Accelerating research through lab notebook prospecting: A chemistry recommender system (final paper number: COMP 141)

Large pharmaceutical companies have a wealth of reaction and chemical structure data, but face a new problem: analyzing that corpus to yield project insights and future directions. One approach would be to have a recommendation system to match drug structures with similar research endeavors across geographically- or organizationally-separated groups. To improve knowledge sharing across our organization, we developed and deployed Chem Recommender, a recommendation system that analyzes work that chemists have recently completed in their electronic lab notebook (ELN) and suggests prior, similar work to them. The goal of the system is to accelerate the drug discovery process by making chemists aware of each other's work, saving both time and money. Since we launched Chem Recommender in July 2017, we've sent over 8500 recommendations to more than 800 unique chemists.

Our algorithm determines that two experiments are similar if at least one product is 90%+ similar and there is 80%+ similarity in at least half of the reactants. To perform our similarity search, we create an index from the ELN by generating n-grams of characters from the SMILES strings of the molecules involved in a reaction. Our similarity search also includes a time component that favors older experiments—when faced with a choice of several candidates, we choose the oldest experiment on the assumption that the chemist is most unlikely to be familiar with older work. At this time, only the top hit is sent to the chemist, the goal being to produce a recommendation that can be consumed quickly, without additional scrolling or clicking. We also ensure that we do not re-recommend the same scientist or experiment to the same person. Although this technique is simple, the results have been positive, with several chemists reporting that the recommendations have aided their molecular syntheses.

In this presentation, we'll describe the various recommendation algorithms available and why we chose to use a case-based approach. Next, we'll describe our reaction similarity algorithm in detail. Finally, we'll provide insights from chemists' feedback.