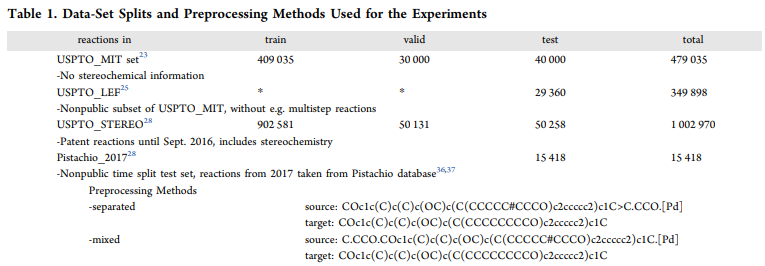
# Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction

* Abstract: - A necessary yet unsolved step in planning synthesis is solving the forward problem: Given reactants and reagents, predict the products.
* The researchers 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.
* 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.
* The researchers model requires no handcrafted rules and accurately predicts subtle chemical transformations. It can also accurately estimate its own uncertainty, with an uncertainty score around 89% accurate in terms of classifying whether a prediction is correct.
* The model also able to handle inputs without a reactant-reagent split and including stereochemistry, which makes researcher’s 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.
* Tackling the challenge of organic synthetic with data-driven approaches is particularly timely as generative models in machine learning for molecules are coming of age.
* There are three salient challenges in predicting chemical reactivity and designing organic synthesis. First, simple combinations would suggest that the space of possible reactions is even greater than the already interactable space of possible molecules.
* 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.
* Third, organic synthesis is almost always a multistep process where one failed step could invalidate the entire synthesis.
* Methods in the literature may be divided into two groups, namely template-based and template-free.
* 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.
* Then it considers 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.
* The problem with the template-based paradigm is the templates themselves are often of questionable validity.
* Recent ML approaches employ template libraries that are automatically extracted from data sets of reactions. 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.
* Atom-mapping is based on template 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 approach many template-free approaches are emerged. Jin et al. characterize chemical reactions by graph edits that lead from the reactants to the products.
* Their reaction prediction is a two-step process. 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.
* The accuracy was significantly improved by excluding the reagents from the reactivity score prediction in the more recent versions. It requires the user to know the identities of the reagents, that means the user already knows the product because the reagent is defined as a chemical species that does not appear in the product!
* Bradshaw et al. separated reactants and reagents and included the reagents only in a context vector for their gated graph neural network.
* A side effect of phrasing reaction prediction as predicting electron flow is that a preprocessing step must be applied to eliminated 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 on language to another language.
* First 2 seq-2-seq methods based on recurrent neural networks (RNNs) for the encoder and the decoder, with one single-head attention layer in between.
* To increase the interpretability of the model, 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 the research “Attention Is All You Need. *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 researcher’s model has 90.4% top-1 accuracy 93.7% in top-2 accuracy on a common benchmark data set. Their 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.
* The researchers’ model 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 researchers focus on four data sets. The USPTO\_MIT data set was filtered and split by Jin et al. The USPTO\_LEF are also used with USPTO\_STEREO dataset.
* The USPTO\_STEREO dataset underwent less filtering, and the stereochemical information was kept. Stereochemistry adds an extra level of complexity because it requires the models to predict not only molecular graph edge changes but potentially also changes in node labels.
* They also use the 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.



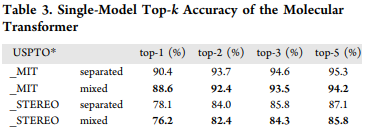
* In the above table the overview of the data sets used in the researchers’ work and points out the two different preprocessing methods. The *separated* reagent preprocessing means that the reactants which contribute atoms to the product, are weakly separated by a > token from the reagents.
* The reagents take part in the reaction but do not contribute any atom to the products. In most of the work the reagents have been separated from the reactants.
* 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.
* Therefore, the researchers decided the to train and test on inputs where the reactants and the reagents were mixed and no distinction was made between the two.
* Molecular Transformer: - 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 is that the RNN components 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 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.
* 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



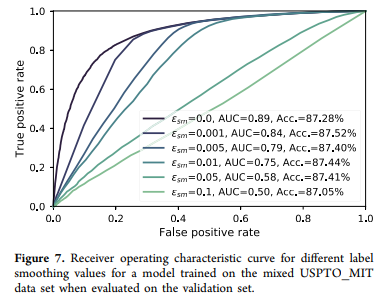
* The dot product of the queries and the keys computes how closely aligned the keys are with the queries. If the query and the key are aligned, then their dot product will be large and vice versa.
* Each key has an associated value vector, which is multiplied by the output of the softmax, through which the dot products were normalized and the largest components were emphasized.
* is a scaling factor depending on the layer size.
* One main advantage of the transformer architecture compared with the seq-2-seq models used in research “The Neural Machine Translation and the Prediction of Organic Chemistry Reactions” and “Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-To-Sequence Models” is the multi-head attention, which allows the encoder and decoder to peek at different tokens simultaneously.
* Positional encodings add position-dependent trigonometric signals to the token embeddings of size and allow the network to know where the different tokens are situated in the sequence.



* The top-*k* outputs are decoded via a beam search. The researchers set the beam size of 5 for all of the experiments.
* The base transformer model had 65 M parameters; the researchers decreased the number of trainable weights to 12 M by going from six layers of size 512 to four layers of size 256.
* A nonzero label smoothing parameter encourage the model to be less confident and therefore negatively affects its ability to discriminate between correct and incorrect predictions.
* They also absorbed that only 4 attention heads were required to achieve peak accuracies.
* For training the researchers used the ADAM optimizer and varied the learning rates 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.
* Results and Discussion: - the researchers double the training data by generating a copy of every reaction in 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, it also increases the training time.
* Ensembling different models is known to increase the performance of NMT models; however, the performance increase is marginal compared with parameter averaging.
* Ensembling two models that contain the weight average of 20 checkpoints two models that contain the weight average of 20 checkpoints of two independently initialized training runs lead to a top-1 accuracy of 91%. Whereas a higher accuracy and better uncertainty estimation can be obtained by model ensembles, they come at additional cost of training or test time.



* Comparison with Previous Work – because all previous works used single models, the researchers also consider only single models trained on the data-augmented versions of the data sets rather than ensembles for remainder of the research they done, to have fair comparison.
* Molecular Transformer clearly outperforms all methods in the literature across the different data sets. It also able to outperform when the reactant and reagents are mixed.
* The researchers 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 by , where *p* is the probability of a single-step prediction being correct.
* The researchers compare their model with Coley et al. although Coley had separated the reagents in their experiment, even though the researchers model able to outperform the across all popularity bins, even their model predict on a mixed reactants-reagents input, and the accuracy gap becomes larger as the template popularity decreases.
* The Molecular Transformer outperforms in virtually every single reaction class. This is because the multihead attention layer in the Molecular Transformer can process long-range interactions between tokens, whereas RNN models impose the inductive bias that tokens far in sequence space are less related.
* S2S RNN model have pitfall because of its erroneous inductive bias of assuming that only tokens close together in the SMILES string are chemically related.
* Even 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 in the SMILES representation.
* Examples of Chemical Challenges That Molecular Transformer Tackles – 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.
* 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.
* Molecular Transformer can help design new synthesis, ultimately saving hours of human labor in the laboratory.
* Molecular Transformer can deal with complicated examples such as the bromination of 7 with *N*-bromosuccinimide, affording 8.
* 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.
* Comparing Molecular Transformer with Quantum-Chemistry-Based Predictors - Molecular Transformer has inferred the physical principles that underline chemical selectivity.
* The common question of distilling interpretable rationales from ML models are still an active area of researcher/study.
* The researchers try to find/address more limited questions: like, Can Molecular Transformer, trained on diverse reactions harvested from patents, make accurate predictions on a specific class of challenging reaction where the state-of-the-art predictors are quantum -chemistry calculations motivated by physical are quantum insights.
* The researchers consider the regioselectivity of electrophilic aromatic substitution reactions in heteroaromatics, a key reaction in medicinal chemistry.
* They state-of-the-art model, RegioSQM, employs quantum-chemistry calculations and achieves a top-1 accuracy of 81% in predicting the site of halogenations. However, the Molecular Transformation also achieves 83% accuracy in top-1 and 91% in top-1 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.
* Molecular Transformer also has significantly less computational expenses than quantum-chemistry calculations.
* The observation that Molecular Transformer correctly predicts those challenging reactions suggest 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 researcher 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%).
* Molecular Transformer is generalizable and remains accurate, even for the less common reactions.
* Even the researchers’ model does not predict the ground truth, it usually predicts a reasonable most likely outcome: In RXN 14, their 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.
* 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.
* Uncertainty Estimation and Reaction Pathway Scoring – to estimate the uncertainty of the Molecular Transformer has its own way that is the model provides a natural way: the product of the probabilities of all predicted tokens can be used as a confidence score.
* The ROC-AUC curve is 0.89 and if the researchers use this confidence score as a threshold to predict whether a reaction is mispredicted.
* To obtain the ROC curves, the researchers used a threshold on the confidence score to decide whether a reaction was mispredicted.
* The researchers plotted the false-positive rate (=FP/(FP+TN)) against the true positive rate (=TP/(TP+FN)) for threshold between 0.0 and 1.0.
* Below figure shows that a subtle change in the training method, label smoothing, has a minimal effect on the accuracy but a surprisingly significant 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 generates higher-scoring translations in terms the accuracy and the BLEU score for human languages and also helps in terms of reaching higher top-1 accuracy in reaction prediction.
* Above figure shows that the small increase in accuracy comes at the cost of no longer being too able to discriminate between a good and a bad prediction.
* Therefore, the researchers do not use any smoothing during the training of their model. The AUC-ROC of their single mixed USPTO\_MIT model measured on the test set was also at 0.89.
* The uncertainty estimation metric allows the researchers to estimate the likelihood of a given reactant-product combination, rather only predicting products given reactant-product combination, rather only predicting products given reactants, and this could be used as a score to rank reaction pathways.
* Within their 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 necessarily imply “difficult” predictions.
* 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.
* It raises an interesting question that is whether the model has learned to avoid alchemy.
* The researchers implemented a constrained beam search, where the probabilities of atomic tokens not observed in the reactants were set of 0.0 and hence not predicted.
* Conclusion: - the researchers how that Molecular Transformer outperforms all the other methods/algorithms. It requires on handcrafted rules and accurately predicts subtle chemical transformations.
* Molecular Transformers can also accurately estimate its own uncertainty, with an uncertainly score that is 89% accurate in terms of classifying whether a prediction is correct.
* They also demonstrated that an accuracy of 88.6% when no distinction is drawn between reactants and reagents in the inputs, and for the noisier USPTO\_STEREO data set, its top-1 accuracies are 78.1% and 76.2% respectively.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee