A freely available crystallisation data set and its application in machine learning

Max Pillong[1], Corinne Marx[1], Philippe Piechon[1], Jerome GP Wicker[2], Richard I Cooper[2] & Trixie Wagner[1]

[1]Global Discovery Chemistry Analytics, Novartis Institutes of Biomedical Research, 4002 Basel, Switzerland

[2]Chemical Crystallography, Chemistry Research Laboratory, Mansfield Road, Oxford, United Kingdom

Supplementary material and methods

ESI content

Accurate mass measurements

Chromatographic separation was performed on a Hypersil GOLD column (Thermo Scientific, Reinach BL, Switzerland) (1.9um, 100x 1.0 mm i.d.). LC–MS was executed using a Thermo Scientifc Ultimate 3000 LC system (Thermo Scientific, Reinach BL, Switzerland). 1 µl sample was injected in microliter pickup mode. The mobile phase solvents for binary gradient elution had the following compositions: water + 0.05% formic acid + 3.75 mM ammonium acetate (solvent A) and acetonitrile + 0.04% formic acid (solvent B). The gradient elution program started with 95% A, increased to 100% B over 5 min, was maintained at 100% B for 2 min, increased to 95% A in 6 seconds, and then kept at 95% A for another 2 minutes for column equilibration. The separation was performed at 40° C using a flow rate of 150 ul/min. The LC system was coupled to a Q Exactive Plus Mass Spectrometer (Thermo Scientific, Reinach BL, Switzerland). Spectra were acquired using electrospray ionisation in positive and negative ion mode, using a resolution setting of 35,000 at m/z 200. High mass accuracy was obtained by using a lock mass.

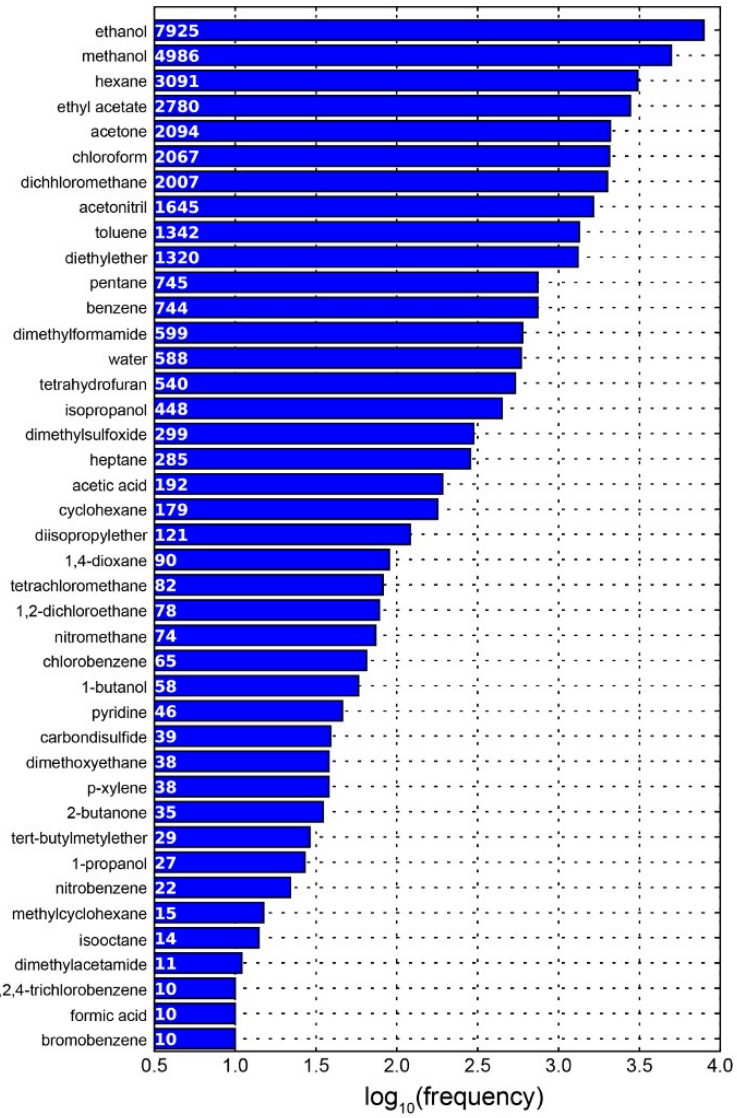

Fig. S1 – Extracted CSD solvents. Shown are the relative frequencies of single solvent experiments in the CSD in white. The x-axis is log-scaled to improve visibility.

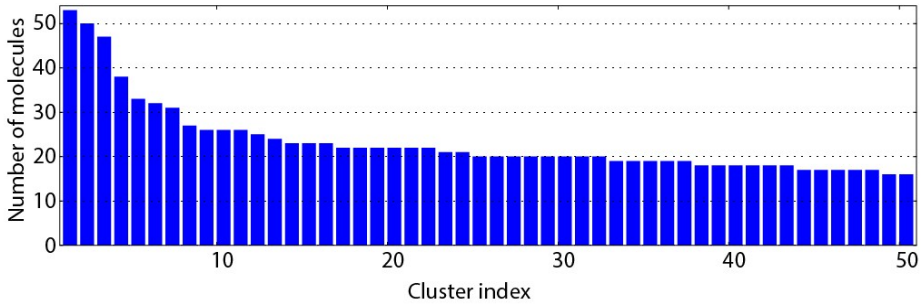

Fig. S2 – Cluster sizes. Shown is the number of molecules found within each of the largest 50 clusters.

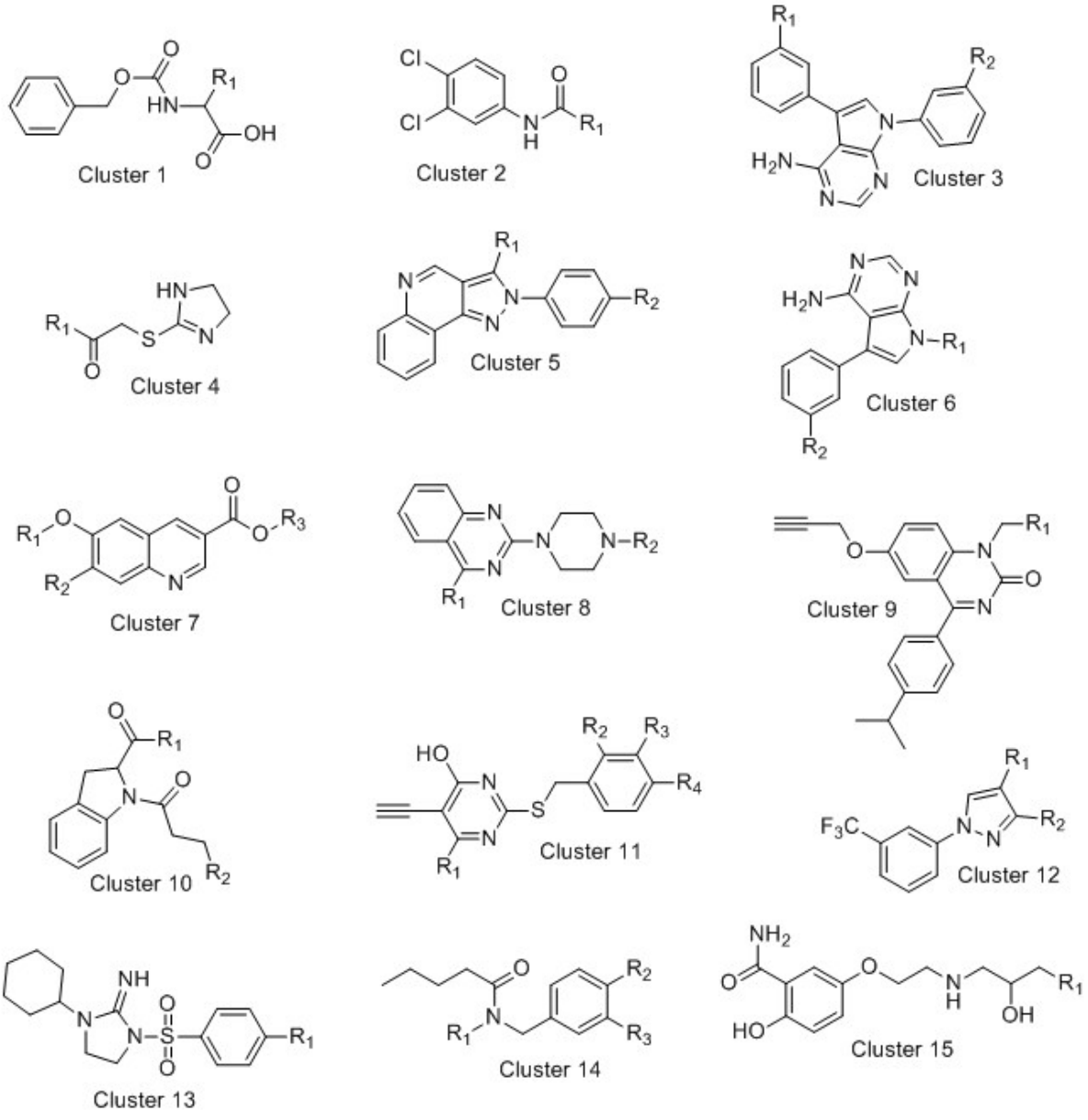

Fig. S3 – Basic scaffolds for public compound clusters. Shown are the general scaffolds of each cluster for the OXDB data set. Compounds in each cluster vary only in the individual R-groups.

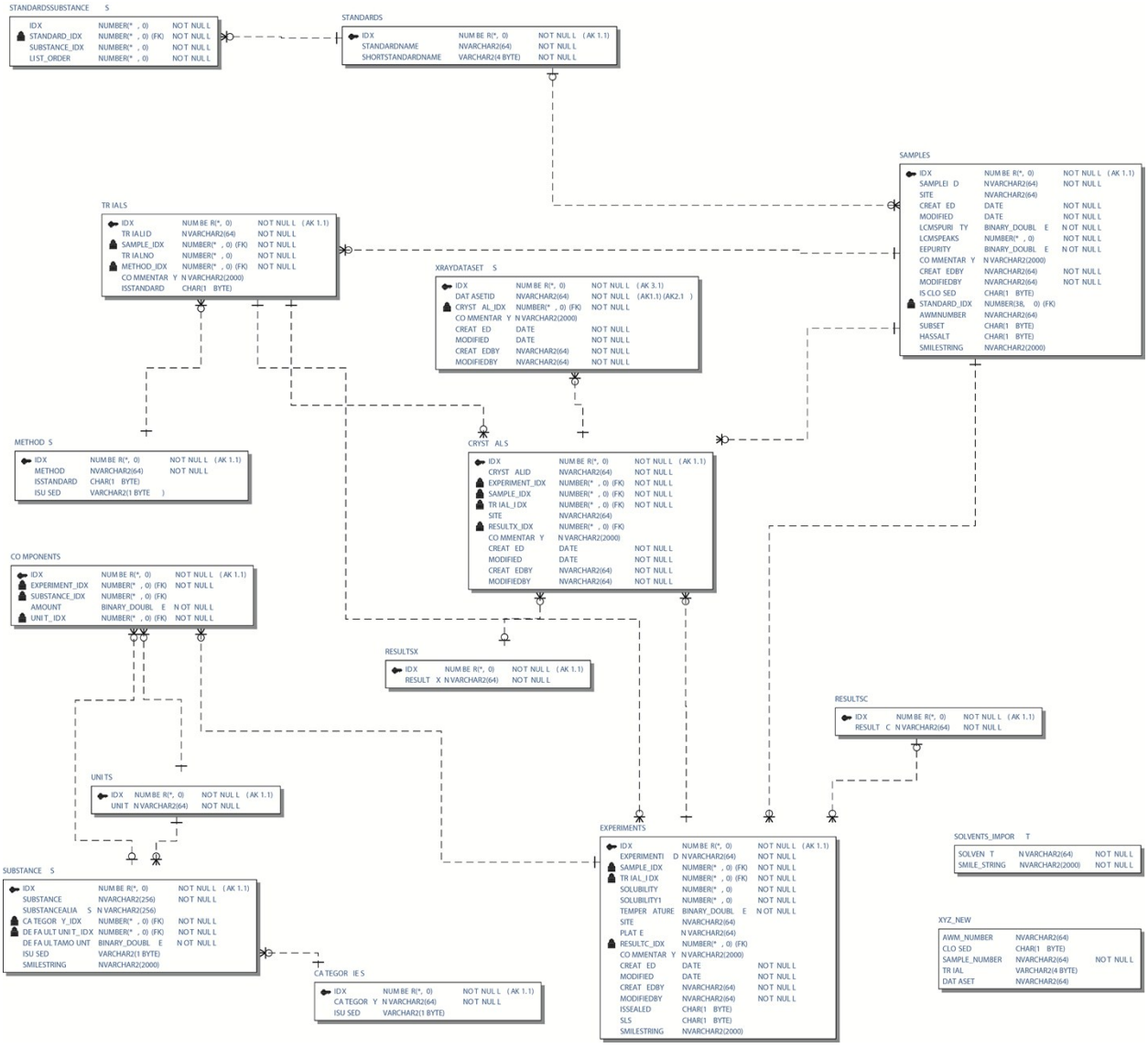

Fig. S4 – ER Diagram. Shown is the entity-relationship diagram for the OracleSQL data base used to store all experiments and their outcomes. This also includes advanced functionality for storage of x-ray diffraction sets and outcomes.

Table S1 – Data base fields for the public compound data. This table lists the individual fields of the data set for the public compounds, supplied with the ESI.

Field	Description
PID	project ID
SMILES	SMILES string for compound
Cluster	assigned cluster
Number	number in cluster
Molweight	molecular weight in g/mol
S-MeOH	solubility in methanol
S-EtOH	solubility in ethanol
S-iPrOH	solubility in iso-propanol
S-ACT	solubility in acetone
S-EMK	solubility in ethyl-methylketone
S-iBMK	solubility in iso-butyl-methylketone
S-EA	solubility in ethyl acetate
S-tBME	solubility in tert-butyl-methylether
S-THF	solubility in tetrahydrofurane
S-CHCl3	solubility in chloroform
S-TOL	solubility in toluene
S-ACN	solubility in acetonitril
S-NiMe	solubility in nitromethane
S-DMF	solutiblity in dimethylformamide
S-MCB	solubility in chlorobenzene
S-DCM	solubility in dichlormethane
S-DEE	solubility in diethylether
S-HEX	solubility in hexane
C-MeOH	crystal propensity in methanol
C-EtOH	crystal propensity in ethanol
C-iPrOH	crystal propensity in iso-propanol
C-ACT	crystal propensity in acetone
C-EMK	crystal propensity in ethyl-methylketone
C-iBMK	crystal propensity in iso-butyl-methylketone
C-EA	crystal propensity in ethyl acetate
C-tBME	crystal propensity in tert-butyl-methylether
C-THF	crystal propensity in tetrahydrofurane
C-CHCl3	crystal propensity in chloroform
C-TOL	crystal propensity in toluene
C-ACN	crystal propensity in acetonitril
C-NiMe	crystal propensity in nitromethane
C-DMF	crystal propensity in dimethylformamide

C-MCB	crystal propensity in chlorobenzene
C-DCM	crystal propensity in dichlormethane
C-DEE	crystal propensity in diethylether
C-HEX	crystal propensity in hexane
Purity	determined LCMS purity of sample

Table S2 – Raw data for statistics of the in-house and the public data. Shown are the solubility and crystallinity class distributions for the combined data from the in-house data set and the public compounds.

		MeOH	EtOH	iPrOH	ACT	EMK	iBMK	EA	tBME	THF	CHCl3	TOL	ACN
Solubility	RS	555	361	240	479	368	282	273	121	538	450	179	343
	KS	970	426	310	764	390	382	336	129	314	274	325	572
	TS	315	320	412	226	190	265	246	67	141	79	251	332
	PS	122	135	130	139	101	183	190	325	70	180	273	175
Crystallinity	XX	575	349	287	430	285	307	321	115	140	150	283	466
	YX	155	110	107	133	104	110	132	120	69	96	163	130
	CT	303	214	210	312	164	150	188	91	113	211	133	230
	DR	255	130	85	131	73	99	106	125	55	142	150	129
	FI	270	232	144	210	88	61	92	47	93	51	49	119
	AM	398	207	258	392	335	385	206	145	593	333	250	348

Exemplary Code

```
# while we cannot provide the full data set for reasons of confidentiality, this code snippet
shows an equivalent approach to model generation based only on the publicly available data. Due
to the limited amount of training data, results will vary in comparison to the main manuscript.

import rdkit
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors
from sklearn import cross_validation
from sklearn.ensemble import RandomForestClassifier

# function used to convert solubility and crystallisation assessments to numeric values

def convertToNumeric(x):
    typeDict={'XX':1,'YX':1,'CT':0,'FI':0,'DR':0,'AM':0,'RS':1,'KS':1,'TS':1,'PS':0,'NA':-1}
return typeDict[x]

mlData=[]
```

```
# read in of molecules in SMILES format

suppl=Chem.SmilesMolSupplier('/home/DB.csv', smilesColumn=1, nameColumn=2, delimiter=',')

# property flags

sProperties=['S-MeOH','S-EtOH','S-iPrOH','S-ACT','S-EMK','S-iBMK','S-EA','S-tBME','S-THF','S-
CHCl3','S-TOL','S-ACN']
cProperties=['C-MeOH','C-EtOH','C-iPrOH','C-ACT','C-EMK','C-iBMK','C-EA','C-tBME','C-THF','C-
CHCl3','C-TOL','C-ACN']

# generate fingerprints and convert solubility and crystallisation results to binary

for m in suppl:
    if not m: continue
    temp=[m,[],[]]
    for i in sProperties:
        temp[1].append(convertToNumeric(m.GetProp(i)))
    for i in cProperties:
        temp[2].append(convertToNumeric(m.GetProp(i)))
    temp.append(rdMolDescriptors.GetMorganFingerprintAsBitVect(m,4,nBits=4096))
    mlData.append(temp)

# test/train split & model generation as well as prediction. The variable s determines the
predicted solvent (MeOH=0, EtOH=1, etc…)

s=0
x_train, x_test, y_train, y_test = cross_validation.train_test_split([x[3] for x in mlData if
x[2][s]>-1],[x[2][s] for x in mlData if x[2][s]>-1],test_size=.50)
treeClassifier=RandomForestClassifier(class_weight='auto',n_estimators=1000, max_features=None,
n_jobs=-1, oob_score=True, criterion='gini')
treeClassifier.fit(x_train,y_train)
pred=treeClassifier.predict(x_test)
```

ESI separate files

Data base

The results for all crystallisation experiments can be found as a CSV file separate to the supplementary. For a description of the denoted fields in the data base, see Table S1.