Finding new medications is a laborious process taking many years to finalize. But before the new drug even starts its clinical trials, a whole procedure of chemical compound discovery and refinement takes place in research laboratories. In the earlier days this meant creating various compounds and testing the effects of their admission on various cell lines. Only when a compound proved to be active, that is initiating a specific cell response, would such compound be even considered to be tested on mice and rats. This process may take many years in order to yield a single promising compound. Since only a fraction of a per mill of tested compounds showed any positive results at all, hundreds of thousands of chemicals had to be created for a single one to reach clinical trials. This implied extremely high costs for pharmaceutical companies and enormous amount of time needed for a drug to enter the market. In late 20th century, with the rise of computational technology, the field of pharmacology devised numerous computer-based methods that significantly aided the overall drug discovery process.

Some of these methods are collectively called virtual screening, and revolve around calculating various properties of compounds based on their atomic structure. These properties are then used to assess the probability that the respective compound will be active towards the target protein. The computational methods can be divided into two major groups: structure based and ligand based. The structure based approaches rely on structures of both chemical compound and the protein it is aimed at and calculate the energies of their complexes. The most popular methods using these properties are ligand docking and molecular dynamics. The ligand-based approach focuses only on the compounds structures and properties, striving to predict their probability of being active based on their similarity to already known drugs or their physicochemical properties. The most popular methods here are fingerprints and pharmacophores. Since they do not require the information about the target proteins' structure, the computational cost of these methods is relatively low.

Fingerprints are a method of describing an object with a string of values, which allow to be analyzed, clustered and classified by various algorithms. They are commonly used in various applications, such as word recognition, database safety and others. In case of chemical fingerprints, the compounds fingerprint may represent its fragmentary structure, physicochemical properties or various mixes of the two. One of such fingerprinting methods are the substructural fingerprints, which depict the occurrences of predefined chemical fragments (substructures) within the analyzed compound. This is a simple method of portraying the approximate structure of a compound, however, it has its flaws. Substructural fingerprints lack the information about the connectivity of the particular substructures, which means that two or more different compounds may share the same fingerprint.

This issue is addressed in the Project "Substructural Connectivity Fingerprint and Extreme Entropy Machines: A New Method of Compound Representation and Analysis". The substructural connectivity fingerprints (SCFP) presented within the Project are substructural fingerprints that contain additional data on the connections of the substructures within the compound, delivering much more accurate data for analysis. The preliminary studies have already shown, that a crude implementation of such methods leads to a significant increase in compound classification accuracy, which translates into an improvement of efficacy of drug design campaigns. The fact that the Project implies use of more sophisticated methods of analysis, may only contribute to even bigger improvement of accuracy measures, and therefore the efficiency of compound classification.

The Project consists of several steps, leading to a final product, that is a complete methodology together with analysis methods and applications available for download and use on any compound set. The first stage of the Project will focus on the construction of SCFP-building algorithm. The following steps will explore different approaches to compound datasets generation and refining, will apply already available machine learning analysis methods, as well as implement novel analysis methods, like extreme entropy machines and graph kernels. The SCFP methodology will be tested against already available substructural methods and various machine learning algorithms. The potency of the fingerprint to discriminate between active and inactive compounds will be assessed with use of multiple chemical compound databases, containing both experimental and computational data. All these studies will be performed for multiple target proteins, including G protein-coupled receptors and protein kinases.

Summarizing, the "Substructural Connectivity Fingerprint and Extreme Entropy Machines: A New Method of Compound Representation and Analysis" Project will result in a new methods of compound representation and analysis, verified through extensive studies on its efficiency in comparison to currently available methods. The Project will significantly enrich current methodology of ligand-based virtual screening, and will enhance the process of searching for new drugs.