Poster

Predicting Protein Crystal Solvent Content with Machine Learning

D. McDonagh, D. Waterman, R. Keegan

CCP4, Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot, OX11 0FA, England

david.mcdonagh@stfc.ac.uk

A key step in the structure determination process of protein crystals is an estimation of the molecular unit cell content, which can be performed after diffraction data has been indexed with a candidate lattice using modern diffraction reduction software. Identifying the potential number of molecules in the unit cell, and the corresponding solvent content, plays a key role in identifying the overall symmetry of the cell, in addition to addressing the phase problem required to calculate the charge density.

The relationship between the solvent content, the volume of the asymmetric unit, and the molecular weight of the protein is well understood, often referred to as the Matthew's Coefficient[1]. However, in the structure solution pipeline, the number of molecules is typically not known. A probability estimator based on Matthews coefficient was developed as a function of resolution of crystal structures in the PDB[2]. This allows probabilities to be assigned to different solvent content estimates, and is now routinely used in the structure solution pipeline in packages like MATTHEWS_COEF in CCP4[3]. This approach works well for small numbers of molecules, but struggles with larger unit cells. In a dataset of 80,691 PDB entries used in this work, we find ~ 20% are identified with the incorrect number of molecules in the unit cell. This leads to a poor initial guess for procedures downstream such as molecular replacement, which results in a significant computational cost.

In this work we investigate a series of machine learning models to predict the solvent content of structures based on Patterson maps, which can be readily calculated after obtaining scaled intensities. We demonstrate an average reduction in prediction error of over 50% compared to MATTHEWS_COEF (Fig. 1, top left), and find negligible reduction in accuracy with respect to the number of molecules in the unit cell (Fig. 1, right). Deploying this approach in the structure solution pipeline should result in much faster structure solution times with fewer computational resources.

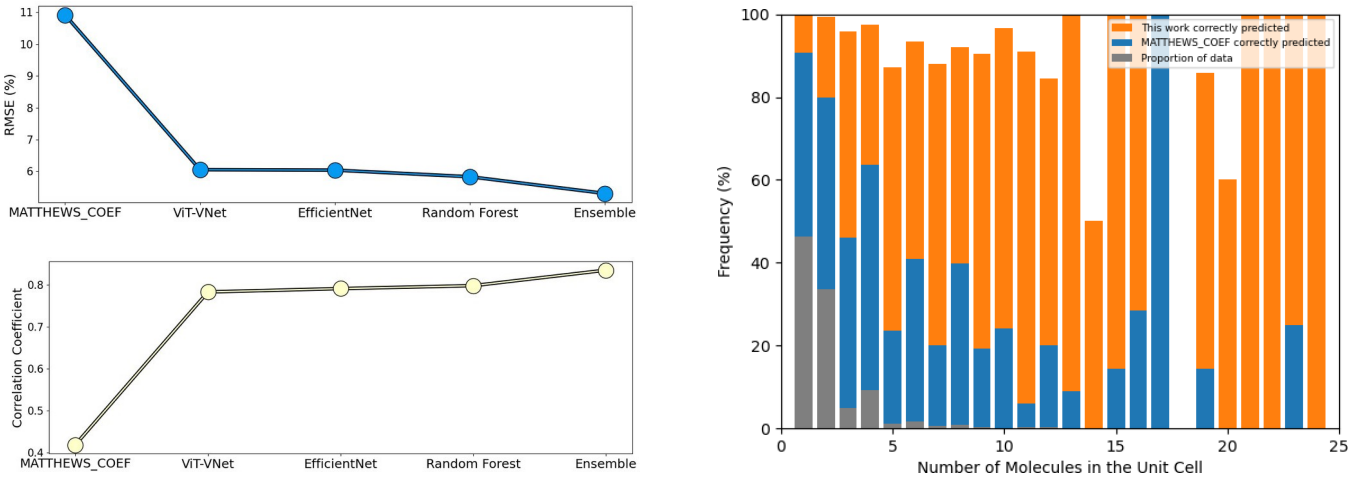

Figure 1. The root mean squared error (RMSE) of predicted solvent content and reported content in the PDB using MATTHEWS_COEF, and the machine learning methods in this work (top left); the Pearson correlation coefficient for the same data (bottom left); and the proportion of data with different numbers of molecules correctly predicted using MATTHEWS_COEF compared to the best method of this work (right). All results are for 16,130 structures randomly assigned to a test set that were not used in training or validation for any of the machine learning models.

[1] Matthews B., (1968). J. *Mol. Biol.* **33**(2), 491-497

[2] Kantardjieff K., Rupp B. (2003). Pro. Sci. **12**(9), 1865-1871

[3] Agirre J. et al, (2023), Actra Crystal. D. **79**(6), 449-461