**Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network**

* The prediction of organic reaction outcomes is a fundamental problem in computational chemistry that is which products form as a result of a chemical reaction.
* A reaction may involve hundreds of atoms, current solution utilizes reaction templates to limit the space, but it suffers from coverage and efficiency issues.
* In this paper the researchers proposed a template-free approach to efficiently explore the space of product molecules by first pinpointing the reaction center - the set of nodes and edges where graph edits occur.
* Only small number of atoms contribute to reaction center; therefore, we can directly enumerate candidate products. The generated candidates are scored by a Weisfeiler-Lehman Difference Network that models high-order interactions between changes occurring at nodes across the molecule.
* Experimentation remains the primary manner in which reaction outcomes are analyzed. It is time consuming, expensive, and requires the help of an experienced chemist.
* The empirical approach is particularly limiting for the goal of automatically designing efficient reaction sequences that produce specific target molecule(s), a problem known as chemical retrosynthesis.
* The researchers propose to formulate the reaction prediction task as a graph transformation problem after viewing molecules as labeled graphs over atoms.
* A chemical reaction transforms input molecules (reactants) into new molecules (products) by performing a set of graph edits over reactant molecules, adding new edges and/or eliminating existing ones.
* The computational challenge is how to reduce the space of possible edits effectively, and how to select the product from among the resulting candidates.
* The solution is “reaction templates”, a reaction template specifies a molecular subgraph pattern to which it can be applied and the corresponding graph transformation.
* Because multiple templates can match a set of reactants, another model is trained to filter candidate products using standard supervised approaches.
* Its drawback of this approach is coverage and scalability.
* A large number of templates is required to ensure that at least one can reconstitute the correct product.
* These templates are either hand-crafted by experts or generated from reaction databases with heuristic algorithms.
* Coley extracts 140K unique templates from a database of 1 million reactions.
* Applying a template involves graph matching and it makes examining large numbers of templates prohibitively expensive.
* The researchers propose a template-free approach by learning to identify the *“reaction center”*, a small set of atoms/bonds that change from reactants to products.
* On average only 5.5% of the reactant molecules directly participate in the reaction.
* The researchers’ forward-prediction approach is divided into two key parts: (1) learning to identify reaction centers and (2) learning to rank the resulting enumerated candidate products.
* The researchers technical approach builds on neural embedding of the *Weisfeiler-Lehman* isomorphism test.
* They incorporate a specific attention mechanism to identify reaction centers while leveraging distal chemical effects not accounted for in related convolutional representations.
* They also propose a novel *Weisfeiler-Lehman Difference Network* to learn to represent and efficiently rank candidate transformations between reactants and products.
* The researcher performs their model performance in two datasets derived from USPTO and compare it with current top performing systems. They find that their system achieves 83.9% and 77.9% accuracy on these two datasets, which outperform the baseline approach by 10% while running 140 times faster.
* Template-based Approach – existing ML models for product prediction are mostly built on reaction templates. This approach differ in the way templates are specified and, in the way, the final product is selected from multiple candidates.
* Wei learns to select among 16 pre-specified, hand-encoded templates, given fingerprints of reactants and reagents.
* Segler and Waller and Coley use a data-driven approach to obtain a large set of templates, and then employ a neural model to rank the candidates.
* The difference between these models are the representation of the reaction.
* In Segler and Waller, molecules are represented based on their Morgen fingerprints, while Coley represents reactions by the features of atoms and bonds in the reaction center.
* The template-based architecture limits both of these methods in scaling up to larger datasets with more diversity.
* Template-free Approach – Kayala also presented a template-free approach to predict reaction outcomes. The researchers approach differs from theirs approach in several ways.
* First, Kayala operates at the mechanistic level - that is identifying elementary mechanistic steps rather than the overall transformations form reactants to products.
* Most reactions consist of many mechanistic steps, their approach requires multiple predictions to fulfill an entire reaction.
* The researchers’ model operates at the graph level-predicting transformation from reactants to products in a single step.
* Second mechanistic descriptions of reactions are not given in existing reaction databases. The Kayala created their training set based on a mechanistic-level template-driven expert system.
* The researchers’ model is learned directly from real-world experimental data.
* Third kayala uses feed-forward neural network where atoms and graphs are represented by molecular fingerprints and additional hand-crafted features.
* The researcher’s approach builds from graph neural networks to encode graph structures.
* Molecular Graph Neural Networks – the molecular graph representation is a key issue in reaction modeling.
* In computational chemistry, molecules are often represented with Morgan Fingerprints, Boolean vectors that reflect the presence of various substructures in a given molecule.
* Duvenaud developed a neural version of Morgen Fingerprints, where each convolution operation aggregates features of neighboring nodes as a replacement of the fixed hashing function.
* It is further expanded by Kearnes into graph convolution models. Dai consider a different architecture where a molecular graph is viewed as a latent variable graphical model. It is a recurrent model which is derived from Belief Propagation-like algorithms.
* Gilmer generalized all previous architecture into message-passing network, and applied them to quantum chemistry.
* Lei models is closest to the researchers Weifeiler-Lehman Network model, the Lei model derived from Weisfeiler-Lehman Kernel that produces isomorphism-invariant representations of molecular graphs.
* The researchers approach bypass’s reaction templates by learning a *reaction center identifier.*
* The researchers train a neural network that operates on the reactant graph to predict a reactivity score for every pair of atoms.
* After that a reaction center is selected by picking a small number of atom pairs with the highest reactivity scores.
* After identifying the reaction center, they generate possible product candidates by enumerating possible bond configurations between atoms in the reaction center subject to chemical constraints.
* They also train another neural network to rank these product candidates so that the correct reaction outcome is ranked highest.
* A chemical reaction is a pair of molecular graphs *(Gr, Gp)*, where *Gr* is called the *reactants* and *Gp* the *products.*
* A molecular graph is described as *G = (V, E),* where *V =* {a1, a2, …, an} is the set of atoms and *E =* {b1, b2, …, bm} is the set of associated bonds of varying types.
* The *Gr* is has multiple connected components since there are multiple molecules comprising the reactants.
* The reaction used for training are *atom-mapped* so that each atom in product graph has unique corresponding atom in the reactants.
* A reaction center is a set of atoms pairs {(ai, aj)}, where the bond type between ai and aj differs from Gr to Gp.
* A reaction center is a minimal set of *graph edits* needed to transform reactants to products.
* Reaction Center Identification – In a given reaction *R = (Gr, Gp)*, each atom pair (au, av) in Gr is associated with a reactivity label yuv {0,1} specifying whether their relation differs between reactants and products.
* The label is determined by comparing *Gr* and *Gp* with the help of atom-mapping. The researchers predict the label on the basis of learned atom representations that incorporate contextual cues from the surrounding chemical environment.
* The WLN is inspired by the Weisfeiler-Lehman isomorphism test for labeled graphs. This architecture is designed to embed the computations inherent in WL isomorphism testing to generate learned isomorphism-invariant representations for atoms.
* WL Isomorphism Test – the idea of isomorphism test is to repeatedly augment node labels by the sorted set of node labels of neighbor nodes and to compress these augmented labels into new, short labels.
* In each iteration, its label is augmented with the element labels of its neighbors. Like a multi-set label is compactly represented as a new label by a hash function.
* Here be the final label of atom av. The molecular graph *G =* (V, E) is represented as a set {(, ) | (u, v) E}, where buv is the bond type between u and v.
* The number of distinct labels grows exponentially with the number of iterations *L*.
* WL Network – The discrete relabeling process does not directly generalize to continuous feature vectors.
* The researchers appeal to neural networks to continuously embed the computations inherent in the WL test. In this let *r* be the analogous continuous relabeling function. After that a node G with neighbor nodes *N(v)*, node features fv, and edge features fuv is “relabeled” according to



* Here (.) could be any non-linear function. The researchers apply this relabeling operation iteratively to obtain context-dependent atom vectors



* Here = fv and U1, U2, V are shared across layers. The final atom representations arise from mimicking the set comparison function in the WL isomorphism test, yielding



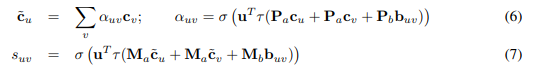
* In the above equation the set comparison is realized by matching each rank-1 edge tensor to a set of reference edges also cast as rank-1 tensors where W[k] is the k-th row of matrix W. In other words, Wq. 3 above could be written as



* The resulting cv is a vector representation that captures the local chemical environment of the atom and involves a comparison against a learned set of reference environments.
* The representation of the whole graph *G* is simply the sum over all the atom representations: **c**g = .
* The researchers present two model to predict reactivity: the *local* and *global* models.
* The researchers’ local model is based directly on the atom representation cu and cv in predicting label yuv.
* The global model, on the other hand, selectively incorporates distal chemical effects with the goal of capturing the fact that atoms outside of the reaction center may be influenced by certain *reagents*.
* The researchers incorporate these distal effects into the global model through an attention mechanism.
* Local Model let cu, cv be the atom representation for atoms u and v, respectively, as returned by WLN. The researcher predicts the reactivity score of (u, v) by passing these through another neural network:



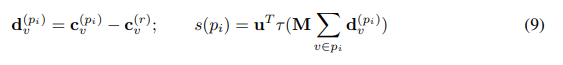
* In the above equation the is the sigmoid function, and buv is an additional feature vector that encodes auxiliary information about the pair such as whether the two atoms are in different molecules or which type of bond connect them.
* Global Model in this model let be the attention score of atom *v* on atom *u*. The global context representation u of atom *u* is calculated as the weighted sum of all reactant atoms where the weight comes from the attention module:



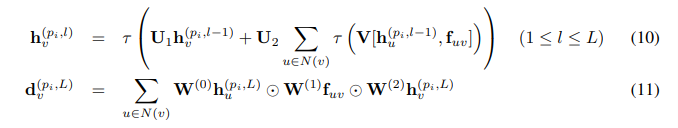
* The attention is obtained with sigmoid rather than softmax non-linearity since there may be multiple atoms relevant to a particular atom *u*.
* Training both models are trained to minimize the following loss function:



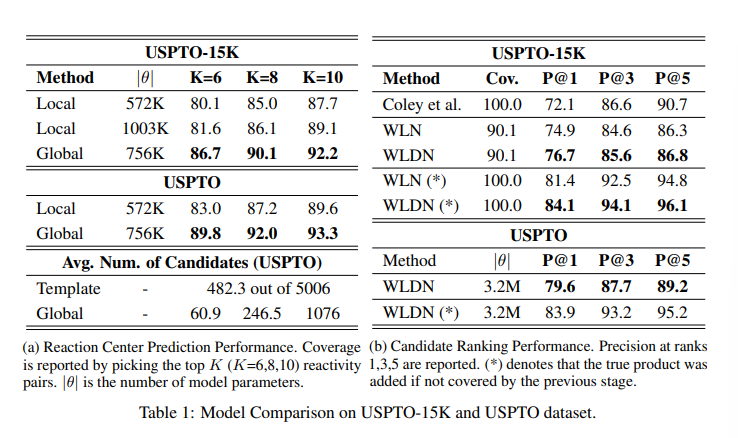
* Here the researchers predict each label independently because of the large number of variables.
* The researchers find that given reaction with *N* atoms, need to predict the reactivity score of *O*(*N2*) pairs.
* The quadratic complexity prohibits them from adding higher-order dependencies between different pairs. They also found that independent prediction yields sufficiently good performance.
* The researchers select the top *K* atom pairs with the highest predicted reactivity score and designate them, collectively, as the reaction center.
* The set of candidate products are then obtained by enumerating all possible bond configuration changes within the set.
* Because every atom has a maximum number of neighbors they can connect to *(valence constraint)*. They also leverage the statistical bias that reaction centers are very unlikely to consist of disconnected components *(connectivity constraint)*.
* The training set for candidate ranking consists of lists = {(r, p0, p1, …, pm)}, where r are the reactants, p0 is the known product, and p1, …, pm are other enumerated candidate products.
* The goal is to learn a scoring function that ranks the highest known product p0.
* We need to learn the represent (r, p) in a manner that can focus on the key difference between the reactants *r* and products *p* while also incorporating the necessary chemical contexts surrounding the changes.
* The researchers again propose two alternative models to score each candidate pair (*r, p*).
* The first model naively represents a reaction by summing difference vectors of all atom representations obtained from a WLN on the associated connected components.
* The second and improved model, called WLDN, takes into account higher order interactions between these differences’ vectors.
* WLN with Sum-Pooling be the learned atom representation of atom *v* in candidate product molecule pi: the researchers define *difference vector*  pertaining to atom *v* as follows:



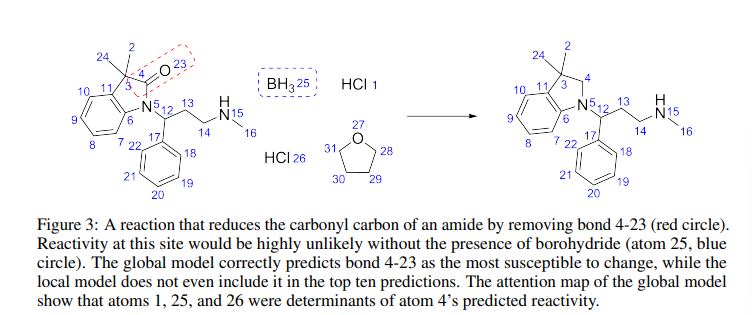
* The reactants and products are atom-mapped so the researchers can use *v* to refer to the same atom. The pooling operation is a simple sum over these difference vectors, resulting in a single vector for each (r, pi) pair.
* This vector is then fed into another neural network to score the candidate product pi.
* Weisfeiler-Lehman Difference Network (WLDN) simply summing all difference vectors, the WLDN operates on another graph called a *difference graph.* A difference graph *D* (r, pi) is defined as a molecular graph which has the same atoms and bonds as pi, with atom *v*’s feature vector replaced by .
* The benefits of using different graphs are first, in *D* (r, pi), atom *v*’s feature vector deviates from zero only if it is close to the reaction center, thus focusing the processing on the reaction center and its immediate context.
* Second, *D* (*r, pi*) explicates neighbor dependencies between difference vectors. The WLDN maps this graph-based representation into a fixed-length vector, by applying a separately parameterized WLN on top of *D* (r, pi):



* Where = . The final score of pi is *s*(pi) = uT(M.
* Training – both models are trained to minimize the softmax log-likelihood objective over the scores {s(p0), s(p1), …, s(pm)} where s(p0) corresponds to the target.
* The researcher uses reactions from USPTO granted patents.
* After filtering the data, they obtained a set of 480K reactions, to which they refer.
* They divide these datasets into 400K, 40K and 40K for training, development, and testing purposes.
* For comparison purposes they report the result on the subset of 15K reaction from the dataset used by Coley.
* They follow the Coley’s split, with 10.5K, 1.5K, and 3K for training, development, and testing.
* Setup for Reaction Center Identification – the output of this component consists of *K* atom pairs with the highest reactivity scores.
* The researchers compute the *coverage* as the proportion of reactions where all atom pairs in the true reaction center are predicted by the model.
* Atom-level features include its elemental identity, degree of connectivity, number of attached hydrogen atoms, implicit valence, and aromaticity.
* Bond-level features include bond type (i.e., single, double, triple, or aromatic), whether it is conjugated, and whether the bond is part of a ring.
* The researcher both model that is global and local model are build upon the Weisfeiler-Lehman Network, where the unrolled depth is 3, the models are optimized with Adam, with learning rate decay factor is 0.9.



* Setup for Candidate Ranking – it is used to determine whether the model can select the correct product from a set of candidates derived from reaction center.
* The researchers compare model accuracy against the top-performing template-based approach developed by Coley.
* This approach employs frequency-based heuristics to construct reaction templates and then uses a neural model to rank the derived candidates.
* For this experiment, the researchers set *K =* 8 for candidate generation, which achieves 90% coverage and yields 250 candidates per reaction.
* For comparison of a standard WLN representation against its counterpart with difference Networks (WLDN), they train them under the same setup on USPTO-15K, fixing the number of parameters to 650K.
* The researcher evaluates their model on USPTO for large scale evaluation. They set the *K* = 6 for candidate generation and report the result of the best model architecture.
* For factorization of the coverage of candidate selection and the accuracy of candidate ranking, they consider two evaluation scenarios: (1) the candidate list as derived from reaction center; (2) the above candidate list augmented with the true product if not found.
* Reaction Center Identification – the coverage depends on the number of atom pairs *K*, with the higher coverage for large values of *K*. These results demonstrate that even for *K* = 8, the model achieves high coverage, above 90%.
* This result demonstrate that global model achieves better performance and consistency across all experiments compare to local model.
* It proves the well-known fact that reactivity depends on more than the immediate local environment surrounding the reaction center.
* The presence of certain *functional groups* far from the reaction center can promote or inhibit different modes of reactivity.
* Reactivity is often influenced by the presence of *reagents*, which are separate molecules that may not directly contribute atoms to the product.
* In the below figure we can see that observed reactivity can be attributed to the presence of a reagent molecule that is completely disconnected from the reaction center itself. The local model fails to anticipate this reactivity, the global one accurately predicts the reaction center. The attention map highlights the reagent molecule as the determinant context.



* In the above table 1a shows that for *K* = 6, the researchers model generates an average of 60.1 candidates and reaches a coverage of 89.8%.
* The template-based baseline requires 5006 templates extracted from the training data to achieve 90.1% coverage with an average of 482 candidates per example.
* Weakness of the baseline model can be explained by the difficulty in defining general heuristics with which to extract templates from reaction examples.
* The different levels of specificity based on the extent to which atoms surrounding the reaction center are included or generalized.
* It introduces an unavoidable trade-off between generality (fewer templates, higher coverage, more candidates) and specificity (more templates, less coverage, fewer candidates).
* The massive number of templates required for high coverage is a serious impediment for the template approach because each template application requires solving a subgraph isomorphism problem. It takes on average 7 seconds to apply the 5006 templates to a test instance, while the researcher model takes less than 50 ms, about 140 times faster.
* The baseline templates were optimized on the test and have 100% coverage, the researchers compare its performance against their models to which the correct product added.
* The researchers’ model outperforms the baseline in top-1 accuracy. It demonstrates that the WLDN model consistently outperforms the WLN model.
* It is consistent with researchers’ intuition that modeling higher order dependencies between the difference vectors is advantageous over simply summing over them.
* The researchers group the test instances according to their frequency and report the coverage of the global model and the mean reciprocal rank (MRR) of the WLDN model on each of them.
* They find that their approach achieves the highest performance for frequent reactions. It also maintains reasonable coverage and ranking accuracy even for rare reactions.
* They randomly selected the 80 reaction examples from the test set, ten from each of the template popularity and then asked 10 chemists to predict the outcomes of each given reactants. The average accuracy across the ten chemists was 48.2% and the model achieves the accuracy of 69.1%.
* The researcher proposed a novel template-free approach for chemical reaction prediction.
* The researcher first predicts a small set of atoms/bonds in reaction center, and then produce candidate products by enumerating all possible bond configuration changes within the set.
* The researchers’ both models that is reaction center identifier and candidate ranking model build from Weisfeiler-Lehman Network and its variants that learn compact representation of graphs and reactions.