Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation

Philippe Schwaller
IBM Research Zurich
University of Cambridge
phs@zurich.ibm.com
ps779@cam.ac.uk

Teodoro Laino
IBM Research Zurich
teo@zurich.ibm.com

Théophile Gaudin
IBM Research Zurich
tga@zurich.ibm.com

Peter Bolgar
University of Cambridge
pb592@cam.ac.uk

Costas Bekas
IBM Research Zurich
bek@zurich.ibm.com

Alpha A Lee
University of Cambridge
aal44@cam.ac.uk

Abstract

Organic synthesis is one of the key stumbling blocks in medicinal chemistry. A necessary yet unsolved step in planning synthesis is solving the forward problem: given reactants and reagents, predict the products. Similar to other works, we treat reaction prediction as a machine translation problem between SMILES strings of reactants-reagents and the products. We show that a multi-head attention Molecular Transformer model outperforms all algorithms in the literature, achieving a top-1 accuracy above 90% on a common benchmark dataset. Our algorithm requires no handcrafted rules, and accurately predicts subtle chemical transformations. Crucially, our model can accurately estimate its own uncertainty, with an uncertainty score that is 89% accurate in terms of classifying whether a prediction is correct. Furthermore, we show that the model is able to handle inputs without reactant-reagent split and including stereochemistry, which makes our method universally applicable.

1 Introduction

Organic synthesis – the making of complex molecules from simpler building blocks – remains one of the key stumbling blocks in drug discovery [1]. Although the number of reported molecules has reached 135 million, this still represents only a small proportion of the estimated 10^{60} feasible drug-like compounds [2, 3]. The lack of a synthetic route hinders access to potentially fruitful regions of chemical space. Tackling the challenge of organic synthesis with data-driven approaches is particularly timely as generative models in machine learning for molecules are coming of age [4, 5, 6, 7, 8, 9, 10]. These generative models enrich the toolbox of medicinal chemistry by suggesting potentially promising molecules that lie outside of known scaffolds.

There are three salient challenges in predicting chemical reactivity and designing organic synthesis. First, simple combinatorics would suggest that the space of possible reactions is even greater than the already intractable space of possible molecules. As such, strategies that involve handcrafted rules quickly become intractable. Second, reactants seldom contain only one reactive functional group. Designing synthesis requires one to predict which functional group will react with a particular reactant and where a reactant will react within a functional group. Predicting those subtle reactivity differences is challenging because they are often dependent on the what other functional groups are nearby. In addition, for chiral organic molecules, predicting the relative and absolute configuration of chiral centers adds another layer of complexity. Third, organic synthesis is almost always a multistep

Preprint. Work in progress.

process where one failed step could invalidate the entire synthesis. For example, the pioneering total synthesis of the antibiotic tetracycline takes 18 steps [11]; even a hypothetical method that would be correct 80% of the time would only have a 1% chance of getting 18 predictions correct in a row (assuming independence). Therefore, tackling the synthesis challenge requires methods that are both accurate and have a good uncertainty estimates. This would crucially allow us to estimate the "risk" of the proposed synthesis path, and put the riskier steps in the beginning of the synthesis, so that one can fail fast and fail cheap.

In this paper, we focus on the forward problem of reaction prediction: "Given reactants and reagents, what are the most likely products?". We build on the multi-head attention Transformer architecture in machine translation [12] and the work of Schwaller et al. [13] in chemical reaction prediction and develop a sequence-to-sequence algorithm based on the SMILES [14, 15] representation. Our algorithm outperforms all reaction prediction algorithms in the literature and reaches 90.4% top-1 accuracy (93.7% top-2 accuracy) on a common benchmark dataset. Importantly, our algorithm does not make use of any handcrafted rules. It can accurately predict subtle and selective chemical transformations, getting the correct chemoselectivity, regioselectivity and, to some extent, stereoselectivity. In addition, our model can estimate its own uncertainty. The uncertainty score predicted by the model has an ROC-AUC of 0.89 in terms of classifying whether a reaction is correctly predicted. Moreover, we show that the model is able to handle inputs without reactant-reagent split – there is no need for atom mapping, reaction templates or even assigning a chemical species as reagent or reactant. Therefore, our method is universally applicable across existing datasets. Our model has been made available since August 2018 in the backend of the IBM RXN for Chemistry [16], a free web-based graphical user interface, and has been used by several thousand organic chemists worldwide to perform more than 10,000 predictions so far.

2 Related work

The long history of computational chemical reaction prediction has been extensively reviewed in [17] and [18]. Methods in the literature may be divided into two different groups, namely, template-based and template-free.

Template-based methods [19, 20, 21] use a library of reaction templates or rules. These templates describe the atoms and their bonds in the neighborhood of the reaction center, before and after the chemical reaction has occurred. Template-based methods then consider all possible reactions centers in a molecule, and enumerate the possible transformations based on the templates together with how likely each transformation is to take place. As such, the key steps in all template-based methods are the construction of templates, and the evaluation of how likely the template is to apply. The focus of the literature has thus far been on the latter question of predicting whether a template applies [20, 21]. However, the problem with the template-based paradigm is that templates themselves are often of questionable validity. Earlier methods generated templates by hand using chemical intuition [22, 23, 24]. Handcrafting is obviously not scalable as the number of reported organic reactions constantly increases and significant time investment is needed to keep up with the literature. Recent machine learning approaches employ template libraries that are automatically extracted from datasets of reactions [20, 21]. Unfortunately, automatic template extraction algorithms still suffer from having to rely on meta-heuristics to define different "classes" of reactions. More problematically, all automatic template extraction algorithms rely on pre-existing atom mapping – a scheme that maps atoms in the reactants to atoms in the product. However, correctly mapping the product back to the reactant atoms is still an unsolved problem [25] and, more disconcertingly, commonly used tools to find the atom-mapping (e.g. NameRXN [26, 27]) are themselves based on libraries of expert rules and templates. This creates a vicious circle – atom-mapping is based on templates and templates are based on atom-mapping, and ultimately, seemingly automatic techniques are actually premised on handcrafted and often artisanal chemical rules.

To overcome the limitations of template-based approaches, several template-free methods have emerged over the recent years. Those methods can in turn be categorized into graph-based and sequence-based. Jin et al. characterize chemical reactions by graph edits that lead from the reactants to the products [28]. Their reaction prediction is a two-step process. The first network takes a graph representation of the reactants as input and predicts reactivity scores. Based on those reactivity scores, product candidates are generated and then ranked by a second network. An improved version, where candidates with up to 5 bond changes are taken into account and multi-dimensional reactivity

matrices are generated, was recently presented [29]. While an earlier version of the model included both reactants and reagents in the reaction center determination step, the accuracy was significantly improved by excluding the reagents from the reactivity score prediction in the more recent versions. This requires the user to know what are the identity of the reagents, which implicitly means that the user must already know the product as the reagent is defined as chemical species that do not appear in the product! Similarly, Bradshaw et al. [30] separated reactants and reagents and included the reagents only in a context vector for their gated graph neural network. They represented the reaction prediction problem as a stepwise rearrangement of electrons in the reactant molecules. A side effect of phrasing reaction prediction as predicting electron flow is that a preprocessing step must be applied to eliminate reactions where the electron flow cannot easily be identified – Bradshaw et al. considered only a subset of the USPTO_MIT dataset, containing only 73% of the reactions with a linear electron flow (LEF) topology, thus by definition excluding pericyclic reactions and other important workhorse organic reactions. A more general version of the algorithm was recently presented in [31]. Perhaps most intriguingly, all graph-based template-free methods in the literature require atom-mapped datasets to generate the ground truth for training, and atom mapping algorithms make use of reaction templates.

Sequence-based techniques have emerged as an alternative to graph-based methods. The key idea is to use a text representation of the reactants, reagents and products (usually SMILES), and treat reaction prediction as machine translation from one language (reactants–reagents) to another language (products). The idea of applying sequence-based models to the reaction prediction problem was first explored by Nam & Kim [32]. Schwaller et al. [13] have shown that using analogies between organic chemistry and human languages sequence-to-sequence models (seq-2-seq) could compete against graph-based methods. Both previous seq-2-seq works were based on recurrent neural networks for the encoder and the decoder, with one single-head attention layer in-between [33, 34]. Moreover, both previous seq-2-seq forward prediction works separated reactants and reagents in the inputs using the atom-mapping, and [13] tokenized the reagent molecules as individual tokens. To increase the interpretability of the model, Schwaller et al. [13] used attention weight matrices and confidence scores that were generated together with the most likely product.

In this work, we present a fully attention-based model adapted from [12], the Molecular Transformer, that outperforms all previous methods while being completely atom-mapping independent and not requiring to split the input into reactants and reagents. Similar to the work of Nam & Kim [32] and Schwaller et al. [13], our current model is trained end-to-end, fully data-driven and is free of chemical knowledge/rules.

3 Data

Most of the publicly available reaction datasets were derived from the patent mining work of Lowe [35], where the chemical reactions were described using a text-based representation called SMILES [14, 15]. In order to compare to previous work we focus on four datasets. The USPTO_MIT dataset was filtered and split by Jin et al [28]. This dataset was also used in [13] and adapted to a smaller subset called USPTO_LEF by Bradshaw et al [30] to make it compatible with their algorithm. In contrast to the MIT and LEF datasets, USPTO_STEREO [13] underwent less filtering and the stereochemical information was kept. Up to date, only seq-2-seq models were used to predict on USPTO_STEREO. Stereochemistry adds an additional level of complexity because it requires the models not only to predict molecular graph edge changes, but potentially also changes in node labels. Additionally, we used a non-public time-split test set, extracted from the Pistachio database [36], to compare the performance on a set containing more diverse reactions against a previous seq-2-seq model [13].

Table 1 shows an overview of the datasets used in this work and points out the two different preprocessing methods. The *separated* reagents preprocessing means that the reactants (educts), which contribute atoms to the product, are weakly separated by a $>$ token from the reagents (e.g. solvents and catalysts). Reagents take part in the reaction, but do not contribute any atom to the product. So far, in most of the work, the reagents have been separated from the reactants to compute t. Jin et al. [28] increased their top-1 accuracy by almost 6%, when they removed the reagents from the first step, where the reaction centers were predicted. In Schwaller et al. [13] the reagents were represented not as individual atoms, but as separate reagent tokens and only included the 76 most common reagents [38]. Bradshaw et. al. passed the reagent information as a context vector to

Table 1: Dataset splits and preprocessing methods used for the experiments

Reactions in	train	valid	test	total
USPTO_MIT set [28] - No stereochemical information	409,035	30,000	40,000	479,035
USPTO_LEF [30] - Non-public subset of USPTO_MIT, without e.g. multi-step reactions	*	*	29,360	349,898
USPTO_STEREO [13] - Patent reactions until Sept. 2016, includes stereochemistry	902,581	50,131	50,258	1,002,970
Pistachio_2017 [13] - Non-public time split test set, reactions from 2017 taken from Pistachio database [37, 36]			15418	15418
Preprocessing methods				
- separated	source: COc1c(C)c(C)c(OC)c(C(CCCCC#CCCO)c2ccccc2)c1C>C.CCO.[Pd] target: COc1c(C)c(C)c(OC)c(C(CCCCCCCCO)c2ccccc2)c1C			
- mixed	source: C.CCO.COc1c(C)c(C)c(OC)c(C(CCCCC#CCCO)c2ccccc2)c1C.[Pd] target: COc1c(C)c(C)c(OC)c(C(CCCCCCCCO)c2ccccc2)c1C			

their model. In [31] it was shown that the model performs better when the reagents are tagged as such. Unfortunately, the separation of reactants and reagents is not always obvious. Different tools classify different input molecules as the reactants and hence the reagents will also differ [38]. For this reason, we decided to train and test on inputs where the reactants and the reagents were mixed and no distinction was made between the two. We called this method of preprocessing *mixed*. The *mixed* preprocessing makes the reaction prediction task significantly harder, as the model has to determine the reaction center from a larger number of molecules.

All the reactions used in this work were canonicalized using RDKit [39]. The inputs for our model were tokenized with the regular expression found in [13]. In contrast to Schwaller et al. [13], the reagents were not replaced by reagents tokens, but tokenized in the same way as the reactants.

4 Model

The model used in this work is based on the transformer architecture [12]. The model was originally constructed for neural machine translation (NMT) tasks. The main architectural difference compared to seq-2-seq models previously used for reaction prediction [13, 32], is that the recurrent neural network component was completely removed and it is fully based on the attention mechanism.

The transformer is a step-wise autoregressive encoder-decoder model comprised of a combination of multi-head attention layers and positional feed forward layers. In the encoder, the multi-head attention layers attend the input sequence and encode it into a hidden representation. The decoder consists of two types of multi-head attention layers. The first is masked and attends only the preceding outputs of the decoder. The second multi-head attention layer attends encoder outputs, as well as the output of the first decoder attention layer. It basically combines the information of the source sequence with the target sequence that has been produced so far [12].

A multi-head attention layer itself consists of several scaled-dot attention layers running in parallel, which are then concatenated. The scaled-dot attention layers take three inputs: the keys K, the values V and the queries Q, and computes the attention as follows:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

The dot product of the queries and the keys computes how closely aligned the keys are with the queries. If the query and the key are aligned, their dot product will be large and vice versa. Each key has an associated value vector, which is multiplied with the output of the softmax, through which the dot-products were normalized and the largest components were emphasized. d_k is a scaling factor depending on the layer size. The encoder computes interesting features from the input sequence, which are then queried by the decoder depending on its preceding outputs [12].

One main advantage of the transformer architecture compared to the seq-2-seq models used in [32, 13] is the multi-head attention, which allows the encoder and decoder to peek at different tokens simultaneously.

Since the recurrent component is missing in the transformer architecture, the sequential nature of the data is encoded with positional encodings [40]. Positional encodings add position-dependent trigonometric signals (see Equations 2) to the token embeddings of size d_{emb} and allow the network to know where the different tokens are situated in the sequence.

$$\text{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{emb}}}}\right), \qquad \text{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{emb}}}}\right) \tag{2}$$

We based this work on the PyTorch implementation provided by OpenNMT [41]. All the components of the transformer model are explained and illustrated graphically on [42].

While the base transformer model had 65M parameters [12], we decreased the number of trainable weights to 12M by going from 6 layers of size 512 to 4 layers of size 256. We experimented with label smoothing [43] and the number of attention heads. In contrast to the NMT model [12], we set the label smoothing parameter to 0.0. As seen below, a non-zero label smoothing parameter encourages the model to be less confident and therefore negatively affects its ability to discriminate between correct and incorrect predictions. Moreover, we observed that at least 4 attention heads were required to achieve peak accuracies. We however, kept the original 8 attention heads, because this configuration achieved superior validation performance. For the training we use the ADAM optimizer [44] and vary the learning rate as described in [12] using 8000 warm up steps, the batch size is set to approximately 4096 tokens, the gradients are accumulated over four batches and normalized by the number of tokens. The model and results can be found online [45].

5 Results & Discussion

Table 2: Ablation study of Molecular Transformer on the USPTO_MIT dataset with separated reagents. Train and test time were measured on a single Nvidia P100 GPU. The test set contained 40k reactions.

	Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]	Training	Testing
Single models						
Baseline	88.8	92.6	93.7	94.4	24h	20m
Baseline augm.	89.6	93.2	94.2	95.0	24h	20m
Baseline augm.	90.1	93.5	94.4	95.2	48h	20m
Augm. av. 20	**90.4**	**93.7**	**94.6**	**95.3**	48h	20m
Ensemble models						
Ens. of 5	90.5	93.8	94.8	95.5	48h	1h25m
Ens. of 10	90.6	93.9	94.8	95.5	48h	2h40m
Ens. of 20	90.6	93.8	94.9	95.6	48h	5h03m
Ens. of 2 av. 20 :	**91.0**	**94.3**	**95.2**	**95.8**	2x48h	32m

Table 2 shows the performance of the model as a function of different training variations. SMILES data augmentation [46] leads to a significant increase in accuracy. We double the training data by generating a copy of every reaction in the training set, where the molecules were replaced by an equivalent random SMILES (augm.) on the range of datasets and preprocessing methods. Results are also improved by averaging the weights over multiple checkpoints, as suggested in [12]. Our best single models are obtained by training for 48 hours on one GPU (Nvidia P100), saving one checkpoint every 10,000 time steps and averaging the last 20 checkpoints. Ensembling different models is known to increase the performance of NMT models [47]; however, the performance increase (Ens. of 5 / 10 / 20) is marginal compared to parameter averaging. Nonetheless, ensembling two models which contains the weight average of 20 checkpoints of two independently initialized training runs leads to a top-1 accuracy of 91%. While a higher accuracy and better uncertainty estimation can be obtained by model ensembles, they come at an additional cost in training and/or test time. The top-5 accuracies

of our best single models (weight-average of the 20 last checkpoints) on the different datasets are shown in Table 3. The top-2 accuracy is significantly higher than Top-1, reaching over 93% accuracy.

Table 3: The single model top-k accuracy of the Molecular Transformer

USPTO*		Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]
_MIT	separated	90.4	93.7	94.6	95.3
_MIT	mixed	**88.6**	**92.4**	**93.5**	**94.2**
_STEREO	separated	78.1	84.0	85.8	87.1
_STEREO	mixed	**76.2**	**82.4**	**84.3**	**85.8**

5.1 Comparison with Previous Work

As all previous works used single models, we consider only single models trained on the data-augmented versions of the datasets rather than ensembles for the remainder of this paper in order to have a fair comparison. Table 4 shows that the Molecular Transformer clearly outperforms all methods in the literature across the different datasets. Crucially, although separating reactant and reagent yields the best model (perhaps unsurprisingly because this separation implies knowledge of the product already), the Molecular Transformer still outperforms the literature when reactant and reagents are mixed. Moreover, our model achieves a reasonable accuracy in the _STEREO dataset, where stereochemical information is taken into account, whereas all prior graph-based methods in the literature cannot account for stereochemistry.

Table 4: Comparison of Top-1 accuracy in [%] obtained by the different single model methods on the current benchmark datasets.

USPTO*		S2S [13]	WLDN [28]	ELECTRO [30]	GTPN [31]	WLDN5 [29]	our work
_MIT	separated	80.3	79.6		82.4	85.6	**90.4**
_MIT	**mixed**		74				**88.6**
_LEF	separated		84.0	87.0	87.4	88.3	**92.0**
_LEF	**mixed**						**90.3**
_STEREO	separated	65.4					**78.1**
_STEREO	**mixed**						**76.2**

A looming question is how the Molecular Transformer perform by reaction type. Table 5 shows that the weakest predictions of the Molecular Transformer are on resolutions (the transformation of absolute configuration of chiral centers, where the reagents are often not recorded in the data), and the ominous label of "unclassified" (where many mis-transcribed reactions will end up). Moreover, the Molecular Transformer outperforms [13] in virtually every single reaction class. This is because the multi-head attention layer in the Molecular Transformer can process long ranged interactions between tokens, whereas RNN models impose the inductive bias that tokens fare in sequence space are less related. This bias is erroneous as the token location in SMILES space bears no relation to the distance between atoms in 3D space.

5.2 Comparison with human organic chemists

Coley et al. [29] conducted a study, where 80 random reactions from 8 different rarity bins were selected from the USPTO_MIT test set and presented to 12 chemists (Graduate students to Professors) to predict the most likely outcome. The predictions of the human chemists were then compared against those of the model. We performed the same test with our model trained on the mixed USPTO_MIT dataset and achieve a top-1 accuracy of 87.5%, significantly higher than the average of the best human (76.5%) and the best graph-based model (72.5%). Additionally, as seen in Figure 1 Molecular Transformer is generalizable and remains accurate even for the less common reactions.

Table 5: Prediction of the augm. mixed STEREO single model on the Pistachio_2017 test set, compared to [13], where the reactants and reagents were separated.

	Count	S2S acc. [13] [%]	Our acc. [%]
Pistachio_2017	15418	60.0	**78.0**
- Classified	11817	70.2	**87.6**
- Heteroatom alkylation and arylation	2702	72.8	**86.6**
- Acylation and related processes	2601	81.5	**90.0**
- Deprotections	1232	69.0	**88.6**
- C-C bond formation	329	55.6	**81.2**
- Functional group interconversion (FGI)	315	54.0	**91.7**
- Reductions	1996	71.6	**86.1**
- Functional group addition (FGA)	1090	71.8	**89.3**
- Heterocycle formation	310	57.7	**90.0**
- Protections	868	52.9	**87.4**
- Oxidations	339	41.3	**85.0**
- Resolutions	35	**34.3**	28.6
- Unrecognized	3601	26.8	**46.3**
With stereochemistry	4103	48.2	**67.9**
Without stereochemistry	11315	64.3	**81.6**
Invalid Smiles		2.8	**0.5**

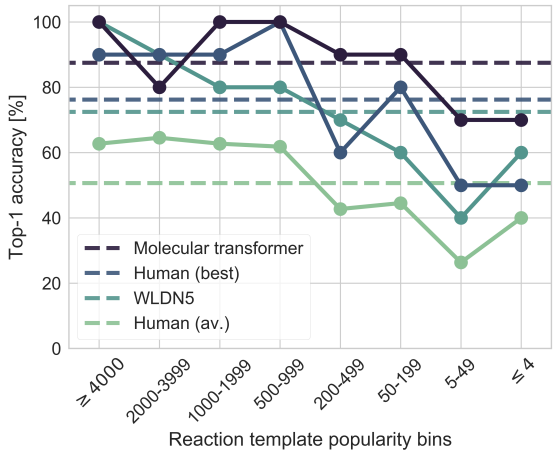

Figure 1: Top-1 accuracy of our model (mixed, USPTO_MIT) on 80 chemical reactions across 8 reaction popularity bins in comparison with a human study and their graph-based model (WLDN5) [29].

Figure 2 shows the 6 of the 80 reactions for which our model did not output the correct prediction in its top-2 choices. Even though our model does not predict the ground truth, it usually predicts a reasonable most likely outcome: In RXN 14, our model predicts that a primary amine acts as the nucleophile in an amide formation reaction rather than a secondary amine, which is reasonable on the grounds of sterics. In RXN 68, the reaction yielding the reported ground truth is via a nucleophilic substitution of Cl^- by OH^- by addition-elimination mechanism, followed by lactim-lactam tautomerism. For the reaction to work there must have been a source of hydroxide ions, which is not indicated among the reactants. In the absence of hydroxide ions, the best nucleophile in the reaction mixture is the phenolate ion generated from the phenol by deprotonation by sodium hydride. In RXN 72, the correct product predicted, but the ground truth additionally reports a by-product (which is mechanistically dubious as HCl will react with excess amine to form the ammonium salt). In RXN 76, a carbon atom is clearly missing in the ground truth. In RXN 61, we predict a SN_2 where the anion of the alcohol of the beta hydroxy ester acts as nucleophile, whereas the mechanism of the

ground truth is presumably ester hydrolysis followed by nucleophilic attack of the carboxylate group. Proton transfers in protic solvents are extremely fast, thus deprotonation of the alcohol OH is much faster than ester hydrolysis. Moreover, the carboxylate anion is a poor nucleophile.

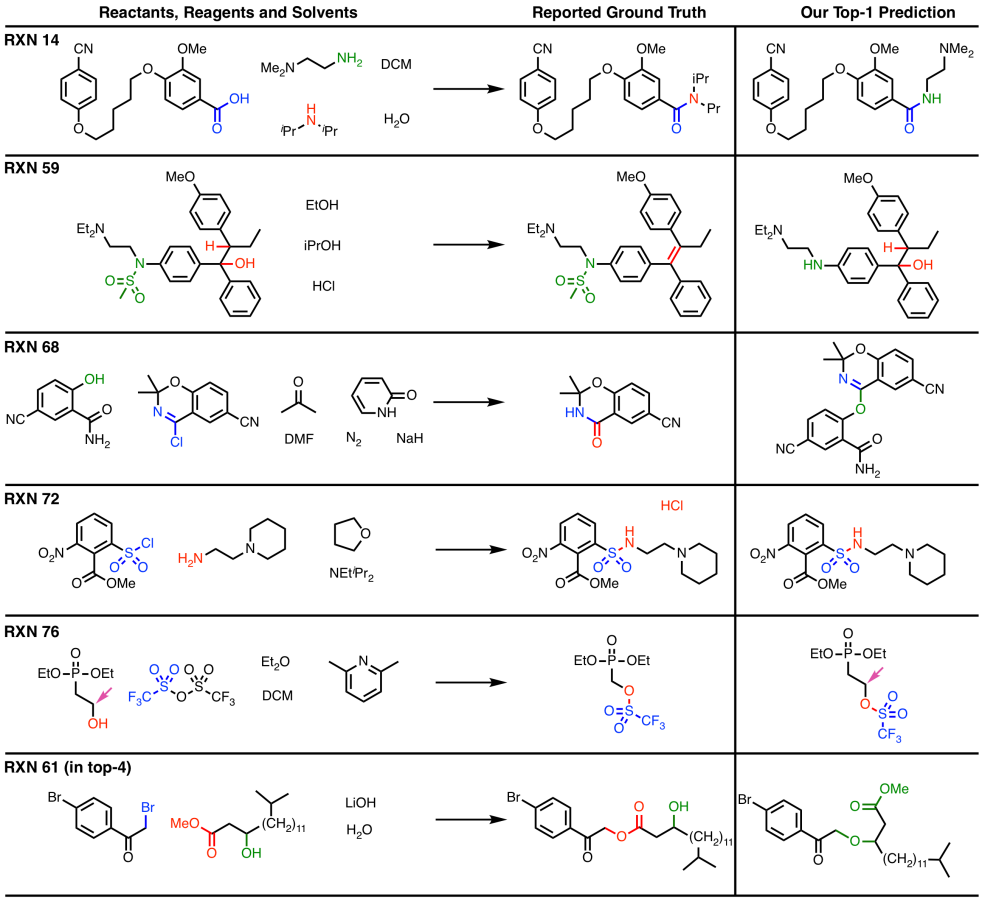

Figure 2: The 6 reactions in the human test set [29] not predicted within top-2 using our model trained on the augmented mixed USPTO_MIT set.

5.3 Uncertainty estimation and reaction pathway scoring

As organic synthesis is a multistep process, in order for a reaction predictor to be useful it must be able to estimate its own uncertainty. The Molecular Transformer model provides a natural way achieve this – the product of the probabilities of all predicted tokens can be used as a confidence score. Figure 3 plots the receiver operating characteristics (ROC) curve and shows that the AUC-ROC is 0.89 if we use this confidence score as a threshold to predict whether a reaction is mispredicted. Interestingly, Figure 3 reveals that a subtle change in the training method, label smoothing, has a minimal effect on accuracy but a surprisingly significant impact on uncertainty quantification. Label smoothing was introduced by Vaswani et al. [12] for NMT models. Instead of simply maximizing likelihood of the next target token at a given time step, the network learns a distribution over all possible tokens. Therefore, it is less confident about its predictions. Label smoothing helps to generate higher-scoring translations in terms of accuracy and BLEU score [48] for human languages, and also helps in terms of reaching higher top-1 accuracy in reaction prediction. The top-1 accuracy on the validation set (mixed, USPTO_MIT) with the label smoothing parameter set to 0.01 is 87.44% compared to 87.28% for no smoothing. However, Figure 3 shows that this small increase in accuracy comes with the cost of not being to able to discriminate between a good and a bad prediction anymore. Therefore, no label smoothing was used during the training of our models. The AUC-ROC of our single mixed

USPTO_MIT model measured on the test set was also at 0.89. The uncertainty estimation metric allows us to estimate the likelihood of a given reactants-product combination, rather only predicting product given reactants, and this could be used as a score to rank reaction pathways.

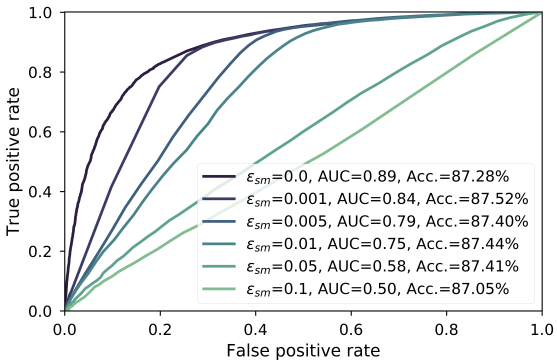

Figure 3: Receiver operatoring characteristic curve for different label smoothing values for a model trained on the mixed USPTO_MIT dataset, when evaluated on the validation set.

6 Conclusion

Building on [12] and [13] , we show that a multi-head attention Transformer network, the Molecular Transformer, outperforms all known algorithms in the reaction prediction literature, achieving 90.4% top-1 accuracy (93.7% top-2 accuracy) on a common benchmark dataset. The model requires no handcrafted rules, and accurately predicts subtle chemical transformations. Moreover, the Molecular Transformer can also accurately estimate its own uncertainty, with an uncertainty score that is 89% accurate in terms of classifying whether a prediction is correct. The uncertainty score can be used to rank reaction pathways. We point out that previous work have all considered an unrealistically generous setting of separated reactants and reagents. We demonstrate an accuracy of 88.6% when no distinction is drawn between reactants and reagents in the inputs, a score that outperforms previous work as well. For the more noisy USPTO_STEREO dataset, our top-1 accuracies are 78.1% (separated) and 76.2% respectively. The Molecular Transformer has been freely available since August 2018 through a graphical user interface on the IBM RXN for Chemistry platform [16], and has so far been used by several thousand organic chemists worldwide for performing more than 10,000 chemical reaction predictions.

Acknowledgements

PS and AAL acknowledges the Winton Programme for the Physics of Sustainability for funding. The authors thank G. Landrum, R. Sayle, G. Godin and R. Griffiths for useful feedback and discussions.

References

[1] Blakemore, D. C. *et al.* Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry* **10**, 383 (2018).

[2] Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews* **16**, 3–50 (1996).

[3] Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery* (2018).

[4] Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar variational autoencoder. *arXiv preprint arXiv:1703.01925* (2017).

[5] Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4**, 268–276 (2018).

[6] Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501* (2017).

[7] Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances* **4**, eaap7885 (2018).

[8] Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of generative autoencoder in de novo molecular design. *Molecular informatics* **37**, 1700123 (2018).

[9] Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364* (2018).

[10] Kang, S. & Cho, K. Conditional molecular design with deep generative models. *arXiv preprint arXiv:1805.00108* (2018).

[11] Charest, M. G., Lerner, C. D., Brubaker, J. D., Siegel, D. R. & Myers, A. G. A convergent enantioselective route to structurally diverse 6-deoxytetracycline antibiotics. *Science* **308**, 395–398 (2005).

[12] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 6000–6010 (2017).

[13] Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. "found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* (2018).

[14] Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31–36 (1988).

[15] Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences* **29**, 97–101 (1989).

[16] IBM RXN for Chemistry. URL `https://rxn.res.ibm.com`.

[17] Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug discovery today* (2018).

[18] Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* **51**, 1281–1289 (2018).

[19] Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS central science* **2**, 725–732 (2016).

[20] Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **3**, 434–443 (2017).

[21] Segler, M. H. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal* **23**, 5966–5971 (2017).

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

[23] Szymkuć, S. *et al.* Computer-assisted synthetic planning: The end of the beginning. *Angewandte Chemie International Edition* **55**, 5904–5937 (2016).

[24] Grzybowski, B. A. *et al.* Chematica: A story of computer code that started to think like a chemist. *Chem* **4**, 390–398 (2018).

[25] Chen, W. L., Chen, D. Z. & Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 560–593 (2013).

[26] Nextmove Software nameRXN. URL `http://www.nextmovesoftware.com/namerxn.html`.

[27] Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of chemical information and modeling* **55**, 39–53 (2015).

[28] Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems*, 2604–2613 (2017).

[29] Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity (2018). URL `https://chemrxiv.org/articles/A_Graph-Convolutional_Neural_Network_Model_for_the_Prediction_of_Chemical_Reactivity/7163189/1`.

[30] Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. & Hernández-Lobato, J. M. Predicting electron paths. *arXiv preprint arXiv:1805.10970* (2018).

[31] Anonymous. Graph transformation policy network for chemical reaction prediction. In *Submitted to International Conference on Learning Representations* (2019). Under review.

[32] Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. *arXiv preprint arXiv:1612.09529* (2016).

[33] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[34] Luong, M.-T., Pham, H. & Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

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

[36] Nextmove Software Pistachio. URL `http://www.nextmovesoftware.com/pistachio.html`.

[37] Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **59**, 4385–4402 (2016).

[38] Schneider, N., Stiefl, N. & Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling* **56**, 2336–2346 (2016).

[39] Landrum, G. *et al.* Rdkit/Rdkit: 2017_09_1 (Q3 2017) Release (2017). URL `https://zenodo.org/record/1004356#.Wd3LDY6l2EI`.

[40] Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[41] Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. M. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL* (2017).

[42] Annotated Transformer. URL `http://nlp.seas.harvard.edu/2018/04/03/attention.html`.

[43] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

[44] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[45] Molecular Transformer. URL `https://github.com/pschwllr/MolecularTransformer`.

[46] Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076* (2017).

[47] Liu, Y. *et al.* A comparable study on model averaging, ensembling and reranking in nmt. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 299–308 (Springer, 2018).

[48] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318 (Association for Computational Linguistics, 2002).