The SCFP construction algorithm transforms the compounds into a graph representation, where atoms are represented as nodes and the bonds between them as edges. Next, the SMARTS patterns of substructure keys are detected within the compound. The graph representation of the compound is then transformed into a semi-substructural one, where particular substructures (hits) and remaining atoms are represented as nodes and the connections between them are represented as edges (Figure 2). The connections between substructures are read using a handful of graph-dedicated algorithms (Iterative Deepening Depth-First Search, Breadth-Frist Search, etc.). The connections are finally translated into a connectivity matrix, and may be stored in a few formats: a matrix, matrix "hit" coordinates and linear notation (Figure 3).

The substructures searched came from the popular predefined sets: SubstructureFP (580 keys), MACCSFP (166 keys), and Klekota-Roth FP1 (4800 keys). The resulting fingerprint can be analyzed using machine learning methods, such as support vector machines, naive bayes, random forest and extreme entropy machines2. Here, the classification results are compared to original, key-based fingerprints and Extended fingerprint, a popular non-key-based substructural fingerprint.