The goal of this project is to **develop universal methods for data representation to improve the quality of machine learning methods with an examplary application in chemistry**. Representation theory attempts to answer the question how to reflect the actual data by a computer. Undoubtedly, a proper selection affects the quality of applied algorithms (machine learning methods). In the present project, we refer to this observation and propose some solutions. As confirmed by experts and preliminary research done on the example of chemical data, despite the large amount currently available and used representations, they take into account the knowledge about the properties of data chemicals in a patchy way, which in turn brings difficulties in computer performed evaluation of the activity of molecules on a target protein. That is why we want to deal with the development of new representation which is expected to take into consideration both structural features and physico-chemical properties of molecules. Furthermore, searching for the active compounds reveals the problem of inappropriate choice of metric for the space. Therefore, the project will provide a new concept to construct it.

The project is essentially in a few parts, but the main stage will be to cover theoretical aspects connected with new representations, and preparing the implementations. Ultimately, it is planned to provide an application with a user-friendly interface. We will use three different methodologies: based on graphs, models using n-gram approach and innovative method of Hu moments. We will also introduce our technique of metric calculation based on entropy. The introduced technology will be examined by assessing methods results quality and comparing the results obtained by visualization algorithms with expert knowledge.

The major reason for addressing this subject is the existing problem in the machine learning field caused by the use of inappropriate representations. Unfortunately, the problem is in fact twofold: this results in too low quality of applied techniques or low efficiency of algorithms, which, as in the example of drugs discovery is vital both in terms of time and financially. Although the solutions provided in the project are shown in use in chemistry, they would be applicable in many fields of machine learning, for instance in computer vision. However, this will provide the improvement of the process of drugs searching.