Teaching the computer reactivity rules and strategies of automated retrosynthetic planning

Sara Szymkuć

Supervisor: Prof. Bartosz A. Grzybowski

Teaching computers to design syntheses of organic molecules has been one of the oldest challenges of computational chemistry. First software packages aiming to solve this problem were developed already in the late 1960s. Many other programs were created in the 1970s and 1980s but none of them met the expectations of organic-synthetic chemists and the effort was largely abandoned by the 2000s. This is quite unfortunate given that, in the meantime, computers have mastered many other analytical skills that had been considered exclusive domains of human intellect and creativity – for example, they can solve complex differential equations in symbolic forms (Mathematica) or can play games of strategy (chess, Go) better than human champions. In my doctoral thesis, I have taken inspiration (and hope) from these advances and for several years have been developing a computational platform known as *Chematica* that could finally plan efficient chemical syntheses.

The first step in teaching computer chemistry was to employ proper machine-readable data format for reactions and molecules. SMILES/SMARTS alphanumeric notation was chosen largely because of the speed with which it can process reaction operations and also because it was possible to augment it with detailed information about stereochemistry.

Equipped with this suitable notation, I undertook the challenge to teach the machine a nearly complete selection of organic reaction types. For each reaction class, I ventured deep into the underlying mechanism and delineated carefully the reaction core encompassing structural motifs and admissible substituents as well as "contextual" information describing incompatible functional groups, need for protection outside the core, and information about typical conditions. The rules and database fields I formalized underlie *Chematica*'s knowledge base of over 60,000 reactions of which I personally coded ca. 15,000.

The ultimate goal of my work was to teach the machine how to plan complete synthetic pathways without any human intervention. The key element here was to define proper scoring function enabling the search algorithm to estimate whether it is "moving" in a promising "synthetic direction," and how far it is from starting materials. To this end, I proposed a dual scoring function that assesses "synthetic positions" based on both molecular complexity and reaction feasibility.

Unfortunately, even correct but logically isolated synthetic steps do not necessarily make up for a sensible pathway. Recognizing this problem, I focused on how to identify the most promising reaction sequences and eliminate those that are unproductive or problematic.

The ultimate value and usability of any retrosynthetic software lies in experimental validation of its predictions. Initially, I performed "paper validation" whereby the program blindly recreated some published synthetic routes. The next step was the wet-lab validation. In this ultimate test, *Chematica* designed eight syntheses that were subsequently executed by chemists at Sigma-Aldrich, Northwestern University, and in our own laboratory at ICHO PAS. All computer designs were confirmed experimentally establishing the first-ever validation of a retrosynthetic software.

In the meantime, my scientific interests have been gradually shifting to new areas, especially to systems chemistry. In this spirit, the last chapter of my thesis describes recent work on the computational discovery of chemical-reaction cycles, akin to those used by biological systems and potentially useful as a means to recycle catalysts or to autoamplify valuable chemicals.