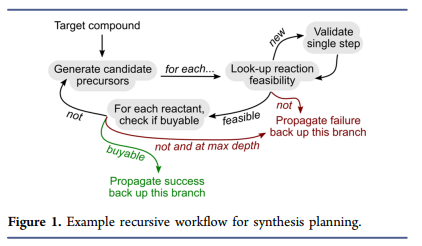
# Machine Learning in Computer-Aided Synthesis Planning

* Conspectus: - computer-added synthesis planning (CASP) is focused on the goal of accelerating the process by which chemists decide how to synthesis small molecule compounds.
* The ideal CASP program would take a molecular structure as input and output a sorted list to detailed reaction schemes that each connect that target to purchased starting materials via a series of chemically feasible reaction steps.
* Now it is possible to construct and validate purely data-driven approaches to synthesis planning. As a result, synthesis planning has been opened to machine learning techniques, and the field is advancing rapidly.
* In this research the researchers focus on two critical aspects of CASP and recent machine learning approaches to both challenges.
* First, they discuss the problem of retrosynthesis planning, which requires a recommender system to propose synthetic disconnections starting from a target molecule.
* They describe how the search strategy, necessary to overcome the exponential growth of the search space with increasing number of reaction steps, can be assisted through a learned synthetic complexity metric.
* They also describe how the recursive expansion can be performed by a straightforward nearest neighbor’s model that makes clever use of reaction data to generate high quality retrosynthesis disconnections.
* Second, they introduce this task in the context of reaction validation, its utility extends to the prediction of side products and impurities, among other applications.
* Machine learning is rapidly transforming CASP, but there are several remaining challenges and opportunities, many pertaining to the availability and standardization of both data and evaluation metrics.
* Introduction: - what is synthesis planning – Synthesis planning is the process of determining how to synthesize a chemical compound from available starting materials through a series of chemically feasible reactions steps.
* E.J. Corey approaches this problem in reverse: beginning with the product and choosing suitable disconnections recursively.
* Anatomy of a CASP Tool – it consists of five major components:
  + A template library containing the rules by which disconnections are proposed.
  + A recursive template application engine that generates candidate reactants for target product molecules.
  + A database containing compounds that do not need to be expanded retrosynthetically (e.g., are commercially available).
  + A strategy to guide the retrosynthesis search toward chemicals in that database.
  + A method for single-step or pathway-level scoring for example, a preference for fewer synthetic steps.



* Role of machine learning – the machine learning techniques have the ability to approximate complex functions where the exact relationship between input and output variables is not easily codified.
* It is generally attributed to a combination of improvements in hardware capability and in data availability.
* Focus – the researchers focus on two major domains within organic synthesis where machine learning can be and has been applied:
  + Retrosynthesis, synthesis route planning
  + Forward synthesis, prediction of reaction outcomes.
* Machine learning is more broadly useful for inferring nonobvious relationships within high-dimensional data, which is increasingly relevant to chemistry as the field’s ability to rapidly generate high-fidelity experimental data improves.
* Route Planning: - Introduction – the researchers summarize contemporary strategies grouped into three categories: template library-based, template-free and focused template application.
* Template library-based – template library-based retrosynthesis involves matching generalized reaction rules to target molecules to yield one or more candidate precursors.
* Contemporary approaches use algorithms template extraction from atom-mapped reaction examples.
* An extracted rule must contain atoms that change connectivity, but the degree to which auxiliary atoms are included is flexible.
* There is an inevitable trade-off between specificity and speed: including too many neighboring atoms leads to large, poorly generalized template libraries, which are computationally expensive to apply; too few neglects the necessary context and leads to unfeasible disconnections.
* The researchers find that using simple heuristic provides an appropriate balance. They also remove the rare templates because it will make the computational expensive.
* Application of the full template library to a target produces candidate precursors, often numbering in the 100s or 1000s.
* Careful handling of stereochemistry is required for faithful preservation, inversion, or destruction of tetrahedral centers and cis/trans chirality, e.g., using the open-source RDChiral wrapper for RDKit.
* To avoid the combinatorial explosion, recursive expansion must focus on only the most promising disconnections that yield easily synthesizable compounds.
* Chematica uses user-definable Chemical Scoring Functions (CSFs): algebraic functions of molecular parameters (e.g., 2). The SA\_Score is a more sophisticated fragment-contribution method that scores rare structural motifs as complex. An alternative to these state-agnostic metrics is the use of proof-number search.
* Unfortunately, these heuristics are not suited to retrosynthesis due to the nonequivalence of *structural* complexity and *synthetic* complexity.
* The retro-deprotections are always considered “uphill” moves, because the molecule becomes larger, when they may actually be “downhill” that is productive.
* The researchers develop the SCScore as a data-driven metric designed to described real syntheses, modeled after the premise that the products of published reactions should, on average, be more *synthetically* complex than each of their reactants.
* A feed-forward neural network model was constructed to computer a synthetic complexity score from an extended-connectivity fingerprint (ECFP) and trained using a hinge loss function to promote separation between scores assigned to the molecules in each pair.
* With this pairwise ranking objective, the model implicitly learns what structures and motifs are more prevalent as reactants.
* Template-Free – it is an alternative attractive for several reasons. First, calculating subgraph isomorphism is computationally expensive, especially subgraph isomorphism is computationally expensive, especially for large libraries.
* Second, the degree of template generality/specificity can lead to either low-quality or incomplete recommendations.
* And third and last is that the template-based methods cannot propose fundamentally novel disconnections.
* Retrosynthesis requires the prediction of reactant molecules. Liu et al. sequence-to-sequence model convert a product SMILES to reactant(s) SMILES. While no significant benefit in accuracy, the ability to propose the true, recorded reactants as high-ranked suggestions, over template library application was reported, neural translation has been successfully applied to the inverse problem of reaction prediction, suggesting that straightforward model improvements might improve its performance.
* Focused Template Application – rather than forego templates entirely, focused template methods select *relevant* templates to apply, mitigating the computational expense of full library application.
* It does not overcome the question of generalization during templates extraction, however, nor the need to exclude rare templates.
* To overcome these challenges, the researcher devised a strategy for retrosynthesis expansion based on the concept of molecular similarity reminiscent of a nearest neighbor model.
* The researchers first calculate the structural similarity between the target and all known products to recall relevant precedents from a reaction corpus.
* For any resulting precursor sets, their structural similarity to the reaction precedent’s reactants is calculated and multiplied by the product similarity to provide an overall score.
* Similarity-based scoring implicitly considers potential functional group conflicts or missing activating groups. This approach outperforms the template-free seq2seq model and extends to full route planning when applied recursively.
* Prediction of Reaction Outcomes: - Introduction – advancements in data-driven retrosynthesis planning have helped avoid manual curation of template libraries and improve computational speed.
* An additional benefit is increased confidence in template applicability without explicitly encoding reactivity conflicts.
* An important goal of recommender systems is this reduction of false positive: proposed reaction incorrectly thought to be chemically feasible.
* To validate retrosynthesis suggestions, one can solve the inverse problem of forward synthesis: given specific experimental parameters, what is the product distribution?
* It limitations are the absence of detailed concentration information in reaction databases necessitates simplification to only reactants and reagents, perhaps also including catalysts, solvents, and reagents, perhaps also including catalysts, solvents, and temperature.
* The other limitation is the absence of side-product information; prediction of the full product distribution must be recast as prediction of the major project.
* Classifying reaction feasibility – Segler et al. describe such a neural network model that classifies reactions as true or false based on their fingerprint representations, trained using true experimental data augmented by synthetic negative data.
* Predicting Mechanistic Steps – the baldi group approaches reaction prediction with a mechanistic perspective. Their ReactionPredictor identifies electron sources and sinks, enumerates possible interactions, and ranks those interactions using graph-based representations of molecules with pseudomolecular orbital considerations.
* Its limitation is the need to manual encode mechanistic rules for artificial data generation.
* Ranking Template – Wei et al. develop a proof-of-concept for using machine learning to predict template applicability. Their model, given reactant and reagents, predicts which of the 16 rules is most relevant, using simulated data generated by these rules.
* Segler and Waller extended this approach, but neither approach directly predicts the major product, as application of the highest-ranked template may yield several nonequivalent product species, for example, when site selectivity is ambiguous as in a halogenation where multiple aromatic C-H bonds could be activated.
* Ranking Products – the researchers work on forward prediction began with a hybrid system that directly predicts product(s) of chemical reactions, rather than templates.
* It has two framework that is template-based forward enumeration and machine learning-based candidate ranking.
* This two-approach overcome the bias toward positive reaction examples. The alternate action outcomes that most help the model learn chemical reactivity are those that seem chemically plausible.
* Applying the reaction templates defines the *scope* of chemically valid outcomes; they are overgeneralized to increase product coverage at the expense of specificity.
* Due to limited performance using standard fingerprints, the researchers devised an “edit-based” representation to enable richer, more explicit encoding of prior chemical knowledge.
* Reactions are represented by atom- and bond-level features of atoms gaining/losing hydrogen atoms and pairs of atoms gaining/losing bonds.
* The features vectors constitute the inputs to a feedforward neural network model, which embeds each edit separately before sum-pooling their latent representations and further transforming that combined representation into one aggregated score.
* The model was trained to maximize the log-probability assigned to the true major product.
* The researchers’ study was the first large-scale demonstration of product prediction using experimental data, but predictions cannot be made outside the template library scope, and using templates for data augmentation limits scalability.
* Generating Products – To overcome the drawbacks of the hybrid template/neural network approach, the researchers’ study forwent templates and used a trainable model for candidate enumeration in addition to ranking.
* Molecules are represented as attributed graphs, with atom- and bond-level features as before. Atom features are iteratively embedded using a Weisfeller-Lehman Network (WLN), a graph convolutional neural network, which incorporates information from neighboring atoms into each atoms’ representation.
* Pairwise reactivity scores are calculated for each atom pair based on their feature vectors, quantifying the propensity of that interaction to change.
* The model is trained to predict which pairs of atoms belong to the reaction center. Top pairs are used to combinatorically enumerate possible bond changes; structural and valence requirements restrict candidates to chemically valid molecules.
* The fully learned approach achieves significantly higher accuracy than the template-based approach and operates order of magnitude faster, enabling its application to a large data set of ca. 400,000 reactions.
* Applying translation models to reaction prediction was first demonstrated by Nam and Kim, who trained a sequence-to-sequence model to predict product SMILES strings from reactant SMILES using a combination of real and synthetic data, tested only on textbook problems.
* Language-based models have unique limitations that warrant mention. Mistakes may be linguistic, rather than chemical, so the second-ranked outcome is not necessarily the next-most chemically plausible.
* The model predictions are of the next SMILES character/token to append to the current sequence, not of reactive atoms and bonds, so inferring quantitative reactivity trends requires additional analysis and atom-to-atom mapping; in contrast, template-based and graph-based model can be directly analyzed via scores assigned to different modes of reactivity or interactions to understand what is perceived as chemically likely.
* Outlook: - Data availability – an important collection is USPTO dataset. Elsevier’s commercial Reaxys database comprises a much larger collection of reactions extracted from the chemical literature.
* However, its most entries do not contain atom mapping or information about reaction conditions or yield, nor is any attempt made to include, concentrations or equivalence ratios, despite these being specified in the original articles.
* Data standardization and competition-style challenges would accelerate further development of this field.
* Data Sharing – the literature bias toward reporting only successful reactions, access to negative reaction data would present tremendous opportunities for future research.
* Failed reaction data has proved useful in materials discovery and could offer greater insight into chemical reactivity than can synthetic negative reaction data.
* Data Applicability – for outcome prediction, existing data is insufficient to accurately capture effects of different reaction conditions. Synthetic negative examples necessarily use the same reactants and conditions as positive examples; this can lead to models that mimic patterns in the data instead of learning actual reactivity trends.
* Reaction run under atypical conditions, even those with poor yields, provide complementary information to those run under standard conditions.
* Evaluation – standardization of evaluation is essential. Forward prediction models have a natural evaluation metric: in the list of predicted outcomes, how highly is the true (recorded) outcome ranked?
* For retrosynthesis, quantitative evaluation is challenging because experiments are required for definitive proof of success. As a single-step evaluation, an effective program should propose and rank highly the reactants of published reactions when their products are treated as target compounds.
* Programs should also not be evaluated on whether or how quickly pathways are found, as this is orthogonal to the quality of suggestions and neglects consideration of proposals feasibilities.
* Reliance on human chemists is problematic in terms of standardization and scalability.
* There is no best pathway: a discovery-focused medicinal chemist might want a disconnection that enables the synthesis of several diverse analogues, while a process chemist might care more strongly about cost, toxicity of side products, E-factors, etc.
* Interpretability – the increased performance associated with deep neural networks compared to simpler models often comes at the cost of interpretability.
* The approaches to rationale extraction in other domains benefit from fundamental changes to network architectures.
* Acceptance and Adoption – many chemists still not accept ML and see it with skeptical. To overcome this requires not only developing useful tools but also communicating the benefits and drawbacks of machine learning. Basic misconceptions can lead to inaccurate claims about limitations of its utility.
* Conclusion – there are opportunities for the application of ML and AI techniques to organic synthesis and synthesis planning, including countless applications beyond those mentioned in research.
* The potential payoff for computer-aided synthesis planning is higher than ever. It has a very good future and the researchers hope one day a fully autonomous synthesis: a true realization of the “robo-chemist” will made.
* But we also need to overcome the few researchers concerns.

Connor W. Coley, William H. Green and Klavs F. Jensen