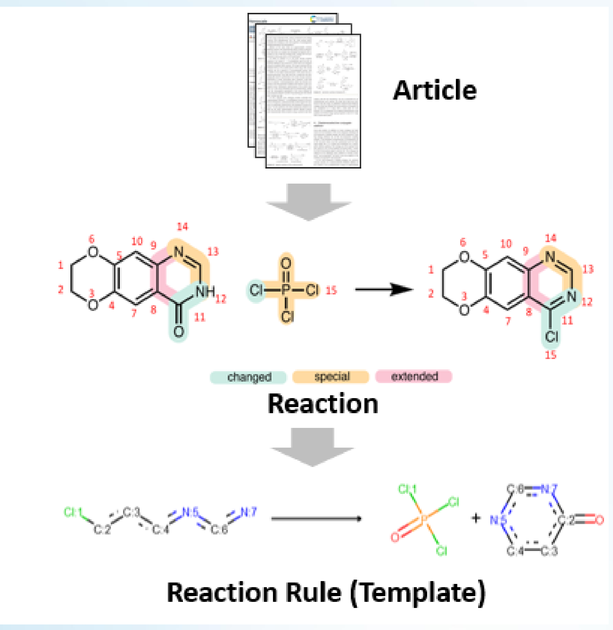
Undergraduate Research Opportunities Program

Deep Learning in Synthesis Planning

Student Name:	MA Wing Yin Timothy	**Project Supervisor:**	Prof. Su Haibin
Major / School:	CHEM	**Dept. / School :**	CHEM

This project seeks to integrate machine learning into chemistry to create a "synthesis planning shop (SPLASH)" to benefit chemists and the public in seeking practical and cost-effective pathways to synthesis all sorts of molecules, from medicines to polymers. SPLASH is a retrosynthesis program with machine learning capabilities, it utilizes the brain-inspired Artificial Neural Networks (ANN), a machine learning algorithm called Monte Carlo Tree Search (MCTS), also notations for chemical molecules, called SMILES and SMARTS and other notations to let the program comprehend and extract templates from the reactions fed into the system, currently the SPLASH system have over 4 million reactions and 200 thousand templates.

This project also seeks to understand the general trend and correlations in Nickel-catalyzed carbon-carbon cross coupling reactions, through compiling a database of reaction data and running PCA and TMAP with chemical descriptors, such as electronegativity, cone angle and buried volume that may contribute huge impacts in the reactivity of the reactants and ligands.

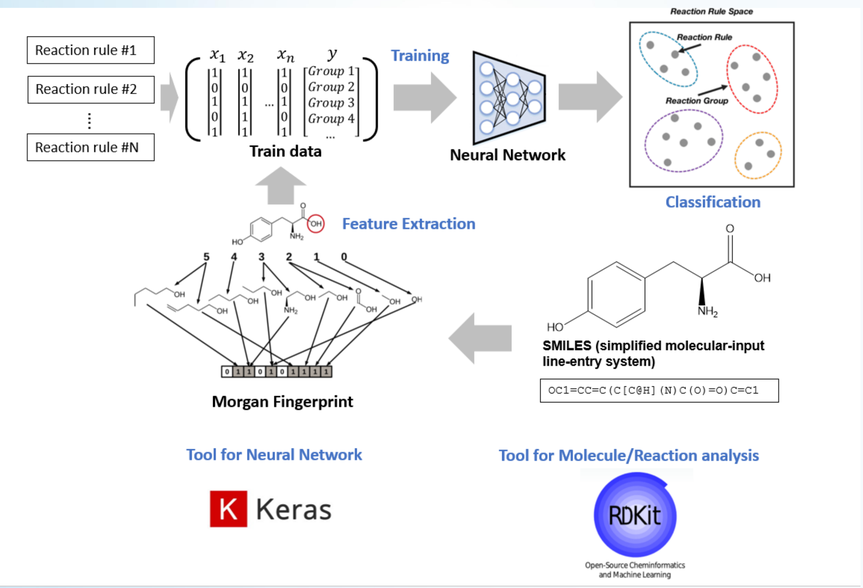


Automatic Data Extraction and Classification of reactions by Machine Learning and Neural Network

From organic chemistry related literature and patents, the reactions within the figures and tables will be extracted automatically through a self-developed Data extraction tool into SMILES, a language that denotes the chemical structure of the molecules within the reaction. SMILES will be translated into SMARTS, which is SMILES with atom-mapping that tracks the changes occurred to the reactants after the reaction has finished.

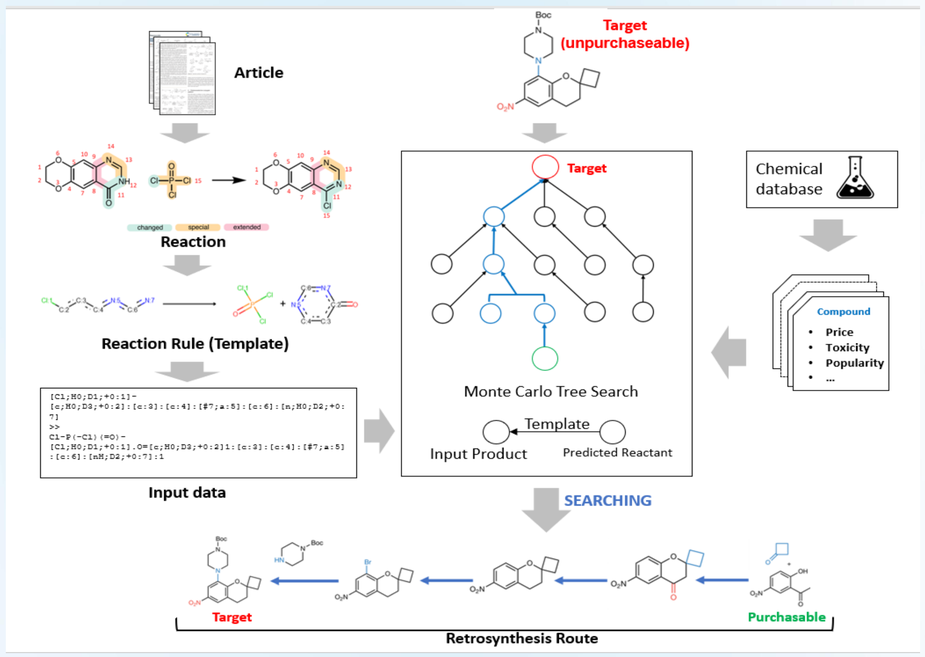
The SMILES from the reactions in literatures will also be transformed into another chemical language and undergoes neural network training.

After extensive training, the network is able to automatically classify the reaction rules and reaction groups suitable for the newly inputted reaction templates. Hence the templates could be classified, and better retrosynthesis routes can be devised with the suitable reaction rules for the reactions.

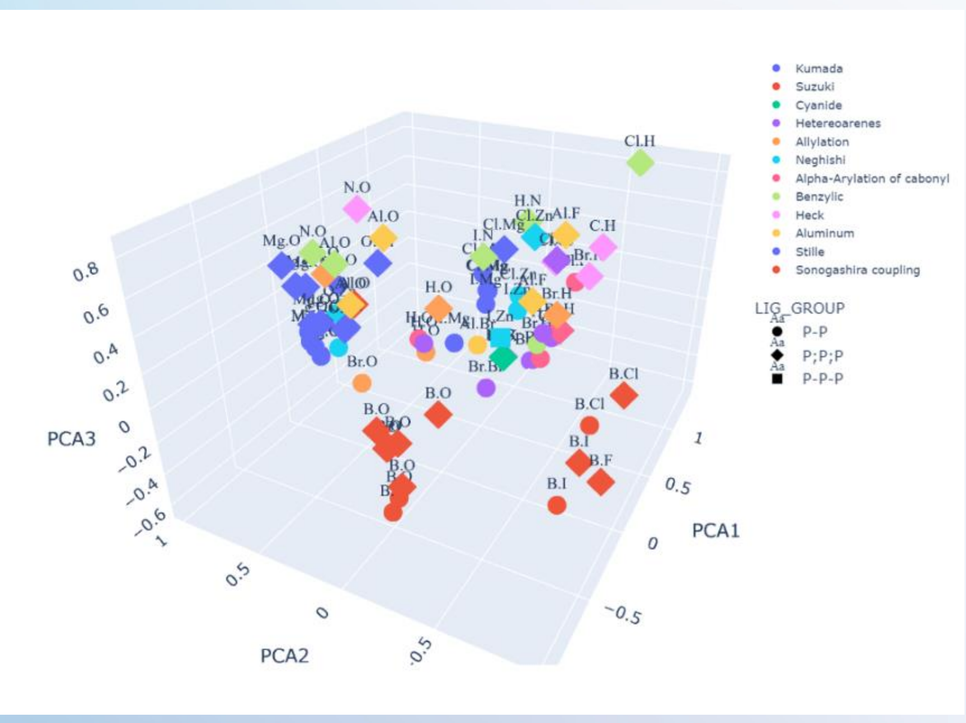


Automatic selection of synthesis pathway by Monte Carlo Tree Search

Once the user have submitted the SMILES notation of the desired chemical molecule into the SPLASH system. The retrosynthetic, synthesis from the product to reactant, Monte Carlo Tree Search will begin with the system's database of 200K templates. As the synthesis plan is retrosynthetic, the system will find a reactant that is 'one step backwards' in the synthesis scheme, eventually the Tree Search will end with a simpler reactant that is more available in terms of price, availability, popularity, etc.



Data-Driven cross-coupling reaction analysis for Nickel catalysis



Nickel catalysis is one of the most state-of-the-art and widely researched section for organic methodology in the recent decades. With its relative cheap price compared to its traditional transition metal counterparts, such as Palladium and Platinum, and useful reactivity for cross-coupling reactions, Nickel is one of the most promising metals for catalysis of organic reaction. In this project, we have filtered and chosen 551 articles on Nickel-catalyzed carbon cross-coupling reactions from high-quality research journals, including nature, science, JACS and extracted reaction information for 7449 reactions. From these reactions, we have sorted them in different categories according to the ligand used, type of scaffold of the reaction, leaving group, etc., which will assist us in plotting graphs and establishing trends between different categories of information. With the further usage of PCA and TMAP coupled with chemical descriptors such as Electronegativity, cone angle, buried volume, etc., trends and insightful correlations within Nickel-catalyzed carbon-carbon reactions can potentially be discovered.