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 3, 2021.

ABSTRACT. We present the new retrosynthesis prediction method RetroTRAE using fragment-based tokenization combined with the Transformer architecture. RetroTRAE represents chemical reactions by using the changes of fragment sets of molecules using the atomic environment fragmentation scheme. Atom environments stand as an ideal, chemically meaningful building blocks together producing 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 conservation of atoms in reactions. Our model achieved a top-1 accuracy of 67.1% within the bioactively similar range for USPTO test dataset, outperforming the other state of the art, translation methods. We investigated the effect of different encoding scenarios on predicting the reactant candidates. We also critically assessed the retrieval process that converts a set of fragments into a molecule with respect to coverage, degeneracy and resolution. Our new template-free model for retrosynthetic prediction provides fast and reliable retrosynthetic route planning for substances whose fragmentation patterns are revealed.

1. INTRODUCTION

Planning the reaction pathways of an organic molecule is the 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]. 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, forward- and backward-reaction pathway predictors in the literature can be divided into those that rely on the construction of reaction templates, and those that 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. The template-free methods can be further subdivided according to the way of molecular representation: (i) graph-based methods [5–8] and (ii) sequence-based methods [9–11, 43]. Sequence-based modeling recasts the reaction pathway planning problem as a language translation problem by using a string representations of molecules. The current state of the art, forward- and backward-reaction predictors are mostly built on the Transformer architecture [13]. The Transformer, developed as a result of a collaborative effort, is a neural machine translation (NMT) model which 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 of making general predictions using different compound databases including drug-like molecules [16] and carbohydrate reactions [17] to examine regio- and stereoselectivity. This success has paved the way for additional publications on retrosynthesis using SMILES [18–23].

SMILES strings are typical inputs of NMT models. Despite its widespread usages, SMILES can easily lead to erroneous predictions due to its grammatical complexity. In other words, SMILES-based prediction methods tend to make grammatically invalid predictions, which deteriorate prediction efficiency. To solve this problem, SCROP [21] included a neural network-based syntax correcter to decrease the invalidity rate. Similarly, Duan et al [19] focused 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, efficient representation of source-target data structure is critical for accurate predictions. In this study, we propose a direct translation approach for retrosynthetic prediction by associating atomic environments of the reactants with the products. Atom environments are topological fragments centered with an atom with a preset radius [36]. The radius is defined by the number of shortest topological distance between atoms via covalent bonds, i.e., the smallest number of covalent bonds. Throughout the study, they are regarded as the basis of molecules and employed in our prediction workflow. Our design enables us to capture the changes of atom environments associated with the chemical reaction. To accurately generate the reactant candidates for a target molecule we use the best performing Transformer architecture as the state of the art, in NMT applications. We show that our model achieves top-1 accuracy of 53.4% for exact matches and 67.1% if bioactively similar predictions are included. These results are comparable to or better than the existing methods without suffering from problems associated with the complex grammar of SMILES.

2. Method

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

An overview of our Transformer model, namely 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 an unique integer value. The lists of AEs are provided as input sequences for RetroTRAE. RetroTRAE is trained to predict proper AE sequences of reactants corresponding to true reactants.

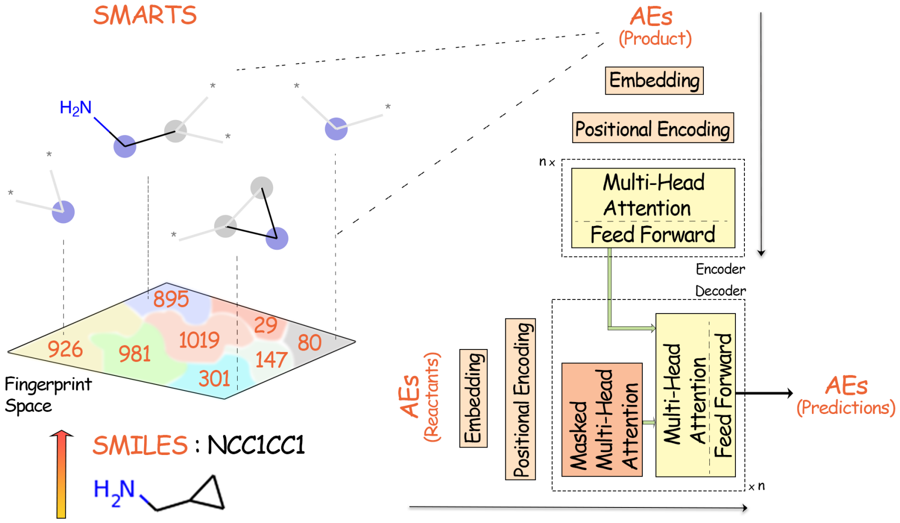


FIGURE 1. A schematic diagram of RetroTRAE including input-output structure.

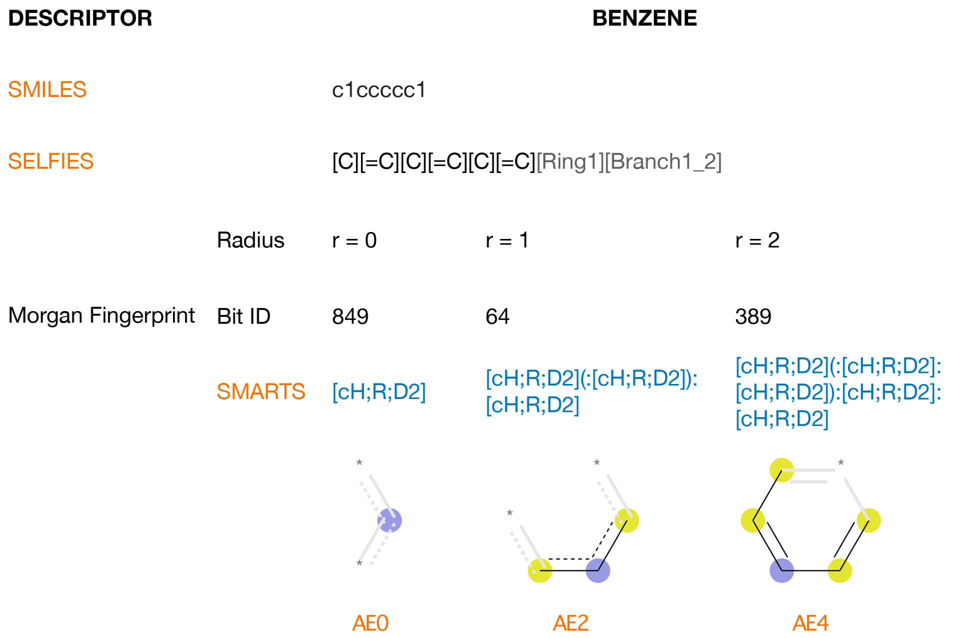
2.2. **Atom Environments.** We employed the concept of circular atom environments to represent molecules in the reaction dataset. Circular environments are defined as topological neighborhood fragments of varying 'radii' containing all bonds between included atoms [36]. They are centered on a particular atom, called the central atom. The 'radius' refers to the maximum allowed topological distance between the center atom and all covalently bonded atoms. The topological distance between two atoms is measured as the number of bonds on the shortest path between them. Thus, an atom environment of radius "r" contains all atoms in the molecule with a topological distance r or smaller to the center 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 are 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] as shown in Figure 2. For example, AEs with radius r = 0 include only the atom type of the center atom. We call the set of all AEs with r = 0 AE0. Atom environments with r = 1 contain the center atom, all atoms adjacent to the center atom (nearest neighbors), and all the bonds between these atoms. The set of all AEs with r = 1 is denoted as AE2.

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 closely resemble regarding the level of information they display. The text parts of the SMARTS description contains two levels of detail. The first detail is about the aromaticity and the H count of the element. The second level of detail includes the number of neighboring heavy atoms and ring information (represented by the "D" and "R" respectively). By definition, an environment of radius 0 corresponds to a single atomic environment, the radius 1 bits all have at least three atoms, and the radius 2 bits each have at least 5 atoms.

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

- AE2 and AE2, indicating atom environments of radius 0 and 1,



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]
		AE0	AE2	AE4

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, aromatic and aliphatic ring atoms are highlighted in yellow and gray. A wildcard [*] is used for any atom.

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

Atom environments of radius 2 (AE4) results in millions of distinct fragments present in large data sets. Due to a vast vocabulary size of AE4, they are not suitable for translation purposes. Thus, only hashed version of the Morgan fingerprint is selected for radius 2. The open-source RDKit module version 2020.03.1 is 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, we used a subset of the filtered US patent reaction dataset, USPTO-Full, which is obtained with a text-mining approach [27,28]. This subset [5] contains 480K atom-mapped reactions after removing duplicates and erroneous reactions from USPTO-Full. For training our models, atom-mapping information was not used. However, we implicitly benefit from the fact that each atom in the product has an unique corresponding atom in the reactants. Also, there are no reaction class information is available in this dataset. The product-reactant pairs are carefully curated in the same manner with our previous study [29]. As a result, we generated two distinct curated datasets consist of unimolecular ($\mathrm{P} \Longrightarrow \mathrm{R}$) and bimolecular ($\mathrm{P} \Longrightarrow \mathrm{R}_1 + \mathrm{R}_2$) reactions, with sizes 100K and 314K respectively. Additionally, we used the PubChem compound database containing 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 9:1 to generate training and testing sets. The validation sets were randomly sampled from training sets (10%). We used the stochastic gradient descent algorithm [40] to train model parameters in combination with negative log-likelihood (NLL) loss function. For each dataset, we performed multiple tests within the range of hyper-parameter space as described in the Supplementary Table 1 to achieve optimal performance. The best hyperparameters are chosen according to the performances on the validating set. With these hyperparameters, the average training speed was approximately 11 min per epoch corresponding to 1000 steps for single reactant dataset. We trained our models for a minimum of 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 hyper-parameters are described in the Supplementary Information.

2.5. **Evaluation.** To evaluate the performance of our translation model, a suitable similarity metric needs to be selected to measure the similarity between predictions and true reactants. The Tanimoto (T_c) and the Sørrensen-Dice coefficient (S) as two of the special cases of Tversky index are the selected metrics for the purpose this study. The exact form of the Tversky Index is given below:

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

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. Tanimoto and Dice coefficients measured between two molecules range between 0 and 1. The value of zero represents the total dissimilarity while 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.

As there are many ways to decompose a molecule, retrosynthetic prediction tools are able to procure large number of possible synthetic routes. However, selecting an appropriate synthetic route is challenging. As a rule of thumb, we used top-1 predictions as the best recommendations to report network performance, as well as of 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 Tanimoto coefficients. This implementation also allows a quantitative comparison of 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 find the best molecular structure encoding. We compared the results of our Transformer models with the previously developed substructure-based retrosynthetic predictor as presented in Table 1. The Transformer model representing molecules with the union of AE0 and AE2 outperformed all other models, achieving an exactly matching accuracy of 53.4%. The relationship between structural similarity and biological activity has been extensively investigated in systematic analyses [48–51]. Molecules are found to have similar biological activities when their similarity is over 0.85. The addition of bioactively similar predictions ($T_c \geq 0.85$) increases the accuracy by 13.7% over the exact matches, resulting in 67.1% overall model accuracy. Using ECFP2 also performed well and showed slightly worse performance than using AEs. From now on, we refer to the model with the union of AE0 and AE2 as RetroTRAE.

The Transformer-based models show marked improvements over the previous bi-LSTM-based method regarding the exact match accuracy. This enhancement represents a substantial overall performance gain by 15-17%. However, when MACCS keys are used for fragmentation, the number of exact and bioactively similar matches are found to be 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).

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 in terms of the exact and bioactively similar matches ($T_c \geq .85$) and the mean Taniomoto coefficient 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	52.9	66.6	**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	**53.4**	**67.1**	**0.88**

Another interesting observation is the low performance of ECFP4. The number of exact matches is dropped by nearly half 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 radius 0, 1 and 2 associated with the activated bits of hashed ECFPs for the unimolecular reaction dataset. With a radius of 0 and 1, each ECFP bit contains less than 10 and 20 unique AEs. However, with a radius of 2, most bits correspond to many unique AEs ranging from 100 to

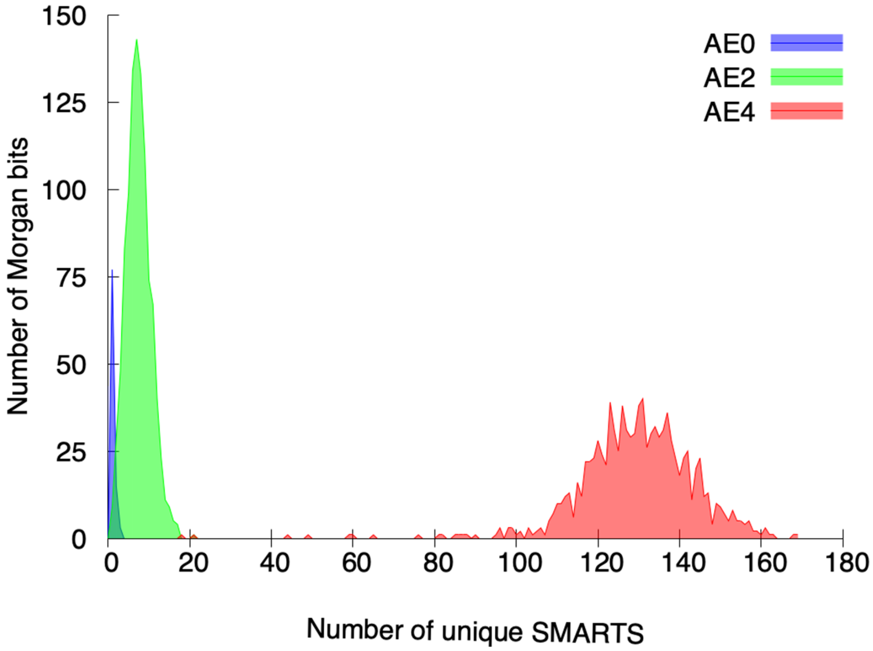


FIGURE 3. The histogram of Morgan bits according to the number of unique SMARTS patterns from AE0 (blue), AE2 (green), and AE4 (red).

160. In other words, ECFP4 has a much higher bit collision rate than ECFP2 or ECFP0. The presence of higher density bits would complicate the relationships between the fragments of a product and true reactants, deteriorating the prediction power of a model. Therefore, finding an optimal set of fragments representing a molecular structure most accurately is a critical factor in improving predictive power for 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	53.4	55.8	60.1	67.1	72.5	0.88	0.94
Bimolecular	61.9	62.7	64.6	67.2	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 to one and two fragments with respect to ground truth. We call these soft thresholds. For unimolecular reactions, the average reactant length is 27. The single and double fragment mutations correspond to $T_c \geq 0.96$ and $T_c \geq 0.92$. The degree of similarity is different for bimolecular reactions, because both reactants have an average length of 17. Detailed description of the similarity scale can be found in Supporting Information for the soft thresholds as a function of reactant fingerprint length (see Supporting Table 6).

Soft thresholds present two clear advantages over hard thresholds particularly when working with close analogs. First, soft thresholds allow us to find the type and number of the fragments deviated from the ground truth easily. In contrast, classifications done by arbitrarily defined thresholds are difficult to grasp intuitively. Because, there is simply no way to envision a molecule knowing in advance the structure and the pair-wise similarity value of a reference molecule. For this reason, similarity maps are developed for better interpretation of the resulting similarity by visualizing the atomic contributions [53]. Second, by using soft thresholds, we avoid any risk of losing high-quality reactant candidates which can be excluded with hard thresholds. The idea of structural complexity is closely associated with the fingerprint length. This suggest that the high-quality predictions with low and medium complexity have a higher chance to be excluded by hard thresholds. As such, a high-quality double mutated prediction with medium complexity represented with 13 atom environments can be overlooked by commonly used bioactively similar threshold ($T_c \geq .85$)

3.2. **Comparison with existing retrosynthesis planning methods.** Table 3 presents a performance comparison of our model with available retrosynthesis models trained without reaction class information. For a fair comparison, we compared with the models that were trained and tested with the MIT-full, large versions of USPTO dataset. Our approach achieves a top-1 exact matching accuracy of 53.4% and 61.9% for unimolecular and bimolecular reactions without reaction class information (Table 2). In general, This level of accuracy is better than most existing non-Transformer and Transformer models. The performance of RetroTRAE is comparable to the best of existing methods, namely Lin's Transformer model [20]. When bioactively similar predictions are considered, the overall accuracy of both datasets increases to 67.1%. This result surpasses all the 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	53.4
RetroTRAE + Bioactive – this work	67.1

The mean Tc of the predictions by the best-performing model is found 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 assessed in chemical terms. While the inset of the figures show the regime with Tc values having a p-value of 0.1, our lowest similarity threshold value ($T_c > 0.8$) has a p-value of 1e-04 or lower. Therefore, the predictions satisfying $T_c > 0.8$ are said to occur in high similarity regime. The statistical equivalence between similarity scores of each fingerprint type we used are shown in Figure 7C. The unified atom environments and ECFP2 share the similar distribution profiles (See

Supporting Figure 7A and 7B). Hence we find that they return almost identical similarity values as presented in Figure 7C. The vertical dashed line corresponds to a p-value of 1e-04. Landrum [52] showed that only 250 of the 25K pairs have a Taniomoto similarity value higher than 0.434 and 0.655 if computed with ECFP2 and MACCS keys respectively. Likewise, our lowest similarity threshold $T_c > 0.8$ corresponds to $T_c > 0.9$ computed with MACCS keys.

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

For single mutant cases, the changes were often associated with misplacement of functional groups at *ortho/meta/para* positions. For double mutant cases, most changes were also observed in *ortho/meta/para* substitution patterns, similar to the single mutation cases. In addition, the length of simple aliphatic chains is often predicted incorrectly because many fragments from a long aliphatic chain are identical. Thus, the length of a aliphatic chain may not be described accurately with the set of unique fragments. As indicated in similarity maps, none of the atoms of the reactant candidates negatively contributes (red) to the similarity value. 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 are correctly predicted. When we utilize hard thresholds the number of altered atomic environments can be more than two. However, they are mainly observed at the core structure, and not affecting the accuracy of reactive sites. More reaction examples with high quality predictions are shown in the Supporting Information.

$$\text{Target} \Longrightarrow \text{Reactant}(\text{R}_1) + \text{Reactant}(\text{R}_2) - -\text{Prediction}(\text{P})$$

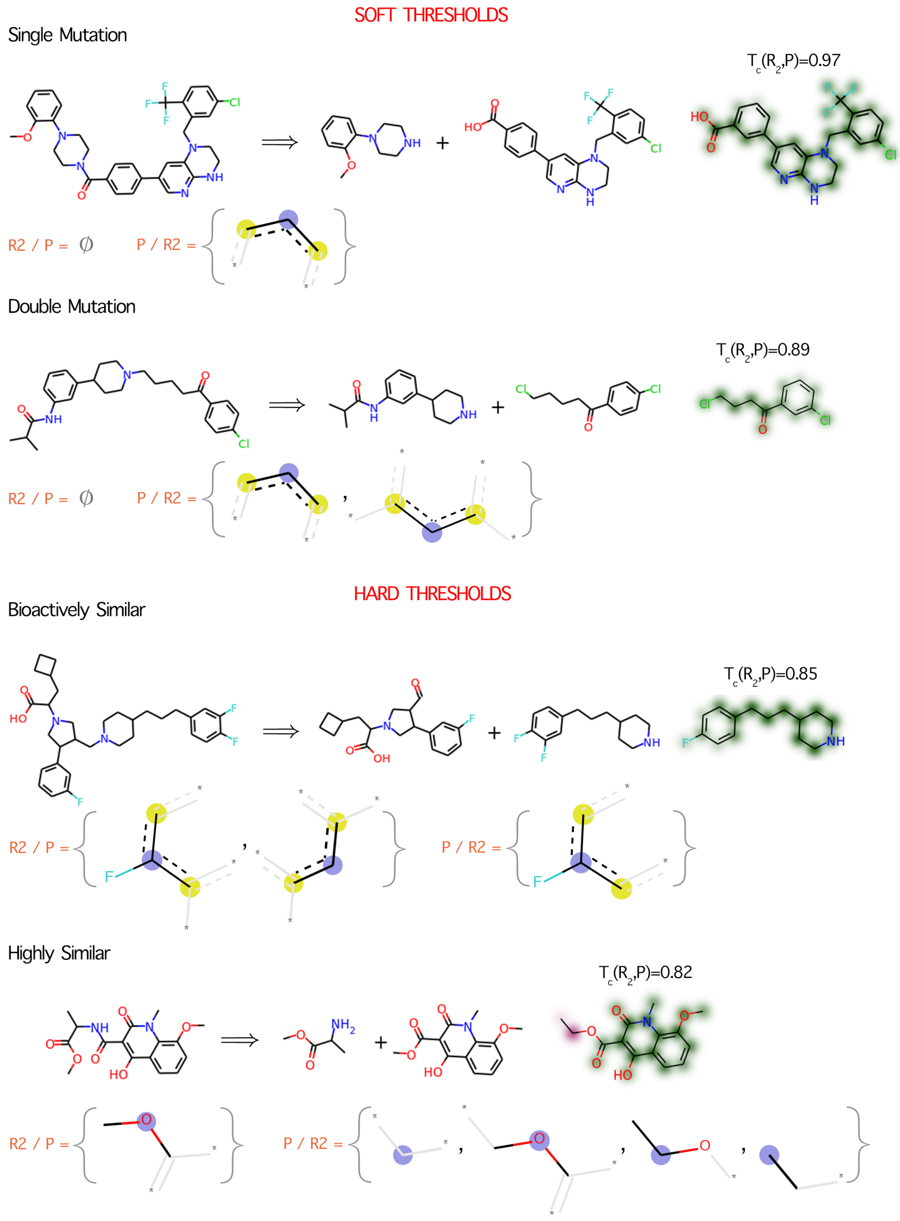


Figure 4. A representative example belong to each threshold level is shown. Distinct fragments are given as SMARTS patterns. Predictions are drawn as similarity maps using the Morgan fingerprints. The first reactant is predicted correctly and the qualities of second reactants are evaluated. The fragments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

3.4. **Covering chemical space with atom environments.** As previously mentioned, AEs can be considered as the basis of molecules. 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 the coverage (Figure 5). The coverage is defined as the chemical space spanned by those unique atom environments. From the area-proportional Euler graph (Figure 5), it is clear that the AEs of the reactants of the USPTO dataset do not span a broad range of chemical space. 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 those unique AEs occurs only once in the whole set. In fact, many AEs from PubChem are found in only one compound record and we call them singletons. The percentages of singletons are 38.5% and 35.2% for the AE0 and AE2 sets generated from PubChem. The cardinality of each set of unique AEs is supplied as supporting information together with their intersections.

3.5. **Retrieving a molecule from atom environments.** After predictions are made by RetroTRAE, the chemical structures of predicted reactants can be retrieved through database search. We investigated the success rate of retrieving a reactant candidate with 1000 USPTO test molecules using PubChem. The retrieval test result shows that more than half of the predictions (55.7% of them) can be retrieved accurately (Figure 6). Allowing single mutations increases retrieval rate by 30 percent. When double mutations are allowed, all 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 in the process of converting fragments into valid molecules. This may correspond to multiple possible reaction pathways. Considering small differences between molecules with high T_c values (Figure 4), multiple molecules generally have differences in stereochemistry, the length of aliphatic chains, and the location of peripheral functional groups, such as ortho/meta/-para positions. Thus, such small differences can easily be corrected by experienced chemists. Last, it is worth mentioning that AEs are less

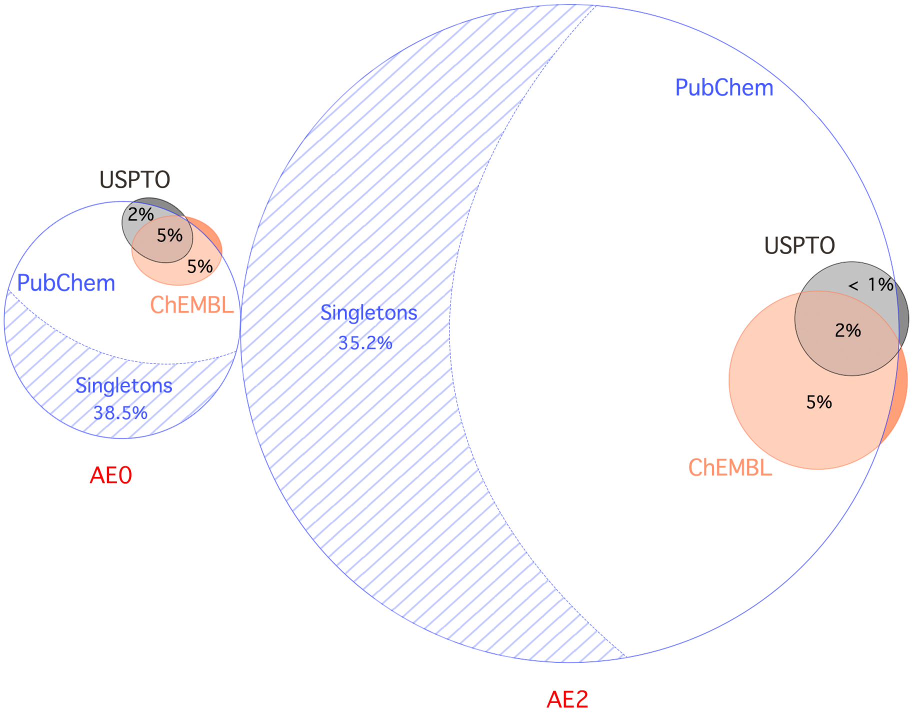


Figure 5. Area-proportional Euler graph representing the space of atomic environments for the following databases; PubChem 110M, ChEMBL 2.08M (ChEMBL v28, as of May 2021) and USPTO-Fully atom-mapped 500K reactions (ca. 1.3M molecules). AE0 is scaled up by 20 times for better visual interpretation.

degenerate, i.e., have fewer reactant candidates corresponding to a prediction, than ECFP fingerprints in retrieval process. Using ECFP bit indices for database search retrieves 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.

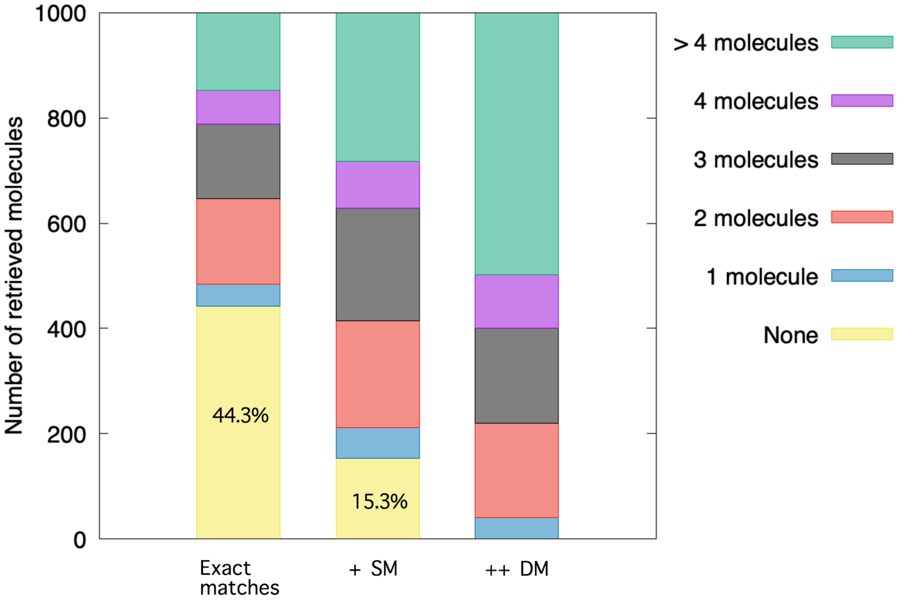


FIGURE 6. Retrieval of reactant candidates via a large PubChem compound search database.

4. Conclusion

We developed a new template-free retrosynthesis prediction model, RetroTRAE, using the transformer architecture and atom environment representation. RetroTRAE showed comparable or improved performances to other state-of-the-art models. The present approach provided reactant candidates with an exact match accuracy of 53.4%. Besides exact match accuracy, high-quality reactant candidates selected by soft and hard thresholds are found to be statistically significant at below 1.0e-04 level. The average prediction accuracy with a threshold of $T_c \geq 0.85$ is observed around 67%, outperforming the current state-of-the-art methods by a large margin. We demonstrated atom environments as promising descriptors for studying reaction route prediction and discovery since they provide highly descriptive representation free from grammatical complexity of SMILES.

5. 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	SGD
Dropout	0.1, 0.2, 0.5	0.1
Number of epoch	600-1500	1000
Validation per epoch	@2—@100	@100
Learning Rate	0.01—2.5	0.1, 0.05, 0.01
Learning Rate Scheduler	Decay, SGDR	SGDR
Cycle per epoch	3/1—1/3	5/4
Decay factor	0.8 - 0.98	0.91

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

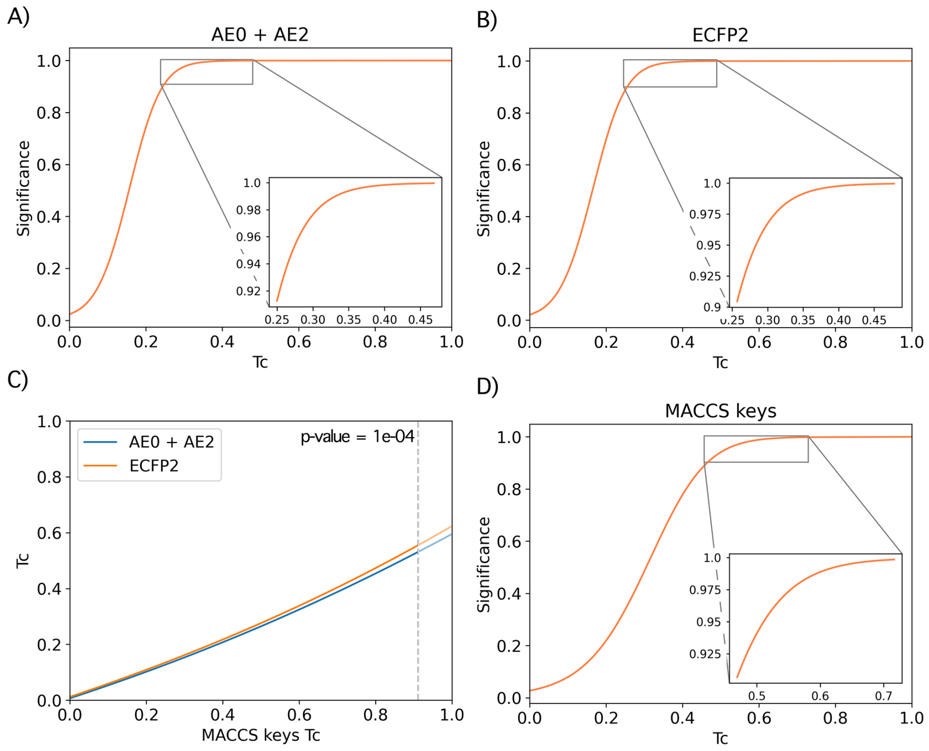


FIGURE 7. Figures A, B and D represent the cumulative distribution function of the reactants in the USPTO DB for the unified atom environments, ECFP2, and MACCS keys respectively. The measure 1 – (p-value) is used to assess significance. P-values has the range 0 to 1 and smaller p-values indicate higher 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] L Bottou, *Stochastic Gradient Learning in Neural Networks*, Proceedings of Neuro-Nımes **91** (1991), no. 8, 12.

[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.