# Substructure Based Fingerprints for Molecular Property Predictions

Mojtaba Haghighatlari[1], Johannes Hachmann[1, 2]

1. Chemical and Biological Engineering, University at Buffalo, SUNY, Buffalo, NY, United States.
2. New York State Center of Excellence in Materials Informatics, Buffalo, NY, United States.

The goal of our work is to recover the rigorous and deterministic quantum chemical mapping from the structure/topology of a molecule to its properties by means of machine learning (ML) and informatics. To recover this structure-property relationship, we need to start by developing a suitable numerical representation of the molecules that are contained in quantum chemical data compilations used as trainings sets. This challenge has recently been approached utilizing a number of new representations. One of these are fingerprints, and they reflect the presence or absence of particular functional groups in a molecule. Molecular fingerprints or the similarity metrics between fingerprint vectors have frequently been used to express ML prediction models. The choice of search keys (substructures/motifs) in the fingerprint plays an important role in the applicability domain of the resulting models. There is always a trade-off in the predictive capabilities of different types of fingerprints for specific applications. For a benchmark data set of electronic properties of more than 2 million organic semiconductors, we could show that ring substructures, which are easily accessible, provide a surprisingly useful fingerprint basis. This limited choice of motifs allowed for the construction of ML models with exceedingly competitive prediction errors. These models can then be used for in virtual high-throughput screening for accelerating materials discovery and as a guide for rational design.