Dictionary-Driven Atom Type Recognition for Aspherical Electron Density Reconstruction in Macromolecules with the MATTS Data Bank

P. M. Rybicka[1], M. Kulik[1], C. Grandgirard[2], M.L. Chodkiewicz[1], P. M. Dominiak[1]

[1]University of Warsaw, Faculty of Chemistry, Biological and Chemical Research Centre, ul. Żwirki i Wigury 101, 02-089 Warsaw, Poland, [2]CESI Graduate School of Engineering, 6 Rue Bois du Chêne le Loup, 54500, Vandœuvre-lès-Nancy, France

p.rybicka@uw.edu.pl

Interpreting X-ray diffraction data and determining the three-dimensional structure of a molecule requires an accurate, reliable, and valid electron density model. The Multipole Model improves simpler spherical approach of the Independent Atom Model by accounting for the deformation of the electron density mostly from chemical bonding and lone electron pairs [1]. However, to observe and measure the aspherical nature of electron density, atomic charges, and atom positions, a high-resolution experimental data, often unavailable for macromolecules, is typically required. At lower resolutions, electron density can be described with atom types using data banks like MATTS [2,3], which compiles transferable electron density parameters for various atom types based on small-molecule structures (multipole parameters can be transferred between chemically equivalent atoms). While multipole modeling is well-established in small-molecule crystallography, its application to macromolecules remains limited, hindered by issues in the atomic model such as missing atoms, clashes, alternative conformations, and distorted geometry as well as the computational cost of the standard atom type assignment procedure based on atomic coordinates.

To improve the accuracy and speed of atom typing using the MATTS data bank, we have been developing a dictionary-driven atom type assignment procedure for macromolecules based on standarized residue and atom names (e.g. HIS ND1), without the need for a coordinate-based matching. The dictionary covers the standard protein residues, RNA and DNA nucleotides, and most common ligands, including different protonation states and terminal variants. We analyzed over 45,000 high-quality, diverse structures from the RCSB Protein Data Bank, processed using our previously developed *DiSCaMB* [4] library adapted for mmCIF files, alongside the open-source *gemmi* [5] library to clean disorder, add hydrogen atoms, and handle protonation states. Atom types from the MATTS2021 data bank were assigned, successfully recognizing 99% of over 562 million atoms and forming the basis for our atom type dictionary. Future work will focus on integrating this atom type recognition approach into tools like *Phenix* [6], making aspherical modeling more accessible for routine macromolecular crystallography.

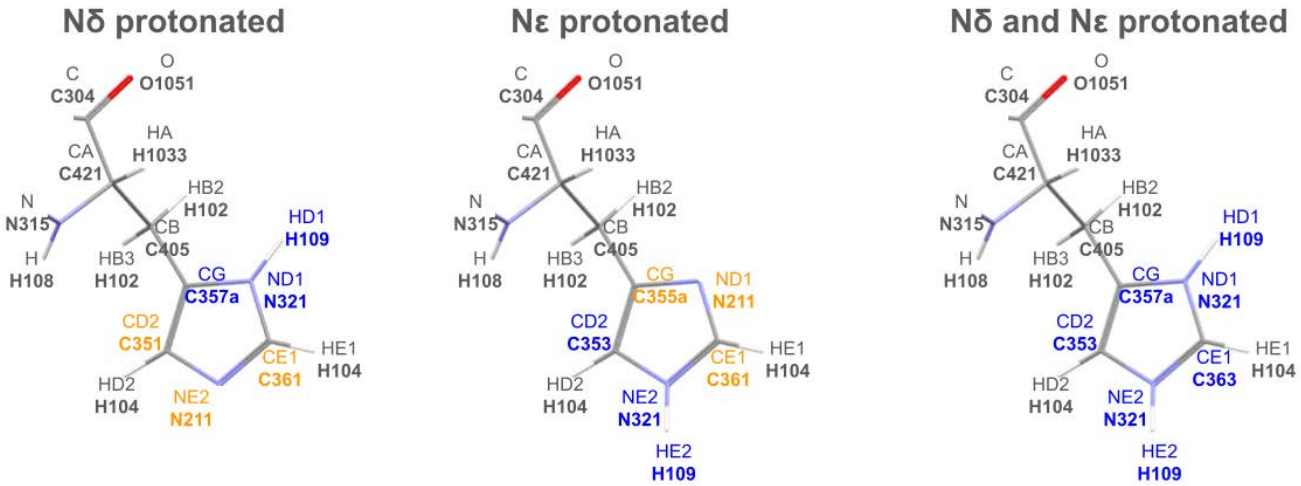

Figure 1. MATTS2021 atom types assigned to histidine with different nitrogen protonation states in the imidazole ring.

[1] Hansen, N.K. *et al.* (1978). *Acta Cryst. A* **34**, 909-921.

[2] Jha, K.K. *et al.* (2022). *J. Chem. Inf. Model.* **62,** 3752-3765.

[3] Rybicka, P.M. *et al.* (2022). *J. Chem. Inf. Model.* **62**, 3766-3783.

[4] Chodkiewicz, M.L. *et al.* (2018). *J. of Appl. Cryst.* **51,** 193-199.

[5] Wojdyr, M. *et al.* (2022). *J. of Open Source Soft.*, **7(73)**, 4200.

[6] Liebschner, D. *et al. (*2019). *Acta Cryst*. **D75**, 861-877.

This research was funded by National Science Centre, Poland UMO-2023/49/N/ST4/02967.