Oral presentation

Using deep learning predictions reveals large number of register errors in PDB deposits

Daniel Rigden

University of Liverpool, UK
drigden@liv.ac.uk

The accuracy of the information in the Protein Data Bank (PDB) is of great importance for the myriad downstream applications that make use of protein structural information. Despite best efforts, the occasional introduction of errors is inevitable, especially where the experimental data are of limited resolution. We have previously established a novel protein structure validation approach based on spotting inconsistencies between the residue contacts and distances observed in a structural model and those computationally predicted by methods such as AlphaFold 2. It is particularly well-suited to the detection of register errors. Importantly, the new approach is orthogonal to traditional methods based on stereochemistry or map-model agreement, and is resolution-independent. Here we identify thousands of likely register errors by scanning 3-5Å resolution structures in the PDB. Unlike most methods, application of our approach yields suggested corrections to the register of affected regions which we show, even by limited implementation, lead to improved refinement statistics in the vast majority of cases. A few limitations and confounding factors such as fold-switching proteins are characterised, but we expect our approach to have broad application in spotting potential issues in current accessions and, through its implementation and distribution in CCP4, helping ensure the accuracy of future deposits.