HIGH THROUGHPUT ANALYSIS OF STRUCTURAL GEOMETRIES MINED FROM THE CSD: HOW TO INTERPRET 3000 HITS IN AN AFTERNOON?

G. Barr, W. Dong, C. J. Gilmore, A. Parkin and C.C. Wilson
Department of Chemistry, University of Glasgow, Glasgow, United Kingdom, G12 8QQ

Speaking on behalf of the authors: A. Kern, Bruker AXS GmbH, Karlsruhe, Germany

With the explosion in high quality structural determinations in the area of small molecule crystallography, the problem of efficient mining of the CSD is very relevant for structural chemists. The CSD represents an enormously powerful resource, but faced with more than 300,000 structures, the attempt to extract meaningful chemical information can be daunting. Many search algorithms and data mining strategies already exist, but here we present a method of extracting chemical information from such datasets using cluster analysis. By exporting the complete geometric information for a fragment of interest - all of the distances and angles - dSNAP provides facilities for automatically classifying and visualising the results of a search. It allows the user to assess those classifications chemically, diminishing the need for tedious manual classification. The real beauty of the method is in its interactivity and scalability - to go from getting an overview of the entire dataset and easily spot any outliers or errors, to the ability to 'drill down' within a cluster, looking at more and more detailed differences between different fragments.

dSNAP is available free of charge and can be downloaded at www.chem.gla.ac.uk/snap/.

HIGH THROUGHPUT ANALYSIS OF STRUCTURAL GEOMETRIES MINED FROM THE CSD: HOW TO INTERPRET 3000 HITS IN AN AFTERNOON?

G. Barr, W. Dong, C. J. Gilmore, A. Parkin and C.C. Wilson
Department of Chemistry, University of Glasgow, Glasgow, United Kingdom, G12 8QQ

Speaking on behalf of the authors: A. Kern, Bruker AXS GmbH, Karlsruhe, Germany

With the explosion in high quality structural determinations in the area of small molecule crystallography, the problem of efficient mining of the CSD is very relevant for structural chemists. The CSD represents an enormously powerful resource, but faced with more than 300,000 structures, the attempt to extract meaningful chemical information can be daunting. Many search algorithms and data mining strategies already exist, but here we present a method of extracting chemical information from such datasets using cluster analysis. By exporting the complete geometric information for a fragment of interest - all of the distances and angles - dSNAP provides facilities for automatically classifying and visualising the results of a search. It allows the user to assess those classifications chemically, diminishing the need for tedious manual classification. The real beauty of the method is in its interactivity and scalability - to go from getting an overview of the entire dataset and easily spot any outliers or errors, to the ability to 'drill down' within a cluster, looking at more and more detailed differences between different fragments.

dSNAP is available free of charge and can be downloaded at www.chem.gla.ac.uk/snap/.