RETROTRAE: RETROSYNTHETIC TRANSLATION OF ATOMIC ENVIRONMENTS WITH TRANSFORMER

UMIT V. UCAK

Department of Chemistry, Division of Chemistry and Biochemistry,
Kangwon National University, Chuncheon, 24341, Republic of Korea

ISLAMBEK ASHYRMAMATOV

Department of Chemistry, Division of Chemistry and Biochemistry,
Kangwon National University, Chuncheon, 24341, Republic of Korea

JUNSU KO

Arontier co. Seoul, 06735, Republic of Korea

JUYONG LEE

Department of Chemistry, Division of Chemistry and Biochemistry,
Kangwon National University, Chuncheon, 24341, Republic of Korea

Arontier co. Seoul, 06735, Republic of Korea

E-mail addresses: umit@kangwon.ac.kr, ashyrmamatov@kangwon.ac.kr, junsuko@arontier.co, juyong.lee@kangwon.ac.kr, junsuko@arontier.co.

Date: August 14, 2021.

Abstract. Herein we present a new retrosynthesis prediction method, viz. RetroTRAE, which uses fragment-based tokenization combined with the Transformer architecture. RetroTRAE mimics chemical reasoning, and predicts reactant candidates by learning the changes of atomic environments associated with the chemical reaction. Atom environments stand as ideal, chemically meaningful building blocks, which together produce a high-resolution molecular representation. Describing a molecule with a set of atom environments establishes a clear relationship between translated product-reactant pairs due to the conservation of atoms in the reactions. Our model achieved a top-1 accuracy of 68.1% within the bioactively similar range for the USPTO test dataset, outperforming other state-of-the-art translation methods. Besides yielding a high level of overall accuracy, the proposed method solves the translation issues arising from the SMILES-based retrosynthesis planning methods effectively. Through careful inspection of reactant candidates, we demonstrated atom environments as promising descriptors for studying reaction route prediction and discovery. RetroTRAE provides fast and reliable retrosynthetic route planning for substances whose fragmentation patterns are revealed. Our methodology offers a novel way of devising a retrosynthetic planning model using fragmental and topological descriptors as natural inputs for chemical translation tasks.

1. Introduction

Planning the reaction pathways of organic molecules is a central component of organic synthesis. The idea of reducing the complexity of a desired organic molecule by considering all logical disconnections forms the basis of the retrosynthetic approach [1–3]. Therefore, the aim of the retrosynthetic approach is therefore to suggest a logical synthetic route to generate a target molecule from a set of available reaction building blocks. The retrosynthetic approach acts recursively on the target molecule until chemically reasonable pathways are identified [4]. From a broader perspective, predictors for forward and backward reactions reported in the literature can be classified into those that rely on the construction of reaction templates and those that are template-free, data-driven networks trained in an end-to-end fashion. Template-free methods have emerged as an effective means of addressing the methodological limitations of the template-based paradigm. These methods can be further subdivided according to the way of molecular representation protocol: (i) graph-based methods [5–8] and (ii) sequence-based methods [9–11, 43].

Sequence-based modeling recasts the problem of reaction pathway planning as a language translation problem by using a string representations of molecules. Current state-of-the-art forward- and backward-reaction predictors are mostly built on the Transformer architecture [13]. The Transformer is a neural machine translation (NMT) model that solely depends upon attention mechanism [12, 13]. Molecular Transformer was the first adaptation of Transformer with SMILES [25] for the forward-reaction prediction task [14, 15]. Further studies demonstrated the ability to make general predictions using different compound databases, including drug-like molecules [16] and carbohydrate reactions [17], to examine regioselectivity and stereoselectivity. This success has paved the way for additional research on retrosynthesis using SMILES and Transformer [18–23].

SMILES strings are typical inputs for retrosynthetic predictors using NMT models. Despite its widespread usage, SMILES can easily lead to erroneous predictions. It is because the SMILES has fragile grammatical structure and is not suitable for tokenization. For this reason, SMILES-based prediction methods tend to make grammatically invalid predictions reducing the prediction efficiency. To solve this problem, SCROP [21] included a neural-network-based syntax corrector to decrease the invalidity rate. Similarly, Duan et al [19] focused on determining the causes of invalid SMILES to improve the prediction accuracy. In addition, grammatically valid SMILES are not guaranteed to be semantically valid or synthetically accessible. In our previous study [29], we demonstrated that representing molecules as the sets of fragments is an effective solution to the aforementioned problems.

Considering the complexity of retrosynthetic analysis, an efficient representation of source-target data structure is critical for accurate predictions. In this study, we show that representing molecules using sets of atom environments (AE) is an efficient alternative approach for devising a retrosynthetic prediction models to conventional SMILES-based approaches. AEs are topological fragments centered on an atom with a preset radius [36], defined by the number of shortest topological distances between atoms via covalent bonds. Unlike SMILES tokens, each AE is chemically meaningful and easily interpretable. NMT models are designed to translate between different pairs of tokens, whereas SMILES-to-SMILES translations require a model to learn the chemical change via rearrangements of regular expressions due to the conservation of atom types in an ideal reaction dataset. On the other hand, AEs in close vicinity of reaction center encapsulate the chemical change. The chemical change becomes observable in associated tokens, fragments, thus can be captured by the model.

Here we propose a direct translation approach for retrosynthetic prediction by associating the AEs of the reactants with the products. Throughout the study, these are regarded as the basis of molecules and employed in our prediction workflow. Our design enables us to capture the changes in molecules that are associated with reactions by focusing on fragments related to the reaction centers. To accurately generate the reactant candidates for a target molecule we use the Transformer architecture [13]. We show that our model achieves a top-1 accuracy of 55.4% for exact matches and 68.1% if bioactively similar predictions are included. These results are better than those of the existing methods, without suffering from problems associated with SMILES representation.

2. Method

2.1. **Model overview.** The main goal of the Transformer architecture is to generate the next word of a target sequence. Transformer uses an encoder unit and a decoder unit to translate between sequences by effectively employing a multi-head attention mechanism on each unit. Input and output sequences for our Transformer model are the lists of fragments. We tested several different schemes to convert molecules into a list of fragments, such as MACCS keys [55], bit vectors of extended circular fingerprint (ECFP) [54], and the atom environments (AEs) [36]. As presented in the next section, we identified that the AE representation resulted in the best performing model. AEs are fragments consisting of a central atom and its covalently bonded neighbors with a predefined radius. They can be considered the basis of constructing molecules, in a manner similar to the pieces of a puzzle. Each AE is described by a simplified molecular-input line-entry system arbitrary target specification (SMARTS) pattern [26].

An overview of our Transformer-based model, viz. RetroTRAE, is depicted in Figure 1. Starting from a product molecule, it is decomposed into a set of unique integer values. Each AE, a SMART pattern, is associated with a unique integer value. The lists of AEs were provided as input sequences for RetroTRAE. RetroTRAE is trained to predict the proper AE sequences of reactants corresponding to the true reactants.

2.2. **Atom Environments.** We employed the concept of circular atom environments to represent the molecules in the reaction dataset. Circular environments are defined as topological neighborhood fragments of varying 'radii' containing all bonds between the included atoms [36]. They are centered on a particular atom, called the central atom. The

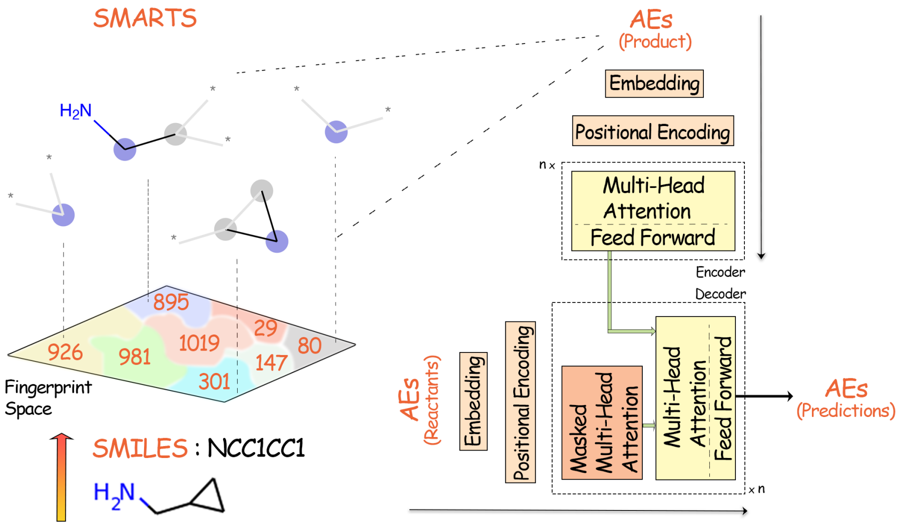
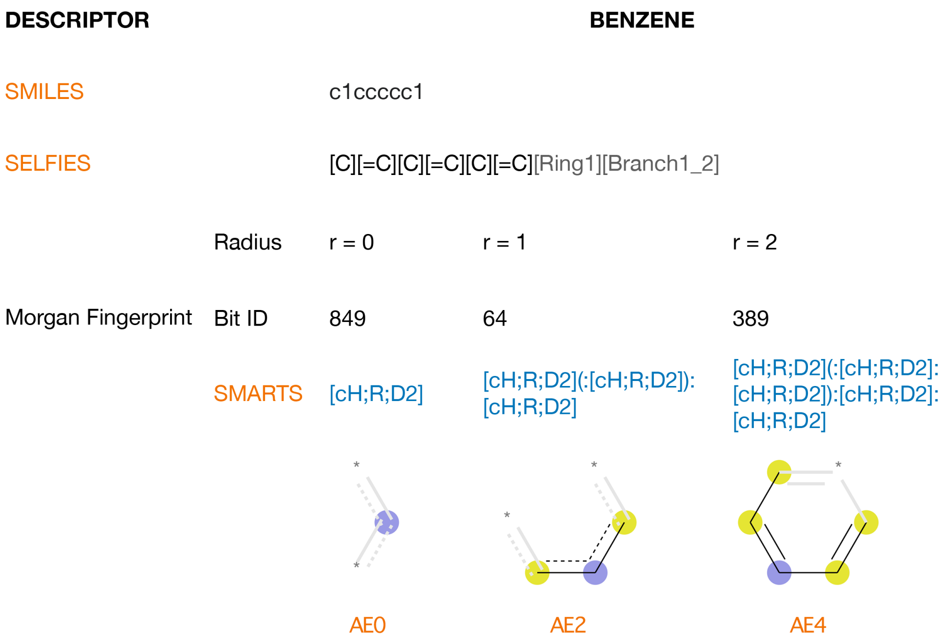


Figure 1. A schematic of RetroTRAE including the input-output structure.

'radius' refers to the maximum allowed topological distance between the central atom and all covalently bonded atoms. The topological distance between two atoms was measured as the number of bonds on the shortest path between them. Thus, an AE of radius "r" contains all the atoms in the molecule with a topological distance r or smaller from the central atom, and all bonds between them.

To construct the AEs, we used ECFPs of varying radii implemented in RDKit. We extracted all unique fragments that were folded into bits of ECFPs. AEs generated by the ECFP algorithm are invariant to rotation and translation and are easily interpretable as SMARTS patterns [32–34]. In Figure 2, the string representation of benzene is given as common SMILES and SMARTS patterns representing the atom environments generated by the ECFP fingerprint, along with the recently developed SELFIES [35] description. SMARTS and SELFIES are similar with respect to the level of information they display. The text sections of the SMARTS description contain two levels of detail: the first level concerns the aromaticity and H count of the element, and the second level includes the number of neighboring heavy atoms and ring information (represented by "D" and "R", respectively).

By definition, AEs with radius r = 0 only include the atoms of the central atom type. We denote the set of all AEs with r = 0 as AE0. AEs with r = 1 contain the central atom, all atoms adjacent to the central atom (nearest neighbors), and all the bonds between these atoms. The set of all AEs with r = 1 is denoted as AE2.



DESCRIPTOR		BENZENE		
SMILES		c1ccccc1		
SELFIES		[C][=C][C][=C][C][=C][Ring1][Branch1_2]		
	Radius	r = 0	r = 1	r = 2
Morgan Fingerprint	Bit ID	849	64	389
	SMARTS	[cH;R;D2]	[cH;R;D2](:[cH;R;D2]):[cH;R;D2]	[cH;R;D2](:[cH;R;D2]:[cH;R;D2]):[cH;R;D2]:[cH;R;D2]

FIGURE 2. String representations of benzene are represented in the form of SMILES, SELFIES and as a combination of SMARTS patterns generated by the Morgan fingerprint. In atom environment renderings, the central atom is highlighted in blue whereas aromatic and aliphatic ring atoms are highlighted in yellow and gray, respectively. A wildcard [*] is used to represent any atom.

We focused on two fragmentation schemes: AEs and ECFPs. A word-based tokenization scheme was applied to both AEs and the indices of the ECFP bit vectors. An ECFP bit vector corresponds to a one-hot encoded vector in the fingerprint space, such as a sentence, which is a one-hot encoded in vocabulary space. In this study, the following representations encoded as bit indices and SMARTS were attempted:

- AE2 and AE2, indicating atom environments of radius 0 and 1,
- ECFP0, ECFP2, and ECFP4 [37] corresponding to the Morgan fingerprints of radius 0, 1, and 2, hashed into a dimension of 1024.

AEs of radius 2 (AE4) result in millions of distinct fragments present in large datasets. Because of the vast vocabulary size of AE4, they are not suitable for translation purposes. Thus, only the hashed version of the Morgan fingerprint was selected for a radius of 2. The open-source RDKit module version 2020.03.1 was utilized to generate ECFPs and AEs.

2.3. **Dataset.** Neural machine translation methods require a large corpus of diverse source-target pairs for successful translation. To evaluate and compare our model with the current state-of-the-art models, we used a subset of the filtered US patent reaction dataset, USPTO-Full, which was obtained using a text-mining approach [27, 28]. This subset [5] contains 480K atom-mapped reactions after removing duplicates and erroneous reactions from USPTO-Full. To train our models, the atom-mapping information was not used. However, we implicitly benefitted from the fact that each atom in the product had a unique corresponding atom in the reactants. In addition, there was no reaction class information available in this dataset.

The product-reactant pairs were carefully curated in the same manner as in our previous study [29]. As a result, we generated two distinct curated datasets consisting of unimolecular ($P \Longrightarrow R$) and bimolecular ($P \Longrightarrow R_1 + R_2$) reactions, with sizes 100K and 314K respectively. Additionally, we used the PubChem compound database, which contains 111 million molecules, and the ChEMBL database, to recover molecules from a list of AEs and compare the space of AEs [30, 31].

2.4. **Training Details.** Our curated datasets were randomly split into a 9:1 ratio to generate the training and testing sets. The validation sets were randomly sampled from the training set (10%). We used the Adam algorithm [40] to train the model parameters in combination with a negative log-likelihood (NLL) loss function. For each dataset, we performed multiple tests within the range of the hyper-parameter space, as described in Supplementary Table 1, to achieve optimal performance. The best hyperparameters were chosen according to their performance on the validating set. With these hyperparameters, the average training speed was approx. 11 min per epoch with a batch size of 100.. We trained our models on average 1000 epochs with the learning rate scheduler stochastic gradient descent with warm restarts (SGDR) [39] and applied a residual dropout with a rate of 0.1 [38]. The details of our key hyperparameters are described in the Supplementary Information.

2.5. **Evaluation.** To evaluate the performance of our translation model, a suitable metric was required to measure the similarity between the predictions and the true reactants The Tanimoto (T_c) and the Søṙrensen-Dice coefficient (S) as two of the special cases of the Tversky index were the similarity metrics used in this study. The exact form of the Tversky index is as follows:

$$S(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|} \tag{1}$$

Here, $\alpha, \beta \geq 0$ are the parameters of the Tversky index. Setting $\alpha = \beta = 1$ leads to the Tanimoto coefficient; setting $\alpha = \beta = 0.5$ leads to the Sørrensen-Dice coefficient. The Tanimoto and Dice coefficients measured between two molecules range between 0 and 1. The value of zero represents the total dissimilarity, whereas a value of 1 represents the exact match. Pairwise similarities between the predicted and correct sequences are calculated at the end of each epoch for every pair present in the validation set using the chosen metrics.

Since there are many ways to decompose a molecule, retrosynthetic prediction tools can procure many different possible synthetic routes. However, the selection of an appropriate synthetic route is challenging. As a general rule, we used top-1 predictions as the best recommendations to report network performance, as well as for molecular search and retrieval. We used the ccbmlib Python package [47] to generate similarity value distributions of the fingerprints and assess the statistical significance of the Tanimoto coefficients. This implementation also allowed for a quantitative comparison of the similarity values between various fingerprint designs.

3. Results and Discussion

3.1. **Performance of RetroTRAE.** We evaluated the retrosynthetic predictor performance of the selected fingerprint variants to determine the best molecular structure encoding. We also compared the results of our Transformer models with those of the previously developed fragment-based retrosynthetic predictor (Table 1). The Transformer model representing molecules with the union of AE0 and AE2 outperformed all other models, achieving an exactly matching accuracy of 55.4%. The relationship between structural similarity and biological activity has been extensively investigated in systematic analyses [48–51]. Molecules found to have similar biological activities when their similarity is over 0.85. The addition of bioactively similar predictions ($T_c \geq 0.85$) increased the accuracy by 12.7% over the exact matches, resulting in an overall model accuracy of 68.1%. The model using

ECFP2 also performed well and showed slightly worse performance than using AEs. Hereafter, we refer to the model with the union of AE0 and AE2 as RetroTRAE.

Table 1. Performance summary of various Transformer-based models trained with different fragmentation schemes and a comparison with the Bi-LSTM-based models. Success rates (%) are given with respect to exact and bioactively similar matches ($T_c \geq .85$) and the mean Tanimoto coefficients of all predictions are listed.

Model	**Unimolecular dataset**		
	$T_c = 1.0$	$T_c \geq .85$	$\overline{T_c}$
Bi-LSTM-based [29]			
MACCS	29.9	57.7	0.84
ECFP2	35.6	50.7	0.80
ECFP4	9.1	28.4	0.66
Transformer-based			
MACCS	30.1	57.5	0.85
ECFP0	50.8	61.2	0.85
ECFP2	54.9	67.6	**0.88**
ECFP4	26.0	50.1	0.73
AE0	47.2	57.4	0.83
AE2	50.9	59.9	0.84
AE0 ∪ AE2	**55.4**	**68.1**	**0.88**

The Transformer-based models demonstrated significant improvements over the previous bi-LSTM-based method with respect to the exact match accuracy. This enhancement represented a substantial overall performance gain of 15-17%. However, when MACCS keys were used for fragmentation, the number of exact and bioactively similar matches were similar. This suggests that the combination of MACCS keys may have limited diversity, i.e., low resolution power. In contrast, AE2 describes the chemical space more precisely and provides 60 times higher resolution power than MACCS keys (Supporting Table 5).

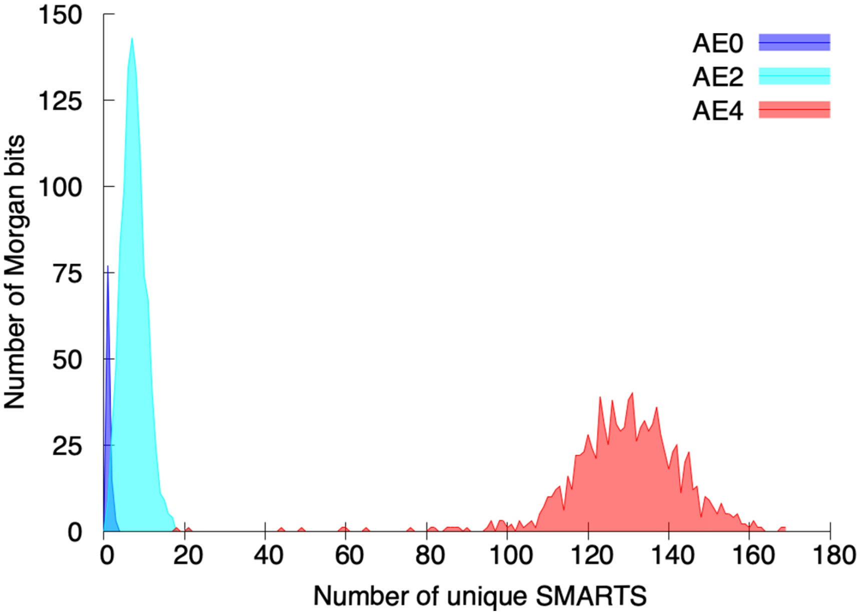


FIGURE 3. The histogram of Morgan bits according to the number of unique SMARTS patterns from AE0 (blue), AE2 (cyan), and AE4 (red).

Another interesting observation is the poor performance of ECFP4. The number of exact matches dropped to nearly half that of ECFP2. This poor performance may be due to a high collision rate of ECFP4 (Figure 3). We investigated the number of unique AEs of radii 0, 1, and 2 that were associated with the activated bits of hashed ECFPs for the unimolecular reaction dataset. With a radii of 0 and 1, each ECFP bit contained fewer than 10 and 20 unique AEs, respectively. However, with a radius of 2, most bits corresponded to many unique AEs, ranging from 100 to 160. In other words, ECFP4 has a much higher bit collision rate than ECFP2 or ECFP0. The presence of higher-density bits complicates the relationships between the fragments of a product and the true reactants, deteriorating the prediction power of the model.

Therefore, finding an optimal set of fragments representing a molecular structure most accurately is a critical factor in improving the predictive power of retrosynthesis planning.

TABLE 2. The accuracy (%) of single and double reactant predictions by using the union of AE0 and AE2.

Datasets	$T_c = 1.0$	SM	DM	$T_c \geq .85$	$T_c \geq .80$	$\overline{T_c}$	$\overline{S}$
Unimolecular	55.4	57.8	62.1	68.1	72.5	0.88	0.94
Bimolecular	61.9	62.7	64.6	67.8	69.7	0.77	0.87

Prediction performance, as a function of different similarity threshold values for the best performing model is shown in Table 2. By using AEs, we can select more reasonable thresholds that are size-dependent, similar to the similarity metrics. Single and double mutations represent changes in one and two fragments with respect to the ground truth. We refer to these as soft thresholds. For unimolecular reactions, the average reactant length is 27. The single and double fragment mutations corresponded to $T_c \geq 0.96$ and $T_c \geq 0.92$, respectively. The degree of similarity was different for bimolecular reactions because both reactants had an average length of 17. A detailed description of the similarity scale can be found in the Supporting Information for the soft thresholds as a function of the reactant fingerprint length (Supporting Table 6).

Soft thresholds present two clear advantages over hard thresholds, particularly when working with close analogs. First, soft thresholds allowed us to easily find the type and number of fragments that deviated from the ground truth. In contrast, classifications made by arbitrarily defined thresholds were difficult to comprehend. This is because there is no way to envision a molecule just by knowing the structure and the pairwise similarity value of a reference molecule in advance. Therefore, similarity maps were developed to provide better interpretation of the resulting similarity by visualizing the atomic contributions [53]. Second, by using soft thresholds, we avoided any risk of losing high-quality reactant candidates that could be excluded with hard thresholds. The idea of structural complexity was closely associated with the fingerprint length. This suggests that high-quality predictions with low and medium complexity had a higher chance of being excluded by hard thresholds. As such, a high-quality double mutated prediction with medium complexity represented with 13 atom environments could be overlooked by a commonly used bioactively similar threshold ($T_c \geq .85$)

The mean $\mathrm{T_c}$ of the predictions by the best-performing model was found to be 0.88, which is highly statistically significant with a p-value $< 10^{-5}$ (Table 2). Figure 7 shows the statistical significance of the selected similarity thresholds above which the quality of non-exact predictions is assessed in chemical terms. The inset of the figures shows the regime with $\mathrm{T_c}$ values having a p-value of 0.1, whereas our lowest similarity threshold value ($\mathrm{T}_c > 0.8$) had a p-value of 1e-04 or lower. Therefore, the predictions satisfying $\mathrm{T}_c > 0.8$ occur in the high similarity regime. The statistical equivalences between the similarity scores of each fingerprint type we used are shown in Figure 7C. The unified AEs and ECFP2 shared similar distribution profiles (Supporting Figures 7A and 7B). Hence, we found that they returned almost identical similarity values, as shown in Figure 7C. Landrum [52] showed that only 250 of the 25K pairs have a Tanimoto similarity value higher than 0.434 and 0.655 if computed with ECFP2 and MACCS keys respectively. Likewise, our lowest similarity threshold $\mathrm{T}_c > 0.8$ corresponded to $\mathrm{T}_c > 0.9$ if computed with MACCS keys.

3.2. Comparison with existing retrosynthesis planning methods.

Overall, Transformer-based models lead to better performances compared to non-Transformer models. Table 3 presents a performance comparison of our model with the available retrosynthesis models trained without reaction class information. Performance differences in the SMILES-based Transformer models can be attributed to improvements in data augmentation (with non-canonical SMILES) [22, 45], tokenization scheme (character or atom level) [18, 20], and postprocessing (by rectifying invalid SMILES) [19, 21]. The better predictive power of our model appears to be due to better reaction representation beyond the standard SMILES. For fair comparison, we compared with models that were trained and tested with large versions of the USPTO dataset, either filtered MIT-full or MIT-fully atom mapped reaction datasets.

Our approach achieved top-1 exact matching accuracies of 55.4% and 61.9% for unimolecular and bimolecular reactions without reaction class information, respectively (Table 2). In general, this level of accuracy was better than existing non-Transformer and Transformer models using SMILES. Lin's Transformer model using character level SMILES tokenization [20] was comparable to the performance of RetroTRAE. When bioactively similar predictions were considered, the overall accuracy of both datasets increased to 68%. This result surpassed all current state-of-the-art approaches by a large margin.

Table 3. Model performance comparison without additional reaction classes. The results are based on either filtered MIT-full or MIT-fully atom mapped reaction datasets.

Model	top-1 accuracy (%)
Non-Transformer	
Coley et al., Similarity, 2017 [42]	32.8
Segler et al., Neuralsym, 2017 [41]	35.8
Segler-Coley,–rep. by Lin, 2020 [20, 41]	47.8
Dai et al., GLN, 2019 [44]	39.3
Liu et al.–rep. by Lin, 2020 [20, 43]	46.9
Transformer-based	
Zheng et al., SCROP, 2020 [21]	41.5
Wang et al., RetroPrime, 2021 [45]	44.1
Tetko et al., AT, 2020 [22]	46.2
Lin et al., 2020 [20]	54.1
RetroTRAE – this work	55.4
RetroTRAE + Bioactive – this work	68.1

3.3. **Examples of high-quality predictions.** As we have stressed in our previous report [29], the similarity score can be considered an effective metric for assessing the retrosynthetic quality of predictions. High similarity scores indicate higher-quality retrosynthetic predictions. Thus, we included single and double fragment mutations, bio-active, and highly similar predictions as high-quality reactant candidates. Figure 4 shows a representative example for each category. These examples help us interpret non-exact, but high-quality, reactant candidates.

For single mutant cases, all the atom types were correct and the changes were often associated with misplacement of a single atom environment (e.g. at the ortho/para/meta position). For double mutant cases, most changes were also observed in *ortho/meta/para* substitution patterns, similar to the single mutation cases. In addition, the

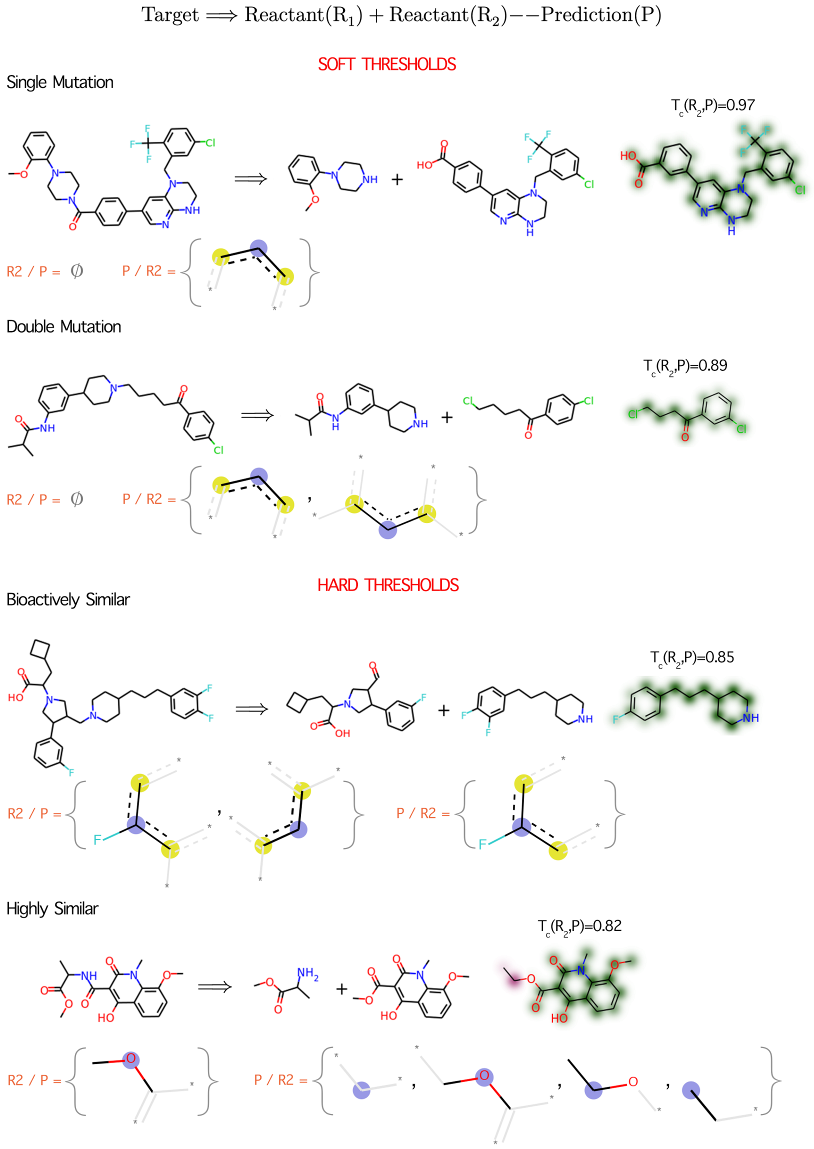


Figure 4. A representative example belonging to each threshold level is shown. Distinct fragments are given as SMARTS patterns. Predictions are drawn as similarity maps using the Morgan fingerprints. The first reactant is predicted correctly and the qualities of the second reactants are evaluated. The fragments only belonging to the prediction or its true counterpart are given as set notation differences, which allows us to describe the chemical change more concretely. Colors indicate atom-level contributions to the overall similarity (green: increases in similarity score, red: decreases in similarity score, uncolored: has no effect).

length of simple aliphatic chains is often incorrectly predicted because many fragments from a long aliphatic chain are identical. Thus, the length of an aliphatic chain cannot be accurately described using a set of unique fragments.

As indicated in the similarity maps, none of the atoms of the reactant candidates negatively contributed (red) to the similarity value (Figure. 4). After inspecting the bioactively similar predictions, we concluded that the most significant aspects of retrosynthetic analysis, such as bond disconnections, reactive functional groups, and core structures, were correctly predicted. In terms of hard thresholds, the number of altered atomic environments could be greater than two. However, they were mainly observed at the core structure, and do not affect the accuracy of the reactive sites. More reaction examples with high-quality predictions are provided in the Supporting Information.

3.4. **Covering the chemical space with atom environments.** Because AEs can be considered the basis of molecules, we investigated how many AEs are required to represent chemical space properly. We generated the AE0 and AE2 sets using all compounds in PubChem (111M), ChEMBL (2.08M), and the USPTO 500K (1.3M) dataset and visualized their diversity and coverage (Figure 5). Coverage was defined as the chemical space spanned by these unique atom environments. The area-proportional Euler graph (Figure 5) demonstrates that the AEs of the reactants in the USPTO dataset is not enough to describe diverse molecules and do not span a broad range of chemical space. This indicates that the current USPTO reaction dataset may not be large enough to train a truly general retrosynthesis predictors. We believe that our model would perform more accurately, if we have more diverse reaction datasets.

USPTO reaction dataset contains 275 (r = 0) and 15,982 (r = 1) unique AEs. ChEMBL and PubChem contain 386 (r = 0), 39,149 (r = 1) and 3450 (r = 0), 533,276 (r = 1) unique AEs, respectively. Although there are large differences in favor of PubChem, a significant portion of these unique AEs occurs only once in the whole set. In fact, many AEs from PubChem were found in only one compound record, which we refer to as singletons. The percentages of singletons were 38.5% and 35.2% for the AE0 and AE2 sets, respectively, generated from PubChem. The cardinality of each set of unique AEs was supplied as supporting information together with their intersections.

3.5. **Retrieving reactant candidates via atom environments.** After predictions are made by RetroTRAE, the chemical structures of the predicted reactants can be retrieved through a database search.

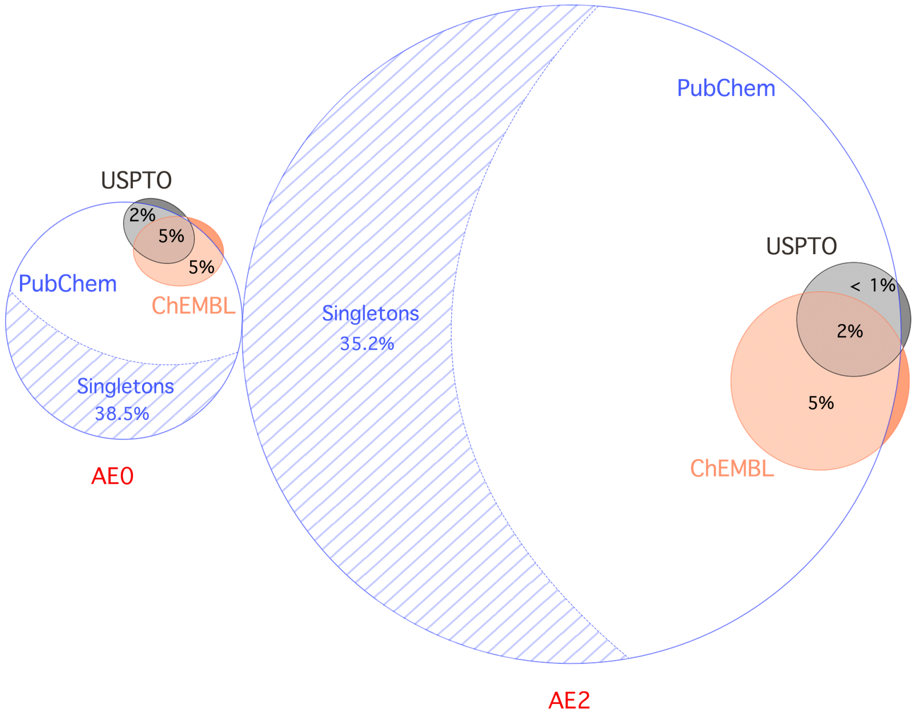


FIGURE 5. Area-proportional Euler graph representing the space of atomic environments for the following databases: PubChem 110M, ChEMBL 2.08M (ChEMBL v28, as of May 2021), and USPTO-Fully atom-mapped 500K reactions ($\sim 1.3M$ molecules). AE0 is upscaled by 20 times for better visual interpretation.

We investigated the success rate of retrieving a reactant candidate with 1000 USPTO test molecules using PubChem. The retrieval test results showed that more than half the predictions (55.7%) could be retrieved accurately (Figure 6). Allowing single mutations increased the retrieval rate by 30%. When double mutations were allowed, all the test molecules could be retrieved successfully. These results suggest that representing and predicting molecules with fragments is a viable and practical approach.

Using the top-1 predictions does not necessarily lead to a single synthetic route considering the degeneracy of the fragment representation. It is always possible to access multiple candidates during the process of converting fragments into valid molecules. This may correspond to multiple possible reaction pathways. Considering the small differences between molecules with high T_c values (Figure 4), multiple molecules generally have differences in stereochemistry, the length of their aliphatic chains, and the location of their peripheral functional groups, such as ortho/meta/para positions. Thus, such small differences can be easily corrected by experienced chemists.

Finally, it is worth mentioning that AEs are less degenerate, i.e., have fewer reactant candidates corresponding to a prediction, than ECFP fingerprints during the retrieval process. Using ECFP bit indices for database searches retrieve 1.7 times more reactant candidates on average. The difference is mainly due to bit collisions that occur during truncation to the bit vector and the absence of stereochemical information in our dataset.

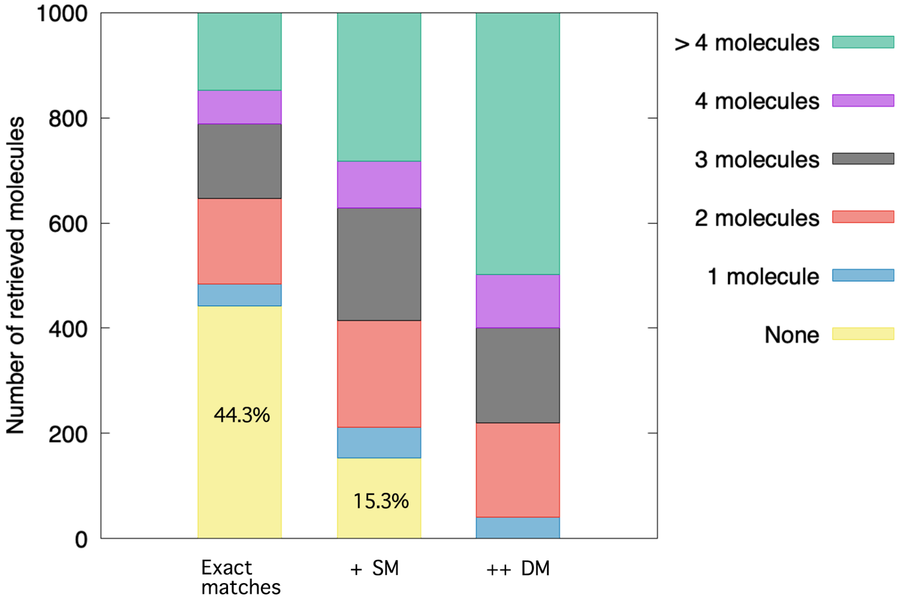


Figure 6. Retrieval of reactant candidates via a large PubChem compound search database. SM and DM represent single mutation and double mutations.

4. Conclusion

We developed a new template-free retrosynthesis prediction model, viz. RetroTRAE, using the Transformer architecture and atom environment representation. We demonstrated that AEs are promising descriptors for studying reaction route prediction and discovery because they provide a highly descriptive representation, free from the grammatical complexity of SMILES. RetroTRAE showed comparable or improved performance compared to other state-of-the-art models. We critically assessed the retrieval process that converts a set of fragments into a molecule with respect to coverage, degeneracy, and resolution. The present approach provided reactant candidates with an exact match accuracy of 55.4%. In addition to the exact match accuracy, high-quality reactant candidates selected by soft and hard thresholds were found to be statistically significant below the 1.0e-04 level. The average prediction accuracy with a threshold of $T_c \geq 0.85$ was $\sim 68\%$, outperforming the current state-of-the-art methods by a large margin. Our approach introduces a novel scheme for fragmental and topological descriptors to be used as natural inputs for retrosynthetic prediction tasks. We believe that our model will open new possibilities for the development of ML models not only for retrosynthetic prediction but also for reaction and property predictions.

5. Availability of data and materials

The datasets supporting the conclusions of this article are available via *https://github.com/knu-lcbc/Transformer_RetroTRAE* repository.

6. Supporting Information

Table 4. Hyper-parameter space and hyper-parameters for the best model.

Parameter	**Possible Values**	**Best Model Parameters**
Number of layers	2-8	4
Number of head	4-12	8
Size of hidden layers	256, 512, 1024	512
Size of intermediates	512, 1024, 2048	2048
Optimizer	Adam or SGD	Adam
Dropout	0.1, 0.2, 0.4	0.1
Learning rate	0.0001—0.01	0.001
Learning rate scheduler	Decay, SGDR	SGDR

Table 5

Representation	Sequence length		Vocabulary Size	
	Source	Target	Source	Target
MACCS	32.30	39.15	130	131
ECFP0	9.95	13.44	79	99
AE0	9.95	13.44	119	118
ECFP2	18.33	21.37	1025	1028
AE2	18.33	21.37	7533	8007
ECFP4	46.39	52.78	2052	2053

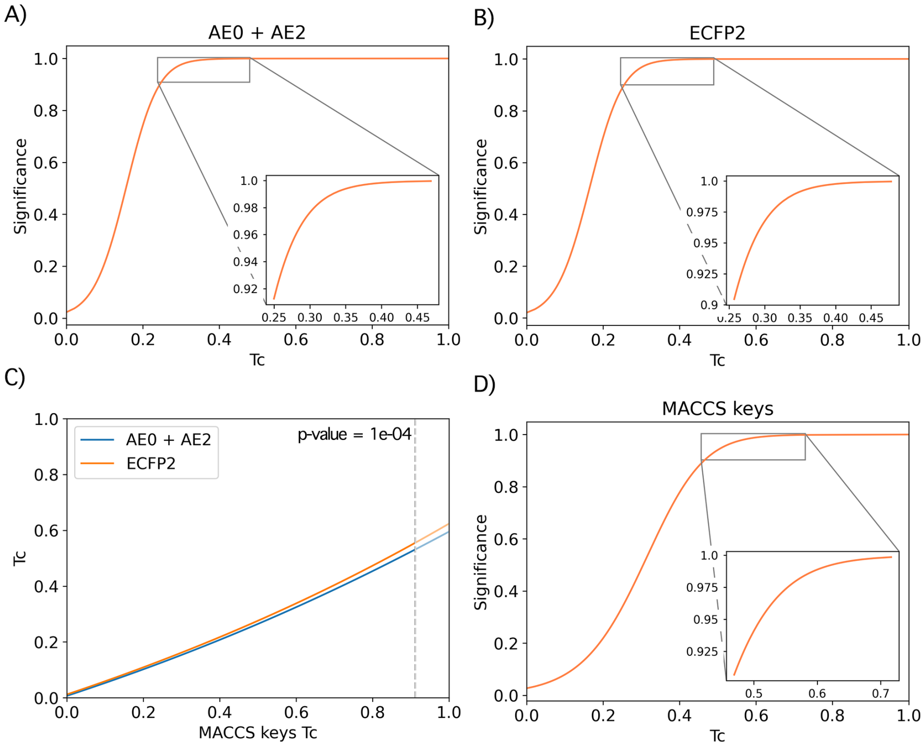


Figure 7. Figures A, B and D represent the cumulative distribution function of the reactants in the USPTO DB for the unified atom environments, ECFP2, and MACCS keys respectively. The measure 1 – (p-value) is used to assess significance. P-values has the range 0 to 1 and smaller p-values indicate higher significance. The Figure D shows the relation of MACCS Tc values to Tc values of unified atom environments and ECFP2. The vertical dashed line corresponds to a significance level of p-value set to 1e-04.

Table 6. The single and double mutant cases as a function of reactant fingerprint length

Length	5	8	11	14	17	20	23	26	29	32
T_c of SM	0.80	0.88	0.91	0.93	0.94	0.95	0.96	0.96	0.97	0.97
T_c of DM	0.60	0.75	0.82	0.86	0.88	0.90	0.91	0.92	0.93	0.94

```
Raw data of Figure 5.

USPTO-AE0 = 275,
ChEMBL-AE0 = 386,
PubChem-AE0 = 3450,
USPTO-AE0 ∩ ChEMBL-AE0 = 171,
USPTO-AE0 ∩ PubChem-AE0 = 250,
ChEMBL-AE0 ∩ PubChem-AE0 = 358,
USPTO-AE0 ∩ ChEMBL-AE0 ∩ PubChem-AE0 = 170,

USPTO-AE2 = 15982,
ChEMBL-AE2 = 39149,
PubChem-AE2 = 533276,
USPTO-AE2 ∩ ChEMBL-AE2 = 10251,
USPTO-AE2 ∩ PubChem-AE2 = 15224,
ChEMBL-AE2 ∩ PubChem-AE2 = 37725,
USPTO-AE2 ∩ ChEMBL-AE2 ∩ PubChem-AE2 = 10232,
```

References

[1] E. J. Corey, *Robert Robinson lecture. Retrosynthetic thinking - Essentials and examples*, Vol. 17, 1988.

[2] E.J. Corey and X.M Cheng, *The Logic of Chemical Synthesis*, Wiley, 1989.

[3] Elias James Corey, *The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture)*, Angew. Chem. Int. Edit. **30** (1991), no. 5, 455–465, DOI 10.1002/anie.199104553.

[4] E. J. and Todd Wipke Corey W., *Computer-assisted design of complex organic syntheses*, Science **166** (1969), no. 3902, 178–192, DOI 10.1126/science.166.3902.178.

[5] Wengong and Coley Jin Connor W. and Barzilay, *Predicting organic reaction outcomes with weisfeiler-lehman network*, Adv. Neur. In. **2017-Decem** (2017), no. Nips, 2608–2617, available at `1709.04555`.

[6] Vignesh Ram and Bunne Somnath Charlotte and Coley, *Learning Graph Models for Retrosynthesis Prediction* (2020), 1–15 pp., available at `2006.07038`.

[7] Chence and Xu Shi Minkai and Guo, *A graph to graphs framework for retrosynthesis prediction*, 37th International Conference on Machine Learning, ICML 2020 **PartF168147-12** (2020), 8777–8786, available at `2003.12725`.

[8] Chaochao and Ding Yan Qianggang and Zhao, *RetroXpert: Decompose Retrosynthesis Prediction like a Chemist*, posted on 2020, DOI 10.26434/chemrxiv.11869692, available at `2011.02893`.

[9] Ilya and Vinyals Sutskever Oriol and Le, *Sequence to sequence learning with neural networks*, Advances in Neural Information Processing Systems **4** (2014), no. January, 3104–3112, available at `1409.3215`.

[10] Juno and Kim Nam Jurae, *Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions* (2016), 1–19 pp., available at `1612.09529`.

[11] Philippe and Gaudin Schwaller Théophile and Lányi, *"Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models*, Chem. Sci. **9** (2018), no. 28, 6091–6098, DOI 10.1039/c8sc02339e, available at `1711.04810`.

[12] Dzmitry and Cho Bahdanau Kyung Hyun and Bengio, *Neural machine translation by jointly learning to align and translate*, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015), 1–15, available at `1409.0473`.

[13] Ashish and Shazeer Vaswani Noam and Parmar, *Attention is all you need*, Adv. Neur. In. **2017-Decem** (2017), no. Nips, 5999–6009, available at `1706.03762`.

[14] Philippe and Laino Schwaller Teodoro and Gaudin, *Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction*, ACS Central Science **5** (2019), no. 9, 1572–1583, DOI 10.1021/acscentsci.9b00576, available at `1811.02633`.

[15] Philippe and Petraglia Schwaller Riccardo and Zullo, *Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy*, Chemical Science **11** (2020), no. 12, 3316–3325, DOI 10.1039/c9sc05704h.

[16] Alpha A. and Yang Lee Qingyi and Sresht, *Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space*, Chemical Communications **55** (2019), no. 81, 12152–12155, DOI 10.1039/c9cc05122h.

[17] Giorgio and Schwaller Pesciullesi Philippe and Laino, *Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates*, Nature Communications **11** (2020), no. 1, 1–8, DOI 10.1038/s41467-020-18671-7.

[18] Pavel and Godin Karpov Guillaume and Tetko, *A Transformer Model for Retrosynthesis*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11731 LNCS** (2019), no. 1, 817–830.

[19] Hongliang and Wang Duan Ling and Zhang, *Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions*, RSC Advances **10** (2020), no. 3, 1371–1378, DOI 10.1039/c9ra08535a.

[20] Kangjie and Xu Lin Youjun and Pei, *Automatic retrosynthetic route planning using template-free models*, Chemical Science **11** (2020), no. 12, 3355–3364, DOI 10.1039/c9sc03666k.

[21] Shuangjia and Rao Zheng Jiahua and Zhang, *Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks*, Journal of Chemical Information and Modeling **60** (2020), no. 1, 47–55, DOI 10.1021/acs.jcim.9b00949.

[22] Igor V. and Karpov Tetko Pavel and Van Deursen, *State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis*, Nature Communications **11** (2020), no. 1, 1–11, DOI 10.1038/s41467-020-19266-y, available at `2003.02804`.

[23] Eunji and Lee Kim Dongseon and Kwon, *Valid, Plausible, and Diverse Retrosynthesis Using Tied Two-Way Transformers with Latent Variables*, Journal of Chemical Information and Modeling **61** (2021), no. 1, 123–133, DOI 10.1021/acs.jcim.0c01074.

[24] David and Hahn Rogers Mathew, *Extended-Connectivity Fingerprints*, J. Chem. Inf. Model. **50** (2010), no. 5, 742–754, DOI 10.1021/ci100050t.

[25] David Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comp. Sci. **28** (1988), no. 1, 31–36, DOI 10.1021/ci00057a005.

[26] Daylight Chemical Information Systems Inc., *Daylight Theory Manual, Chapter 4: SMARTS—A Language for Describing Molecular Patterns.*, `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`. Accessed January 2021.

[27] D. M. Lowe, *Extraction of chemical structures and reactions from the literature*, University of Cambridge, 2012.

[28] Daniel Lowe, *Chemical reactions from US patents (1976-Sep2016)*, posted on 2017, DOI 10.6084/m9.figshare.5104873.v1.

[29] Umit V. and Kang Ucak Taek and Ko, *Substructure-based neural machine translation for retrosynthetic prediction*, Journal of Cheminformatics **13** (2021), no. 1, 1–15, DOI 10.1186/s13321-020-00482-z.

[30] Evan E. and Wang Bolton Yanli and Thiessen, *Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities*, Vol. 4, Elsevier B.V., 2008.

[31] *The ChEMBL database in 2017*, Nucleic Acids Research **45** (2017), no. D1, D945–D954, DOI 10.1093/nar/gkw1074.

[32] Greg Landrum, *RDKit: Open-Source Cheminformatics Software*, 2016.

[33] Universität Hamburg. Center for Bioinformatics, *SMARTSviewer* (2010), `http://smartsview.zbh.uni-hamburg.de/`. Accessed Februrary 2021.

[34] Karen and Ehrlich Schomburg Hans Christian and Stierand, *Chemical pattern visualization in 2D - The SMARTSviewer*, Journal of Cheminformatics **3** (2011), no. SUPPL. 1, 2–3, DOI 10.1186/1758-2946-3-S1-O12.

[35] Mario and Häse Krenn Florian and Nigam, *Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation*, Machine Learning: Science and Technology **1** (2020), no. 4, 045024, DOI 10.1088/2632-2153/aba947, available at `1905.13741`.

[36] Volker D. and Bolton Hähnke Evan E. and Bryant, *PubChem atom environments*, Journal of Cheminformatics **7** (2015), no. 1, 1–37, DOI 10.1186/s13321-015-0076-4.

[37] David and Hahn Rogers Mathew, *Extended-Connectivity Fingerprints*, Journal of Chemical Information and Modeling **50** (2010), no. 5, 742–754, DOI 10.1021/ci100050t.

[38] Nitish and Hinton Srivastava Geoffrey and Krizhevsky, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, J. Mach. Learn. Res. **15** (2014), no. 1, 30.

[39] Ilya and Hutter Loshchilov Frank, *SGDR: Stochastic gradient descent with warm restarts*, 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2017), 1–16, available at `1608.03983`.

[40] Diederik P. and Ba Kingma Jimmy Lei, *Adam: A method for stochastic optimization*, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015), 1–15, available at `1412.6980`.

[41] Marwin H.S. and Waller Segler Mark P., *Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction*, Chemistry - A European Journal **23** (2017), no. 25, 5966–5971, DOI 10.1002/chem.201605499.

[42] Connor W. and Rogers Coley Luke and Green, *Computer-Assisted Retrosynthesis Based on Molecular Similarity*, ACS Central Science **3** (2017), no. 12, 1237–1245, DOI 10.1021/acscentsci.7b00355.

[43] Bowen and Ramsundar Liu Bharath and Kawthekar, *Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models*, ACS Central Science **3** (2017), no. 10, 1103–1113, DOI 10.1021/acscentsci.7b00303, available at `1706.01643`.

[44] Hanjun and Li Dai Chengtao and Coley, *Retrosynthesis prediction with conditional graph logic network*, Advances in Neural Information Processing Systems **32** (2019), no. NeurIPS, 1–11, available at `2001.01408`.

[45] Xiaorui and Qiu Wang Jiezhong and Li, *RetroPrime : A Chemistry-Inspired and Transformer-based Method for Retro- synthesis Predictions.*

[46] Vipul and Venkatasubramanian Mann Venkat, *Retrosynthesis Prediction using Grammar-based Neural Machine Translation : An Information-Theoretic Approach.*

[47] Martin and Bajorath Vogt Jürgen, *Ccbmlib - A python package for modeling tanimoto similarity value distributions*, F1000Research **9** (2020), DOI 10.12688/f1000research.22292.1.

[48] Yvonne C. and Kofron Martin James L. and Traphagen, *Do structurally similar molecules have similar biological activity?*, Journal of Medicinal Chemistry **45** (2002), no. 19, 4350–4358, DOI 10.1021/jm020155c.

[49] Steven W. and Debe Muchmore Derek A. and Metz, *Application of belief theory to similarity data fusion for use in analog searching and lead hopping*, Journal of Chemical Information and Modeling **48** (2008), no. 5, 941–948, DOI 10.1021/ci7004498.

[50] Jürgen and Jasial Bajorath Swarit and Hu, *Activity-relevant similarity values for fingerprints and implications for similarity searching*, F1000Research **5** (2016), no. 0, DOI 10.12688/f1000research.8357.1.

[51] Mathias and Günther Dunkel Stefan and Ahmed, *SuperPred: drug classification and target prediction.*, Nucleic acids research **36** (2008), no. Web Server issue, 55–59, DOI 10.1093/nar/gkn307.

[52] Greg Landrum, *Thresholds for "random" in fingerprints the RDKit supports* (2021), `https://greglandrum.github.io/rdkit-blog/fingerprints/similarity/reference/2021/05/18/fingerprint-thresholds1.html`. Accessed May 2021.

[53] Sereina and Landrum Riniker Gregory A., *Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods*, Journal of Cheminformatics **5** (2013), no. 9, 1–7, DOI 10.1186/1758-2946-5-43.

[54] David and Hahn Rogers Mathew, *Extended-Connectivity Fingerprints*, J. Chem. Inf. Model. **50** (2010), no. 5, 742–754, DOI 10.1021/ci100050t.

[55] Joseph L and Leland Durant Burton A and Henry, *Reoptimization of MDL Keys for Use in Drug Discovery*, J. Chem. Inf. Comp. Sci. **42** (2002), no. 6, 1273–1280, DOI 10.1021/ci010132r.