# **Circular fingerprint inversion: an algorithmic approach**

Dehaen Wim1,Čmelo Ivan1,Karlova Andrea2,Svozil Daniel1

1 Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technicka 5, 166 28, Prague, Czech Republic

2 Department of Computer Science University College London London, United Kingdom

Molecular structures can be conveniently and concisely converted to molecular fingerprints, which allow fast database searching, model building and other tasks. However, they are not a full molecular representations, and typically there is some degree of information loss when converting a molecular structure to its corresponding fingerprint. In other words, converting a fingerprint back to its parent structure is a non-trivial and often impossible task if a structure fingerprint dictionary lookup is not possible.

One of the most commonly used families of fingerprints are circular fingerprints, such as the Morgan fingerprint, also referred to as the Extended-Connectivity Fingerprints(ECFP). These fingerprints encode chemical environments within a defined topological distance from each atom in the structure. Conversion to a circular fingerprint was considered to be a one-way operation, and even a way to „encrypt“ proprietary structures, but recently Le *et al* used a neural network based method to show a significant percentage (10-50%) of ECFP can be inverted. This surprising and powerful result does have some limitations: the neural architecture used is quite involved and requires a non-trivial amount of training resources, including a training set of 1.4 million drug-like molecules. Additionally, a different model needs to be trained for different fingerprint settings, such as radius and bitvector length.

To address these limitations, we investigated an algorithmic approach to fingerprint reconstruction. Briefly, the unknown molecule is built atom per atom in a depth-first type search, at every stage matching the intermediate molecule to the fingerprint and backtracking when there is an on-bit in the fingerprint that does not match the target fingerprint. This strategy managed to recover a large percentage of structures and had the advantage of not requiring any training data, meaning biases in a training set (such as only drug-like molecules, only synthesizable molecules etc) are not a factor. No training time is required, so inversion at given settings can immediately be started. To test this idea, we reconstructed molecules from the GDB, a database that encodes all possible structures of a given heavy atom count satisfying certain minimal criteria (only heavy atoms C,N,O,F, valence rules, stability rules etc). Our prototype approach reconstructed 100% of gdb-7, 99.3% of gdb-8, 95.1% gdb-9, 88.5% of gdb-10 and 82.4% of gdb-11 from their ECFP6(2048).