Determining the structure of macromolecules, proteins and their complexes is one of the basic research questions posed by biology and structural chemistry. Proteins are complex molecules consisting of a large number of amino acids linked together in a chain by peptide bonds. They are responsible for the structure, function and regulation of tissues and organs inside the body, and do most of the work in the cells. Their structure can be determined on the basis of the electron density distribution obtained from the X-ray diffraction experiment. One of the values describing data quality is resolution - the better it is, the more details of the structure can be observed.

Obtaining high resolution X-ray diffraction data for macromolecules is difficult but necessary to reconstruct the electron density using an aspheric multipole model. Such a model surpasses the commonly used spherical model in terms of describing electron density deformation from chemical bonds and lone electron pairs. Unfortunately, the overwhelming majority of the available data contains only information about structure, not electron density. In order to use them, banks of aspherical atomic types are implemented. Such banks allow the use of an average description of similar atoms to reconstruct the electron density, rather than describing each atom separately. However, current aspherical data banks are not suitable enough to reconstruct the electron density of most macromolecules directly from the file that contains their structure - mmCIF.

In this project, we propose a solution to this problem. For this, we will use a new approach to recognizing atom types tailored specifically to proteins, RNA, DNA and the most common ligands. Using the methods previously created by our research group, we will create a new bank based on data from the world's largest crystallographic protein database - the Protein Data Bank. We propose to introduce an algorithm to recognize atomic types based on the names of atoms strictly defined in dictionaries and present in proteins, RNA, DNA and the most common ligands. The project will involve significant amounts of data and combining knowledge of chemistry, structural biology and programming.

The implementation of this project will improve the quality of molecular structure determination and reconstruction of the electron density distribution of proteins, RNA, DNA and the most common ligands. This is critical for structural biology and chemistry as charge-density analysis can identify the functions of macromolecules, predict how structural changes might affect them, or understand the interactions that occur in molecular complexes. We predict that this concept should also solve problems with hydrogen atoms, which are difficult to locate even at high resolution, because the electron density distribution from their single electron is deformed towards the bond. The introduction of our bank that recognizes atom types based on the names of atoms in proteins, RNA, DNA and the most common ligands will significantly affect areas such as pharmacology, structural biology and nanotechnology.

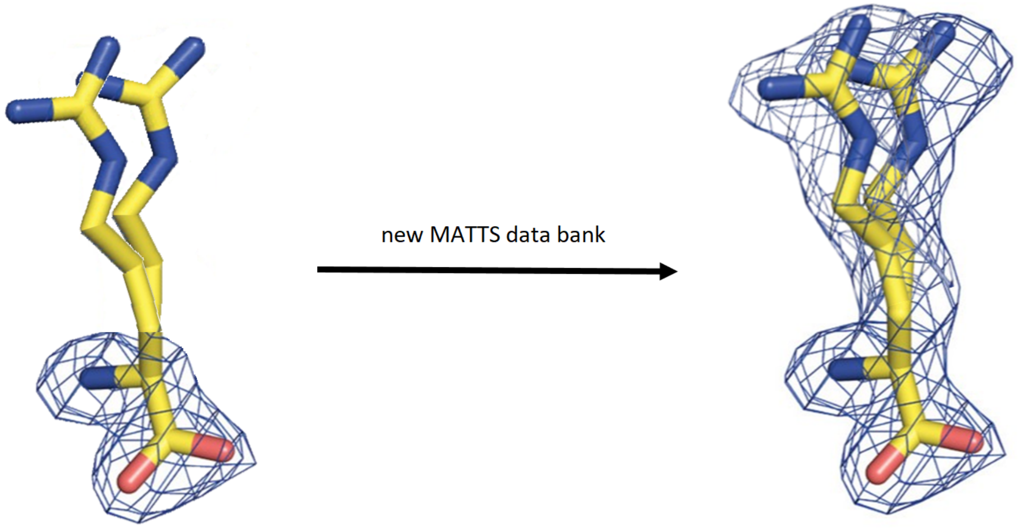

Figure 1. Structure of arginine that is a part of a protein. Atoms are colored according to the scheme: blue - nitrogen, red - oxygen, yellow - carbon. Hydrogen atoms are not shown. Left panel: the reconstructed electron density (contour shown as a net) is incomplete due to a lack of correct recognition of atom types which are present in two conformers. Right panel: all atom types are correctly recognized, and electron density can be properly reconstructed.