JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

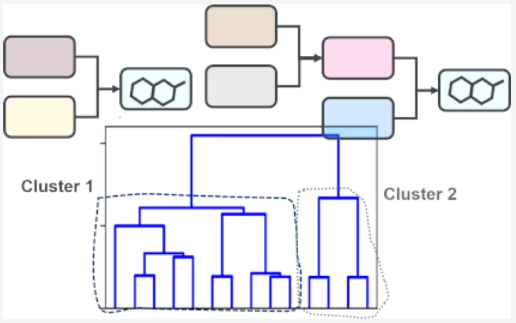
Clustering of Synthetic Routes Using Tree Edit Distance

Samuel Genheden,* Ola Engkvist, and Esben Bjerrum

Cite This: J. Chem. Inf. Model. 2021, 61, 3899−3907 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present a novel algorithm to compute the distance between synthetic routes based on tree edit distances. Such distances can be used to cluster synthesis routes generated using a retrosynthesis prediction tool. We show that the clustering of selected routes from a retrosynthesis analysis is performed in less than 10 s 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 and that the routes in a cluster tend to use similar chemistry. The algorithm is included in the latest version of open-source AiZynthFinder software (https://github.com/MolecularAI/aizynthfinder) and as a separate package (https://github.com/MolecularAI/route-distances).



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) and how to optimize the reaction conditions and evaluate the feasibility of a reaction. The rise of deep learning methods and data-driven approaches in the last decade has led to an increased interest in the field,[1−4] although expert or hybrid systems have been around for long and have recently also shown impressive results.[5,6] The synthesis of a compound can be described as a route or reaction tree, showing what reactions need to be carried out in order to synthesize the final compound. The precursors of the target compound are typically smaller molecules (building blocks) which are either readily available in storage or can be synthesized from molecules in stock.

The exploration of the enormous synthesis space with a synthesis prediction tool typically produces more than one route such that selecting what routes are present to the chemists and what routes to proceed with to synthesis is an important task. The routes present to the chemists should be as diverse as possible; it may not be so interesting to present routes that differ, for instance, only in the kind of halogen substituent on one of the precursors or two pairs of precursors producing the same two products. A well-designed scoring function could aid in 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[7] or the aggregation of single-step likelihood functions.[8,9] However, it is unclear if such scoring functions are sufficient to produce a diverse set of routes. In Chematica software, the similarity problem is overcome by a path retrieving algorithm that iteratively selects routes.[10] At each iteration, reactions present in the already retrieved routes are penalized, which increases the diversity of the retrieved routes. The routes are scored based on a recursive price estimator.[7,10] A similar strategy is employed in the CompRet algorithm to enumerate a reaction network.[11] An alternative is clustering the predictions, which could help in deciding which route to proceed with by reducing the set of predicted routes and give a better overview of the suggestions. Recently, Mo et al. suggested a long short-term memory (LSTM)-based neural network architecture to encode routes[12] and used it to distinguish between predicted routes from ASKCOS[13] and human-designed routes. Relevant to our aim with route clustering, they also suggested using the latent space encoding of the routes as the basis for distance calculations and clustering.

Here, we will take another approach and introduce a novel clustering algorithm that is based on a tree edit distance (TED) calculation.[14] The algorithm was implemented in the AiZynthFinder retrosynthesis tool,[15,16] and here, we show that it produces intuitive clusters within a reasonable timeframe.

METHODS

Route Predictions. We selected 5000 random compounds from ChEMBL.[17] The tautomeric states were calculated with

Received: February 28, 2021
Published: August 3, 2021

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

the RDKit.[18] The simplified molecular-input line-entry system strings of the compounds were used as input to AiZynthFinder software to predict synthetic routes. The expansion and filter policies used in the search were derived from the USPTO data set, as discussed previously.[15,19] In-house 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[15] of the routes.

Distance Calculations. A synthetic route or a reaction tree is a bipartite tree consisting of molecules and reaction nodes, with the target molecule as the root. The distance between two trees (T_1 and T_2) can be computed using a tree edit algorithm.[14] 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(see below). The TED is then defined as a minimum-cost sequence of such operations that transform T_1 into T_2. To compute TED, we use the APTED (all path TED) algorithm,[20,21] which is available as a Python package.[22] APTED only guarantees an optimal solution for an ordered tree, that is, a tree where the children of a node have an inherent order. A reaction tree is however an unordered tree, and finding a solution to such a tree is NP-complete. Specialized algorithms to compute TED for unordered trees have been suggested,[14,23,24] but we found none, which were appropriate for our task and had a reference implementation. 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. N_T is calculated from $\Pi_i N_i^C!$, where N_i^C is the number of children of node i. If the product of $N_T(T_1)$ and $N_T(T_2)$, that is, 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. 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 the original 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 random enumerations were created by randomly ordering the children nodes for any node in the tree. A full enumeration of all trees was carried out by first precomputing the possible permutations of all children nodes and then looping over all possible products of children node permutations. For each such product, a new tree was constructed from the original tree by reordering the children nodes.

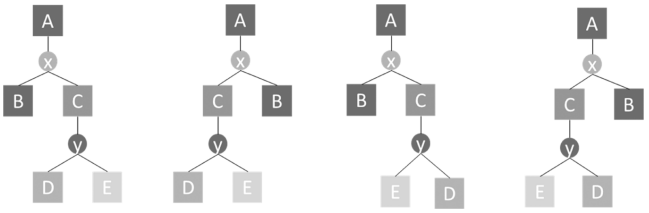


Figure 1. Four possible enumerations of the same reaction tree. Molecule nodes are A, B, C, D, and E, and reaction nodes are x and y.

The insertion and deletion cost is set to unity, and the substitution cost is set to the Jaccard distance[25] between the fingerprints of the nodes. The fingerprint of a molecule node was set as 2048-bit fingerprint (ECFP4, computed using the Morgan algorithm in the RDKit[18,26]). The fingerprint of a reaction node was then set as the difference between the fingerprints of the reactants and products.

Clustering. Clustering is based on the TED matrix. Because TED is not a distance in cartesian space, we used hierarchical clustering with single linkage as implemented in scikit-learn.[27] The optimal cluster size was determined by the silhouette method.[28] Because the number of analyzed routes ($\leq$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.[15]

LSTM-Based Clustering. We downloaded the LSTM-based neural network model of Mo et al. from GitHub.[12,29] 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 5000 compounds selected from ChEMBL. We will then show some illustrative examples of the synthetic route clustering for a few selected compounds. Furthermore, we will compare the TED-based clustering approach to the deep learning method of Mo et al.[12] and investigate the approach for patented routes.

Cost Functions in the TED Calculations Are Chemically Motivated. TED is a graph theoretical metric and as such does not inherently take chemical knowledge into consideration. However, the cost function used in the calculations was chosen to make chemical sense. The deletion and insertion costs were set to unity as is common in the literature.[21] This is the cost of substituting two nodes with maximal dissimilarity, and it would be hard to understand why a deletion or insertion should cost more or less than this and what that cost would be. The substitution cost was set to the Jaccard distance, equivalent to one of the most common similarity metrics in chemistry,[25] the Tanimoto coefficient. Although, other metrics have been used in chemistry,[25] we were satisfied with the performance when using the Jaccard distance (see below), and we did not pursue any other option. Because we do not have a ground truth to compare to, an evaluation of different substitution costs can at most tell us that they are different.

The substitution cost could potentially be augmented with additional chemical knowledge. One avenue would be to assign a substitution cost of zero to chemically equivalent (de)-protection reactions. For a chemist, it would not matter if, for example, different protection reagents were used as long as

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

they give the same product. To accomplish this, the reaction database on which we based the retrosynthesis predictions would have to be augmented with additional annotations. Although subjective to a certain degree, chemists tend to regard certain reactions as key steps, as an example, a reaction where the central core of the scaffold is created. If these key steps can be robustly and objectively identified, they can be weighted higher in the scoring, leading to possible improvements in the perceived outcome of the clustering.

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 5000 compounds in this study, the average number of reactions and molecules is 3.8 and 7.5, respectively. As shown in Figure 2, the number of possible trees (N_T) is less than 100 for a majority of these routes. At the cut-off we used for the distance calculation, 20, we captured 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 a semi-exhaustive strategy was used. We also tried to use a limit of 40 instead of 20, but the correlation coefficient, r, between the distances for all routes was 0.99, and the mean absolute difference was 0.1 units when comparing the two cut-offs. Furthermore, the clustering similarity was 0.99, indicating that the small differences in distance did not affect the clustering. As an alternative to these heuristic strategies, we also explored an option to form an ordered tree by sorting the molecule nodes on their InChI keys. 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.

Clustering Algorithm is Sufficiently Fast for Small Route Collections but Does Not Scale Well. 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, so compared to this, the average clustering time is fast. On average, 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); although the median time is 2 s, the worst time is 214 s. The clustering time is naturally correlated with the number of routes, but the correlation is not clear as shown in Figure 3 because the shape of the route is also an important factor. There are compounds for which 25 routes were clustered in less than 1 s, yet there are also compounds for which 25 routes were clustered in 3 min. However, for 95% of the compounds, the clustering was carried out in less than 30 s, which is acceptable considering the average route prediction time.

To investigate the scaling of the method, we also calculated the distances of the top 100 routes for each compound. The average clustering time for these routes was 377 s, which is an average increase in time of 513 s with an average increase in the number of routes of 12. These timings render the approaches prohibitively large to be used in a high-throughput setting. Therefore, the TED method should be used predominantly to cluster routes selected by other means (in our case, the Monte Carlo tree search score). The average number of completely solved routes is only 37 for these compounds, and, as such, there is only a weak argument for clustering hundreds or thousands of compounds. It would be sufficient to extract a few highly scored routes and then cluster them.

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 few routes. 85% of the clusters contain at most 5 routes, and 96% of the clusters contain at most 10 routes.

Clustering Preserves the Distribution of Route Shapes. 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

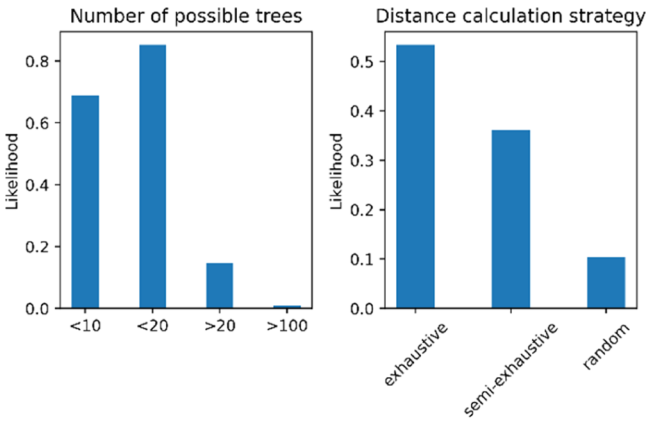


Figure 2. Statistics from the distance calculations. (Left) The bar chart showing parts of the distribution of the number of possible trees (N_T). (Right) The likelihood of using a particular strategy when computing the distance between two routes.

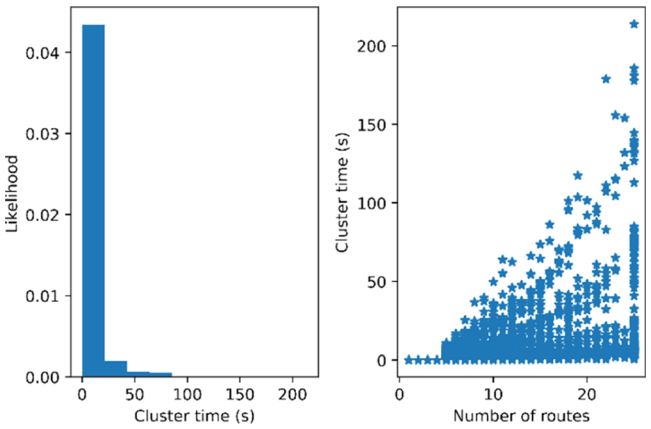


Figure 3. Timings of the clustering. (Left) The distribution of the clustering time over all 5000 ChEMBL compounds; (right) the relationship between number of analyzed routes and the cluster time.

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

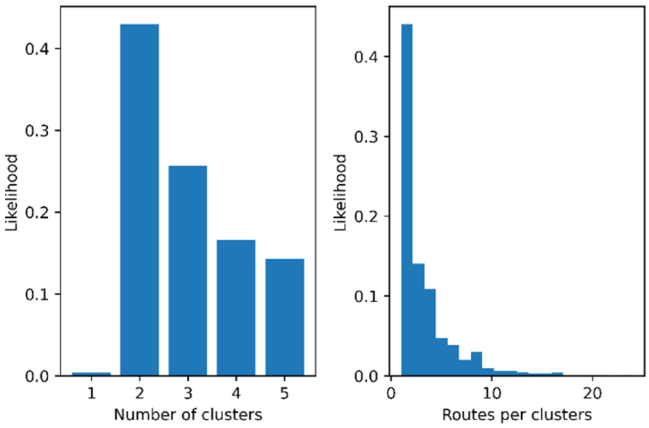


Figure 4. Statistics of the cluster optimization. (Left) The likelihood of forming a particular number of clusters; (right) the distribution of the number of routes per cluster.

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 route representatives is very similar. It seems that clustering leads to a selection of slightly shorter routes, both in terms of the 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 either all routes or only representative routes. 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 of molecules. However, the overall shape of the distributions is preserved using the clustering algorithm. Thus, we can conclude that representative routes from the clustering can be used to reduce the set of predicted routes.

TED-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 show the cluster representatives for the compounds (Figures 6−8), and in the Supporting Information, we show all of the routes (Tables S1−S3).

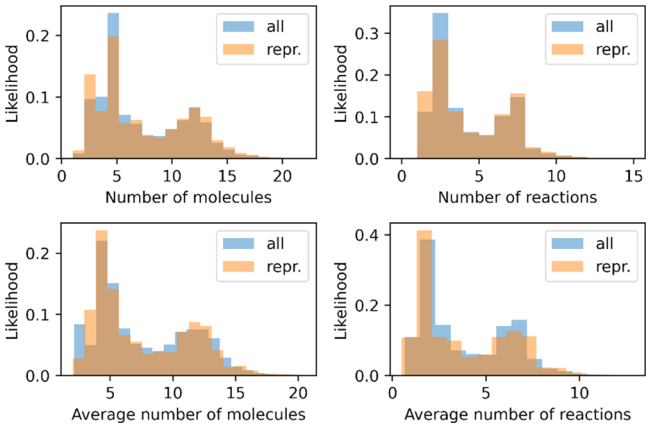


Figure 5. Distribution of all routes or representative routes (repr.). (Top) The number of molecules and reactions in the routes. (Bottom) The average number of molecules and reactions for a compound.

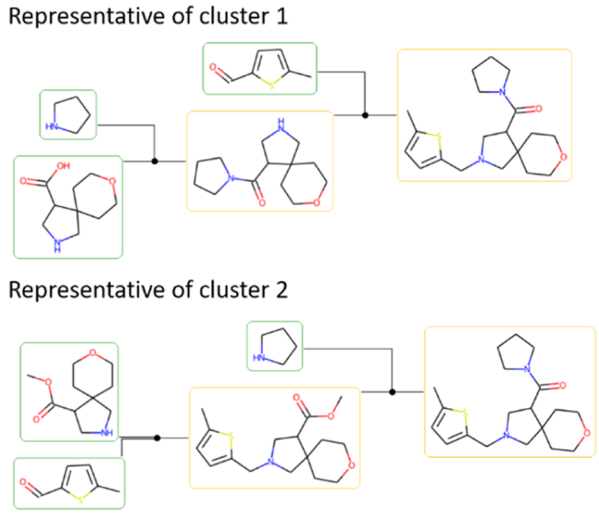


Figure 6. Cluster representatives for the first example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

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, that is, whether pyrrolidine is attached first or second (see Table S1). The other difference lies in the substitution on the methylthiophene molecule and 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 pyrrolidine is attached either first or second—a natural and intuitive grouping.

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 with the two non-aromatic cycles and one Suzuki coupling that forms 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 uses 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 arylation reactions. 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 subsequent tosyloxy alkylation reaction. The third reaction is Suzuki coupling, followed by a deprotection step and finally arylation. Again, it is clear that clustering helps to group similar routes based on the order of reactions and complexity of the route.

The third and final example highlights a convergent route where in the last step, an ether bond is formed by a substitution reaction. The difference between the clusters lies in how the two molecules forming the ether bond are synthesized. The first molecule is formed by Suzuki coupling followed by reduction in cluster 1 and cluster 2, but in cluster 3, the reduction is not necessary. The second molecule, which will form an ether bond with the first, is synthesized by forming a substituted 2,5-pyrroledione. In cluster 1 and 2, pyrroledione is formed from an anhydride and a substituted cyclopentane followed by reduction, whereas in cluster 2, it is formed from a ring-forming reaction. The routes within the clusters differ

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

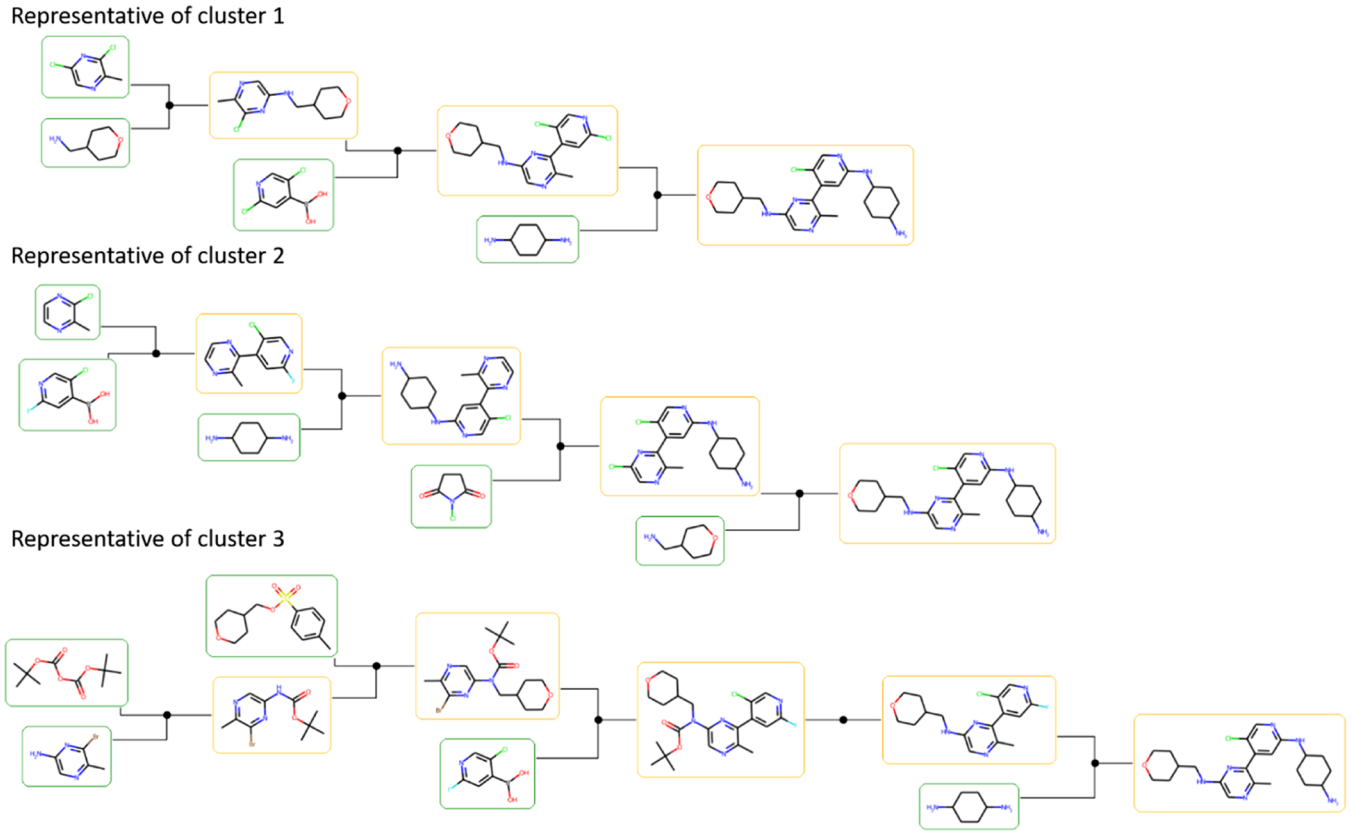


Figure 7. Cluster representatives for the second example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

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.

TED-Based Clusters Tend to Have Routes with Similar Chemistry. The above highlighted examples of the TED-based clusters are encouraging as they show that clustering appears to be intuitive. However, it is unfeasible to manually inspect the full set of clusters produced for the 5000 compounds. Instead, we devised a simple analysis to capture how similar the chemistry is within a cluster. For each pair of routes in a cluster, we compared the NextMove reaction class[15,30] of the first and last step of the routes. For convergent routes, we only compared the reaction class of the last reaction and we did not count reactions that were unclassified. Using this analysis, we found that there is 70% likelihood that routes in the same cluster have the same reaction class for the first step and 68% likelihood that the reaction class for the last step is the same. There is also 49% likelihood that the reaction classes of both of the last and first step are the same. Furthermore, because this analysis was only based on comparing the reaction class labels exactly, the likelihood would probably increase if we could group together similar reaction classes. If we instead compare routes that are not in the same clusters, the likelihood is 28% that the reaction class for the first step is equal, 36% that the reaction class for the last step is equal, and only 10% that both the last and the first steps have the same reaction class. This simple analysis further confirms that the TED-based procedure gives clusters that are intuitive and share the same chemistry.

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 encodings for a sample of routes. There is a clear but weak correlation; the correlation coefficient, r, is only 0.51. 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 also shown in Figure 9; 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 the 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 0.22. If we only do the similarity calculation for routes where the number of clusters is the same, the average similarity is 0.79 and the lowest cluster similarity is 0.37. This indicates that some of the differences in the cluster assignments could be attributed to the silhouette method used to determine the optimal number of clusters. However, the correlation between the difference in the number of clusters and cluster similarity is week ($r = -0.10$). Overall, the LSTM-based clustering leads to fewer clusters, each with more routes as seen in Figure S1. For the example compounds 2 and 3, discussed above and shown in Figures 7 and 8, the LSTM-based clustering produced four clusters for both compounds instead of three.

Cluster Assignments Can be Rationalized but the Optimal Number of Clusters Can be Subjective. We analyzed the difference between the TED-based and LSTM-based clustering closely by selecting an example compound

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

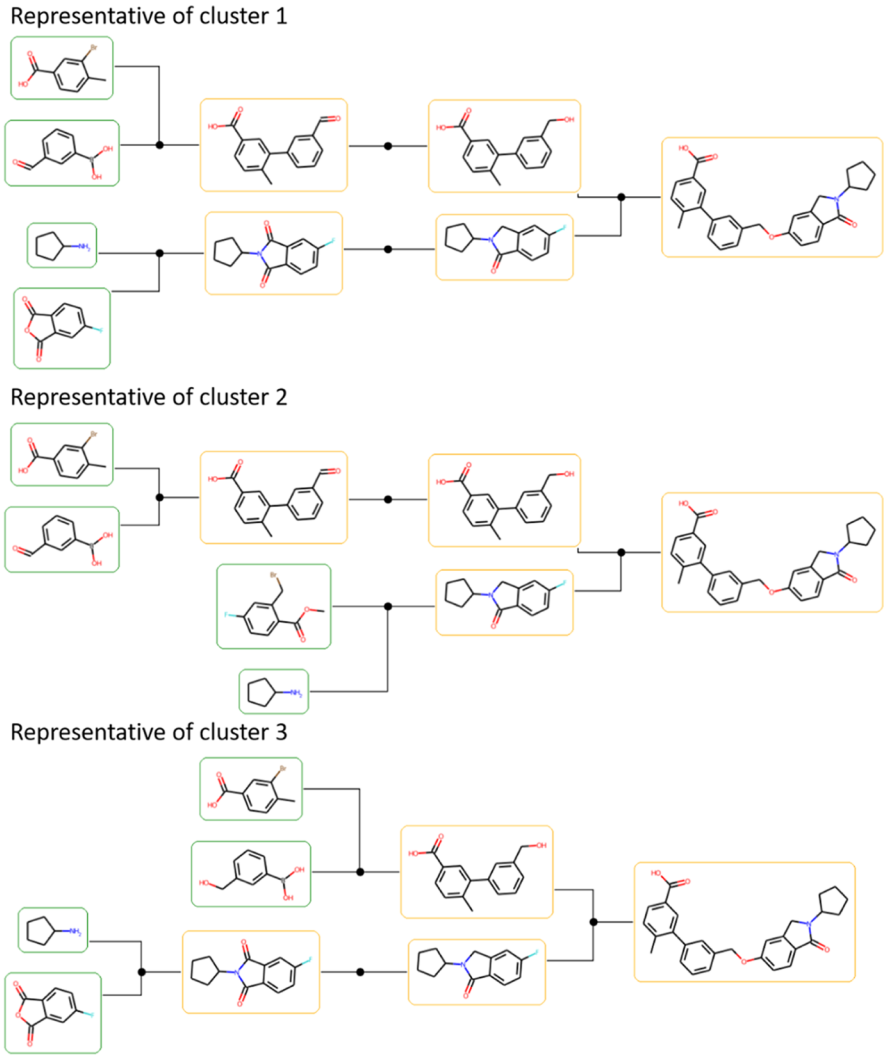


Figure 8. Cluster representatives for the third example compound. Molecules framed in an orange rectangle are not available in stock, whereas the compounds framed in a green rectangle are.

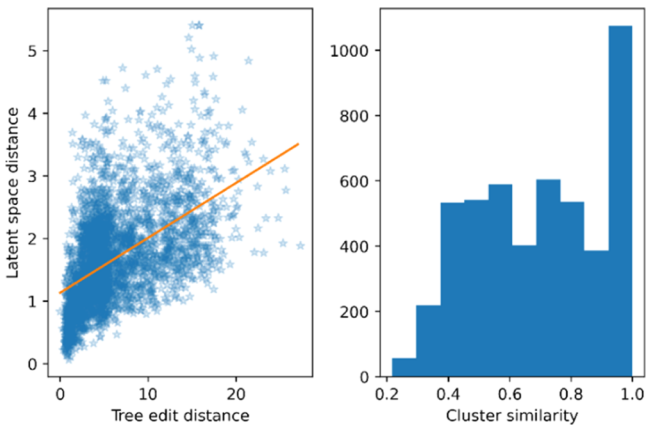


Figure 9. Similarity between the TED-based and LSTM-based clustering. (Left) Correlation between the underlying distances. A sub-sample of all the routes is shown as scatter, together with a linear regression line based on all routes. (Right) The distribution of the cluster similarity for all compounds.

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 10a,b. Additionally, in Table S4, we show all the routes as clustered using the TED matrix. TED-based clustering leads

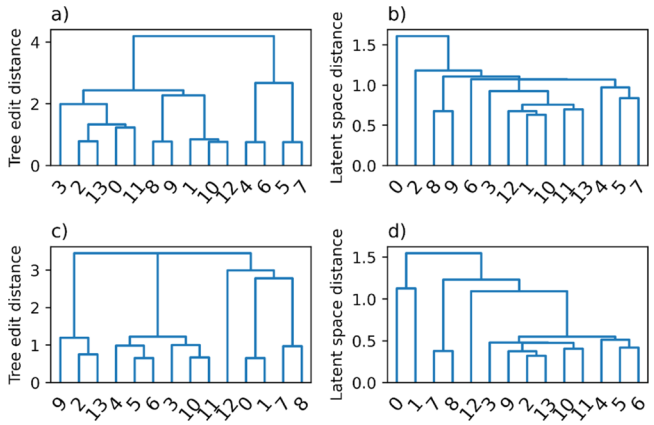


Figure 10. Dendrogram from distance matrices giving very different clusters. (a,b) Dendrograms for an example compound for which cluster similarity is 0.23. (c,d) Dendrograms for an example compound for which cluster similarity is 0.70.

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

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 to form an N−S bond. The second TED-based cluster is formed from routes 1, 10, and 12, and this cluster differs from cluster 1 in the first step which is acylation, forming an N−C bond. The second step is the same as in cluster 1. The remaining three clusters contain two routes each. Cluster 3 formed from routes 4 and 6 is characterized by an initial deprotection step followed by an 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 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 have to form only two clusters from the TED matrix, one cluster would be formed from routes 4 to 7 and one cluster would be formed from other routes, as is clear from the dendrogram in Figure 10a. Furthermore, it is clear from the analysis of reaction classes that the LSTM-based clusters are more chemically dissimilar than the TED-based clusters. There is only 56% likelihood that routes in the same cluster have the same reaction class for the first step, 54% likelihood that the reaction class of the last step is the same, and 36% likelihood that both reaction classes agree. These likelihoods are considerably lower than those for the TED-based clustering. To be fair, it should be pointed out that the LSTM network was trained to discriminate between human-made and predicted routes and not to produce intuitive clusters.[12] Therefore, it is likely that we can find examples where the LSTM-based clustering gives sub-optimal solutions.

For the example compound where the cluster similarity is average, we show the dendrograms in Figure 10c,d 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 the 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 are joined with clusters 3 and 4 from the TED-based clustering. Cluster 3 is 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 10c), whereas in the dendrogram of the latent space distance matrix, it is closer to, for instance, route 3 (see Figure 10d), 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.[31,32] For instance, in Figure S2, we show dendrograms for a compound where the silhouette optimization of the number of clusters arguably fails for the TED-based clustering. For this compound, route 19 was so different from all other routes when considering the TED matrix that this route is placed in a singleton cluster and all the other routes are placed in a second cluster. The latent space distance matrix on the other hand is more evenly spaced so that the silhouette optimization suggests four clusters. In this case, it would be more fruitful to decide the number of TED-based clusters manually. However, because we are only discussing a few examples here, 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 the TED-based clustering is generally more discriminative than the LSTM-based clustering and leads to more clusters that are more chemically similar.

TED-Based Clustering Can Separate Patent-like Routes from Patent-Evading Routes. Molga et al. used Chematica software to find patent-evading (PE) routes for the three drugs linezolid, sitagliptin, and panobinostat.[33] They were able to identify novel routes that did not contain the key reactions of the patented routes by implementing different search constraints. We manually extracted several of the routes from the paper and calculated the TED matrices to investigate if our clustering approach can separate patent-like (PL) routes from PE routes. The dendrograms of the distances matrices are shown in Figure 11. We see that for all three drugs, the PL routes are well-separated from the PE routes. If one would create two clusters for each drug, one cluster would contain all the PE routes and the other cluster would contain all the PL routes. This further highlights that the TED approach can create meaningful clusters and that it might be used to identify PE routes.

CONCLUSIONS

We have introduced a novel algorithm to compute distances between synthetic routes, and this has been used to cluster predicted routes. The clusters appear to be intuitive as they can

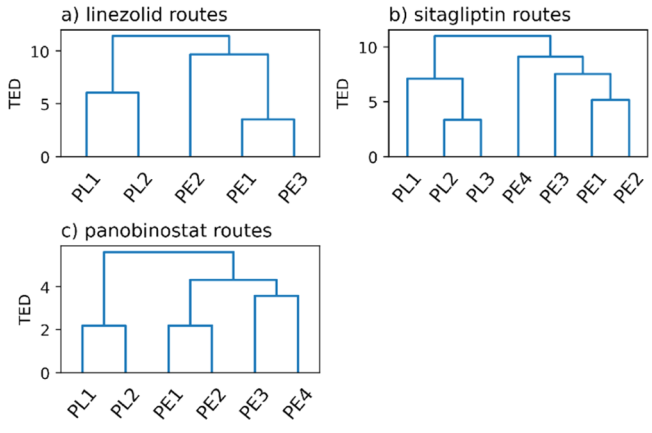


Figure 11. Dendrogram from distance matrices of routes identified using Chematica for three drugs. The routes are either PL or PE as detonated in ref 33. (a) PL1 and PL2 correspond to the routes in Figure 4C and PE1−PE3 correspond to Figure 4E in the original reference. (b) PL1 corresponds to Figure 5C, PL2 corresponds to Figure 5E, and PE1−4 corresponds to Figure 5F−I in the original reference. (c) PL1 and PL2 correspond to Figure 6C, and PE1−4 corresponds to Figure 6E−G in the original reference. Small variations in the same routes have been omitted from the analysis.

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907

be easily rationalized and we showed that the routes in a cluster belong to similar reaction classes. This implies that the clustering algorithm can reduce the number of predicted routes and thereby aid in the selection of routes for wet-laboratory synthesis. In contrast, the approach of Mo et al. sometimes gives unintuitive clusters, which could probably be explained by the training task.[12] The TED-based clustering is designed to be generally used but is particular useful for medicinal chemists in discovery chemistry, where the goal is to analyze a small number of possible synthetic routes. Because the clustering algorithm is on average fast for small route collections but does not scale very well for large sets of routes, it currently has limited use for process chemists who inspect thousands of routes. We have included the TED-based clustering algorithm in the latest release of AiZynthFinder software and a separate library (https://github.com/MolecularAI/route-distances).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00232.

Routes in figure 11 are provided (ZIP)

All routes corresponding to the representative routes in Figures 6−8; routes for which the dendrogram is shown in Figure 10; and additional plots and tables for the comparisons between the TED-based and LSTM-based clustering (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Samuel Genheden − *Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0002-7624-7363; Email: samuel.genheden@astrazeneca.com

Authors

Ola Engkvist − *Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0003-4970-6461

Esben Bjerrum − *Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0003-1614-7376

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00232

Notes

The authors declare no competing financial interest.

The software is available on Github: https://github.com/MolecularAI/aizynthfinder and https://github.com/MolecularAI/route-distances, and all the output from AiZynthFinder is available on Figshare: https://figshare.com/articles/dataset/Clustering_of_synthetic_routes_using_tree_edit_distance/13372505. Some of the data have been removed for proprietary reasons as we were using internal building blocks in the tree search, but the distances and clustering results are available.

■ ACKNOWLEDGMENTS

Rocío Mercado is acknowledged for proofreading the article.

■ REFERENCES

(1) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(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*, 1572−1583.

(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*, 434−443.

(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.

(5) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094−1106.

(6) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588*, 83−88.

(7) 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*, 4640−4651.

(8) 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. **2019**, arXiv:1910.08036.

(9) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic Retrosynthetic Route Planning Using Template-Free Models. *Chem. Sci.* **2020**, *11*, 3355−3364.

(10) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522−532.

(11) Shibukawa, R.; Ishida, S.; Yoshizoe, K.; Wasa, K.; Takasu, K.; Okuno, Y.; Terayama, K.; Tsuda, K. CompRet: A Comprehensive Recommendation Framework for Chemical Synthesis Planning with Algorithmic Enumeration. *J. Cheminf.* **2020**, *12*, 52.

(12) 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**, *12*, 1469−1478.

(13) 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* **2019**, *365*. DOI: 10.1126/science.aax1566

(14) Bille, P. A Survey on Tree Edit Distance and Related Problems. *Theor. Comput. Sci.* **2005**, *337*, 217−239.

(15) 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*, 154−168.

(16) 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*, 70.

(17) 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. *Nucleic acids res.* **2017**, *45*, D945−D954.
(18) RDKit *Open-source Cheminformatics*. http://www.rdkit.org.
(19) Genheden, S.; Engkvist, O.; Bjerrum, E. J. A Quick Policy to Filter Reactions Based on Feasibility in AI-Guided Retrosynthetic Planning. **2020**. ChemRxiv Prepr. https://doi.org/10.26434/CHEMRXIV.13280495.V1.
(20) Pawlik, M.; Augsten, N. Tree Edit Distance: Robust and Memory-Efficient. *Inf. Syst.* **2016**, *56*, 157−173.
(21) Pawlik, M.; Augsten, N. Efficient Computation of the Tree Edit Distance. *ACM Trans. Database Syst.* **2015**, *40*, 1−40.
(22) https://github.com/JoaoFelipe/apted (version 1.0.0).
(23) 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**, *79*, 52−59.
(24) Yoshino, T.; Higuchi, S.; Hirata, K. A Dynamic Programming A* Algorithm for Computing Unordered Tree Edit DistanceIIAI-AAI. *Proceedings—2nd IIAI International Conference on Advanced Applied Informatics*, 2013; pp 135−140.
(25) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(26) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(27) 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.
(28) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.
(29) https://github.com/moyiming1/Retrosynthesis-pathway-ranking (accessed November 24, 2020).
(30) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451−3479.
(31) Estivill-Castro, V. Why so Many Clustering Algorithms. *ACM SIGKDD Explor. Newsl.* **2002**, *4*, 65−75.
(32) Budka, M. Clustering as an Example of Optimizing Arbitrarily Chosen Objective Functions. *Stud. Comput. Intell.* **2013**, *457*, 177−186.
(33) Molga, K.; Dittwald, P.; Grzybowski, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, *5*, 460−473.

https://doi.org/10.1021/acs.jcim.1c00232
J. Chem. Inf. Model. 2021, 61, 3899−3907