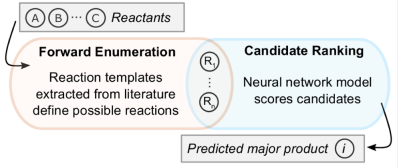
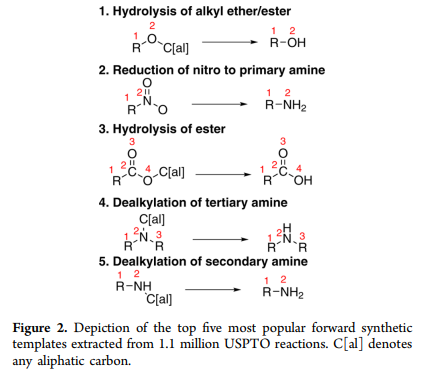
# Prediction of Organic Reaction Outcomes Using Machine Learning

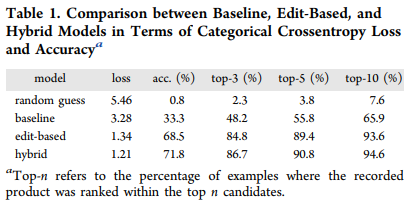
* Abstract: - one critical challenge in developing high-quality pathway suggestions is that proposed reaction steps often fail when attempted in the laboratory, despite initially seeming viable.
* The researchers report a framework for anticipating reaction outcomes that combines the traditional use of reaction templates with the flexibility in pattern recognition afforded by neural networks.
* Candidate reactions are represented using a unique edit-based representation that emphasizes the fundamental transformation from reactants to products, rather than the constituent molecule’ overall structures.



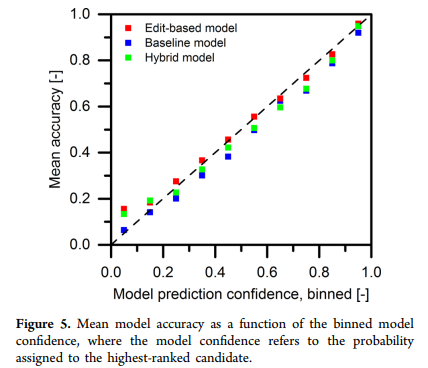
* Introduction: - the computer-aided synthesis design was introduced over 40 years ago to help the chemists in the form of retrosynthesis planning software.
* It was first introduced by Coley and Wipke, Coley approach us to codifying the retrosynthesis involved the explicit identification of molecular structures which lend themselves to disconnection or, rather can be produced by known reaction templates – sub-molecular patterns that encode changes in atom connectivity.
* Applying retrosynthesis templates to a target molecule produces a candidate synthesis tree. However, a synthetic route based on retrosynthesis templates does not always lead to a successful forward synthesis.
* Templates are locally defined pattern-matching rules, inherently naïve to what is present in the rest of the molecule.
* Forward analysis is such an important part of pathway evaluation that even the very first retrosynthesis program, Corey’s LHASA could identify functional group conflicts which might lead a lack of specificity or selectivity.
* Other programs like SPOHIA and Eros identify potentially reactive functional groups using manually curated reactivity rules and empirical calculations.
* Manual encoding of these rules has some disadvantages. First, it relies on the intuition and experience of a small number of chemists.
* Second, it is not scalable-it is not realistic to exhaustively define the full substrate scope and incompatibilities for every possible reaction.
* Third, conflicting reactivity is rarely black and white; incompatibility depends on the exact nature of the reacting molecules.
* The use of graph-based representation of molecules and assign approximate molecular orbitals to each so that a mechanistic step can be considered an interaction between a donor and an acceptor orbital.
* Even, it shows great result, the need of manual encoding of mechanistic rules to generate training data could be problematic and limit scalability.
* Wei at el. describe the use of NNs to predict the outcome of reactions based on reactant fingerprints, but limit the outcomes of reactions based on reactant fingerprints, it limits their study to 16 types of reactions covering a very narrow scope of possible alkyl halide and alkene reactions.
* Segler and Waller describe two approaches to forward synthesis prediction. First, is a knowledge-graph approach that uses the concept of half reactions to generate possible products given exactly two reactants participates.
* The second approach uses neural networks to rank reaction templates given reactant fingerprints.
* The researchers’ approach is highly biased toward reactions with high yields, and reactions with negligible or zero yields are rarely reported except for illustrative purpose, e.g., highlighting the necessity of a catalyst.
* One cannot train a model directly on literature data to classify a certain reaction as productive or unproductive, since there are almost no unproductive examples available.
* The researchers describe a model that learns to predict the major products of chemical reactions given a set of reactant molecules by combining rigid reaction templates and machine learning.
* The researchers report the following contributions: (1) a data augmentation strategy whereby reaction databases are supplemented with chemically plausible negative reactions examples;
* (2) the successful application of that strategy using automatically extracted forward synthesis templates, where poor specificity is not a hindrance and no manual curation is required;
* (3) a new reaction representation focused on the fundamental transformation at the reaction site rather than constituent reactant and product fingerprints;
* (4) the implementation and validation of a neural network-based model that learns when certain modes of reactivity are more or less likely to occur than other potential modes.
* Approach: - Overview – the researchers’ model predicts the outcome of a chemical reaction in a two-step manner: (1) applying overgeneralized forward reaction templates to a pool of reactants to generate a set of chemically plausible products, and
* (2) estimating which candidate product is the major product as a multiway classification problem using machine learning.
* In the first way the researchers apply a library of forward synthetic templates to define which products could be produced based on the initial reactants.
* In the second stage, each candidate reaction is scored individually by the machine learning model. This model assesses from competing reactions. The scores from all candidates are compared in a softmax network layer to generate probabilities describing which product is predicted to be most abundant.
* A key component of their approach is this two-step formalization. In this first step, existing reaction databases are augmented with negative reaction examples. This circumvents the limitations of only having high-yielding reaction data.
* Including the alternate products in the training set allows the researchers to extract more information from each reaction entry.
* The recorded product of a reaction in the patent database is the “true” product that the model learns to predict, while the chemically plausible alternate products generated via templates are the “false” products which were not reported in the literature.
* Data – forward templates were extracted from the prefiltered set of 1,122,662 atom-mapped reaction SMILES in 1976-2013\_USPTOgrants\_reactionSmiles\_feb2014filters.rsmi. For training and test major product identification, a 15 000-member subset of the full set of reactions in 1976-2013\_USPTOgrants\_SML.7z was used.
* No two reaction examples in the data set have an identical reactant pool.
* Forward Enumeration – to build the database of forward templates the researchers use a heuristics-driven algorithm inspired by Law et al. and Bogevig et al.
* The reaction core is expanded to include adjacent unmapped leaving groups and immediately neighboring atoms. Neighboring atoms are fully generalized into any non-hydrogen substituent for maximal generality to achieve high coverage at the expense of low specificity.
* The five most popular templates are shown in below figure with *ex post facto* labels. Although these five happen to be unimolecular functional group conversions, the researchers have not predefined common functional groups or common transformations; moreover, the model does not rely on manual curation, labeling, or sorting of these extracted templates.
* Application of templates is computationally expensive, and the marginal coverage benefit of including additional templates decreases rapidly with rank. To focus the model on the most prevalent reaction types, only templates with more than 50 precedents were included in subsequent steps; this corresponds to the top 1689 templates.



* The removal of annotations labeling species’ roles was motivated by the fact that they were originally assigned using information about the recorded product.
* The candidate products were each reduced to the product fragment with the longest SMILES strings to neglect any byproduct salts and approximate the “major product”.
* The inclusion of reagents, catalysts, and solvents makes the task of product prediction more challenging due to the competing candidate products they produce.
* The 15 000-member data set used for training and testing consists only of reactions where the recorded product was found within the candidate set generated by this template set.
* Prior to the subsequent candidate ranking step, during training, the researchers determine which candidates match the “true” recorded reaction outcome.
* When multiple candidates match the recorded product, the candidate corresponding to the most popular template is kept and the remaining matching candidate are discarded so only one “true” candidate exists.
* Candidate Ranking – the parametrized reaction scoring function, a central component of the workflow is challenging to design for many reasons, one of which is representation.
* A reaction is a complex data structure for which there is no universal vector-based description.
* An atom-mapped reaction candidate is parsed into four different types of edits:
  + An atom loses a hydrogen
  + An atom gains a hydrogen
  + Two atoms, and lose a connecting bond
  + Two atoms, and gain a connecting bond
* Changes in bond order are marked as a loss of the original bond order and a gain of the new bond order.
* Loss or gain of a hydrogen is represented by 32 easy-to-compute features of the reactant atom alone, .
* Loss or gain of a bond is represented by a concatenation of the features of the atoms involved and four features of the bond .
* Because edits occur at the reaction center by definition, the overall representation of a candidate reaction depends *only* on the atoms and bonds at the reaction core.
* The researchers added rapidly calculable structural and electronic features of the reactants’ atoms that reflects the local chemical environment first and foremost, but also reflect the surrounding molecular context.
* The design of the neural network is motivated by the likelihood of a reaction being a function of the atom/bond changes that are required for it to occur.
* Individual edits are first analyzed in isolation using a neural unique to the corresponding edit type. Three fully connected dense layers are used to embed initial features into an intermediate features vector representation.
* The intermediate features vectors of all edits are summed and passed through a final neural network to produce a scalar score.
* The absolute likelikhood scores of all candidates are compared in a final softmax layer, which produces a vector of probabilities from a vector of numbers by testing their values as pseudoenergies in a Boltzmann distribution with.
* When trained with this architecture, the four distinct neural networks become tailored to assessing their corresponding edit type.
* Layer size were fixed before the cross-validation tests after initial screening to ensure sufficiently flexibility to describe the training data.
* All the hidden layers are fully connected with bias and tanh activation. The final model required 144 001 parameters.
* The researchers also implement a baseline model, which attempts to rank candidate products based on the products alone; no consideration is given to the reactants or corresponding reaction edits.
* Product molecules are represented by radius-2 Morgan circular fingerprints of length 1024 for this model.
* One hidden layer with tanh activation is used prior to the linear output layer. The baseline scoring model is shown in S5 with its 51 301 parameters.
* At last, the researchers implement a hybrid model, which train the full edit-based and baseline architectures simultaneously and uses the sum of their scores for each candidate reaction.
* Result: - the model was trained and tested using a 5-fold cross-validation with the Adadelta optimizer and early stopping.
* Each fold used a 70%***/***10%***/***20% training/validation/testing split and creased training once the validation loss did not improve the five epochs.
* The edit-based model achieves a test accuracy of 68.5%, averaged across all folds.
* The hybrid model, combining the edit-based representation with the proposed products’ fingerprint representations, achieves an accuracy of 71.8%.



* The actual objective during training is minimization of the categorical cross entropy loss, -log *p*( the average of the negative natural logarithm of the probability assigned to the true candidate.
* This reflects both model accuracy and model confidence, the edit-based model (1.34) is vastly superior to the baseline model (3.28); the hybrid model (1.21) offers an additional improvement.
* The true outcome was assigned a near-zero probability in a majority of examples; in comparison, the edit-based model exhibits a more favorable distribution shifted higher probabilities – the hybrid model even more so.
* The distribution of assigned ranks is short-tailed in the edit-based and hybrid models and long-tailed in the baseline model. The shape of this tail affects overall model success as the success criterion is relaxed from rank = 1 to rank n.
* It is important to understand the significance of the probability the model assigns to each candidate.
* The below figure depicts the performance of each model as a function of that model’s “confidence”. In this context, confidence refers to the probability assigned to the high-ranked candidate.
* There is very strong correlation of the model’s accuracy with the prediction confidence; this signifies that the probability assigned to a candidate product does indicate the actual likelihood of that being the major product; this is a highly desirable characteristic for a predictive model.
* Prediction Examples – analysis of individual predictions gives insight into model behavior than statistical measures.
* In the recorded outcome, a methyl ester is hydrolyzed to form a carboxylic acid. On the same substrate, there is a *t*-butyldimethyl silyl ether, which is a commonly employed alkoxy protecting group highly resistant to hydrolysis.



* The outcomes are assigned similar probabilities, however, indicating that the model believes either of these two outcomes to be likely.
* It highlights the importance of capturing sterics in candidate representations; in researchers’ model atom-level featurization, two relevant attributes are included: the Labute Approximate Surface Area contribution and the Total Polar Surface Area contribution.
* Input Feature Analysis – Average values for each set of indices are calculated from the true candidates of the 10 500 reaction examples in the training set.
* The resulting decrease in model performance measures how strongly the trained model relies on values of those attributes to discriminate candidates, although it does not reflect how the model might have developed had that attribute been missing during training.
* The worst performance observed when masking a single input feature only represents a 5.2% decrease in accuracy.
* For loss and gain of hydrogen, the researchers see that the model relies on “functional” descriptors for evaluation, Estate index, Crippen contribution to molar refractivity (MR), and the Gasteiger partial charge.
* Loss and gain of a bond mainly dependent on the “structural” features describing the atomic number, number of neighbors, and number of hydrogens.
* Aromaticity is another important feature for assessing bond gain, likely due to the prevalence of aryl halide coupling reactions.
* Discussion: - Forward Template Quality – An inherent shortcoming in the automatic extraction of reaction templates is their locality.
* To address the issue of locality, Soh et al. describe an analysis of functional group occurrence in reaction examples to infer which groups promote/inhibit certain types of reactivity.
* Extracted templates are also highly reliant to atom-mapping to describe the correspondence between reactant and product atoms.
* The researchers’ approach, inaccurate atom-mapping is a less significant issue, as their templates are designed to be over general to achieve high coverage of potential product species. The overall model performance does not depend strongly on atom mapping quality.
* The coverage afforded by the forward templates affects the generalizability of the overall two-step model: for a yet-unseen reaction example, the model must be able to identify the true product as a candidate; otherwise, the true product cannot be evaluated by the neural network.
* The researchers’ model can evaluate *any* candidate reaction, even if the corresponding template and/or substrates have never been seen before. They also do not observe a strong dependence of model performance on the number of candidates.
* Candidate Reaction Representation – the majority of existing machine learning algorithms require fixed-length vectors or one-dimensional sequences as inputs.
* Reaction could be represented as (a) the concatenation of reactant and product fingerprints (b) the difference between reactant and product fingerprints, previously used for a reaction classification task, or (c) some other representation with greater focus on the reaction core.
* In general, the performance of a neural network depends strongly on the choice of input representation.
* These types of molecule-level representation of reactants are insufficient for the types of real, multifunctional between different combinations of substructures must be considered.
* The edit-based reaction representation has its own flaws, from researchers’ analysis of which features the trained model is most dependent on, they find that the “functional” descriptors obtained through rapid molecule-level calculations are important in addition to the structural descriptors.
* Context Awareness – the current model assesses candidates only in terms of the molecules that contribute atoms to the final product, i.e., the reaction core.
* In its current form, the lack of context means that the model weights reaction candidates under implicitly defined typical conditions.
* The unavailability of contextual information also means that the model learns to limit the confidence with which it makes some predictions.
* Reaction within a specific family is often reported under similar conditions, so extrapolation to atypical conditions would be challenging. Further work on context-aware models must address the question of how to train on and extrapolate from this limited data.
* Reagent representation, a subproblem of context-awareness, poses other significant challenge. Representing a highly reactive reagent, e.g., a brominating agent, as a fingerprint or through a generic molecular representation is a poor use of information.
* From the reaction literature the researchers find/identify the most popular agents for different additions, but it is not clear how to apply that knowledge in a scalable and fully automated way and capture that historical information in a vector representation.
* Conclusion: - using both a novel model framework for generating and ranking candidate reaction outcomes and a novel edit-based representation, the researchers are able to reproduce in silico the qualitative results of actual experimental reactions.
* Through expanded atom – and bond-level featurization of the reaction core, especially descriptors with more direct relevance to chemical reactivity, this model framework can be expanded upon to achieve even higher predictive performance.
* There is a tremendous role for machine learning to play in computer assisted synthesis design, not only as a key component of automated retrosynthesis planning, but as a standalone tool for chemists to assess reaction viability.

Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green and Klavs F. Jensen