Electronic Supplementary Material (ESI) for CrystEngComm.

Supporting Information

Listing 1 Example code for training and evaluating accuracy of a model from training and test data with known crystallinity labels. The code runs in Python2.7 and uses the following packages: RDKit version 2012.12.1; SciKit Learn version 0.14.1; and NumPy 1.8.0

```
from rdkit import Chem
from rdkit.Chem import Descriptors
from rdkit.ML.Descriptors import MoleculeDescriptors
from sklearn import preprocessing, svm, metrics
from sklearn.ensemble import RandomForestClassifier
import numpy as np



# Create list of descriptor names from rdkit list of names (one per line).
# This descriptor list method only works for rdkit version 2012.09 and earlier,
# as newer versions contain extra descriptors which give non-numerical values
# for some molecules. These are:
# MinPartialCharge, MaxPartialCharge, MinAbsPartialCharge and MaxAbsPartialCharge.
# These descriptors must be removed from the descriptor list manually for later
# RDKit versions

names = [x[0] for x in Descriptors._descList]
calc = MoleculeDescriptors.MolecularDescriptorCalculator(names)

train_desc_unscaled = []
train_labels = []
test_desc_unscaled = []
test_labels = []

def add_molecules (filename, iscryst, unscaled_descriptors, labels):
    molecule_supplier = Chem.SmilesMolSupplier(filename)
    for molecule in molecule_supplier:
        if molecule is not None:
            descriptors = calc.CalcDescriptors(molecule)
            unscaled_descriptors.append(descriptors)
            labels.append(iscryst)

# Generate descriptors and labels for all data (training and test, cryst and non-cryst)
add_molecules( 'non_crystalline_train_file.smi', 0, train_desc_unscaled, train_labels )
add_molecules( 'crystalline_train_file.smi', 1, train_desc_unscaled, train_labels )
add_molecules( 'non_crystalline_test_file.smi', 0, test_desc_unscaled, test_labels)
add_molecules( 'crystalline_test_file.smi', 1, test_desc_unscaled, test_labels )

# Scale descriptors for use with SVM
train_desc_unscaled = np.array(train_desc_unscaled)
train_labels = np.array(train_labels)
scaler = preprocessing.StandardScaler().fit(train_desc_unscaled)
train_desc = scaler.transform(train_desc_unscaled)

# Train a Support Vector Machine predictor
SVM_classifier = svm.SVC(gamma=0.001, C=100., probability = True)
SVM_classifier = SVM_classifier.fit(train_desc,train_labels)
```

```
#Train a Random Forest Classifier (on unscaled descriptors)
RF_classifier = RandomForestClassifier(n_estimators=100, max_depth=5,random_state=0,n_jobs=1)
RF_classifier = RF_classifier.fit(train_desc_unscaled,train_labels)

# Scale test descriptors
test_desc_unscaled = np.array(test_desc_unscaled)
test_labels = np.array(test_labels)
test_desc = scaler.transform(test_desc_unscaled)

# Output confusion matrix and percentage accuracy on test sets
print 'SVM'
SVM_predictions = SVM_classifier.predict(test_desc)
print metrics.confusion_matrix(test_labels,SVM_predictions)
SVM_accuracy = SVM_classifier.score(test_desc,test_labels)
print SVM_accuracy

print "Random Forest"
RF_predictions = RF_classifier.predict(test_desc_unscaled)
print metrics.confusion_matrix(test_labels,RF_predictions)
RF_accuracy = RF_classifier.score(test_desc_unscaled,test_labels)
print RF_accuracy

#Calculate probability of belonging to either class for each test molecule
SVM_probabilities = SVM_classifier.predict_proba(test_desc)
```

Table 1 Descriptor definitions

RDKit Descriptors	Paper
MolWt, HeavyAtomMolWt, NumRadicalElectrons, NumValenceElectrons, HeavyAtomCount, NumHeteroatoms, NumRotatableBonds, RingCount	Self-explanatory; the implementation can be found in the open source RDKit version 2012.12.1 descriptor module
Chi0v, Chi1v, Chi2v, Chi3v, Chi4v, ChiNv, HallKierAlpha , Kappa1, Kappa2, Kappa3	Rev. Comp. Chem. vol 2, 367-422, (1991)
Chi0n, Chi1n, Chi2n, Chi3n, Chi4n, ChiNn	Similar to Hall Kier ChiXv, but uses nVal instead of valence
BalabanJ	Chem. Phys. Lett. vol 89, 399-404, (1982)
BertzCT	J. Am. Chem. Soc., vol 103, 3599-601 (1981)
Ipc	J. Chem. Phys., vol 67, 4517-33 (1977)
LabuteASA PEOE-VSA1 – PEOE-VSA14 SMR-VSA1 – SMR-VSA10 SlogP-VSA1 – SlogP-VSA12	J. Mol. Graph. Mod., vol 18, 464-77 (2000)
TPSA	J. Med. Chem., vol 43, 3714-7, (2000)
MolLogP, MolMR	J. Chem. Inform. Comput. Sci., vol 39, 868-73 (1999)
EState-VSA1 – EState-VSA11 VSA-EState1 – VSA-EState10	MOE-type descriptors using electrotopological state indices and surface area contributions developed at RD from J. Chem. Inform. Comput. Sci., vol 31, 76-81 (1991)

Table 2 Fragment definitions

Fragment name	Definition
NHOHCount	Number of NHs and OHs
NOCount	Number of Nitrogen and Oxygen atoms
NumHAcceptors	Number of Hydrogen Bond Acceptors
NumHDonors	Number of Hydrogen Bond Donors
fr-Al-COO	Number of aliphatic carboxylic acids
fr-Al-OH	Number of aliphatic hydroxyl groups
fr-Al-OH-noTert	Number of aliphatic hydroxyl groups excluding tert-OH
fr-ArN	Number of N functional groups attached to aromatics
fr-Ar-COO	Number of Aromatic carboxylic acids
fr-Ar-N	Number of aromatic nitrogens
fr-Ar-NH	Number of aromatic amines
fr-Ar-OH	Number of aromatic hydroxyl groups
fr-COO	Number of carboxylic acids
fr-COO2	Number of carboxylic acids
fr-C-O	Number of carbonyl
fr-C-O-noCOO	Number of carbonyl O, excluding COOH
fr-C-S	Number of thiocarbonyl
fr-HOCCN	Number of C(OH)CCN-Ctert-alkyl or C(OH)CCNcyclic
fr-Imine	Number of Imines
fr-NH0	Number of Tertiary amines
fr-NH1	Number of Secondary amines
fr-NH2	Number of Primary amines
fr-N-O	Number of hydroxylamine groups
fr-Ndealkylation1	Number of XCCNR groups
fr-Ndealkylation2	Number of tert-alicyclic amines (no heteroatoms, not quinine-like bridged N)
fr-Nhpyrrole	Number of H-pyrrole nitrogens
fr-SH	Number of thiol groups
fr-aldehyde	Number of aldehydes
fr-alkyl-carbamate	Number of alkyl carbamates
fr-alkyl-halide	Number of alkyl halides
fr-allylic-oxid	Number of allylic oxidation sites excluding steroid dienone
fr-amide	Number of amides
fr-amidine	Number of amidine groups
fr-aniline	Number of anilines
fr-aryl-methyl	Number of aryl methyl sites for hydroxylation
fr-azide	Number of azide groups
fr-azo	Number of azo groups
fr-barbitur	Number of barbiturate groups
fr-benzene	Number of benzene rings
fr-benzodiazepine	Number of benzodiazepines with no additional fused rings
fr-bicyclic	Number of bicyclic rings
fr-diazo	Number of diazo groups
fr-dihydropyridine	Number of dihydropyridines
fr-epoxide	Number of epoxide rings
fr-ester	Number of esters
fr-ether	Number of ether oxygens (including phenoxy)
fr-furan	Number of furan rings
fr-guanido	Number of guanidine groups
fr-halogen	Number of halogens
fr-hdrzine	Number of hydrazine groups

Fragment name	Definition
fr-hdrzone	Number of hydrazone groups
fr-imidazole	Number of imidazole rings
fr-imide	Number of imide groups
fr-isocyan	Number of isocyanates
fr-isothiocyan	Number of isothiocyanates
fr-ketone	Number of ketones
fr-ketone-Topliss	Number of ketones excluding diaryl, a,b-unsat.
fr-lactam	Number of beta lactams
fr-lactone	Number of cyclic esters (lactones)
fr-methoxy	Number of methoxy groups -OCH3
fr-morpholine	Number of morpholine rings
fr-nitrile	Number of nitriles
fr-nitro	Number of nitro groups
fr-nitro-arom	Number of nitro benzene ring substituents
fr-nitro-arom-nonortho	Number of non-ortho nitro benzene ring substituents
fr-nitroso	Number of nitroso groups, excluding NO2
fr-oxazole	Number of oxazole rings
fr-oxime	Number of oxime groups
fr-para-hydroxylation	Number of para-hydroxylation sites
fr-phenol	Number of phenols
fr-phenol-noOrthoHbond	Number of phenolic OH excluding ortho intramolecular Hbond substituents
fr-phos-acid	Number of phosphoric acid groups
fr-phos-ester	Number of phosphoric ester groups
fr-piperdine	Number of piperdine rings
fr-piperzine	Number of piperzine rings
fr-priamide	Number of primary amides
fr-prisulfonamd	Number of primary sulfonamides
fr-pyridine	Number of pyridine rings
fr-quatN	Number of quarternary nitrogens
fr-sulfide	Number of thioether
fr-sulfonamd	Number of sulfonamides
fr-sulfone	Number of sulfone groups
fr-term-acetylene	Number of terminal acetylenes
fr-tetrazole	Number of tetrazole rings
fr-thiazole	Number of thiazole rings
fr-thiocyan	Number of thiocyanates
fr-thiophene	Number of thiophene rings
fr-unbrch-alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)
fr-urea	Number of urea groups

Table 3 RDKit descriptor list and single variable classifier accuracy

Index	Descriptor	Accuracy	Index	Descriptor	Accuracy
8	Chi0v	0.7721138	53	SlogP VSA3	0.6307864
3	NumValenceElectrons	0.7721138	73	VSA EState10	0.6287786
20	Kappa1	0.7695482	122	fr aryl methyl	0.628667
21	Kappa2	0.7691021	60	TPSA	0.6277747
11	Chi1v	0.7660904	70	EState VSA8	0.6275516
23	LabuteASA	0.7658673	108	fr NH1	0.6255438
0	MolWt	0.7649749	55	SlogP VSA5	0.6246514
7	Chi0n	0.7641941	48	SlogP VSA1	0.6207474
6	Chi0	0.7618516	103	fr C O noCOO	0.6170664
82	HeavyAtomCount	0.7618516	121	fr aniline	0.6158394
1	HeavyAtomMolWt	0.76029	49	SlogP VSA10	0.6152817
9	Chi1	0.7596207	31	PEOE VSA3	0.6122699
10	Chi1n	0.7556051	28	PEOE VSA13	0.6079197
12	Chi2n	0.7525934	54	SlogP VSA4	0.606135
91	MolMR	0.7525934	64	EState VSA2	0.6033463
13	Chi2v	0.7472393	39	SMR VSA10	0.6
4	BalabanJ	0.7380926	24	PEOE VSA1	0.597546
15	Chi3v	0.735527	45	SMR VSA7	0.5940881
14	Chi3n	0.7331846	102	fr C O	0.5931958
88	NumRotatableBonds	0.7296152	51	SlogP VSA12	0.592638
5	BertzCT	0.7284997	33	PEOE VSA5	0.5881762
22	Kappa3	0.7269381	90	MolLogP	0.5881762
17	Chi4v	0.7216955	69	EState VSA7	0.5843837
16	Chi4n	0.7175683	58	SlogP VSA8	0.5839375
87	NumHeteroatoms	0.7133296	25	PEOE VSA10	0.5837144
119	fr amide	0.6940323	30	PEOE VSA2	0.5833798
84	NOCount	0.6890128	68	EState VSA6	0.5814835
41	SMR VSA3	0.6890128	37	PEOE VSA9	0.581372
89	RingCount	0.6790853	67	EState VSA5	0.5795873
19	Ipc	0.6775237	136	fr halogen	0.5781372
52	SlogP VSA2	0.670831	32	PEOE VSA4	0.5743447
107	fr NH0	0.6699387	34	PEOE VSA6	0.5706637
44	SMR VSA6	0.6662577	56	SlogP VSA6	0.5684328
80	VSA EState8	0.6624651	83	NHOHCount	0.5649749
97	fr Ar N	0.6612381	128	fr bicyclic	0.5607362
43	SMR VSA5	0.6606804	47	SMR VSA9	0.5587284
36	PEOE VSA8	0.6569994	86	NumHDonors	0.5573898
65	EState VSA3	0.6539877	38	SMR VSA1	0.556609
27	PEOE VSA12	0.6484105	42	SMR VSA4	0.5552705
35	PEOE VSA7	0.6468489	62	EState VSA10	0.552705
85	NumHAcceptors	0.6452872	126	fr benzene	0.5492471
66	EState VSA4	0.6381484	161	fr piperdine	0.5484663
18	HallKierAlpha	0.6345789	112	fr Ndealkylation2	0.5472393
81	VSA EState9	0.6314557	168	fr sulfonamd	0.5422197

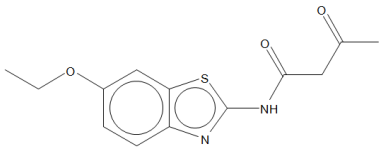
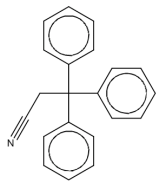
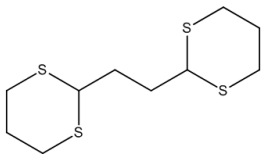
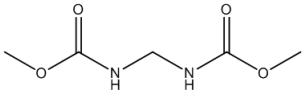
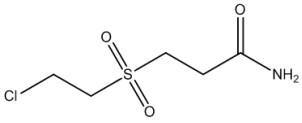
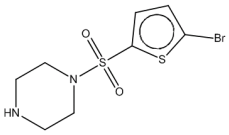
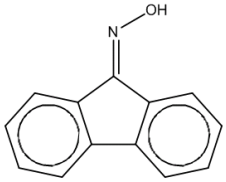
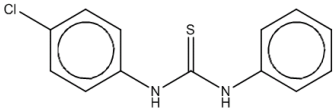
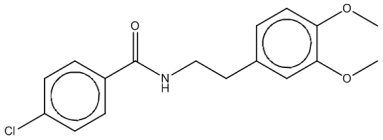
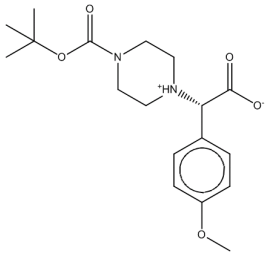
Index	Descriptor	Accuracy
143	fr ketone	0.5419967
133	fr ether	0.5408812
157	fr phenol	0.5406581
162	fr piperzine	0.5403235
158	fr phenol noOrthoHbond	0.5393196
167	fr sulfide	0.539208
26	PEOE VSA11	0.5372002
99	fr Ar OH	0.5368656
57	SlogP VSA7	0.5339654
101	fr COO2	0.5338539
100	fr COO	0.5338539
118	fr allylic oxid	0.5329615
174	fr thiophene	0.5320692
165	fr pyridine	0.5320692
172	fr thiazole	0.5282766
50	SlogP VSA11	0.5278305
93	fr Al OH	0.5261573
92	fr Al COO	0.5252649
176	fr urea	0.5248187
138	fr hdrzone	0.5248187
71	EState VSA9	0.524261
144	fr ketone Topliss	0.5219186
61	EState VSA1	0.52058
132	fr ester	0.5196877
150	fr nitro	0.5189069
98	fr Ar NH	0.5175683
113	fr Nhpyrrole	0.5175683
29	PEOE VSA14	0.5172337
109	fr NH2	0.5171221
94	fr Al OH noTert	0.5168991
147	fr methoxy	0.516676
111	fr Ndealkylation1	0.5163413
151	fr nitro arom	0.5155605
152	fr nitro arom nonortho	0.5151143
134	fr furan	0.5145566
139	fr imidazole	0.5144451
104	fr C S	0.5139989
148	fr morpholine	0.5131065
40	SMR VSA2	0.5110987
156	fr para hydroxylation	0.5109872
95	fr ArN	0.5109872
169	fr sulfone	0.5102064
117	fr alkyl halide	0.5100948
149	fr nitrile	0.5099833
96	fr Ar COO	0.5083101

Index	Descriptor	Accuracy
175	fr unbrch alkane	0.508087
154	fr oxazole	0.5075293
146	fr lactone	0.50686
171	fr tetrazole	0.50686
163	fr priamide	0.5055215
106	fr Imine	0.5054099
115	fr aldehyde	0.5042945
155	fr oxime	0.5037368
140	fr imide	0.5029559
131	fr epoxide	0.5028444
130	fr dihydropyridine	0.5028444
120	fr amidine	0.5023982
105	fr HOCCN	0.5023982
110	fr N O	0.501952
137	fr hdrzine	0.5018405
114	fr SH	0.5016174
170	fr term acetylene	0.5013943
116	fr alkyl carbamate	0.5009481
141	fr isocyan	0.5006135
159	fr phos acid	0.5006135
160	fr phos ester	0.5006135
135	fr guanido	0.5006135
63	EState VSA11	0.5006135
145	fr lactam	0.500502
127	fr benzodiazepine	0.5003904
46	SMR VSA8	0.5002789
79	VSA EState7	0.5002789
129	fr diazo	0.5002789
59	SlogP VSA9	0.5002789
164	fr prisulfonamd	0.5002789
123	fr azide	0.5002789
166	fr quatN	0.5002789
76	VSA EState4	0.5002789
153	fr nitroso	0.5002789
78	VSA EState6	0.5002789
72	VSA EState1	0.5002789
77	VSA EState5	0.5002789
2	NumRadicalElectrons	0.5002789
75	VSA EState3	0.5002789
173	fr thiocyan	0.5002789
74	VSA EState2	0.5002789
142	fr isothiocyan	0.5000558
125	fr barbitur	0.4998327
124	fr azo	0.4988288

Table 4 Lipinski rule–of–5

Physicochemical property	Value
Molecular weight (amu)	$150 \leq x \leq 500$
Total polar surface area (Å^2)	< 150
Hydrogen bond donors	≤ 5
Hydrogen bond acceptors	≤ 10
CLogP	≤ 5
Rotatable bonds	≤ 7

Table 5 Blind test numbering

Vial	Compound Name	Structure	Crystallinity label (NC or C)
1	2-acetoamido-6-ethoxybenzothiazole		NC
2	3,3,3 triphenylpropionitrile		C
3	2-[2-(1,3-dithian-2-yl)ethyl]-1,3-dithiane		C
4	*N*,*N*-methylenebis(methyl carbamate)		NC
5	3-(2-chloro-ethanesulfonyl-propionamide)		NC
6	1-(5-bromo-2-thienyl)sulfonyl)piperazine		NC
7	9-fluorenone oxime		C
8	1-(4-chlorophenyl)-3-phenylthiourea		C
9	4-chloro-*N*-(3,4-dimethoxyphenethyl)-benzamide		NC
10	(4-(tert-butoxycarbonyl)-1-piperazinyl)(4-methoxyphenyl)acetic acid		NC

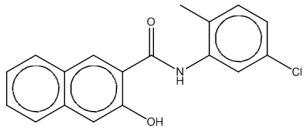
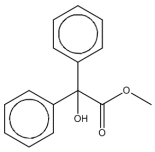
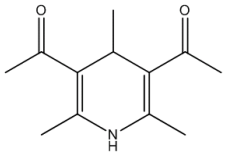
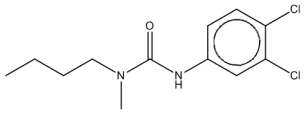
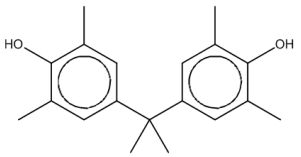
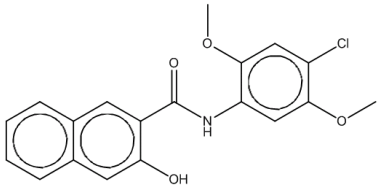
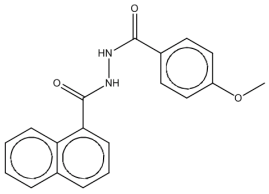
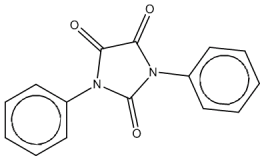
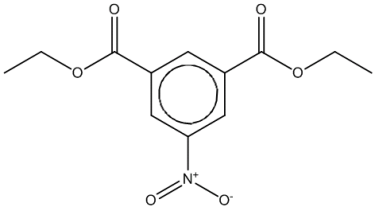
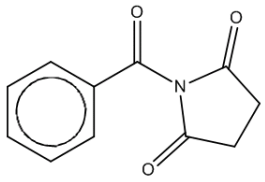
Vial	Compound Name	Structure	Crystallinity label (NC or C)
11	5’-chloro-3-hydroxy-2’-methyl-2-naphtanilide		NC
12	Methyl benzilate		C
13	3,5-diacetyl-2,4,6-trimethyl-1,4-dihydropyridine		C
14	1-butyl-3-(3,4-dichlorophenyl)-methylurea		NC
15	2,2-bis-(3,5-dimethyl-4-hydroxyphenyl)-propane		C
16	4’-chloro-3-hydroxy-2’,5’-dimethoxy-2-naphthoic acid anilide		NC
17	1-(4-methoxybenzoyl)-2-(1-naphthoyl)-hydrazine		NC
18	1,3-diphenylparabanic acid		C
19	5-nitro-isophthalic acid diethyl ester		C
20	1-benzoyl-pyrrolidine-2,5-dione		C

Table 6 Blind recrystallisation success table

Compound number	Ethanol	DCM	DMF	Ethyl acetate	Chloroform	Diethyl ether	Model prediction
1	N	N	N	N	N	N	NC
2	N	N	Y	–	–	–	C
3	N	N	Y	–	–	–	C
4	N	N	N	N	N	N	NC
5	N	N	N	N	N	N	NC
6	N	N	N	N	N	N	NC
7	N	N	N	Y	–	–	C
8	N	N	N	N	N	N	C
9	N	N	N	N	N	N	NC
10	N	N	N	N	N	N	NC
11	N	N	N	N	N	N	NC
12	Y	–	–	–	–	–	C
13	N	N	Y	–	–	–	C
14	N	N	N	N	N	N	NC
15	Y	–	–	–	–	–	C
16	N	N	N	N	N	N	NC
17	N	N	N	N	N	N	NC
18	–	–	–	–	–	–	–
19	N	N	N	N	N	N	C
20	N	N	N	Y	–	–	C