Poster Presentation

MS112.P08

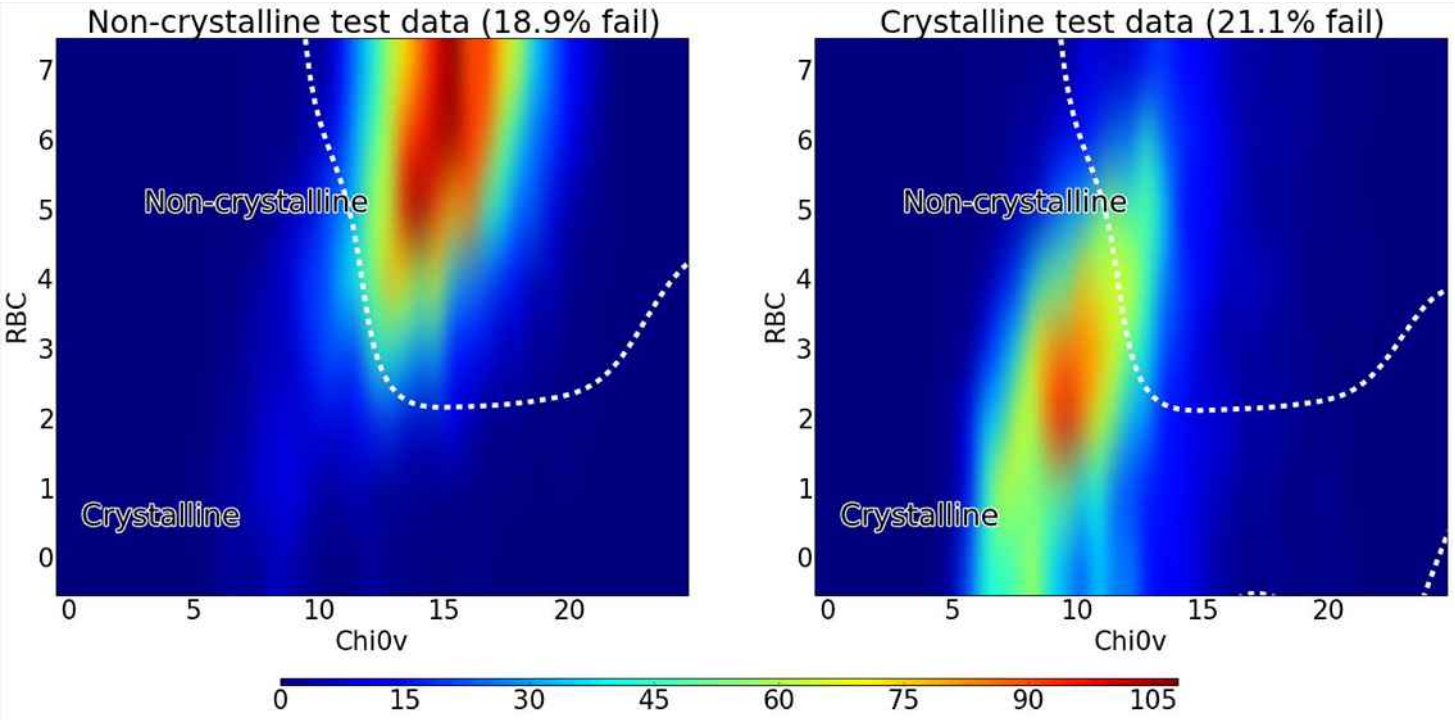
Predicting crystallisation propensity of small molecules

J. Wicker[1], R. Cooper[1], W. David[2]

[1]*University of Oxford, Chemical Crystallography, Oxford, UK,* [2]*ISIS Facility, Rutherford Appleton Laboratory, Chilton, UK*

We show that suitably chosen machine learning algorithms can be used to predict the "crystallisation propensity" of classes of molecules with a promisingly low error rate, using the Cambridge Structural Database and ZINC database to provide training examples of crystalline and non-crystalline molecules. Supervised learning tasks involve using machine learning algorithms to infer a function from known training data which allows classification of unknown test data. Such algorithms have been successfully used to predict continuous properties of compounds, such as melting point[1] and solubility[2]. Similar methods have also been applied to protein crystallinity predictions based on amino acid sequences[3], but little has previously been done to attempt to classify small organic molecules as crystalline or non-crystalline due to the difficulty in finding descriptors appropriate to the problem. Our approach uses only information about the atomic types and connectivity, leaving aside the confounding effects of solvents and crystallisation conditions. The result is reinforced by a blind microcrystallisation screening of a sample of materials, which confirmed the classification accuracy of the predictive model. An analysis of the most significant descriptors used in the classification is also presented, and we show that significant predictive accuracy can be obtained using relatively few descriptors.

[1] *A. Varnek, N. Kireeva, I. V Tetko et al, J. Chem. Inf. Model. 2007, 47, 1111–22,* ***[2]*** *B. Louis, J. Singh, B. Shaik et al, Chem. Biol. Drug Des., 2009, 74, 190–5,* ***[3]*** *G. Babnigg, A. Joachimiak, Journal of structural and functional genomics, 2010, 11, 71–80*



Keywords: crystallisation propensity, machine learning