# Efficient Data Undersampling for Rule-Based Retrosynthetic Planning

**MIN SIK PARK** ( mspark91@gmail.com )
Samsung Advanced Institute of Technology
**Dongseon Lee**
Samsung Advanced Institute of Technology
**Youngchun Kwon**
Samsung Advanced Institute of Technology
**Eunji Kim**
Samsung Advanced Institute of Technology
**Youn-Suk Choi**
Samsung Advanced Institute of Technology

**Research Article**

**Keywords:** retrosynthetic planning, organic molecules, synthetic database, undersampling

**Posted Date:** May 18th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-526435/v1

# Abstract

Computer-aided retrosynthetic planning for organic molecules, which is based on a large synthetic database, is a significant part of the recent development of an autonomous robotic chemist. As in other AI fields, however, the class imbalance problem in the dataset affects the prediction performance of retrosynthetic paths. Here, we demonstrate that applying undersampling methods to the imbalanced reaction dataset can improve the prediction of retrosynthetic rules for target molecules. We report improvements in the top-1 and top-10 prediction accuracies by 13.8% (13.1, 5.4%) and 8.8% (6.9, 2.4%) for the undersampling based on the similarity (random, dissimilarity) clustering of molecular structures of products, respectively. These results demonstrate the importance of a deep understanding of the statistical distribution, internal structure, and sampling for the training dataset. For practical application, the target-oriented undersampling method is proposed and confirmed by the improved prediction performance of 9.3 and 4.2% for top-1 and top-10 accuracies, respectively.


# Introduction

Traditional retrosynthetic planning for complex organic molecules has been the domain of trained and experienced chemists. In this task, a target molecule is transformed into precursors with simpler molecular structures. When the criteria of affordable and easy synthesis are met, the planning can be finished. Unlike the forward reaction prediction, which is mainly fixed by reactants, the retrosynthetic prediction can have numerous synthetic paths. Theoretically, the number of paths is roughly proportional to the factorial of the number of atoms (i.e., molecular complexity). Nevertheless, synthesis experts can narrow down this large possibility to few pathways using their experience and domain knowledge (such as considerations for short pathways, low prices, etc.). Therefore, time and computing resources will be wasted if the algorithm considers all possible paths. On the contrary, if the analysis is conducted by the synthesis experts alone, the experience and knowledge of human experts are limited in personal scope.

Researchers have published synthetic research results in journals and patents. Since 90’s, these data have been digitized in the form of huge databases. The advantages of these chemistry databases are as follows. First. all data have been experimentally verified (compared with the algorithmic consideration of all possible paths). Second, the amount of data is much greater than the domain knowledge of individual researchers. Therefore, with the recent rapid progress of deep learning, the use of machine learning algorithms to learn the latent space of retrosynthetic reaction rules has become a very active research area [1–11]. A rule-based model is a machine learning algorithm that learns the reaction rules that correspond to the input target molecules. Reaction rules consist of a pair of reaction centers for reactants and products. The reaction center means changing parts (atom and bonding) before and after the chemical reaction. Hence, the strength of rule-based models [1–4] is that it is easy to identify the selected reaction rules for the target molecule compared with the molecular transformer based on sequence-to-sequence models and the attention mechanism [5–9].

The class imbalance problem occurs in many databases, including chemical reaction databases, and impairs the performance of prediction models. In a reaction database, the frequency of reaction rules refers to the number of reaction samples covered by each rule. Furthermore, the reaction database we used in this study shows the imbalanced frequency distribution. This class imbalance may cause misclassification for minority reaction rules with low frequencies, which can reduce the diversity of predictions (i.e., exploration). To increase prediction accuracy (i.e., exploitation), we can reduce the number of rules using a high-frequency cut-off limit. However, there is a trade-off between the prediction accuracy and the diversity of reaction rules. In other words, a reduced number of rules can make the retrosynthesis planning too simple, which can make the rule prediction of the target molecules with high synthetic difficulty unsuccessful. To increase the prediction accuracy while maintaining the number of reaction rules (rule diversity), we can consider the oversampling (data augmentation), undersampling (data reduction), consolidating (integrating data from multiple data sources), collapsing (creating dataset with reduced size, which represents statistics of original dataset), and generalizing (creating summary data from original dataset) techniques. Among them, the oversampling can be performed by duplication of reaction datasets included in minor reaction rules with low frequency [12–14]. Contrastingly, the undersampling can be performed by removing a partial reaction dataset included in major reaction rules with high frequency [15, 16]. For this reason, the undersampling is suitable in cases of large imbalanced datasets; it can save time and computational resources required for training.

To solve the class imbalance problem in the retrosynthesis planning, multiscale reaction classification [17] and data augmentation [18] methods have been studied, which have improved the prediction accuracy by 3–6% (average balanced accuracy) and 4–6% (top-1–top-10 accuracy, and only small dataset showed an improved accuracy between top-1 and top-10), respectively. In these studies, the USPTO dataset [19] was used, and the number of extracted reaction rules was approximately 7,700–187,000. However, the effect of undersampling on large reaction datasets remains unclear. Here, we present the effect of undersampling to improve the prediction accuracy in the retrosynthesis planning and propose the target-oriented undersampling method for practical application in an autonomous synthesis system.

# Results

### Characterization of Extracted Chemical Reaction Rules

We used the reaction database showing the imbalanced frequency distribution, where the reaction rules having more than 10 frequency occupied 0.63% (61,234 rules) only among the rules with frequency ≥ 1. On the contrary, the covered reaction data (frequency ≥ 10) occupied 21.3% (3,395,642 reaction data) among the corresponding reaction data (frequency ≥ 1) (Table S1), which indicates a high imbalance of our database. To improve the prediction accuracy by reducing the number of classes (reaction rules, 9,672,940 à 61,234, Figure S1) while maintaining most of the reaction database (15,930,914 à 3,395,642), the reaction rules having frequency under 10 were removed (Table S1). As in the following discussion, the reaction rules with low frequency have large size compared to the ones with high frequency, therefore, it is

expected that the reaction rules with the larger size (frequency < 10) is normally subset of the reaction rules with the smaller size (frequency ≥ 10). Therefore, the dataset having frequency more than 10 were selected as the test bed for undersampling study without losing generality, which also shows high class imbalance (Figure 1, bottom).

The top-10 reaction rules selected by frequency are presented in Figure 2. The rank 1 reaction rule has a maximum frequency of 82,906, which is a chemical reaction between carboxylic acid and ester functional groups. Among top-10 reaction rules, six rules are related to the functional groups containing oxygen (such as ester, carboxylic acid, alcohol, and ether), and the other four rules are related to the nitrogen functional groups (such as amine, nitro, and azide). Furthermore, the protecting groups of tert-butyldimethylsilyl, tert-butoxycarbonyl, tert-butyl, and acetyl for amine and alcohol functional groups are involved. Although the part of the Reaxys database [20] is used, we expect that these top-10 reaction rules are also major reactions historically identified by chemists. The explanation is that the database that we used (approximately 16M) corresponds to more than half of the reaction database with one product and one- and two-reactants (approximately 28M) in the Reaxys [20].

High-ranking reaction rules have a relatively small size (i.e., short length). Except for the protecting groups, center atoms in reaction centers have one (radius=1) or two (radius=2) adjacent heavy atoms before and after the reaction (Figure 2). On the contrary, low-ranking reaction rules include multiple (radius ≥ 3) heavy atoms near the center atom. Therefore, we expect that most excluded rules having a frequency less than 10 can be a subset of the selected rules having frequency over 10, where the size and details are different, but the reaction center is the same.

Higher frequency cut-offs can be used to further improve the prediction accuracy of machine learning models. However, the elimination of additional large size reaction rules can affect the reaction yield depending on the specific atomic environment near the reaction center, where specific large reaction rules can provide higher yields than general small reaction rules. Hence, the trade-off between the diversity of the rule affecting yields and the improvement of prediction accuracy by the frequency cut-off should be carefully considered when generating datasets.

### Quantitative Performance on the Undersampling Datasets

Our strategy of undersampling is to improve the prediction accuracy while maintaining the number of reaction rules. In other words, our approach is to reduce data imbalance, which increases the prediction accuracy for minor frequency rules. In Figure 3, structural clusters based on random, similarity, and dissimilarity undersampling are shown for rank 1 and rank 3 reaction rules corresponding to the top rank rules for oxygen (Figure 3(a)) and nitrogen (Figure 3(b)) functional groups, respectively. In Figure 3(a), the similarity cluster (blue box) includes organic molecules with similar scaffolds (conjugated rings) and carboxylic acid group. However, the dissimilarity cluster (green box) has organic molecules with dissimilar scaffolds of rings and chains. Finally, the organic molecules having conjugated and saturated rings and a side-chain having carboxylic acid group appears in a random cluster (red box), which is in between similarity and dissimilarity clusters. In addition, a similar relationship is demonstrated in the

molecular clusters having an amine functional group (Figure 3(b)). All amine functional groups are located next to the conjugated ring and, especially, cellobiose ($C_{12}H_{22}O_{11}$) and 2H-pyran with an acetoxymethyl and triyl triacetate are mainly included in random cluster (red box).

Similarly, for each of the 61,234 reaction rules, three undersampling (random, similarity, and dissimilarity) datasets having the size of 612,340 were prepared using the Taylor–Butina clustering algorithm [21, 22]. Neural network models were trained on using those random, similarity, and dissimilarity datasets. The 5% prediction data randomly separated from each dataset were used to measure the prediction accuracy, in which the prediction dataset sizes are 239,877 and 30,617 for baseline and undersampling models, respectively. Figure 4(a) presents the top-10 prediction accuracies of a baseline and three undersampling trained models for each prediction dataset. In particular, the prediction accuracies of three undersampling models were averaged by using both prediction datasets independently and commonly sampled from three undersampling datasets for statistical robustness, therefore, standard errors were marked on each averaged undersampling graph in Figure 4(a) and the overall averaged undersampling graph in Figure S3. For the top-1 accuracy, the similarity and random models show a higher accuracy compared with the dissimilarity and baseline models. In particular, the similarity model shows a higher accuracy in all ranges of top-k among all models due to the characteristic of the prediction dataset with structural similarity. By analogy with the one-sided selection [16] technique, the selected 10 samples for each reaction rule may show the obvious representativeness by removing data having noise and near the boundary between classes (reaction rules). Indeed, the large molecules in the dissimilar datasets (Figure 3) have a chance to match multiple reaction rules having a similar probability. Thus, the prediction accuracy of the dissimilarity model is relatively low. In addition, all undersampling models show higher-than-baseline accuracies for all top-k, which can be understood as mitigating the data imbalance problem through undersampling.

Despite these improvements, the undersampling method results in information loss, as some data are discarded. To examine the effect of this information loss on the prediction accuracy, the same prediction dataset was used for all undersampling models. Figure 4(b) depicts the prediction accuracy obtained from all undersampling models using the same random dataset. In all ranges of top-k, the similarity model shows the highest prediction accuracy among all models. Unlike in the previous results, the dissimilarity model shows the second-highest accuracy, and the random model shows the lowest performance for all top-k. Besides, the top-1 accuracies for similarity and dissimilarity models have been increased (Table S2). This is probably due to that some prediction data in the random dataset may have structural similarity with the similarity and dissimilarity training datasets. On the contrary, the training and prediction data in the random dataset may not be structurally similar. Indeed, as shown in Figure 3, the undersampled data do not match exactly, and some samples in each dataset are structurally similar. Figure 4(c) shows the prediction accuracy obtained from all undersampling models using the same similarity dataset. The random model shows the highest prediction accuracy. Furthermore, the dissimilarity model indicates a higher prediction accuracy than the similarity model. In Figure 4(d), the dissimilarity prediction dataset was used for all undersampling models. Similarity and random models

show similar tendencies for all top-k and overlap with the dissimilarity model after top-7. These results stem from the characteristic of the dissimilar prediction dataset. As shown in Figure 3, the molecular structures of dissimilarity datasets are heterogeneous compared with the random and similarity datasets. The convergence of prediction accuracies in all undersampling models after top-7 can be explained by such a highly heterogeneous dataset. These results confirm that the dissimilarity undersampling was successful.

To get more extended picture for this information loss, the two oversampling experiments based on random and synthetic minority oversampling (SMOTE) [12, 13] samplings were performed, where the same frequency cut-off of 10 was used for reaction rules. To reduce the burden on the vast amount of oversampled data, the hybrid method was applied, in which the undersampling and oversampling were applied for the dataset of reaction rules over frequency 20 (36% of the reaction rules) and under frequency 20 (62% of the reaction rules), respectively (Figure S4(a)). Therefore, the size of oversampling dataset is twice the size of undersampling dataset. The prediction accuracies of oversampling models are shown in Figure S4(b). Both random and SMOTE samplings show very similar prediction accuracy overall range of top-k. Comparing to the accuracy of random undersampling method (Figure 4(a)), the 6% for top-1 and 2% for top-10 were improved, which seems to be due to a doubling of the dataset size. It is also expected that the prediction accuracy of the model trained with both oversampled data under frequency 20 and original data over frequency 20 is lower than the hybrid model we used due to data imbalance. This shows that the oversampling model can improve the prediction accuracy and the hybrid model using oversampling for minor rules and undersampling for major rules can be an alternative model between oversampling model with a training burden and undersampling model with some information loss.

Furthermore, higher frequency cut-off models that combine undersampling methods can be considered. In this case, the reaction data per reaction rule increases, and the number of reaction rules (i.e., classes) decreases by applying a higher-frequency cut-off. In other words, the overall prediction accuracy increases, but the diversity of reaction rules decreases. Hence, the number of reaction samples corresponding to survived reaction rules also increases, and this may remedy the loss of information by undersampling. Compared to previous studies of multiscale reaction classification [17] and data augmentation [18] methods, the effect of data sampling on prediction accuracy can be relatively large. In other words, inappropriate sampling in the undersampling model may result in low prediction performance.

### Qualitative Analysis based on Detailed Examples

For the qualitative analysis of the baseline model and the three undersampling (random, similarity, and dissimilarity) models, we selected three target molecules from three reaction rules with different frequencies. In addition, target-oriented undersampling experiments based on these target molecules were performed by using the training dataset sampled by structural similarity with each target molecule, where the structural similarity was calculated for all product molecules in each reaction rule using the

Tanimoto coefficient [23]. The 10 reaction samples with small pairwise distance for each reaction rule were selected for the training and prediction dataset. As shown in Figure 5, the first molecule ( 2,6-dimethylphenanthridine) was selected from the reaction rule with a frequency of 10, and, therefore, all reaction samples containing the first molecule are included in all datasets of both the baseline and three undersampling (random, similarity, and dissimilarity) models. Hence, the prediction of reaction rule for this molecule using the baseline and four undersampling (random, similarity, dissimilarity, and target-oriented) models represents one example of whether the imbalance problem is mitigated by undersampling. Figure 5(c) shows the results of a single-step retrosynthesis for 2,6-dimethylphenanthridine, in which the predicted reactants between top-1 and top-5 are represented for the baseline model and four undersampling models. Due to the dataset imbalance, the baseline model predicted a ground truth in top-3, where the ground-truth rule corresponds to the minor rules with frequency 10. Contrastingly, the similarity, dissimilarity, and target-oriented models result in higher-ranked top-1, top-2, and top-2 predictions for the ground truth, respectively, which is higher than the rank of the baseline model. Meanwhile, the random model predicts the ground truth at top-4, which is lower than the rank of the baseline model. In addition, the predicted reaction rules between top-1 and top-5 are the same for both the baseline and four undersampling models, where the only difference is in the ranks of the predicted reaction rules. Although this target molecule was included in individual datasets for all prediction models, the baseline model predicts the ground truth in the top-3, which ranks relatively low compared with undersampling models. This example confirms that the undersampling method can mitigate the imbalance problem.

Figures 5c and  present how the four undersampling models work for the target molecules not included in all training datasets except for the baseline model (Figure 5(a)). The two target molecules were selected from the reaction samples of major reaction rules with frequencies of 208 and 373 (Figure 5(b)). In Figure 5(c), the prediction results for the target molecule of 7H-benzo[c]phenothiazine reveal that the baseline model ranks the ground truth at the top-1, which may be attributed to both the reaction rule with a high frequency of 208 and the dataset containing the target molecule. Among the undersampling models, the dissimilarity and target-oriented models predict the ground truth as top-1, and the other two models failed to predict the ground truth within top-5. The degree of overlap in the predicted reaction rules between the baseline model and each undersampling model is only 20% each. In contrast, the degrees of overlap are 60, 80, and 60% between similarity and dissimilarity models, similarity and random models, and dissimilarity and random models, respectively. In other words, the three undersampling models show similar prediction tendencies compared with the baseline model, and it is mainly related to the differences between the training datasets with distinct statistical distributions. However, the target-oriented model shows only 0, 20, 0 % overlap with similarity, dissimilarity, and random models, respectively.

Finally, an additional selection criterion for the target molecule of methyl (E)-3-(2-methoxy-5-(1-(3,4,5-trimethoxyphenyl)vinyl)-phenyl)acrylate is a large molecule, which can contain multiple reaction centers (Figure 5(c)). As a result, only the random undersampling model predicts the ground truth in the top-1. In contrast, the baseline model predicts it in the top-4, even though the reaction rule of the ground truth has a higher frequency of 373 compared with 10 frequency used in the random undersampling model. The

degree of overlap in the predicted reaction rules between the baseline model and the undersampling models (similarity, dissimilarity, random, and target-oriented) is 0, 20, 40, and 40%. The degree of overlap between the undersampling models is 0 ~ 40%. Hence, all prediction models had difficulties with predicting the reaction rule of the ground truth due to multiple reaction centers in the large target molecules regardless of the statistical distribution of the training datasets.

For these three target molecules, the prediction accuracy of target-oriented models was shown in Figure 5(d), where top-k accuracies are very similar for all target molecules. This result represents the robustness of target-oriented models for the prediction accuracy irrespective of target molecules. In addition, the top-1 and top-10 accuracies are higher than those of baseline model by 9.3 and 4.2%, respectively. Hence, this target-oriented model can be an alternative model overcoming inappropriate sampling problem in undersampling methods.

# Discussion

In this work, we have studied the undersampling approach for improving the prediction of minority reaction rules extracted from an imbalanced chemical reaction dataset for the retrosynthesis path planning. Four datasets—with random, similarity, dissimilarity, and target-oriented undersampling—were prepared using a clustering algorithm for grouping the product molecules according to their structural similarity. We found that the data imbalance, the statistical distribution of datasets, and the size of the target molecules can affect the prediction accuracy for the reaction rules.

Based on our results, we expect that an optimized model showing both a higher prediction accuracy and a prediction of diverse reaction rules can be found by balancing the number of reaction rules through a frequency cut-off and the number of reaction samples for each reaction rules through undersampling. In other words, the trade-off between the diversity of rule affecting yields and the improvement of prediction accuracy by the frequency cut-off should be carefully considered when generating datasets. A baseline model trained on a dataset following the statistical distribution of a published database can be appropriate for the exploitation of reaction rules. In contrast, the undersampling models trained with the datasets following a uniform distribution can be suitable for the exploration of reaction rules. In addition, the extracted reaction rules can be split or merged. The major reaction rules with high frequency can be split into multiple minor rules using dissimilarity clustering, which can result in a database of more reaction rules with smaller frequencies. On the other hand, the minor reaction rules with low frequency can be merged into major rules, resulting in a database of less reaction rules with higher frequencies. These databases can affect the prediction accuracy for reaction rules. In this regard, further research on the undersampling method through the optimization process for a given problem and an experimental verification of the rule-based undersampling model will be conducted using an autonomous robotic synthesis system currently under development.

# Methods

## Dataset Preparation

The reaction dataset used in this work was obtained from the Reaxys database [20]. The dataset contains the reaction data as strings in the simplified molecular input line-entry system (SMILES) [24], where the reaction pair SMILES of reactants and products were transformed from the extended connection tables (V3000), which were downloaded from the Reaxys database, using an RDKit Python module [25]. The reactions having the form of “Reactant1_SMILES.Reactant2_SMILES. … >> Product1_SMILES.Products2_SMILES. …” were generated. The reaction data were constrained by the conditions of single and two reactants only, single-product and single-step only, no reagents. Finally, we used 16 million reactions to extract chemical reaction rules. (Figure S1)

## Extraction of Chemical Reaction Rules

A chemical reaction rule consists of changing parts, including atoms and the bonds between reactants and a product. These reaction rules are expressed in the form of SMARTS [26], which can express the partial structure in a molecule. The reaction rules were extracted from the prepared reaction dataset as follows. First, the atomic mapping between reactants and a product was applied using an RDKit [25] (Supporting Information (SI) Section S1). Second, the reaction rules having the form of SMARTS were generated using an RDChiral Python module which can handle stereochemistry for the introduction, destruction, retention, and inversion of tetrahedral centers and the cis/trans chirality of double bonds [27]. Third, all extracted reaction SMARTS rules were checked for error by reproducing reactants from each product. Finally, we obtained a dataset of 61,234 reaction rules (Figure S1) and 3,395,642 reactions after doing a cut-off over 10 frequency in 16 million reactions (baseline; Figure 1, top).

## Training and Prediction Models

Neural networks can learn and predict reaction rules. The dimension of an input vector is determined by the size of a Morgan fingerprint transformed from the SMILES of a product molecule using an RDKit [25]. We used a bit vector with a size of 8,192; hence, the dimension of an input layer was the same. For hidden layers, we used a single layer with 256 nodes. The output layer has the same number of reaction rules (i.e., 61,234 dimensions). For training, a rectified linear unit activation [28] and a dropout [29] ratio of 0.5 were applied. In addition, the Adam optimizer [30] (learning rate = 0.001), cross-entropy loss, and mini-batch with a size of 128 were used. At the output layer, a log-softmax activation was applied. PyTorch was used for training all models [31]. Training of neural network models was performed on 4 NVIDIA Tesla V100 GPUs and 4 M40 GPUs for the baseline and undersampling models, respectively.

## Similarity Clustering and Undersampling

The structural similarity was clustered as follows. First, the Morgan fingerprint with length 1,024 bit was used for encoding the product molecules. Second, the pairwise distance between the fingerprints of the product molecules was calculated using the Tanimoto coefficient [23]. Finally, the structural clustering

was performed using the Taylor–Butina algorithm [21, 22] with the inputs of the calculated pairwise distance.

The undersampling was performed from the baseline dataset (Figure 1, bottom). The first case is a random sampling, in which 10 reaction samples were randomly selected for each reaction rule, regardless of structural similarity or dissimilarity. The second one is a similarity sampling, in which 10 reaction samples were selected, as many as possible from a single cluster, including structurally similar products, if available. For this, a distribution threshold of 0.7 was used for the structural clustering. The third one is a dissimilarity sampling, in which 10 reaction samples with structurally dissimilar products were uniformly chosen from different clusters, where a distribution threshold of 0.3 in the clustering was used. The last one is a target-oriented sampling, in which 10 reaction samples were selected based on their structural similarity to the target molecule by using the Tanimoto coefficient for each reaction rule (Figure S2).

# Declarations

### Code availability

An executable file and example USPTO datasets for the atomic mapping are available for testing purposes only at http://github.com/mspark91/Atomic_Mapping/tree/master.

### Author contributions

M.S.P. implemented the models and performed analysis. Y-S.C. supervised the project. D.S.L., Y.K., and E.K. helped with the data collection. M.S.P. wrote the manuscript. All authors contributed to discussions and revisions.

### Competing interests

The authors declare no competing financial interest.

### Additional information

**Supplementary information** is available for this paper at

**Correspondence and requests for materials** should be addressed to M.S.P.

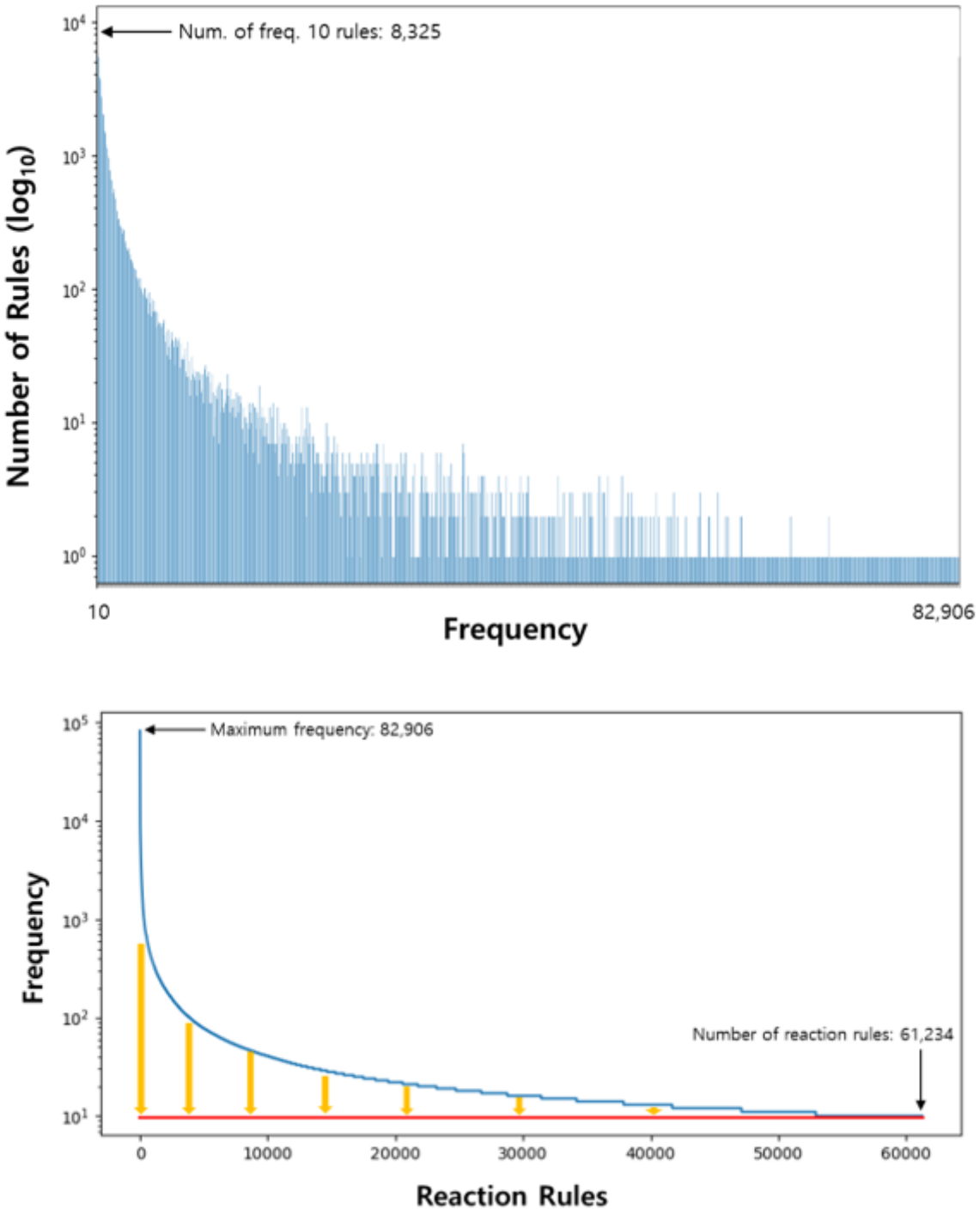
# References


1. Segler, M. H. S. & Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **23**, 5966–5971 (2017).
2. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature.* **555**, 604–610 (2018).

3. Badowski, T., Gajewska, E. P., Molga, K. & Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem., Int. Ed.* **59**, 725–730 (2020).
4. Fortunato, M. E., Coley, C. W., Barnes, B. C. & Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. *J. Chem. Inf. Model.* **60**, 3398–3407 (2020).
5. Liu, B. *et al.* Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
6. Chen, B., Barzilay, R. & Jaakkola, T. S. Path-Augmented Graph Transformer Network. Preprint at https://arxiv.org/abs/1905.12712 (2019).
7. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
8. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).
9. Kim, E., Lee, D., Kwon, Y., Park, M. S. & Choi, Y-S. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *J. Chem. Inf. Model.* **61**, 123–133 (2021).
10. Ishida, S., Terayama, K., Kojima, R., Takasu, K. & Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* **59**, 5026–5033 (2019).
11. Dai, H., Li, C., Coley, C. W., Dai, B. & Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. Proc. Adv. Neural Inf. Process. Syst. 8870–8880(2019).
12. Chawla, N. V., Hall, L. O., Bowyer, K. W. & Kegelmeyer, W. P. SMOTE: Synthetic minority oversampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
13. Han, H., Wang, W. & Mao, B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proc. Int. Conf. Intell. Comp. 878–887(2005).
14. He, H., Bai, Y., Garcia, E. A. & Li, S. A. D. A. S. Y. N. Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proc. Int. J. Conf. Neural Netw. 1322–1328(2008).
15. Liu, X. Y., Wu, J. & Zhou, Z. H. Exploratory Undersampling for Class-Imbalance Learning. Proc. Int. Conf. Data Mining 965–969(2006).
16. Kubat, M. & Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. Proc. Int. Conf. Mach. Learn. 179–186(1997).
17. Baylon, J. L., Cilfone, N. A., Gulcher, J. R. & Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **59**, 673–688 (2019).
18. Fortunato, M. E., Coley, C. W., Barnes, B. C. & Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. *J. Chem. Inf. Model.* **60**, 3398–3407 (2020).

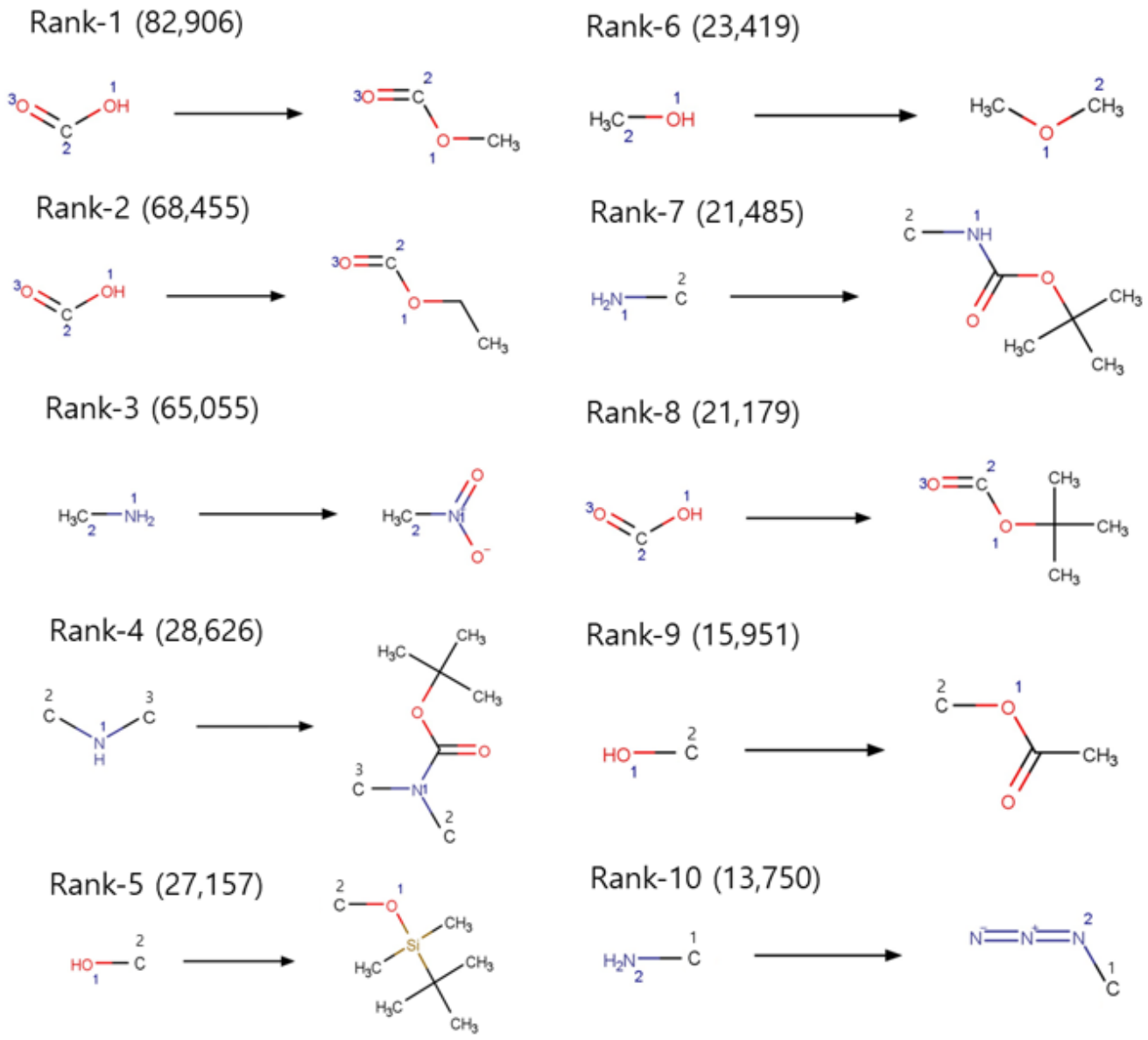
19. Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature; University of Cambridge(2012).
20. *Reaxys* http://www.reaxys.com (Elsevier Life Sciences, 2019, accessed Apr. 2019).
21. Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Model.* **35**, 59–67 (1995).
22. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Model.* **39**, 747–750 (1999).
23. Rogers, D. J. & Tanimoto, T. T. A Computer Program for Classifying Plants. *Science.* **132**, 1115–1118 (1960).
24. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
25. Landrum, G. A. & RDKit Open-Source Cheminformatics Software. http://www.rdkit.org, (2019).
26. Sayle, R. 1st-class SMARTS patterns.EuroMUG97, (1997).
27. Coley, C. W., Green, W. H., Jensen, K. F. & RDChiral An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inf. Model.* **59**, 2529–2537 (2019).
28. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on International Conference on Machine Learning 807–814(2010).
29. Srivastava, N. *et al.* A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
30. Kingma, D. P., Ba, J. & Adam A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
31. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Proc. Adv. Neural Inf. Process. Syst. 8026–8037(2019).

# Figures



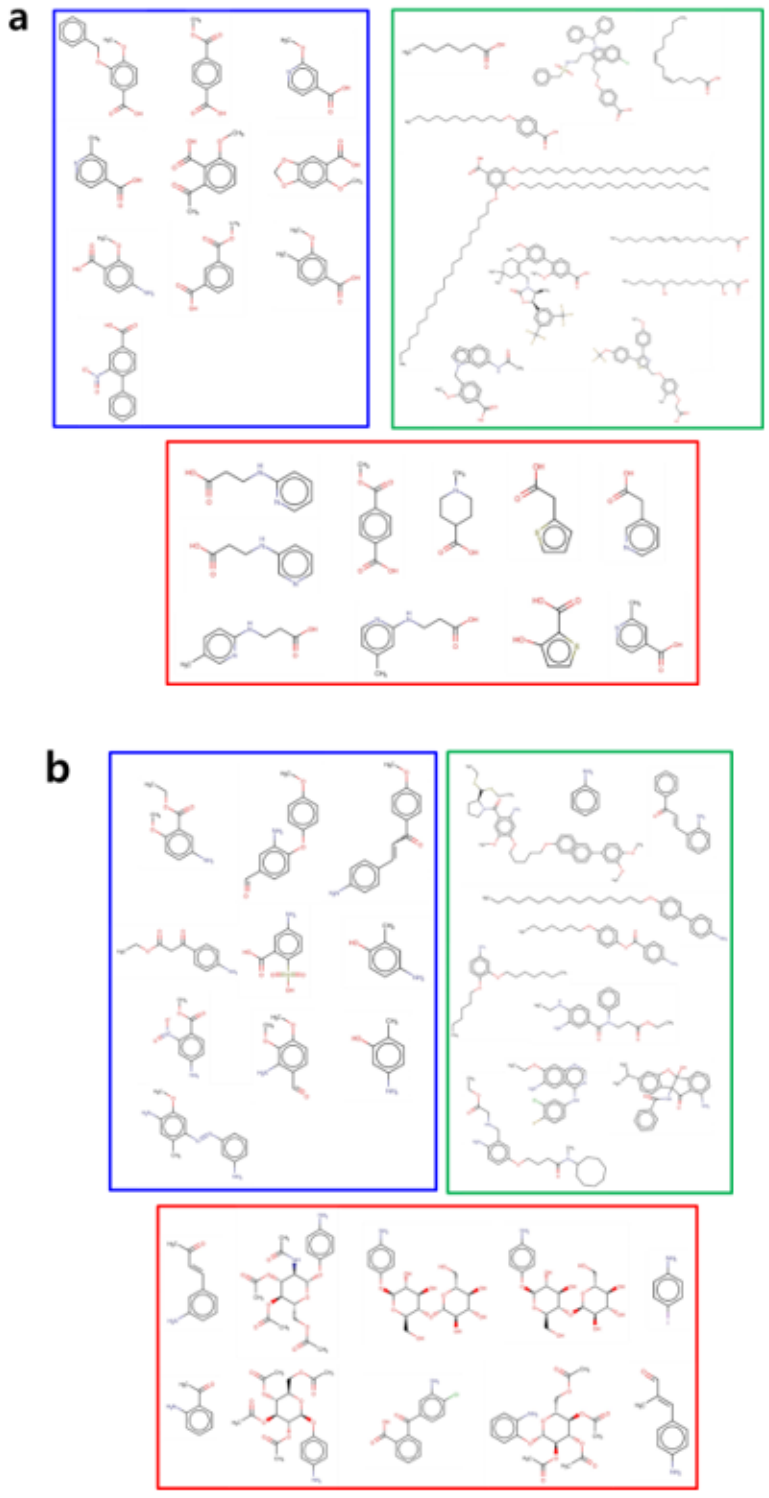

## Figure 1

Distribution of the reaction database following the frequency and representative reaction rules. Top: number of rules depending on the frequency over 10 in the baseline dataset. Bottom: distribution of frequency for the reaction rules over 10 cut-off, where orange downward arrows represent the dataset undersampling. Thus, all reaction rules have 10 frequency (red line).



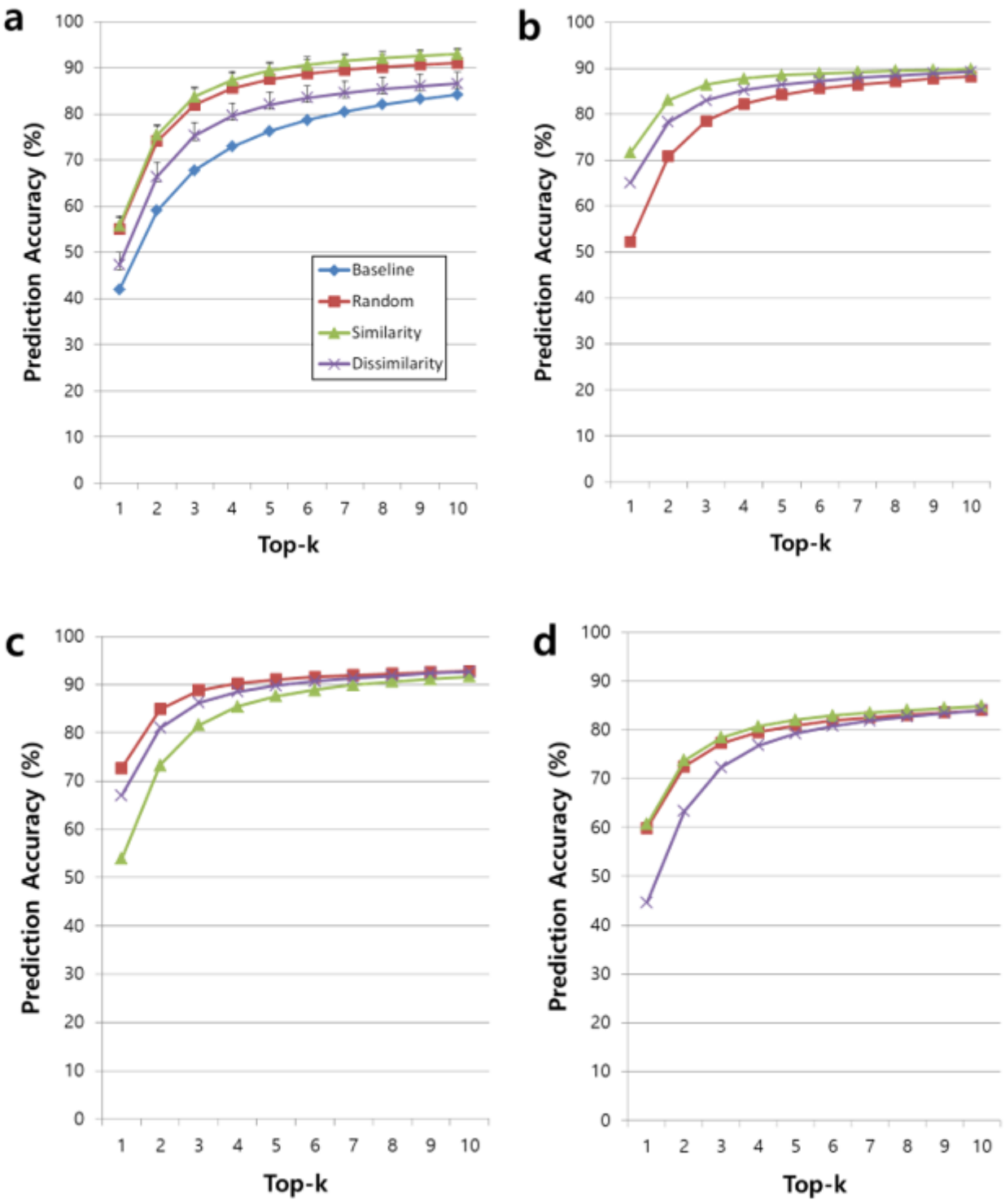

## Figure 2

Top-10 reaction rules selected by frequency (numbers in parentheses), which are extracted from the baseline dataset.




## Figure 3

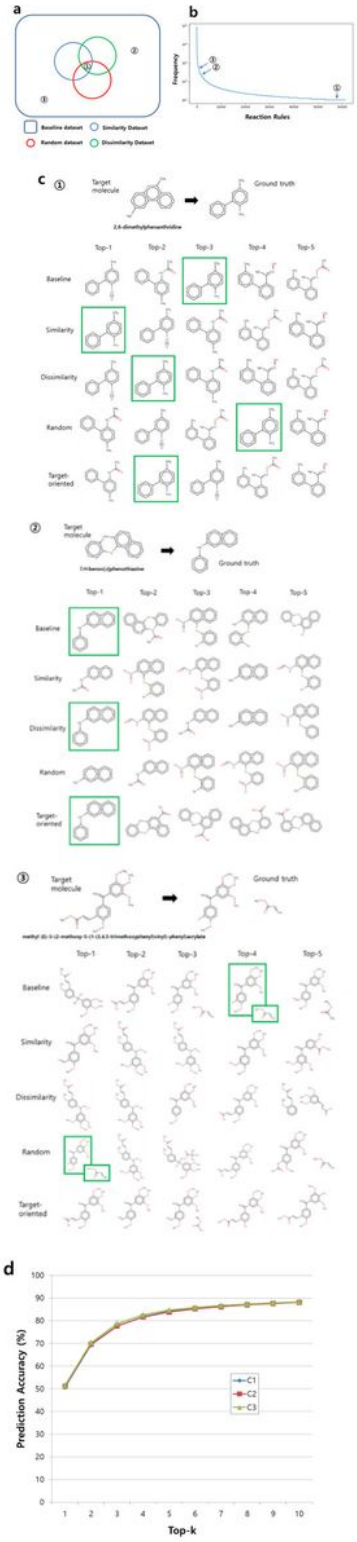
Structural clustering of product molecules by random, similarity, and dissimilarity in undersampling; clustering for (a) rank 1 rule and (b) rank 3 rule in Figure 2. The red, blue, and green boxes correspond to random, similarity, and dissimilarity clusters, respectively.




## Figure 4

Prediction accuracy for the baseline model and random, similarity, dissimilarity undersampling models using (a) prediction datasets separated from their own database, respectively (standard errors are marked on each undersampling graph); Prediction accuracy for the analysis of information loss effect in the undersampling models using (b) a prediction dataset separated from the random dataset; (c) a prediction

dataset separated from the similarity dataset; (d) a prediction dataset separated from the dissimilarity dataset.




## Figure 5

Top-5 single-step results of retrosynthesis planning for three target molecules with different frequencies: (a) a diagram for the relative inclusion relation between the datasets and the location of three target molecules; (b) positions of three target molecules in the graph of the frequency distribution; (c) top-5

retrosynthesis paths for  2,6-dimethylphenanthridine, where a reactant inside the green box denotes the ground truth. Top-5 retrosynthesis paths for  7H-benzo[c]phenothiazine and  methyl (E)-3-(2-methoxy-5-(1-(3,4,5-trimethoxyphenyl)vinyl)-phenyl)acrylate, where reactants inside the green box denote the ground truth. d) Prediction accuracy for target-oriented models of target molecules 2,6-dimethylphenanthridine (C1), 7H-benzo[c]phenothiazine (C2), and methyl (E)-3-(2-methoxy-5-(1-(3,4,5-trimethoxyphenyl)vinyl)-phenyl)acrylate (C3) in Figure 5(c).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingInformation.pdf