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

* A necessary yet unsolved step in planning synthesis is solving the forward problem: Given reactants and reagents, predict the products.
* The researcher treats reaction prediction as a machine translation problem between simplified molecular-input line-entry system (SMILES) strings of reactants, reagents, and the products.
* The researcher’s model that is 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.
* Introduction: - 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.
* Second, reactants seldom contain only one reactive functional group. Predicting subtle reactivity differences is challenging because they are often dependent on the what other functional groups are nearby.
* Third, organic synthesis is almost always a multistep process where one failed step could invalidate the entire synthesis.
* Tackling the synthesis challenges requires methods that are both accurate and have good uncertainty estimates.
* Template-based methods 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.
* 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 problem with the template-based paradigm is that templates themselves are often of questionable validity.
* All automatic template extraction algorithms rely on pre-existing atom mapping, a scheme that maps atoms in the reactants to atoms in the reactant atoms is the reactants to atoms in the product.
* 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 template-based approaches, many template-free methods have used in recent years, this approach can be categorized into graph-based and sequence-based.
* Jin characterizes chemical reactions by graph edits that lead from the reactants to the products.
* There 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.
* A new improved version, where candidates with up to five bond changes are taken into account and multidimensional reactivity matrices are generated.
* Bradshaw 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.
* One side effect of the 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.
* 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 and treat reaction prediction as machine translation from one language to another language.
* Previously seq-2-seq works were based on recurrent neural networks (RNN)s for the encoder and decoder, with one single-head attention layer in between.
* Schwaller et al. used attention weight matrices and confidence scores, that were generated together with the most likely product.
* The researchers show that a fully attention-based model adapted from *Advances in Neural Information Processing Systems* with the SMILES representation, the Molecular Transformer, outperforms all previous methods while being completely atom-mapping independent and not requiring splitting the input into reactants and reagents.
* The researchers model reaches 90.4% top-1 accuracy (93.7% in top-2 accuracy) on a common benchmark data set. It can also 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.
* Data: - the researcher focus on four data sets they are USPTO\_MIT, USPTO\_LEF, USPTO\_STEREO and a nonpublic time-split test set, extracted from the Pistachio database, to compare the performance on a set containing more diverse reactions against a previous seq-2-seq model.



* The *separated* reagent preprocessing means that the reactants (educts), which contribute atoms to the product, are weakly separated by a > token from the reagents.
* Jin increased their top-1 accuracy by 6% when they removed the reagents from the first step, where the reaction centers were predicted. It was shown that the model performs better when the reagents are tagged.
* 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.
* Therefore, the researchers mixed the reactants and reagents into the training and test data set so no distinction was made between the two.
* Molecular Transformer: - the model is based on the transformer architecture. The main architectural difference compared with seq-2-seq models previously used for the reaction prediction, 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.
* 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



* If the query and the key are aligned, their dot product will be large and vice versa. Each key has as 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.
* *dk* is a scaling factor depending on the layer size.
* The main advantage of the transformer architecture compare with seq-2-seq models is the multihead attention, that allows the encoder depending on its preceding outputs.
* The sequential nature of the data is encoded with positional encodings. Positional encodings add position-dependent trigonometric signals to the token embeddings of size *demb* and allow the network to know where the different tokens are situated in the sequences.



* The top-*k* outputs are decoded via a beam search. The researchers decrease the parameter to 12M by going from six layers of size 512 to four layers of size 256.



* The label smoothing parameter is set to 0.0 because it has been observed that a nonzero label smoothing parameter encourages the model to be less confident and therefore negatively effects its ability to discriminate between correct and incorrect predictions.
* The researchers observed that at least four attention heads are required to achieve accuracies. They use eight head and for training they use the ADAM optimizer and varied the learning rate using 8000 warm up steps.
* Results and Discussion: - the researchers double the training set by generating a copy of every reaction in the training set, in which molecules are replaced by an equivalent random SMILES.
* The results also improved by averaging the weights over multiple checkpoints, as well as increasing the training time.
* 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%.
* The top-2 accuracy is significantly higher than the top-1 accuracy, reaching >93% accuracy.
* Comparison with Previous Work – below table shows that the Molecular Transformer clearly outperforms all methods in the literature across the different data sets.



* They find 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 *pN*, where *p* is the probability of a single-step prediction being correct.
* The results also shows that Molecular Transformer is not simply memorizing the data and can leverage information inferred from more common reactions to make predictions on rarer reactions.
* Transformer can process long-range interactions between tokens, whereas RNN models impose the inductive bias that tokens far in sequence space are less related.
* Examples of Chemicals Challenges That Molecular Transformer Tackles – at first the researchers take chemoselectivity, Molecular Transformer predicts acid results in 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.
* Carefully thinking about the role of each reagent and the reactivity of the cyclic anhydride intermediate suggests the esterification of the unconjugated carboxylic acid. It is observed and Molecular Transformer predicts.
* Molecular Transformer can furnish the right product by inferring the reactivity of this complex pattern of distant functional groups. The reduction of 5 using excess DIBAL-H was expected to lead to the unselective reduction of the secondary and the tertiary amides.



* The next challenge is regioselectivity – Molecular Transformer can deal with complicated examples such as the bromination of 7 with *N*-bromosuccinimide, affording 8.
* It can successfully deal with transition-metal-catalyzed reactions as well. It can predict the reactive reactivity of the different C-CI bonds in 2,4,5-trichloropyrimidine 9 in the successive Suzuki coupling reactions with phenylboronic acid.
* At last, they use the Molecular Transformer to predict the stereoselectivity of organic reactions, the Molecular Transformer successfully predicted the fused bicyclic ketone 13 by lithium aluminum hydride gives the major diastereoiser 14.
* It also correctly predicts the formation of the (*E*)-alkene in 16 by the treatment of 15 with tosyl chloride and lithium *tert*-butoxide is also successfully predicted.
* Comparing Molecular Transformer with Quantum-Chemistry-Based Predictors – the researcher attempts to address a limited question that is, 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.
* They consider that the regioselectivity of electrophilic aromatic substitution reactions in heteroaromatics, a key reaction in medicinal chemistry.
* the regioselectivity is controlled by a subtle balance of electronic and steric effects of substituents.
* The state-of-the-art model RegioSQM, employs quantum-chemistry calculations and achieves a top-1 accuracy of 81% in predicting the site of halogenation.
* The Molecular Transformer also achieve top-1 accuracy of 83% and top-2 accuracy of 91% on the same 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.
* It is also less computationally expensive. 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 – the Molecular Transformer 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%).
* Molecular Transformer is generalizable and remains accurate, even for the less common reactions.
* The model does not predict the ground truth, it usually predicts a reasonable most likely outcome: In RXN 14, the researchers model predicts that a primary amine acts as the nucleophile in an amide formation reaction rather than a secondary amine.
* In RXN 68, the reaction yielding the reported ground truth is via a nucleophilic substitution of CI־ by OH־ by the addition-elimination mechanism, followed by lactim-lactam tautomerism.
* In RXN 72, the correct product is predicted, but the ground truth additionally reports a byproduct.
* In RXN 76, a carbon atom is clearly missing in the ground truth, In RXN 61, they predict a SN2 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.
* Uncertainty Estimation and Reaction Pathway Scoring – the organic synthesis takes multistep to process, for a reaction predictor to be useful it must estimate its uncertainty. The Molecular Transformer model provides a natural way to achieve this: the product of the probabilities of all predicted tokens can be used as a confidence score.
* To obtain the ROC curves, they used a threshold on the confidence score to decide whether a reaction was mispredicted.
* They counted the prediction using true-positive, true-negative, false-negative and false-positive.
* The below figure reveals that a subtle change in the training method, label smoothing, has a minimal effect on the accuracy but a surprisingly impact on the uncertainty qualification.



* 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.
* Label smoothing helps to generate higher-scoring translations in terms of the accuracy and the BLEU score for human languages and also helps in terms of reaching top-1 accuracy in reaction prediction.
* The smoothing slightly increases the performance of the model, there no label smoothing was used during the training of their model.
* The uncertainty estimation metric allows them 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.
* 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.
* To this the researchers 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.
* Conclusions: - 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 Molecular Transformer also predict its uncertainty, with a score of 89% in terms of classifying whether a prediction is correct.
* The uncertainty score can be used to rank reaction pathways.
* For the noisy USPTO\_STEREO data set, their top-1 accuracies are 78.1 and 76.2%, respectively.