ADDOPT

ADVANCED DIGITAL DESIGN OF PHARMACEUTICAL THERAPEUTICS

The Cambridge Crystallographic Data Centre

Hartree Centre
Science & Technology Facilities Council

UNIVERSITY OF LEEDS

AstraZeneca

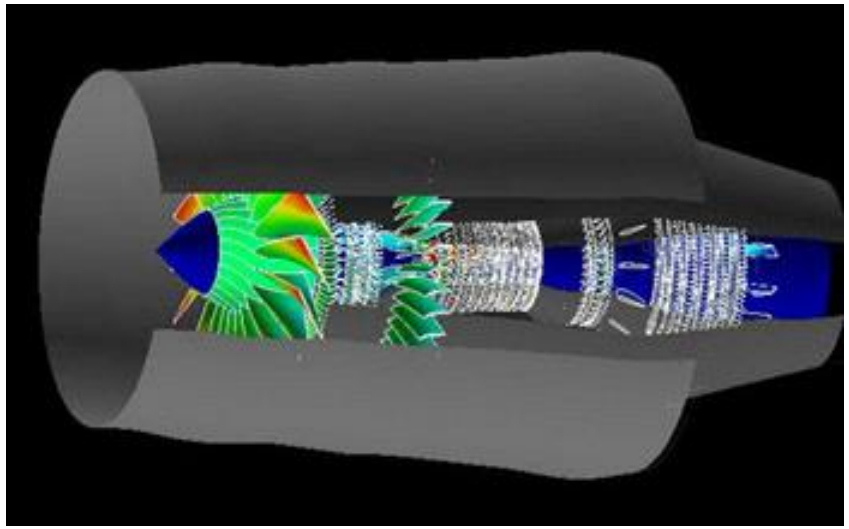
Pfizer

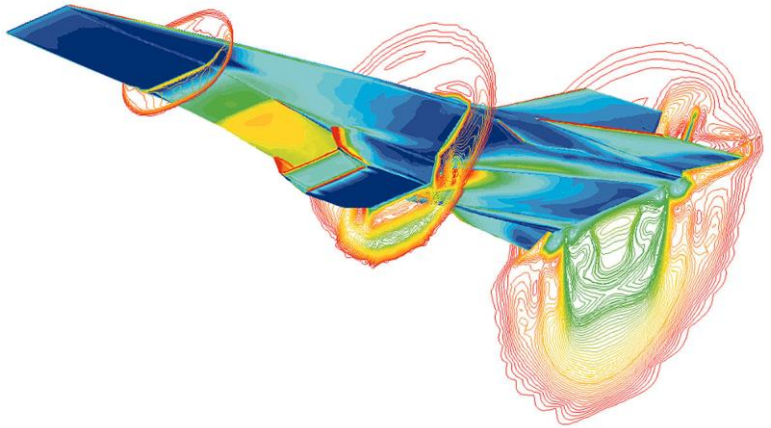
From a large dataset of crystal structures to an understanding of the chemical space of pharmaceuticals

Andy Maloney (CCDC)
Rebecca Mackenzie (STFC)
Kevin Roberts (U. Leeds)
Helen Blade (AZ)
Bob Docherty (Pfizer)

The importance of data

What if we designed drugs like we design aeroplanes?



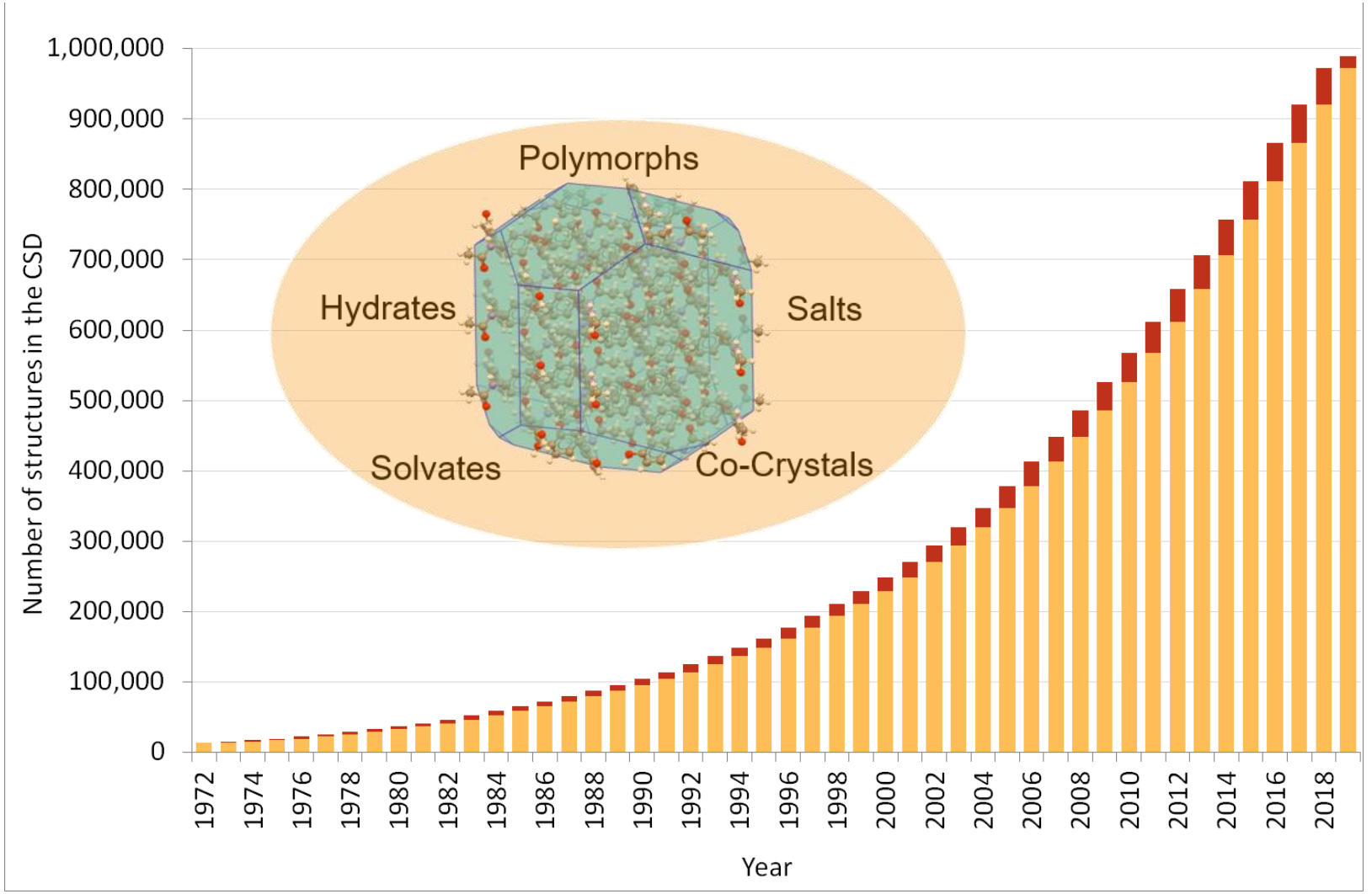


W.S Woltosz, *J. Comput. Aided Mol. Des.* (2012), **26**, 159-163

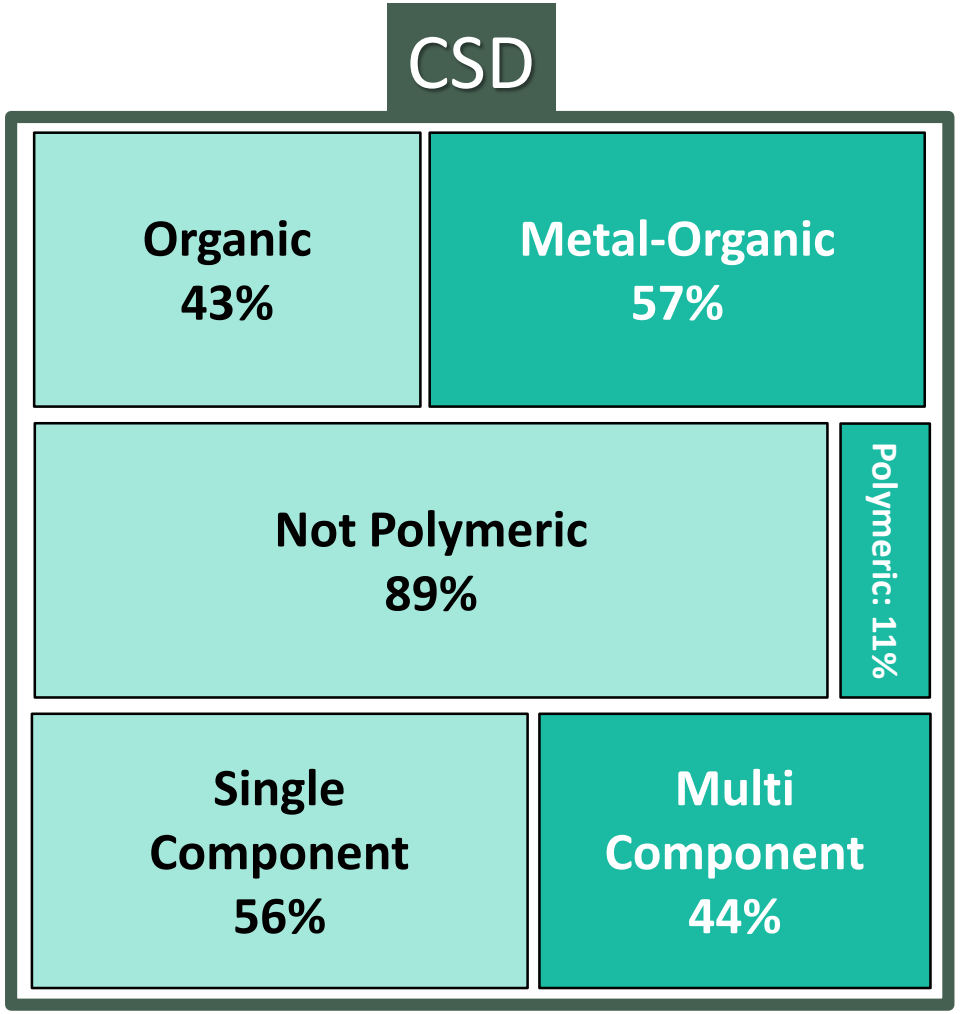
The evolution of digital design

The Cambridge Structural Database (CSD)

The world's repository of experimentally determined organic and metal-organic crystal structures



The Cambridge Structural Database (CSD)



- Drug discovery and development
- Materials design
- Crystal engineering
- Data validation

The CSD Drug Subset

Drug definition taken from a database of approved drugs

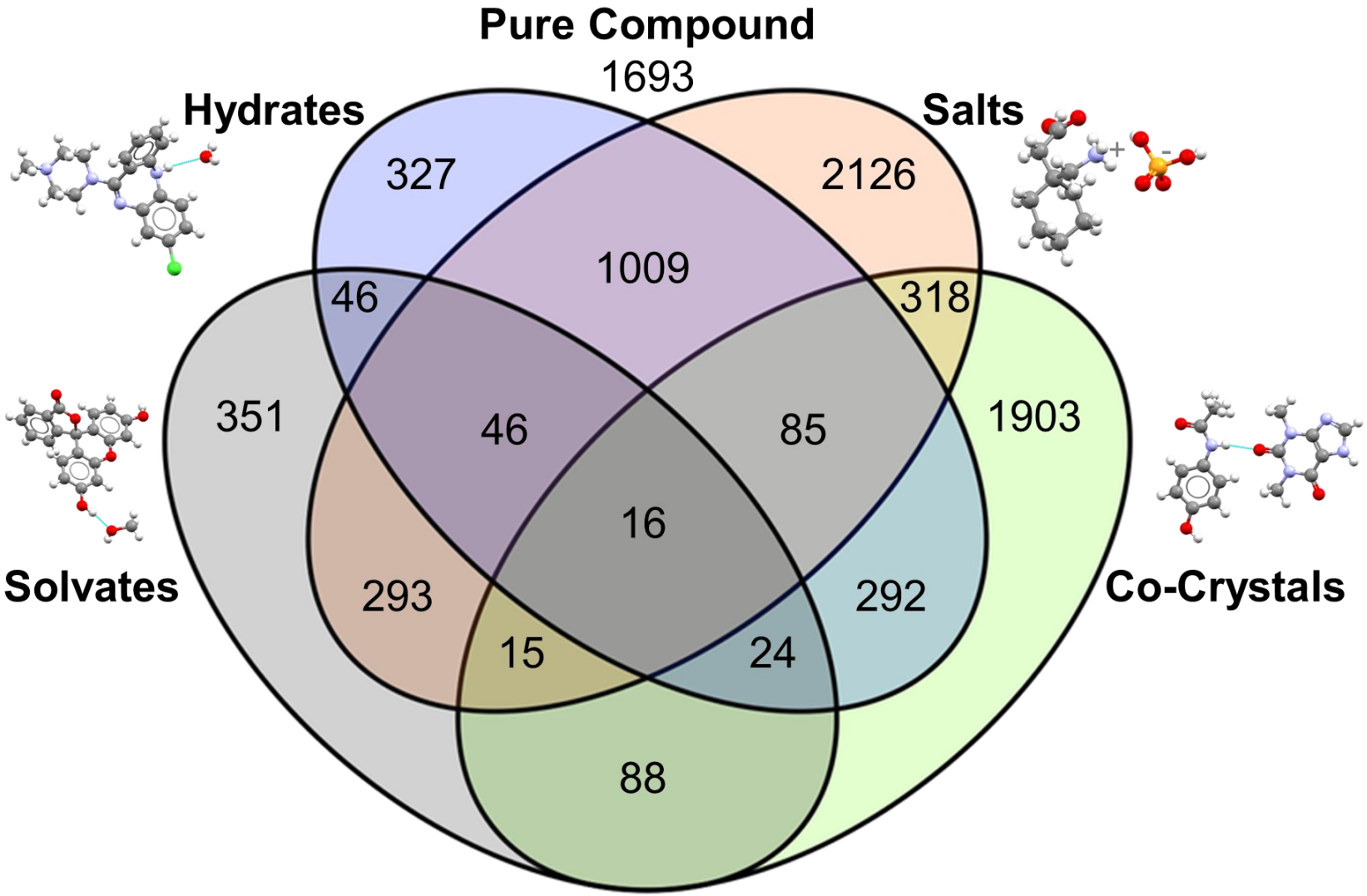
DrugBank
Open Data Drug & Drug Target Database

Extract these drugs from the CSD and create a subset that is searchable and sortable by categories such as solid form

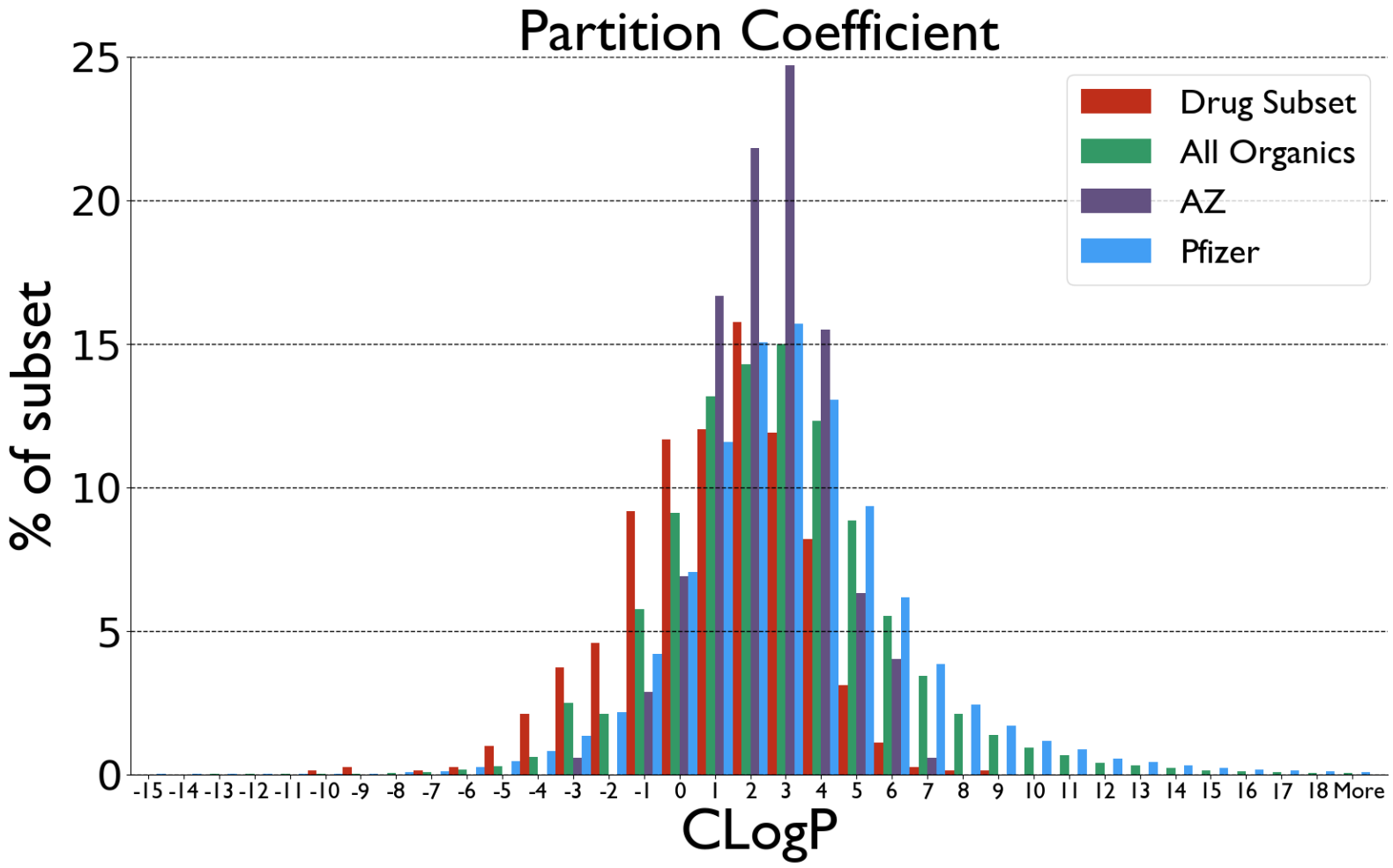
Understand the chemical space by calculating molecular and crystallographic descriptors

Repeat that process for the in-house databases of Pfizer and AZ

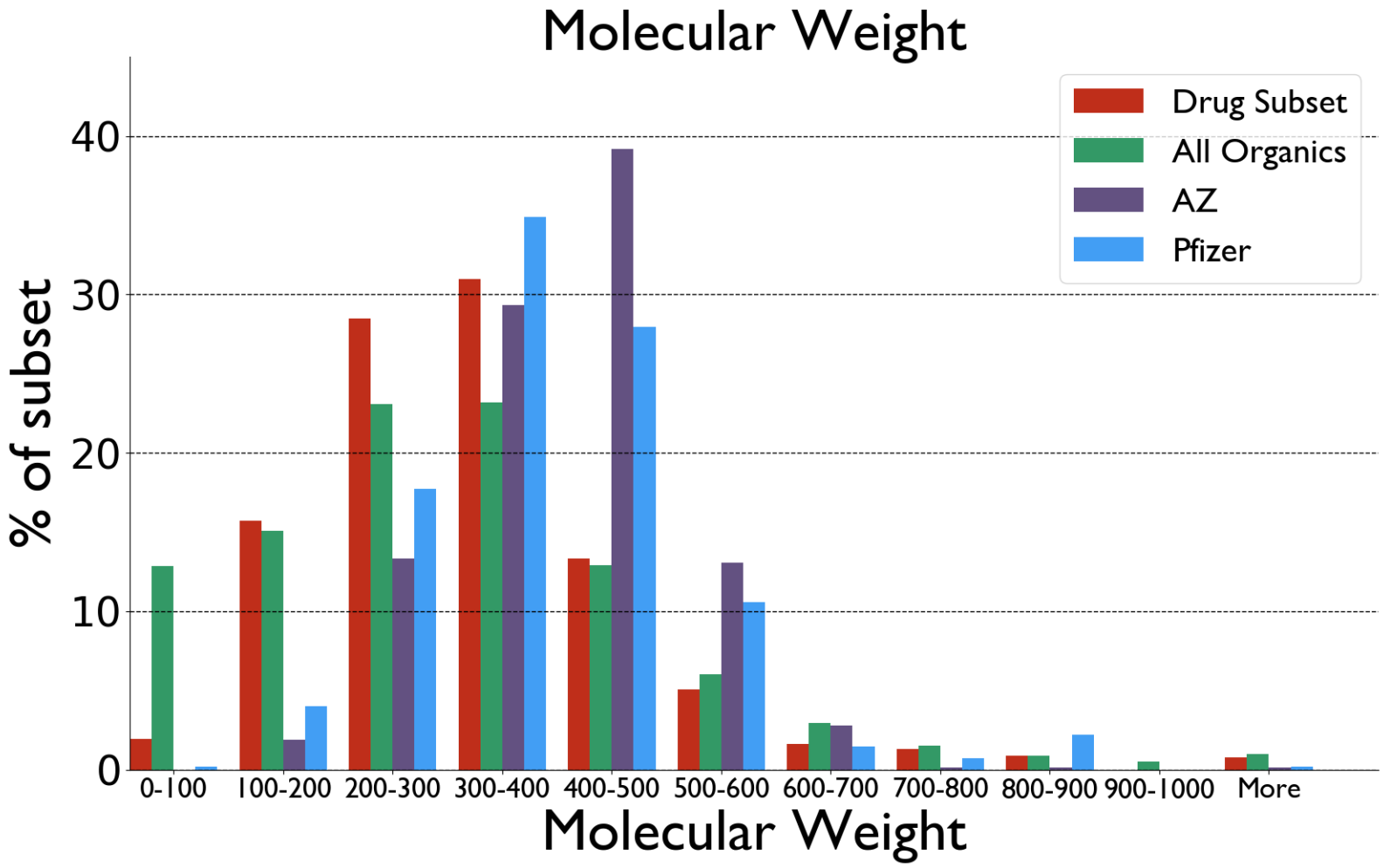
The CSD Drug Subset



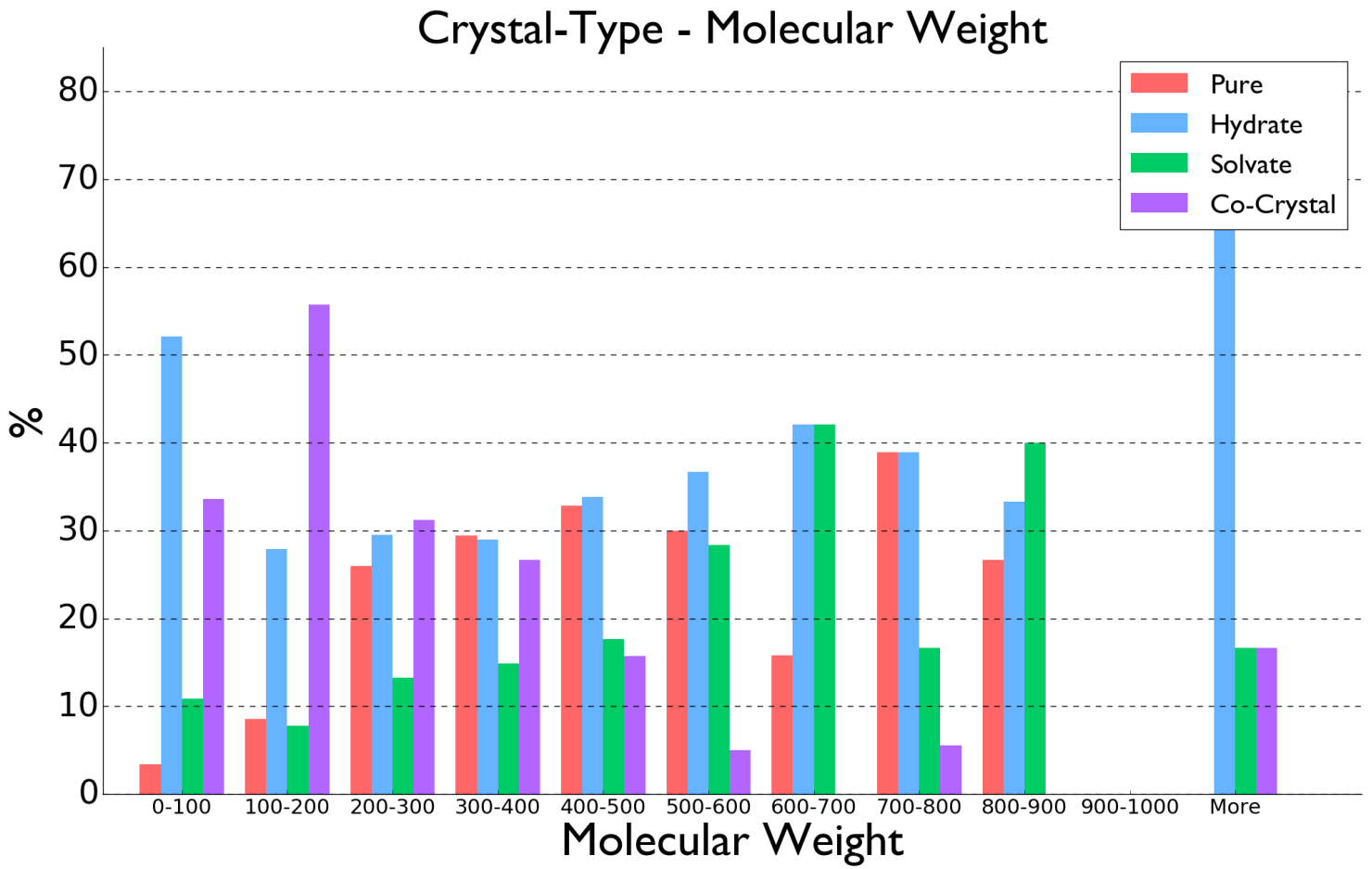
The CSD Drug Subset – Molecular descriptors



The CSD Drug Subset – Molecular descriptors

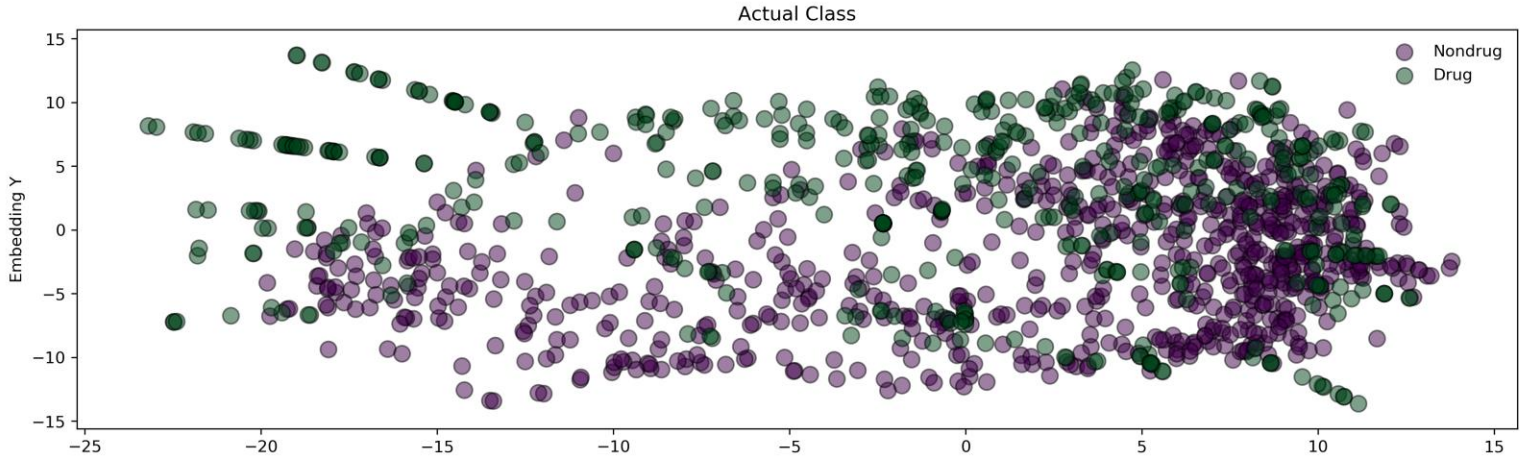


Linking molecular descriptors and solid forms

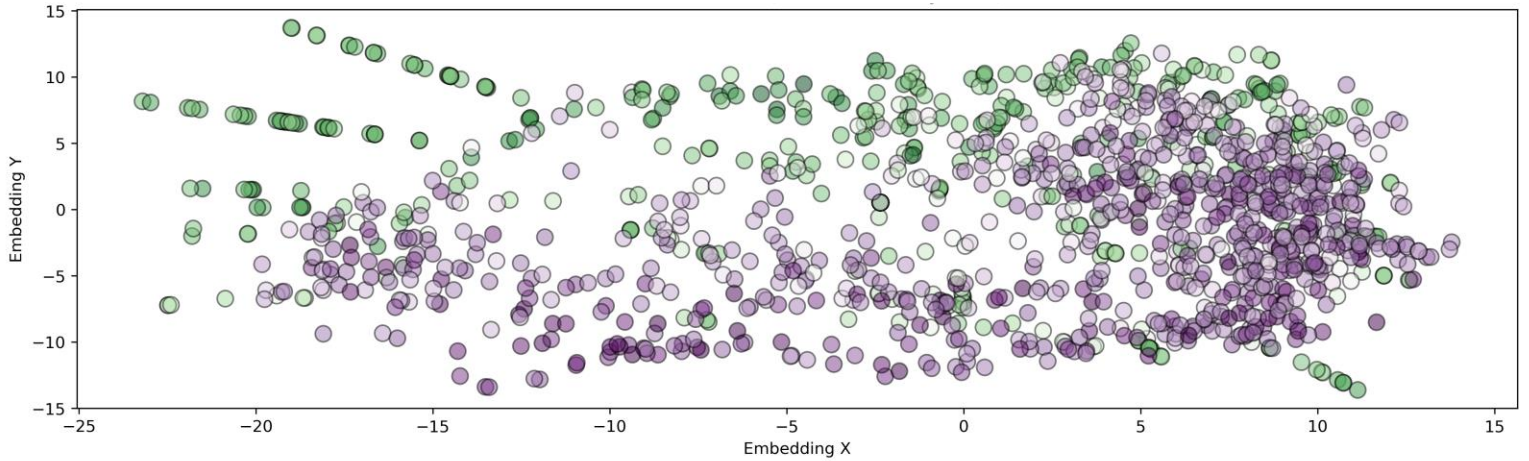


New definitions of “drug-like”?

Actual classes:



Predicted classes:



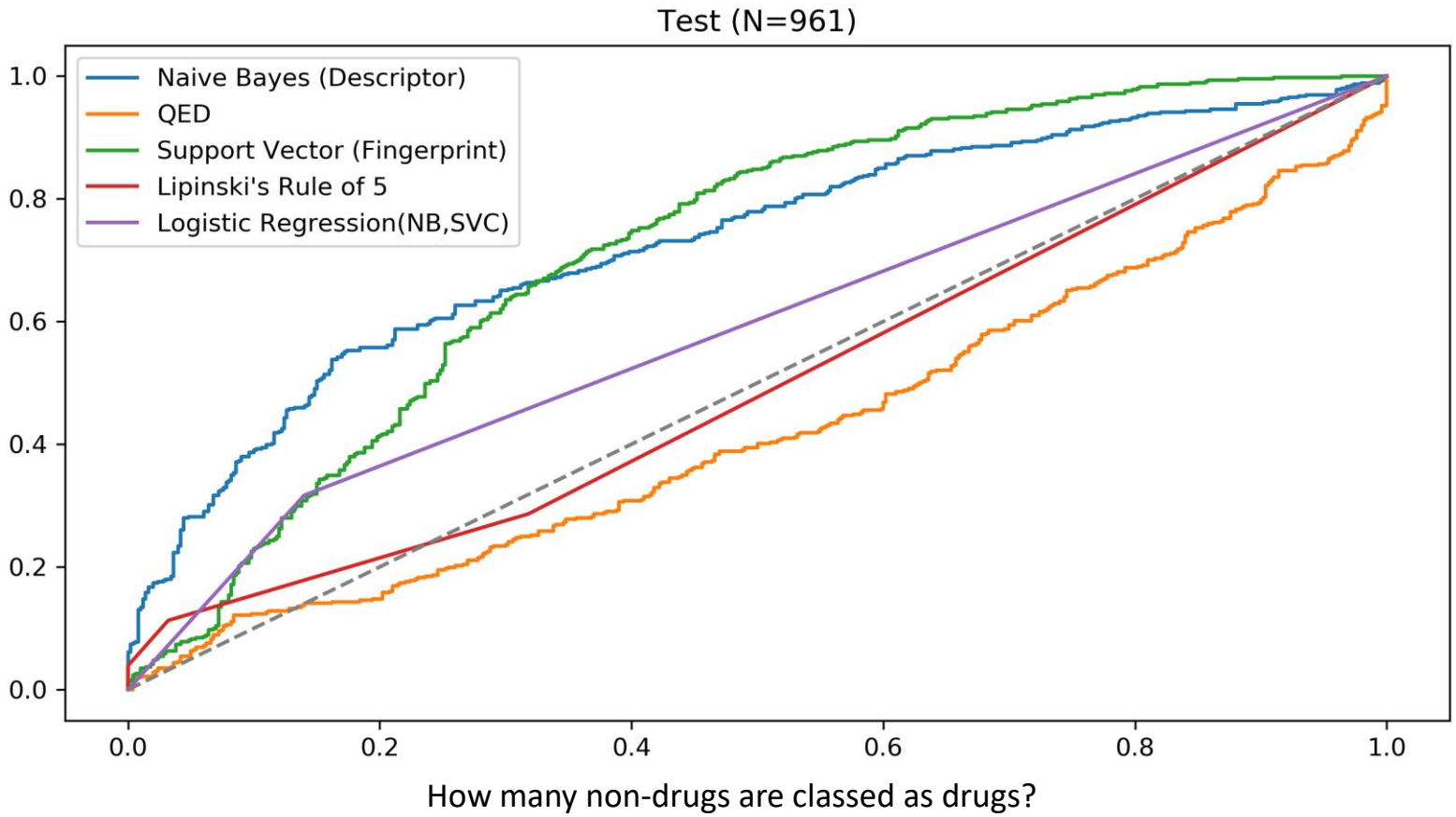
The less intense the colour, the less sure the classifier is.

85% accuracy

Embedding allows us to show a 2D representation of the feature space.
Molecules very ‘alike’ are close together in this space.

Predicting drug-likeness

How many drugs are classed as drugs?

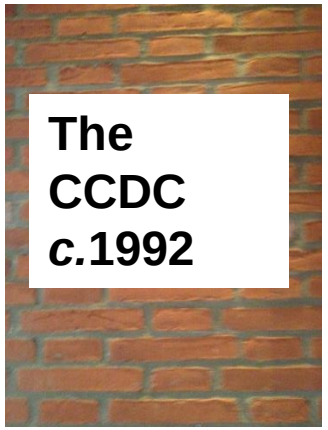


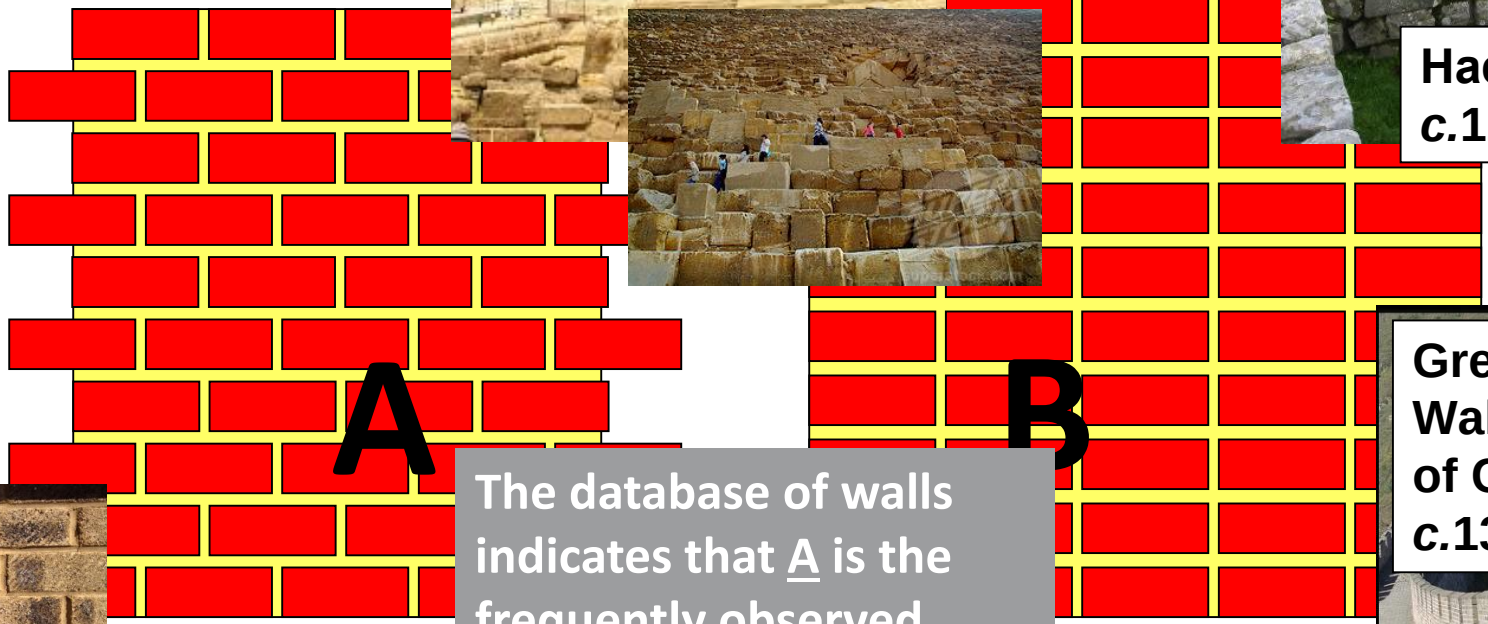
The dotted line indicates random classifications.
QED and Lipinski's do worse than random on this test set.

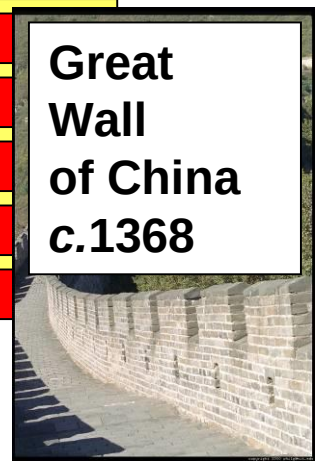
The higher and to the left, the better the model is at finding **known drugs**. Naïve Bayes is very good at this.
The higher and to the right, the better the model is at finding '**drug-like**' – (non-drugs as drugs, whilst still finding as many known drugs). Support vector is very good at this.

What else can the data tell us?

Which is the stable wall?

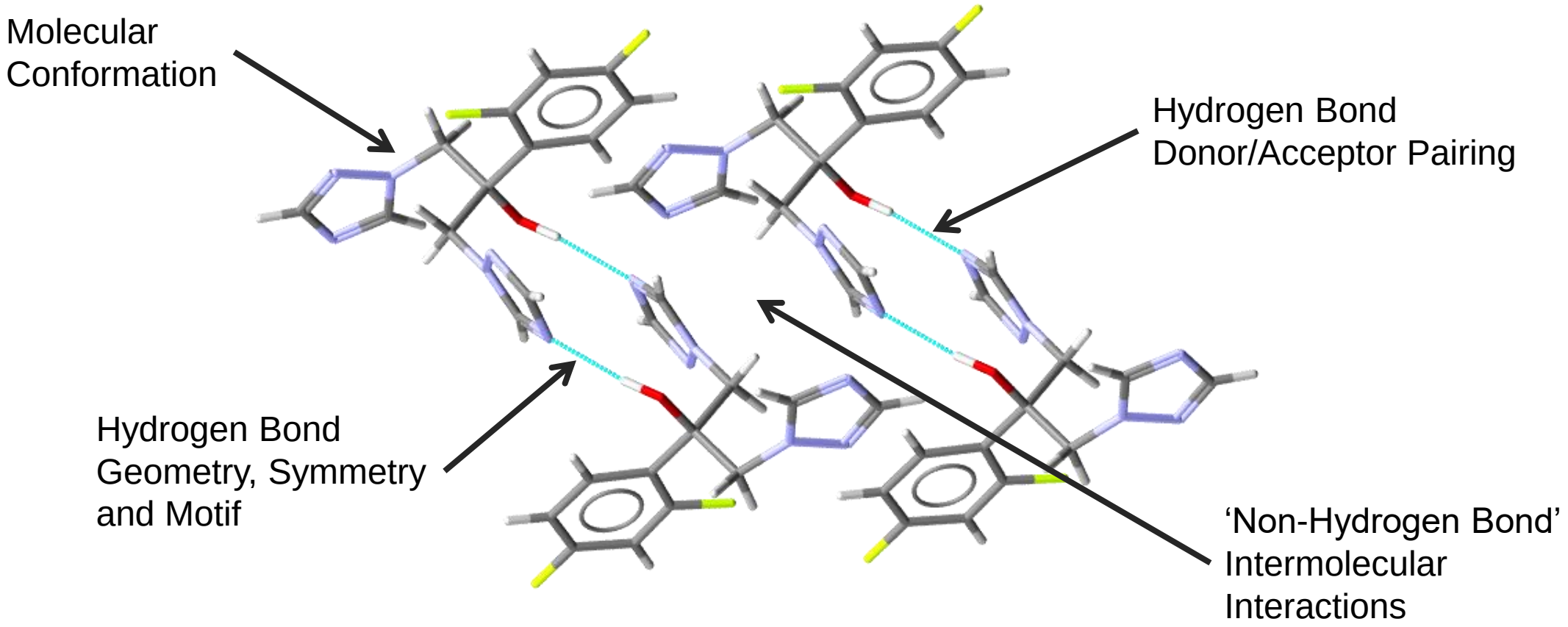




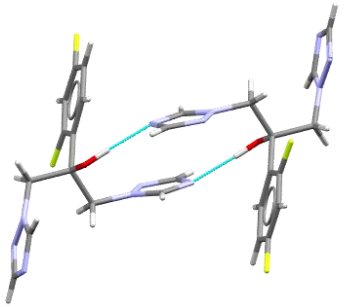


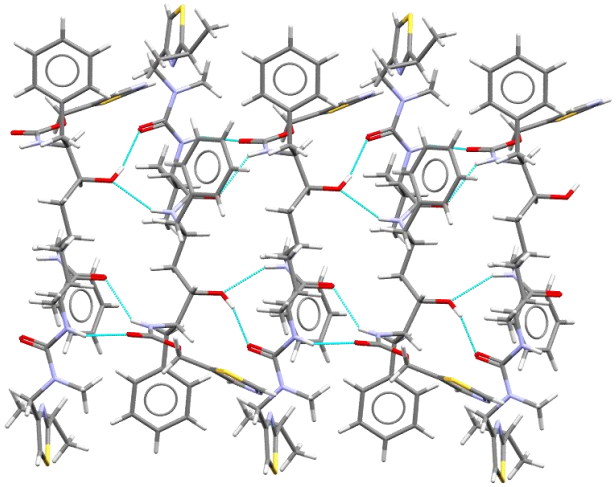
The database of walls indicates that A is the frequently observed arrangement and therefore the one that achieves stability.

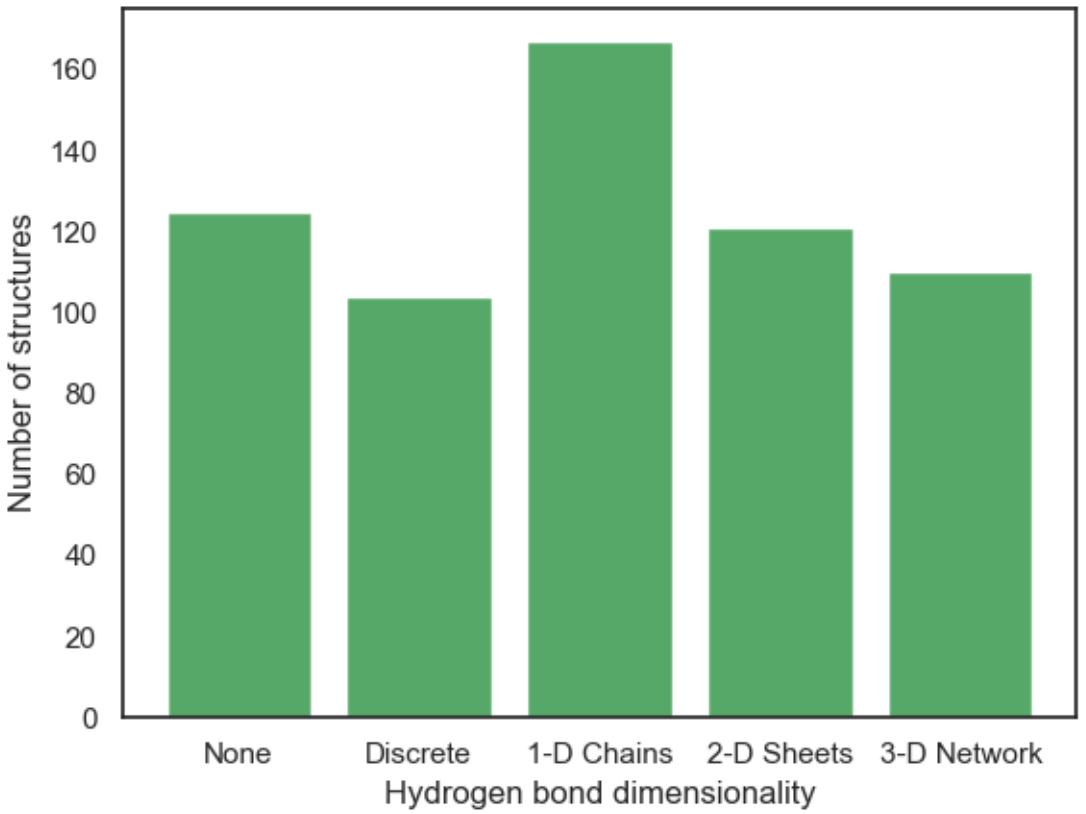
Utilising the data – “solid form informatics”

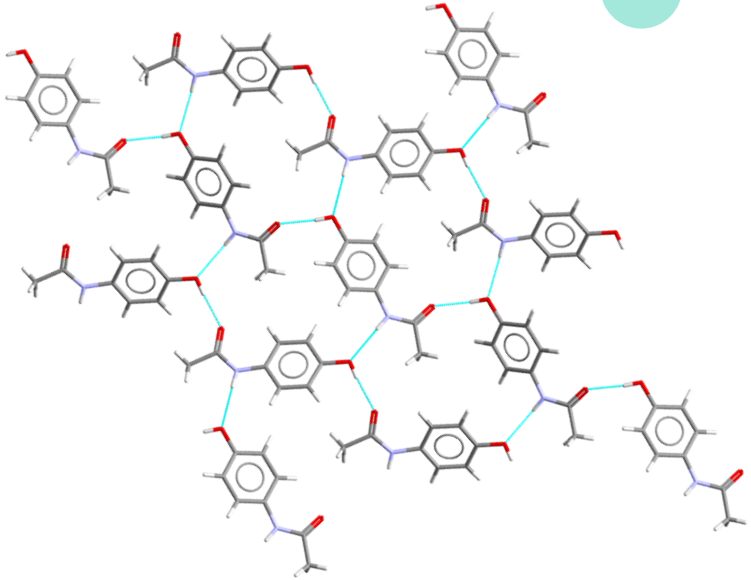


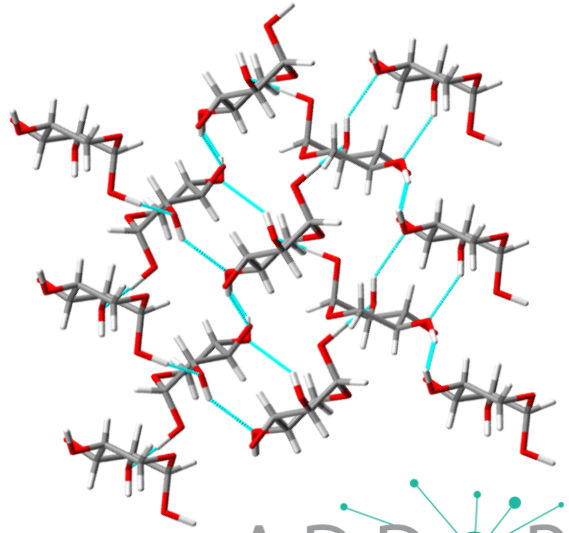
Trends in structural features



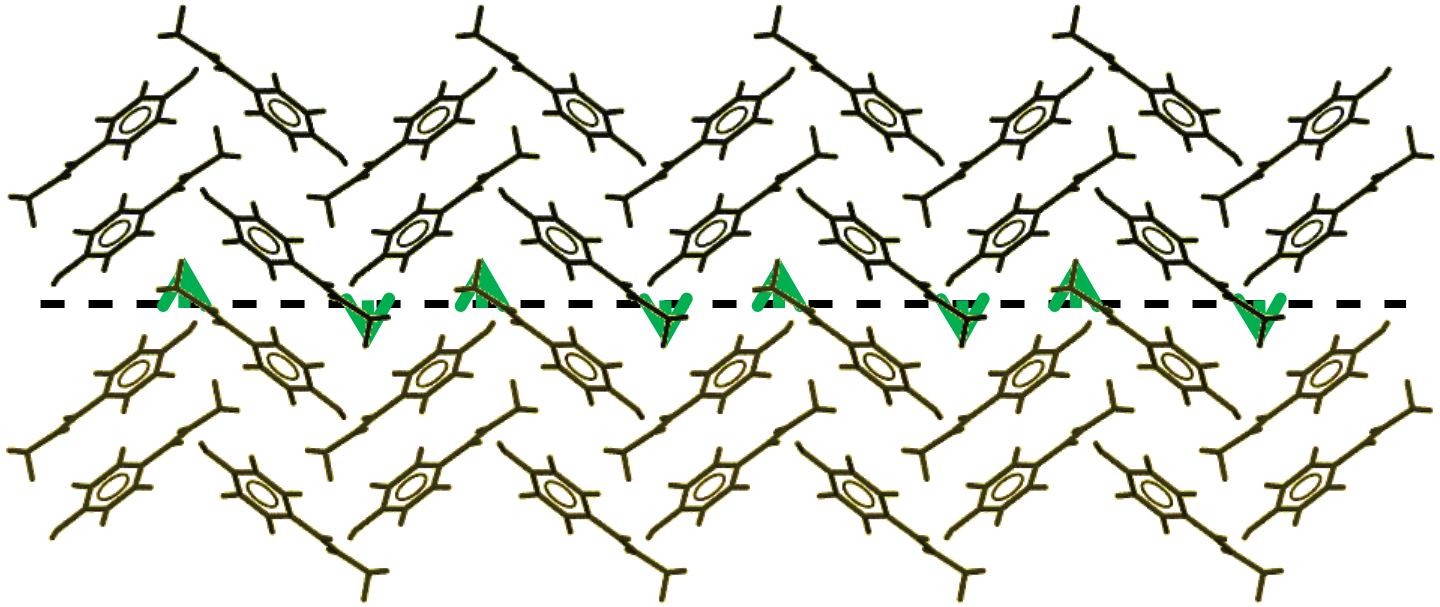


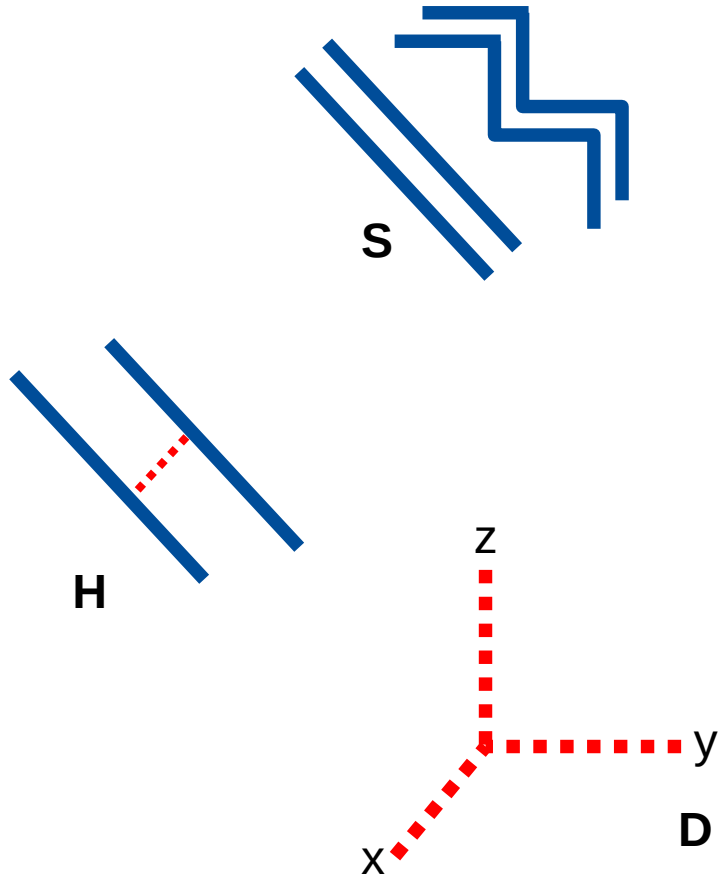




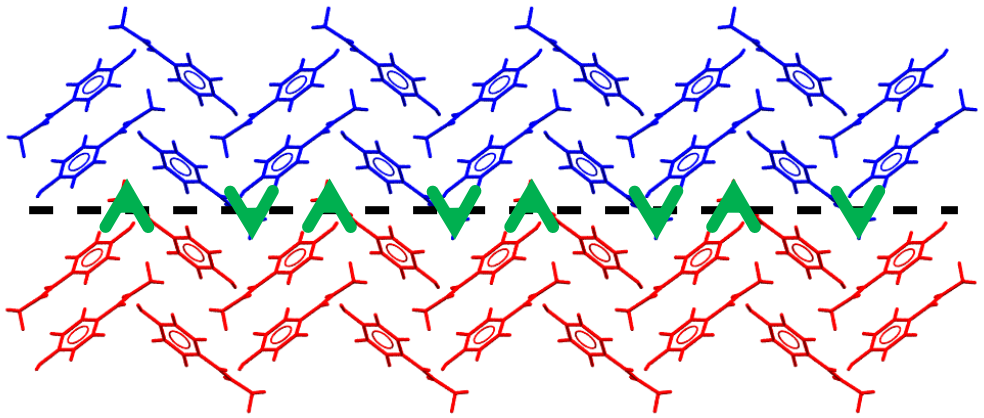


Combining structural features

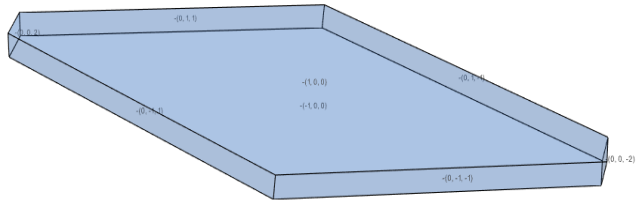




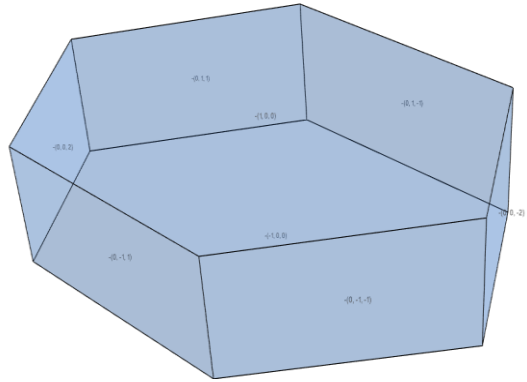
Building on the data – “particle informatics”

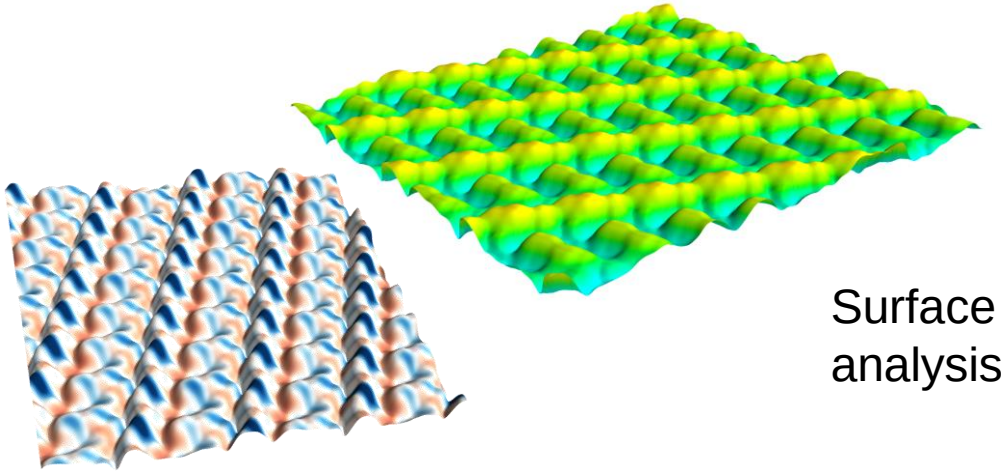


Mechanical properties



Morphology prediction



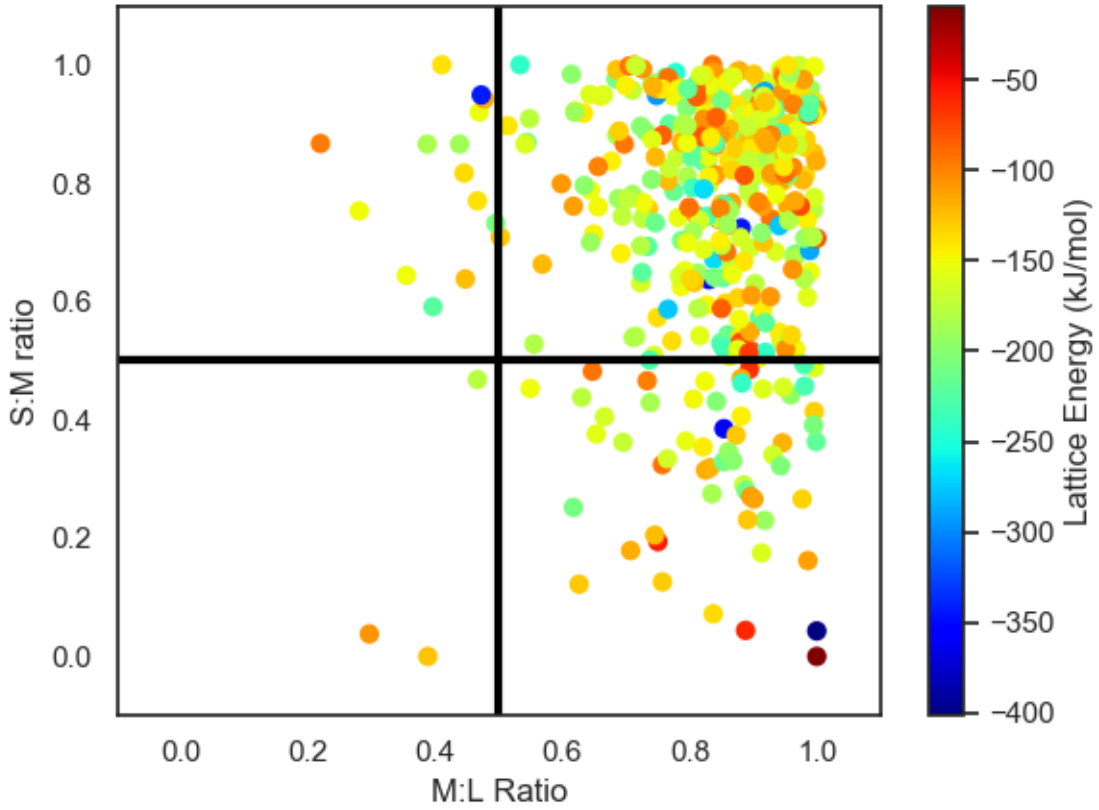


Surface analysis

Trends in particle properties

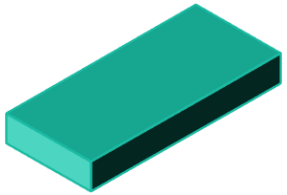
"needle"

Lattice energies and predicted morphologies

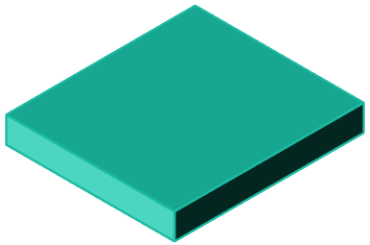


"block"

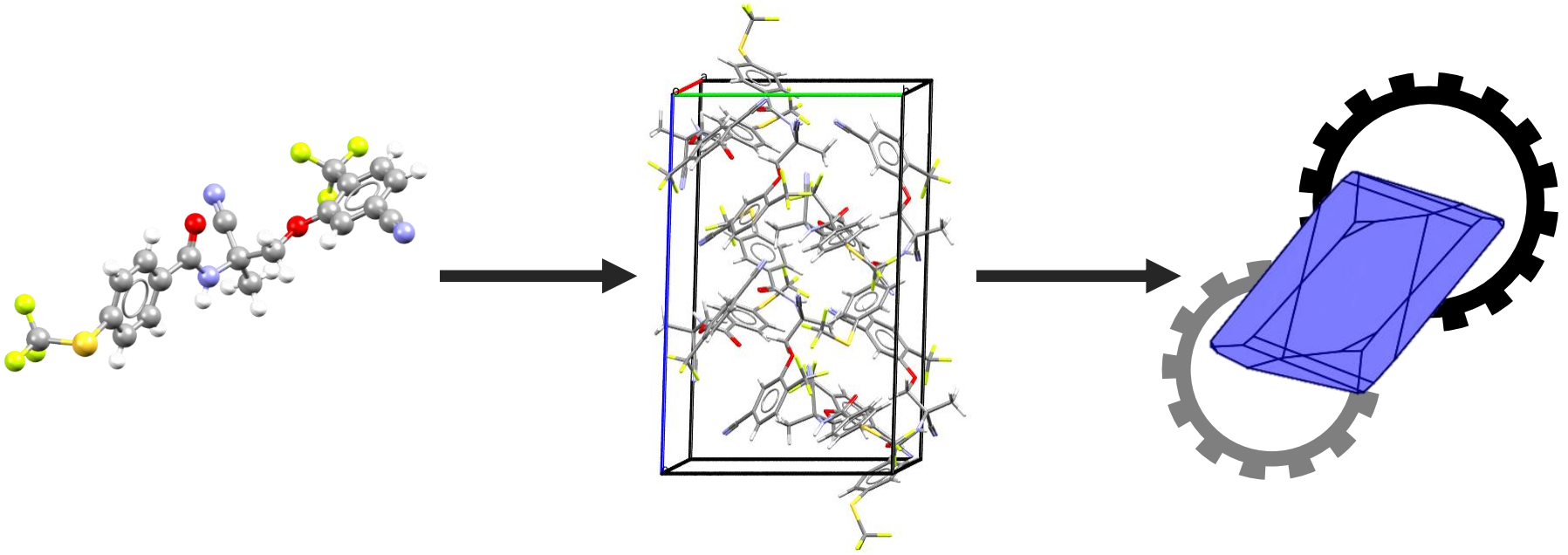
"lath"



"plate"

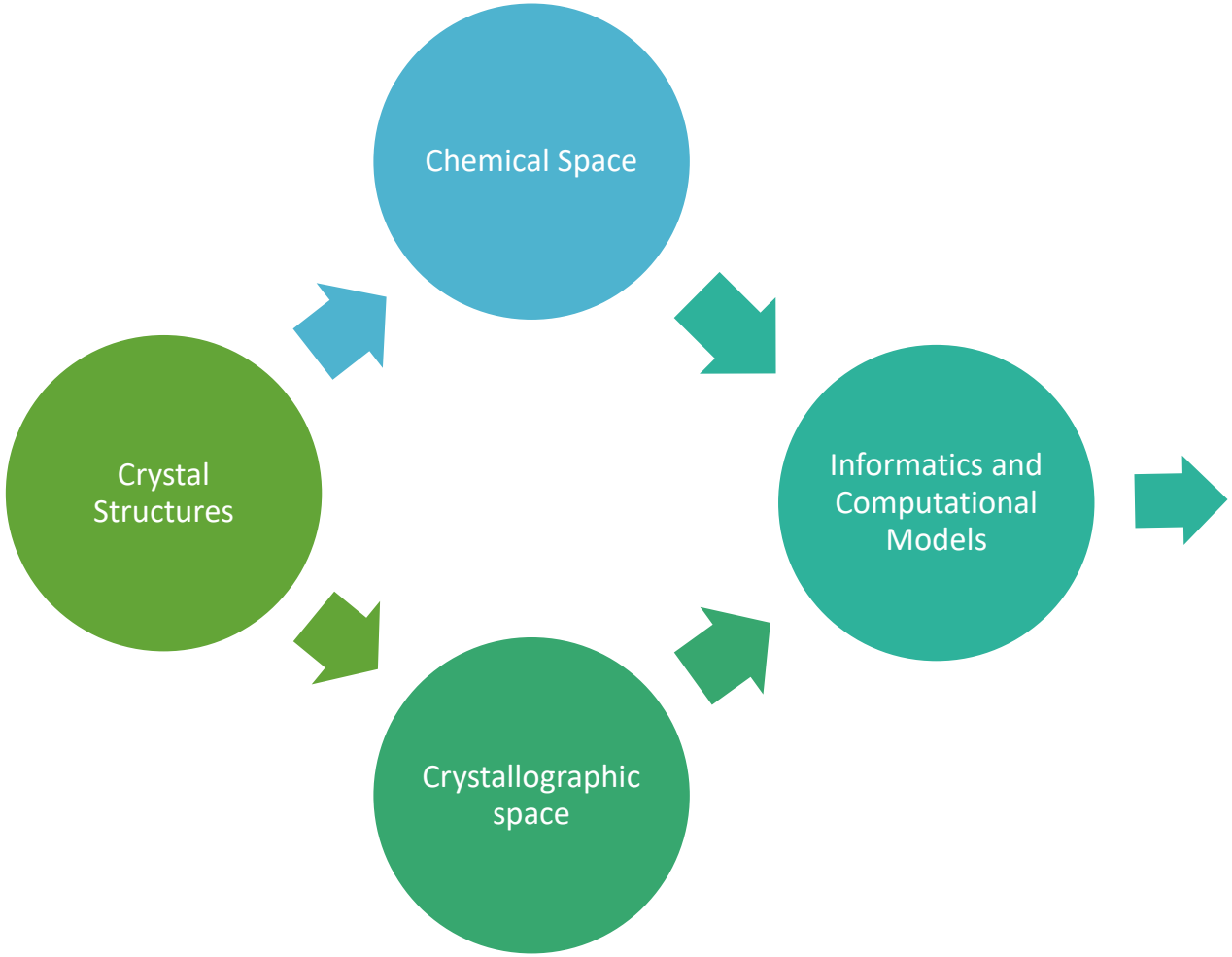


From molecules to materials



Molecule **Form** **Properties**

Summary



Industrial Impact

Digital design workflows to understand the interplay between molecules and materials

Acknowledgements

Mat Bryant, Ian Bruno (CCDC)
Dawn Geatches, Chris Morris (STFC)
Jonathan Pickering, Ian Rosbottom, Robert Hammond (U. Leeds)
Stefan Taylor (AZ)
Kevin Back (Pfizer)