Praxis der Forschung

Chemical object recognition – Machine learning based translation from images to molecular structures

Topic

Artificial intelligence and machine learning are becoming powerful tools for research in natural sciences, in particular in chemistry and materials science. Building complex machine learning workflows often requires the extraction and analysis of large amounts of data that is published in literature and patents. Automated extraction of molecular structures from images using rule based algorithms or end-to-end learning approaches remains challenging. Significant progress in this task will allow the extraction of high-quality data sets from scientific literature and patents.

In this project, you will work on a diverse set of challenges: Identification of previously published methods for image to structure conversion, generation and publication of a benchmark dataset for such algorithms, development and implementation of a new algorithm combining multiple convolutional neural networks for object recognition and a graph based algorithm for structure extraction and representation. The goal is to publish the benchmark dataset as well as the newly developed extraction method in form of a peer-reviewed publication and to develop a versatile software tool for the extraction of chemical structures from large amounts of publications.

Requirements

- Interest in interdisciplinary work at the interface between computer science and chemistry
- Basic knowledge of machine learning and object recognition using convolutional neural networks
- No chemistry knowledge is required!

Group

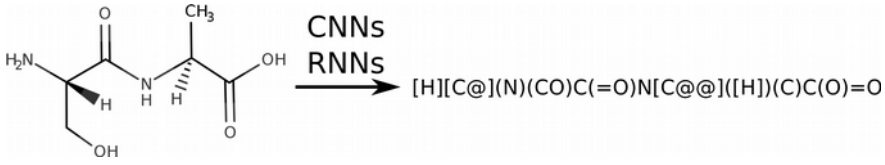
The AiMat group recently started at KIT and works on interdisciplinary research topics at the interface between computer science and material sciences. We develop and use AI and machine learning models to improve materials simulation methods and to accelerate the design of new materials.

Contact

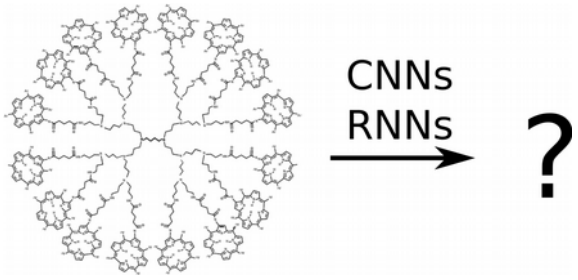
Jun.-Prof. Pascal Friederich (pascal.friederich@kit.edu)

Details

State-of-the-art in image to structure translation is a CNN/RNN based end-to-end learning approach that works for small molecules and simple structures but becomes unreliable for more complicated chemicals. This appraoch is illustrated in the image on the right, where a molecular structure is translated to a string based graph representation called SMILES. The state-of-the-art approch suffers from multiple problems that will be addressed in this project:



- Semantical and syntactical constraints make the SMILES representation fragile and hard to reliably learn by machine learning models
- The end-to-end learning approach cannot be extended to larger and more complicated chemical structures as frequently encountered in real world scenarios (see second illustration)



In this project you will work on the development of a combination of state-of-the-art object recognition algorithms trained on specialised datasets and a graph based representation of molecules that will solve the aforementioned challenges and enable the reliable extraction of molecular structures from publications and patents (see third illustration).