AUGMENTED TRANSFORMER ACHIEVES 97% AND 85% FOR TOP5 PREDICTION OF DIRECT AND CLASSICAL RETRO-SYNTHESIS

A PREPRINT

Igor V. Tetko*
Institute of Structural Biology
Helmholtz Zentrum München,
and BigChem GmbH,
Germany, Munich
itetko@bigchem.de

Pavel Karpov
Institute of Structural Biology
Helmholtz Zentrum München,
and BigChem GmbH,
Germany, Munich
carpovpv@gmail.com

Ruud Van Deursen
Firmenich International SA,
Research&Development Division,
Switzerland, Geneva
ruud.van.deursen@firmenich.com

Guillaume Godin
Firmenich International SA,
Research&Development Division,
Switzerland, Geneva
guillaume.godin@firmenich.com

ABSTRACT

We investigated the effect of different augmentation scenarios on predicting (retro)synthesis of chemical compounds using SMILES representation. We showed that augmentation of not only input sequences but also, importantly, of the target data eliminated the effect of data memorization by neural networks and improved their generalization performance for prediction of new sequences. The Top-5 accuracy was 85.4% for the prediction of the largest fragment (thus identifying principal transformation for classical retro-synthesis) for USPTO-50k test dataset and was achieved by a combination of SMILES augmentation and beam search. The same approach also outperformed best published results for prediction of direct reactions from the USPTO-MIT test set. Our model achieved 90.4% Top-1 and 96.5% Top-5 accuracy for its most challenging mixed set and 97% Top-5 accuracy for the USPTO-MIT separated set. The appearance frequency of the most abundantly generated SMILES was well correlated with the prediction outcome and can be used as a measure of the quality of reaction prediction.

1 Introduction

To synthesize an organic compound is to solve a puzzle with many pieces and potentially several pieces missing. Here, the pieces are single reactions, and finding their sequential combination to create a final product is the retrosynthesis task.

The success of the logic of organic synthesis developed by E.J. Corey [1] triggered the development of computer programs aiming to find appropriate ways to synthesize a molecule. The first retrosynthesis program LHASA [2] utilizes a template-based approach. Every template (rule, synthon) in a curated database of known transformations is sequentially applied to a target molecule, and then sets of reagents are selected according to a specified strategy. Reagents, in turn, undergo the same decompositions until a set of commercially available compounds is found. Retrosynthesis always has multiple routes – a retrosynthetic tree – ending with different starting materials. Thus, a practical algorithm for retrosynthesis has to solve not only the rule acquisition and selection problem but also has

*Helmholtz Zentrum München – Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

capabilities to effectively navigate in this tree, taking into account different strategies. These tasks relate directly to artificial intelligence strategies [3].

Due to the difficulty of maintaining the template databases, most projects dependent on them, including LHASA, did not become widely used tools. The only major exception is, perhaps, the program Synthia™ (previously CHEMATICA [4]) which is a successful commercial product. In the Synthia™ program, rules are automatically extracted from atom-mapped reaction examples [5]. However, there is an ambiguity in the mapping definition and, more importantly, the automatic rule does not take into account other undefined possible reactive centers in a molecule. Applying such transformations may result in molecules that fail to react as predicted, e.g., 'out-of-scopes' and special care to filter out these cases has to be taken [6]. Alternative way for extraction of these rules is to apply data-driven deep learning technique that corresponds to a machine learning approach where an algorithm (usually in the form of a neural network) is trained on the raw data. After the training finishes, the network contains all the implicitly encoded features (rules) of the corresponding input via its parameters. Works on reaction prediction outcomes [7] and retrosynthesis [8, 9] showed the feasibility of a symbolic approach where reactions are written as SMILES [10] strings as in a machine translation. The product is written in the "source language", whereas the set of reactants is written in the "target language". Both languages, however, have the same alphabet and grammar. First works on symbolic (retro)synthesis [8, 11] were carried out with Seq2Seq [12] models following robust and more easy to train Transformer approach [13] that bring to the-state-of-the-art results [7, 14]. Meanwhile other approaches based on similarity [15], convolutional [16], and conditional graph logic networks [17] showed promising results.

SMILES representation of molecules is ambiguous. Though the canonicalization procedure exists [18], it has been shown that models benefit from using a batch of random SMILES (augmentation) during training and inference [19, 20, 21, 22]. Recently, such augmentation was also applied to reaction modeling [7, 23, 24, 25].

In this article, we scrutinize the various augmentation regimes and show that augmentation leads to better performance compared to the standard beam search inference or evaluation of the model under different temperatures.

2 Methods

2.1 Model architecture

Following our previous study [9] we used the Transformer [13] architecture to train all the models. The key component of the Transformer architecture is a self-attention block equipped with internal memory and attention. During the training phase the block extracts and structures the incoming data, splitting it into memory keys and associated values. Thus, the block resembles a library, where all the books (values) are referred by an index (keys). On a new request the model calculates the attention to the known keys and then extracts the knowledge from the values proportionally. The Transformer shows excellent results not only on (retro) synthesis [7, 9, 14] tasks but also on ordinary classification and regression QSAR studies [21].

The performance of the Transformer was estimated for the prediction of the whole training set after each epoch. The five models with the highest fraction of correctly predicted training set SMILES were stored. As a rule, the stored models correspond to the latest epochs of training. The weights of five stored models were averaged to form the final model, which was used to predict reactions from the test sets.

After several trials, we decided to use the Transformer architecture with 6 layers and 8 heads (6x8), which was used in the original work [7].We found that using a smaller architecture with 3 layers and 8 heads (3x8), which was used in our previous study [9], required more epochs to converge and thus longer overall training time to achieve the same performance. We restricted training of the model to 100 epochs to perform model development in a reasonable time and preserve the possibility to compare different augmentation approaches. For the final optimal architectures, we further investigated the effect of training time. The beam search with n=5 beams was used to predict the test set.

2.2 Datasets

The same training set filtered from USPTO database [26] containing 50k reactions classified into 10 reaction types was used. We used splitting proposed by [8] and divided it into 40,029, 5,004, and 5,004 reactions for the training, validation, and test sets, respectively. As in the previous study [9], after observing that early stopping using validation set did not improve model test accuracy, we combined training and validation sets into the combined training set. The 5'004 test reactions were predicted only once the model training was finished and were not used at any stage of the model development.

2.3 Augmentation

The datasets used in this study were composed of both canonical and so-called augmented SMILES. Both canonical and augmented SMILES were generated using RDKit [27]. We introduced this SMILES free augmentation method into RDKit at the end of 2018 [19, 20]. The augmented SMILES were all valid structures with an exception that starting atom and direction of graph enumerations were selected by chance. The augmentation increased the diversity of the training set.

The baseline dataset contained only canonical SMILES. The other datasets also contained SMILES augmented as summarized. Four different scenarios were used to augment training set sequences. Sequences were augmented using increasingly complex datasets as shown in Tables 1 and 2. Namely, we used augmentation of products only (xN), augmentation of products and reactant/reagents (xNF), augmentation of products and reactants/reagents followed by shuffling of the order of reactant/reagents (xNS), and finally mixed forward/reverse reactions, where each retrosynthesis reaction from xNS was followed by the inverse (forward synthesis) reaction. One more analysis was performed where the Transformer was asked to predict a fixed random SMILES string.

Only xN were used for augmentations of the test sets because no information about reactant/reagents could be used for the retrosynthesis prediction.

Table 1: Augmentations of analyzed training datasets.

Dataset	Description
xN	For N=1 the dataset contains canonical SMILES for reactants and products. For N>1 in addition to one canonical SMILES, the dataset contains (N-1) instances of the same reaction with augmented SMILES for the products (input data). The SMILES of reactants were canonical.
xNR	Products are encoded as canonical SMILES, but for reactants only one of possible random SMILES was chosen.
xNF	The first instances of each reaction contained canonical SMILES while other (N-1) instances were augmented for both input (products) and output (reactants) data. The order of SMILES in output data was not changed.
xNS	Same as xNF but the order of SMILES in reactants was randomly shuffled.
xNM	The same as xNS but also contained the same number of inverted (forward synthesis) reactions. The forward reactions started with "." to distinguish them from retro-synthetic ones.

Table 2: Examples of data augmentations for two reactions. Canonical SMILES are shown in bold.

Dataset	Input (product), output (reactants) data	Example
x0	canonical, canonical	**CC(c1ccc(Br)nc1)N(C)C,CC(=O)c1ccc(Br)nc1.CNC** **O=Cc1cncc(Br)c1,O=C(O)c1cncc(Br)c1**
x2	canonical,canonical random, canonical	**CC(c1ccc(Br)nc1)N(C)C,CC(=O)c1ccc(Br)nc1.CNC** n1c(Br)ccc(c1)C(N(C)C)C,**CC(=O)c1ccc(Br)nc1.CNC** **O=Cc1cncc(Br)c1,O=C(O)c1cncc(Br)c1** c1(cncc(Br)c1)C=O,**O=C(O)c1cncc(Br)c1**
x2R	canonical, fixed random random, fixed random	**CC(c1ccc(Br)nc1)N(C)C, c1cc(Br)ncc1C(=O)C.CNC** n1c(Br)ccc(c1)C(N(C)C)C, c1cc(Br)ncc1C(=O)C.CNC **O=Cc1cncc(Br)c1**, c1c(cncc1C(=O)O)Br c1(cncc(Br)c1)C=O, c1c(cncc1C(=O)O)Br
x2F	canonical, canonical random, random	**CC(c1ccc(Br)nc1)N(C)C, CC(=O)c1ccc(Br)nc1.CNC** n1c(Br)ccc(c1)C(N(C)C)C, CC(=O)c1ccc(nc1)Br.CNC **O=Cc1cncc(Br)c1, O=C(O)c1cncc(Br)c1** c1(cncc(Br)c1)C=O, c1c(cncc1C(=O)O)Br

Table 2: Examples of data augmentations for two reactions. Canonical SMILES are shown in bold.

Dataset	Input (product), output (reactants) data	Example
x3S	canonical, canonical random, shuffled random, shuffled	**CC(c1ccc(Br)nc1)N(C)C,CC(=O)c1ccc(Br)nc1.CNC** n1c(Br)ccc(c1)C(N(C)C)C,CNC.CC(=O)c1ccc(nc1)Br CN(C(c1ccc(Br)nc1)C)C,CNC.c1cc(Br)ncc1C(O)C **O=Cc1cncc(Br)c1,O=C(O)c1cncc(Br)c1** c1(cncc(Br)c1)C=O,c1c(cncc1C(=O)O)Br n1cc(cc(c1)C=O)Br,OC(=O)c1cncc(c1)Br
x2M	canonical, canonical .canonical, canonical random, shuffled .shuffled. random	**CC(c1ccc(Br)nc1)N(C)C,CC(=O)c1ccc(Br)nc1.CNC** .CC(=O)c1ccc(Br)nc1.CNC,CC(c1ccc(Br)nc1)N(C)C n1c(Br)ccc(c1)C(N(C)C)C,CNC.CC(=O)c1ccc(nc1)Br .CNC.CC(=O)c1ccc(nc1)Br,n1c(Br)ccc(c1)C(N(C)C)C **O=Cc1cncc(Br)c1,O=C(O)c1cncc(Br)c1** .O=C(O)c1cncc(Br)c1,O=Cc1cncc(Br)c1 c1(cncc(Br)c1)C=O,c1c(cncc1C(=O)O)Br .c1c(cncc1C(=O)O)Br,c1(cncc(Br)c1)C=O

2.4 Analysis of predicted SMILES

The beam search was used to infer five reactant sets from the model for each entry in the test file. The SMILES predicted within the same beam search were sorted in the decreasing order of their probabilities. Predictions containing erroneous SMILES representation, which could not be processed by RDKit, were discarded. The remaining predictions were converted to canonical SMILES. In cases where the predicted reaction contained several disconnected SMILES, they were sorted to have the same representation. If two or more identical predictions were found for the same input only the first prediction was kept: in this way we deduplicated reactions predicted for the same input data. For augmented test datasets, SMILES predicted for the same reaction were accumulated and those with the largest number of occurrences were selected as the top-ranked. If exactly the same number of predictions were found for two or more SMILES, the weights of the SMILES were set to be inversely proportional to their relative position in the respective beam search. Precisely, to rank predictions we used the following formula:

$$rank(SMILES) = \sum_{n \in [0, augmentations)} \sum_{i \in [1, beam]} \frac{\delta(SMILES_{n,i}, TARGET)}{1.0 + 0.001 * i} \tag{1}$$

where the first sum was over canonical (n=0) and augmented SMILES for the same input reaction. When the target canonicalized SMILES was equal to the predicted canonicalized SMILES at position i of the beam search for augmentation n, δ =1. Otherwise, if predicted and target SMILES did not coincide, δ=0. The term 0.001*i was used to weight the predicted SMILES to be inversely proportional to its position in the beam search (see also Table A1).

The SMILES strings with the largest weights and thus those appeared most frequently amidst the first sequences within the beam predictions were selected as the top-ranked. The Top-1 and Top-5 SMILES were used to estimate the prediction performances of models.

2.5 Analysis of stereochemistry free datasets

About 20% of the reactions in the training and test sets had molecules with stereochemistry. The stereochemistry was encoded in SMILES with "/","\","@" and "@@" characters. However, a number of practical projects have relaxed stereochemistry requirements. Therefore, we separately reported the performance of the models for datasets with and without stereo-chemical information.

2.6 Prediction of the largest reactant

The prediction of SMILES for retro-synthesis includes exact prediction of the reactants. However, the same reaction performed using different reactants can result in a similar yield. Therefore, a prediction of only the main (the largest) reactant can be considered more relevant for retro-synthesis predictions, since we need to first identify the reaction

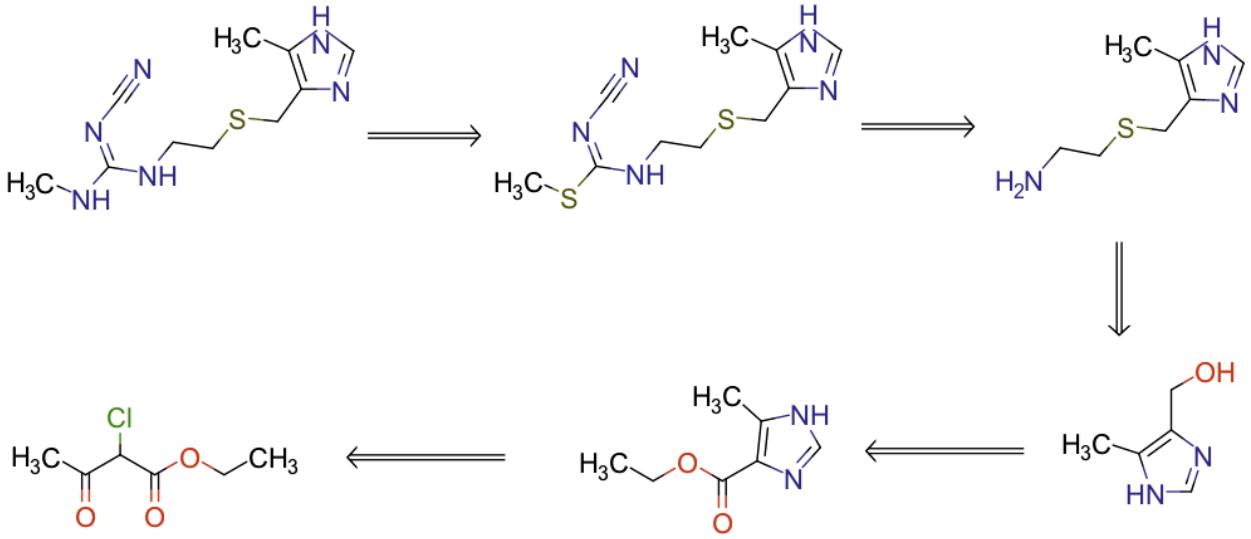

Figure 1: Classical representation of the retrosynthesis of cimetidine focusing on the principal transformations, as is typically written by synthetic chemists (adapted from https://de.wikipedia.org/wiki/Cimetidin under CC BY-SA 3.0 license).

type (Fig. 1). The exact procedure and conditions of the reactions can be considered a subsequent task. An analysis of the performance of models using such relaxed requirements was also performed. The currently used Top-n accuracy measures also include prediction of other reactants [8, 9, 14, 15, 17, 24], which may not be necessary for classical methodical retrosynthesis planning.

2.7 Character and exact sequence performance during training

During the model training, we calculated character-based performance, which corresponded to the number of exactly predicted characters for the target SMILES, as well as exact sequence accuracy indicating the number of correctly predicted exact sequences. Both of these measures were approximations of the final accuracy, for which the predicted SMILES were first converted to canonical ones and only after were compared to the target values.

3 Results and discussion

3.1 Analysis of canonical datasets

The development of a model with canonical SMILES (x1) as the training set provided 40.9% accuracy for prediction of canonical test set. An attempt to use this model to predict augmented test set (x5, x10) resulted in much lower Top-1 predictions, of 23.3% and 18.4%, respectively. This result was expected, because the model trained with only canonical sequences was not able to generalize and predict augmented SMILES, which used different styles of molecule representation.

3.2 Augmentation of products only (xN)

The augmentation of the products (input data), with just one additional augmented SMILES x2, increased Top-1 accuracy to 43.7% for the test data composed of canonical sequences. Increasing the number of augmentations in the training set did not increase the Top-1 prediction accuracy. Thus, the augmentation of the training set with just one random SMILES contributed the best performance. This result is in concordance with another study where only one random SMILES was used to augment data [25].

3.3 Analysis of character and exact sequence based prediction accuracy

To better understand the model training, we also developed several models where approximately 10% of the dataset did not participate in training but was used to monitor its prediction performance (aka validation set). Contrary to the

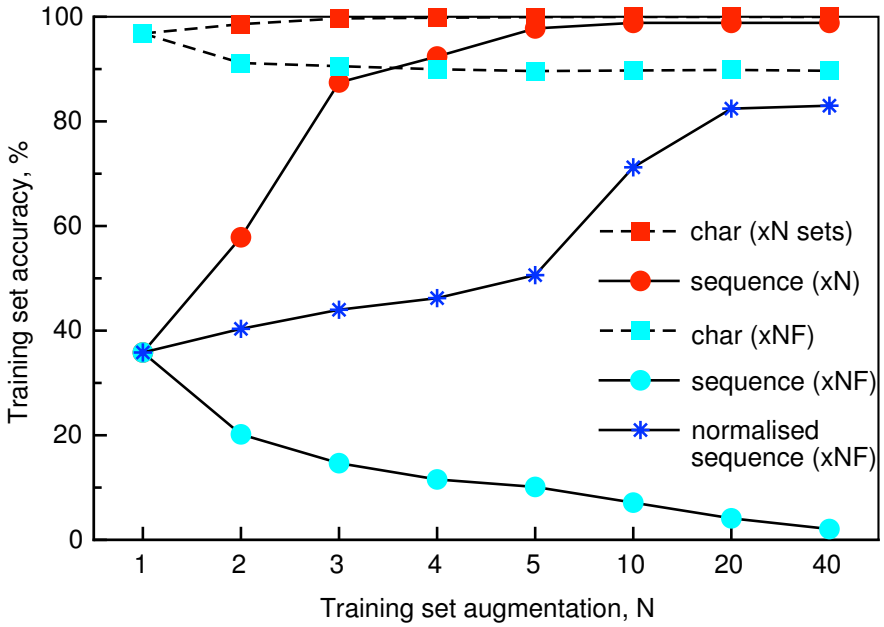

Figure 2: Character and exact sequence based accuracies calculated for the validation set. Transformer memorized the target sequences if the target sequences were all canonical SMILES (red dots). It also reasonably predicts sequence composition for randomized target SMILES (cyan rectangle, dashed) but the performance decreased for prediction of exact full SMILES. The performance normalized on percentage of canonical sequences increased with number of augmentations N, since some of random sequences were canonical ones.

test set sequences, which were used to test performance and thus did not participate in the training, SMILES used for validation were different random SMILES generated according to the respective augmentation protocol for the training set sequences. Thus, different from the test set, which tested performance of models to predict a new reaction, the validation set tested the ability of the Transformer to predict different SMILES generated for the same reaction.

The Transformer was able to recognize different representations of the same reaction. For example, when training x1, the character and exact sequence based accuracy for prediction of the validation sequences were 96.5% and 34.5%, respectively. The final performance for the test set, 40.9%, was higher because some reaction products from the Transformer provided non-canonical SMILES, which were correctly matched after the transformation to canonical ones. When using augmented training sequences (x10), the accuracies increased to 99.97% and 98.9%, for character and exact sequence-based accuracy, respectively (see Fig. 2).

The Transformer recognized different representations of SMILES for reactants and reagents for the same training set reaction and was able to exactly restore the target products, which were memorized. It was also able to memorize any random sequences. To demonstrate this, we used a random SMILES sequence (xNR set in Tables 1, 2 and Fig. 3) instead of the canonical sequences as the target for prediction. While this task was more difficult and took more epochs to train, the Transformer was able to perfectly memorize random sequences. Since the SMILES prediction target was random, the Transformer was not able to learn canonicalization rules on how to write the target. It calculated Top-1 prediction accuracy of 26.8%, for the test set which was significantly lower compared to 42.2% achieved using x10 dataset with canonical sequences as the target.

3.4 Augmentation of reactants and reagents

A boost of the Transformer performance was observed when in addition to products, i.e. the inputs SMILES, we also augmented the target SMILES, i.e. reactants and reagents. This task was more difficult for the Transformer, which resulted in a drop in both character and sequence based scores for validation sequences during the training stage. For example, when using the training dataset with one augmented SMILES, x2F, the character based accuracy dropped to 91.3%, which was lower than 98.6% calculated with the x2 dataset composed of canonical product SMILES (Fig. 2). For a larger number of augmentations, the character-based accuracy converged to a plateau, e.g. 89.96% and 89.67% for the x5F and x20F training sets, respectively. Thus despite the fact that the Transformer faced a prediction of random SMILES, it was still able to provide a reasonable prediction of their character composition.

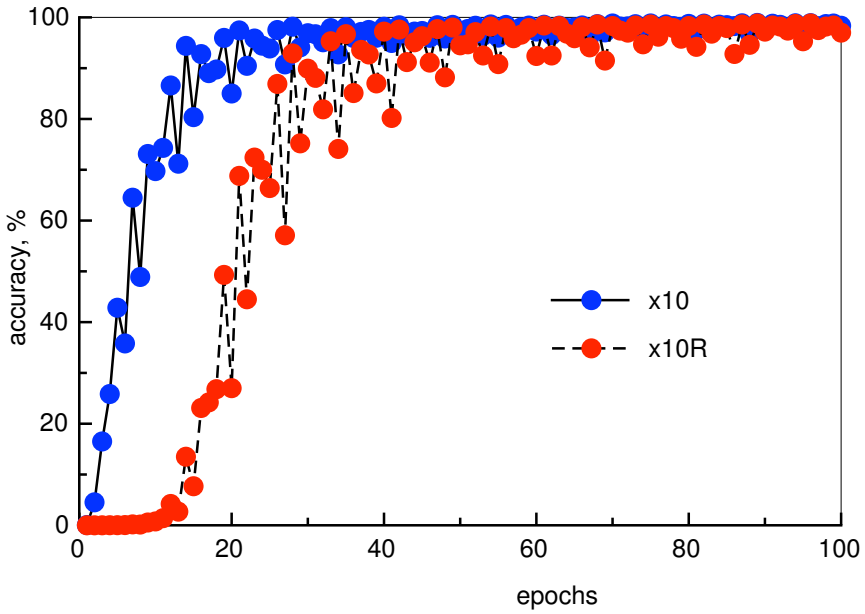

Figure 3: Validation accuracy of Transformer for prediction of canonical (x10) and random (x10R) SMILES (see Table 1 and 2 for explanation of used abbreviations).

However, of course, the Transformer was not able to predict the exact random product SMILES. This resulted in a decrease in sequence-based accuracy based on the number of augmentations for xNF training datasets (cyan circle). Still the Transformer was able to predict some of the sequences, which corresponded to the subset of canonical sequences in the validation set. Interestingly, the sequence accuracy normalized to the percentage of canonical SMILES in the validation set increased with the number of augmentations since some randomly generated sequences were canonical SMILES.

3.5 Top-1 performance analysis

For augmentations with 1 or 2 random SMILES, the Top-1 prediction performance of the models trained with augmentation of reactants and reagents only, xN, and full reaction augmentation, xNF, were similar. For a larger number of augmentations the models trained with xNF sets had systematically better performances than those developed with xN sets (Fig. 4).

The training with the x20F set provided the highest Top-1 performance of 52.1% when this model was applied to the test set generated with the same number of augmented sequences. An additional increase in the augmentations in the training or test set did not improve this Top-1 performance.

3.6 Shuffling order of reactants

In addition to augmenting the full reaction, we also randomly shuffled the orders of reactants (see xNS set description in Table 1 and 2). The effect of this additional shuffling improved Top-1 performance to 53.1% for the x20S training dataset applied to the test set with the same number of augmentations (Fig. 5). Further increasing the number of augmentations resulted in the loss of Top-1 prediction accuracy.

3.7 Top-5 performance analysis

This measure provided a relaxed estimation of the performance of the model by allowing the correctly predicted reaction listed amid Top-5 best reactions. It is as a measure of curiosity of models, by allowing them to report diverse reactions.

For each number of augmentations, the Top-5 performance generally increased with the number of augmented sequences. The highest Top-5 value was consistently calculated across different scenarios for training sets with 4-5 augmentations only (Fig. 6). The highest accuracy, 78.9%, was calculated for the mixture dataset using the x5M training set augmentation. This number had approximately 1% higher accuracy than that calculated using the x5S training set (Fig.

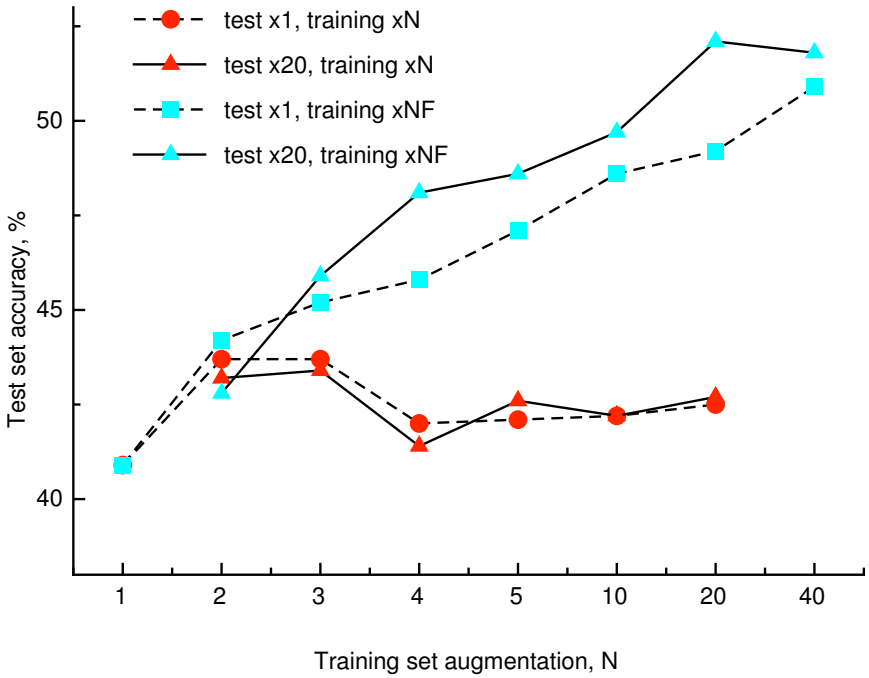

Figure 4: Top-1 performance of models developed with different number of augmentation (shown on x axis) and different augmentation scenarios applied to both test and training sets (red colour: only products were augmented; cyan colour: full reactions were augmented). The use of the large number of augmentations for the test set (solid lines) improved prediction accuracy for models developed with augmentation of full reactions but did not influence the performance of models where only input data were augmented.

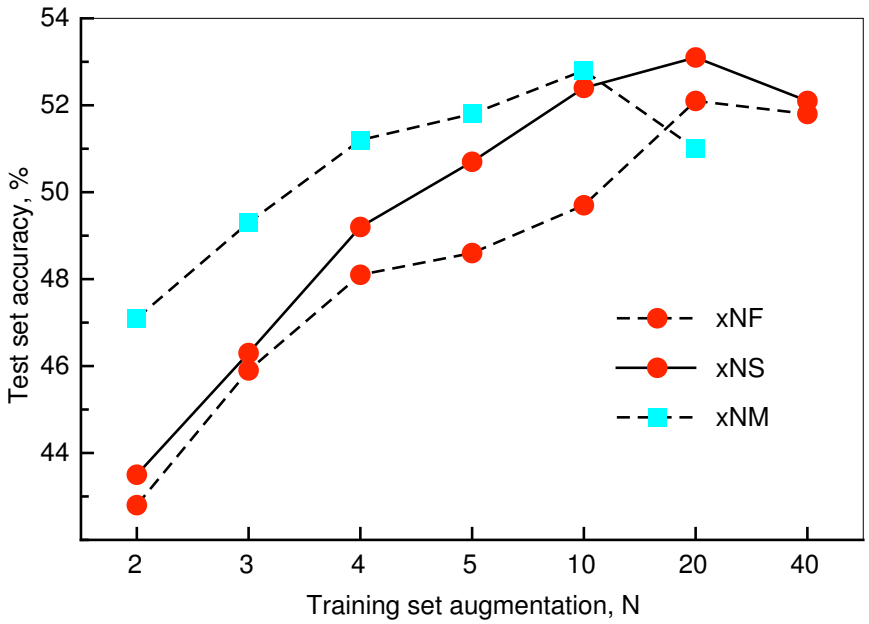

Figure 5: Top-1 accuracies calculated for models developed with different augmentation scenarios. All models were applied to the x20 test set.

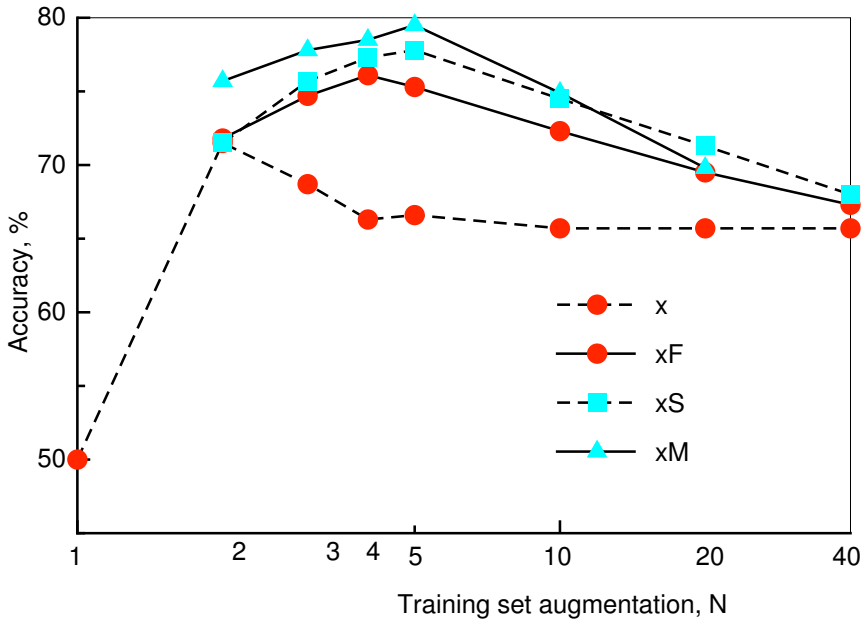

Figure 6: Top-5 performance of transformer models developed with different training set augmentation protocol (See Table 1 & 2) for prediction of the x20 test set.

6). For several analyses, we also reported the Top-10 performance for the best models to facilitate comparison with other reported methods.

3.8 Etalone model

For all studies we used a fixed number of epochs N=100. However, we needed to confirm that this was a sufficient number of epochs and to determine if we could calculate better results by training for longer. We selected the x5M training set, which provided the highest performance for Top-5 accuracy, and trained it for an additional 1000 iterations. This additional training improved Top-1 accuracy to 53.3% while Top-5 performance increased to 79.4%. This model, which is reported in Table 3[2], was used as a reference/etalone for other analyses.

Further improvement was achieved by using a large number of augmentations, and x100 as the test set. With this setting the model achieved accuracy of 53.6% and 80.8% for Top-1 and Top-5 predictions, respectively.

[2]The etalone model was built using 1000 iterations for x5M training set. Its performance was evaluated using beam size = 5, temperature =1. The altered parameters are shown for several other application scenarios. For beam = 1 and x1000 augmentations the model calculated 53.7, 80 and 84.3 for Top-1, Top-5 and Top-10 predictions, respectively. This augmentation as well as the one with beam size=10 applied to x100 analysed the same number of predicted sequences.

Table 3: Analysis of the reference model performance depending on the parameters of the application protocol.

Apply model setting	Test set x1		Test set x20		Test set x100		
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-10
etalone model predictions	48.3	72.4	53.3	79.4	53.6	80.8	85
temperature, t=1.3	49.1	67.7	52.7	77.7	53.3	78.4	83.2
no beam search, i.e. beam = 1	47.7	47.7	53.3	75.3	53.8	78.8	81.7
beam size, beam= 10	48.3	73.4	53.5	80	53.5	81	85.7
beam size, beam = 44	48.3	72.5	53.5	80	53.5	80.5	85.8

3.8.1 Influence of temperature

In our previous study [9] we observed that using higher temperature during beam size increased accuracy of the models for Top-1 prediction. It should be mentioned that no augmentation was used in that study. Under the same experimental setup with no augmentation, i.e. when predicting test set composed of only canonical sequences, x1, Top-1 accuracy of the model increased from 48.3% to 49.1% and 49.2% when using temperatures 1.3 of 1.5, respectively. However, the Top-1 and Top-5 performances for the augmented data (x20) decreased from 53.3% to 52.7% and 52.4%, respectively. For the same test set the Top-5 accuracies also decreased from 79.4% to 77.7% and 77.4% for both temperatures, respectively. Thus, while higher temperatures increased the variability of predictions and thus performance for prediction of canonical sequences, its effect was negative for the augmented data. In particular, it resulted in the lower accuracy of Top-5 predictions, i.e., the heating decreased curiosity of the neural network predictions.

3.8.2 Influence of beam search

In the above studies we consistently used beam size 5 for all analyses. The goal of beam search was to generate multiple predictions for the same data and thus to better explore variability of predictions. For example, when using the x20 test set and a beam size of 5, we obtained up to 100 individual predictions, which were used to select the most frequently appearing Top-1 and Top-5 sequences. Increasing the beam size to 10 further increased Top-1 by 0.1 to 53.3% but provided an increase of Top-5 by 0.5% to 79.9% for Top-5 predictions for this test set. The decrease of the beam size to 3 provided a slightly higher Top-1 score of 53.4% but decreased the Top-5 to 78.5% for the same test set. The use of beam size 1 decreased Top-1 and Top-5 performance to 53% and 75.4%, respectively (Table 3). These results were expected: the variation of the beam size slightly influenced the identification of the highest ranked sequence but its smaller number reduced an exploration of the space of other top-reactions for larger n.

Both beam search and augmentation increased the number of predicted SMILES which in turn led to better accuracy of model predictions. Thus both of these methods could contribute to generation of multiple predictions to be used to identify top-ranked sequences. The maximum size of the beam was restricted by the size of the target vocabulary (number of characters in the target SMILES), which was 44 characters for our dataset. Because of the design of the beam search and because we explicitly excluded duplicated predictions (see section 2.4, p. 4 as well as Table A1 on

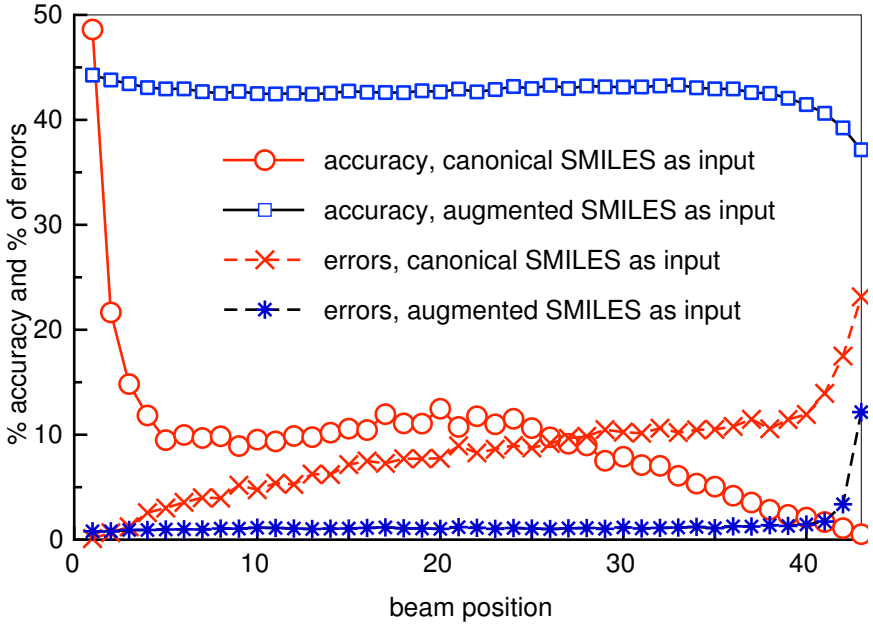

Figure 7: Accuracy of prediction of SMILES generated at the respective position of the beam search using the largest beam size=44. The use of canonical SMILES as input produced the highest accuracy (48.3%) for the first beam, which degraded for other positions of the beam while use of augmented SMILES provides about 44% correct predictions which is slowly decreasing with the increase of the beam position. The number of erroneous SMILES is increasing with the beam position for both types of SMILES, but it significantly higher for predictions when using canonical SMILES as input.

p. 17), the dataset used for analysis did not generate duplicated sequences for the same beam search. However, such sequences were indeed generated at different positions of the beam as different representation of the same SMILES. The number of non-unique sequences generated within the same beam search increased with the length of the beam. Interestingly, the use of canonical SMILES as input data contributed to the largest number of unique SMILES, which were 86%, 82% and 78% for beam search of size 5, 10 and 44, respectively. The use of augmented random SMILES as input contributed smaller numbers of unique sequences, e.g., 42%, 28% and 13% for beam search of size 5, 10 and 44, respectively. For both types of SMILES some generated SMILES were erroneous and could not be correctly converted by RDKit. For large positions of beam, canonical SMILES produced a much bigger percentage of incorrect SMILES, as compared to the use of random SMILES (see Fig. 7). The large difference in the results generated when starting from canonical and random SMILES was also observed for analysis of the percentage of correct predictions for each beam position.

The use of canonical SMILES provided a higher accuracy for the first beam position, but its accuracy was much lower for other beams. This was because the Transformer generated canonical SMILES for the canonical input sequences (e.g., 91% of valid SMILES produced at the position 1 of the beam search for input canonical SMILES were canonical ones) and since only one valid canonical SMILES could be produced, it failed to generate new correct SMILES. Indeed, during the training phase, the Transformer always had a pair of canonical SMILES as input and target sequences. Contrary to that, using augmented SMILES allowed more freedom and allowed it to contribute valid but not necessarily canonical SMILES (e.g., only 33% of generated SMILES at the position one of the beam search were canonical ones if augmented SMILES were used as input).

The decrease in performance of SMILES generated when using canonical SMILES was one of the main reasons to implement deduplication of data and retain only the first SMILES for the prediction of reactions (see section "Analysis of predicted SMILES"). When deduplication was not performed and all SMILES generated during the beam search were used to rank predictions (compare Tables A1 and A2), the Top-1 performances of models were most significantly affected when using only few augmentations, e.g. for the etalone model its accuracy dropped from 48.3% to 47% but did not change for, e.g. Top-5 performance. In principle, the analysis retaining multiple predicted sequences is based on more data and thus was more stable. Therefore, it could be used when several augmentations and/or large values of Top-n are used for analysis.

As it was mentioned above, both data augmentation and beam search could be used to generate multiple predictions. For the same number of generated sequences, 1000 per SMILES, using a beam = 10 search for the x100 set produced lower accuracy, 53.5% compared to 53.7% using augmented data with the x1000 test set without any beam search. The performance of both methods were the same and equal to 53.7% when the deduplication procedure was not used. However, the beam search contributed to better accuracy, i.e., 81% vs 80% and 85.7% vs 84.3% compared to the use of augmententation alone for Top-5 and Top-10, respectively. Thus, using beam search allowed better exploration of data when suggesting several alternative reactions. In any case the augmentation was a very important part of the beam search and for the best performance, both these approaches should be used simultaneously. We also do not exclude that optimisation of the augmentation may improve its results in the future. Moreover, data augmentation used alone without a beam search contributed superior models to the beam search used without any data augmentation.

3.8.3 Accuracy of prediction

For some reaction predictions without the use of augmented sequences or position at the beam search majority of predicted sequences were identical, while for other reactions the Transformer generated many different SMILES as possible reactants (see Table A3). The frequency of the appearance of the most frequent SMILES could, therefore, indicate the confidence of the Transformer in the prediction. Fig. 8 indicated that such frequency very well be correlated with the accuracy of prediction and could be used as a confidence score for the chemist. For about 20% of the most confident predictions, the accuracy of the retrosynthesis prediction was above 80%. Of course, the same approach can be used for Top-5 predictions by suggesting one or more plausible pathways for retrosynthesis.

3.9 Analysis of prediction accuracy for different scenarios

The accuracy of the etalone model was about 5% to 7% (Top-1) and 10% (Top-5) higher for reactions without stereochemistry than for those with it. 20% of the reactions in the test set contained molecules with stereochemistry. An increase in the number of augmentations of the test set increased the accuracy of both stereo and no-stereochemical reactions. Stereochemical reactions in the dataset may also suffer from a larger number of annotation errors or/and can have lower prediction since such data were underrepresented in the training set. Additionally, for some reactions despite the relative stereochemistry was conserved it may still define confusing information for the model due to the reactant satellite effect. The R/S could be also affected by the way the SMILES was written, e.g. from A to Z or Z to A.

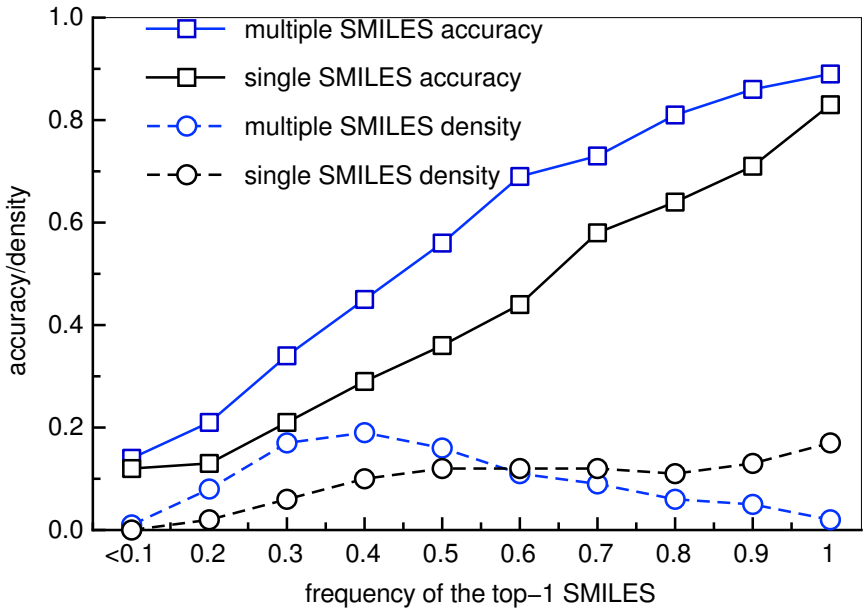

Figure 8: Accuracy and density (fraction of predictions) of the Transformer for Top-1 analysis as a function of the frequency of appearance of the Top-1 SMILES in the output of the Transformer. Results are shown for the etalone model with beam = 10, which was applied to 100x test dataset. The use of multiple SMILES in predictions achieved the higher accuracy for the same frequency of the Top-1 SMILES compared to the use of deduplicated SMILES but had the distribution density shifted towards lower frequencies.

3.10 Analysis of prediction accuracy for the largest fragment

A chemist generally performs a retrosynthesis by decomposing a target molecule into pieces. These pieces are the minimal information required to propose a retrosynthesis and knowledge of the largest fragment is frequently sufficient to start planning the reaction (see 1). That is why we decided to consider this requirement as a new measure of the model performance: accuracy of prediction of the largest fragment. By considering only the largest fragment during the evaluation of a prediction, the accuracy increased by about 5% and reached 85.4% for a pragmatic methodical scenario (see 1) of Top-5 reaction prediction. Indeed, the largest fragment is the most important entity chemists need to identify the correct reaction class. The chemists can subsequently select the best available reagents. Such strategy allows us to create a system that can automatically deduce the correct reaction class instead of explicitly providing it as input to a model used elsewhere [17]. Adding the reaction class as prior information is equivalent to getting a hint on an exam, which should not be allowed. Additionally, it also reduces the chance to propose alternate feasible reactions. Thus, in our opinion, the prior information should not be used in the metric for retrosynthesis prediction.

Table 4: Prediction accuracy of the etalone model for different subsets of the test set using beam search of size 10. For settings in bold font the accuracy was estimated for prediction of the largest fragment only.

Test set augmentation	Top-1			Top-5			Top-10		
	all	stereo (20%)	no stereo (80%)	all	stereo (20%)	no stereo (80%)	all	stereo (20%)	no stereo (80%)
x1	48.3	44.7	49.2	73.4	67.3	74.9	77.4	71	79
x20	53.4	47.3	55	80	73.3	81.9	84.2	79.2	85.4
x100	53.5	47.1	55.1	81	74.6	82.6	85.7	81.2	86.8
x1	53.5	48.7	54.7	79.2	72.7	80.9	81.6	75.1	83.3
x20	58.5	52	60.1	84.7	79	86.1	88.6	83.6	89.8
x100	58.5	51.2	60.3	85.4	79.4	86.9	90	85.1	91.2

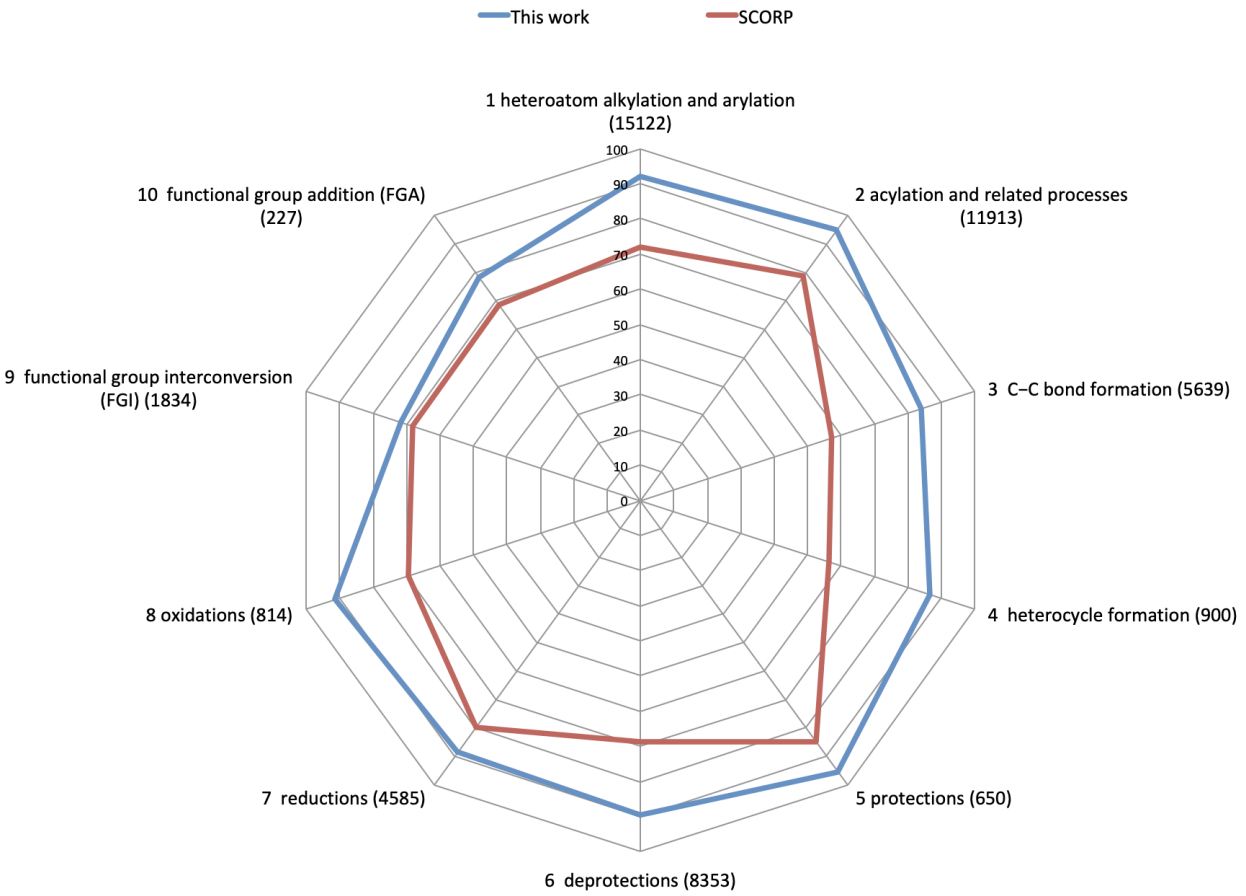

Figure 9: Accuracy of prediction of different classes of reactions.

3.11 Analysis of prediction accuracy for different classes

The original dataset [8] provides a reaction type label for every reaction. In total, 10 reaction classes ranging from protection/deprotection to carbon-carbon bond, and heterocycle formation present the most common reactions in organic synthesis. The comparison of accuracy for each class of reactions was presented in Fig. 9. Our best model showed excellent results, outperforming the state-of-the-art Self-Corrected Transformer (SCORP) [14]. Functional group interconversion and addition, as well as carbon-carbon bond formation were the most difficult for the models to predict. It was not surprising, due to diverse possibilities for choosing reactions and corresponding reactants for C-C bond creation compared to more straightforward oxidation or protection where the set of groups and reactants is more narrow.

3.12 Prediction of direct reactions

The same strategy described in this work was applied to predict the direct reactions. We used 439k reactions as the training set and predicted 40k reactions from the test set by training the Transformer with the same architecture and parameters. The separated and mixed sets were used. In the separated set reactants and reagents were separated with ">" sign while in mixed set all ">" are substituted with "." and the order of reactants and reagents was additionally shuffled. The mixed set was more difficult for training since the Transformer had to identify the reaction center from a larger number of molecules. However, such a set better reflected a practical application since separation of data on reactants and reagents in some cases would not be possible without a knowledge of the target product and thus it did provide a hint to the Transformer about the reaction center. We have removed 316 reactions where the largest products had length smaller than 5 characters. The Transformer was training using the x5N augmentation protocol for the separated set as well as the x5S and x5M protocols for the mixed set. Since it would be impractical to predict all reagents and reactants for the retrosynthesis task, which was used to additionally augment data in x5M protocol, only the largest fragment was

retained as a target for the reverse reactions. Augmented test sets were predicted using beam size 10 (Table 5). For the mixed test set the order of reactants and reagents was shuffled.

Table 5: Prediction accuracy for direct reaction from USPTO-MIT test set using beam size = 10.

Training set	Test set x1			Test set x20			Test set x100		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
x5N (separated)	91.1	96.3	96.7	91.8	96.9	97.3	91.9	97	97.4
x5S (mixed)	90	95.8	96.2	90.4	96.4	96.9	90.4	96.5	97
x5M (mixed)	90	95.5	95.7	90.2	96.1	96.5	90.2	96.2	96.8

As in previous studies, separation of reagent and reactants with ">" symbols contributed to a model (x5N) with higher prediction than for models with mixed sets (x5S and x4M). The additional augmentation of data using retrosynthesis reactions (x5M) did not improve the model. This could be due to the fact that the data for direct reactions were much larger and already contained sufficient information to develop accurate models. While using x100 test set still contributed better prediction accuracy than using x20, the improvements were in order 0.1% or no improvement at all. Thus the effect of using larger augmentations on model performance reached saturations for x100 test set.

3.12.1 Comparison with other published models

The proposed augmentation protocol achieved the best published results on the USPTO-50k dataset (Table 6) as well as on the USPTO-MIT sets (Table 7). It is interesting that the model provided the highest gain in performance for prediction of the more challenging mixed dataset. Since the model was trained with randomly shuffled augmented data, it was able to very well generalise them and provide excellent prediction for the new mixed data.

Table 6: Comparison of retrosynthesis recently published methods for retrosynthesis prediction on USPTO-50k.

Model	Top-1	Top-5	Top-10	Reference
Seq2Seq	37.4	57.0	61.7	[8]
Transformer (3*6)	42.7	69.8	–	[9]
Transformer (6*8),(self corrected)	43.7	65.2	68.7	[14]
Transformer, augmentation	44.8	57.7	79.4	[24]
Similarity-based	37.3	63.3	74.1	[15]
Graph Logic Network	52.5	75.6	83.7	[17]
The etalone model applied to x100 augmented dataset,beam size = 10	53.5	81	85.7	This work
The accuracy of the etalone model estimated for prediction of the largest fragment only	58.5	85.4	90	This work

Table 7: Comparison of recently published methods for direct synthesis prediction on USPTO-MIT set.

Model	Top-1, separated	Top-1, mixed	Top-5, separated	Top-5, mixed	Reference
Transformer (single)	90.4	88.6	95.3	94.2	[25]
Transformer (ensemble)	91		95.8		[25]
Seq2Seq	80.3		87.5		[7]
WLDN	79.6		89.2		[24]
GTPN	83.2		86.5		[28]
WLDN5	85.6		93.4		[29]
This work (x100, beam 10)	91.9	90.4	97	96.5	

4 Conclusions and outlook

This study showed that careful design of the training set was of paramount importance for the performance of the Transformer. Training the model to learn different representations of the same reaction by distorting the initial canonical data eliminated the effect of memorisation and increased the generalisation performance of models. These ideas are intensively used, e.g. for image recognition [30], and have been already successfully used in the context of several chemical problems [19, 20, 21, 22], including reaction predictions [23, 25], but were limited to the input data. For the first time we showed that application of augmentation to the target data significantly boosts the quality of the reaction prediction. We also showed that frequency of predicted SMILES could be used as a confidence metric for retrosynthesis prediction.

SMILES random augmentation had the ability to stabilize the model's learning by adding more data and adding more randomness and freedom into the network. Remarkably, this augmentation functioned similarly to ensemble learning, allowing for better statistics and improving the performance of the model. Beam search and augmentation were complementary and our final model by essence got better results than model developed using graphs representation of molecules [17], for which the use of similar data augmentation technique currently is not possible. Our next step is to develop a retrosynthesis model based on the full USPTO dataset.

5 Availability of data and materials

The training sets, models and model predictions are available at https://github.com/bigchem/synthesis.

6 Funding

This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 676434, "Big Data in Chemistry" and ERA-CVD "CardioOncology" project, BMBF 01KL1710. The article reflects only the author's view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

7 Acknowledgments

The authors thank NVIDIA Corporation for donating Quadro P6000, Titan Xp, and Titan V graphics cards for this research work.

References

[1] Corey, E. J. & Cheng, X.-M. *The Logic of Chemical Synthesis* (Wiley-Interscience, 1995).

[2] Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **228**, 408–418 (1985).

[3] Baskin, I. I., Madzhidov, T. I., Antipin, I. S. & Varnek, A. A. Artificial intelligence in synthetic chemistry: achievements and prospects. *Russ. Chem. Rev.* **86**, 1127 (2017).

[4] Klucznik, T. *et al.* Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522 – 532 (2018). URL http://www.sciencedirect.com/science/article/pii/S2451929418300639.

[5] Law, J. *et al.* Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **49**, 593–602 (2009).

[6] Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai (2018). URL https://doi.org/10.1038/nature25978.

[7] Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).

[8] Liu, B. *et al.* Retrosynthetic reaction prediction using neural Sequence-to-Sequence models. *ACS Central Science* **3**, 1103–1113 (2017).

[9] Karpov, P., Godin, G. & Tetko, I. V. A transformer model for retrosynthesis. In Tetko, I. V., Kůrková, V., Karpov, P. & Theis, F. (eds.) *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, 817–830 (Springer International Publishing, Cham, 2019).

[10] Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

[11] Nam, J. & Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv e-prints* arXiv:1612.09529 (2016). 1612.09529.

[12] Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q. (eds.) *Advances in Neural Information Processing Systems 27*, 3104–3112 (Curran Associates, Inc., 2014).

[13] Vaswani, A. *et al.* Attention is all you need. *ArXiv* (2017).

[14] Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using Self-Corrected transformer neural networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).

[15] Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-Assisted retrosynthesis based on molecular similarity. *ACS Cent Sci* **3**, 1237–1245 (2017).

[16] Ishida, S., Terayama, K., Kojima, R., Takasu, K. & Okuno, Y. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. *J. Chem. Inf. Model.* **59**, 5026–5033 (2019).

[17] Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8872–8882 (Curran Associates, Inc., 2019).

[18] Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).

[19] Tetko, I. V., Karpov, P., Bruno, E., Kimber, T. B. & Godin, G. Augmentation is what you need! In Tetko, I. V., Kůrková, V., Karpov, P. & Theis, F. (eds.) *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, 831–835 (Springer International Publishing, Cham, 2019).

[20] Kimber, T. B., Engelke, S., Tetko, I. V., Bruno, E. & Godin, G. Synergy Effect between Convolutional Neural Networks and the Multiplicity of SMILES for Improvement of Molecular Prediction. *arXiv e-prints* arXiv:1812.04439 (2018). 1812.04439.

[21] Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Fast and Reliable tool for QSAR. *arXiv e-prints* arXiv:1911.06603 (2019). 1911.06603.

[22] Jannik Bjerrum, E. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv e-prints* arXiv:1703.07076 (2017). 1703.07076.

[23] Fortunato, M., Coley, C. W., Barnes, B. & Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning (2020).

[24] Chen, B., Shen, T., Jaakkola, T. S. & Barzilay, R. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis. *arXiv e-prints* arXiv:1910.09688 (2019). 1910.09688.

[25] Schwaller, P. *et al.* Molecular transformer: A model for Uncertainty-Calibrated chemical reaction prediction. *ACS Cent Sci* **5**, 1572–1583 (2019).

[26] Lowe, D. M. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis (2012).

[27] Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

[28] Do, K., Tran, T. & Venkatesh, S. Graph transformation policy network for chemical reaction prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 750–760 (2019).

[29] Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).

[30] Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6**, 60 (2019).

Supporting Information

Table A1: Illustration of a procedure used to rank predicted reactions.

Step	Input SMILES	Beam1,Beam2,Beam3
Initial prediction	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CC(C),C(C)CC,C(N)N CNN,CCC,CC= CC.CCC,CCC.CC,C#
Canonicalisation, sorting and error detection	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC,CCCC,CNN CNN,CCC,error CC.CCC,CC.CCC,error
Elimination of duplicates and erros	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC,CCCC,CNN CNN,CCC CC.CCC
Enumeration	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC(0),CCCC(1),CNN(2) CNN(0),CCC(1) CC.CCC(0)
Ranks, see Eq. 1.		CCC = [1] +[1/(1+1./1000)] + [0] = **1.999** CNN = [1/(1+2./1000)] + [1] + [0] = 1.998 CC.CCC =[0] + [0] + [1] = 1 CCCC = [1/(1+1./1000)] + [0] + [0] = 0.999 The Top-2 ranked predictions are CCC and CNN.

Table A2: Illustration of procedure used to rank predicted reactions when using multiple predictions within the same beam.

Step	Input SMILES	Beam1,Beam2,Beam3
Initial prediction	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CC(C),C(C)CC,C(N)N CNN,CCC,CC= CC.CCC,CCC.CC,C#
Canonicalisation, sorting and error detection	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC,CCCC,CNN CNN,CCC,error CC.CCC,CC.CCC,error
Elimination of duplicates and erros	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC,CCCC,CNN CNN,CCC CC.CCC
Enumeration	SMILES_CAN SMILES_AUG1 SMILES_AUG2	CCC(0),CCCC(1),CNN(2) CNN(0),CCC(1) CC.CCC(0)
Ranks, see Eq. 1.		CCC = [1] +[1/(1+1./1000)] + [0] = **1.999** CNN = [1/(1+2./1000)] + [1] + [0] = 1.998 CC.CCC =[0] + [0] + [1] + [1/(1+1./1000)] = **1.999** CCCC = [1/(1+1./1000)] + [0] + [0] = 0.999 The Top-2 ranked predictions are CCC and CC.CCC.

Table A3: Illustration of procedure used to rank predicted reactions when using multiple predictions within the same beam.

Reaction	Frequency of SMILES	Ratio of the most frequent to all SMILES
CCOC(=O)C1CCCN(C(=O)COc2ccc(-c3ccc(C#N)cc3)cc2)C1»CCOC(=O)C1CCCNC1.N#Cc1ccc(-c2ccc(OCC(=O)O)cc2)cc1	926* 51 7 6 2 1 1 1 1 1 1 1	926/999 = 0.93
CCCCC(=O)O»CCCCC(=O)OC(=O)CCCC	203 112 107 98 57 19 16 13 12 12 12 12 11 11 11 11 11 11 11 11 11 8 8 8 8 6 6 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1	203/999 = 0.2

Star indicates the correctly predicted reaction. For the first reaction the most frequent SMILES was predicted 926 times or 93% of all predictions. For this SMILES the model was very confident in the outcome of retrosynthesis, which it correctly predicted. For the second reaction Transformer generated 78 different SMILES and the Top-1 SMILES appeared only in 20% of all predictions. The model failed to predict correct SMILES at all.