ACS central science

Research Article

Cite This: ACS Cent. Sci. XXXX, XXX, XXX−XXX

http://pubs.acs.org/journal/acscii

Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction

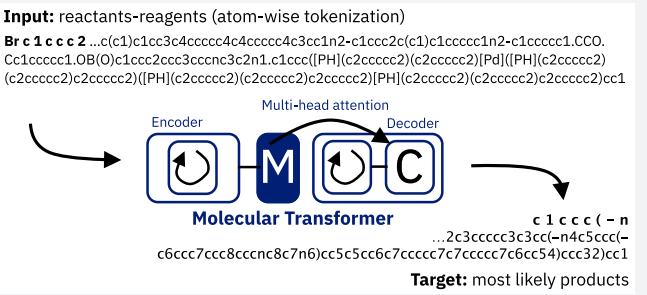
Philippe Schwaller,*[,†,‡] Teodoro Laino,[†] Théophile Gaudin,[†] Peter Bolgar,[§] Christopher A. Hunter,[§] Costas Bekas,[†] and Alpha A. Lee*[,‡]

[†]IBM Research − Zurich, Rüschlikon 8803, Switzerland

[‡]Department of Physics, University of Cambridge, Cambridge CB3 0HE, United Kingdom

[§]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

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 work, we treat reaction prediction as a machine translation problem between simplified molecular-input line-entry system (SMILES) strings (a text-based representation) of reactants, reagents, and the products. We show that a multihead attention Molecular Transformer model outperforms all algorithms in the literature, achieving a top-1 accuracy above 90% on a common benchmark data set. Molecular Transformer makes predictions by inferring the correlations between the presence and absence of chemical motifs in the reactant, reagent, and product present in the data set. Our model 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 a reactant−reagent split and including stereochemistry, which makes our method universally applicable.



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.[4−10] 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 a 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 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 have only 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 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.

The long history of computational chemical reaction prediction has been extensively reviewed in refs 12 and 13. Methods in the literature may be divided into two different groups, namely, template-based and template-free.

Template-based methods[14−16] 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

Received: June 11, 2019

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

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.[15,16] However, the problem with the template-based paradigm is that templates themselves are often of questionable validity. Previous methods generated templates by hand using chemical intuition.[17−19] Handcrafting is obviously not scalable because the number of reported organic reactions constantly increases, and a significant time investment is needed to keep up with the literature. Recent machine-learning approaches employ template libraries that are automatically extracted from data sets of reactions.[15,16] 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,[20] and, more disconcertingly, commonly used tools to find the atom mapping (e.g., NameRXN[21,22]) 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.[23] Their reaction prediction is a two-step process. The first network takes a graph representation of the reactants as input and predicts reactivity scores. On the basis of those reactivity scores, product candidates are generated and then ranked by a second network. An improved version, where candidates with up to five bond changes are taken into account and multidimensional reactivity matrices are generated, was recently presented.[24] Whereas a previous 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 the identities of the reagents, which implicitly means that the user must already know the product because the reagent is defined as a chemical species that does not appear in the product! Similarly, Bradshaw et al.[25] 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 data set, 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 ref 26. Perhaps most intriguingly, all graph-based template-free methods in the literature require atom-mapped data sets 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 simplified molecular-input line-entry system (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 and Kim.[27] Schwaller et al.[28] have shown that using analogies between organic chemistry and human language 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 (RNNs) for the encoder and the decoder, with one single-head attention layer in between.[29,30] Moreover, both previous seq-2-seq forward prediction works separated reactants and reagents in the inputs using the atom mapping, and ref 28 tokenized the reagent molecules as individual tokens. To increase the interpretability of the model, Schwaller et al.[28] used attention weight matrices and confidence scores that were generated together with the most likely product.

In this work, we focus on the question of predicting products given reactants and reagent. We show that a fully attention-based model adapted from ref 31 with the SMILES[32,33] representation, the Molecular Transformer, outperforms all previous methods while being completely atom-mapping independent and not requiring splitting the input into reactants and reagents. Our model reaches 90.4% top-1 accuracy (93.7% top-2 accuracy) on a common benchmark data set. Importantly, our model 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. Our model has been made available since August 2018 in the backend of the IBM RXN for Chemistry,[34] a free web-based graphical user interface, and has been used by several thousand organic chemists worldwide to perform more than 40 000 predictions so far.

DATA

Most of the publicly available reaction data sets were derived from the patent mining work of Lowe,[35] where the chemical reactions were described using a text-based representation called SMILES.[32,33] To compare to previous work, we focus on four data sets. The USPTO_MIT data set was filtered and split by Jin et al.[23] This data set was also used in ref 28 and adapted to a smaller subset called USPTO_LEF by Bradshaw et al.[25] to make it compatible with their algorithm. In contrast with the MIT and LEF data sets, USPTO_STEREO[28] underwent less filtering, and the stereochemical information was kept. 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 to predict not only molecular graph edge changes but potentially also changes in node labels. Additionally, we used a nonpublic 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.[28]

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

Table 1. Data-Set Splits and Preprocessing Methods Used for the Experiments

reactions in	train	valid	test	total
USPTO_MIT set[23]	409 035	30 000	40 000	479 035
-No stereochemical information				
USPTO_LEF[25]	*	*	29 360	349 898
-Nonpublic subset of USPTO_MIT, without e.g. multistep reactions				
USPTO_STEREO[28]	902 581	50 131	50 258	1 002 970
-Patent reactions until Sept. 2016, includes stereochemistry				
Pistachio_2017[28]			15 418	15 418
-Nonpublic time split test set, reactions from 2017 taken from Pistachio database[36,37]				
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			

Table 1 shows an overview of the data sets used in this work and points out the two different preprocessing methods. The *separated* reagent 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. Jin et al.[23] 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.,[28] the reagents were represented not as individual atoms but as separate reagent tokens and included only the 76 most common reagents.[38] Bradshaw et al. passed the reagent information as a context vector to their model. In ref 26, 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 because the model has to determine the reaction center from a larger number of molecules.

All of the reactions used in this work were canonicalized using RDKit.[39] The inputs for our model were tokenized with the regular expression found in ref 28. In contrast with Schwaller et al.,[28] the reagents were not replaced by reagents tokens but tokenized in the same way as the reactants.

MOLECULAR TRANSFORMER

The model used in this work is based on the transformer architecture.[31] The model was originally constructed for neural machine translation (NMT) tasks. The main architectural difference compared with seq-2-seq models previously used for the reaction prediction[27,28] is that the RNN component was completely removed, and it is fully based on the attention mechanism.

The transformer is a stepwise autoregressive encoder−decoder model composed of a combination of multihead attention layers and positional feed forward layers. In the encoder, the multihead attention layers attend the input sequence and encode it into a hidden representation. The decoder consists of two types of multihead attention layers. The first is masked and attends only the preceding outputs of the decoder. The second multihead 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.[31]

A multihead 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

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

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, then their dot product will be large and vice versa. Each key has an associated value vector, which is multiplied by 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.[31]

One main advantage of the transformer architecture compared with the seq-2-seq models used in refs 27 and 28 is the multihead attention, which allows the encoder and decoder to peek at different tokens simultaneously.

Because 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 eqs 2) to the token embeddings of size d_{emb} and allow the network to know where the different tokens are situated in the sequence.

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

The top-k outputs are decoded via a beam search. We set the beam size to 5 for all of the experiments. We based this work on the PyTorch implementation provided by OpenNMT.[41] All of the components of the transformer model are explained and illustrated graphically in ref 42.

Whereas the base transformer model had 65 M parameters,[31] we decreased the number of trainable weights to 12 M

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

Table 2. Ablation Study of Molecular Transformer on the USPTO_MIT Data Set with Separated Reagents[a]

	top-1 (%)	top-2 (%)	top-3 (%)	top-5 (%)	training	testing
			Single Models			
baseline	88.8	92.6	93.7	94.4	24 h	20 min
baseline augm.	89.6	93.2	94.2	95.0	24 h	20 min
baseline augm.	90.1	93.5	94.4	95.2	48 h	20 min
augm. av. 20	**90.4**	**93.7**	**94.6**	**95.3**	48 h	20 min
			Ensemble Models			
ens. of 5	90.5	93.8	94.8	95.5	48 h	1 h 25 min
ens. of 10	90.6	93.9	94.8	95.5	48 h	2 h 40 min
ens. of 20	90.6	93.8	94.9	95.6	48 h	5 h 3 min
ens. of 2 av. 20	**91.0**	**94.3**	**95.2**	**95.8**	2 × 48 h	32 min

[a]Training and test times were measured on a single Nvidia P100 GPU. The test set contained 40k reactions.

by going from six layers of size 512 to four layers of size 256. We experimented with label smoothing[43] and the number of attention heads. In contrast with the NMT model,[31] we set the label smoothing parameter to 0.0. As seen below, a nonzero 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 four attention heads were required to achieve peak accuracies. We, however, kept the original eight attention heads because this configuration achieved superior validation performance. For the training, we used the ADAM optimizer[44] and varied the learning rate as described in ref 31 using 8000 warm up steps, the batch size was set to ∼4096 tokens, and the gradients were accumulated over four batches and normalized by the number of tokens. The model and results can be found online.[45]

RESULTS AND DISCUSSION

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 data sets and preprocessing methods. Results are also improved by averaging the weights over multiple checkpoints, as suggested in ref 31, as well as increasing the training time. Our best single models are obtained by training for 48 h 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 with parameter averaging. Nonetheless, ensembling two models that contain the weight average of 20 checkpoints of two independently initialized training runs leads to a top-1 accuracy of 91%. Whereas a higher accuracy and better uncertainty estimation can be obtained by model ensembles, they come at an additional cost of training or test time. The top-5 accuracies of our best single models (weight average of the 20 last checkpoints) on the different data sets are shown in Table 3. The top-2 accuracy is significantly higher than the top-1 accuracy, reaching >93% accuracy.

Comparison with Previous Work. Because all previous works used single models, we consider only single models trained on the data-augmented versions of the data sets rather than ensembles for the remainder of this paper to have a fair comparison. Table 4 shows that the Molecular Transformer clearly outperforms all methods in the literature across the different data sets. 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 data set, where stereochemical information is taken into account, whereas all prior graph-based methods in the literature cannot account for stereochemistry. We note that if one was to use a reaction prediction algorithm to plan an N-step synthesis, then the probability of getting the scheme right would be p^N, where p is the probability of a single-step prediction being correct (assuming independence). Therefore, the performance gap between models becomes exponentially amplified when one deploys it to solve synthesis planning problems.

Table 3. 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**

Coley et al.[24] published their performance predictions by dividing the reactions of the USPTO_MIT test set into template popularity bins. The template popularity of the test set reactions was computed by counting how many times the corresponding reaction templates were observed in the training set. In Figure 1, we compare the top-1 accuracy of our USPTO_MIT models with the model of Coley et al.[24] Although Coley et al. had separated the reagents in this experiment, we outperform them across all popularity bins, even with our model predicting on a mixed reactants−reagents input, and the accuracy gap becomes larger as the template popularity decreases. These findings suggest that the Molecular Transformer is not simply memorizing the data and can leverage information inferred from more common reactions to make predictions on rarer reactions.

A looming question is how the Molecular Transformer performs 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 mistranscribed reactions will end up). Moreover, the Molecular Transformer outperforms[28] in virtually every single reaction class. This is because the multihead attention layer in the Molecular

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

Table 4. Comparison of Top-1 Accuracy (in %) Obtained by the Different Single-Model Methods on the Current Benchmark Data Sets

USPTO*		S2S[28]	WLDN[23]	ELECTRO[25]	GTPN[26]	WLDN5[24]	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**

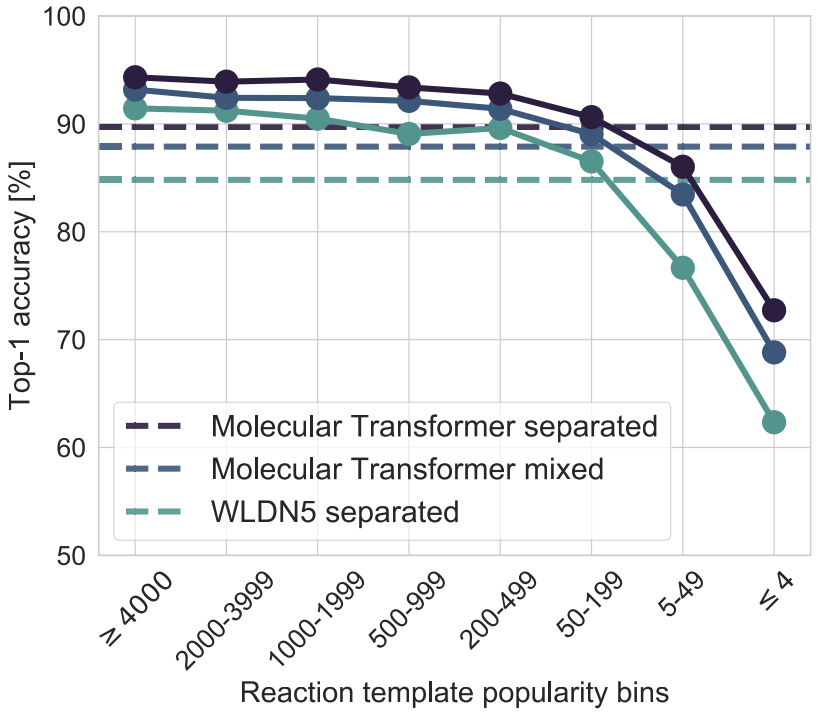


Figure 1. Molecular Transformer outperforms the state-of-the-art model across both common and rare reactions. The figure shows the top-1 accuracy of our augmented mixed and separated USPTO_MIT single model compared with the model from ref 24 on the USPTO_MIT test set, divided into template popularity bins. (The number of times a particular reaction type is seen in the data set.) The dashed lines show the average across all bins.

Transformer can process long-range interactions between tokens, whereas RNN models impose the inductive bias that tokens far in sequence space are less related. This bias is erroneous because the token location in SMILES space bears no relation to the distance between atoms in 3D space.

Figure 2 qualitatively illustrates the systematic pitfalls of the S2S RNN model[28] because of its erroneous inductive bias of assuming that only tokens close together in the SMILES string are chemically related. Figure 2A is a nucleophilic substitution. Although the reaction is simple, the RNN model predicts an erroneous product that makes little chemical sense where distal groups are joined together, an artifact of the location of those groups in the SMILES representation. Figure 2B is a simple Buchwald−Hartwig coupling reaction. RNN again predicts a chemically nonsensible product with chemically unreasonable bonds.

Examples of Chemical Challenges That Molecular Transformer Tackles. In the following section, we demonstrate the ability of Molecular Transformer to predict the outcome of a wide range of organic reactions with nontrivial selectivity involved. For some of the reactions discussed below, an organic chemist familiar with that particular class of reaction could predict the outcome after thorough reasoning. However, Molecular Transformer can immediately provide us with the ground-truth answer. All of the reactions discussed in this section and shown in Figure 3 are not in the training set.

We first consider challenges in chemoselectivity. As Molecular Transformer predicts, the treatment of the fused polycycle **1** with peracetic acid results in the epoxidation of the alkene and not the Baeyer−Villiger oxidation of the ketone.[48] Molecular Transformer also successfully predicts the stereochemistry around the two newly forming stereocenters in **2**. Selective esterification of the dicarboxylic acid **3** is possible by the sequential addition of acetyl chloride and an alcohol.[49,50] Careful thinking about the role of each reagent and the reactivity of the cyclic anhydride intermediate suggests the esterification of the unconjugated carboxylic acid. This is indeed what is observed and what Molecular Transformer predicts. The outcome of this reaction is the consequence of the 1,5-relationship between the two carboxylic acids and the presence of the conjugated double bond. Whereas it takes time and experience for an organic chemist to recognize the concurrent presence of these functional groups as their implication on the reaction outcome, Molecular Transformer can furnish the right product by inferring the reactivity of this complex pattern of distant functional groups. The reduction of

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

Table 5. Prediction of the Augm. Mixed STEREO Single Model on the Pistachio_2017 Test Set Compared with Ref 28, Where the Reactants and Reagents Were Separated

	count	S2S acc. (%)[28]	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**

5 using excess DIBAL-H was expected to lead to the unselective reduction of the secondary and the tertiary amides.[51] However, **6** was observed as the major product, in agreement with the prediction of Molecular Transformer. This shows how Molecular Transformer can help design new syntheses, ultimately saving many hours of human labor in the laboratory.

We next consider challenges in regioselectivity. Predicting the regioselectivity of electrophilic aromatic substitutions is straightforward in many cases. However, the concomitant presence of multiple directing groups and steric crowding can sometimes make human predictions ambiguous. Molecular transformer can deal with complicated examples such as the bromination of **7** with *N*-bromosuccinimide, affording **8**.[52] Molecular Transformer successfully deals with transition-metal-catalyzed reactions as well. It can predict the relative reactivity of the different C−Cl bonds in 2,4,5-trichloropyrimidine **9** in the successive Suzuki coupling reactions with phenylboronic acid.[53]

Our last examples illustrate the power of Molecular Transformer in predicting the stereoselectivity of organic reactions. The reduction of the fused bicyclic ketone **13** by lithium aluminum hydride gives the major diastereoisomer **14**, successfully predicted by Molecular Transformer.[54] The formation of the (*E*)-alkene in **16** by the treatment of **15** with tosyl chloride and lithium *tert*-butoxide is also successfully predicted.[55]

Comparing Molecular Transformer with Quantum-Chemistry-Based Predictors. Having qualitatively discussed a series of challenging examples of chemical selectivity that Molecular Transformer successfully predicts, we next turn to quantitatively explore whether Molecular Transformer has inferred the physical principles that underlie chemical selectivity. The general question of distilling interpretable rationales from machine-learning models is still an active area

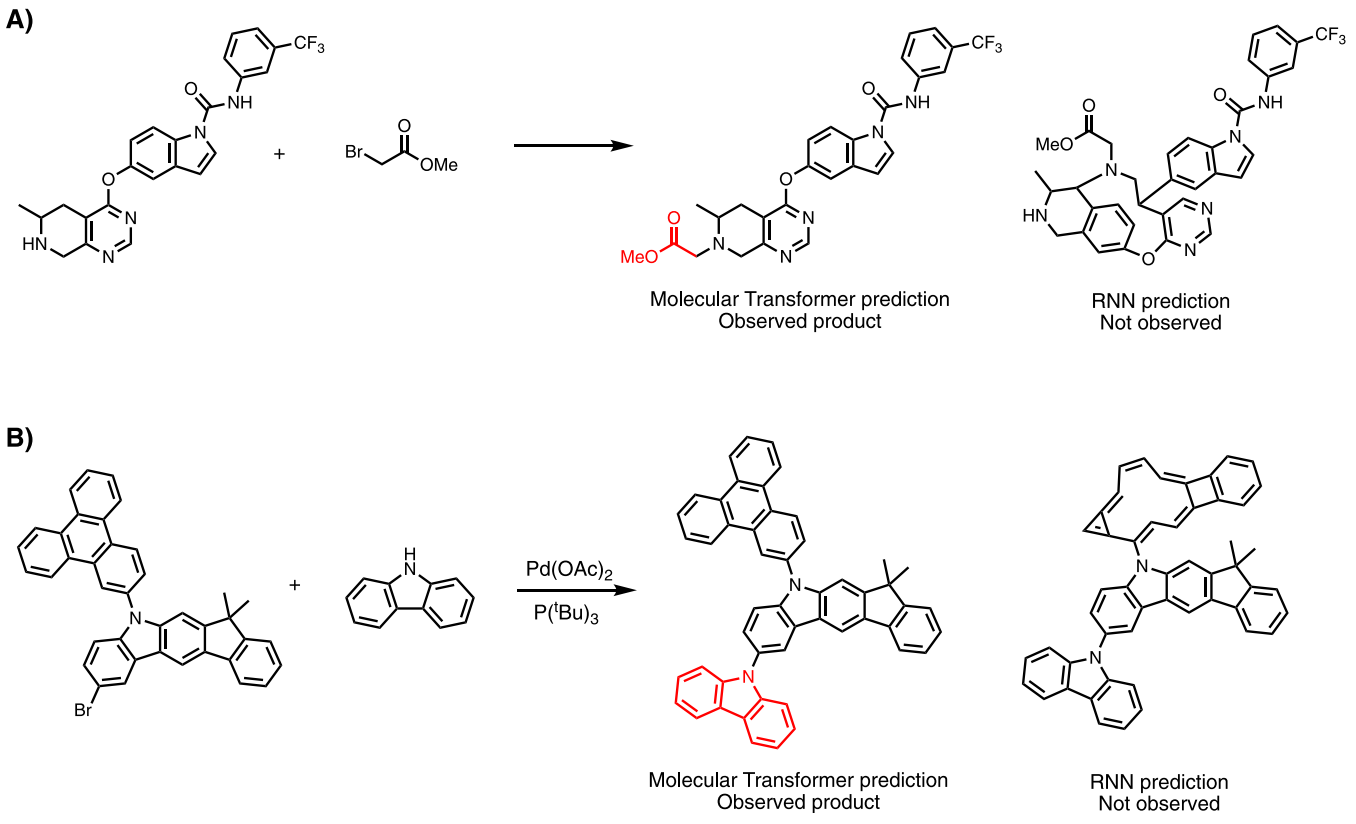


Figure 2. Erroneous inductive bias of the S2S RNN model[28] of assuming that only tokens close together in the SMILES string are chemically related leads to systematic pitfalls for reagents with a long SMILES representation. Molecular Transformer correctly predicts the product for both (A) and (B), whereas the RNN model predicts a product that is not only incorrect but also chemically unreasonable.

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

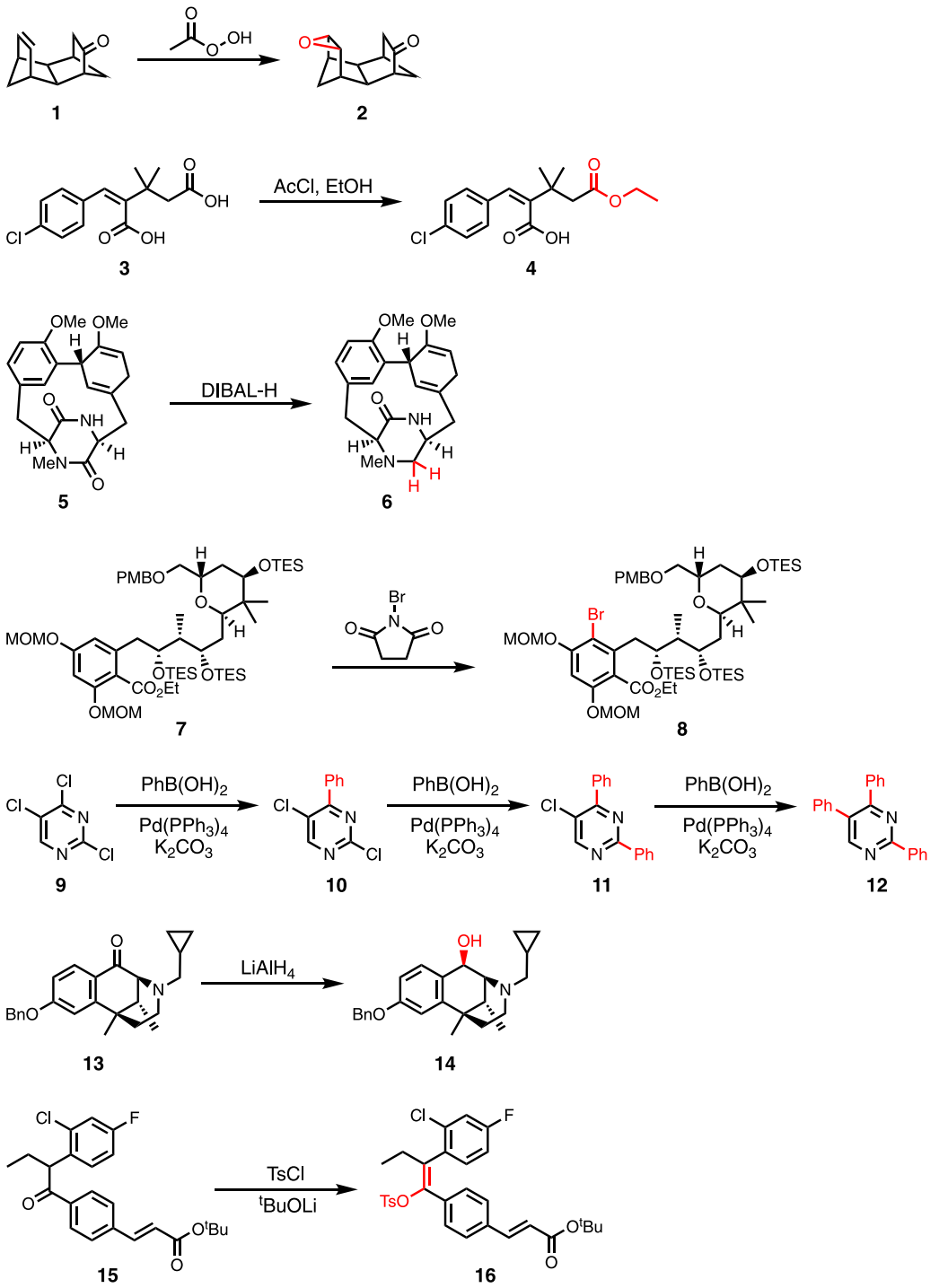


Figure 3. Examples of challenging chemo-, regio-, and stereoselective transformations that Molecular Transformer successfully predicts. Although the figure separates reactants and reagents for clarity, the predictions were done without making this distinction using the model hosted on IBM RXN.[34]

of research. As such, we attempt to address a more limited question: Can Molecular Transformer, trained on diverse reactions harvested from patents, make accurate predictions on a specific class of challenging reactions where the state-of-the-art predictors are quantum -chemistry calculations motivated by physical organic chemistry insights.

To this end, we consider the regioselectivity of electrophilic aromatic substitution reactions in heteroaromatics, a key reaction in medicinal chemistry. Although the reaction mechanism is simple, regioselectivity is controlled by a subtle balance of electronic and steric effects of substituents. We also focus on this reaction because recent pioneering work has systematically curated a large set of examples of halogenation of heteroaromatics from the literature and developed a quantum-chemistry model that quantitatively predicts selectivity,[56] and thus there is a clear benchmark. The state-of-the-art model, RegioSQM,[56] employs quantum-chemistry calculations and achieves a top-1 accuracy of 81% in predicting the site of halogenation. Surprisingly, Molecular Transformer achieves a top-1 accuracy of 83% and top-2 accuracy of 91% on the same

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

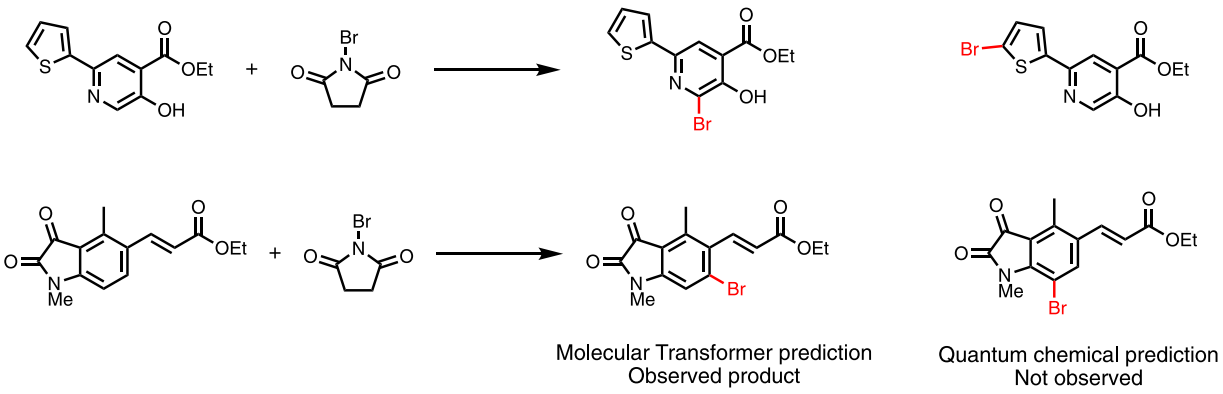


Figure 4. Molecular Transformer achieves a higher accuracy than quantum-chemistry calculations in predicting the regioselectivity of electrophilic aromatic substitution reactions in heteroaromatics. The figure shows examples where RegioSQM,[56] the state of the art, fails, whereas Molecular Transformer makes the correct prediction.

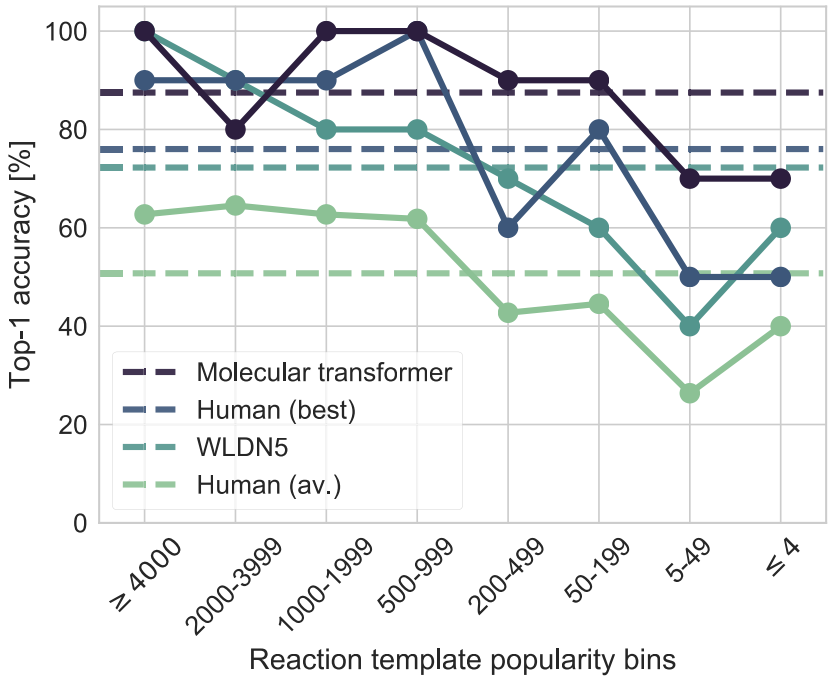


Figure 5. Top-1 accuracy of our model (mixed, USPTO_MIT) on 80 chemical reactions across eight reaction popularity bins in comparison with a human study and their graph-based model (WLDN5).[24]

data set when predicting on the 445 reactions that are not in the training set of the Molecular Transformer and have a single reactive site. Molecular Transformer is also significantly less computationally expensive than quantum-chemistry calculations. Figure 4 shows examples where quantum-chemistry calculations fail to predict the correct site of bromination, whereas Molecular Transformer makes the correct prediction.

The observation that Molecular Transformer correctly predicts those challenging reactions suggests that it might have distilled specific physical chemistry principles from an assortment of diverse reactions, a necessary condition underlying a successful chemical modeling framework.

Comparison with Human Organic Chemists. Coley et al.[24] conducted a study where 80 random reactions from eight different rarity bins were selected from the USPTO_MIT test set and presented to 11 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 data set 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 5, Molecular Transformer is generalizable and remains accurate, even for the less common reactions.

Figure 6 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 the 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

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

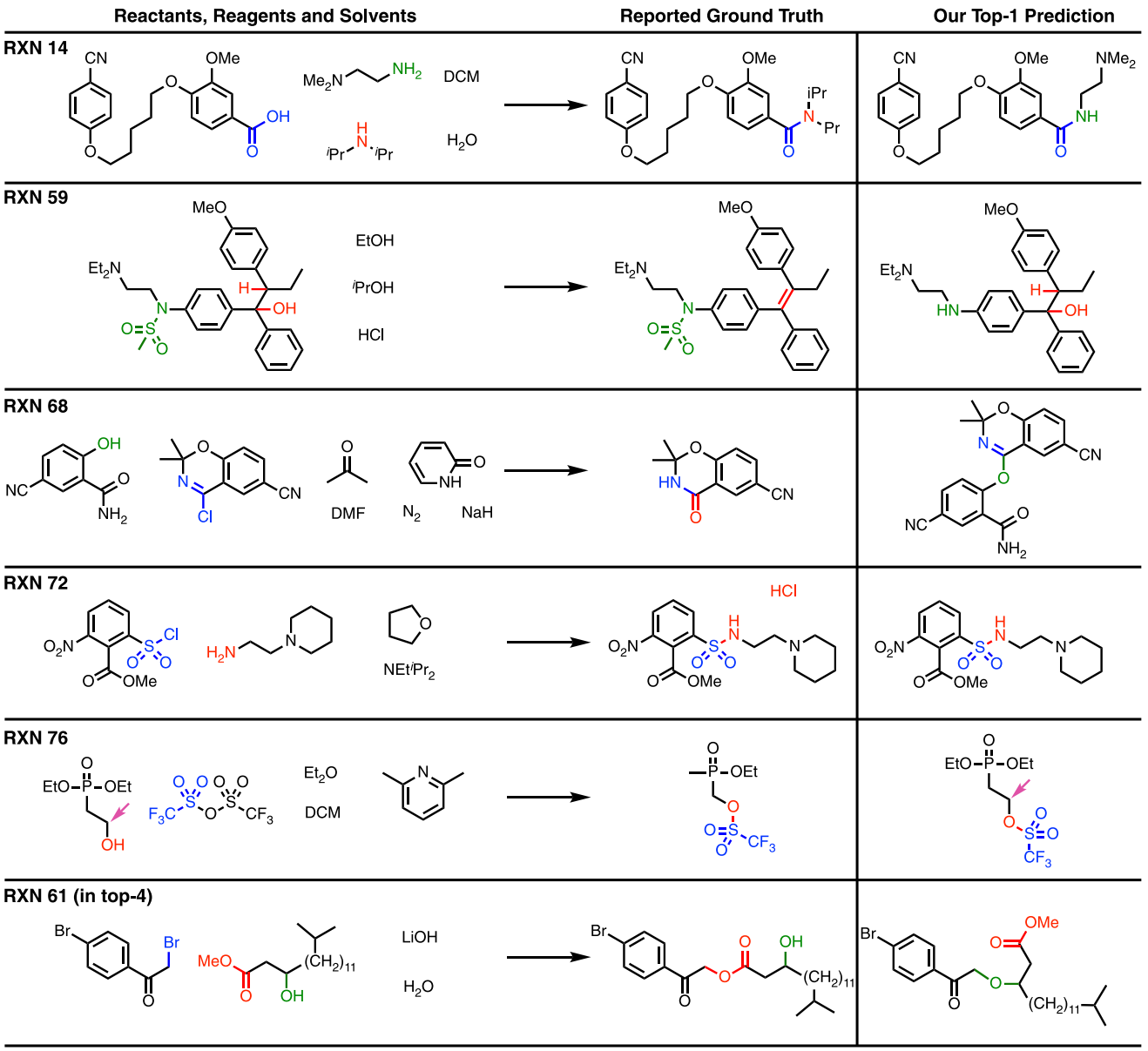


Figure 6. Six reactions in the human test set[24] not predicted within top-2 using our model trained on the augmented mixed USPTO_MIT set.

reaction mixture is the phenolate ion generated from the phenol by deprotonation by sodium hydride. In RXN 72, the correct product is predicted, but the ground truth additionally reports a byproduct (which is mechanistically dubious because 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 S_N2 reaction where the anion of the alcohol of the beta hydroxy ester acts as a nucleophile, whereas the mechanism of the ground truth is presumably ester hydrolysis, followed by the nucleophilic attack of the carboxylate group. Proton transfers in protic solvents are extremely fast, and thus deprotonation of the alcohol OH is much faster than ester hydrolysis. Moreover, the carboxylate anion is a poor nucleophile.

Uncertainty Estimation and Reaction Pathway Scoring. Because organic synthesis is a multistep process, 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 7 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. To obtain the ROC curves, we used a threshold on the confidence score to decide whether a reaction was mispredicted. We counted the predictions that matched the products reported in the patent with a confidence score above the threshold as true-positives (TPs), the predictions that did not match the reported products and were below the threshold as true-negatives (TNs), the predictions that matched the reported products but were below the threshold as false-negatives (FNs), and finally, the predictions that did not match the reported products but were above the threshold as false-positives (FPs). Then, we plotted the false-positive rate (= FP/(FP + TN)) against the true-positive rate (= TP/(TP + FN)) for thresholds between 0.0 and 1.0. Interestingly, Figure 7 reveals that a subtle change in the training method, label

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

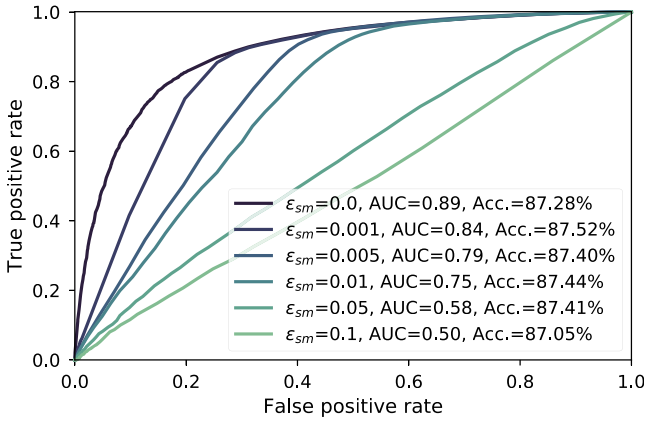


Figure 7. Receiver operating characteristic curve for different label smoothing values for a model trained on the mixed USPTO_MIT data set when evaluated on the validation set.

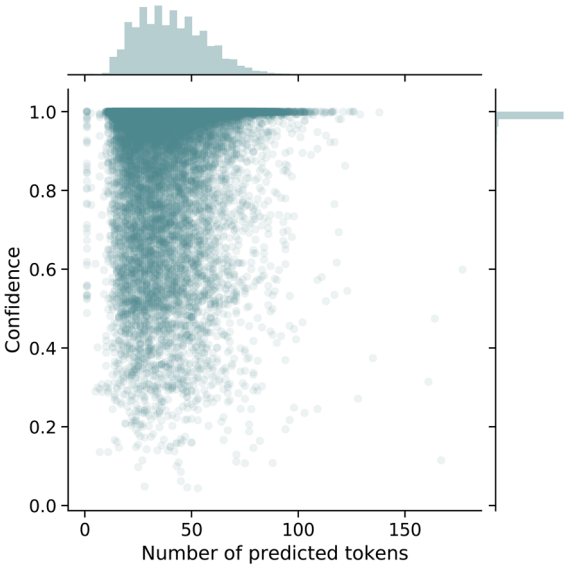


Figure 8. Length of the predicted sequences plotted against the confidence score of the sequence for a model trained on the mixed USPTO_MIT data set with a label smoothing parameter of 0.0. The Pearson product moment correlation coefficient between the length and the confidence score is 0.06.

smoothing, has a minimal effect on the accuracy but a surprisingly significant impact on the uncertainty quantification. Label smoothing was introduced by Vaswani et al.[31] for NMT models. Instead of comparing the output of the model at a given time step during training with a one-hot encoded target vector, label smoothing reduces the mass of the correct token in the target vector and distributes the smoothing mass across all other tokens in the vocabulary. Therefore, the model learns to be less confident about its predictions. Label smoothing helps to generate higher-scoring translations in terms of the accuracy and the BLEU score[57] 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 with 87.28% for no smoothing. However, Figure 7 shows that this small increase in accuracy comes at the cost of no longer being to able to discriminate between a good and a bad prediction. 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 reactant−product combination, rather only predicting products given reactants, and this could be used as a score to rank reaction pathways.[58,59]

Within our uncertainty estimation framework, which is based on the product of probabilities of all predicted tokens, a potential unwanted bias is a bias against long-product SMILES; a large molecule should not necessary imply "difficult" predictions. Figure 8 provides reassuring empirical evidence that this bias is absent. There is no correlation between the confidence score and the length of the SMILES string.

Chemically Constrained Beam Search. Because no chemical knowledge was integrated into the model, technically, the model could perform "alchemy", for example, turning a fluoride atom in the reactants into a bromide atom in the products, which was not in the reactants at all. As such, an interesting question is whether the model has learned to avoid alchemy. To this end, we implemented a constrained beam search, where the probabilities of atomic tokens not observed in the reactants were set to 0.0 and hence not predicted. However, there was no change in accuracy, showing that the model had successfully inferred this constraint from the examples shown during training.

CONCLUSIONS

We show that a multihead 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 data set. 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 has 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 data set, 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[34] and has so far been used by several thousand organic chemists worldwide to perform more than 40 000 chemical reaction predictions.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: phs@zurich.ibm.ch (P.S.).
*E-mail: aal44@cam.ac.uk (A.A.L.).

ORCID

Philippe Schwaller: 0000-0003-3046-6576
Alpha A. Lee: 0000-0002-9616-3108

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

P.S. and A.A.L. acknowledge the Winton Programme for the Physics of Sustainability for funding. We thank G. Landrum, R. Sayle, G. Godin, and R. Griffiths for useful feedback and discussions.

■ REFERENCES

(1) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic Synthesis Provides Opportunities to Transform Drug Discovery. *Nat. Chem.* **2018**, *10*, 383.

(2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(3) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. Expanding the Medicinal Chemistry Synthetic Toolbox. *Nat. Rev. Drug Discovery* **2018**, *17*, 709−727.

(4) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, 2017; pp 1945−1954.

(5) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(6) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian Optimization for Automatic Chemical Design. 2017, arXiv:1709.05501. arXiv.org e-Print archive. https://arxiv.org/abs/1709.05501 (accessed July 29, 2019).

(7) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Science advances* **2018**, *4*, No. eaap7885.

(8) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.

(9) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018; pp 2323−2332.

(10) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43−52.

(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* **2005**, *308*, 395−398.

(12) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23*, 1203−1218.

(13) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(14) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.

(15) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(16) Segler, M. H.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966−5971.

(17) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408−418.

(18) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.

(19) Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wołos, A.; Klucznik, T. Chematica: A Story of Computer Code That Started to Think like a Chemist. *Chem.* **2018**, *4*, 390−398.

(20) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic Reaction Mapping and Reaction Center Detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3*, 560−593.

(21) NextMove Software. *NameRXN*. http://www.nextmovesoftware.com/namerxn.html (accessed July 29, 2019).

(22) 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. *J. Chem. Inf. Model.* **2015**, *55*, 39−53.

(23) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *Advances in Neural Information Processing Systems*, 2017; pp 2604−2613.

(24) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chemical science* **2019**, *10*, 370−377.

(25) Bradshaw, J.; Kusner, M. J.; Paige, B.; Segler, M. H. S.; Hernández-Lobato, J. M. A Generative Model for Electron Paths. In *International Conference on Learning Representations*, 2019.

(26) Do, K.; Tran, T.; Venkatesh, S. Graph Transformation Policy Network for Chemical Reaction Prediction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **2019**, 750−760.

(27) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. 2016, arXiv:1612.09529. arXiv.org e-Print archive. https://arxiv.org/abs/1612.09529 (accessed July 29, 2019).

(28) 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**, *9*, 6091−6098.

(29) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations*, 2015.

(30) Luong, M.-T.; Pham, H.; Manning, C. D. Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* **2015**, 1412−1421.

(31) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems* **2017**, 6000−6010.

(32) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(33) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.

(34) *IBM RXN for Chemistry*. https://rxn.res.ibm.com (accessed July 29, 2019).

(35) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Thesis, University of Cambridge, 2012.

(36) NextMove Software. *Pistachio*. http://www.nextmovesoftware.com/pistachio.html (accessed July 29, 2019).

(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.* **2016**, *59*, 4385−4402.

(38) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56*, 2336−2346.

(39) Landrum, G.; et al. *rdkit/rdkit: 2017_09_1 (Q3 2017) Release*, 2017. https://zenodo.org/record/1004356#.Wd3LDY6l2EI (accessed July 29, 2019).

(40) Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y. N. Convolutional Sequence to Sequence Learning. In *International Conference on Machine Learning*, 2017; pp 1243−1252.

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX

(41) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations* **2017**, 67−72.
(42) *Annotated Transformer*. http://nlp.seas.harvard.edu/2018/04/03/attention.html (accessed July 29, 2019).
(43) Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2016**, 2818−2826.
(44) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.
(45) *Molecular Transformer*. https://github.com/pschwllr/MolecularTransformer (accessed July 29, 2019).
(46) Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. 2017, arXiv:1703.07076. arXiv.org e-Print archive. https://arxiv.org/abs/1703.07076 (accessed July 29, 2019).
(47) Liu, Y.; Zhou, L.; Wang, Y.; Zhao, Y.; Zhang, J.; Zong, C. A Comparable Study on Model Averaging, Ensembling and Reranking in NMT. *CCF International Conference on Natural Language Processing and Chinese Computing* **2018**, *11109*, 299−308.
(48) Woodward, R. B.; Fukunaga, T.; Kelly, R. C. Triquinacene. *J. Am. Chem. Soc.* **1964**, *86*, 3162−3164.
(49) El-Newaihy, M. F.; Salem, M. R.; Enayat, E. I.; El-Bassiouny, F. A. The Condensation of Some Aromatic Aldehydes with Diethyl beta,beta-dimethylglutarate. *J. Prakt. Chem.* **1982**, *324*, 379−384.
(50) Nahmany, M.; Melman, A. Chemoselectivity in reactions of esterification. *Org. Biomol. Chem.* **2004**, *2*, 1563−1572.
(51) He, C.; Stratton, T. P.; Baran, P. S. Concise Total Synthesis of Herqulines B and C. *J. Am. Chem. Soc.* **2019**, *141*, 29−32.
(52) Yu, J.; Yang, M.; Guo, Y.; Ye, T. Total Synthesis of Psymberin (Irciniastatin A). *Org. Lett.* **2019**, *21*, 3670−3673.
(53) Ceide, S. C.; Montalban, A. G. Microwave-assisted, efficient and regioselective Pd-catalyzed C-phenylation of halopyrimidines. *Tetrahedron Lett.* **2006**, *47*, 4415−4418.
(54) Duarte, F. J.; Anand, N. K.; Sharma, P.; Singh, D.; Nektar Therapeutics. Opioid Agonists and Uses Thereof. U.S. Patent 0,022,167 A1, 2017.
(55) Savage, S.; McClory, A.; Zhang, H.; Cravillion, T.; Lim, N.-K.; Masui, C.; Robinson, S. J.; Han, C.; Ochs, C.; Rege, P. D.; Gosselin, F. Synthesis of Selective Estrogen Receptor Degrader GDC-0810 via Stereocontrolled Assembly of a Tetrasubstituted All-Carbon Olefin. *J. Org. Chem.* **2018**, *83*, 11571−11576.
(56) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jorgensen, M. Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chemical Science* **2018**, *9*, 660−665.
(57) Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* **2002**, 311−318.
(58) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System-Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Model.* **1995**, *35*, 34−44.
(59) Segler, M. H.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604.

DOI: 10.1021/acscentsci.9b00576
ACS Cent. Sci. XXXX, XXX, XXX−XXX