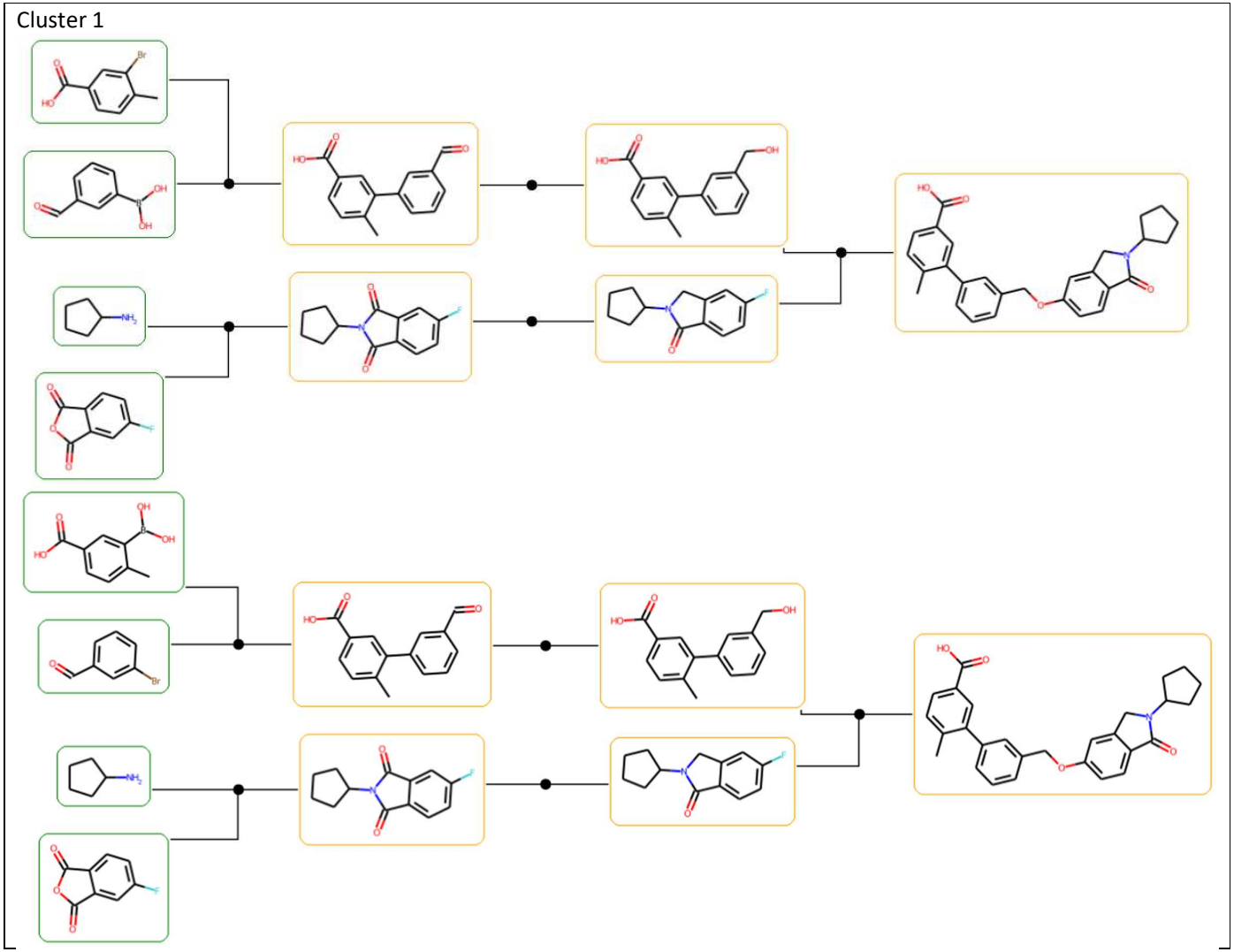
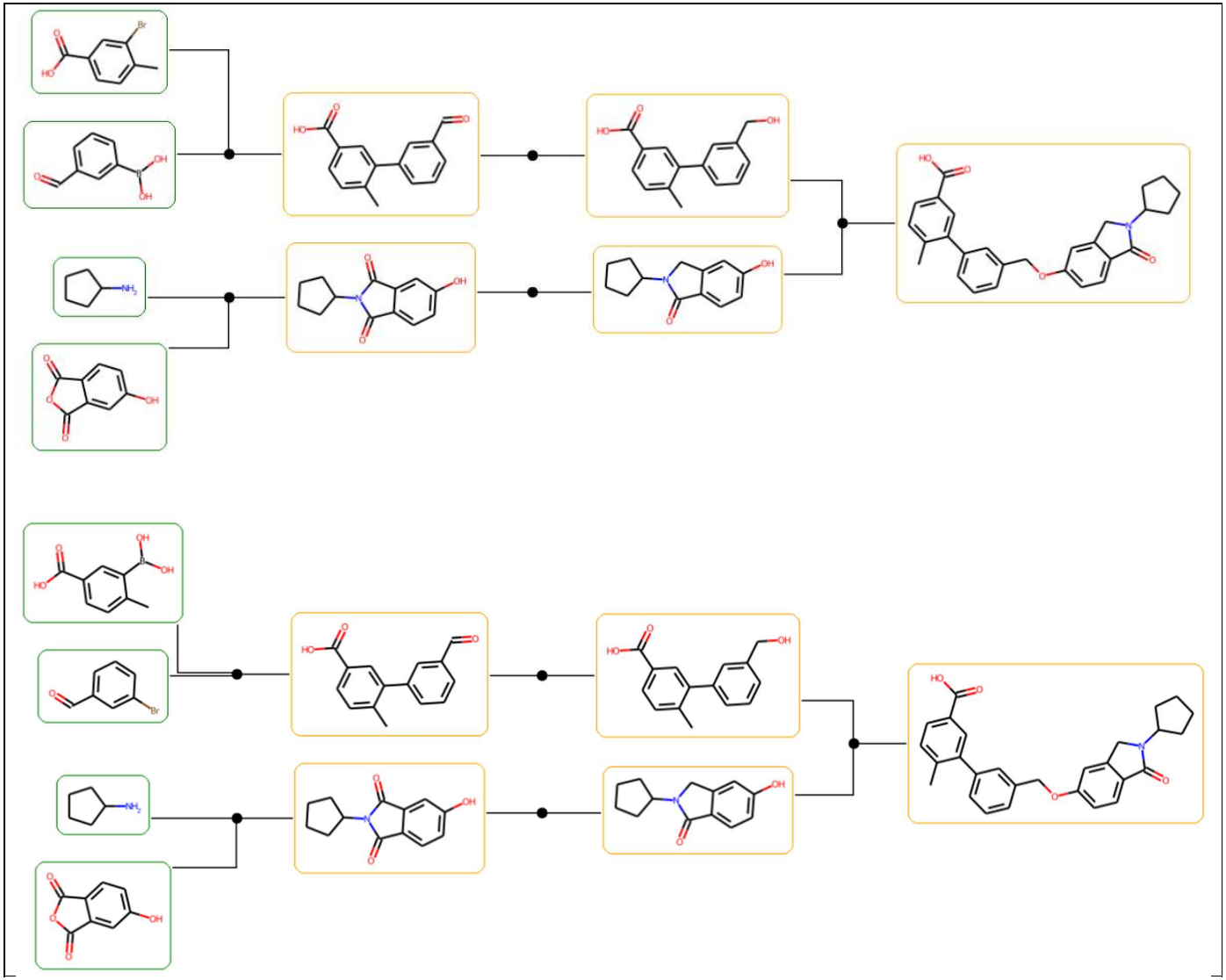
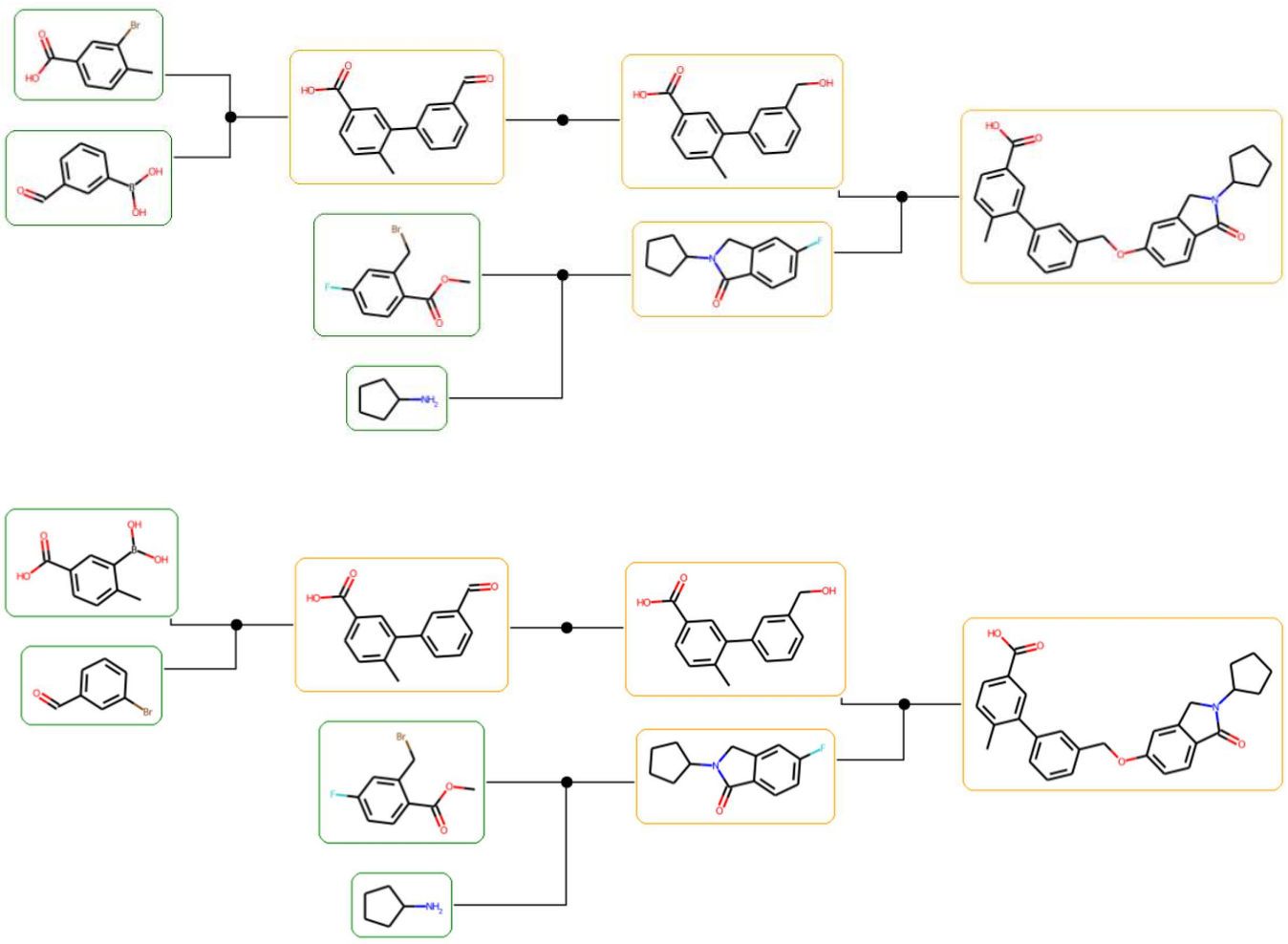


Table S3 – the routes for the third example compound

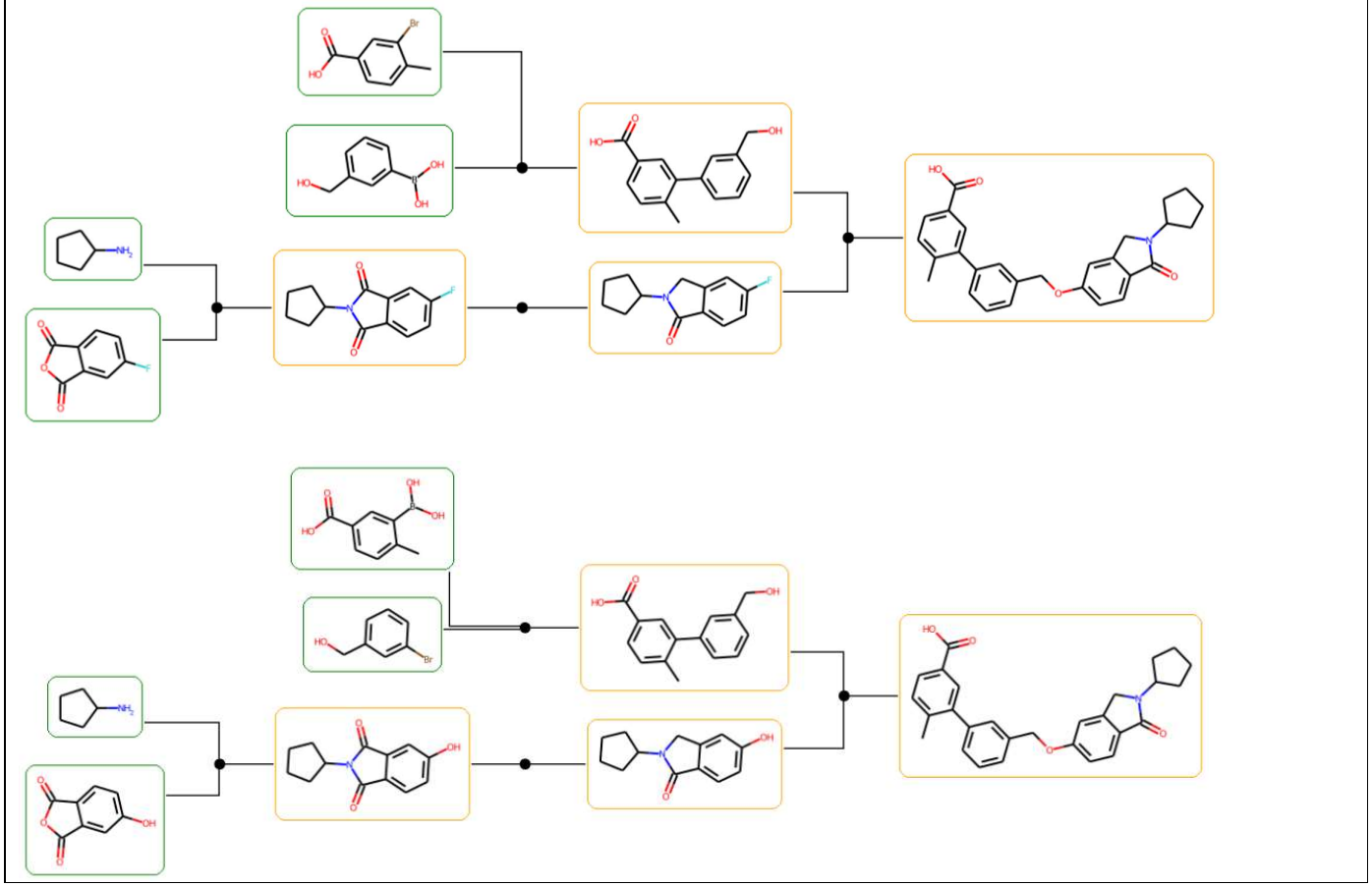




Cluster 2



Cluster 3



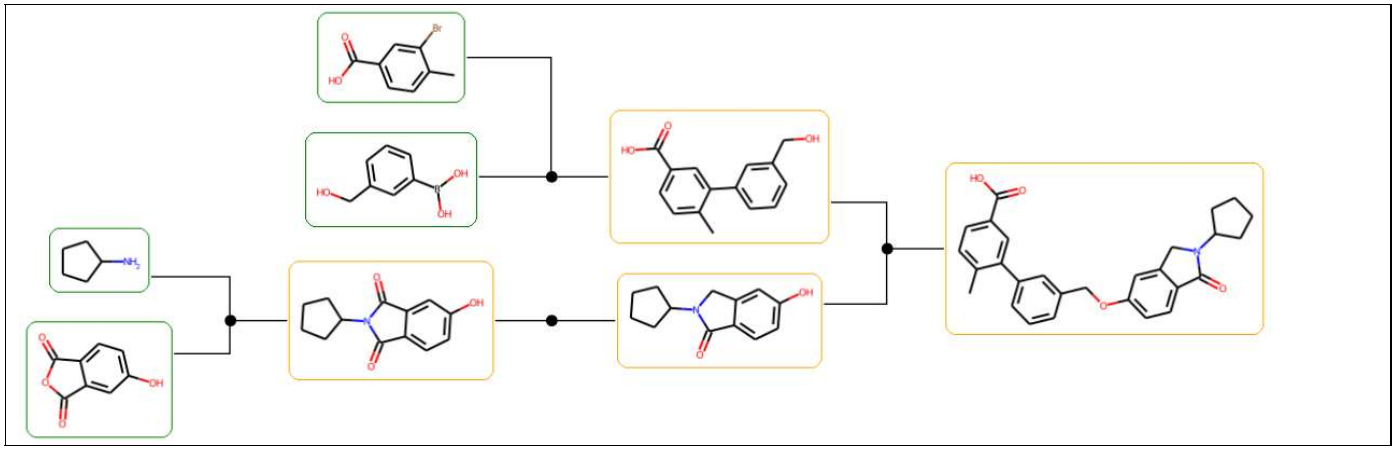
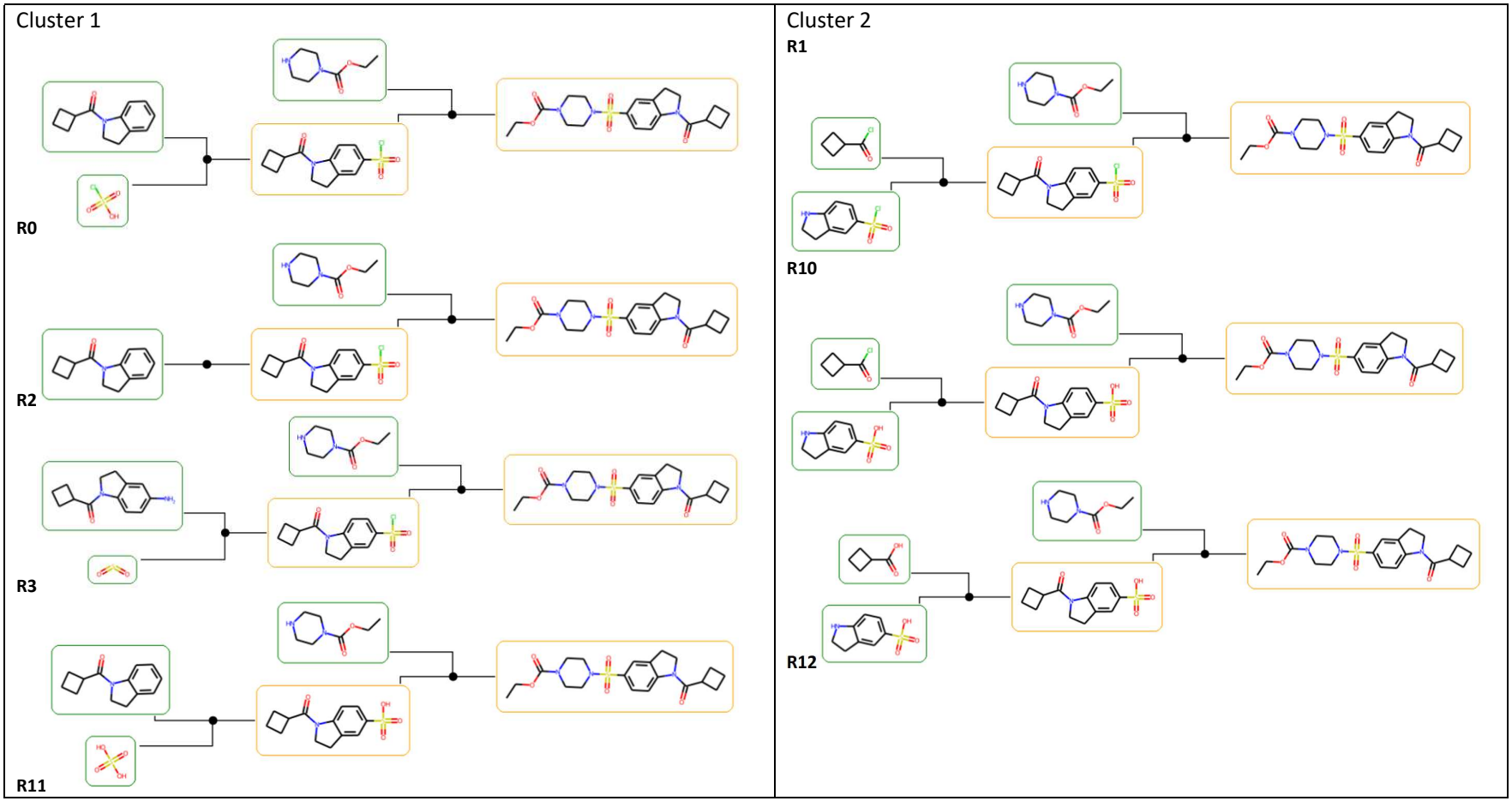
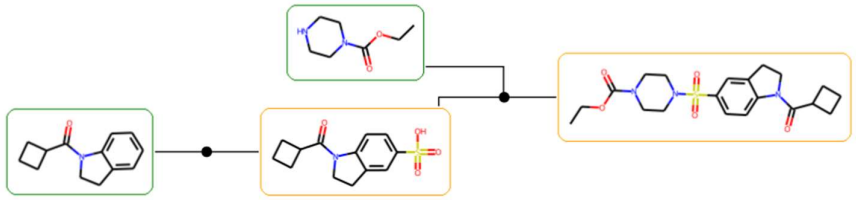


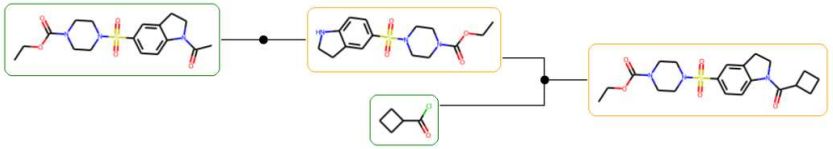
Table S4 – the routes for the example compound for which the cluster similarity was 0.23. Clusters labels are given by TED-based clustering.



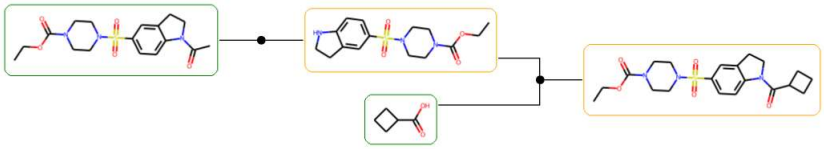


R13

Cluster 3



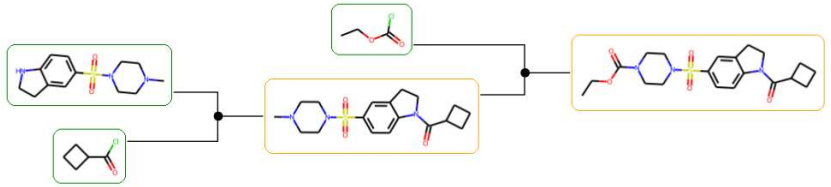
R4



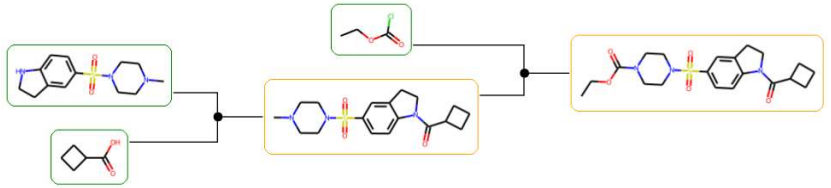
R6

Cluster 4

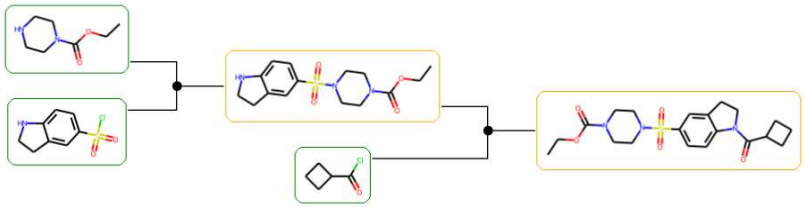
R8



R9



Cluster 5



R5

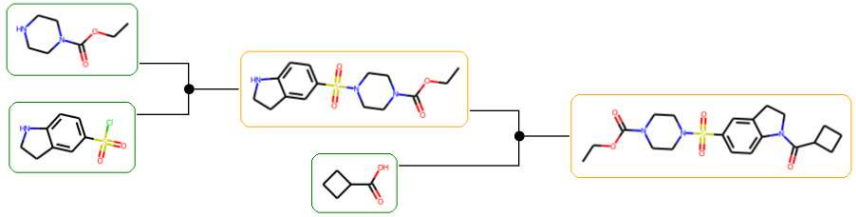
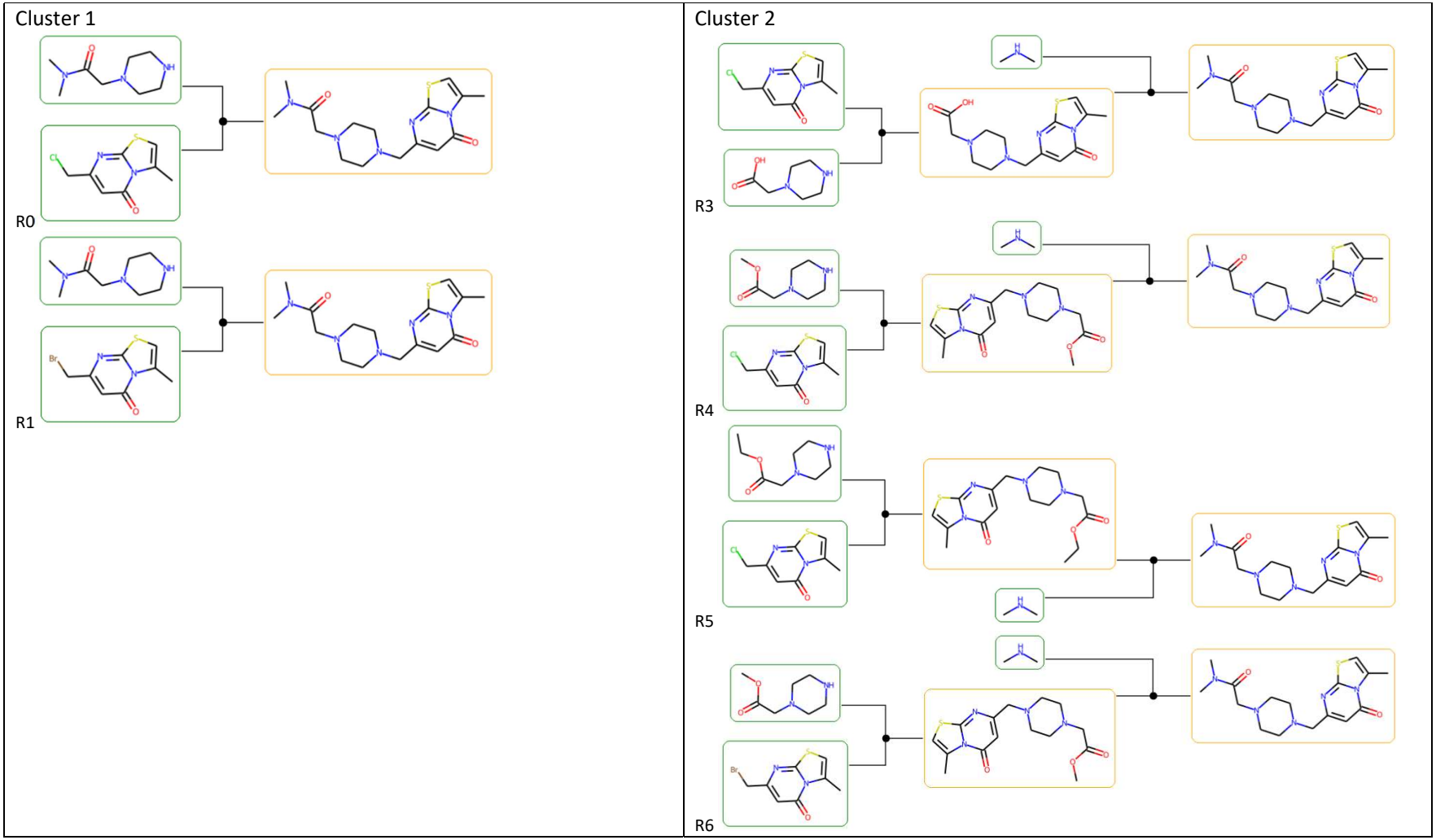
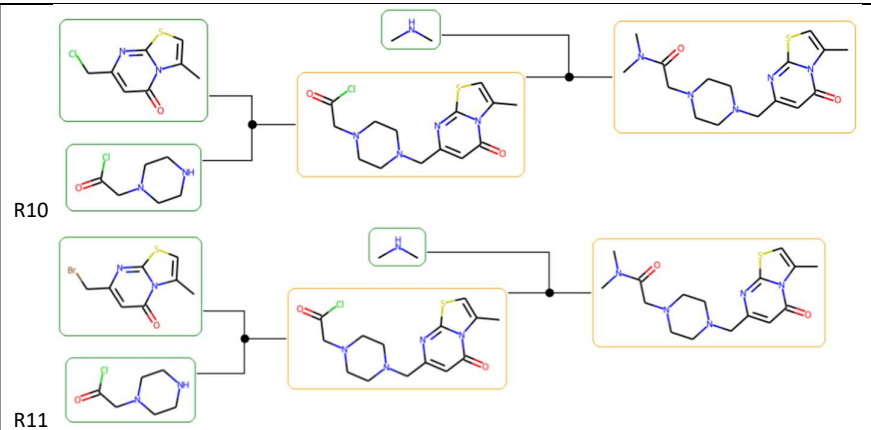
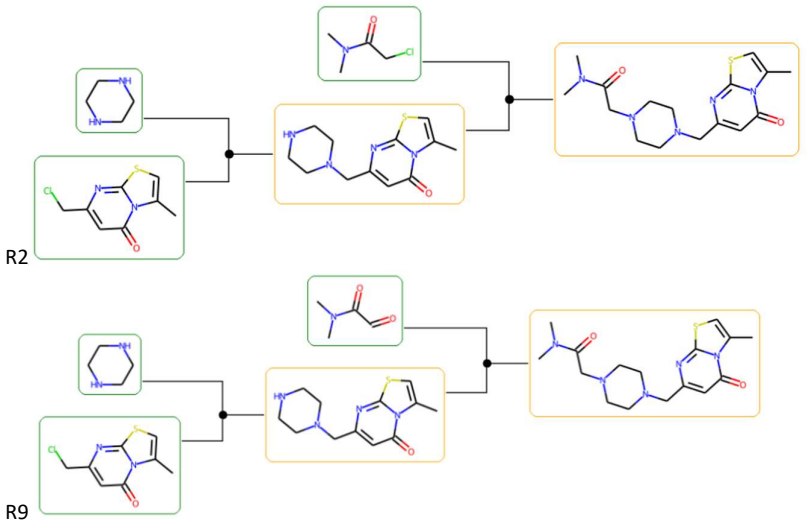
 R7	

Table S5 – the routes for the example compound for which the cluster similarity was 0.70. Clusters labels are given by TED-based clustering.

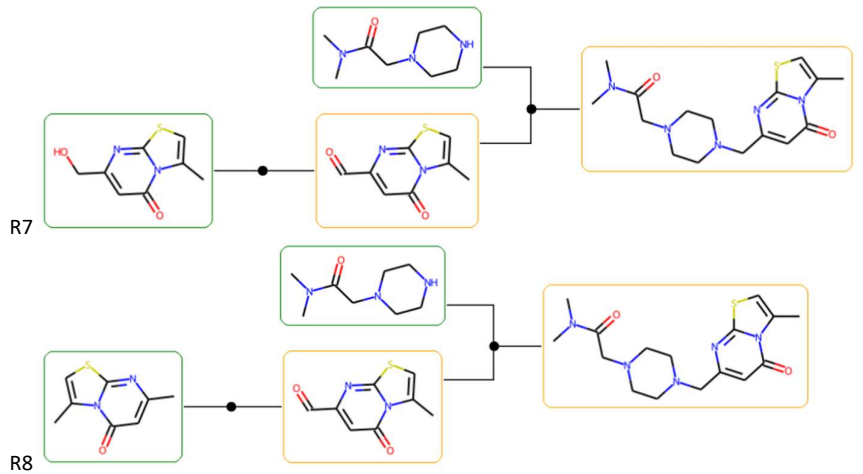


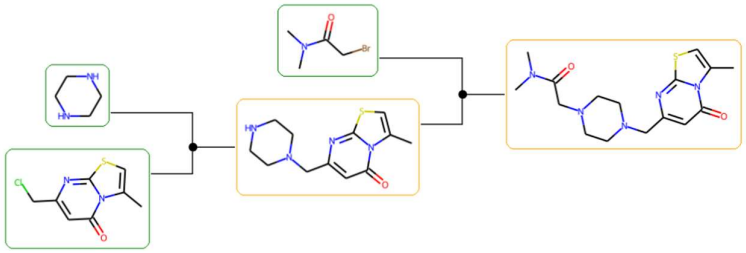
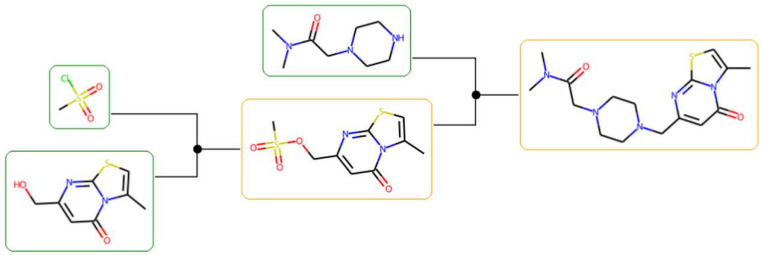


Cluster 3



Cluster 4



 R13	
Cluster 5  R12	