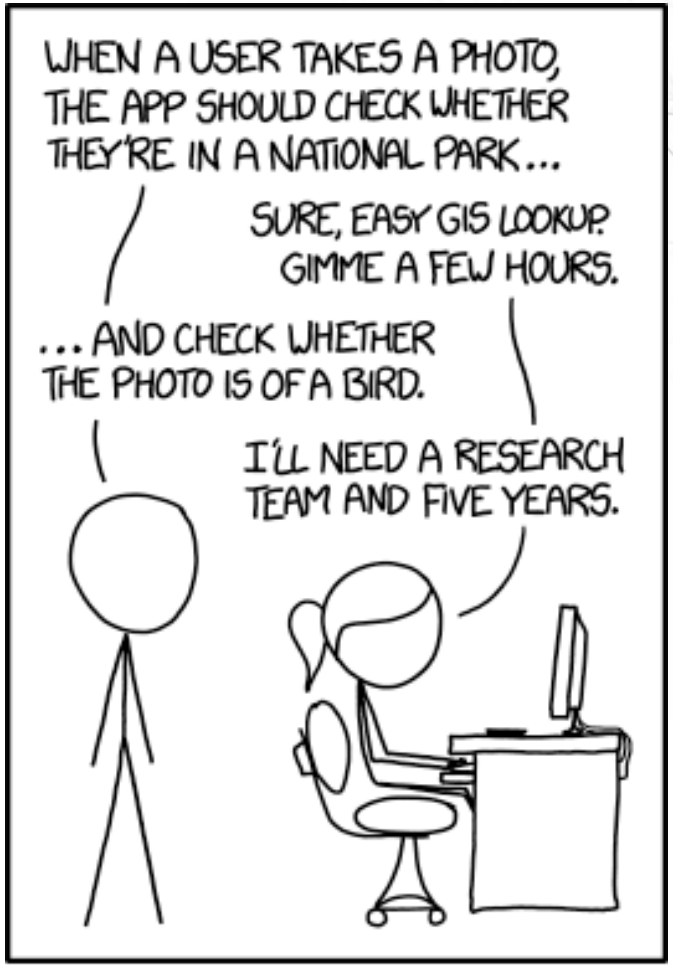
Collection and Organisation of Crystallisation Data

David Lowe

Wednesday 6th March 2024

CCDC
advancing structural science

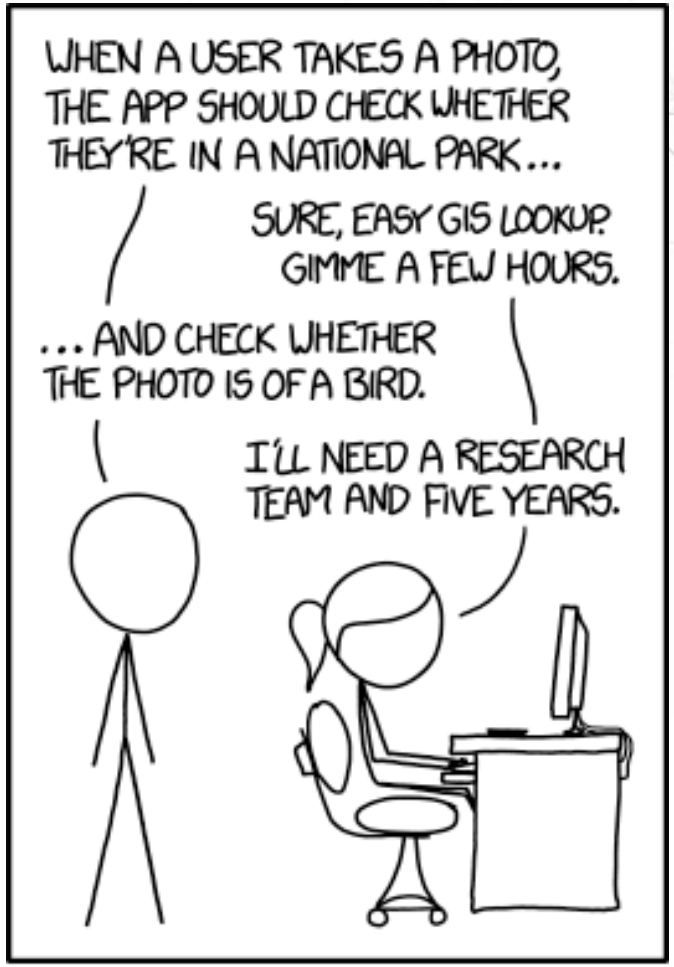
Introduction



https://xkcd.com/1425/

Introduction

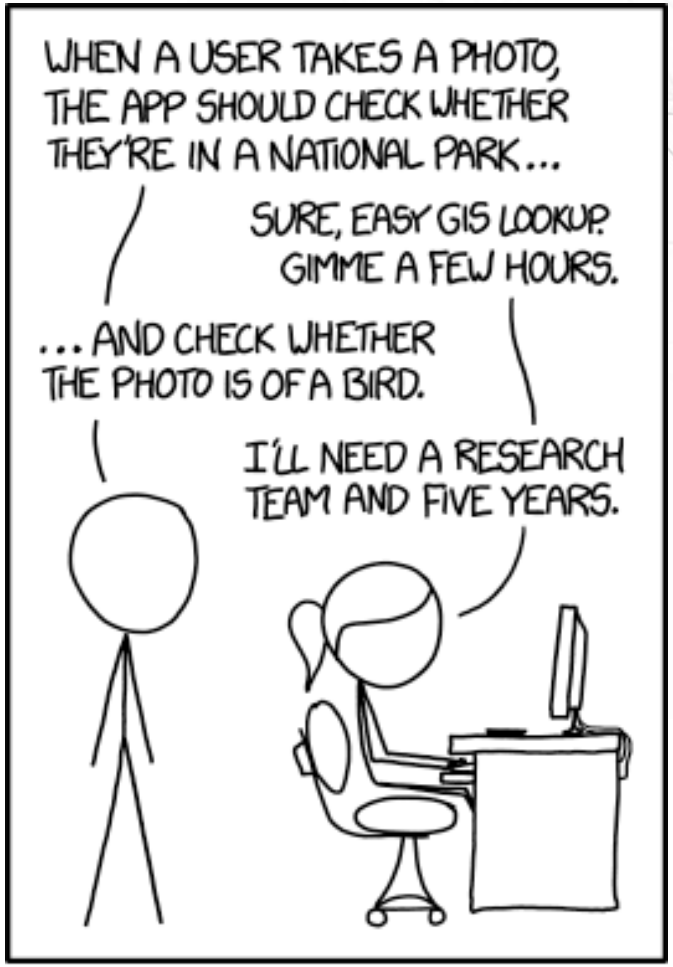
- Technology is advancing incredibly quickly
- Things that might seem impossible now could be feasible soon…



https://xkcd.com/1425/

Introduction

- Technology is advancing incredibly quickly
- Things that might seem impossible now could be feasible soon…
- … but only if there is training data available



https://xkcd.com/1425/

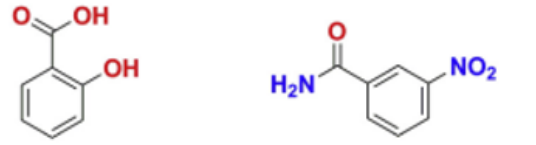
Introduction

- The CSD contains complete and curated data about crystal structures
- Limited information is recorded about their creation
  - Only recrystallisation solvent
- Past conversations indicate crystallisation processes are of industrial interest
- CSD entries have associated publications that can have knowledge extracted and formalised

Polymorphic Cocrystal Preparation

Salicylic acid: 3-nitrobenzamide (SA-3NBZ) (1:1): Cocrystallization can be conducted by 2 methods: liquid-assisted grinding (LAG) and the slow evaporation solution method (SESM). LAG: A mixture of SA (15.8 mg) and 3NBZ (16.6 mg) with a 1:1 molar ratio were weighed, mixed with drops of ethanol, then the mixture was ground for 30 s in an agate mortar. The solids were collected and dried at 40°C for 1 h. SESM: A mixture of SA (15.8 mg) and 3NBZ (16.6 mg) were dissolved in an ethanol solution, heating to 50°C for 30 min. Then, the solution was filtered by a 0.45 μm filtrate membrane in a vial. The vial was sealed with parafilm with several holes. The single crystals were obtained after 3 days.

SA-3NBZ (2:2): The cocrystal was obtained by the SESM method. A mixture of SA (15.8 mg) and 3NBZ (16.6 mg) were dissolved in a mixed ethanol: chloroform solution (1:1, v/v) and heated to 50°C for 30 min. The solution was filtered with a 0.45 μm filtrate membrane and transferred into another vial. The vial was sealed, poking several holes on the surface. The crystals and solids were harvested after 3 days.



Salicylic acid, SA — 3-Nitrobenzamide, 3NBZ

Scheme. 1. Chemical structures of SA and 3NBZ.

Knowledge Extraction

- Feasibility study investigating the availability of complementary data for current CSD entries
  - We looked at 1,706 entries from the CSD Drug Subset
- The following data was targeted:
  - Internal paper IDs
  - Information about crystallisation conditions
    - Components, solvents, techniques, conditions
  - Non-crystallisation data
    - Melting points and phase transitions
    - Solubility measurements
    - Bioactivity

CCDC

How much data was found?

- From **1706** CSD entries:

Property	Found	Already in CSD	Total coverage in sample
Crystallisation process	1591	0	93%
Bioactivity	854	430	50%
Melting point	543	91	32%
Solubility	196	0	11%
Phase transition temp.	53	0	3.1%

Recording Crystallisation Data

- The initial intention was to maximise information by recording the crystallisation process as individual steps
- A preliminary investigation of 100 entries found that the effort required was out of proportion to the value
- Limit coverage to the key features:
  - Identities and amounts of component, solvents, additives
  - Initial and equilibrium temperatures
  - Solvent evaporation?
  - Specialised technique used e.g. sublimation, vapour diffusion, LAG
- Enough information to get the same product

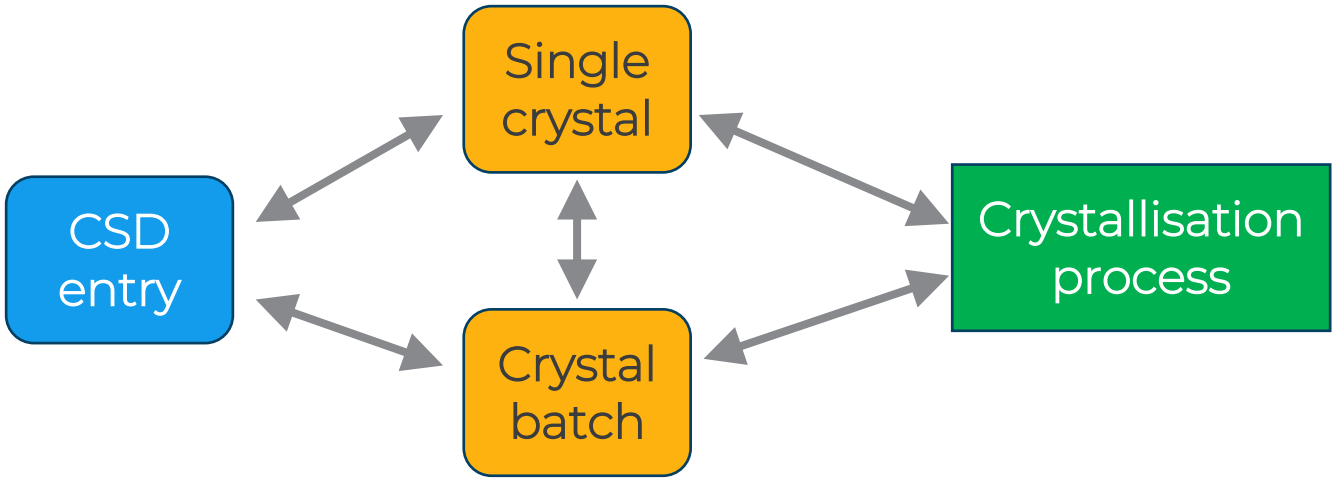
CCDC

Structuring Crystallisation Data

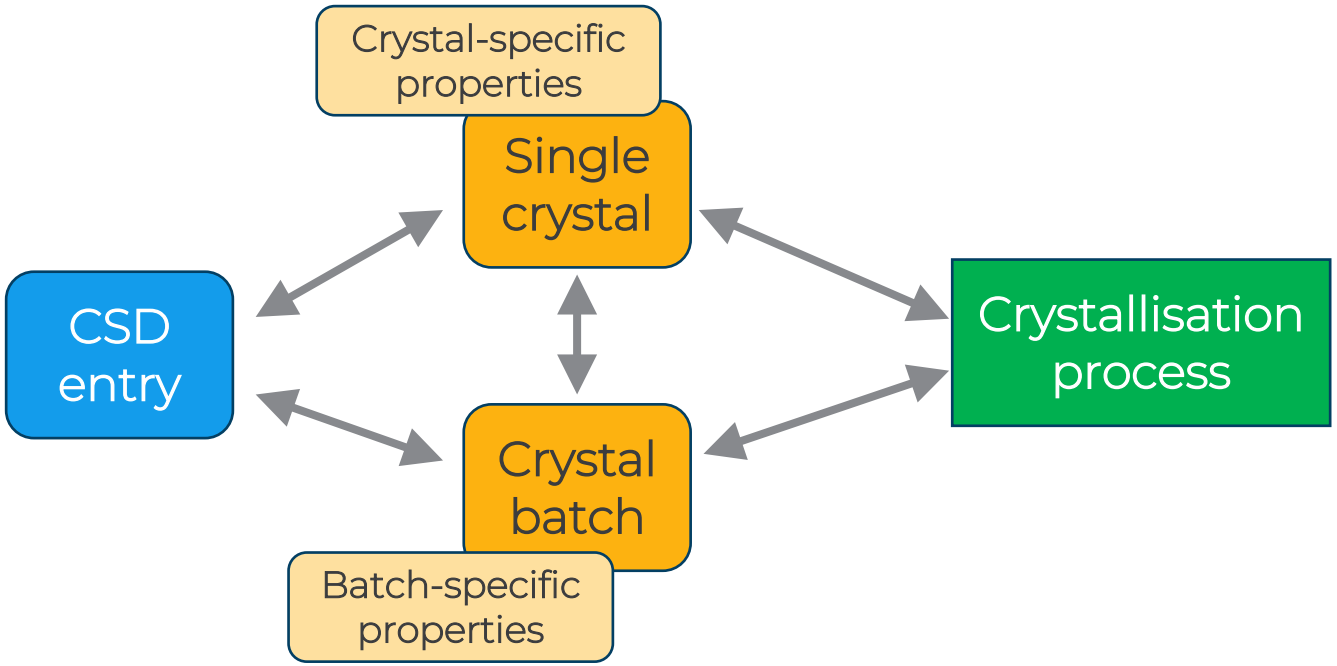
CSD entry

CCDC

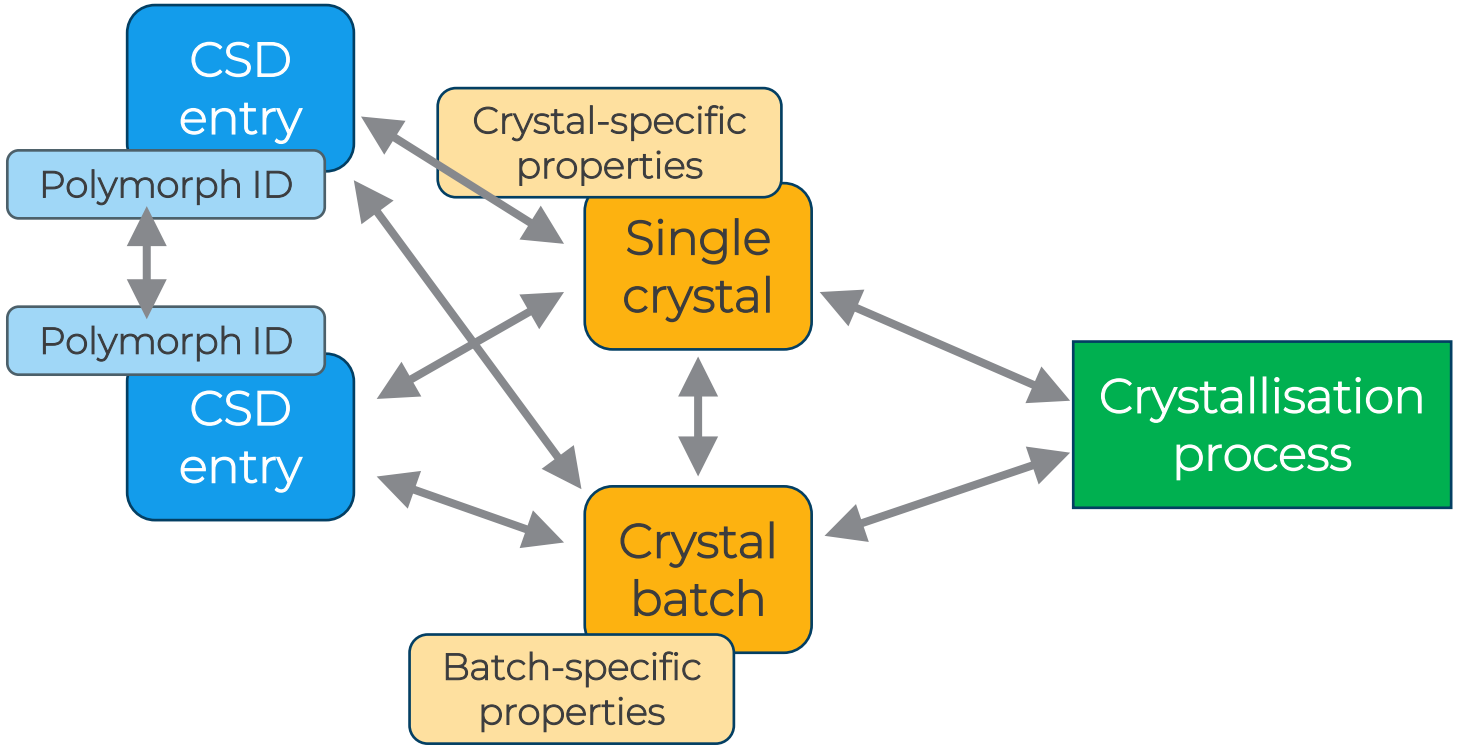
Structuring Crystallisation Data



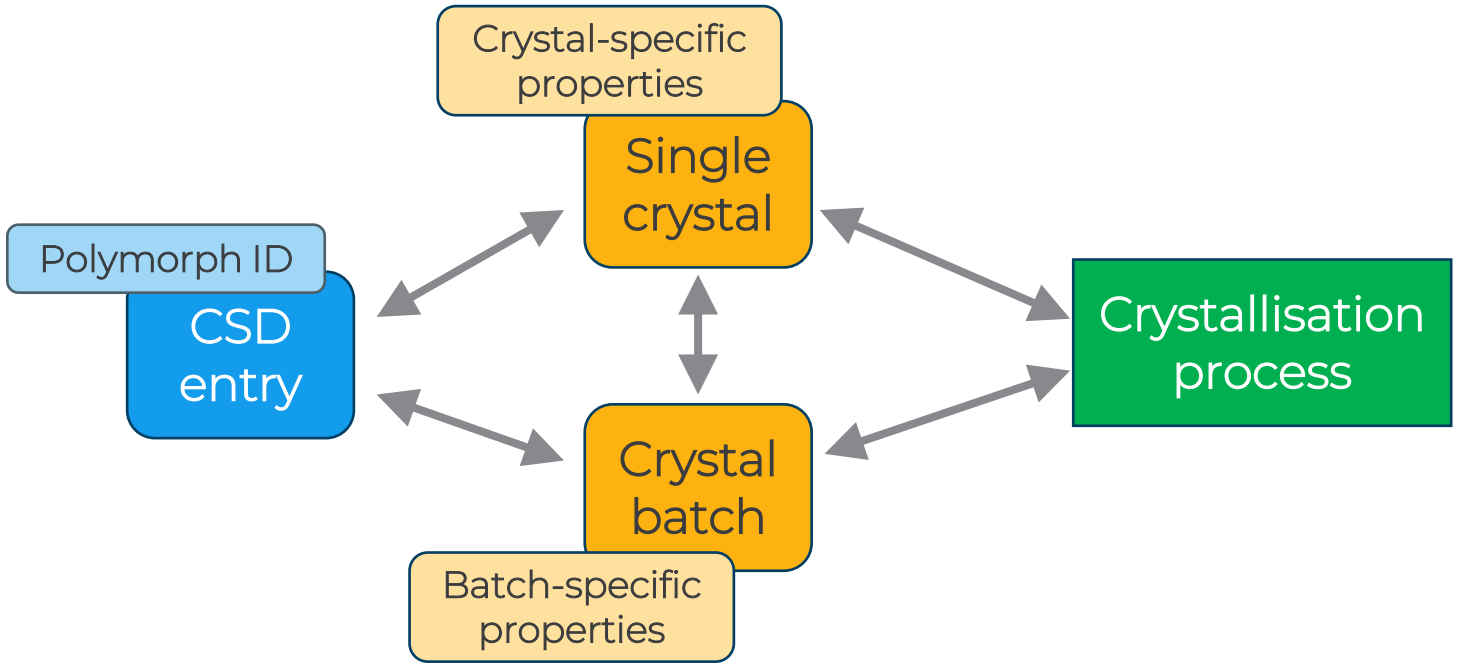
Structuring Crystallisation Data



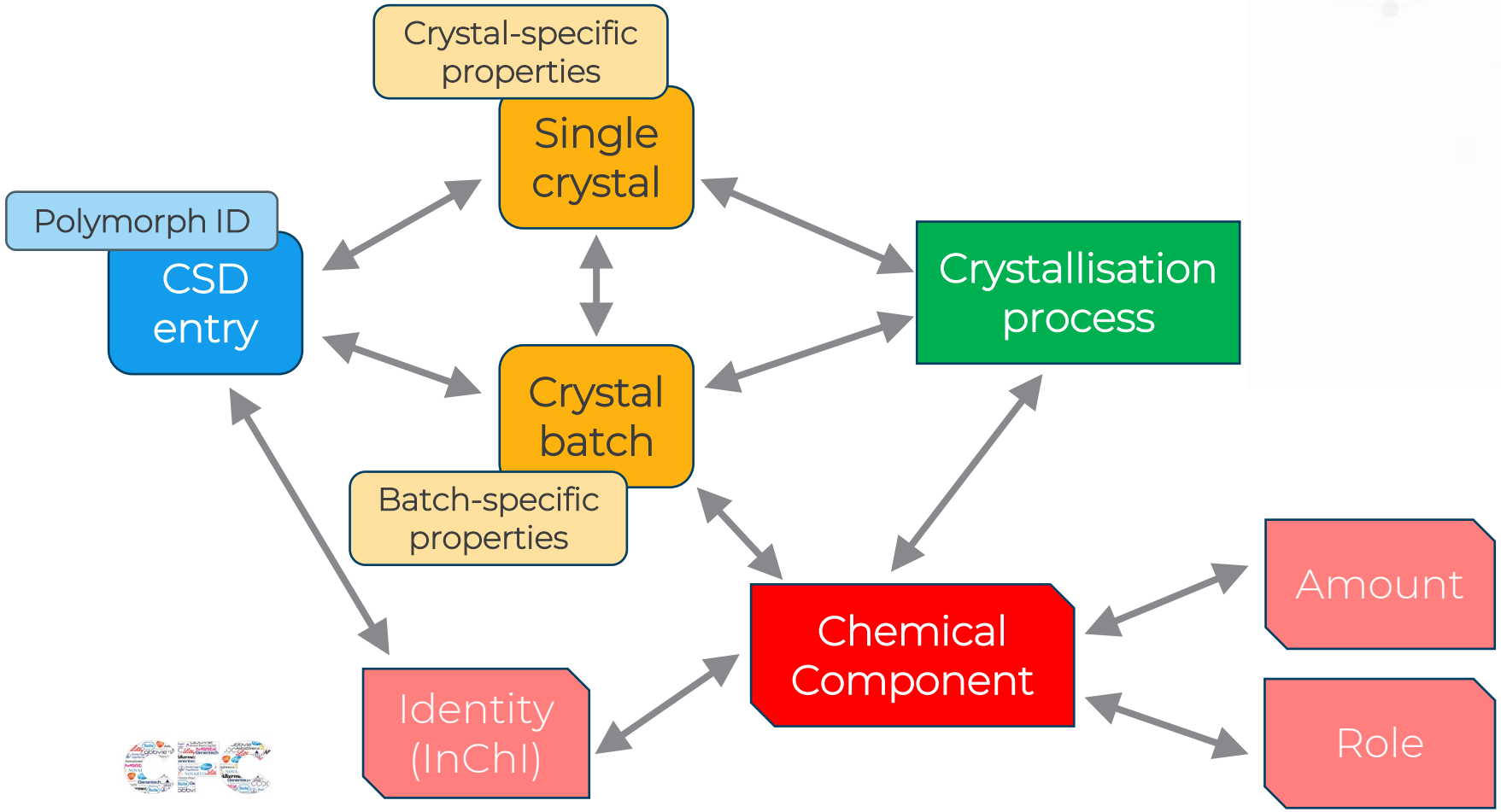
Structuring Crystallisation Data



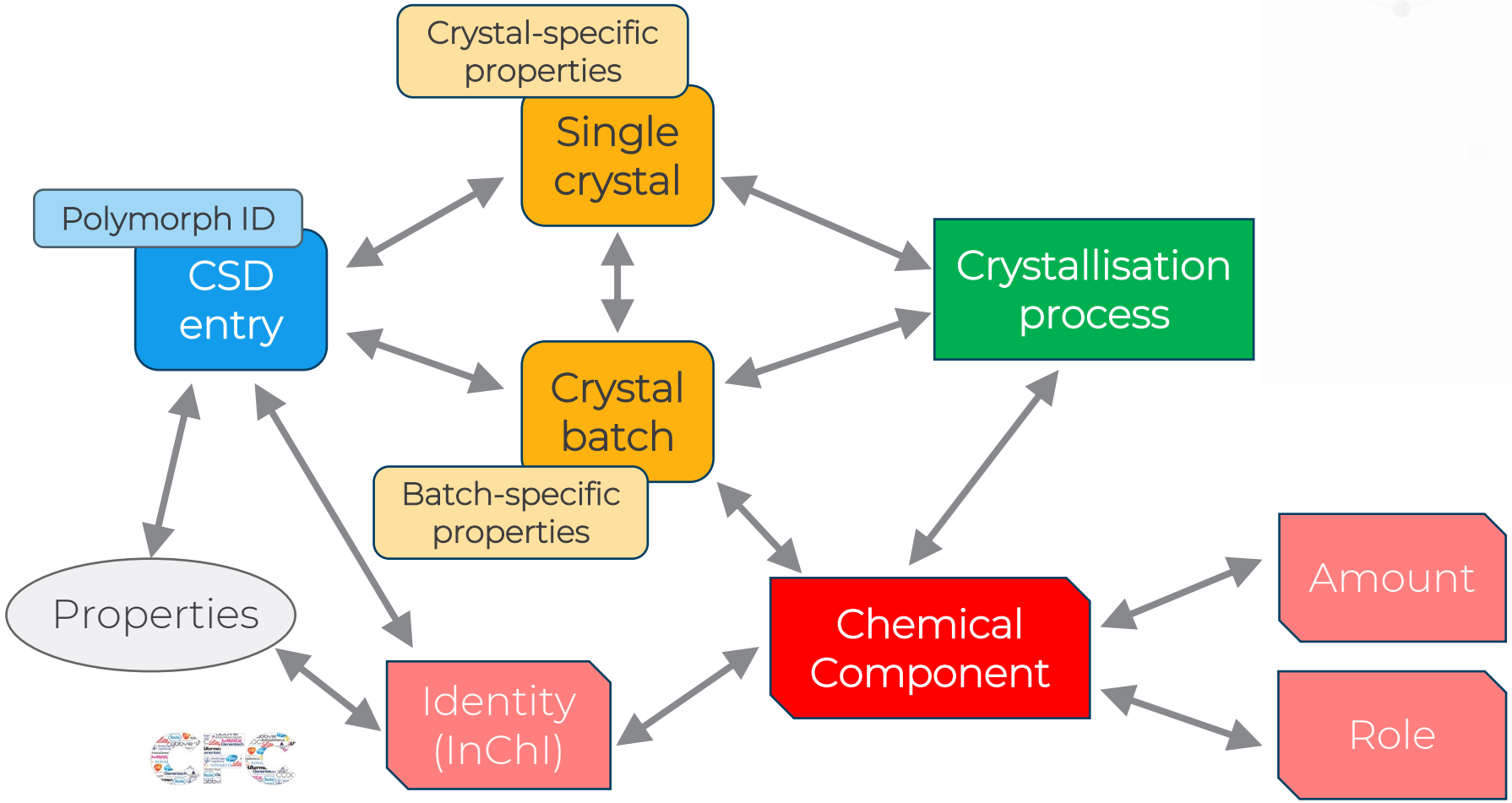
Structuring Crystallisation Data



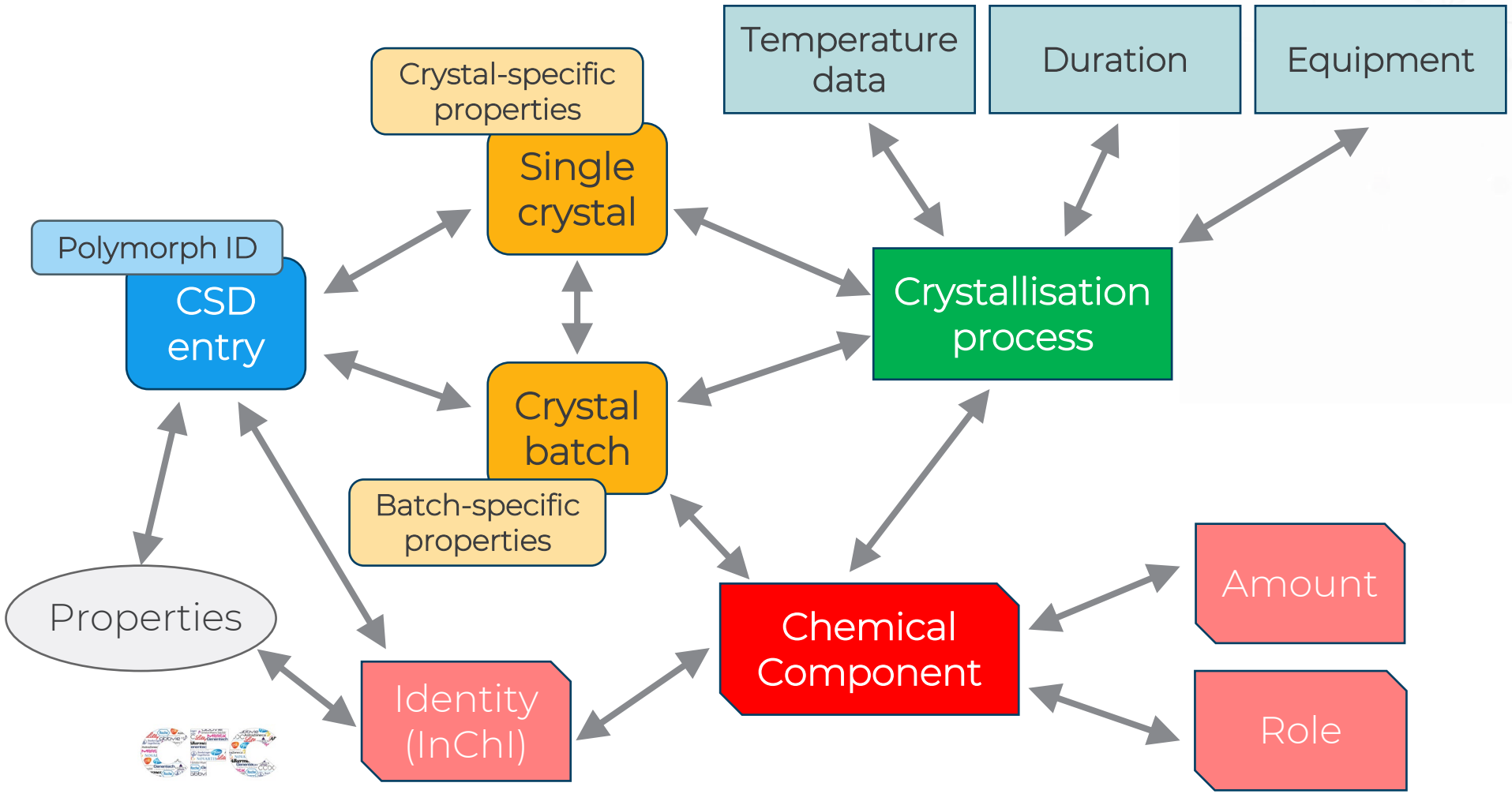
Structuring Crystallisation Data



Structuring Crystallisation Data

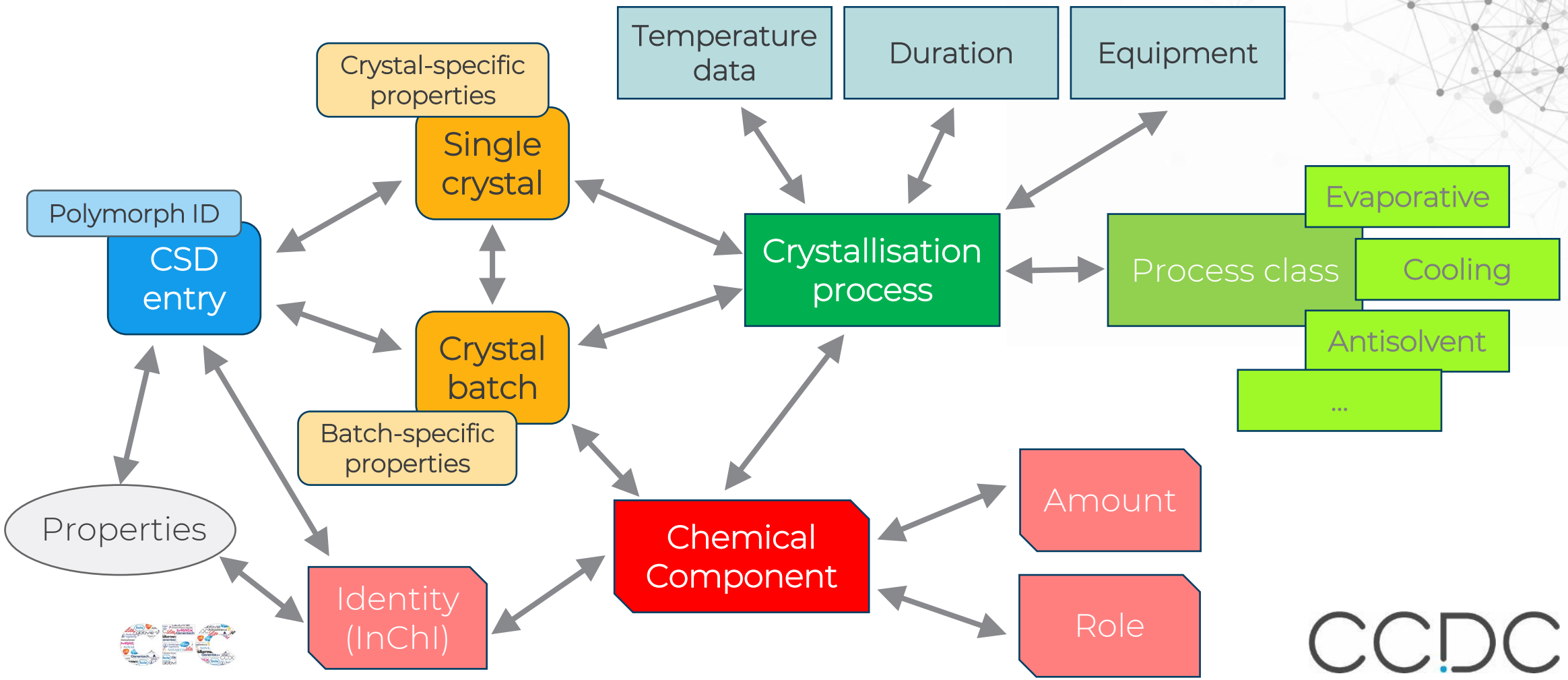


Structuring Crystallisation Data



CCDC

Structuring Crystallisation Data



Interrogating the Data

- 2080 methods for 1591 entries from 1163 families
- How many different process classes?

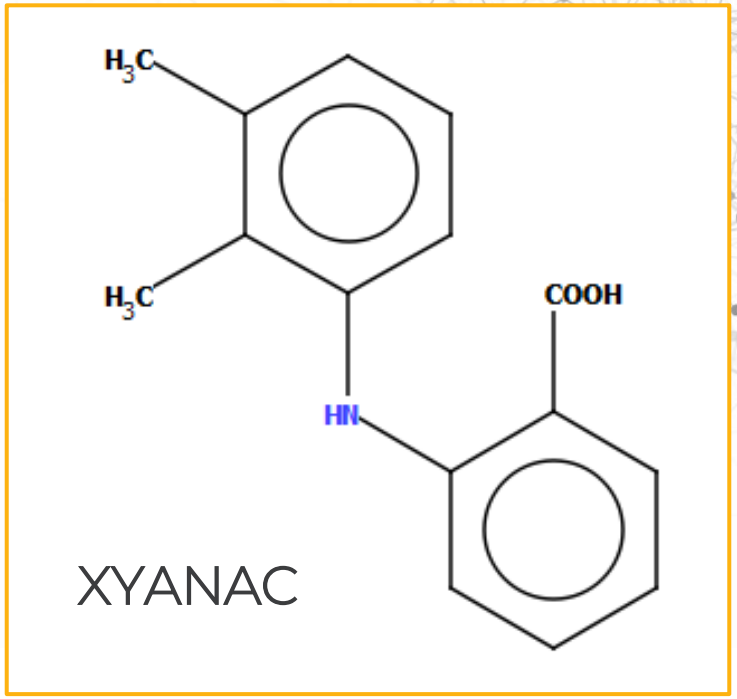
	Methods	Entries	Families
Evaporative	1220	1076	826
Cooling	281	236	194
Grinding	169	151	147
Antisolvent	116	63	57
Sublimation	86	85	19
Melt	30	26	18
Seeded	22	22	18

Polymorphs

- 45 families with two or more polymorphs in the dataset
- 25 families with two polymorphs made using the same solvent
  - 15 with both polymorphs formed by evaporation
  - 7 with both polymorphs formed by cooling
  - 4 with evaporation / no evaporation pairs
- 15 families with polymorphs with diff. solvents/same conditions

Example: Mefenamic acid

- Present in 70 organic CSD entries
- 18 entries in the Crystallisation Dataset
  - 40 methods
- Used with 8 different solvents
- How many of each process class
  - 14 evaporation
    - 6 entries that used grinding also tried separate evaporative method
  - 1 melt crystallisation
  - 1 cooling (the only single component compound)
  - 2 unclassified



Potential Applications

- Training set for Natural Language Processing of chemical literature
  - Transfer learning?
- Predicting crystallisation conditions
  - Input structural info of desired crystal components
  - Output predicted crystallisation conditions
- Aiding CSP
  - Adding features to CSP system to predict polymorph formed given specified crystallisation conditions

CCDC

How to Progress

- Invest more resources in knowledge extraction from literature?
  - How valuable is this data?
  - What might you want to do with it?
  - Any accessible data sources other than papers?
- Focus so far is initial crystallisation – very small scale
  - Want to cover scale-up stage and industrial scale also
  - Need examples to ensure ontology covers all methods/attributes
- Expand the Ontology to cover adjacent/relevant properties
  - Attributes that measure the "success" of the crystallisation process

CCDC

Thank you

Questions?