Warszaw, 2.10.2012

Katarzyna Jarzembska

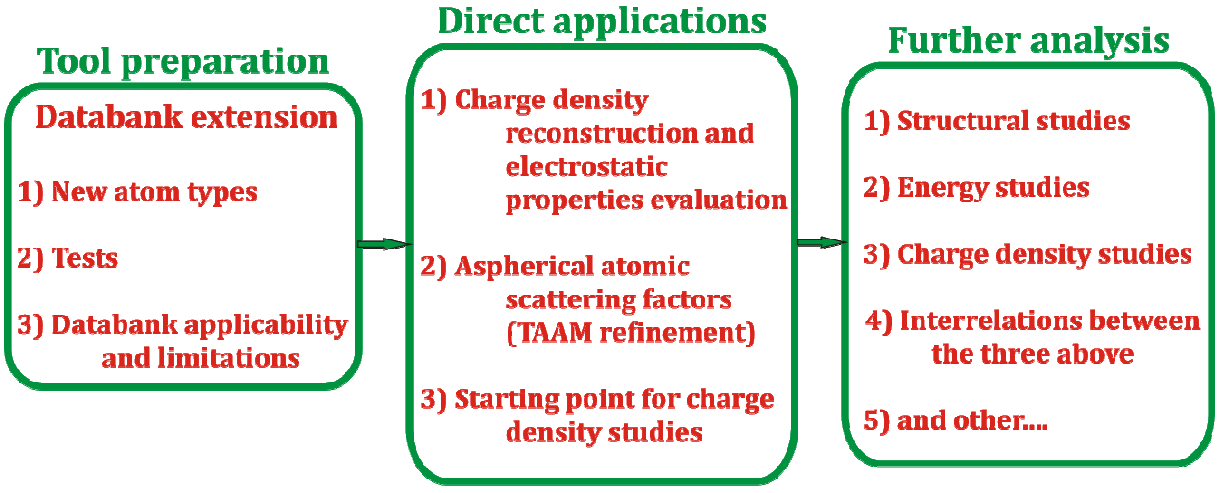
Chemistry Department

University of Warsaw

„EXTENSION OF THE ASPHERICAL PSEUDOATOM DATABANK TOWARDS NUCLEIC ACIDS AND ITS APPLICATION IN STRUCTURAL, CHARGE DENSITY, AND ENERGY STUDIES."

Advisor: Dr Paulina M. Dominiak
Supervisor: Prof. dr hab. Krzysztof Woźniak

My PhD project consists of two clearly distinguishable parts. The first part includes modifications and extension of the aspherical pseudoatom databank, the UBDB (University at Buffalo Databank), and the related *LSDB* program. It also contains extensive tests of the databank applicability to the estimation of electrostatic properties of molecules. The second part of my Thesis is based on the use of the extended UBDB for the following scientific purposes: (1) reconstruction of the charge density distribution of macromolecules, which may constitute foundation for further estimation of electrostatic properties; (2) as a source of aspherical atomic scattering factors, which could be used in the structure refinement (so-called Transferable Aspherical Atom Model (TAAM) refinement);(3) as a starting model in the experimental charge density studies. The study is not just limited to databank-related scientific issues, but it also addresses subjects regarding the interrelations between geometry, energy, charge density and crystal morphology. Selected biochemical aspects were additionally investigated. The idea of my PhD thesis is schematically illustrated below (Scheme 1).



Scheme 1. The scope of my PhD thesis.

Consequently, my study resulted in a new modified version of the UBDB databank extended with atoms required for modelling RNA and DNA molecules. The enhanced bank, the UBDB2011, now contains over 200 atom types present in the most relevant biochemical systems. Multipolar parameters stored in the databank were evaluated on the basis of over 600 organic molecules. During my work on the databank, I also modified the UBDB-related *LSDB* program, so as to include adequate atom type definitions, local coordinate systems, and updated X–H bond distances according to Allen *et al*. The *LSDB* now also writes automatically input files suitable for the *MOPRO* suite of programs, which is an alternative for the well-known *XD* package.

Once the UBDB2011 version of the databank was ready, I extensively tested its applicability to estimating the electrostatic properties of chemical systems. The verification was based on the electrostatic interaction energy evaluation for a set of nucleic acid base and amino acid complexes, for which it was possible to run the reference *ab initio* calculations. I showed that the UBDB2011+EPMM (Exact Potential Multipole Method (EPMM)) method satisfactorily reproduces electrostatic interaction energies for a set of nucleic acid base complexes with respect to *ab initio* and/or DFT results (R^2 > 0.9, RMSD = 3.7 kcal·mol^{-1}). Correlations were found to be high, while energy trends were preserved. What is important, the UBDB databank enables taking the asphericity of atoms into consideration, therefore, unlike point-charge models, it addresses the directionality of atom–atom interactions.

However, the UBDB+EPMM approach has a number of limitations. The databank parameters do not reproduce conformational variety, and do not include the crystal field influence, and some other subtle effects. The UBDB data is also constrained by the accuracy of atom type definitions and the charge density parameter computation method (basis set used, DFT method, Fourier truncation error, Hansen-Coppens model limitations, atom type definition adequacy, *etc.*). Consequently, when applied for electrostatic interaction energy evaluation, the UBDB2011+EPMM accuracy is of the order of 5 kcal·mol^{-1}. Furthermore, the databank is not appropriate for describing charge density distribution of non-standard molecules, or systems containing alternated double bond fragments, which I showed using the example of Amphotericin B, and its iodoacetyl derivative.

The prepared and tested UDBD2011 new databank opened up the possibilities for further studies. Thus, as mentioned in the preface, the second part of my Thesis can be divided into three separate sections. The first one is devoted to the electrostatic interaction investigations of influenza neuraminidase complexes with a series of inhibitors, with particular attention paid to the role of water molecules in the active site. The bank was used for charge density distribution reconstruction. In the second section, I focused on a set of 10 uracil derivatives. Here, the databank was applied as a source of atomic scattering factors in the TAAM refinement. Finally, I dedicated my attention to charge density studies, where I managed to obtain high resolution data, *i.e.* of 6-methyl-2-thiouracil (**6m2tU**), and a co-crystal of a complementary nucleic acid base pair, 1-methylthymine and 9-methyladenine (**9mA:1mT**).

The studies of the neuraminidase-inhibitor complex revealed the databank potential in the analysis of biochemical systems and drug activities. Thanks to the UBDB+EPMM method, it was possible to estimate the partial contributions to the total electrostatic interaction energy coming from single fragments of ligand molecules, or from any selected aminoacid residue of the protein part. This way it is easy to identify the active site region (strongest interacting residues), or crucial atomic groups, limiting or stimulating the protein-inhibitor binding. My contribution was, though, directly related to the water molecule role in the neuraminidase active site. I succeeded to estimate the solvent influence on the complex stability. The results helped to explain some differences in the inhibition activity of selected ligands, which in certain cases were not correlated directly with the strength of electrostatic binding energy of ligand- protein complex. On the whole, such electrostatic interaction energy studies are most justified in the case of systems, where electrostatic interactions outweigh other interaction types. Influenza neuraminidase was a perfect case for such considerations.

Subsequent the analysis of 10 uracil derivatives (*i.e.*, uracil, 1-methyluracil, 1,5-dimethyluracil, 2-thiouracil, 4-thiouracil, 2,4-dithiouracil, 1-methyl-4-thiouracil, 2-thio-5-methylouracil, 2-thio-6-methyluracil, and 5-fluorouracil) revealed that the UBDB2011 databank works also very well as a source of atomic aspherical scattering factors. It occurred that TAAM refinement provides significantly more accurate molecular geometries than the standard IAM (Independent Atom Model) refinement. The energy and geometry studies showed that TAAM-derived molecular geometries are closer to those obtained from the experimental charge density studies and neutron measurement, but also to the periodically optimised structures within the *CRYSTAL* package. The TAAM cohesive energy and motif interaction energy trends are also in perfect agreement with the ones observed for optimised structures, whereas the IAM results differ significantly. My studies provided seven high quality structures, among which the 4-thio-uracil structure had not been earlier deposited in the CSD. All the investigated systems, apart from 2,4-dithiouracil, form layered architectures. The molecules within the layers are held by hydrogen bonds, whereas the molecular layers interact with one another via more dispersive in nature, however, significant contacts. The cohesive energy difference between the most thermodynamically stable 5-fluorouracil, and the least advantageous crystal lattice of 1-methyl-4-thiouracil, amounts to 40 kJ·mol^{-1}. The analysis of structural motifs revealed the methyl group stimulation of stronger interlayer contacts, whereas the S4 substituent usually affects the stability and quality of the formed crystal. These two factors most substantially influence the aromaticity index values, *HOMA* and *NICS*. *HOMA* and *NICS* do not agree, though. The calculated deformation molecule energy between the isolated molecule and the molecule in a crystal lattice, depends on the strength and number of intermolecular interactions, in which this particular molecule participates, and can reach about 10 kJ·mol^{-1}. I have additionally analyzed the interesting motif created by the fluorine atoms in the 5-fluorouracil (**5fU**) crystal lattice. Four **5fU** molecules create a tetrameric motif with all the four fluorine atoms pointing into each other. On the basis of the databank, I have reconstructed the electron density of molecular fragments contributing to such a pattern. The derived deformation electron

density indicated the potential stabilizing character of the F…F contact, of the 'lump to hole' type.

The analysis of charge density distribution of nucleic acid base crystals constituted the last subject undertaken in my PhD Thesis. The project resulted in two high-resolution datasets, *i.e.*, for 6-methyl-2-thiouracil and for the co-crystal structure of 1-methylthymine and 9-methyladenine. The **6m2tU** studies showed that the databank can not only be applied as a starting charge density model, but can also be helpful in handling disorder in a crystal lattice. In my case, the UBDB2011 was used to model the problematic sulphur atom, and indicated the hypothesis of some content of the oxo-thiol **6m2tU** tautomer exiting in the crystal lattice. This could explain systematic errors, which I observed during the charge density data analysis, and peculiar residual peaks visible in the vicinity of the sulphur atom. The study showed also that one has to be very careful with the modelling and data interpretation. On the other hand, regular and well-diffracting **6m2tU** crystals allowed me to find the interrelations between charge density, energy and crystal morphology. It occurred that there is a connection between the crystal architecture features, and the macroscopic shape of the crystal. I have also found relations between crystal thermodynamic characteristics and crystal face stability, and crystal formation. Finally, good agreement was observed between the calculated hydrogen bond energy with that derived from the charge density in the Espinosa's approach.

The **9mA:1mT** co-crystal constituted the last subject of my studies. Interestingly, the **9mA:1mT** base pair does not form the standard Watson-Crick configuration in the crystal lattice, but the Hoogsteen-Watson-Crick base pair motif. It was also found out, that generally adenine derivatives when forming crystals with uracil species, tend to bind in this particular manner. At the same time, this kind of purine-pyrimidine orientation is rarely encountered even in the relatively labile RNA structures. Nonetheless, computational analysis results support slightly better stabilisation of the cHW-type dimers, than the standard cWW ones. The topological investigations gave very similar results to those obtained for the analogous molecules. What is more, the databank parameters reconstructed reasonably well the experimental charge density evaluated for **9mA:1mT**, while the model satisfactorily reproduced the intermolecular interaction region. The **9mA:1mT** is yet another crystal structure of layered architecture characterised by quite strong hydrogen bonds and effective π-stacking interactions. An interesting feature of the crystal network is the presence of close H…H contacts, which is not the case in the mono-component **1mT** and **9mA** crystals.

To recapitulate, my PhD thesis provides extended and improved tools for charge density modelling, which can be successfully applied to electrostatic property study of macromolecules, for deriving more accurate structure geometries, or used in charge density analysis. Furthermore, my studies indicate the interrelations between geometry, structure motifs, lattice energy and crystal morphology. Most of the presented results have already been subjects of scientific articles (to date, I am a co-author of 11 publications). Last manuscripts, regarding these topics, are currently under preparation.

The further development in the directions indicated by the presented studies concern the databank application to macromolecular systems, which are of biological importance. The UBDB databank and the idea of TAAM refinement should be popularised, as only the IAM refinement is widely applied, at present. Currently, the databank is being combined with the *PHENIX* package to enable the application of the TAAM procedure also to proteins. Investigations show that there should be some complementary methods provided to supplement the electrostatics derived on the basis of the databank, among which, also, solvent effects should be taken into account. On the other hand, one could consider alternative ways of accurate charge density modelling, free from the limitations of the UBDB databank, and also, the multipolar model.

A comprehensive geometry, energy, and environment analysis of the non-standard nucleic acid base interactions, and also of other important factors, such as interactions with sugar moieties, phosphate groups, water or ion species seems interesting, too. This could give a better insight into the RNA binding properties, and may also lead to some pharmaceutical outcome.