Universal molecular syntax for sustainable machine learning applications

Initiative: Momentum - Förderung für Erstberufene

Bewilligung: 23.03.2021

Laufzeit: 7 Jahre

Projekt-Website: https://www.bioac.ac.rwth-aachen.de/cms/~ictk/BIOAC/lidx/1/

In the last years, machine learning (ML) has been introduced as new tool in many fields in chemistry, e.g. to optimize catalysts and develop new drugs. All previous chemical databases rely on the usage of SMILES codes which work fine in organic chemistry with the majority of covalent bonds but totally fail in inorganic chemistry. Data-driven ML methods use the information from data bases to plan new syntheses, optimize reaction conditions and identify substances. Here is the largest hurdle for structure-based ML applications in inorganic chemistry. Most inorganic ML studies thus used an empirical approach but for the ML-supported synthesis planning in inorganic chemistry, the innovative development of a universal structure code would be a fundamental revolution which can be used worldwide. It is envisioned to conceptualise a new structure code and utilize it in ML studies on recent topics from bioinorganic chemistry and sustainable chemistry. The new structure syntax must be versatile, robust in its grammar and biunique, thus allowing conversion of structures into syntax and vice versa. Multiple bonds, ring systems and unusual bonding modes will be included by syntax extension.

Projektbeteiligte

Prof. Dr. Sonja Herres-Pawlis
Rheinisch-Westfälische
Technische Hochschule Aachen
Fakultät 1
Fachgruppe Chemie
Institut für Anorganische Chemie
Aachen

Open Access-Publikationen

TUCAN: A molecular identifier and descriptor applicable to the whole periodic table from hydrogen to oganesson
Making the InChI FAIR and sustainable while moving to Inorganics
Making the InChI FAIR and sustainable by moving to open-source on GitHub