Learning the language of organic chemistry: developing artificial intelligence models using existing knowledge

Teodoro Laino, PhD
IBM Research -- Zurich

Reaction prediction is a fundamental task in organic chemistry. What if one could learn from all reactions published in articles and patents? What if this knowledge could be used to build a system that would assist researchers in the synthesis route design process? The use of data-based models combined with Artificial Intelligence (AI) strategies is emerging as a valuable and robust solution to address these fundamental questions. In this presentation, I will introduce an AI model in which forward reaction prediction and retrosynthesis are treated as a machine translation problem between SMILES strings of reactants-reagents and the products. I will show that a multi-head attention Molecular Transformer model outperforms all algorithms in the literature, achieving a top-1 accuracy above 90% on a common benchmark dataset. Our algorithm requires no handcrafted rules, and accurately predicts subtle chemical transformations. Furthermore, I will show that the model is able to handle inputs without reactant-reagent split and including stereochemistry, which makes our method universally applicable across existing datasets. All models are freely accessible world-wide as cloud applications: https://www.research.ibm.com/ai4chemistry