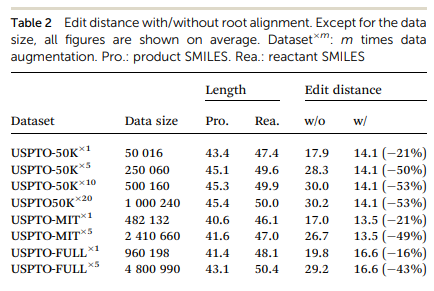
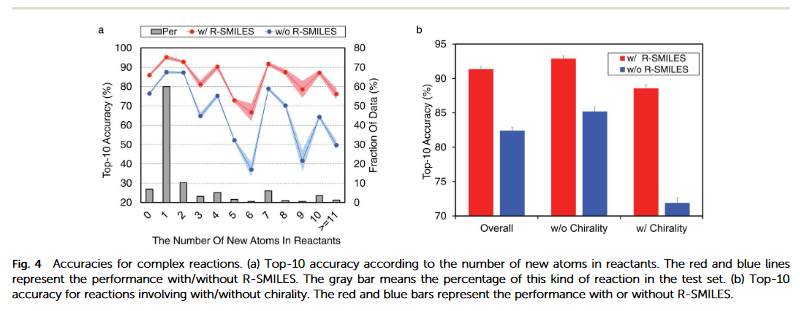
# Root-aligned SMILES: a tight representation for chemical reaction prediction

* Abstract: - A popular computational paradigm formulates synthesis prediction as a sequence-to-sequence translation problem, where the typical SMILES is adopted for molecule representations.
* The molecular graph topology is largely unaltered from reactants to products, resulting in the suboptimal performance of SMILES if straightforwardly applied.
* The researchers propose the root-aligned SMILES (R-SMILES), which specific a tightly aligned one-to-one mapping between the product and the reactant SMILES for more efficient synthesis prediction.
* Because strict one-to-one mapping and reduced edit distance, the computational model is largely relieved from learning the complex syntax and dedicated to learning the chemical knowledge for reactions.
* Introduction: - the first computer-aided synthesis planning program LHASA was formally proposed by Corey et al. and showed great potential. Since then, many rule-based organic synthesis systems have come out, like SYNLMA, WODCA, and Synthia.
* The current literature can be roughly categorized into two schools: selection-based methods and generation-based methods.
* Selection-based methods turn synthesis prediction into a ranking or classification problem, where the goal is to rank the matched reaction templates or target molecules higher than those unmatched for the input molecule.
* The selection-based methods are unable to predict templates that are not in the training set, which makes it suffer from poor generalization on new target structures and reaction types.
* Generation-based methods, however, address the synthesis prediction with a generative model where target compounds are generated, which significantly alleviates the poor generalization issue of selection-based methods.
* The first and one of critical step is to select the appropriate representation forms of both the product and the reactants.
* A molecular graph explicitly describes the topological structure of the molecule, upon which the recently well-developed GNNs can be directly leveraged.
* The graph-based representation has one problem that is it involve a graph generation problem, which is challenging and usually solved by sequential graph edit operation predictions.
* The other popular paradigm to represent molecules is using strings that are generated following some predefined chemical notation systems, of which the simplified molecule-input system (SMILES) is most widely used currently.
* In this paper the researchers argue that the general-purpose SMILES is deficient for the synthesis prediction problem.
* The one-to-many mapping between input SMILES and output SMILES renders synthesis prediction extremely challenging as the computational model should learn not only the chemical rules for chemical reactions but also the SMILES that ensures a one-to-one mapping between molecules and SMILES.
* The large input-output SMILES discrepancy leaves the search space of reactants huge, degrading the performance of synthesis prediction models.
* The canonical SMILES is incompatible with some data augmentation techniques where multiple SMILES are needed for one molecule to bypass the data scarcity issue, as the concept of “canonical SMILES” is violated by multiple SMILES for one molecule.
* The molecular graph topology is in fact largely unaltered from reactants to products as the molecular charges usually occur locally during the chemical reactions.
* The R-SMILES adopts the same atom as the root of the SMILES strings for both the products and the reactants, which makes the input and the output SMILES maintain a one-to-one mapping and highly similar to each other.
* The researchers proposed a transformer-based autoencoder for synthesis prediction.
* The researchers first pretrain the proposed autoencoder with the cheaply available unlabeled molecular data for extracting the compact molecular representations and mastering essential SMILES syntax in the decoder.
* After that the model is finetuned with the reaction data, where the model is largely relieved from learning the complex syntax and can be dedicated to learning the chemical knowledge for reactions.
* The researchers’ method successfully predicted several multistep retrosynthesis, that illustrates its great potential in complicated synthesis planning tasks.
* Methods: - the researchers implement their method on different synthesis tasks, including, reactant-to-product, product-to-reactant, product-to-synthon, and synthon-to-reactant.
* The first two methods can be defined as template-free method and the other two can be defined as semi-template method.
* For simplicity, the product is abbreviated as P and the reactant as R. The direct transformation between products and reactants is denoted by P2R and R2P.
* Semi-template methods decompose retrosynthesis into two stages: (1) first identify intermediate molecules called synthons, and then (2) complete synthons into reactants.
* The researchers use S for synthons, and therefore the P2S and S2R represent the two stages respectively.
* The all four tasks are all formulated as end-to-end seq2seq problem and solved by the same model architecture to make comparisons with state-of-the-art (SOTA) methods.
* Because the reaction type is not always available in real-world scenarios, all experiments in this research are carried out without this information.
* Datasets and data preprocessing – the researchers used the USPTO-50K, USPTO-MIT and USPTO-FULL datasets.
* The retrosynthesis prediction, reactions that contain multiple products are duplicates into multiple reactions to ensure every reaction in data has only one product.
* During the pretraining stage, depending on whether it is a forward or retrosynthesis prediction, products or reactants in the training set of USPTO-FULL are used for self-supervised training, where molecules in the test set of USPTO-50K and USPTO-MIT are removed.
* Root-aligned SMILES – they follow Schwaller et al.’s regular expression to tokenize SMILES to meaningful tokens.
* The researchers used atom mapping in the reactions to find the common structures.
* The root alignment operation is effortless in the P2R stage, where the input is only a single product.
* The researcher first selects a root atom from the product randomly, then set it as root atom to get the product SMILES.
* The researchers remove all atom mapping from the final input and output to avoid any information leak.
* The researcher put the product and synthon SMILES together as input, separated by a special token that does not exist in the SMILES syntax.
* They also choose to align reactants to synthons to minimize the difference between the input and the output since there is a one-to-one mapping between synthons and reactants.
* In the R2P stage, the researchers align the product SMILES to the largest reactant. After root alignment, the input and output are highly similar to each other, which helps the model to reduce the search space and makes cross-attention strongest.
* Data augmentation with R-SMILES – they apply 20x augmentation at training and test sets of USPTO-50K and 5x augmentation at training and test sets of USPTO-MIT and USPTO-FULL.
* During the training phase, by enumerating different atoms as the root of SMILES, they can obtain multiple input-output pairs as the training data.
* After that in the inference stage, they input several different SMILES representing the same input to obtain multiple sets of outputs.
* Results and discussion: - Statistical analysis of the minimum edit distance with R-SMILES – the minimum edit distance between two strings is defined as the minimum number of editing operations needed to transform one into the other.
* The researcher adopts this to measure the discrepancy between input and output SMILES. Without R-SMILES, the average minimum edit distance between product and reactant SMILES is 17.9 on USPTO-50K, 17.0 on USPTO-MIT, and 19.8 on USPTO-FULL.
* But the proposed distance is 14.1, 13.5 and 16.6, decreasing by 21%, 21% and 16% respectively.



* To alleviate the over-fitting problem, data augmentation with randomized SMILES is critical and widely used in existing methods, but it would inevitably lead to a significant increase in the edit distance.
* The larger discrepancy and one-to-many mapping of randomized SMILES make the learning problem more difficult, hindering the performance of synthesis prediction.
* Comparisons with SOTA methods – top-*K* exact match accuracy, which represents the percentage of predicted reactants that are identical to the ground truth, is adopted as the metric to evaluate the performance.
* The researchers also used the “maximal fragment accuracy” to evaluate the performance of P2R. It requires the exact match of only the largest reactant.
* The top-*K* exact match accuracy is used as the main metric to report the performance.
* The researchers conduct the experiments in two settings: (1) separated and (2) mixed.
* The researchers’ method obtains better results with the exception of top-1 accuracy.
* The results of the retrosynthesis prediction have three main conclusions:
  + (1) the proposed P2R variant consistently outperforms SOTA competitions by a large margin. The P2S and S2R variants also achieve the best results except for the top-1 accuracy on the USPTO-50K dataset. They also combine two phases together to get their product-to-synthon-to-reactant method that outperforms the current best semi-template method that outperforms the current best semi-template method.
  + (2) the Levenshtein augmentation ensures the high similarity between the input and the output SMILES as the researchers do, it cannot guarantee the one-to-one mapping between them, which largely inhibits its performance. By specifying the root atom of input and output SMILES, the researchers’ method can effectively guarantee the one-to-one mapping between them.
  + (3) the researchers’ P2R variant achieves superior or at least comparable performance to the current SOTA template-based method LocalRetro on the USPTO-50K dataset. The template-based approaches does not generalize well to new reactions templates. It strongly demonstrates the limitations of template-based methods.
* Superiority of the proposed R-SMILES with data augmentation – the researchers adopt the vanilla transformer, that is popular language translation model, as the retrosynthesis model.
* In retrosynthesis prediction, data augmentation can be applied to both the training and the test data, or only one of them.
* To test the performance of the R-SMILES with data augmentation, different times of augmentation are conducted on training and test data.
* They find that the performance with R-SMILES is consistently superior to the widely used canonical SMILES in the same data augmentation scenario.
* They also find that if no training data augmentation is applied, doing augmentation on the test data usually lowers the performance with the canonical SMILES.
* At last, the researchers observed that by making plot-level comparisons, they find that with more training data augmentation, the proposed R-SMILES yield higher accuracy.
* If no data augmentation is applied at test time, 5x and 20x data augmentation of the training set increase the top-10 accuracy from 76.2% to 82.4% and 83.0%, respectively, and without R-SMILES it performs lower.
* Thus, the researchers conclude that if too many training data augmentation is applied without R-SMILES, the retrosynthesis task becomes a one-to-many problem.
* Which leads to model to very difficult situation to learn useful chemical knowledge for retrosynthesis. And if no training data augmentation is used, the model may easily suffer from the overfitting problem, which leaves a trade-off issue regarding the data augmentation.
* Visualization of cross-attention mechanism in transformer with R-SMILES – to further illustrate the transformer works with R-SMILES, the researcher selects the 4 reactions and displays the visualization of the cross-attention maps in the retrosynthesis prediction.
* The adopted transformer is an autoregressive model, where the last predicted token is taken as input for predicting the next token.
* By feeding the same canonical SMILES and averaging the attention of each attention head in the last layer if the Transformer Decoder, the researchers get these attention maps to make a direct comparison.
* The attention of the output tended to pay much attention to some input tokens related to the SMILES syntax like ‘)’, and this problem exists in all maps obtained by the model trained with canonical SMILES.
* In the contract the proposed R-SMILES, the model gave the attention that is paid more on corresponding tokens and also succeeded.
* The model trained with R-SMILES not only obtained a well-aligned attention map but also correctly predicted the target R-SMILES, where the target R-SMILES is also the canonical SMILES.
* The model trained with canonical SMILES was unable to find alignment and had to focus on the global information, which ultimately led to the disordered attention maps and the failure of the predictions.
* The R-SMILES gave ordered attention maps and succeeded to predict the target R-SMILES.
* It also shows that the researchers’ proposed R-SMILES effectively allows the model to focus on learning chemical knowledge for reactions and thus improves the accuracy of the model prediction.
* Evaluating R-SMILES in more aspects of retrosynthesis – the researchers investigate the performance of R-SMILES with some more complex reactions in the USPTO-50K, including reactions involving many new atoms in the reactants and chirality.
  + The number of new atoms in reactants – the researchers illustrate the top-10 accuracy with and without R-SMILES with the number of new atoms and they find the similar result, that is, the red line always above the blue line, which shows that the performance with R-SMILES surpasses the other by a large margin.
  + For the reactions whose numbers of new atoms are 9, the improvement is impressively 39.3% demonstrating the R-SMILES remains robust even with small amounts of data.
  + Chirality – Chirality is a property of asymmetry and is important in drug discovery and stereochemistry. It can be represented by ‘@’ or ‘@@’ in SMILES sequences.
  + When chirality exists in the reaction, the accuracy without R-SMILES drops 13.3%. Whereas the researchers’ model drops only 4.3%.
  + To be specific the researchers believe that R-SMILES helps that chiral reaction mainly in two ways: (1) the reduction of editing distance of the chiral reaction is more significant than the overall one. (2) for USPTO datasets, the chiral signatures of the input and output tend to be identical after alignment, which makes the model usually only need to maintain the chiral consistency.



* Multiple retrosynthesis prediction by (our) researchers’ method – the researchers verify their method with several multistep retrosynthesis examples like febuxostat, salmeterol, an allosteric activator for GPX, and a 5-HT receptor ligand.
* Febuxostat is a novel anti-gout drug as the non-purine selection inhibitor of xanthine oxidase.
* The researchers’ first step is hydrolysis of the ester, which is exactly the same as reported. For the remaining steps, their method provides two different synthesis routes.
* The first one is the same as reported, in which 3-cyano-4-isobutoxyphenyl boronic acid and ethyl 2-bromo-4-methylthiazole-5-carboxylate are taken as the reactants of the Suzuki cross-coupling reaction.
* The last steps of them both involve borylation, where the second one is reported by Ishiyama et al. the researchers can make a detailed comparison between these two pathways in terms of yield and price: (1) there are two main findings for the researchers in Urawa et al.’s study (a) boronic acid is thermally less stable than the corresponding boronic ester. Therefore, the ester is more likely to be better for avoiding possible thermal decomposition. (b) the introduction of pinacol boronate can effectively reduce the generation of side reactions, i.e., reductive dehalogenation reactions, which helps to afford the desired product quantitatively.
* The second synthetic pathway is consistent with these findings, which shows the second is likely to have higher yields. (2) from the Reaxys database, it is found that the building block of the second pathway is much cheaper compared to the first path.
* Guo et al. proposed a reaction pathway for it based on the asymmetric Henry reaction.
* The first three steps used by researchers are as follows: - the first step reports the hydrolysis of cyclic acetal, where cyclic acetal has been proved to be stable. The formation of the cyclic acetal can effectively prevent the occurrence of side reactions, which illustrates the model has distinguished the properties of protection groups and preserved them to the starting compound.
* The second step involves the amination of halohydrocarbon, and the third step involves the reduction of the nitro group.
* The final step, which is the core reaction, is the asymmetric Henry reaction, where the researchers’ method has successfully reproduced the generation of new chiral centers at the rank-1 prediction.
* The synthetic pathway of the GPX activator compound is reported by Lin et al. they predicted the synthetic pathway with a template-free model by enumerating different reaction types.
* The researchers’ model can succeed for all five reaction steps within the top-2 predictions without the reaction type.
* Nirogi et al. proposed a benzopyran sulfonamide derivative as an antagonist of 5-HT receptor. Its pathway has seven reaction steps, and the researcher’s method succeed all the rank-1 prediction except the sixth one prediction at rank-6.
* The second and fourth steps have attracted the researchers, they have Hinsberg reaction and nucleophilic aromatic substitution reaction (SNAr).
* In the Hinsberg reaction, primary amines are able to react with benzenesulfonyl chloride. In the SNAr, the meta-nitro group reduces the density of electron cloud, which is conducive to the occurrence of reaction.
* The researchers’ method proposes a novel synthetic pathway for febuxostat that is more consistent with experimental experience.
* Limitations: - the accuracy of the R-SMILES is not so high as that of other reactions. To make it clearer, the researchers also calculated the edit distance between the input and the output SMILES for these reactions.
* Compared with that of non-ring reaction R-SMILES, the edit distance of ring reactions is significantly larger.
* The distance between input and output strings will degrade the reaction prediction performance.
* The atom mapping annotations in the dataset may also be a limitation of the proposed method.
* The reported results on the USPTO-FULL dataset in the researchers’ manuscript, all the R-SMILES are generated with the Indigo toolkit. The proposed method outperforms other competitions at any top-*K* accuracy.
* Conclusions: - R-SMILES specifies a tightly aligned one-to-one mapping between the input and output SMILES, which decreases the edit distance significantly.
* The synthesis prediction model is largely relaxed from learning the complex syntax and can be dedicated to learning the chemical knowledge for reactions.
* The synthetic pathways of some organic compounds are successfully predicted to showcase the effectiveness of the proposed method.
* Since R-SMILES maintains the high similarity of the input and the output, retrosynthesis can be formulated as a grammatical error correction problem rather than a translation from scratch.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shalun Yao, Min Wu, Tingjun Hou, and Mingli Song