

# Polymorphism from a Solution Perspective: Rationalisation at the Molecular Level

A thesis submitted to the University of Manchester  
for the degree of Doctor of Philosophy in the  
Faculty of Engineering and Physical Sciences

**Vicky M. Fawcett**

School of Chemical Engineering and Analytical Science  
The University of Manchester  
2011

# TABLE OF CONTENTS

1.	INTRODUCTION .....	24
1.1.	A CRYSTALLINE SUBSTANCE .....	24
1.2.	POLYMORPHISM.....	24
1.2.1.	SOLVATES AND HYDRATES .....	25
1.2.2.	RELATIVE STABILITY OF POLYMORPHS.....	25
1.2.3.	POLYMORPHISM IN THE PHARMACEUTICAL INDUSTRY .....	27
1.2.4.	POLYMORPH PREDICTION.....	28
1.3.	CRYSTALLISATION .....	29
1.3.1.	FACTORS AFFECTING CRYSTALLISATION.....	29
1.3.2.	POLYMORPH SCREEN.....	32
1.4.	NUCLEATION .....	32
1.4.1.	PRIMARY NUCLEATION .....	33
1.4.2.	SECONDARY NUCLEATION .....	36
1.4.3.	THE NUCLEATION OF POLYMORPHS.....	37
1.4.4.	OSTWALD'S RULE OF STAGES .....	38
1.4.5.	DETECTION OF NUCLEATION .....	39
1.4.6.	CURRENT AND PREVIOUS RESEARCH INTO NUCLEATION .....	40
1.5.	GROWTH OF POLYMORPHIC CRYSTALS .....	42
1.5.1.	POLYMORPHIC PHASE TRANSITIONS .....	44
1.6.	SUMMARY .....	44
2.	METHODS AND MATERIALS .....	48
2.1.	X-RAY POWDER DIFFRACTION.....	48
2.1.1.	XRPD INSTRUMENTS .....	49
2.2.	POLYMORPH SCREEN.....	49
2.3.	CARBAMAZEPINE EXPERIMENTAL DETAILS .....	50
2.3.1.	FORM I.....	50
2.3.2.	FORM II .....	51
2.3.3.	FORM III .....	51
2.3.4.	FORM IV.....	52
2.3.5.	DIHYDRATE .....	52
2.3.6.	DMSO SOLVATE .....	53
2.4.	ROY EXPERIMENTAL DETAILS.....	53
2.5.	TOLBUTAMIDE EXPERIMENTAL DETAILS .....	54
2.5.1.	FORM I.....	54
2.5.2.	FORM II .....	55

2.5.3.	FORM III .....	55
2.5.4.	FORM IV.....	56
2.5.5.	FORM V .....	57
2.6.	THEORETICAL CALCULATIONS.....	57
2.6.1.	MOLECULAR MODELLING .....	57
2.6.2.	GEOMETRY OPTIMISATION .....	58
2.6.3.	MODELLING SOFTWARE USED.....	62
2.6.4.	MOLECULAR REPRESENTATION .....	62
2.6.5.	MOLECULAR DESCRIPTORS.....	62
2.6.6.	MOLECULAR DESCRIPTOR SOFTWARE USED.....	63
2.7.	ARTIFICIAL NEURAL NETWORKS .....	63
2.7.1.	MULTILAYER PERCEPTRON (MLP) .....	67
2.7.2.	TRANSFER FUNCTIONS .....	70
2.7.3.	ARTIFICIAL NEURAL NETWORK SOFTWARE USED .....	70
2.8.	NEUROFUZZY LOGIC .....	70
2.8.4.	FUZZY LOGIC SOFTWARE USED .....	72
2.9.	PRINCIPAL COMPONENT ANALYSIS .....	72
2.10.	PARTIAL LEAST SQUARES ANALYSIS.....	75
2.10.1.	CHEMOMETRIC DATA ANALYSIS SOFTWARE.....	77
3.	SYSTEMS STUDIED .....	82
3.1.	CARBAMAZEPINE.....	82
3.1.1.	WHY CARBAMAZEPINE FOR THIS RESEARCH? .....	82
3.1.2.	THE CARBAMAZEPINE POLYMORPHS.....	83
3.2.	5-METHYL-2-[(2-NITROPHENYL)AMINO]-3-THIOPHENECARBONITRILE.....	90
3.2.1.	USING ROY FOR POLYMORPHIC INVESTIGATION .....	90
3.2.2.	THE ROY POLYMORPHS.....	91
3.3.	TOLBUTAMIDE.....	94
3.3.1.	WHY TOLBUTAMIDE FOR THIS RESEARCH? .....	94
3.3.2.	THE TOLBUTAMIDE POLYMORPHS .....	95
3.3.3.	STABILITY .....	102
3.4.	SUMMARY .....	104
4.	ANALYSIS METHODOLOGY .....	108
4.1.	MOLECULAR MODELLING .....	108
4.1.1.	HYPERCHEM™ .....	108
4.1.2.	GAUSSIAN 03.....	111
4.2.	BULK AND MOLECULAR DESCRIPTORS .....	113
4.2.1.	DESCRIPTOR REDUCTION METHODS.....	113
4.3.	ARTIFICIAL NEURAL NETWORK (ANN) INPUT FILE.....	114
4.4.	FORMRULES ANALYSIS.....	115
4.5.	INFORM ANALYSIS.....	121

4.6.	COMBINED FORMRULES AND INFORM ANALYSIS .....	123
4.6.1.	RAPID ANALYSIS OF COMBINED DATA.....	123
4.6.2.	DETAILED ANALYSIS OF COMBINED DATA USING THE 3D EXPLORER.....	123
4.7.	SUMMARY OF ANALYSIS METHODOLOGY.....	129
5.	RESULTS AND DISCUSSION – MANUAL ANALYSIS .....	131
5.1.	COMPLETE DATASET ANALYSIS.....	131
5.1.1.	COMPLETE DATASET ANALYSIS – FORMRULES .....	133
5.1.2.	COMPLETE DATASET ANALYSIS – INFORM .....	137
5.1.3.	COMPLETE DATASET ANALYSIS – FORMRULES AND INFORM .....	140
5.1.4.	OPTIMISATION OF THE HIGH PERFORMING SETS .....	143
5.2.	LINEAR CORRELATIONS ANALYSIS.....	146
5.2.1.	LINEAR CORRELATION ANALYSIS – FORMRULES .....	149
5.2.2.	LINEAR CORRELATION ANALYSIS – INFORM .....	153
5.2.3.	DESCRIPTOR OVERLAPS IN FORMRULES AND INFORM ANALYSIS.....	158
5.2.4.	LINEAR CORRELATION ANALYSIS – FORMRULES AND INFORM .....	160
5.2.5.	OPTIMISATION OF THE LINEAR CORRELATION BEST SET .....	163
5.3.	OVERALL OPTIMISATION OF BEST DESCRIPTOR SET .....	164
5.4.	DISCUSSION OF THE DESCRIPTORS IN THE OPTIMISED SET .....	171
5.4.1.	THE PREDICTION OF FORM I .....	172
5.4.2.	THE PREDICTION OF FORM II.....	175
5.4.3.	THE PREDICTION OF FORM III .....	180
5.4.4.	THE PREDICTION OF THE DIHYDRATE .....	187
5.4.5.	THE PREDICTION OF SOLVATES.....	190
5.4.6.	SUMMARY OF THE OPTIMISED DESCRIPTORS .....	193
5.5.	VALIDATION OF OPTIMISED SET.....	195
5.5.1.	CROSS VALIDATION RESULTS .....	195
5.5.2.	EXTERNAL VALIDATION RESULTS .....	198
5.6.	CONCLUSION OF MANUAL DATA ANALYSIS.....	201
6.	RESULTS AND DISCUSSION-PLS ANALYSIS .....	204
6.1.	DATA REDUCTION USING PLS .....	204
6.1.1.	ANALYSIS OF SCORE VALUES (FORM II MODEL) .....	204
6.1.2.	ANALYSIS OF SCORE VALUES (FORM III MODEL).....	206
6.1.3.	ANALYSIS OF THE LOADING VALUES (FORM II MODEL) .....	208
6.1.4.	ANALYSIS OF THE LOADING VALUES (FORM III MODEL) .....	210
6.1.5.	ANALYSIS OF THE VARIABLE IMPORTANCE VALUES (FORM II MODEL) ...	212
6.1.6.	ANALYSIS OF THE VARIABLE IMPORTANCE VALUES (FORM III MODEL) .	214
6.1.7.	ANALYSIS OF THE VARIABLE IMPORTANCE DESCRIPTOR OVERLAP .....	215
6.2.	OPTIMISATION OF PLS RESULTS .....	217
6.3.	CONCLUSION OF PLS WORK.....	218
7.	RESULTS AND DISCUSSION - PCA ANALYSIS.....	220



7.1.	DATA REDUCTION USING PCA .....	220
7.1.1.	ANALYSIS OF SCORE VALUES .....	220
7.1.2.	ANALYSIS OF LOADING VALUES.....	222
7.1.3.	SELECTION OF THE MOST VALUABLE DESCRIPTORS .....	227
7.2.	OPTIMISATION OF PCA RESULTS .....	232
7.2.1.	ANALYSIS OF THE OPTIMISED SET .....	234
7.3.	DISCUSSION OF THE DESCRIPTORS IN THE OPTIMISED SET .....	239
7.3.1.	THE PREDICTION OF FORM I .....	240
7.3.2.	THE PREDICTION OF FORM II.....	242
7.3.3.	THE PREDICTION OF FORM III .....	248
7.3.4.	THE PREDICTION OF THE DIHYDRATE .....	254
7.3.5.	THE PREDICTION OF SOLVATES .....	257
7.3.6.	SUMMARY OF THE OPTIMISED DESCRIPTORS .....	258
7.4.	VALIDATION OF THE OPTIMISED SET .....	260
7.4.1.	CROSS VALIDATION RESULTS .....	260
7.4.2.	EXTERNAL VALIDATION RESULTS .....	262
7.5.	CONCLUSIONS .....	265
8.	FINAL OPTIMISATION AND DISCUSSION OF RESULTS .....	268
8.1.	INTRODUCTION .....	268
8.2.	FINAL OPTIMISATION OF THE DESCRIPTOR SETS .....	270
8.2.1.	OPTIMISATION BASED UPON CORRELATION RESULTS.....	270
8.2.2.	OPTIMISATION BASED UPON BEST SET ANALYSIS .....	272
8.2.3.	OPTIMISATION BASED UPON DESCRIPTOR TYPES.....	273
8.2.4.	OPTIMISATION BASED UPON VALIDATION RESULTS.....	280
8.2.5.	CONCLUSION OF THE FINAL OPTIMISATION WORK .....	281
8.3.	DISCUSSION OF THE DESCRIPTORS IN THE FINAL SET .....	282
8.3.1.	THE PREDICTION OF FORM I .....	283
8.3.2.	THE PREDICTION OF FORM II.....	285
8.3.3.	THE PREDICTION OF FORM III .....	288
8.3.4.	THE PREDICTION OF THE DIHYDRATE .....	290
8.3.5.	THE PREDICTION OF SOLVATES .....	292
8.3.6.	SUMMARY OF THE OPTIMISED DESCRIPTORS .....	293
8.4.	VALIDATION OF OPTIMISED SET .....	293
8.4.1.	CROSS VALIDATION RESULTS .....	293
8.4.2.	EXTERNAL VALIDATION RESULTS .....	296
8.5.	CONCLUSION OF THE FINAL OPTIMISATION ANALYSIS .....	298
9.	RESULTS AND DISCUSSION OF ANALYSIS WITH DIFFERENT TARGET MOLECULES .....	301
9.1.	ANALYSIS OF DESCRIPTORS HIGHLIGHTED IN CBZ ANALYSIS.....	301
9.1.1.	TBA RESULTS .....	302

9.1.2.	ROY RESULTS .....	304
9.1.3.	CONCLUSION OF HIGHLIGHTED DESCRIPTOR ANALYSIS .....	307
9.2.	DESCRIPTOR SELECTION USING PCA .....	307
9.2.1.	TBA RESULTS.....	308
9.2.2.	ROY RESULTS .....	312
9.2.3.	SUMMARY .....	316
9.3.	OVERALL CONCLUSIONS AND SUMMARY OF CHAPTER.....	316
10.	CONCLUSIONS.....	319
10.1.	MANUAL ANALYSIS .....	319
10.2.	PARTIAL LEAST SQUARES ANALYSIS.....	320
10.3.	PRINCIPAL COMPONENT ANALYSIS .....	320
10.4.	FINAL COMBINED OPTIMISATION .....	321
10.5.	OVERALL CONCLUSIONS DRAWN FROM THE INVESTIGATION OF THE CARBAMAZEPINE SYSTEM.....	322
10.6.	CONCLUSIONS FROM ANALYSIS OF TBA AND ROY SYSTEMS .....	324
10.7.	SUMMARY OF CONCLUSIONS .....	325
11.	FURTHER WORK.....	328
12.	APPENDIX.....	331
12.1.	CBZ POLYMORPH SCREEN EXPERIMENTAL RESULTS .....	331
12.2.	MOLECULAR AND BULK DESCRIPTOR MEANINGS .....	334
12.3.	TOLBUTAMIDE STABILITY XRPD TRACES.....	347
12.4.	CBZ OPLS FORCEFIELD RESULTS.....	349
12.5.	ROY PM3 CONFORMATIONAL SEARCH RESULTS .....	350
12.6.	SETS OF DESCRIPTORS USED IN THE MANUAL ANALYSIS WORK.....	351
12.7.	LINEAR CORRELATION BETWEEN THE DESCRIPTORS - SCHEMATIC .....	352
12.8.	RULES OF FORM II LOADING VALUE ANALYSIS.....	353
12.9.	PCA ANALYSIS – PLOTS OF MOLECULAR SURFACE AREA AGAINST BULK PROPERTIES.....	354
12.10.	FINAL CBZ ANALYSIS – RULES OF OPT. E DESCRIPTOR SET.....	355

The main body of this thesis contains no more than 80,000 words.

## LIST OF TABLES

Table 2.1 XRPD instrument details .....	49
Table 2.2 TLU example inputs .....	67
Table 2.3 Example of the rules generated by FormRules <sup>[75]</sup> using neurofuzzy logic (please note, this is not a real result and has only been used for illustration purposes) .....	72
Table 3.1 Selected parameters of CBZ anhydrous polymorphs and dihydrate, taken from Grzesiak et al. <sup>[2]</sup> unless stated.....	89
Table 3.2 Select parameters of the ROY polymorphs, taken from Yu <sup>[36]</sup> unless stated. .....	92
Table 3.3 Selected parameters of TBA anhydrous polymorphs .....	101
Table 3.4 Summary of stability orders from literature.....	103
Table 3.5 Summary of form I slurry in different solvents .....	103
Table 4.1 Comparison of OPLS geometry optimised CBZ with literature values <sup>[4]</sup>	109
Table 4.2 Input parameters for Gaussian calculations .....	112
Table 4.3 Examples of ANN input data .....	114
Table 4.4 Example of the rules generated for analysed data in FormRules <sup>[10]</sup> .....	116
Table 4.5 The methods used to generate the data in the model statistics table. Taken from FormRules manual <sup>[14]</sup> .....	119
Table 4.6 Model statistics produced by FormRules <sup>[10]</sup> .....	120
Table 4.7 Summary of the network architecture used in this analysis.....	121
Table 4.8 Model statistics generated in INForm <sup>[9]</sup> .....	122
Table 4.9 A rule for form II prediction from FormRules <sup>[10]</sup> .....	124
Table 4.10 Summary of the rule to predict form III.....	126
Table 5.1 The final set and first and second most successful descriptor sets in FormRules analysis .....	133
Table 5.2 Linear correlations between the best set of descriptors and the final and second best set in the FormRules analysis. Number in brackets are the correlation coefficients .....	134
Table 5.3 Optimisation of the FormRules (FR) descriptor sets. X denotes the presence of the descriptor in the set .....	136
Table 5.4 The final set and first and second most successful descriptor sets in FormRules analysis .....	137
Table 5.5 Linear correlations between the best set of descriptors and the final and second best set in the INForm analysis. Number in brackets are the correlation coefficients .....	138
Table 5.6 Optimisation of the INForm (IN) descriptor sets. X denotes the presence of the descriptor in the set .....	139
Table 5.7 The final set and first and second most successful descriptor sets in FormRules analysis .....	140
Table 5.8 Linear correlations between the best set of descriptors and the final and second best set. Number in brackets are the correlation coefficients .....	141
Table 5.9 Optimisation of the FormRules descriptor sets. X denotes the presence of the descriptor in the set .....	142

Table 5.10 The final set and first and second most successful descriptor sets in the high performing sets analysis.....	143
Table 5.11 Linear correlations between the best set of descriptors and the final and second best set. Number in brackets are the correlation coefficients .....	144
Table 5.12 Optimisation of the top performing descriptor sets. X denotes the presence of the descriptor in the set .....	145
Table 5.13 The 40 descriptors selected for analysis from the correlations analysis	147
Table 5.14 The descriptors and results of the second group of 4 sets.....	147
Table 5.15 The descriptors and results of the third group of 4 sets .....	148
Table 5.16 The descriptors and results of the fourth group of 4 sets .....	148
Table 5.17 The top set from each group based on the results of FormRules analysis .....	149
Table 5.18 Monitoring the repeat occurrence of descriptors in the top sets of FormRules analysis. X denotes the presence of the descriptor in the set .....	150
Table 5.19 Further analysis of FormRules results. X denotes the presence of the descriptor in the set .....	152
Table 5.20 The top set from each group based on the results of INForm analysis..	153
Table 5.21 Monitoring the repeat occurrence of descriptors in the top sets of FormRules analysis .....	154
Table 5.22 Analysis of descriptors highlighted in the INForm analysis.....	155
Table 5.23 Analysis of the effect of d82, dsolv71 and density. X denotes the presence of the descriptor in the set .....	157
Table 5.24 Comparison of the descriptors in the most successful FormRules and INForm sets. X denotes the presence of the descriptor in the set .....	158
Table 5.25 Analysis of overlapping descriptors with those found in INForm and FormRules best sets .....	159
Table 5.26 The top performing set from each group, based on the average result from FormRules and INForm analysis .....	160
Table 5.27 Comparison of the descriptors in the top performing sets using average FormRules and INForm results. X denotes the presence of the descriptor in the set .....	161
Table 5.28 Optimisation of the descriptors highlighted in the combined analysis. X denotes the presence of the descriptor in the set .....	162
Table 5.29 Optimisation of the best sets from the linear correlation analysis. X denotes the presence of the descriptor in the set .....	164
Table 5.30 The two most successful sets from linear correlation and all descriptor analysis.....	165
Table 5.31 Linear correlations between the two most successful sets from linear correlation and all descriptor analysis. Number in brackets is the correlation coefficient.....	165
Table 5.32 Reduction of descriptors in All Best Set. X denotes the presence of the descriptor in the set .....	166
Table 5.33 Linear correlations between the two most successful sets from linear correlation and all descriptor analysis.....	168
Table 5.34 Rules generated in FormRules when the Corr Best Set is analysed .....	169
Table 5.35 The comparison of Corr Best Set with and without d75.....	170
Table 5.36 Summary of the descriptors in the most successful set.....	171
Table 5.37 Rules generated in FormRules for form I prediction .....	172
Table 5.38 Rules generated in FormRules for form II prediction.....	175
Table 5.39 Rules generated in FormRules for form III prediction .....	181

Table 5.40 Solvent Gutmann donor and acceptor numbers .....	183
Table 5.41 Dimethylacetamide Gutmann donor and acceptor numbers .....	184
Table 5.42 Solvent-solute and solvent-solvent interactions based on DN and AN ratios.....	184
Table 5.43 Rules generated in FormRules for Dihydrate prediction .....	187
Table 5.44 Rules generated in FormRules for solvate prediction.....	191
Table 5.45 Summary of the descriptors involved in the CBZ predictive rules.....	194
Table 5.46 Cross validation results .....	196
Table 5.47 External validation results summary.....	199
Table 6.1 Results of ANN analysis of PLS form II score values from seven components .....	205
Table 6.2 Results of ANN analysis of PLS form II score values from two components .....	206
Table 6.3 Results of ANN analysis of PLS form III score values from six components .....	207
Table 6.4 Results of ANN analysis of PLS form III score values from two components .....	207
Table 6.5 Brief description of the eight descriptors used in this analysis.....	209
Table 6.6 Results of ANN analysis of PLS form II using the loading values from two most positive and negative descriptors from two components (eight descriptors) ..	209
Table 6.7 Results of ANN analysis of PLS form II using the loading values from the most positive and negative descriptors from two components (four descriptors) ...	210
Table 6.8 Brief description of the seven descriptors used in this analysis.....	211
Table 6.9 Results of ANN analysis of PLS form III using the loading values from the two most positive and negative descriptors from two components (seven descriptors) .....	211
Table 6.10 Results of ANN analysis of PLS form III using the loading values from the most positive and negative descriptors from two components (three descriptors) .....	212
Table 6.11 Brief description of the ten most important descriptors for form II prediction.....	213
Table 6.12 Results of ANN analysis of PLS form II using variable importance values .....	213
Table 6.13 Brief description of the ten most important descriptors for form III prediction.....	214
Table 6.14 Results of ANN analysis of PLS form III using variable importance values.....	215
Table 6.15 Results of ANN analysis of the overlapping descriptors from the top ten form II and III variable importance values .....	216
Table 6.16 Results of ANN analysis of the unique descriptors from the top ten form II and III variable importance values .....	216
Table 6.17 Optimisation of the model using previously highlighted informative descriptors and the most successful set.....	217
Table 7.1 FormRules and INForm results of PCA score analysis .....	221
Table 7.2 Results of PC1-5 score analysis .....	222
Table 7.3 The most significant descriptors in PC1 .....	223
Table 7.4 Most significant descriptors in PC2.....	224
Table 7.5 Most significant descriptors in PC3 .....	225
Table 7.6 Most significant descriptors in PC4.....	226
Table 7.7 Most significant descriptors in PC5.....	227

Table 7.8 Most positively and negatively loaded descriptors from PC1-13 (PCA-26)	228
Table 7.9 FormRules and INForm results of PCA-26	229
Table 7.10 FormRules and INForm results of PCA-10 analysis	229
Table 7.11 Top and bottom two descriptors from PC1-5 (PCA-20)	230
Table 7.12 FormRules and INForm results of PCA-20	231
Table 7.13 Summary of descriptor reduction results. The number in brackets is the number of descriptors used in the ANN	231
Table 7.14 FormRules and INForm results of PCA-8 analysis	232
Table 7.15 Optimisation results	233
Table 7.16 The ten most successful descriptors for polymorphic form prediction as determined by PCA analysis	234
Table 7.17 Summary of linear correlations of the optimised set	236
Table 7.18 Summary of correlation optimisations	238
Table 7.19 Summary of the descriptors in the most successful set	239
Table 7.20 Rules generated in FormRules for form I prediction	240
Table 7.21 Rules generated in FormRules for form II prediction	242
Table 7.22 Rules generated in FormRules for form III prediction	249
Table 7.23 Rules generated in FormRules for dihydrate prediction	254
Table 7.24 Rules generated in FormRules for solvate prediction	257
Table 7.25 Summary of the descriptors involved in the CBZ predictive rules	259
Table 7.26 Cross validation results summary	261
Table 7.27 External validation results summary	264
Table 8.1 The top sets of descriptors found in the manual and PCA analysis	269
Table 8.2 Linear correlations between the two most successful sets. Number in brackets is the correlation coefficient	270
Table 8.3 Optimisation of the PCA and Corr. best sets based upon the linear correlation results. X denotes the presence of the descriptor in the set	271
Table 8.4 Further optimisation of the Best sets. X denotes the presence of the descriptor in the set	272
Table 8.5 Optimisation of the descriptor set. X denotes the presence of the descriptor in the set	273
Table 8.6 Determination of the effect of removing the target molecules descriptors. X denotes the presence of the descriptor in the set	274
Table 8.7 Descriptors grouped based upon their physical meaning	276
Table 8.8 Optimisation results based upon the physical meanings of the descriptors. X denotes the presence of the descriptor in the set	277
Table 8.9 Optimisation results based upon predictions made by each descriptor. X denotes the presence of the descriptor in the set	279
Table 8.10 Further optimisation of the descriptor set. X denotes the presence of the descriptor in the set	281
Table 8.11 Summary of the descriptors involved in the final set	282
Table 8.12 Rules generated in FormRules for form I prediction	283
Table 8.13 Rules generated in FormRules for form II prediction	285
Table 8.14 Rules generated in FormRules for form III prediction	288
Table 8.15 Rules generated in FormRules for dihydrate prediction	290
Table 8.16 Rules generated in FormRules for solvate prediction	292
Table 8.17 Cross validation results summary	295
Table 8.18 External validation results summary	297

Table 9.1 Analysis of the TBA descriptors based upon previous CBZ research. X denotes the presence of the descriptor in the set .....	302
Table 9.2 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting only TBA form I .....	303
Table 9.3 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting only TBA form II .....	304
Table 9.4 Analysis of the ROY descriptors based upon previous CBZ research. X denotes the presence of the descriptor in the set .....	305
Table 9.5 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting ROY, Y and R forms .....	305
Table 9.6 The most positively and negatively loaded descriptors taken from the TBA principal components .....	309
Table 9.7 Results of the PCA data reduction analysis. The number in brackets is the number of descriptors used in the ANN.....	309
Table 9.8 Results of the PCA data reduction analysis when the data for the third polymorph is removed. The number in brackets is the number of descriptors used in the ANN .....	310
Table 9.9 Descriptors grouped based upon their physical meaning.....	311
Table 9.10 The most positively and negatively loaded descriptors taken from the ROY principal components.....	313
Table 9.11 Results of the PCA data reduction analysis. The number in brackets is the number of descriptors used in the ANN.....	313
Table 9.12 The descriptors that featured in the rules of PC1-4 (8).....	314
Table 9.13 Descriptors grouped based upon their physical meaning.....	315
Table 10.1 Summary of the descriptors involved in the final set .....	323
Table 12.1 CBZ polymorph screen experimental results.....	331
Table 12.2 Comparison of OPLS geometry optimised CBZ .....	349
Table 12.3 The Conformations of the ROY molecule and their associated energy. ....	350
Table 12.4 The descriptor sets used in the manual analysis .....	351
Table 12.5 Rules from the optimised CBZ descriptor set.....	355

## LIST OF FIGURES

Figure 1.1 Solubility curves in a) monotropic and b) enantiotropic systems <sup>[1]</sup> .....	27
Figure 1.2 Solubility curve, adapted from Davey et al. <sup>[1]</sup> .....	30
Figure 1.3 Diagram of how molecules at the crystals surface do not have a full compliment of intermolecular interactions, adapted from Davey et al. <sup>[1]</sup> .....	33
Figure 1.4 The free energy change versus the cluster size, adapted from Davey et al. <sup>[1]</sup> .....	35
Figure 1.5 2D-crystal growth. Dashed lines show the potential growth based on different growth rates of the crystal faces, with <i>a</i> growing faster than <i>b</i> , leading to a decreased <i>a</i> surface. ....	42
Figure 1.6 3D-crystal growth. Schematic of kink, step and flat sites of intermolecular binding to the crystal surface .....	43
Figure 2.1 Diagram highlighting the d spacing and angles of diffraction of the X- rays, adapted from Byrn <sup>[1]</sup> .....	48
Figure 2.2 Radleys greenhouse blowdown head <sup>[9]</sup> .....	50
Figure 2.3 XRPD of carbamazepine form I used as the standard diffraction pattern in this research.....	50
Figure 2.4 XRPD of carbamazepine form II used as the standard diffraction pattern in this research.....	51
Figure 2.5 XRPD of carbamazepine form III used as the standard diffraction pattern in this research.....	51
Figure 2.6 XRPD of carbamazepine form IV used as the standard diffraction pattern in this research. Taken from the theoretical powder patter from CSD reference CBMZPN12 .....	52
Figure 2.7 XRPD of carbamazepine dihydrate used as the standard diffraction pattern in this research.....	52
Figure 2.8 XRPD of carbamazepine DMSO solvate used as the standard diffraction pattern in this research .....	53
Figure 2.9 XRPD of ROY form Y used as the standard diffraction pattern in this research .....	53
Figure 2.10 XRPD of ROY form R used as the standard diffraction pattern in this research .....	54
Figure 2.11 XRPD of TBA form I used as the standard diffraction pattern in this research .....	54
Figure 2.12 XRPD of TBA form II used as the standard diffraction pattern in this research .....	55
Figure 2.13 XRPD of TBA form III used as the standard diffraction pattern in this research, taken from the supporting information of Thirunahari et al. <sup>[15]</sup> .....	56
Figure 2.14 XRPD of TBA form IV used as the standard diffraction pattern in this research, adapted from Sonoda et al. <sup>[16]</sup> .....	56
Figure 2.15 XRPD of TBA form V used as the standard diffraction pattern in this research, taken from Nath et al. <sup>[17]</sup> .....	57
Figure 2.16 Simplified representation of a potential energy surface showing both global and local minimum.....	59
Figure 2.17 A simplified biological neuron .....	64



Figure 2.18 Representation of a Threshold Logic Unit.....	65
Figure 2.19 graphical representation of a threshold function .....	66
Figure 2.20 the pattern space for the two input TLU showing the threshold which determines whether the output is 1 or 0 .....	67
Figure 2.21 Simplified two hidden layer neural network, with light blue shapes representing the nodes.....	68
Figure 2.22 a) hyperbolic tangent sigmoid, tanh and b) logistic nonlinear functions	70
Figure 2.23 A schematic of a simple neurofuzzy system, adapted from Shao et al <sup>[76]</sup> . .....	71
Figure 2.24 An example of a scree plot .....	75
Figure 3.1 The molecular structure of carbamazepine.....	82
Figure 3.2 Highlighting bond rotation in CBZ molecule.....	83
Figure 3.3 CBZ hydrogen-bonded dimer.....	84
Figure 3.4 Packing diagram for form I, taken from the CSD (reference CBMZPN11 <sup>[2]</sup> ).....	84
Figure 3.5 a) packing diagram of form II, taken from the CSD (reference CBMZPN03 <sup>[7]</sup> ) b) space filled model highlighting the possible site for solvent inclusion .....	86
Figure 3.6 Packing diagram for CBZ form III, taken from the CSD (reference CBMZPN01 <sup>[23]</sup> ) .....	86
Figure 3.7 Packing diagram of form IV, taken from the CSD (reference CBMZPN12 <sup>[5]</sup> ).....	87
Figure 3.8 Packing of CBZ dihydrate, taken from the CSD (reference FEFNOT02 <sup>[18]</sup> ) .....	88
Figure 3.9 Packing of the CBZ DMSO solvate, taken from the CSD (reference UNEYIV <sup>[12]</sup> ).....	88
Figure 3.10 Packing of the CBZ acetone solvate, taken from the CSD (reference CRBMZA01 <sup>[12]</sup> ) .....	89
Figure 3.11 Structure of ROY .....	90
Figure 3.12 Taken from a paper by Yu, this diagram shows the different morphologies and colours of the ROY polymorphs <sup>[35]</sup> .....	91
Figure 3.13 ROY crystal structures, a) form R (CSD reference QAXMEH02 <sup>[36]</sup> ) and b) ORP (CSD reference QAXMEH05 <sup>[36]</sup> ) .....	92
Figure 3.14 ROY crystal structures, a) form Y (CSD reference QAXMEH01 <sup>[36]</sup> ), b) YN (CSD reference QAXMEH04 <sup>[36]</sup> ) and c) YT04 (CSD reference QAXMEH12 <sup>[37]</sup> ) .....	93
Figure 3.15 ROY crystal structures, a) form ON (CSD reference QAXMEH <sup>[36]</sup> ) and b) OP (CSD reference QAXMEH03 <sup>[36]</sup> ) .....	93
Figure 3.16 a) intramolecular H-bond in ORP b) intermolecular H-bond in Y .....	94
Figure 3.17 The molecular structure of tolbutamide.....	94
Figure 3.18 U and chair type configurations of TBA polymorphs taken from Thirunahari et al. <sup>[46]</sup> .....	95
Figure 3.19 Packing diagram for form I, taken from the CSD (reference ZZZPUS02 <sup>[50]</sup> ) .....	96
Figure 3.20 DSC of form I, run at 2°C/min .....	96
Figure 3.21 DSC of small endotherm, run at 10°C/min.....	97
Figure 3.22 DSC of small endotherm, run at 1°C/min.....	97
Figure 3.23 Taken from Hasegawa et al. <sup>[51]</sup> the molecular structure and torsion angles of the two forms, a)Form I <sup>L</sup> b) form I <sup>H</sup> .....	98

Figure 3.24 Hot-stage XRPD of TBA form I above and below small endotherm seen in DSC analysis. Blue line is the form I sample at 25°C and the pink is the form I sample at 50°C .....	98
Figure 3.25 Packing diagram for form II, taken from the CSD (reference ZZZPUS03 <sup>[42]</sup> ) .....	99
Figure 3.26 C and D chair type conformations seen in the form II crystal structure, taken from Thirunahari et al. <sup>[46]</sup> .....	99
Figure 3.27 U type packing motif of TBA form III, taken from Thirunahari et al. <sup>[46]</sup> .....	100
Figure 3.28 U type packing motif of form IV taken from Thirunahari et al. <sup>[46]</sup> .....	101
Figure 4.1 Assignments of CBZ bond lengths and angles .....	109
Figure 4.2 The rotatable bonds in ROY (highlighted in green with a red line) .....	110
Figure 4.3 Highlighted torsion angles (Tx) within the TBA molecule that were subjected to a conformational search .....	110
Figure 4.4 Summary of the flow of work carried out in Hyperchem <sup>TM</sup> <sup>[1]</sup> .....	111
Figure 4.5 Summary of Gaussian 03 <sup>[6]</sup> work flow .....	112
Figure 4.6 Summary of the molecular descriptor calculation process .....	113
Figure 4.7 Screen shot of how the inputs and outputs are identified .....	115
Figure 4.8 Two examples of the graphical representation of the rules for different output predictions.....	118
Figure 4.9 Screen shot of how the starting descriptor values are selected.....	124
Figure 4.10 Screen shot from the 3D explorer feature showing a plot of d73 and rate against form II. The blue colouring on the plot indicates a high prediction value, and the red region is a low predictive value .....	125
Figure 4.11 The panel used to change the axis and descriptor values in the 3D explorer .....	125
Figure 4.12 The plot of d73, rate and form II when dsolv75 is at its highest value (initially it was a very low value).....	126
Figure 4.13 3D explorer plot of dsolv47 and VSA against form III.....	127
Figure 4.14 Shows the difference in plot shape when dsolv13 is set to its maximum value .....	127
Figure 4.15 Shows the difference in plot shape when dsolv13 is set to its minimum value .....	128
Figure 4.16 The differences in plots when dsolv13 is changed. From left to right: Maximum value, original value (mid-range) and minimum value .....	128
Figure 4.17 Summary of the overall analysis process.....	129
Figure 5.1 Summary of how the descriptor sets were created .....	132
Figure 5.2 Summary of the number of descriptor sets within each generation of analysis.....	132
Figure 5.3 Crystallisation solvents in which form I is produced plot against the three rule descriptors, rate (blue), temperature (red) and dsolv65 (cream). The shaded area highlights the most favourable descriptor values for form I production.....	173
Figure 5.4 The pure form II experiments plot against the normalised E_vdw values. The green shaded area highlights the most favourable descriptor values for form II production. ....	176
Figure 5.5 Diagram of the van der Waals interactions represented as a molecular surface. H-bonding region (pink), hydrophobic region (green) and mild polar region (blue) coloured upon the surface <sup>[22]</sup> .....	176

Figure 5.6 The experimental results that produced a mixture of form II and another form plot against the normalised E_vdw values. The shaded area highlights the most favourable descriptor values for form II production. ....	177
Figure 5.7 All crystallisation solvents plot against E_vdw normalised values. The shaded area highlights the most favourable descriptor values for form II production. ....	178
Figure 5.8 correlation of E_vdw and the dielectric constant of the solvents. Most favourable form II producing region is represented by the shaded area .....	178
Figure 5.9 The pure form II experiments plot against the normalised MNDO_dipole values. The shaded area highlights the most favourable descriptor values for form II production. ....	179
Figure 5.10 The pure form II experiments plot against the normalised rate values. The shaded area highlights the most favourable descriptor values for form II production. ....	180
Figure 5.11 The pure form III experiments plot against the normalised MNDO_dipole (purple) and Gutmann donor number values (pink). The shaded area highlights the most favourable descriptor values for form III production .....	182
Figure 5.12 Hydrogen bonding between formamide and dichloromethane (left) and nitromethane (right) .....	183
Figure 5.13 Hydrogen bonding in formamide.....	183
Figure 5.14 The pure form III experiments plot against the normalised E_vdw values. The shaded area highlights the most favourable descriptor values for form III production. ....	185
Figure 5.15 E_vdw plot against the dielectric constants of the solvents used in the crystallisations. The shaded area on the graph represents the potentially form III producing values .....	186
Figure 5.16 The pure form III experiments plot against the normalised rate values. The shaded area highlights the most favourable descriptor values for form III production. ....	186
Figure 5.17 Calculation of the total molecular surface area using van der Waals radii, adapted from <sup>[44]</sup> .....	189
Figure 5.18 The dihydrate producing experiments plot against the normalised values of rate (blue), temperature (purple), dsolv71 (green) and dsolv65 (cream).....	190
Figure 5.19 The solvate producing experiments (9 examples of DMSO solvent) plot against the normalised values of MNDO_dipole (purple) and dsolv57 (green) .....	191
Figure 5.20 Plot of all crystallisation solvents normalised MNDO_dipole (purple) and dsolv57 (green) values.....	192
Figure 5.21 Plot of the normalised descriptor values of EtOAc (E) and n-butanol (B) .....	201
Figure 7.1 Scree plot generated from the PCA results of carbamazepine (CBZ) descriptor analysis .....	221
Figure 7.2 Linear correlations of the optimised set with other descriptors.....	235
Figure 7.3 Crystallisation solvents in which form I is produced plot against the three rule descriptors, rate (blue), temperature (purple) and dsolv74 (yellow). The green section highlights the optimal value area for form I production.....	241
Figure 7.4 Calculation of the total molecular surface area using van der Waals radii, adapted from Stanton <sup>[13]</sup> .....	242
Figure 7.5 Crystallisation solvents in which form II is the pure product plot against normalised d69 descriptors values. The green section highlights the optimal value area for form II production.....	243

Figure 7.6 Original plot for form II prediction (left), effect on plot when a high E_ang value is used (right).....	244
Figure 7.7 Crystallisation solvents in which form II is the pure product plot against normalised rate values. The green section highlights the optimal value area for form II production.....	245
Figure 7.8 the axis of an single molecule, adapted from Atkins <sup>[25]</sup> .....	245
Figure 7.9 Crystallisation solvents in which form II is the pure product plot against normalised d68 values. The green section highlights the optimal value area for form II production.....	246
Figure 7.10 Original plot for form II prediction (left), effect on plot when a mid E_ang value (centre) and effect on plot when high E_ang value is used (right) .....	247
Figure 7.11 Original plot for form II prediction (left), and effect on plot when high dsolv76 value is used (right) .....	247
Figure 7.12 Crystallisation solvents in which form III is the pure product plot against normalised d68 values. The green section highlights the optimal value area for form III production. ....	249
Figure 7.13 Crystallisation solvents in which form III is the pure product plot against normalised d84 values. The green section highlights the optimal value area for form III production. ....	250
Figure 7.14 Crystallisation solvents in which form III is the pure product plot against normalised d77 and boiling point values. The green section highlights the optimal value area for form III production.....	251
Figure 7.15 Plot of d77 and boiling point in the prediction of form III.....	252
Figure 7.16 Plot of d77 and boiling point in the prediction of form III with mid range (left) and high (right) values of E_ang.....	253
Figure 7.17 Plot of d77 and boiling point in the prediction of form III with low (left), mid range (centre) and high (right) values of dsolv76.....	253
Figure 7.18 Crystallisation solvents in which form III is the pure product plot against normalised rate values. The green section highlights the optimal value area for form III production. ....	254
Figure 7.19 Crystallisation solvents in which dihydrate is part of the product plot against normalised rate (blue), temperature (purple), doslv74 (orange) and dsolv69 (red) values.....	256
Figure 7.20 Crystallisation solvents plot against normalised dsolv43 (blue) and boiling point (orange) values. ....	257
Figure 7.21 The distribution of the validation solvents descriptor values. E represents the ethyl acetate values and B the n-butanol values.....	262
Figure 8.1 Schematic of the different property regions found for the solvent descriptors .....	275
Figure 8.2 Schematic of the different property regions found for the CBZ descriptors .....	275
Figure 8.3 Schematic of the forms in which each descriptor influences prediction	278
Figure 8.4 Crystallisation solvents in which form I is produced plotted against the three rule descriptors, rate (blue), temperature (purple) and dsolv65 (cream). The green shaded area highlights the most favourable descriptor values for form I production. ....	284
Figure 8.5 Effect of dsolv69 upon the form II predictions based upon the rules generated. Low dsolv69 values (left), mid range values (centre) and high dsolv69 values (right) .....	286

Figure 8.6 Effect of dsolv69 upon the form II predictions based upon the rules generated. Low dsolv69 values (left) and high dsolv69 values (right) .....	287
Figure 8.7 Effect of dsolv69 upon the form III predictions based upon the rules generated. Low dsolv69 values (left), mid range values (centre) and high dsolv69 values (right) .....	289
Figure 8.8 Crystallisation solvents plot against normalised rate (blue) , temperature (purple), dsolv65 (cream) and dsolv43 (green/blue) values.....	292
Figure 8.9 The distribution of the validation solvents descriptor values. E represents the ethyl acetate values and B the n-butanol values.....	296
Figure 9.1 Normalised descriptor values plot for R form producing experiments. Rate (blue), temperature (purple) and E_ang (green).....	307
Figure 9.2 Scree plot based upon the TBA PCA data.....	308
Figure 9.3 Descriptors in the most successful set grouped based upon their physical meaning .....	310
Figure 9.4 Scree plot based upon the ROY PCA data .....	312
Figure 9.5 ROY PCA descriptors grouped based on their physical meaning.....	314
Figure 12.1 the axis of a single molecule, adapted from Atkins <sup>[6]</sup> .....	334
Figure 12.2 Calculation of the total molecular surface area using van der Waals radii, adapted from <sup>[15]</sup> .....	335
Figure 12.3 The methanol slurry to measure the stability of TBA. Commercial form I (black), two samples of the MeOH slurry after 2 days (red and green) and the MeOH slurry after 7 days (blue) .....	347
Figure 12.4 The ethanol slurry to measure the stability of TBA. Commercial form I (black), the EtOH slurry after 2 days (green) and the MeOH slurry after 2 days (red) for comparison .....	347
Figure 12.5 The dichloromethane slurry to measure the stability of TBA. Commercial form I (black), the DCM slurry after 2 days (blue), the DCM slurry seeded with form II after 2 days (green), the DCM slurry seeded with form II after 3 days (brown), the DCM slurry seeded with form II after 5 days (black trace above the red) and the MeOH slurry after 2 days (red) for comparison .....	348
Figure 12.6 The acetone slurry to measure the stability of TBA. Commercial form I (black), the acetone slurry after 2 days (blue), the acetone slurry after 7 days (green), the acetone slurry after 8 days (brown), the acetone slurry after 10 days (light blue), and the MeOH slurry after 2 days (red) for comparison.....	348
Figure 12.7 Assignments of CBZ bond lengths and angles .....	349
Figure 12.8 The torsion angle used in the ROY conformational analysis .....	350
Figure 12.9 Plot of molecular surface area descriptor (d69) against the dielectric constant of the solvents .....	354
Figure 12.10 Plot of molecular surface area descriptor (d69) against the dipole moment of the solvents .....	354

## ABSTRACT

A polymorphic substance is capable of forming a number of different crystalline phases that are referred to as its polymorphs. The critical process that determines the outcome of a crystallization process in a polymorphic system is thought to be the nucleation state, which is the self-assembled stage just prior to the formation of crystals with long-range order. While nucleation is well known to be influenced by macroscopically measurable parameters such as temperature, supersaturation and solvent choice our understanding of the underlying molecular self-assembly processes is very limited. The research described in this thesis explores a new approach to extending our knowledge in this area by the use of a combination of medium throughput crystallisation experiments together with the computation of a range of molecular and solute/solvent descriptors of the system under study.

The main objective of the work was to develop a protocol for relating experimental and computational data via artificial neural network (ANN) analysis, to identify significant links between experimental polymorphic outcomes and molecular properties. By creating a model that can predict the polymorphic form in a given experiment it is anticipated that our understanding of links between nucleation and crystallisation will be enhanced through the determining the pivotal properties of a molecule that cause it to form one polymorph over another.

The ANN method was developed in the context of the carbamazepine system, applying several statistical techniques to the results of 88 crystallisation experiments, featuring 13 solvents, 3 evaporation rates and 4 temperatures. The results show that this approach allows the formulation of further research hypotheses through examination of the physical meaning of the set of descriptors identified by the ANN approach. Crucially, principal component analysis (PCA) was found to be able to efficiently narrow down large sets of computationally derived descriptors to a manageable set by removing redundancy through strongly cross-correlated parameters. The best ANN model generated in this research was capable of predicting the major polymorphic form in 89 % of cross-validation experiments.

The optimised set of descriptors included both solute and solvent properties, which predominantly described the intermolecular interactions in solution. The physical meanings of the descriptors and their impact on the molecular processes during nucleation has been considered and their cross correlation has been examined. Initial results from further experimentation with the tolbutamide and ROY systems indicate that the methodology is also transferable to other polymorphic systems.

## **DECLARATION**

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning. This thesis is the result of my own work except where reference is made to other sources. The main body of this thesis contains no more than 80,000 words.

Vicky Fawcett

## COPYRIGHT STATEMENT

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.manchester.ac.uk/library/aboutus/regulations>) and in The University’s policy on presentation of Theses.



## ACKNOWLEDGEMENTS

Firstly, I would like to thank Prof. Roger Davey, Dr. Sven Schroeder and Dr. Jim McCabe for all their help and support throughout my PhD. This research would not have progressed as well without all of your expertise, so thank you all very much for your time and ideas.

Also, thank you to Guy Hembury, Ben Sattelle, Olof Svensson, Elizabeth Colbourn and Cristina Porro for their scientific support throughout my research.

There are so many other people that I would like to thank, but I won't mention them all by name, because I'm sure that someone will be forgotten. Huge thanks to all of the past and present office members for making the last few years brilliant! I've had a great time and it's a shame it has to end! Thank you also to the Chem. Eng Netball team. I've got to mention them because I loved being part of the team! We've had some success and a little less success, but it's always been fun!

I have to say an extra special thank you to Dad, Mum and Stephen. They have put up with the trail of destruction that has followed me for a number of years, and especially all the books and papers that have taken over our flat recently! Maybe one day I will throw something away (but it's unlikely)! Thank you so so so much!!!

Finally, I would also like to thank EPSRC and AstraZeneca for funding my PhD, without this it would not have been possible to carry out this work.

## THESIS OVERVIEW

**CHAPTER 1 - INTRODUCTION** – Presents an introduction to polymorphism, nucleation and crystallisation and also the current research in these areas.

**CHAPTER 2 – METHODS AND MATERIALS** – Provides information about the experimental techniques used and polymorphic systems analysed. Background theory and software information for all the computational aspects of this research is also provided.

**CHAPTER 3 – SYSTEMS STUDIES** – Literature review of the polymorphic systems studied in this research (carbamazepine (CBZ), tolbutamide (TBA) and 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY)).

**CHAPTER 4 – ANALYSIS METHODOLOGY** – An overview of the molecular modelling, descriptor calculation and Artificial Neural Network (ANN) analysis in this research.

**CHAPTER 5 – RESULTS AND DISCUSSION – MANUAL ANALYSIS** – Presentation of the different analyses carried out using the CBZ dataset. The aim was to reduce the number of descriptors that could make a successful prediction of CBZ polymorphic form from 167 to approximately 10. The results during the development of a predictive ANN are presented and also an overview of the physical meanings of the successful descriptor set that led to a prediction.

**CHAPTER 6 – RESULTS AND DISCUSSION – PLS ANALYSIS** – Another method of descriptor dataset reduction is presented, using partial least squares (PLS) analysis. ANN models were built by selecting important features from the data and attempting to predict the polymorphic outcome of crystallisation results. The initial PLS analysis was carried out by Dr. O. Svensson at AstraZeneca, who then provided the data for further analysis.

**CHAPTER 7 – RESULTS AND DISCUSSION – PCA ANALYSIS** – A further method of descriptor dataset reduction, using principal component analysis (PCA). As with the PLS analysis, this was also carried out by Dr. O. Svensson at AstraZeneca, who then provided the data for further analysis. Descriptors were selected and predictive models built in the ANN. This method led to a successful model, allowing the discussion of the physical meanings of these descriptors. It also demonstrated a transferrable method for selecting important descriptors, which can lead to a successful predictive ANN.

**CHAPTER 8 – FINAL OPTIMISATION AND DISCUSSION OF RESULTS**– The predictive models created in chapters 5 and 7 were subject to further analysis in order to determine if further optimisation could occur. The descriptors in the final model for CBZ prediction were also discussed with regards to how they relate to nucleation and crystallisation.

**CHAPTER 9 – RESULTS AND DISCUSSION OF ANALYSIS WITH DIFFERENT TARGET MOLECULES** – Two additional polymorphic systems were examined, TBA and ROY. The descriptors highlighted in the CBZ analysis (chapter 8) and also the method of descriptor selection highlighted in chapter 7 were assessed to see if the descriptors and the methods were transferrable to different polymorphic target molecules.

**CHAPTER 10 – CONCLUSIONS** – Conclusions based upon the success of the CBZ predictive model for polymorphism were discussed. The preliminary results based on the different polymorphic target molecules were also presented.

**CHAPTER 11 – FURTHER WORK** –Further work that may lead to an improved predictive model, which is also transferrable to other molecules, was presented.

**CHAPTER 12 – APPENDIX** – Experimental polymorph screen results, descriptor definitions, modelling results, XRPD traces and predictive rules presented.

## 1. INTRODUCTION

The crystallisation of polymorphic molecules is an area of pivotal interest to the pharmaceutical industry. The ability to theoretically predict crystallisation outcomes for a given set of experimental conditions just from a set of molecular descriptors would be extremely beneficial to product development, intellectual property generation and its protection. This thesis aims to highlight a method whereby experimental data and theoretically calculated molecular descriptors are brought together through artificial neural network methodologies, in order to predict the polymorphic forms generated by crystallisation processes.

This introductory chapter aims to provide information on the principles of the nucleation and crystallisation of different polymorphic forms. By using knowledge of the molecular descriptors gained from successful polymorph prediction, new insights into the nucleation of polymorphic forms may be generated.

### 1.1. A Crystalline Substance

Solids in which molecules are packed in a regular way with long range order are known as crystals. If however there is only short range order the solid is said to be amorphous<sup>[1]</sup>. In this work only crystals will be investigated and it is the subtle molecular packing arrangements which are of interest.

Due to the highly ordered and symmetric nature of such molecular packings, crystals grow in regular shapes, for example, needle-like, plates and prisms, this is known as the crystals morphology<sup>[2-5]</sup>. In some situations, such as polymorphism (section 1.2) the morphologies can be different between the forms, which can be used as a means of identification. However, morphology differences are not a necessary condition of polymorphism.

### 1.2. Polymorphism

Polymorphism is a phenomenon that occurs in a variety of different fields of science. Within chemistry, a polymorph is defined as a substance that can exist in more than one different crystalline phases that display different chemical and physical properties, for example melting points, solubility, density and bioavailability<sup>[6-9]</sup>. The

crystals of a polymorphic substance may have conformational differences<sup>[10]</sup> or different molecular packings within the unit cell<sup>[11]</sup>. In the example of carbamazepine and 4-chlorophenol<sup>[7]</sup> all of the polymorphs have identical molecular structures within the unit cell, but it is the way in which they are arranged that creates the different forms<sup>[12]</sup>. This is not the case in all polymorphic systems such as L-glutamic acid<sup>[13]</sup> and 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY)<sup>[10]</sup>; in some systems the molecules within the unit cell are also conformationally different, these systems are known as conformational polymorphs.

Closely related to conformational polymorphism is polychromism, demonstrated by the ROY polymorphs, *N*-(*p*-chlorobenzylidene)-*p*-chloroaniline<sup>[14]</sup> and dimethyl-3, 6-dichloro-2,5-dihydroxyterephthalate<sup>[15]</sup>, but this phenomenon of coloured crystals is not commonly seen in organic molecules. Research suggests that the colour of ROY may arise from electron delocalisation between the phenyl and thiophene rings<sup>[16]</sup>. Yu et al.<sup>[17]</sup> suggested that the colour is due to the conformational differences that alter the “ $\pi$ -conjugation between the *o*-nitroaniline chromophore and the thiophene ring”,<sup>[17]</sup>.

### 1.2.1. Solvates and Hydrates

Solvates and hydrates, although not strictly referred to as polymorphs, are closely related species that are important in crystallisation. A solvate is a crystal that regularly incorporates a solvent molecule within its crystal structure<sup>[7]</sup>, similarly a hydrate incorporates a water molecule<sup>[5]</sup>. There are no clear reasons as to why a solvate will form, but it may be due to an increased level of stability in the crystal structure<sup>[7, 18]</sup>. Hydrates may form due to the small size of the water molecule that can easily fit within a molecule and they also have multidirectional potential for hydrogen-bonding,<sup>[7]</sup> which can stabilize a crystal structure.

### 1.2.2. Relative Stability of Polymorphs

In a polymorphic system (for example with two different forms) in equilibrium the two forms conform to Gibbs phase rule (Equation 1.1), with  $F$  representing the degrees of freedom,  $P$  being the number of phases and  $C$  the number of components<sup>[19]</sup>.

$$F = C - P + 2 \quad \text{Equation 1.1}$$

In an example with only two forms, there is only one degree of freedom ( $F=1$ ), which means that if either the temperature or pressure is changed; the corresponding value that allows equilibrium to be retained can be found. If however the system is trimorphic,  $F=0$ , there is only one value of temperature and pressure at which equilibrium can be retained. This temperature value is known as the transition temperature ( $T_t$ ), and in a polymorphic systems is represented by the Clansius-Clapeyron equation (Equation 1.2), in which  $\Delta V$  represents the difference in molar volume between the polymorphs;  $\Delta H_t$  is the latent heat of transition, and  $T$  and  $P$  represent temperature and pressure<sup>[1]</sup>.

$$\frac{dT}{dP} = \frac{T_t \Delta V}{\Delta H_t} \quad \text{Equation 1.2}$$

Going back to a dimorphic example (forms I and II), the more stable form (II) has a lower solubility (independent of solvent) than form I. When a crystal of each form is placed into a supersaturated solution, both chemical potentials ( $\mu$ ) with regard to the solid and liquid phase become equal (Equation 1.3) and by assuming ideality, the solubility ( $x_{eq}$ ) of each form can be represented (Equation 1.4).

$$\begin{aligned} \mu_{solid}(II) &= \mu_{eq}(II) = \mu^0 + RT \ln x_{eq}(II) \\ \mu_{solid}(I) &= \mu_{eq}(I) = \mu^0 + RT \ln x_{eq}(I) \end{aligned} \quad \text{Equation 1.3}$$

$$\begin{aligned} \mu_{solid}(II) &< \mu_{solid}(I) \\ x_{eq}(II) &< x_{eq}(I) \end{aligned} \quad \text{Equation 1.4}$$

Density is also important in polymorphism and can be related to the stability of the form by the density rule<sup>[7, 20]</sup>. This rule states that the lower the density of a form, the less stable the polymorph is, due to a more dense solid having stronger intermolecular interactions<sup>[20, 21]</sup>.

There are two key relationships between sets of polymorphs, these are monotropism and enantiotropism. For a polymorphic pair with a monotropic relationship, the solubilities of the forms are independent of temperature. Whereas if they are enantiotropically related, the relative solubilities are dependent on temperature, with

transformations between forms being a reversible process<sup>[6]</sup>. These concepts are most easily highlighted in Figure 1.1.

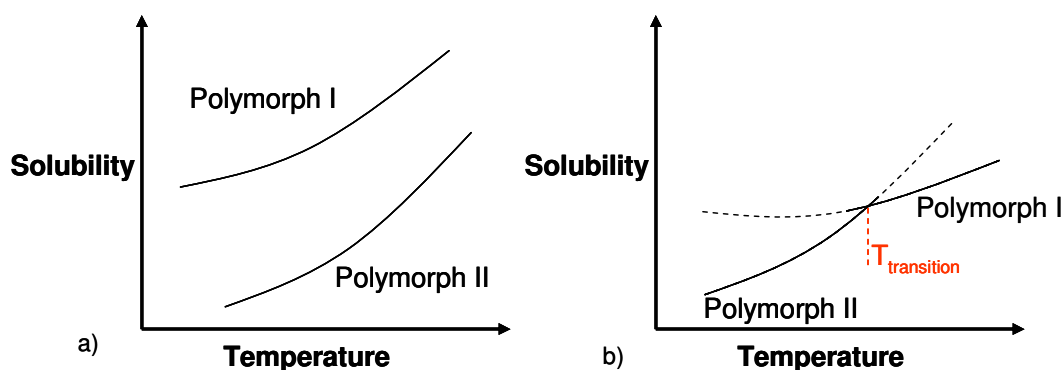


Figure 1.1 Solubility curves in a) monotropic and b) enantiotropic systems<sup>[1]</sup>

### 1.2.3. Polymorphism in the Pharmaceutical Industry

The problem of polymorphism in the pharmaceutical industry has been demonstrated in a number of cases. Perhaps one of the most famous examples of polymorphism and the impact on the pharmaceutical industry is that of Ritonavir, a protease inhibitor for the treatment of Acquired Immunodeficiency Syndrome (AIDS)<sup>[22, 23]</sup>. Throughout development only one crystal form was ever identified, however two years after the drug went to market, the capsules in a new batch failed dissolution testing<sup>[22]</sup>. Using microscopy, a new form was identified, as it had a different crystal habit than observed previously. It was later revealed that this form had a much lower solubility and would affect the efficacy of the drug. According to Ostwald's rule of stages (section 1.4.4) the least stable form crystallises first and in the case of Ritonavir this is what had been seen throughout manufacture.

This is an exceptional example that has led to tighter restrictions on pharmaceutical drugs, however the reason polymorphism was not investigated more thoroughly in this case was because it was a semi-solid formulation. ICH (international conference on harmonisation) guidelines state that “for a drug product that is in solution, there is little scientific rationale for polymorphic control”<sup>[23]</sup>, however this statement has now been proven incorrect, and the criteria amended.

Another feature of polymorphism of interest to the pharmaceutical industry is their ability to disappear and in some cases reappear years later<sup>[24-27]</sup>. Dunitz and Bernstein<sup>[26]</sup> detail a number of examples (1,2,3,5-tetra-*O*-acetyl- $\beta$ -D-ribofuranose,

benzocaine:picric acid, melibiose and mannose to name but a few) in which initially one form was crystallised, but after this it could not be made again. In the case of benzocaine:picric acid further research has been carried out<sup>[28]</sup> and the form that had once disappeared can now be crystallised and observed using thermal microscopy methods.

#### **1.2.4. Polymorph Prediction**

The ability to predict crystal structures based on only the molecular structure, by calculating the global minimum lattice energy<sup>[29]</sup>, is an area of increasing interest in the scientific community. Since 1999 the Cambridge Crystallographic Data Centre (CCDC) has regularly sent test sets of molecules to the leading predictive crystal structure groups by as a challenge to the latest predictive methods<sup>[30-32]</sup>. The 5<sup>th</sup> blind test is being completed in 2010<sup>[33]</sup>. Advances have been made in this area which has then led onto the prediction of polymorphic crystal structures<sup>[34]</sup>. By calculating the global minimum lattice energy, other energetically feasible structures can be highlighted that could indicate polymorphism within a given molecule<sup>[29]</sup>. Work by Price<sup>[30, 34]</sup> discusses how computational calculations can be used to predict polymorphic outcome and states that it is important to consider kinetic factors for successful predictions<sup>[30]</sup>.

Della Valle et al.<sup>[35]</sup> introduced a predictive method that combined experimental Raman spectroscopy data with theoretical energy minimisation of structures. For their model systems, sexithiophene, tetracene and pentacene, the method predicted experimentally known polymorphs well; however the results were inconclusive for many other molecules. This research highlights that there is potential for polymorph prediction to occur, however due to the complex nature of the task, it is very difficult.

Previous work by McCabe<sup>[8]</sup> utilised bulk solvent properties to predict the polymorphic outcome of a carbamazepine polymorph screen using artificial neural networks. This work proved the concept of using solvent properties to predict polymorphic outcomes and led to this current study into the molecular level parameters. Bulk property understanding is very useful on a practical scale, especially in the pharmaceutical industry where efficiency is the key, it does not however improve our understanding of what is occurring at the molecular level when one polymorph is crystallised over another in solution.



### 1.3. Crystallisation

Crystallisation is a heavily researched field due to its widespread application in many of the chemical industries<sup>[36-39]</sup>; but it is within the context of the pharmaceutical industry that is of importance to this research. There are a number of ways to crystallise a product, with the two key methods being suspension and solidification. Suspension requires the sample to be in solution and at supersaturation. If the solubility curve for the solute is known, the exact method by which a product can be obtained is more easily determined. For example, if the solubility is heavily affected by temperature then cooling crystallisation would be most suitable<sup>[40, 41]</sup>. Alternatively, when the temperature has minimal impact on the solubility, evaporation is preferred as it is more likely to generate an increased yield. If however the sample is presented as a liquid that is above its melting point (i.e. a melt) then spray drying can be used, whereby the solution is sprayed and droplets solidify on cooling<sup>[1]</sup>.

In this research cooling and evaporative crystallisation techniques have been employed, because of equipment for multiple simultaneous crystallisations by these methods was available.

#### 1.3.1. Factors Affecting Crystallisation

Controlling the crystallisation process can allow purer products to be generated and potentially polymorphic form selection. The concentration of the sample<sup>[42]</sup>, availability of seeds<sup>[43]</sup>, rate of agitation<sup>[7]</sup>, temperature<sup>[41]</sup> and evaporation or cooling rate<sup>[40]</sup> may all contribute to the crystallisation process. Therefore it is essential to understand these factors in order to improve efficiency and reduce the costs of crystallisation in industry. A discussion of the key factors that affect crystallisation in this work follows.

### 1.3.1.1. Supersaturation

Within this research a supersaturated solution was created in two ways, which then led to the crystallisation of different polymorphic forms. Crystalline material was dissolved in a given solvent at a set temperature and nitrogen blown down onto it to control the evaporation rate. By evaporating the solvent at a fixed temperature the overall volume is reduced and therefore the concentration in solution is increased, creating a supersaturated solution that will then crystallise. The second method is to again dissolve crystalline material in a given solvent and then reduce the temperature. The reduction in temperature decreases the solubility and therefore moves the solution into the supersaturated region (Figure 1.2).

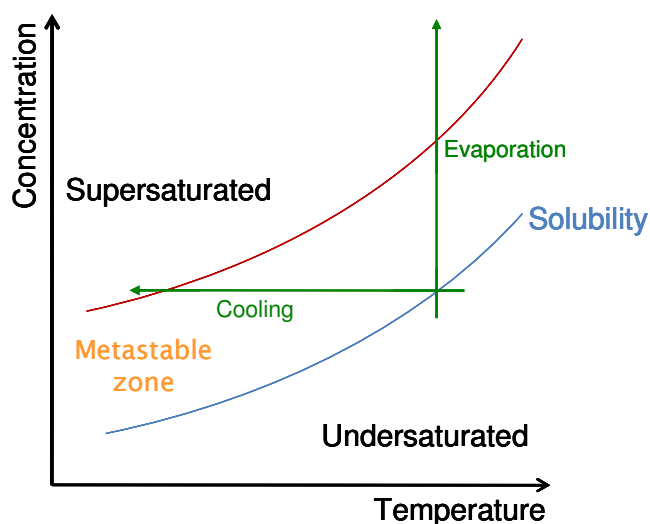


Figure 1.2 Solubility curve, adapted from Davey et al.<sup>[1]</sup>

Crystallisation will not occur when the solution is undersaturated (all of the crystalline material is dissolved<sup>[1]</sup>) and therefore one of the two mentioned techniques must be used in order to put the solution into a region of the solubility diagram where crystallisation can occur. When the metastable zone is exceeded, crystallisation is spontaneous, but if a solution falls into the metastable zone, it is unlikely that crystals will form spontaneously. If a seed was introduced to the solution, growth may occur in the metastable region<sup>[6]</sup>.

For crystallisation to occur the solution needs to be supersaturated, therefore the amount of dissolved solute exceeds equilibrium<sup>[1]</sup>. Supersaturation ( $\sigma$ ) can be defined in thermodynamic terms (Equation 1.5), as the chemical potential difference

between the equilibrium ( $\mu_{eq}$ ) and the supersaturated ( $\mu_{ss}$ ) state of a solute<sup>[1]</sup>, with  $k$  being the Boltzmann constant and  $T$  the temperature of the system in kelvin.

$$\sigma = \frac{(\mu_{ss} - \mu_{eq})}{kT} \quad \text{Equation 1.5}$$

If the solution is ideal then the equation can be reduced further to enable supersaturation to be calculated using the composition of the solution (Equation 1.6).

$$\sigma = \ln\left(\frac{x_{ss}}{x_{eq}}\right) \quad \text{Equation 1.6}$$

It is important to note that the degree of supersaturation can affect the outcome of a crystallisation experiment. By creating a saturated solution at varied high temperatures with filtration to remove all excess solid each solution can be cooled to a single specific temperature, which therefore achieves different supersaturation concentrations<sup>[44]</sup>.

#### **1.3.1.2. Seeding**

Seeding can allow control of morphology, particle size and the polymorphic form crystallised<sup>[45]</sup> and can also promote crystallisation in the metastable region of the solubility diagram<sup>[6]</sup>. Controlling the crystallisation of a particular polymorphic form can be influenced by the addition of seeds of the desired form, which was demonstrated in the literature<sup>[43, 46, 47]</sup>. He et al.<sup>[46]</sup> discussed that if the concentration is only a little above saturation, the form of the seeds may not always result in the same form being crystallised<sup>[46]</sup>, demonstrating the importance of knowing the solubility diagram for a system.

#### **1.3.1.3. Temperature**

Temperature ( $T$ ) can affect the solubility ( $x_{eq}$ ) of a crystal in solution<sup>[40, 41]</sup>, with this relationship shown in Equation 1.7, where  $a$ ,  $b$  and  $c$  are constants derived from experimental data (Equation 1.8)<sup>[1, 40]</sup>.

$$\ln x_{eq} = a + \frac{b}{T} + c \ln T \quad \text{Equation 1.7}$$

$$\begin{aligned} a &= (\Delta S_{S \rightarrow L} - C_p) / R \\ b &= (\Delta H_{S \rightarrow L} - \Delta C_{p S \rightarrow L} T_{ref}) / R \\ c &= (\Delta C_{p S \rightarrow L}) / R \end{aligned} \quad \text{Equation 1.8}$$

$C_p$  is the molar heat capacity, with  $\Delta C_{p S \rightarrow L}$  representing the difference between the heat capacities of the solid and solute in solution.<sup>[40]</sup>  $\Delta H_{S \rightarrow L}$  the change in molar enthalpy and  $\Delta S_{S \rightarrow L}$  the molar entropy change of the solid to the solute in liquid.  $R$  is the gas constant. Previous research has shown that varying the temperature of crystallisation can alter the polymorph crystallised<sup>[44, 48, 49]</sup>; which is an important factor in this research.

### 1.3.2. Polymorph Screen

Polymorph screening is a standard technique used in the pharmaceutical industry to determine if there are any polymorphic forms of a molecule<sup>[50, 51]</sup>. The screens usually take place early in the development of a new drug<sup>[52]</sup> so that any polymorphs can be included in the patent of the molecule. By varying the solvent, temperature and rates in different types of crystallisation, a wide range of experimental conditions can be covered. This allows the identification of any new polymorphic forms to be conducted. Characterisation techniques often include X-ray powder diffraction (XRPD), differential scanning calorimetry (DSC), infrared (IR) and Raman spectroscopies<sup>[52]</sup>. It is now becoming more common that these polymorph screens are automated so that a larger experimental space can be interrogated<sup>[48, 52]</sup>.

In the research discussed in this thesis the polymorph screens were carried out to generate input data for the neural network analysis (detailed in 4.3).

## 1.4. Nucleation

Prior to crystallisation there is an aggregation of molecules in solution. Once such aggregations have reached a critical size, it becomes energetically favourable for them to grow into crystals<sup>[53]</sup>. Research is ongoing into how many molecules are involved when reaching a critical size nucleus. Yau et al.<sup>[54]</sup> used atomic force microscopy (AFM) to observe the number of molecules of protein apoferritin present

in a critical size nucleus. 50, 20 and 10 molecules were observed at a number of different supersaturations (1.1, 1.6 and 2.3 respectively)<sup>[54]</sup>. These aggregates in solution are held together by intermolecular interactions, such as hydrogen-bonds, van der Waals and coulombic interactions<sup>[55]</sup>. In some systems it is unknown as to whether nucleation has occurred spontaneously or due to other factors such as agitation, foreign particles or crystals<sup>[6]</sup>. There are two key types of nucleation, primary and secondary, both of which will be explained in the following section.

### 1.4.1. Primary Nucleation

There are a number of factors that need to be considered in the nucleation of crystals, the first of which is the free energy of cluster formation. When a cluster is formed some of the molecules are in the bulk of the crystal ( $Z_b$ ) and others are on the surface ( $Z_s$ ). The surface molecules of the crystal are under stress as they lack nearest neighbour molecules<sup>[1]</sup>, therefore do not have a full compliment of intermolecular bonds (depicted in Figure 1.3). This stress encourages growth in order to satisfy the bonding requirements of the molecules, and exerts a pressure on the cluster and hence raises its chemical potential.

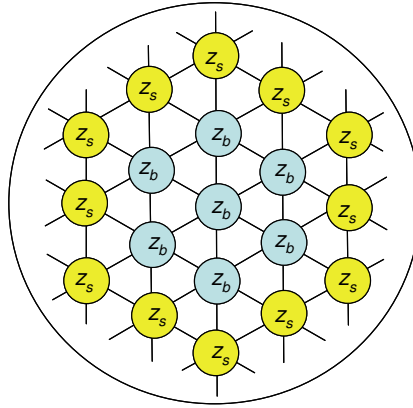


Figure 1.3 Diagram of how molecules at the crystals surface do not have a full compliment of intermolecular interactions, adapted from Davey et al.<sup>[1]</sup>

The free energy of the cluster is represented by Equation 1.9 with  $g_b$  and  $g_s$  being the solid bulk and surface free energies respectively.

$$g_z = (Z_b + Z_s)g_b + (g_s - g_b)Z_s \quad \text{Equation 1.9}$$

Between the cluster and the solution the interfacial tension ( $\gamma$ ) plays a role, as does the surface area of the cluster ( $A$ ) which can be included in the free energy equation (Equation 1.10 and Equation 1.11).

$$\gamma = \frac{(g_s - g_b)Z_s}{A} \quad \text{Equation 1.10}$$

hence

$$g_z = Zg_b + \gamma A \quad \text{Equation 1.11}$$

If a spherical cluster is formed with  $Z$  molecules then the surface area becomes proportional to the number of molecules (Equation 1.12).

$$A \propto Z^{2/3} \quad \text{Equation 1.12}$$

When this relationship is included in the clusters free energy calculation and is written in terms of chemical potentials of the molecules in the bulk of the cluster ( $\mu_b$ ), Equation 1.13 applies. This equation also introduces the area shape factor of the nucleus ( $\beta$ ).

$$g_z = Z\mu_b + \beta\gamma Z^{2/3} \quad \text{Equation 1.13}$$

If the cluster is made up of molecules  $A$  which are in the bulk liquid phase at mole fraction  $x_{ss}$  then nucleation can be represented by a quasi equilibrium which occurs between the cluster and monomers in solution (Equation 1.14).



The free energy change per mole of  $A_z$  upon nucleation is given by Equation 1.15.

$$\Delta G = g_z - Z\mu \quad \text{Equation 1.15}$$

This equation includes a chemical potential term for the monomers ( $\mu$ ), and since it is known that  $\mu = \mu^0 + kT \ln x_{ss}$ , then Equation 1.16 can be formed, which shows the free energy change upon nucleation.

$$\Delta G = (Z\mu_b + \beta\gamma Z^{2/3}) - Z(\mu^0 + kT \ln x_{ss}) \quad \text{Equation 1.16}$$

When the solution is saturated,  $x = x_{eq}$ , therefore  $\mu_b = \mu^0 + kT \ln x_{eq}$  and thus the equation for free energy can be re-written (Equation 1.17). This equation also includes a term for supersaturation,  $\ln(x_{ss}/x_{eq})$ .

$$\Delta G = -ZkT \ln \frac{x_{ss}}{x_{eq}} + \beta\gamma Z^{2/3} \quad \text{Equation 1.17}$$

Equation 1.17 states the relationship between free energy change and supersaturation, which is depicted in Figure 1.4. When the nuclei have reached the critical size ( $Z_c$ ) the free energy begins to decrease. As is apparent from the diagram, the degree of supersaturation affects the height of the free energy barrier.

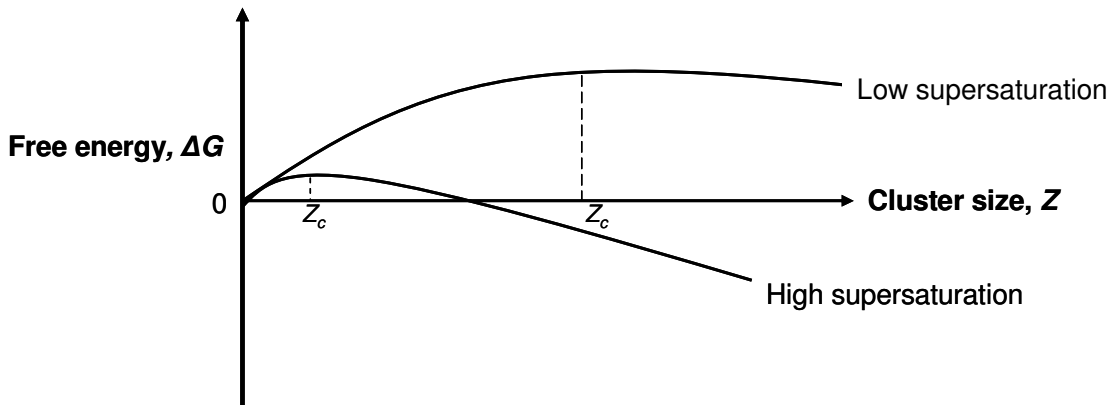


Figure 1.4 The free energy change versus the cluster size, adapted from Davey et al.<sup>[1]</sup>

High supersaturations require less energy to form critically sized nuclei, which upon a further decrease in barrier height can lead to spontaneous nucleation. The rate at which clusters form and grow to their critical size is defined as the nucleation rate.

In primary nucleation it is assumed that the formation of clusters containing the nuclei of molecule A is a stepwise process until the critical size ( $Z_c$ ) is reached<sup>[1, 55]</sup>.



The equilibrium constant ( $K_z$ ) for the formation of the critical nucleus<sup>[55]</sup> of the system is shown in Equation 1.19.

$$K_z = [A_c] / [A]^{Z_c}$$

$$\ln K_z = -\Delta G_c / RT \quad \text{Equation 1.19}$$

By relating the equilibrium constant to the activation free energy for nucleation ( $\Delta G_c$ ), Equation 1.20 can be derived, and by assuming the nucleus to be spherical with a radius ( $r$ ), interfacial tension ( $\gamma$ ) the equation can be rewritten (Equation 1.21).

$$[A_c] = [A]^{z_c} \exp(-\Delta G_c / RT) \quad \text{Equation 1.20}$$

$$\Delta G_c = (4\pi r_c^3 / 3)\Delta G_b + 4\pi r_c^2 \gamma \quad \text{Equation 1.21}$$

Overall the rate of nucleation ( $J$ ) is represented by Equation 1.22, which emphasises how essential the supersaturation ( $\sigma$ ), temperature ( $T$ ), molar volume in the crystal ( $v_c$ ) and interfacial tension ( $\gamma$ ) are to the crystallisation process<sup>[1]</sup>.

$$J = K_J \exp(-B_J \gamma^3 / T^3 \sigma^2) \quad \text{Equation 1.22}$$

where  $B = 16\pi \gamma^3 v_c^2 / 3R^3 T^3$

#### 1.4.1.1. Homogeneous Nucleation

Homogeneous nucleation assumes a stepwise aggregation of molecules, starting spontaneously from a supersaturated solution without influence from other factors. Classical nucleation theory is based on work by Gibbs, Volmer, Becker and Döring, where crystallisation from solution is compared to the condensation of a vapour into liquid<sup>[6]</sup>. Homogeneous nucleation is a rare occurrence in systems with a volume over 100  $\mu\text{L}$ , and due to the presence of impurities that may induce nucleation, heterogeneous nucleation is more commonly observed<sup>[45]</sup>.

#### 1.4.1.2. Heterogeneous Nucleation

When a foreign body is found in a supersaturated solution it may inhibit or accelerate the nucleation rate, and is referred to as heterogeneous nucleation. Whether this substance was added intentionally<sup>[56]</sup> or is a contaminant in the system it can lead to nucleation of a crystal product. The presence of a surface allows adsorption of the molecules that lowers the activation free energy for nucleation<sup>[1, 45]</sup>.

#### 1.4.2. Secondary Nucleation

When a seed of the desired product is presented to a solution in order to induce nucleation, this is termed secondary nucleation. This method is employed in industry as a means to control polymorphic outcome, morphology and particle size<sup>[45]</sup>, and



will often play a dominant role in large scale crystallisations<sup>[56]</sup>. However, the secondary nucleation rate ( $B$ ) induced by the addition of seeds is not only affected by the concentration of the seeds ( $M_T$ ), but also on the stirrer speed ( $N$ ) which affects the solution-crystal interactions and supersaturation ( $\sigma$ ) (Equation 1.23).

$$B = kM_T^j N^k \Delta c^b \quad \text{Equation 1.23}$$

### 1.4.3. The Nucleation of Polymorphs

An expression for the rate of nucleation has been presented in Equation 1.22, but it can also be expressed in the following way when addressing the issue of more than one polymorphic form (Equation 1.24).  $J_I$  represents the nucleation rate of metastable form I,  $J_{II}$  the nucleation rate of stable form II and  $K$  the equilibrium constants.

$$\begin{aligned} J_I &= K_{J,I} \exp[-B_I / (\sigma_i - \sigma_x)^2] \\ \text{and} \quad J_{II} &= K_{J,II} \exp[-B_{II} / \sigma_i^2] \end{aligned} \quad \text{Equation 1.24}$$

These equations come from the supersaturations of the solution initially ( $\sigma_i$ ) (Equation 1.25) and when the solution is saturated with form I,  $\sigma_x$  (Equation 1.26)<sup>[1]</sup>.

$$\sigma_i = (x_i - x_{II}) / x_{II} \quad \text{Equation 1.25}$$

$$\sigma_x = (x_I - x_{II}) / x_{II} \quad \text{Equation 1.26}$$

When  $B$  is defined (Equation 1.27) dimensionless variables can also be defined (Equation 1.28), leading to the two rate equations being solved.

$$B = 16\pi\gamma^3 v^2 / 3R^3 T^3 \quad \text{Equation 1.27}$$

$$\begin{aligned} a &= \sigma_x B_{II}^{1/2} \\ b &= (B_I / B_{II}) \\ c &= [a / \ln(K_{J,II} / K_{J,I})]^{1/3} \end{aligned} \quad \text{Equation 1.28}$$

The solved rate equations for a polymorphic systems highlight under what conditions form I or II have the higher nucleation rate.

If  $K_{J,I} > K_{J,II}$  then above a certain supersaturation value the metastable form I has the higher nucleation rate, but below this value form II would nucleate more quickly. This was shown in work by Cornel et al.<sup>[57]</sup> with the polymorphs of D-mannitol, whereby low supersaturations nucleated the most stable form initially.

If  $K_{J,II} > K_{J,I}$  and  $(1 - a/c)^3 < b$ , then the more stable form II has the higher nucleation rate across the whole range of supersaturation values.

If  $K_{J,II} > K_{J,I}$  and  $(1 - a/c)^3 > b$ , then the metastable form I has the higher nucleation rate over the whole range of supersaturation values.

These three conditions go against what is known as Ostwald's rule of stages, which shall now be discussed.

#### 1.4.4. Ostwald's Rule of Stages

Ostwald's rule of stages formulated in 1897<sup>[55]</sup> suggested that the change from supersaturation to equilibrium, involves a number of steps. Each of these steps represent the smallest change in free energy possible<sup>[1]</sup>. To apply this to a polymorphic system, the changes in free energy represent the movement from the least stable polymorphic form through to the thermodynamically stable form, going through all other possible forms. By following this rule of stages, the least stable form will always be crystallised first<sup>[1]</sup> and a slurry should allow full conversion to the thermodynamically stable form.

Ostwald acknowledged that there would be exceptions to this rule<sup>[6]</sup>, and current research supports these deviations, but, in many cases this rule is true.

Concomitant polymorphism deviates from Ostwald's rule of stages by suggesting that two or more forms can simultaneously crystallise<sup>[21, 58]</sup>. However, the major deviation that concerns the research discussed in this thesis is the phenomenon of cross-nucleation, whereby seeds of one form nucleate another<sup>[59]</sup>. There are relatively few examples of cross-nucleation in the literature; ROY (discussed in section 3.2) is one, another D-Mannitol<sup>[59-63]</sup>. Tao et al.<sup>[62]</sup> highlight that the early nucleating form "does not consume the entire liquid"<sup>[62]</sup>, and can go on to nucleate a more or less thermodynamically stable form. This highlights that the stability of a form does not affect whether it can cross-nucleate<sup>[63]</sup>. Desgranges et al.<sup>[61]</sup> suggested that cross-nucleation is governed by kinetic factors and that it occurs between stable and metastable forms that have very similar free energies. It has also been suggested that there is a need for the cross-nucleated form to grow comparably or faster than the

form that initially nucleated<sup>[36, 60, 62, 63]</sup>, but this does not guarantee that it will occur. Chen et al.<sup>[60]</sup> stated that it is important that the nucleation of the new form is a rapid process, otherwise it may not occur. There is little literature on this topic, and only a few examples of cross-nucleation exist, but it is an important process to note within this research.

Research has been conducted by Cornel et al.<sup>[57]</sup> into the nucleation of the three polymorphs of D-mannitol. They have observed that the degree of supersaturation affects which form is crystallised first, highlighting another deviation from Ostwald's rule of stages. The observation that at a low supersaturation the most stable  $\gamma$  polymorph nucleated directly<sup>[57]</sup> rather than initially nucleating as the less stable form, was made experimentally.

#### 1.4.5. Detection of Nucleation

There are a number of macroscopic parameters that can be monitored in order to confirm that nucleation has occurred in a system, these are temperature, light transmittance and concentration. Nucleation causes a reduction in free energy and therefore heat is produced, DSC or simply a thermometer can be used to monitor this<sup>[1]</sup>. The optical transmittance can be monitored as nucleation causes a change in solution clarity due to increased number of particles. Density, refractive index and conductivity can also all be monitored in order to assess the concentration changes, and it is expected that the concentration of the solution will decrease upon nucleation<sup>[1]</sup>.

The induction time can also be characterised by measuring the time difference between the establishment of supersaturation and the occurrence of nucleation ( $\tau_{ind}$ ), which is useful in determining the rate of nucleation (Equation 1.29)<sup>[1]</sup>.

$$J \propto \frac{1}{\tau_{ind}} \quad \text{Equation 1.29}$$

$$\ln \frac{1}{\tau_{ind}} \propto \frac{\gamma^3}{T^3 \sigma^2} \quad \text{Equation 1.30}$$

The induction time is heavily influenced by the degree of supersaturation, levels of agitation, viscosity and presence of impurities or seeds<sup>[6]</sup>.

#### 1.4.6. Current and Previous Research into Nucleation

Detection and analysis of the nucleation state is an important research area that could give new insight into the mechanism of crystallisation. Particularly relevant to this work would be the discovery of why one polymorph is crystallised over another under certain experimental conditions, and to highlight whether there is order within the solution that is then carried through to the crystallised product. Experimental analysis and theoretical modelling has been carried out in order to learn more about nucleation, with a number of examples presented here.

Experimental interrogation of the nucleation state has met with some success in the protein<sup>[54]</sup> and colloidal<sup>[53]</sup> fields due to the larger size of particles involved. However moving to small molecules presents further problems. Yau et al.<sup>[54]</sup> observed using atomic force microscopy (AFM) that the molecular arrangement of proteins within the nuclei was similar to that found in the crystal structure.

An insight into the nucleation of small molecules was generated by Banerjee et al.<sup>[64]</sup> by determining the crystal structure of Na(saccharinate).*n*H<sub>2</sub>O. Although this molecule has been known for over 100 years, the crystal structure was never solved. In solving the structure they uncovered a number of unusual features not commonly found in a crystal, earning it the title of a “model for nucleation”<sup>[64]</sup> as they believe it to be a close representation of a crystal nucleus. The unusual features were a large asymmetric unit in which there are areas of order and disorder and the presence of extra solvent, suggesting full crystallisation has not yet occurred<sup>[64]</sup>.

Prior to the work of Banerjee et al.<sup>[64]</sup> there have been a number of studies into nucleation, using a variety of techniques and molecules. Davey et al.<sup>[55]</sup> analysed the polymorphic systems of sulfathiazole and 2,6-dihydroxybenzoic acid to probe molecular self-assembly in solution. This was monitored by introducing an additive that inhibits nucleation of one of the polymorphs. This work suggested that molecular assembly in solution does indeed direct nucleation, and that by using different solvents the nucleation of different forms may be obtained<sup>[55]</sup>.

Different experimental techniques have also been utilised in the study of nucleation for example NMR<sup>[49, 65]</sup>, Infrared (IR) spectroscopy<sup>[66, 67]</sup> and neutron scattering<sup>[68]</sup>. Spitaleri et al.<sup>[65]</sup> use <sup>1</sup>H-NMR to study the nucleation of sulfamerizine with a degree of success. They suggested that by using the changing chemical shift influenced by the change in concentration, combined with genetic algorithm and NMR analysis

software shifty, insight can be gained into the packing of the molecules in solution<sup>[65]</sup>. Davey et al.<sup>[66]</sup> assessed the difference in molecular structure of carboxylic acids in solution and when crystalline using IR spectroscopy. Benzoic and tetrolic acids have evidence of a dimeric structure in solution that is also evident in the solid form, a conclusion also expressed by Parveen et al.<sup>[66]</sup>. In contrast, racemic mandelic acid was more complex. Features known to be present in the crystal structure were not observed in solution and also no distinction could be made between racemic and pure enantiomers in solution<sup>[66]</sup>. More recent work by Burton et al.<sup>[68]</sup> generated structural information on the metastable solution of urea using neutron scattering combined with molecular simulations<sup>[68]</sup>. It was found that hydration played an important role in this metastable solution, and that a urea molecule was surrounded by seven other urea and eight water molecules. These findings are in contrast to molecular packing observed in the solid state, in which fourteen urea molecules are found at equivalent radial distances<sup>[68]</sup>.

Directly relevant to this current research is the investigation of the nucleation of polymorphic systems, for example inosine<sup>[49]</sup> and tetrolic acid<sup>[67]</sup>. Chiarella et al.<sup>[49]</sup> worked with inosine which has two forms and a dihydrate, and interrogated the saturated solution using NMR spectroscopy. In solution there was evidence of molecular stacking and also the presence of dimer-like structures that translated into the solid  $\alpha$  form of inosine. It was also noted that the temperature affected the polymorphic outcome and below 10°C the formation of the dihydrate was favoured, but with no differences seen in the solution structure. This raises the question as to the relationship between the solution structure and the solid form. Parveen et al.<sup>[67]</sup> investigated dimorphic tetrolic acid in different solvents using IR spectroscopy. Dimers were observed in solution, but in polar solvents dimerisation was disrupted and the catemeric form was formed<sup>[67]</sup>. Their research highlights the important role solvent plays in polymorphic crystallisation, which will be addressed in the research discussed in this thesis.

Understanding the nucleation state has also led to computational modelling of the phenomenon. Browning et al.<sup>[69]</sup> interrogated different levels of undercooling and found that with a small level of undercooling, metastable crystals were theoretically observed, whereas for strong undercooling it was the most stable crystal<sup>[69]</sup>. Work has also been conducted using small molecules, namely yellow isoxazolone dye, paracetamol and L-Glutamic acid by Deij et al.<sup>[70]</sup> using molecular dynamic

simulations. These three molecules are polymorphic and the structures were built up assuming that the correct molecular orientation are determined at the nucleation stage<sup>[70]</sup>. Deij et al.<sup>[70]</sup> highlighted the importance of interfacial energy, and commented upon how the use of different solvents can alter this energy and direct the formation of the desired polymorph<sup>[70]</sup>. To further consolidate the effect of solvent on polymorph selection, Sharma et al.<sup>[71]</sup> modelled the effect of solvents on polymorphic form and noted that there is clearly an influence on the crystalline product.

## 1.5. Growth of Polymorphic Crystals

Once the nuclei have reached a critical size and crystallisation has begun, the crystals in solution continue to grow. Within the supersaturated environment the number of molecules joining the crystal surface exceeds the number leaving and therefore leads to crystal growth. How able a crystal is to capture these molecules and incorporate them into the crystal lattice is determined by the strength of the interactions between the surface molecules and molecules in solution<sup>[1]</sup>. Figure 1.5 highlights the growth of a two-dimensional crystal and demonstrates the differing growth rates of the crystal surfaces.

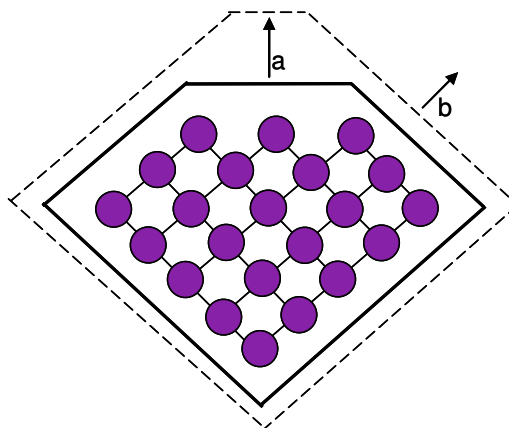


Figure 1.5 2D-crystal growth. Dashed lines show the potential growth based on different growth rates of the crystal faces, with *a* growing faster than *b*, leading to a decreased *a* surface.

There are two intermolecular interactions on face *a* and only one on face *b*. When molecules join face *a*, more energy is provided to the system by binding to the site with most interactions and therefore it will grow faster<sup>[4]</sup>.

The situation is slightly more complex with respect to a three-dimensional crystal (Figure 1.6). A maximum of three intermolecular interactions can occur, which are found at *kinked* (K) sites. When only two intermolecular interactions are possible the attachment site is *stepped* (S) and if there is only one possible intermolecular interaction, the site is *flat* (F).

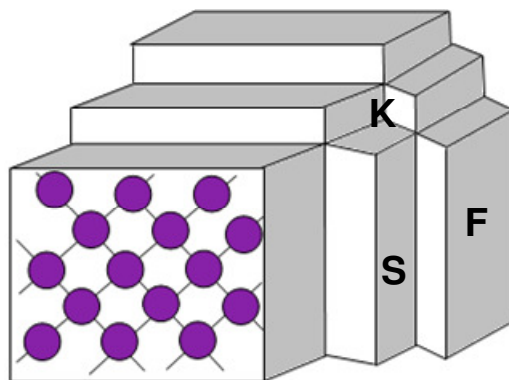


Figure 1.6 3D-crystal growth. Schematic of kink, step and flat sites of intermolecular binding to the crystal surface

Equation 1.31 shows the relationship between the linear growth rates at the different sites of binding on a three-dimensional crystal.

$$v_K > v_S > v_F \quad \text{Equation 1.31}$$

As discussed previously (section 1.1), many crystals have different morphologies. These morphological differences are determined by the growth speed of the crystal faces. In cases of polymorphism, where solvent plays a determinant role in which form is crystallised, the morphological differences that are sometimes seen may be due to the way in which the solvent interacts with the growth face<sup>[3, 72]</sup>.

The growth of a crystal can also be affected by the simultaneous growth of other crystals,<sup>[5]</sup> something that is highlighted in polymorphic system ROY (see section 1.4.4). There is also a well documented affect of additives on crystal growth, which can inhibit or accelerate the growth of polymorphic forms by mimicking the conformation of the molecule and inhibiting growth<sup>[38, 73, 74]</sup>.

### 1.5.1. Polymorphic Phase Transitions

There are both solvent-mediated and solid state phase transitions. In solvent-mediated phase transitions the crystallised metastable form is dissolved into the solution allowing the renucleation and recrystallisation of the stable form<sup>[1, 75]</sup>. As the activation energy for this process is lower<sup>[56]</sup> than that of a solid state transformation, this method is more favourable when the transition is taking place below the melting point of the polymorph<sup>[1]</sup>. Solution-mediated phase transformations are observed in numerous systems, for example 2,6-dihydroxybenzoic acid<sup>[56]</sup>, paclobutrazol<sup>[75]</sup>, D-mannitol<sup>[57]</sup> but most importantly to this research carbamazepine<sup>[76]</sup>.

In a solid state phase transition the stable phase is nucleated and grows in the unstable crystals in a reversible reaction<sup>[1]</sup>. This is often seen between enantiotropic polymorphs, and therefore by changing the temperature the forms can interconvert and recrystallise. Anwar et al.<sup>[77]</sup> observed the phase transformation of form IV to form I of sulfathiazole upon heating. They monitored this transition using time-resolved powder x-ray diffraction and observed the decrease in form IV peaks and the increase in those attributed to form I. Li et al.<sup>[78]</sup> have observed a similar transition in a derivative of ROY<sup>[78]</sup>.

## 1.6. Summary

This chapter aimed to highlight the principles of the nucleation and crystallisation of different polymorphic forms and why polymorph investigation is relevant to the pharmaceutical industry. By increasing knowledge surrounding the molecular level factors that may impact which polymorph is crystallised, more controlled experimental work can be carried out, insights into nucleation may be generated and predictions of crystallisation experiments could be made.



- [1] R. Davey, J. Garside, *From Molecules to Crystallizers: An Introduction to Crystallization*, Oxford University Press, Oxford, **2000**.
- [2] P. Taulelle, J. P. Astier, C. Hoff, G. Pere, S. Veessler, *Chemical Engineering & Technology* **2006**, 29, 239.
- [3] M. M. Parmar, O. Khan, L. Seton, J. L. Ford, *Crystal Growth and Design* **2007**, 7, 1635.
- [4] A. S. Myerson, *Handbook of Industrial Crystallization*, 1st ed. ed., Butterworth-Heinemann, Woburn, MA, **2002**.
- [5] J. K. Haleblan, *Journal of Pharmaceutical Sciences* **1975**, 64, 1269.
- [6] J. W. Mullin, *Crystallization*, Third ed., Reed Educational and Professional Publishing Ltd, Oxford, **2000**.
- [7] S. R. Byrn, R. R. Pfeiffer, J. G. Stowell, *Solid State Chemistry of Drugs*, Second ed., SSCI, Inc., West Lafayette, **1999**.
- [8] J. F. McCabe, *CrystEngComm* **2010**, 12, 1110.
- [9] J. W. Chew, S. N. Black, P. S. Chow, R. B. H. Tan, K. J. Carpenter, *CrystEngComm* **2007**, 9, 128.
- [10] L. Yu, G. A. Stephenson, C. A. Mitchell, C. A. Bunnell, S. V. Snorek, J. J. Bowyer, T. B. Borchardt, J. G. Stowell, S. R. Bryn, *Journal of the American Chemical Society* **2000**, 122, 585.
- [11] D. J. W. Grant, in *Polymorphism in Pharmaceutical Solids*, Vol. 95, first ed. (Ed.: H. G. Brittain), Marcel Dekker, Inc., New York, **1999**, pp. 1.
- [12] A. Grzesiak, M. Lang, K. Kim, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2003**, 92, 2260.
- [13] J. Bernstein, *Acta Crystallographica, Section B: Structural Science* **1991**, 47, 1004.
- [14] J. Bernstein, I. Izak, *J. C. S. Perkin II* **1976**, 429.
- [15] D. Y. Curtin, S. R. Byrn, *J. Am. Chem. Soc* **1969**, 91, 6102.
- [16] G. A. Stephenson, T. B. Borchardt, S. R. Byrn, J. Bowyer, C. A. Bunnell, S. V. Snorek, L. Yu, *Journal of pharmaceutical sciences* **1995**, 84, 1385.
- [17] L. Yu, *Journal of Physical Chemistry A* **2002**, 106, 544.
- [18] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Chemical Communications* **2007**, 1600.
- [19] Y. Li, P. S. Chow, R. Tan, B. H., S. N. Black, *Organic Process Research and Development* **2008**, 12, 264.
- [20] A. Burger, R. Ramberger, *Mikrochimica Acta* **1979**, II, 259.
- [21] J. Bernstein, R. J. Davey, J.-O. Henck, *Angewandte Chemie, International Edition* **1999**, 38, 3440.
- [22] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, J. Morris, *Pharmaceutical Research* **2001**, 18, 859.
- [23] S. R. Chemburkar, J. Bauer, K. Deming, H. Spiwek, K. Patel, J. Morris, R. Henry, S. Spanton, W. Dziki, W. Porter, J. Quick, P. Bauer, J. Donaubauer, B. A. Narayanan, M. Soldani, D. Riley, K. McFarland, *Organic Process Research and Development* **2000**, 4, 413.
- [24] I. Barsky, J. Bernstein, P. W. Stephens, K. H. Stone, E. Cheung, M. B. Hickey, J.-O. Henck, *Crystal Growth and Design* **2008**, 8, 63.
- [25] J.-O. Henck, J. Bernstein, A. Ellern, R. Boese, *J. Am. Chem. Soc* **20014**, 123, 1834.
- [26] J. D. Dunitz, J. Bernstein, *Accounts of Chemical Research* **1995**, 28, 193.
- [27] N. Blagden, R. J. Davey, R. Rowe, R. Roberts, *International Journal of Pharmaceutics* **1998**, 172, 169.

- [28] J.-O. Henck, J. Bernstein, A. Ellern, R. Boese, *Journal of the American Chemical Society* **2001**, *123*, 1834.
- [29] T. Beyer, T. Lewis, S. L. Price, *CrystEngComm* **2001**, *44*, 1.
- [30] S. L. Price, *Advanced Drug Deliver Reviews* **2004**, *56*, 301.
- [31] A. Asmadi, M. A. Neumann, J. Kendrick, P. Girard, M.-A. Perrin, F. J. J. Leusen, *Journal of Physical Chemistry B* **2009**, *113*, 16303.
- [32] M. A. Neumann, F. J. J. Leusen, J. Kendrick, *Angewandte Chemie* **2008**, *120*, 1.
- [33] [http://www.ccdc.cam.ac.uk/about\\_ccdc/science\\_profile/structure\\_prediction/](http://www.ccdc.cam.ac.uk/about_ccdc/science_profile/structure_prediction/), Viewed on 29/11/10.
- [34] C. L. Price, *Accounts of Chemical Research* **2009**, *42*, 117.
- [35] R. G. Della Valle, E. Venuti, A. Brillante, *Journal of Physical Chemistry A* **2008**, *112*, 6715.
- [36] L. Yu, *Accounts of Chemical Research* **2010**, *43*, 1257.
- [37] A. Gavezzotti, *CrystEngComm* **2002**, *4*, 343.
- [38] R. J. Davey, N. Blagden, G. D. Potts, R. Docherty, *Journal of the American Chemical Society* **1997**, *119*, 1767.
- [39] R. J. Davey, J. Richards, *Journal of Crystal Growth* **1985**, *71*, 597.
- [40] F. L. Muller, S. Black, *Organic Process Research and Development* **2009**, *13*, 1315.
- [41] S. Black, F. Muller, *Organic Process Research and Development* **2010**, *14*, 661.
- [42] S. Datta, D. J. W. Grant, *Crystal Research and Technology* **2005**, *40*, 233.
- [43] Y. Suzuki, K. Hara, *Bulletin of the Chemical Society of Japan* **1974**, *47*, 2551.
- [44] A. Getsoian, R. M. Lodaya, A. C. Blackburn, *International Journal of Pharmaceutics* **2008**, *348*, 3.
- [45] N. Rodriguez-Hornedo, D. Murphy, *Journal of Pharmaceutical Sciences* **1999**, *88*, 651.
- [46] X. He, U. J. Griesser, J. G. Stowell, T. B. Borchardt, S. R. Bryn, *Journal of Pharmaceutical Sciences* **2001**, *90*, 371.
- [47] P. G. Vekilov, *Crystal Growth and Design* **2004**, *4*, 671.
- [48] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, *95*, 1918.
- [49] R. A. Chiarella, A. L. Gilon, R. C. Burton, R. J. Davey, G. Sadiq, A. Auffret, M. Cioffi, C. A. Hunter, *Faraday Discuss.* **2007**, *136*, 179.
- [50] G. P. Stahly, *Crystal Growth and Design* **2007**, *7*, 1007.
- [51] M. Alleso, F. Van Den Berg, C. Cornett, F. S. Jorgensen, B. Halling-Sorensen, H. Lopez De Diego, L. Hovgaard, J. Aaltonen, J. Rantanen, *Journal of Pharmaceutical Sciences* **2008**, *97*, 2145.
- [52] R. Hilfiker, J. Berghausen, F. Blatter, A. Burkhard, S. M. De Paul, B. Freiermuth, A. Geoffroy, U. Hofmeier, C. Marcolli, B. Siebenhaar, M. Szelagiewicz, A. Vit, M. Von Raumer, *Journal of Thermal analysis and Calorimetry* **2003**, *73*, 429.
- [53] U. Gasser, E. R. Weeks, A. Schofield, P. N. Pusey, D. A. Weitz, *Science* **2001**, *292*, 258.
- [54] S.-T. Yau, P. G. Vekilov, *Journal of the American Chemical Society* **2001**, *123*, 1080.
- [55] R. J. Davey, K. Allen, N. Blagden, W. I. Cross, H. F. Lieberman, M. J. Quayle, S. Righini, L. Seton, G. J. T. Tiddy, *CrystEngComm* **2002**, *4*, 257.

- [56] R. J. Davey, N. Blagden, S. Righini, H. Alison, E. S. Ferrari, *Journal of Physical Chemistry B* **2002**, *106*, 1954.
- [57] J. Cornel, P. Kidambi, M. Mazzotti, *Ind. Eng. Chem. Res.* **2010**, *49*, 5854.
- [58] A. Singh, I. S. Lee, A. S. Myerson, *Crystal Growth and Design* **2009**.
- [59] L. Yu, *Journal of the American Chemical Society* **2003**, *125*, 6380.
- [60] S. Chen, H. Xi, L. Yu, *Journal of the American Chemical Society* **2005**, *127*, 17439.
- [61] C. Desgranges, J. Delhommelle, *Journal of the American Chemical Society* **2006**, *128*, 10368.
- [62] J. Tao, L. Yu, *Journal of Physical Chemistry B* **2006**, *110*, 7098.
- [63] L. Yu, *CrystEngComm* **2007**, *9*, 847.
- [64] R. Banerjee, P. M. Bhatt, M. T. Kirchner, G. R. Desiraju, *Angewandte Chemie, International Edition* **2005**, *44*, 2515.
- [65] A. Spitaleri, C. A. Hunter, J. F. McCabe, M. J. Packer, S. L. Cockroft, *CrystEngComm* **2004**, *6*, 489.
- [66] R. J. Davey, G. Dent, R. K. Mughal, S. Parveen, *Crystal Growth and Design* **2006**, *6*, 1788.
- [67] S. Parveen, R. J. Davey, G. Dent, R. G. Pritchard, *Chemical Communications* **2005**, 1531.
- [68] R. C. Burton, E. S. Ferrari, R. J. Davey, J. Hopwood, M. J. Quayle, J. L. Finner, D. T. Bowron, *Crystal Growth and Design* **2008**, *8*, 1559.
- [69] A. R. Browning, M. F. Doherty, G. H. Fredrickson, *Physical Review E* **2008**, *77*.
- [70] M. A. Deij, J. H. ter Horst, H. Meekes, P. Jansens, E. Vlieg, *Journal of Physical Chemistry B* **2007**, *111*, 1523.
- [71] S. Sharma, T. P. Radhakrishnan, *Journal of Physical Chemistry B* **2000**, *104*, 10191.
- [72] J. H. Ter Horst, R. M. Geertman, G. M. van Rosmalen, *Journal of Crystal Growth* **2001**, *230*, 277.
- [73] R. Dowling, R. J. Davey, R. A. Curtis, G. Han, S. K. Poornachary, P. S. Chow, R. Tan, B. H., *Chemical Communications* **2010**, *46*, 5924.
- [74] R. M. Ginde, A. S. Myerson, *Journal of Crystal Growth* **1993**, *126*, 216.
- [75] R. J. Davey, P. T. Cardew, D. McEwan, D. E. Sadler, *Journal of Crystal Growth* **1986**, *79*, 648.
- [76] D. Murphy, F. Rodriguez-Cintron, B. Langevin, R. C. Kelly, N. Rodriguez-Hornedo, *International Journal of Pharmaceutics* **2002**, *246*, 121.
- [77] J. Anwar, P. Barnes, *Kinetics of Phase Transformations in Crystals of Drug Compounds Using Time-Resolved Powder X-ray Diffraction*, Vol. 39, Gordon and Breach Science Publishers, S. A., **1992**.
- [78] H. Li, J. G. Stowell, X. He, K. R. Morris, S. R. Bryn, *Journal of Pharmaceutical Sciences* **2007**, *96*, 1079.

## 2. METHODS AND MATERIALS

The research presented in this thesis utilises both experimental and theoretical methods in order to understand and predict the formation of polymorphic forms. This chapter provides an overview over the techniques and materials in this work.

### 2.1. X-Ray Powder Diffraction

X-ray powder diffraction (XRPD) has been used in this research to identify which polymorph has been formed in the crystallisation work. XRPD uses X-rays of a known wavelength, 1.5406 Å, to probe the crystal structure of a molecule. The monochromatic x-ray beam is directed at the crystalline sample, which then produces a diffraction pattern, giving information about the crystal lattice. With the knowledge of  $d$  spacing and the angle at which the peak was generated, the structure can be characterised. This information is generated by understanding the Bragg equation (Equation 2.1), in which  $n$  represents the order of diffraction,  $\lambda$  is the wavelength and  $d$  is the distance between planes in the crystal<sup>[1, 2]</sup>.

$$n\lambda = 2d \sin \theta \quad \text{Equation 2.1}$$

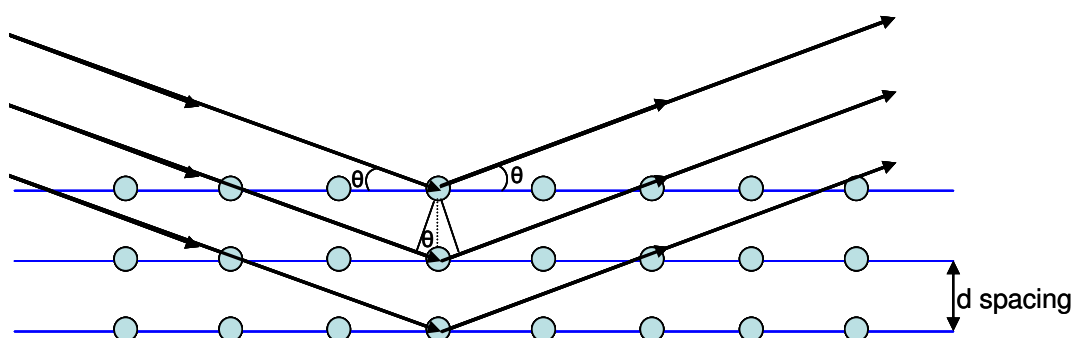


Figure 2.1 Diagram highlighting the  $d$  spacing and angles of diffraction of the X-rays, adapted from Byrn<sup>[1]</sup>

XRPD is useful for polymorphic identification as different crystal lattices, which polymorphs have, will produce unique diffraction patterns. This technique has been noted as a valuable tool for identification of polymorphs<sup>[2-5]</sup>, and in some cases can be used in quantitative analysis<sup>[1, 6, 7]</sup>, which may be useful in future work.

Three different powder diffraction instruments were used in this research for analysis of the different polymorphs formed. The main analyses were carried out using the Bruker D8 instruments at AstraZeneca and within the Chemistry department. CBZ samples were flatted onto a wafer using a microscope slide prior to analysis and ROY samples were lightly ground before flattening onto the wafer due to their needle-like morphology.<sup>[8]</sup>

### 2.1.1. XRPD Instruments

All samples were analysed over  $2^{\circ}$  to  $40^{\circ}$   $2\theta$ , at  $25^{\circ}\text{C}$  with slightly different step sizes and speeds depending upon the instrument.

Table 2.1 XRPD instrument details

XRPD Instrument	Location of Instrument	Step Size ( $^{\circ}$ )	Speed (Seconds per step)	Number of scans
Bruker D8	AstraZeneca	0.014	0.2	1
Bruker D8	The University of Manchester	0.014	0.2	3*
Bruker D4	AstraZeneca	0.0057	0.03	1
Rigaku miniflex benchtop	The University of Manchester	0.03	1.2	1

\*Three scans were required to achieve comparable resolution

## 2.2. Polymorph Screen

The evaporative crystallisations were conducted on a small scale simultaneously (maximum of 24 at once) using a Radleys greenhouse blowdown head (Figure 2.2). This apparatus allowed control of temperature and evaporation rate, with the experimental conditions of 5, 15 and 25 L/min of nitrogen at temperatures of 25, 50, 75 and  $100^{\circ}\text{C}$  being used in all possible combinations. The results of the carbamazepine polymorph screen can be found in Appendix 12.1.



Figure 2.2 Radleys greenhouse blowdown head<sup>[9]</sup>

## 2.3. Carbamazepine Experimental Details

From the commercially available carbamazepine (CBZ), forms I, II, III, Dihydrate and the DMSO solvate can be readily made. Form IV requires addition of a second component but can be produced successfully as shown by Lang et al.<sup>[10]</sup>.

### 2.3.1. Form I

Form I is the triclinic form of CBZ and can be made by heating commercial CBZ at 150°C for 3 hours<sup>[3]</sup>. The XRPD trace obtained from this material was used as the standard form I diffraction pattern in this work (Figure 2.3).

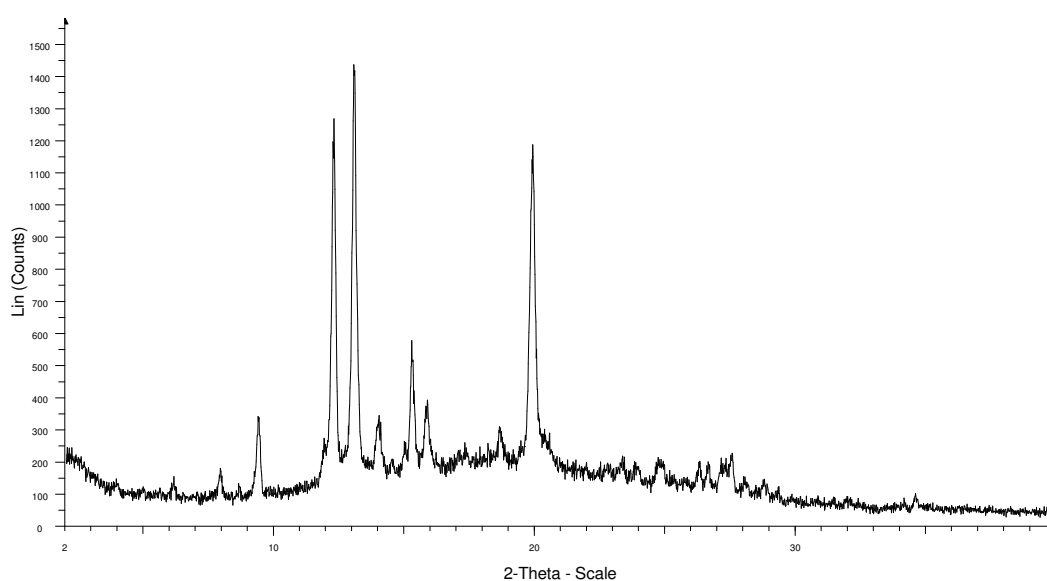


Figure 2.3 XRPD of carbamazepine form I used as the standard diffraction pattern in this research

### 2.3.2. Form II

Form III was dissolved into chloroform aided by heat ( $\frac{1}{3}$  total volume of vessel). The vessel was then filled with petroleum ether and vacuum filtered immediately. This preparation resulted in the crystallisation of form II, with the standard XRPD trace highlighted in Figure 2.4.

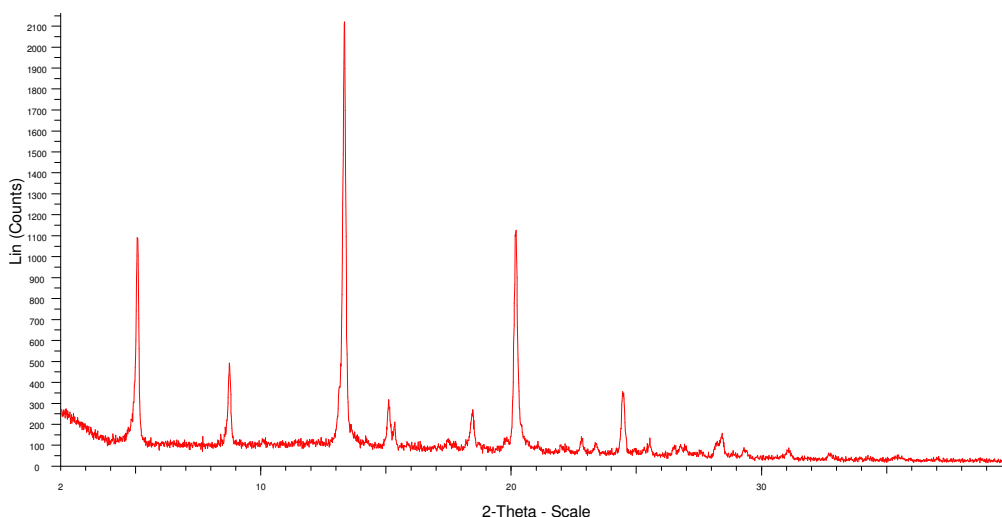


Figure 2.4 XRPD of carbamazepine form II used as the standard diffraction pattern in this research

### 2.3.3. Form III

Form III is commercially available from Sigma Aldrich, but can also be crystallised by slow evaporation from ethanol<sup>[3]</sup>. The XRPD trace of this thermodynamically stable form can be seen in Figure 2.5.

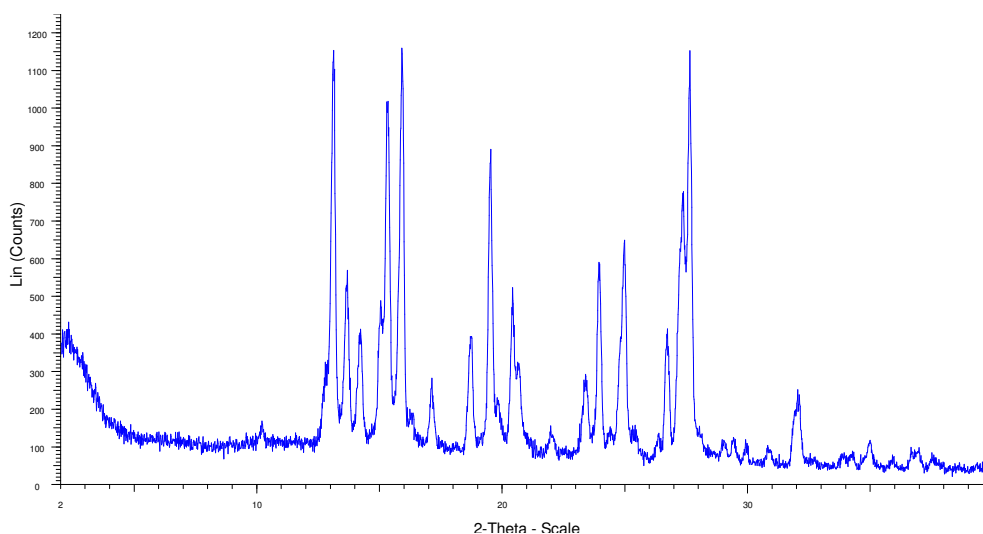


Figure 2.5 XRPD of carbamazepine form III used as the standard diffraction pattern in this research

#### 2.3.4. Form IV

The crystallisation of form IV requires the addition of hydroxypropyl cellulose to a methanol solution, which upon slow evaporation will produce form IV<sup>[3, 10]</sup>. Pure form IV was never obtained in this research; therefore the theoretical XRPD trace from the CSD (ref code CBMZPN12<sup>[10]</sup>) has been used as the standard (Figure 2.6).

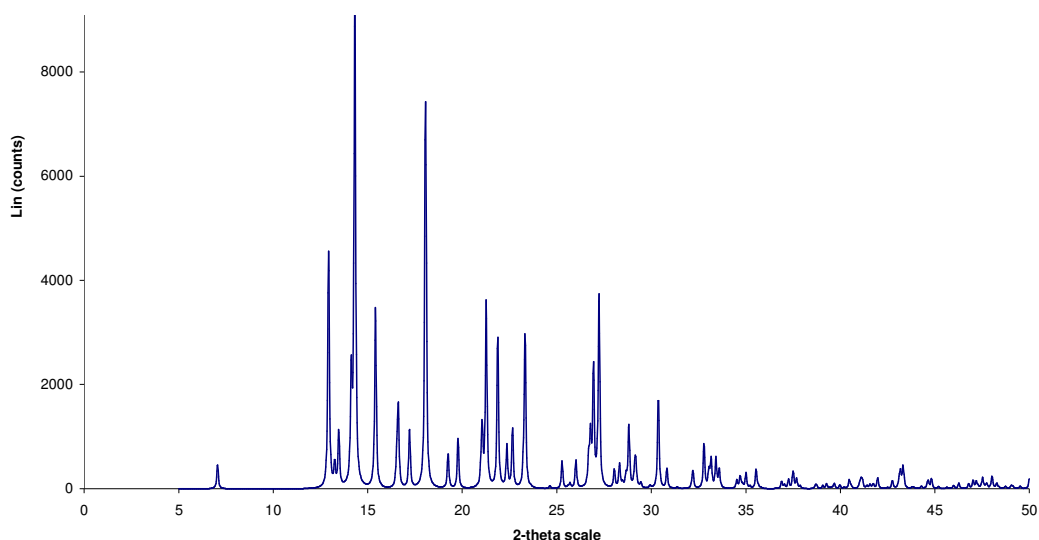


Figure 2.6 XRPD of carbamazepine form IV used as the standard diffraction pattern in this research. Taken from the theoretical powder pattern from CSD reference CBMZPN12

#### 2.3.5. Dihydrate

Evaporative crystallisation from water produced the CBZ dihydrate, with the XRPD trace in Figure 2.7.

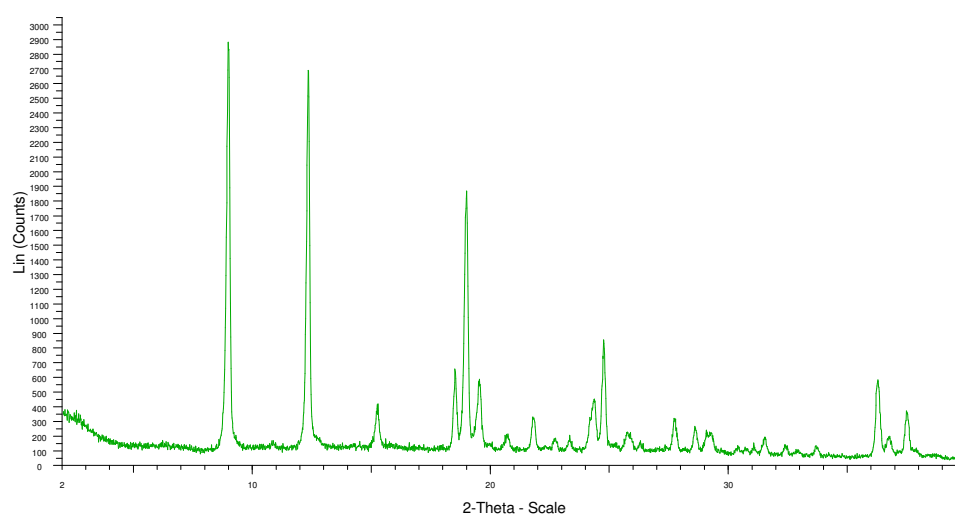


Figure 2.7 XRPD of carbamazepine dihydrate used as the standard diffraction pattern in this research



### 2.3.6. DMSO Solvate

Slow evaporation of a stoichiometric mixture of CBZ and DMSO will result in the DMSO solvate<sup>[11]</sup>. The standard XRPD trace used is shown in Figure 2.8.

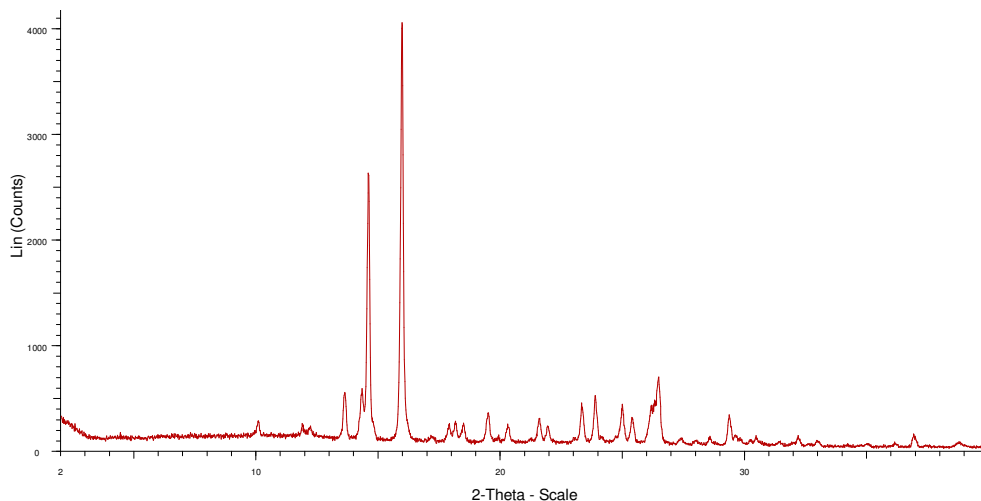


Figure 2.8 XRPD of carbamazepine DMSO solvate used as the standard diffraction pattern in this research

## 2.4. ROY Experimental Details

The thermodynamically stable form of ROY (Y) was received as a gift from Dr. C. Nicholson of The University of Durham. The standard XRPD traces for the Y and R forms are presented here (Figure 2.9 and Figure 2.10 respectively), as they were the only forms crystallised in this research. The XRPD traces of the other polymorphs of ROY may be found in Electronic Appendix, chapter 2, file 1.1.

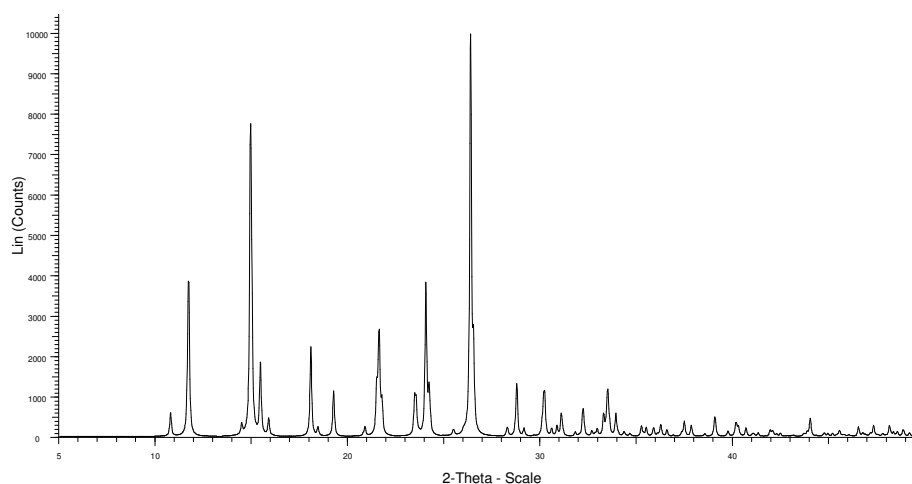


Figure 2.9 XRPD of ROY form Y used as the standard diffraction pattern in this research

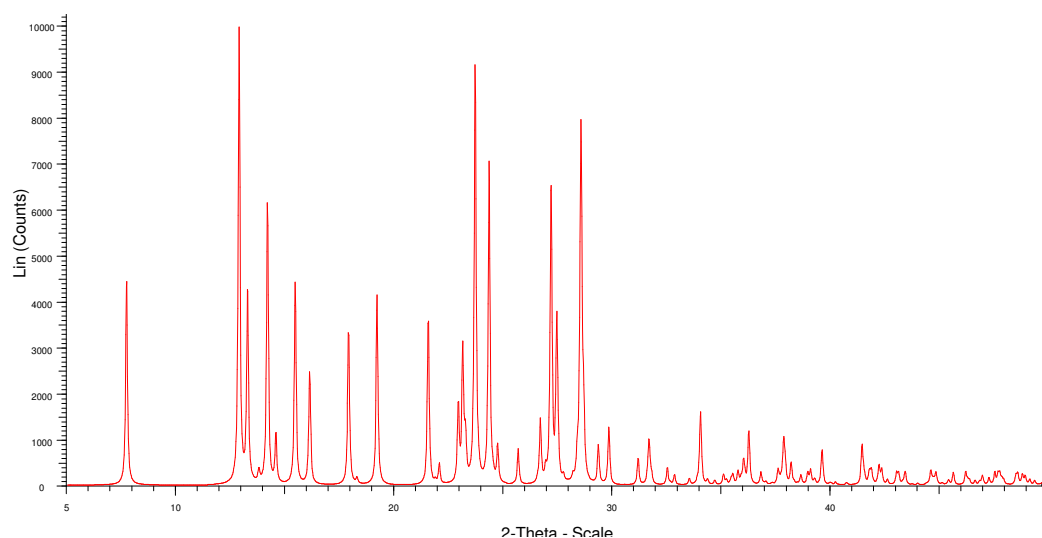


Figure 2.10 XRPD of ROY form R used as the standard diffraction pattern in this research

## 2.5. Tolbutamide Experimental Details

### 2.5.1. Form I

The most common method for the preparation of form I is using the method of Simmons et al.<sup>[12]</sup>, whereby 25 g of tolbutamide (TBA) was dissolved in 50 mL of benzene at 50°C. To this solution 25 mL of a 40°C hexane solution was added slowly, and crystallisation occurred at room temperature. The XRPD trace generated for this form can be seen in Figure 2.11.

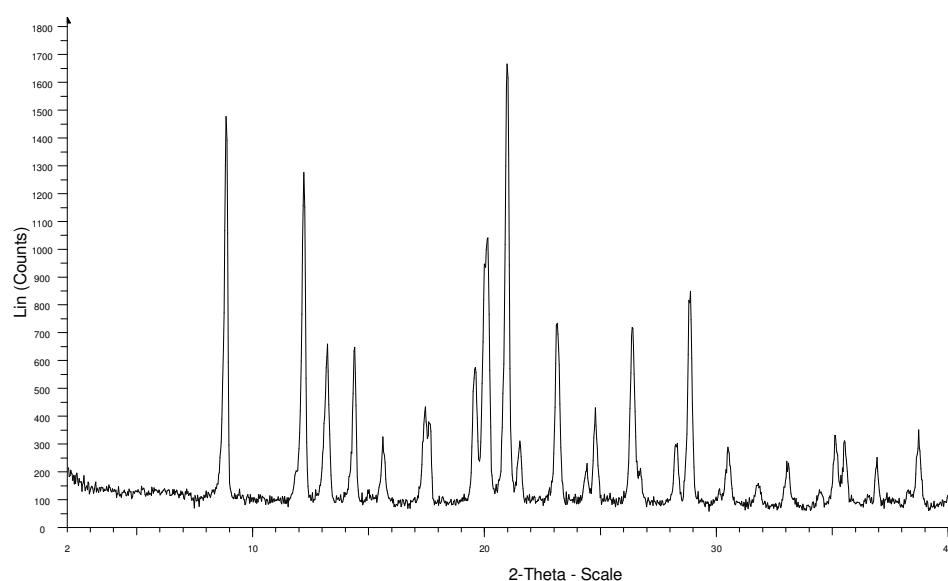


Figure 2.11 XRPD of TBA form I used as the standard diffraction pattern in this research

### 2.5.2. Form II

The method of preparing form II used by Burger<sup>[13]</sup> was to dissolve TBA in boiling carbon tetrachloride and slowly cool. Al-Saieq et al.<sup>[14]</sup> produced form II by slow evaporation from methanol or ethanol. The standard form II XRPD trace is shown in Figure 2.12.

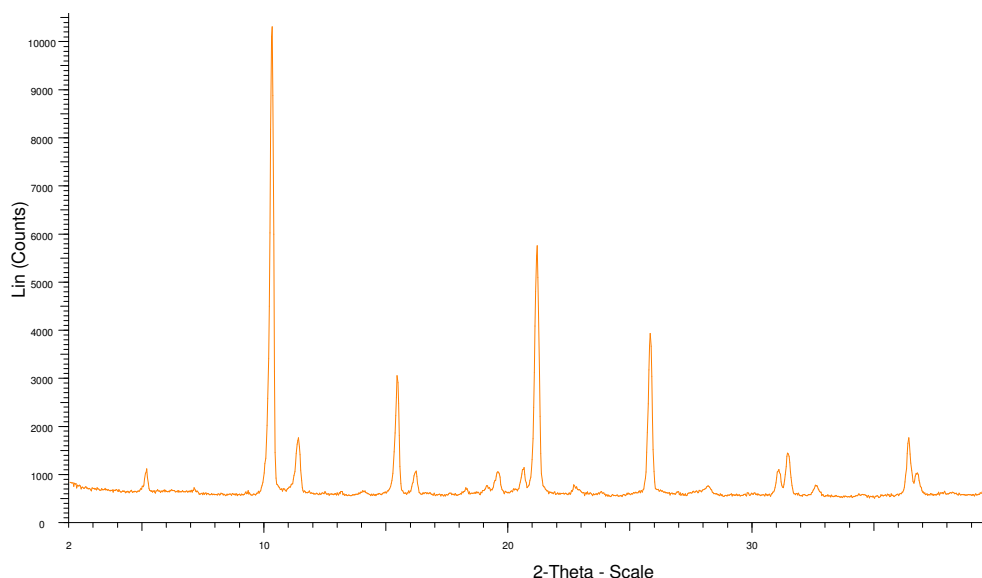


Figure 2.12 XRPD of TBA form II used as the standard diffraction pattern in this research

### 2.5.3. Form III

20 g of TBA was added to 20 mL of water and 40 mL of ethanol at a temperature of 40 °C, as presented in research by Simmons et al.<sup>[12]</sup>. This solution was then crystallised at room temperature. Thirunahari et al.<sup>[15]</sup> reported that this method generated a mixture of form I and III crystals and in order to produce pure form III they carried out the same method at RT and left the crystals to grow for one day. No pure form III was obtained in this research therefore the XRPD trace in the work by Thirunahari et al.<sup>[15]</sup> was used as the standard and is presented in Figure 2.13.

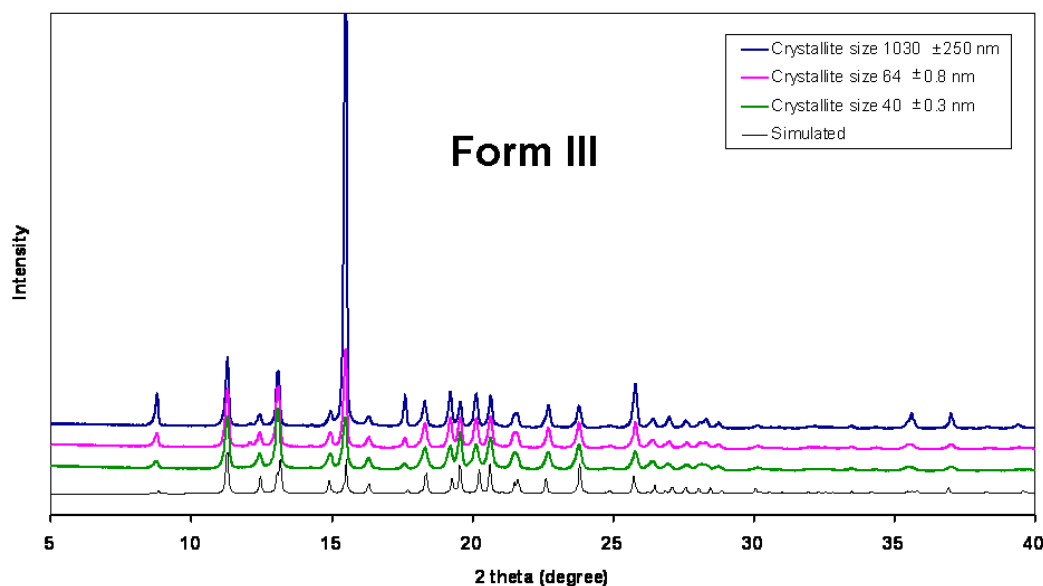


Figure 2.13 XRPD of TBA form III used as the standard diffraction pattern in this research, taken from the supporting information of Thirunahari et al.<sup>[15]</sup>

#### 2.5.4. Form IV

Sonoda et al.<sup>[16]</sup> crystallised form IV in the presence of 2,6-Di-*O*-methyl- $\beta$ -cyclodextrin that inhibits solution mediated transformation to form I. The crystallisation takes place in pH 8.0 sodium phosphate buffer. It was noted in Sonoda et al.<sup>[16]</sup> research that form IV crystallises initially in this buffer, even when the 2,6-Di-*O*-methyl- $\beta$ -cyclodextrin is not present. However, it quickly transforms to form I.<sup>[16]</sup> Thirunahari et al.<sup>[15]</sup> found that slow evaporation from an acetonitrile solution at room temperature also produced form IV. Pure form IV was never obtained in this work, therefore the XRPD trace from the research of Sonoda et al.<sup>[16]</sup> was used as a standard.

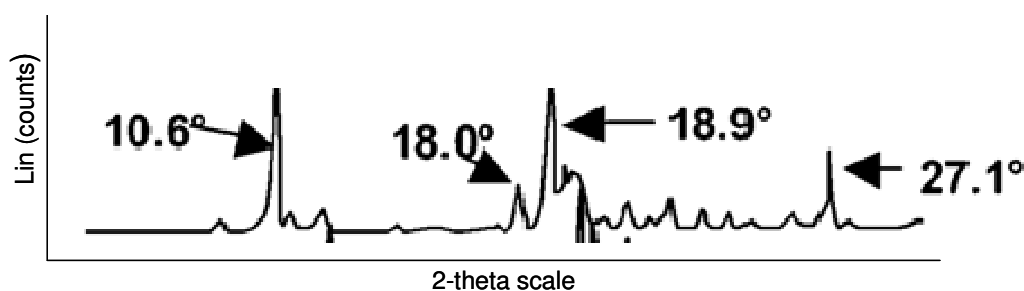


Figure 2.14 XRPD of TBA form IV used as the standard diffraction pattern in this research, adapted from Sonoda et al.<sup>[16]</sup>

### 2.5.5. Form V

Form V is a very new polymorph of tolbutamide discovered by Nath et al.<sup>[17]</sup> and was crystallised whilst searching for cocrystals. The addition of 1 mL of a cooled (-20°C) solution of 0.5 mL conc. HNO<sub>3</sub> and 10 mL methanol, to a tolbutamide solution (-20°C, 30 mg TBA in 10 mL methanol) that was then allowed to crystallise, formed this new polymorph. Form V has not been seen in this research, but for reference the XRPD trace has been presented in Figure 2.15.

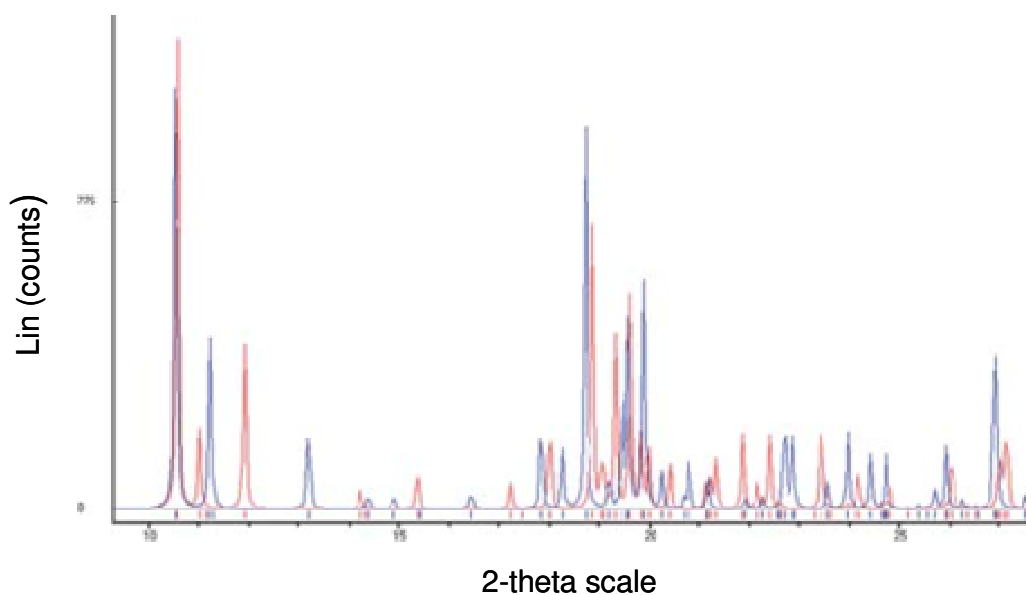


Figure 2.15 XRPD of TBA form V used as the standard diffraction pattern in this research, taken from Nath et al.<sup>[17]</sup>

## 2.6. Theoretical Calculations

The polymorphic and solvent molecules were modelled within solvent force fields, and from these optimised structures, molecular descriptors were calculated.

### 2.6.1. Molecular Modelling

Two different software packages were used in this research to compute molecular models; these were Hyperchem<sup>TM</sup><sup>[18]</sup> and Gaussian 03<sup>[19]</sup>. Hyperchem<sup>TM</sup><sup>[18]</sup> was used primarily for the early calculations with relatively low theory levels (OPLS and PM3). OPLS (Optimized Potentials for Liquid Simulations) is a molecular mechanics force field that represents the simplest form of molecular modelling and was initially

devised to model proteins and nucleic acids<sup>[20]</sup>. Molecular mechanics force fields allow the prediction of molecular geometry using previously derived data from related molecules.<sup>[21]</sup> The derived data used is a combination of different parameters and equations that define the energy and was demonstrated in early work using a collection of hydrocarbons to refine the model<sup>[22]</sup>.

PM3 (Parametric Method 3) is a semi-empirical calculation method that is an improvement upon the MNDO (Modified Neglect of Diatomic Overlap) and AM1 (Austin Model 1)<sup>[23]</sup> models. Semi-empirical methods consider the valence electrons within the system and incorporate core electrons by implementing a reduction in nuclear charge<sup>[24, 25]</sup>. These calculations are also less computationally expensive than higher level calculations because they are fitted to experimental results, and do not calculate every parameter used. The key assumption in semi-empirical methods is the zero-differential overlap (ZDO) that sets the overlap between pairs of different orbitals, commonly the s and p-orbitals to zero<sup>[24]</sup>.

B3LYP/6-31G\* is the basis set used in the high level calculations in Gaussian 03<sup>[19]</sup>. B3LYP stands for Becke's correlation exchange combined with Lee, Yang and Parr's correlation function that allows the full computation of the correlation energy of a system<sup>[26]</sup>.

6-31G\* is a split valence method that means there are 6 core orbitals considered, in which 3 are inner and 1 is an outer valence orbital, with a Gaussian function, and a single point geometry optimisation (\*).

To visualise all of the models, Molden<sup>[27]</sup> and Gaussview<sup>[28]</sup> were used.

## **2.6.2. Geometry Optimisation**

Geometry optimisation was used within this research to find the minimum energy structure that is thought to be the most stable molecular conformation<sup>[29]</sup>. A three-dimensional structure is presented that requires optimisation and its energy is calculated,  $V(0)$ . Once the starting energy is known, the atomic coordinates are moved and the energy recalculated until the lowest energy compared to the initial structure is found (depicted in Figure 2.16). The derivative of the energy with respect to the molecular coordinates ( $q_k$ ) is known as the gradient ( $g_k$ ).<sup>[29]</sup> At the minimum

point the gradient is equal to zero, and therefore optimisation occurs until this is achieved. This process is linearly represented by Equation 2.2<sup>[30]</sup>.

$$V(q_k) = V(0) + \sum_k g_k q_k \quad \text{Equation 2.2}$$

Within these types of calculation there can often be a number of minimum energy points. However, these are often a local minima and an extensive optimisation is required to obtain the global minimum (Figure 2.16).

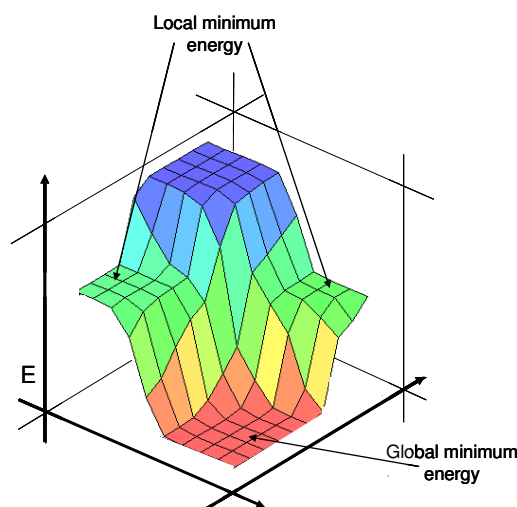


Figure 2.16 Simplified representation of a potential energy surface showing both global and local minimum

#### 2.6.2.1. Solvent Force Fields in DFT Calculations

In this research it is important to consider the interaction of the molecule with the solvents used in crystallisation. In Gaussian 03<sup>[19]</sup>, a number of pre-formed implicit solvent force fields are available, which have been utilised within this work. Solvent force fields have to be explicitly included in the Gaussian 03<sup>[19]</sup> calculations, with a polarisable continuum model (PCM) being used in this research. The PCM model creates a cavity in which the molecule is placed and exerts bulk effects (dielectric properties most importantly) of the solvent upon the molecule, slightly affecting its geometry<sup>[24, 31]</sup>.

The background to this approach is based on the solvation free energy (represented by Equation 2.3) which takes into account electrostatic components,  $\Delta G_{elec}$ , van der

Waals interactions of the solvent and solute,  $\Delta G_{vdw}$  and also the energy required to form the solute cavity,  $\Delta G_{cav}$ <sup>[24]</sup>.

$$\Delta G_{sol} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cav} \quad \text{Equation 2.3}$$

The electrostatic component is particularly important in charged or highly polar molecules, as the PCM method will use an average value for the dielectric constant<sup>[31]</sup>. This component is based on the Born model whereby a charge is placed into a solvent cavity and it is the work required to transfer an ion from the vacuum into the medium that is the electrostatic component<sup>[24]</sup>. This is shown in Equation 2.4 where  $q$  is the ions charge,  $a$  is the radius of the solvent cavity and  $\epsilon$  is the dielectric constant.

$$\Delta G_{elec} = -\frac{q^2}{2a} \left( 1 - \frac{1}{\epsilon} \right) \quad \text{Equation 2.4}$$

This is a relatively simplistic overview of the electrostatic component, whereas in quantum mechanical calculations the process is more complex. One of the key features within these advanced calculations is the shape of the solvent cavity. In PCM calculations the cavity is calculated using the van der Waals radii of the solute molecule, rather than a sphere<sup>[24, 32]</sup>. To improve the accuracy of these calculations the cavity surface is divided into many sections, each with an associated point charge. Polar coordinates for each atom are determined at the centre of their van der Waals sphere, therefore placing a charge at each point upon the surface of the cavity. This leads to more accurate determinations of the solute-solvent electrostatic interactions.

The non-electrostatic terms, van der Waals interaction component and the free energy of cavity formation, are also important in the free energy of solvation calculations, especially when the solvent is not highly charged or polar<sup>[24]</sup>. These terms can be combined in Equation 2.5, where  $\gamma$  and  $b$  are constants within this equation and  $A$  is the total solvent accessible surface area of the solute molecule.

$$\Delta G_{vdw} + \Delta G_{cav} = \gamma A + b \quad \text{Equation 2.5}$$



The free energy of cavity formation refers to the work required to create the solute cavity,<sup>[33]</sup> whilst acting against the solvent pressure and the entropy change due to the reorganisation of the solvent molecules around the molecule being analysed. The solvents most affected by the reorganisation are those closest to the solute molecule (within the first solvation shell), which is proportional to the solvent accessible surface area of the solute,  $A$ . The van der Waals interactions between solute and solvent are also affected by the number of solvent molecules within the first solvation shell, as these interactions are heavily affected by the distance between the molecules of interest. These factors highlight why both terms are proportional to the solvent accessible surface area of the solute. The constants  $\gamma$  and  $b$  are often experimentally determined, with  $\gamma$  having the value 7.2 cal/(molÅ<sup>2</sup>) and  $b$  often set as zero<sup>[24]</sup>.

Although Gaussian 03<sup>[19]</sup> has many predefined solvent force fields, it is also possible for the user to define force fields of solvents that are not represented. Four different variables are required for the created of a solvent force field; these are the static dielectric constant (EPS), the dielectric constant at infinite frequency (EPSINF), the solvent radius (RSOLV) and the density (DENSITY) of the solvent. The dielectric constants are accessible within the literature and the other values can be calculated simply.

EPSINF ( $\epsilon_\infty$ ) can be calculated by using the refractive index of the solvent ( $n_d$ ), calculated using the relative permittivity ( $\epsilon_r$ ) and permeability ( $\mu_r$ ) of the material (Equation 2.6).  $\mu_r$  is often close to 1, which therefore means  $n_d$  approximately equals  $\sqrt{\epsilon_r}$ , leading to Equation 2.7<sup>[34]</sup>.

$$n_d = \sqrt{\epsilon_r \mu_r} \quad \text{Equation 2.6}$$

$$\epsilon_\infty = (n_d)^2 \quad \text{Equation 2.7}$$

RSOLV is a value that relates the molar volume( $\bar{V}$ ) to the radius of the solvent and is based on research by Stearn and Eyring<sup>[35]</sup> (Equation 2.8).  $\bar{V}$  can be calculated (Equation 2.9) using the molecular weight ( $MW$ ) and density ( $\rho$ , g mL<sup>-1</sup>) of the solvent and Avogadro's constant ( $N_A$ ).

$$RSOLV = \left( \frac{\bar{V}}{8N_A} \right)^{\frac{1}{3}} \times 10^{-10} \quad \text{Equation 2.8}$$

$$RSOLV = \left( \frac{MW}{8N_A \rho} \right)^{\frac{1}{3}} \times 10^8 \quad \text{Equation 2.9}$$

DENSITY values are available in the literature for common solvents, but Equation 2.10 needs to be used in order to calculate the values in terms of Å<sup>-3</sup>.

$$DENSITY = \left( \frac{\rho N_A}{MW \times 10^{24}} \right) \quad \text{Equation 2.10}$$

### 2.6.3. Modelling Software Used

Hyperchem<sup>TM</sup><sup>[18]</sup> version 8 Student edition by Hypercube Ltd. was used in this research to model the molecules, geometry optimise and perform conformational searches at low levels of theory. This was followed by higher level calculations in Gaussian 03, Revision D.01<sup>[19]</sup>.

### 2.6.4. Molecular Representation

In order to calculate molecular descriptors a numerical representation of the molecular structure needs to be created. In some software, such as Hyperchem<sup>TM</sup><sup>[18]</sup> a molecule can be drawn by hand and from this, based on average bond lengths and angles, the molecular structure is optimised. Cartesian coordinates were created in Hyperchem<sup>TM</sup><sup>[18]</sup> that were then used as the starting geometry for Gaussian 03<sup>[19]</sup> optimisations.

### 2.6.5. Molecular Descriptors

Molecular descriptors are a numerical representation of molecular properties. In this research both the polymorphic molecule and the solvents used in crystallisation have been modelled to allow descriptors to be calculated. Two different pieces of software have been used in these calculations (MOE<sup>[36]</sup> and software from the book Molecular Descriptors in QSAR/QSPR<sup>[37]</sup>) to allow a large range of descriptors to be generated. There are two types of descriptor used in this research, these are *empirical* and *theoretical*. Within these types there are different classes of descriptor, for

example structural descriptors which are mainly concerned with intramolecular interactions and solvational descriptors which explain intermolecular interactions in solution<sup>[37]</sup>. Within the *theoretical* class of descriptors there are constitutional, topological and geometrical descriptors which are the most simple. These describe the atoms, bonding and shape of the molecule, whereas there are the more complex charge-distribution, molecular-orbital and thermodynamic descriptors<sup>[37]</sup>. *Empirical* descriptors can be generated experimentally, which in many cases is impractical due to the molecule or time and cost of experiments. *Theoretical* descriptors on the other hand are mathematical representations, based on fundamental physical equations.

Details of all the molecular descriptors used in this research can be found in Appendix section 12.2.

#### **2.6.6. Molecular Descriptor Software Used**

Molecular descriptors were calculated using software from a book entitled, Molecular descriptors in QSAR/QSPR<sup>[37]</sup> and also by commercially available software, MOE<sup>[38]</sup> (molecular operating environment) from the Chemical Computing Group.

### **2.7. Artificial Neural networks**

Artificial Neural Networks (ANN) have found relatively new applications within the chemical sciences<sup>[39, 40]</sup>, whereas previously they have been more heavily used in other fields such as biological sciences<sup>[41-44]</sup>, economics<sup>[45]</sup> and physics<sup>[46]</sup>, highlighting their diversity as a analytical tool. However, ANNs are now becoming more common in the chemical literature with various applications, from predicting physicochemical values<sup>[47-53]</sup> to determining chemical shifts in <sup>13</sup>C NMR data<sup>[54]</sup>. Not only can ANNs predict outputs from a given set of numerical data, but they can also be applied to imaging. A good example of this application is seen in work that monitors receptor binding with UV/Vis spectroscopy<sup>[55]</sup>. The network monitors specific wavelengths and creates an output based on this. Earlier examples use an ANN to predict structure-activity relationships of carboquinone derivatives<sup>[56]</sup>, and

others use modelled molecular descriptors to predict aqueous solubility of organic molecules<sup>[49-52]</sup>.

ANNs are based on the human brain and the concept of learning through experience. As a child you learn to read and write and this is through looking at letters and remembering shapes and patterns. ANNs do something similar to the brain but in the case of this research, with numbers. Figure 2.17 shows a simplified diagram of the human neuron and it is this architecture that is the foundations of ANNs.

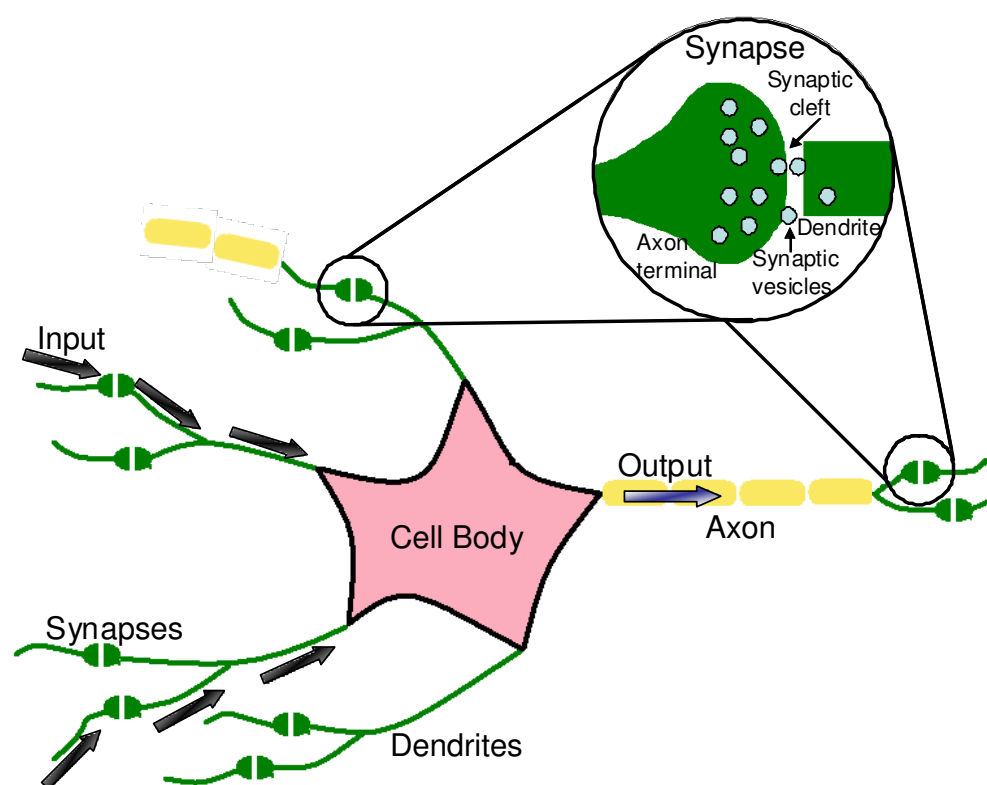


Figure 2.17 A simplified biological neuron

In human brains there are billions of neurons or nerve cells that relay impulses of information to other neurons within the body by the transportation of ions across synapses. The dendrites of one neuron provide the cell body with chemical signals that can act to either excite or inhibit<sup>[57]</sup> the action potential. Within the cell body there is a potential threshold. If the impulses received by the dendrites are above the threshold, an impulse or action potential is passed along the axon (of which there is only one per neuron). As shown in Figure 2.17 the axon can branch at the end and can transfer the chemical impulses to the dendrites of other neurons. The area at

which these impulses are passed from the axon of one neuron to the dendrites of another is called a synapse, which is diagrammatically represented at the top of Figure 2.17. Across a synapse there is a potential difference and as an impulse is passed along the axon, if the impulse is great enough it will be passed through the synapse by neurotransmitters. This impulse travels into the dendrite of the next neuron, creating a postsynaptic potential. However, if the impulse can not overcome the potential difference of the synapse, the signal is passed no further, demonstrating the “all-or-nothing”<sup>[57, 58]</sup> impact of these impulses. It is important to note that the synapses only operate in one direction, and therefore the signals can not pass back through the axon to the cell body<sup>[59]</sup>.

The artificial neuron is based on the human neuron, with the threshold logic unit (TLU) being one of the earliest examples<sup>[60]</sup>. Early work by McCulloch and Pitts<sup>[60]</sup> is the basis of modern artificial neural networks, with further advances being made by Hopfield<sup>[58, 61]</sup> some years later, introducing the idea of nonlinearity between the input and output values<sup>[62, 63]</sup>.

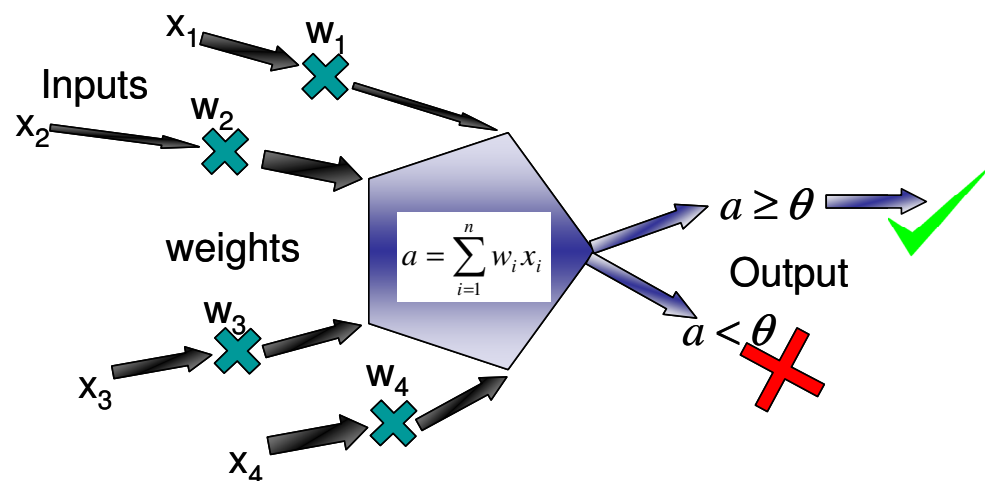


Figure 2.18 Representation of a Threshold Logic Unit

Figure 2.18Figure 2.18 represents a TLU, where  $x$  is an input, resembling the chemical impulse passed through the dendrites in the human neuron. The size of  $x$  can vary; for example your input data is melting points so they are mainly different values. The differences visible in this diagram compared to the human neuron are the weights ( $w$ ) which are similar to the synapses in the biological neuron. All inputs are

multiplied by the corresponding weights, with the results of each input line being summed. Similarly to the action potential in the human neuron it is only when a specific threshold ( $\theta$ ) is reached that an impulse is fired, in this case,  $y$ . TLUs are binary systems and therefore if the threshold is reached,  $y = 1$ , but if the value is below, the output is 0 (Equation 2.11)<sup>[57]</sup>.

$$a = \sum_{i=1}^n w_i x_i$$

Equation 2.11

$$y = \begin{cases} 1 & \text{if } a \geq \theta \\ 0 & \text{if } a < \theta \end{cases}$$

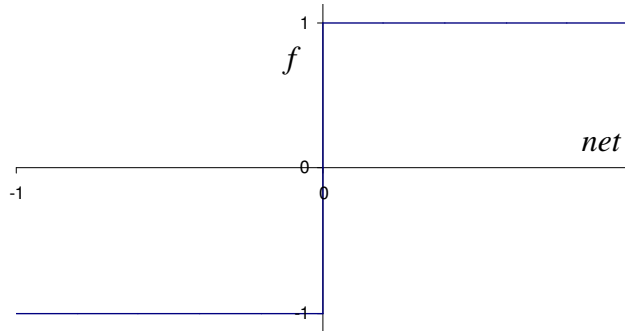


Figure 2.19 graphical representation of a threshold function

In an ANN the weights can be modified either manually or by the network whilst training. This is done to reduce the error between the predicted and desired output. Both the inputs and outputs are given to the network, allowing the adjustment of the weights automatically during training using an algorithm<sup>[64]</sup>. Once the error meets the convergence criterion the training is complete.

In the simplest example of a TLU with only two inputs, a linear decision surface is created to allow the classification of the networks outputs. For example in Table 2.2, weights have been set to 1 and therefore the summation of the inputs determines whether the output is 1 or 0. In this example the threshold is set to 1.75, therefore all inputs that add up to less than this value will show an output of 0, highlighted in the plot (Figure 2.20).

Table 2.2 TLU example inputs

$X_1$	$X_2$	Sum of inputs	Output
0	0	0	0
0	1	1	0
1	1	2	1
1	0	1	0
0	1	1	0
1	1	2	1

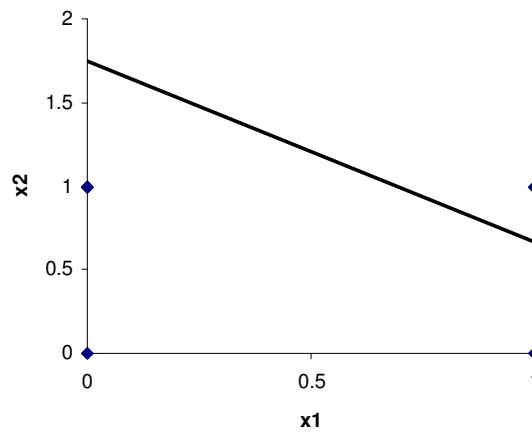


Figure 2.20 the pattern space for the two input TLU showing the threshold which determines whether the output is 1 or 0

Clearly this is a very simplified case but it highlights the idea of a decision surface. When the network becomes complicated it is more favourable to represent the values in terms of vectors, of which there are two of interest, the *weight* and *input*. Also, as one moves towards more complex systems, the decision surface will no longer be linear in 2D space<sup>[64]</sup>. With this, having such a simple threshold characterisation method is not practical and there is a move towards nonlinear functions.

### 2.7.1. Multilayer Perceptron (MLP)

A perceptron “is an enhancement of the TLU”<sup>[57]</sup> with a single layer and is concerned with a non-linear neuron (unlike the TLU). External bias can be placed on each input, and the learning is iterative rather than continuous<sup>[65]</sup>.

Multilayer perceptrons (MLPs) are a combination of a number of perceptrons often trained using a back-propagation algorithm. Hidden layers are not input or output

layers (see Figure 2.21) but are commonly used in complex networks as they have the ability to extract the relevant information from the inputs that can train the network. They consist of different nodes that are connected and pass information between them<sup>[66]</sup>, leading to the development of complex relationships. At every node there is a summation of information as every connection is weighted.

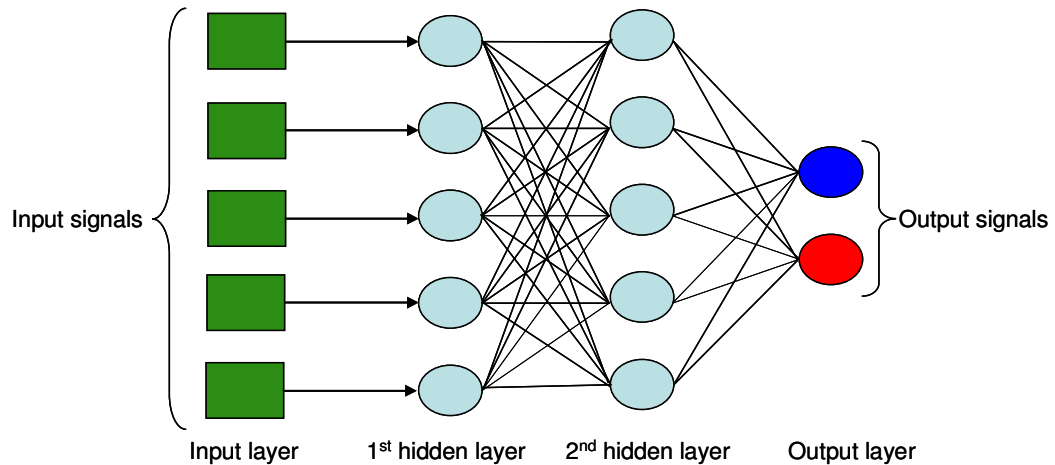


Figure 2.21 Simplified two hidden layer neural network, with light blue shapes representing the nodes

The number of hidden layers to use within a network is an area of much research, with a number of guidelines in the literature<sup>[62, 67-69]</sup>. If there are too few, the learning process may be hindered, too many and the network may overtrain and not generalise well<sup>[62]</sup>. Bourquin et al.<sup>[62]</sup> and Plumb et al.<sup>[69]</sup> both state that using Kolmogorov's theorem, which suggests twice the number of inputs plus one, would be a good starting value.

The INForm<sup>[70]</sup> software automatically detects the number of hidden layers required for the dataset and implements that value. In future, a full interrogation of the effects of the number of hidden layers on the data should be attempted.

There are basically two steps in the training process of a MLP; initially the inputs are propagated through the network, layer by layer to produce an output<sup>[65]</sup>. If this output does not correspond to the actual output given to the network, then the information is fed back through the network to propagate through once again. When the process is restarted the weights are altered automatically by the network to try and reduce the error on the output, which is known as backpropagation.



There are two key features to the backpropagation algorithm, the learning rate ( $\eta$ ) and the momentum ( $\mu$ ). An often used method in backpropagation is the delta-rule, which states that a change in any of the parameters should be proportional to the input and the output layer error<sup>[59]</sup>. The proportionality constant used in this rule is  $\eta$ , the learning rate, which controls the average size of the change in the weights. The momentum term decreases training times in regions of constant error values<sup>[71]</sup>, allowing escape from local minima and prevents sudden changes in the direction of the weight values<sup>[59]</sup>.

$$\Delta w_{ji}^l = \eta \delta_j^l out_j^{l-1} + \mu \Delta w_{ji}^{l(previous)} \quad \text{Equation 2.12}$$

Equation 2.12 shows the calculation of the weight to be used in the next iteration based on the previous change in weight ( $\Delta w_{ji}^{l(previous)}$ ), the network output ( $out_j^{l-1}$ ) and the learning rule and momentum terms. The  $l$  term signifies the layer,  $i$  the input source and  $j$  the current neuron with  $\delta_j^l$  being the error in that neuron<sup>[59]</sup>.

The error term is treated differently if it is within a hidden layer (Equation 2.13) or in the output layer (Equation 2.14).

$$\delta_j^l = \left( \sum_{k=1}^r \delta_k^{l+1} w_{kj}^{l+1} \right) out_j^l (1 - out_j^l) \quad \text{Equation 2.13}$$

$$\delta_j^{last} = (y_j - out_j^{last}) out_j^{last} (1 - out_j^{last}) \quad \text{Equation 2.14}$$

If the two equations shown in Equation 2.13 and Equation 2.14 are substituted into Equation 2.12 the full equation for weight correction in a hidden layer can be presented (Equation 2.15)<sup>[59]</sup>.

$$\Delta w_{ji}^l = \eta \left( \sum_{k=1}^r \delta_k^{l+1} w_{kj}^{l+1} \right) out_j^l (1 - out_j^l) out_j^{l-1} + \mu \Delta w_{ji}^{l(previous)} \quad \text{Equation 2.15}$$

This equation highlights that information from three different layers is used in order to correct the weights<sup>[59]</sup>, the previous, current and next weight. Based on this information it is therefore very difficult to gather information about the inputs influence on prediction based on the weight values<sup>[59]</sup>.

### 2.7.2. Transfer Functions

As depicted in Figure 2.18, a threshold function is one transfer function method of determining the output of the ANN. However, this function may be too simplistic in many cases and therefore by using a nonlinear transfer function, a range of outputs can be determined rather than either 0 or 1. In this research a sigmoidal nonlinearity transfer function is utilised and is defined by Equation 2.16 and graphically represented in Figure 2.22, where the shape of the sigmoid is controlled by  $\rho$ <sup>[66]</sup>.

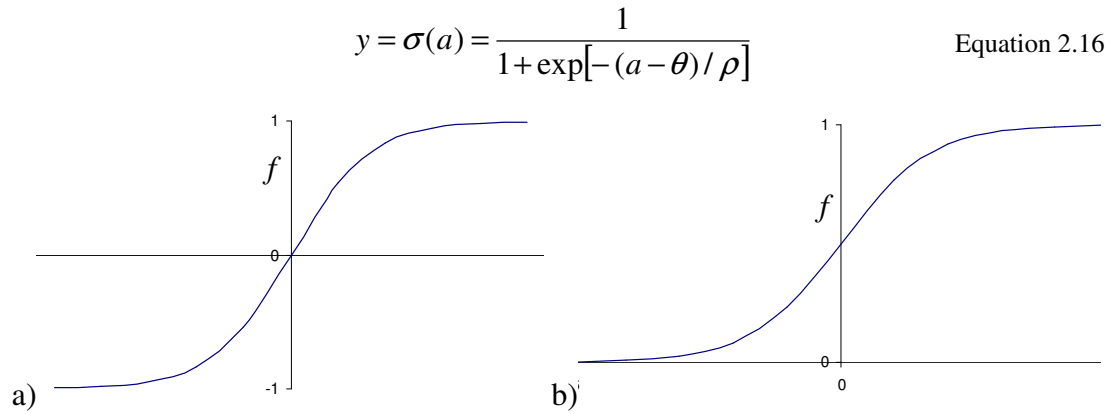


Figure 2.22 a) hyperbolic tangent sigmoid, tanh and b) logistic nonlinear functions

### 2.7.3. Artificial Neural Network Software Used

Two different pieces of software were used in this research, these were INForm<sup>[70]</sup> version 3.7, an Intelligensys Ltd. Product and also Neurosolutions<sup>TM</sup><sup>[72]</sup> version 5 from NeuroDimensions.

## 2.8. Neurofuzzy logic

Fuzzy logic is an alternative method to the traditional logic methods, which have binary outputs, utilising intermediate positions for its output<sup>[73, 74]</sup>. In an example taken from the FormRules manual<sup>[73]</sup>, temperature control is an ideal way to explain the difference between what is now known as ‘crisp’ logic and fuzzy logic. If ideal room temperature is 20°C you would not say that 19°C or 21°C was too cold or hot respectively. Crisp logic would draw these conclusions, whereas fuzzy logic allows intermediate values such as cool or warm to be determined. A thermostat would be constantly turning on and off if ‘crisp’ logic was used to control room temperature,

whereas when fuzzy logic is used, the thermostat would have predetermined rules to decide if a little heating was required to keep the room temperature stable, or if it should be left off to cool down.

This is a very simplistic view of fuzzy logic, but, it can be very useful if you know the rules to generate outputs you desire i.e. temperature control. When the system has no predefined rules, neurofuzzy logic needs to be employed to understand the “cause and effect relationships”<sup>[73]</sup> of the input data and the desired output. Neurofuzzy logic uses the learning abilities of an ANN in combination with the linguistic output of fuzzy logic to generate “IF, THEN” rules for a problem<sup>[73]</sup>. In FormRules<sup>[75]</sup> the network architecture used in the training is predefined in the software as an associative memory network<sup>[73]</sup>. This network structure differs from a MLP network, which has been used in this work also, by having different types of nodes. Figure 2.23 shows the simplified structure of a neurofuzzy system taken from Shao et al.<sup>[76]</sup>. FormRules<sup>[75]</sup> employs the ASMOD (Adaptive Spline Modelling of Data) algorithm whereby multiple training models are generated and tested within the software to see which matches the data most closely<sup>[76]</sup>.

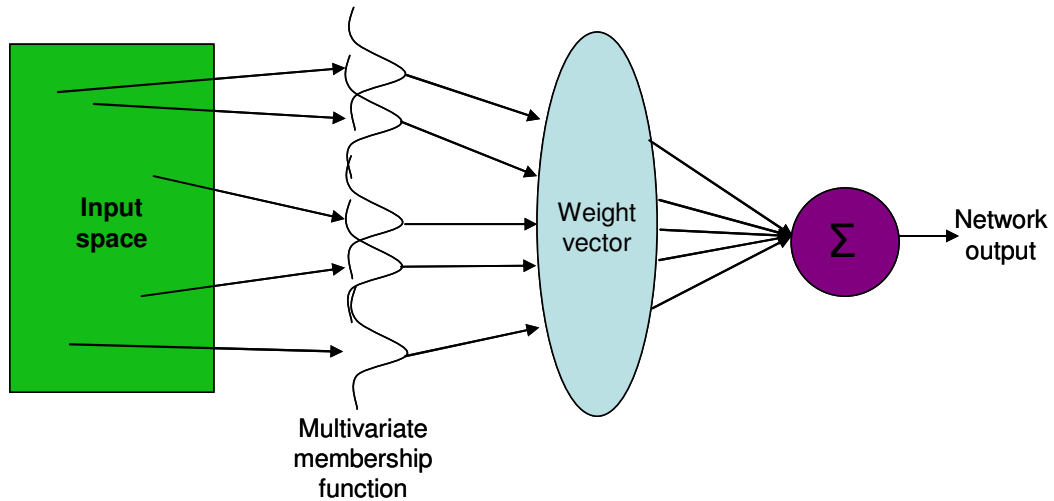


Figure 2.23 A schematic of a simple neurofuzzy system, adapted from Shao et al<sup>[76]</sup>.

The transfer functions shown here are the Gaussian type, there are also triangular functions and on-off type functions,<sup>[73]</sup> which are automatically assigned to the data by the software. These functions determine where the high and low boundaries in the data are and are used in the output of the rules.

The result of using neurofuzzy logic is a set of simple rules based on the input data that highlight if a certain input variable is set at a particular level, what the output will be. In this research, the descriptors have been input and FormRules<sup>[75]</sup> determines the rules that lead to the different polymorphic outcomes. Structure Risk Management (SRM) is the training method used in FormRules<sup>[75]</sup>. SRM uses the bias (number of free parameters) and variance (training data error) to look at the prediction error<sup>[73]</sup>, and is the most effective method for data sets of this size.<sup>[73]</sup>

Table 2.3 Example of the rules generated by FormRules<sup>[75]</sup> using neurofuzzy logic (please note, this is not a real result and has only been used for illustration purposes)

Rule	Output
IF rate is LOW	THEN Form III is HIGH (1.00)
IF rate is MID	THEN Form III is LOW (1.00)
IF rate is HIGH	THEN Form III is LOW (1.00)

#### 2.8.4. Fuzzy Logic Software Used

FormRules<sup>[75]</sup> version 3.3, an Intelligensys Ltd. product has been used in this result to produce fuzzy logic rules.

### 2.9. Principal Component Analysis

Principal component analysis (PCA) has been applied to numerous different scientific problems,<sup>[77-81]</sup> but most importantly in this research as a data reduction technique. PCA is a multivariate data analysis method that aims to reduce the dimensionality of a dataset. It constructs linear combinations of the variables to account for as much of the total variation in the whole dataset as possible<sup>[82]</sup>. Therefore reducing the amount of variables, but still containing all of the information presented within the dataset.

A square correlation matrix,  $\mathbf{R}$ , of  $n$  variables ( $X_1 \dots X_n$ ) is generated based upon all of the correlations in the data. Entry  $R_{ij}$  in the matrix represents the correlation of the variables ( $X_i, X_j$ ). Equation 2.17 shows the covariance in the variables, where  $E$  is the expectation function<sup>[64]</sup> and  $\mu$  represents the expected value of  $X_i$ , also equal to  $E(X_i)$ .

Equation 2.18 shows the correlation between the variables  $(X_i, X_j)$ , with  $\sigma$  representing the standard deviation of the corresponding variables.<sup>[64, 83]</sup>

$$\text{covariance}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \quad \text{Equation 2.17}$$

$$R_{ij} = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_{X_i} \sigma_{X_j}} \quad \text{Equation 2.18}$$

Based on these equations, the square correlation matrix is a combination of the diagonal matrix of eigenvalues presented as **L**, and a matrix of eigenvectors, with one vector per eigenvalue, **V**, along with its transpose **V'**, shown in Equation 2.19

$$\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}' \quad \text{Equation 2.19}$$

The square root of the eigenvalues are taken, leading to Equation 2.20.

$$\begin{aligned} \mathbf{R} &= \mathbf{V}\sqrt{\mathbf{L}}\sqrt{\mathbf{L}}\mathbf{V}' \\ \mathbf{R} &= (\mathbf{V}\sqrt{\mathbf{L}})(\sqrt{\mathbf{L}}\mathbf{V}') \end{aligned} \quad \text{Equation 2.20}$$

If  $\mathbf{V}\sqrt{\mathbf{L}}$  is represented by **A** and  $\sqrt{\mathbf{L}}\mathbf{V}'$  by **A'** then the equation can be further simplified to that in Equation 2.21<sup>[84]</sup>, with **A** representing the factor loading matrix.

$$\mathbf{R} = \mathbf{A}\mathbf{A}' \quad \text{Equation 2.21}$$

The factor loading matrix is a very useful tool as it shows the correlations between the variable and component of which it is a part. When a variable has a highly positive or large negative value, it signifies the strongest correlation with that component.

As each component is a representation of single linear combination identified within the data, with the first component ( $\text{PC}_{(1)}$ ) contains the largest variation and subsequent components contain less. For every input there is an associated score values for each component. This score value incorporates all of the information within the variables for that component. This score is generated by multiplying the loading matrix values (**A**) with the inversed correlation matrix variables (**R**<sup>-1</sup>) to generate a component score correlation matrix (**B**). This value (**B**) is then multiplied

by the variable matrix (**Z**) in order to give the component score matrix (**F**)<sup>[84]</sup>, shown in Equation 2.22 and Equation 2.23.

$$\mathbf{B} = \mathbf{R}^{-1} \mathbf{A} \quad \text{Equation 2.22}$$

$$\mathbf{F} = \mathbf{ZB} \quad \text{Equation 2.23}$$

When the second component is calculated, it is uncorrelated to  $PC_{(1)}$  and contains the maximum variation in the remaining data<sup>[84]</sup>. Equation 2.24 and Equation 2.25 show the linear equations for  $PC_{(l)}$  and the other components ( $PC_{(m)}$ ), with  $w$  representing the weighting placed on each variable and  $X$  representing the variable<sup>[82]</sup>.

$$PC_{(1)} = w_{(1)1}X_1 + w_{(1)2}X_2 + \dots + w_{(1)x}X_x \quad \text{Equation 2.24}$$

$$PC_{(m)} = w_{(m)1}X_1 + w_{(m)2}X_2 + \dots + w_{(m)x}X_x \quad \text{Equation 2.25}$$

Determining the number of principal components to extract is an importation feature of PCA, with it being possible to extract the same number of components as variables. The goal of PCA within this research was to be a data reduction method, therefore the number of components needed to be selected. One of the most straight forward ways to do this is by looking at the amount of variance of the data each component represents. When the variance is small, no additional information about the dataset is being communicated. Scree plots may be utilised to identify when most of variance in the dataset has been represented. The eigenvalues for each component are plot against the components number, thus illustrating the amount of variance contained within each PC. By drawing a straight line through the points from the highest PC, when the eigenvalues deviate from the line, this is taken to be the point when the components should be retained<sup>[82]</sup>. In the example in Figure 2.24 the eigenvalues deviate from the line at  $PC_{(3)}$  and therefore the first three components contain most of the variation in the dataset.

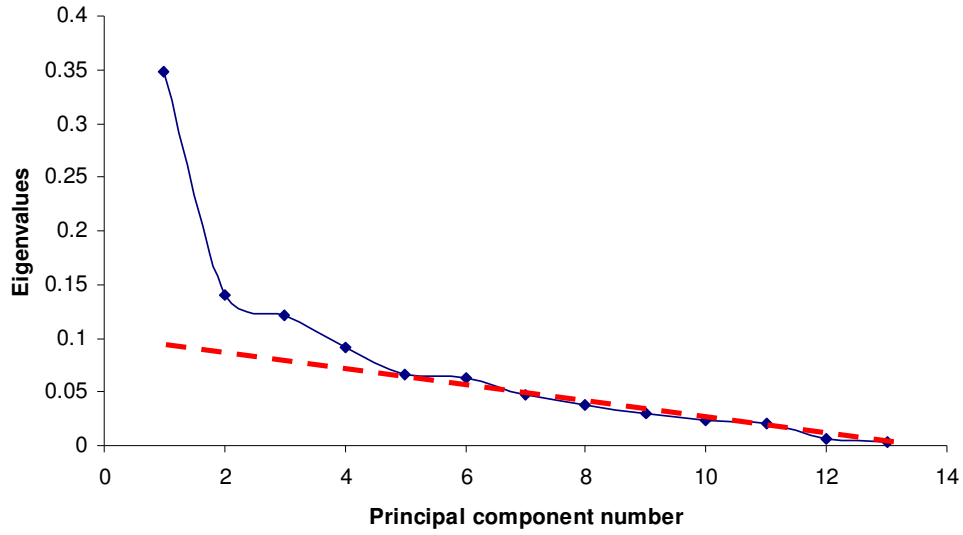


Figure 2.24 An example of a scree plot

PCA is a highly effective data reduction method that not only allows components that represent all of the information within the dataset to be produced, but also allow the selection of the most influential variables in each of the PCs. This was highly valuable within this research as it is the specific information that individual variables (or descriptors) contain that is of most interest.

## 2.10. Partial Least Squares Analysis

Partial least squares (PLS) analysis, like PCA can be used as a data reduction method and has been utilised within this research to highlight important descriptors. The overall aim of this technique is to determine independent linear correlations within the data and highlight the most important variables<sup>[85]</sup>, an approach that has been used in other scientific research<sup>[85-90]</sup>.

A matrix ( $\mathbf{X}$ ) of  $N$  experiments with  $J$  variables and matrix ( $\mathbf{Y}$ ) of  $N$  experiments with  $M$  outputs are produced and components ( $a$ ) are created.  $(x_{i1}, x_{i2}, \dots, x_{ij})$ ,  $(y_{i1}, y_{i2}, \dots, y_{im})$ ,  $(w_{1a}, w_{2a}, \dots, w_{ja})$  and  $(c_{1a}, c_{2a}, \dots, c_{ma})$  can be represented by the vectors  $\mathbf{x}_i$ ,  $\mathbf{y}_i$ ,  $\mathbf{w}_a$  and  $\mathbf{c}_a$  respectively for input, output, input weighting and output weighting values<sup>[86]</sup>. With the aim of PLS being to determine linear correlations ( $t_a$ ) within the input data and to generate fewer correlations than variables ( $J$ ), where  $a = (1, \dots, A)$ , Equation 2.26 shows the construction of these linear combinations for input and output ( $u_{ia}$ ) variables.

$$\begin{aligned} t_{ia} &= w_{1a}x_{i1} + w_{2a}x_{i2} + \dots + w_{ja}x_{ij} = \mathbf{w}'_a \mathbf{x}_i \\ u_{ia} &= c_{1a}y_{i1} + c_{2a}y_{i2} + \dots + c_{ma}y_{im} = \mathbf{c}'_a \mathbf{y}_i \end{aligned} \quad \text{Equation 2.26}$$

The weight values placed upon each variable are there to maximise the covariance between the input and output variables, and can highlight a variables influence upon the components ( $a$ ), with larger values having the most effect.

Determining the number of components to calculate presents a similar problem to that seen in the PCA analysis, and like PCA the first component contains the most information about the dataset, with subsequent components containing less.

For every component there is a loading vector ( $\mathbf{p}_a$ ) which is equal to  $(p_{1a}, p_{2a}, \dots, p_{ja})$ . These loadings contain the regression equations for the columns of matrix  $\mathbf{X}$  and the linear correlations,  $(t_a)^{[86]}$  and are useful to determine which variable has the most influence on a component. Coefficient  $b_{mj}$  takes into account the correlations between input variables, producing better predictions, and uses a modified weight term,  $w_{ja}^*$ , which improves the prediction further (Equation 2.27).

$$\begin{aligned} y_{im} &= b_{mj}x_{ij} \\ b_{mj} &= \sum_{a=1}^A c_{ma} w_{ja}^* \end{aligned} \quad \text{Equation 2.27}$$

The importance of each variable for the projection (VIP) in PLS analysis can be calculated, allowing the removal of those that do not provide additional information<sup>[86]</sup>. Each variable ( $v_j$ ) is given a number between 0 and 1, with higher values indicating importance within the dataset. The variable importance values can be calculated based on the regression coefficient ( $b_{mj}$ ), the weights ( $w_{ja}$ ), the modified weights ( $w_{ja}^*$ ), the weight and fraction of variance in  $\mathbf{Y}$ , and the loadings  $(p_{ja})^{[86]}$ . In this research the VIP values were calculated based upon the weights, using Equation 2.28.

$$v_j^{(r)} = \frac{\sum_{a=1}^A |w_{ja}| R_{Ya}^2}{\max_{j \in J} \left( \sum_{a=1}^A |w_{ja}| R_{Ya}^2 \right)} \quad j = 1, \dots, J \quad \text{Equation 2.28}$$



$|w_{ja}|$  represents the absolute value of the weight on variable  $j$  and  $R_{Y_a}^2$  is the fraction of variance explained by the component with regards to the output  $\mathbf{Y}$ <sup>[86]</sup>.

### **2.10.1. Chemometric data analysis software**

The PCA and PLS analysis was conducted in collaboration with Dr. O. Svensson at AstraZeneca. All chemometric analysis was carried out using SIMCA-P+<sup>[91]</sup>, version 12.01, a Umetrics AB product. The number of components used in the PCA analysis was set to the number of solvents used in the experimental work. The number of components used in PLS analysis are variable and dependent on the output. Initially two components were created, and further components added until the percentage variance of the data was high in order to retain most of the information from the descriptors. In the CBZ analysis only form II and III were predicted with seven and six components respectively. PLS analysis was not carried out for the ROY and TBA data.

- [1] S. R. Byrn, R. R. Pfeiffer, J. G. Stowell, *Solid State Chemistry of Drugs*, Second ed., SSCI, Inc., West Lafayette, **1999**.
- [2] H. G. Brittain, *Polymorphism in Pharmaceutical Solids*, Vol. 95, first ed., Marcel Dekker, Inc., New York, **1999**.
- [3] A. Grzesiak, M. Lang, K. Kim, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2003**, 92, 2260.
- [4] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, 95, 1918.
- [5] T. L. Threllfall, *Analyst* **1995**, 120, 2435.
- [6] P. McArdle, K. Gilligan, D. Cunningham, A. Ryder, *Applied Spectroscopy* **2005**, 59, 1365.
- [7] L. Malpezzi, G. A. Magnone, N. Masciocchi, A. Sironi, *Journal of Pharmaceutical Sciences* **2005**, 94, 1067.
- [8] R. J. Roberts, R. S. Payne, R. C. Rowe, *European Journal of Pharmaceutical Science* **2000**, 9, 277.
- [9] [www.radleys.com](http://www.radleys.com), viewed on 09/08/08.
- [10] M. Lang, J. W. Kampf, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2002**, 91, 1186.
- [11] S. G. Fleischman, S. S. Kuduva, J. A. McMahon, B. Moulton, R. D. Bailey Walsh, N. Rodriguez-Hornedo, M. J. Zaworotko, *Crystal Growth and Design* **2003**, 3, 909.
- [12] D. L. Simmons, R. J. Ranz, N. D. Gyanchandani, P. Picotte, *Canadian Journal of Pharmaceutical Sciences* **1972**, 7, 121.
- [13] A. Burger, *Scientia Pharmaceutica* **1975**, 43, 161.
- [14] S. S. Al-Saieq, G. S. Riley, *Pharmaceutica Acta Helvetiae* **1981**, 56, 125.
- [15] S. Thirunahari, S. Aitipamula, P. S. Chow, R. B. H. Tan, *Journal of Pharmaceutical Sciences* **2010**, 99, 2975.
- [16] Y. Sonoda, F. Hirayama, H. Arima, Y. Yamaguchi, W. Saenger, K. Uekama, *Crystal Growth and Design* **2006**, 6, 1181.
- [17] K. Nath, A. Nangia, *CrystEngComm* **2011**, 13, 47.
- [18] Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, pp. HyperChem(TM).
- [19] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople, Gaussian, Inc. Wallingford CT, **2004**.
- [20] W. L. Jorgensen, J. Tirado-Rives, *Journal of the American Chemical Society* **1988**, 110, 1657.

- [21] A. Hinchliffe, *Molecular Modelling for Beginners*, 1st ed., John Wiley & Sons Ltd., Chichester, **2003**.
- [22] R. G. Snyder, J. H. Schachtschneider, *Spectrochimica Acta* **1965**, 21, 169.
- [23] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *Journal of the American Chemical Society* **1985**, 107, 3902.
- [24] A. R. Leach, *Molecular modelling: Principles and Applications*, Addison Wesley Longman Limited, Essex, **1996**.
- [25] F. Jensen, *Introduction to Computational Chemistry*, 1st ed., John Wiley & Sons Ltd., Chichester, **1999**.
- [26] C. J. Cramer, *Essential of Computational Chemistry, Theories and Models*, 1st ed., John Wiley & Sons Ltd., Chichester, **2002**.
- [27] G. Schaftenaar, CMBI, The Netherlands, p. MOLDEN.
- [28] Gaussian Inc., Pittsburgh, PA, USA p. Gaussview 3.0.
- [29] *Hyperchem 8 manual*.
- [30] J. Simons, *An Introduction to Theoretical Chemistry*, Cambridge University Press, Cambridge, **2003**.
- [31] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *Journal of Physical Chemistry B* **2009**, 113, 6378.
- [32] M. Cossi, V. Barone, R. Cammi, J. Tomasi, *Chemical Physics Letters* **1996**, 225, 327.
- [33] J. Tomasi, B. Mennucci, E. Cancès, *Journal of Molecular Structure* **1999**, 464, 211.
- [34] P. Atkins, J. de Paula, *Atkins' Physical Chemistry*, 7th ed., Oxford University Press, Oxford, **2002**.
- [35] A. E. Stearn, H. Eyring, *Journal of Chemical Physics* **1937**, 5, 113.
- [36] MOE, Chemical Computing Group, p. Molecular Operating Environment.
- [37] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [38] M. O. Environment, Chemical Computing Group, p. Molecular Operating Environment.
- [39] C. Klanner, D. Farrusseng, L. Baumes, M. Lengliz, C. Mirodatos, F. Schuth, *Angewandte Chemie, International Edition* **2004**, 43, 5347.
- [40] L. X. Chen, J. Gasteiger, *Angewandte Chemie, International Edition* **1996**, 35, 763.
- [41] Y. Chen, D. Xu, *Bioinformatics* **2005**, 21, 575.
- [42] R. C. Wade, H. Bohr, P. G. Wolynes, *Journal of the American Chemical Society* **1992**, 114, 8284.
- [43] M. H. Fatemi, M. Jalali-Heravi, E. Konuze, *Analytica Chimica Acta* **2003**, 486, 101.
- [44] M. De Matas, Q. Shao, M. F. Biddiscombe, S. Meah, H. Chrystyn, O. S. Usmani, *European Journal of Pharmaceutical Science* **2010**, In Press.
- [45] H. White, in *IEEE International Conference on Neural Networks Vol. 2*, San Diego, CA, USA, **1988**, pp. 451.
- [46] B. Denby, *Computer Physics Communications* **1988**, 49, 429.
- [47] M. Karelson, D. A. Dobchev, O. V. Kulshyn, A. R. Katritzky, *Journal of Chemical Information and Modelling* **2006**, 46, 1891.
- [48] A. U. Bhat, S. S. Merchant, S. S. Bhagwat, *Industrial & Engineering Chemistry Research* **2008**, 47, 920.
- [49] V. Tantishaiyakul, *Journal of Pharmaceutical and Biomedical Analysis* **2005**, 37, 411.

- [50] O. Engkvist, P. Wrede, *Journal of Chemical Information and Computer Sciences* **2002**, 42, 1247.
- [51] J. Huuskonen, *Journal of Chemical Information and Computer Sciences* **2000**, 40, 773.
- [52] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, 113, 9480.
- [53] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tine, *Journal of Chemical Information and Modeling* **2006**, 46, 2030.
- [54] K. A. Blinov, Y. D. Smurnyy, M. E. Elyashberg, T. S. Churanova, M. Kvasha, C. Steinbeck, B. A. Lefebvre, A. J. Williams, *Journal of Chemical Information and Modeling* **2008**, 48, 550.
- [55] S. L. Wiskur, P. N. Floriano, E. V. Anslyn, J. T. McDevitt, *Angewandte Chemie, International Edition* **2003**, 42, 2070.
- [56] T. Aoyama, Y. Suzuki, H. Ichikawa, *Journal of Medicinal Chemistry* **1990**, 33, 905.
- [57] K. Gurney, *And Introduction to Neural Networks*, Routledge, London, **2001**.
- [58] J. J. Hopfield, *Proceedings of the National Academy of Sciences of the United States of America* **1982**, 79, 2554.
- [59] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Second ed., WILEY-VCH, Weinheim, **1999**.
- [60] W. S. McCulloch, W. Pitts, *Bulletin of Mathematical Biophysics* **1943**, 5, 115.
- [61] J. J. Hopfield, *Proceedings of the National Academy of Sciences of the United States of America* **1984**, 81, 3088.
- [62] J. Bourquin, H. Schmidli, P. van Hoogevest, H. Leuenberger, *Pharmaceutical Development and Technology* **1997**, 2, 95.
- [63] J. Zupan, J. Gasteiger, *Analytica Chimica Acta* **1991**, 248, 1.
- [64] J. C. Principe, N. R. Euliano, W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals through Simulations*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [65] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Second ed., Prentice-Hall, Inc., New Jersey, **1999**.
- [66] C. M. Handley, P. L. A. Popelier, *Journal of Physical Chemistry A* **2010**, 114, 3371–3383.
- [67] N. Rizkalla, P. Hildgen, *Drug Development and Industrial Pharmacy* **2005**, 31, 1019.
- [68] G. Hanrahan, *Analytical Chemistry* **2010**, 82, 4307.
- [69] A. P. Plumb, R. C. Rowe, P. York, M. Brown, *European Journal of Pharmaceutical Science* **2005**, 25, 395.
- [70] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [71] *INForm: Intelligent Formulation manual v3.7*, Intelligensys Ltd., **2009**.
- [72] NeuroSolutions™, NeuroDimension, Inc., Gainesville, Florida.
- [73] *FormRules: Formulating Rules software manual v3.3*, Intelligensys Ltd., **2007**.
- [74] H. M. Cartwright, L. M. Stztandera, *Soft Computing Approaches in Chemistry, Vol. 120*, first ed., Springer-Verlag, Heidelberg, **2003**.
- [75] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [76] Q. Shao, R. C. Rowe, P. York, *European Journal of Pharmaceutical Sciences* **2006**, 28, 394.

- [77] M. Alleso, F. Van Den Berg, C. Cornett, F. S. Jorgensen, B. Halling-Sorensen, H. Lopez De Diego, L. Hovgaard, J. Aaltonen, J. Rantanen, *Journal of Pharmaceutical Sciences* **2008**, 97, 2145.
- [78] R. C. Schweitzer, J. B. Morris, *Analytica Chimica Acta* **1999**, 384, 285.
- [79] M. S. Bhatia, K. B. Ingale, P. B. Choudhari, N. M. Bhatia, R. L. Sawant, *Bioorganic and Medicinal Chemistry* **2009**, 17, 1654.
- [80] Y. Ren, H. Liu, X. Yao, M. Liu, *Journal of Chromatography A* **2007**, 1155, 105.
- [81] M. Ringner, *Nature Biotechnology* **2008**, 26, 303.
- [82] W. R. Dillon, M. Goldstein, *Multivariate Analysis: Methods and Applications*, 1st ed., John Wiley & Sons, Inc., New York, **1984**.
- [83] C. Seaton, **2011**, pp. Personal email correspondence.
- [84] B. G. Tabachnick, L. S. Fidell, *Using Multivariate Statistics*, 5th ed., Pearson Education, Inc., Boston, **2007**.
- [85] H. L. Zhai, X. G. Chen, Z. D. Hu, *Chemometrics and Intelligent Laboratory Systems* **2006**, 80, 130.
- [86] M. J. Anzanello, S. L. Albin, W. A. Chaovalitwongse, *Chemometrics and Intelligent Laboratory Systems* **2009**, 97, 111.
- [87] D. T. Stanton, P. J. Madhav, L. J. Wilson, T. W. Morris, P. M. Hershberger, C. N. Parker, *Journal of Chemical Information and Computer Sciences* **2004**, 44, 221.
- [88] K. Kipouros, K. Kachrimanis, I. Nikolakakis, S. Malamataris, *Analytica Chimica Acta* **2005**, 550, 191.
- [89] H. Golmohammadi, *Journal of Computational Chemistry* **2009**, 30, 2455.
- [90] A. Hoskuldsson, *Chemometrics and Intelligent Laboratory Systems* **2001**, 55, 23.
- [91] SIMCA-P+, Umetrics AB, Umeå, Sweden, **12.01**.

### 3. SYSTEMS STUDIED

Three different polymorphic systems were examined by the research described in this thesis: carbamazepine (CBZ), tolbutamide (TBA) and 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY). Most of the work focused on the generation of a large set of carbamazepine data and its analysis; ROY and tolbutamide were used at an advanced stage of the project as validation systems. This chapter summarises previously published research on these polymorphic systems.

#### 3.1. Carbamazepine

5H-Dibenz[b,f]azepine-5-carboxamide (Figure 3.1) more commonly known as carbamazepine (CBZ), has been selected for analysis within this study. Carbamazepine is administered as a treatment to epilepsy and bipolar disorder worldwide due to its “analgesic and anticonvulsion properties”<sup>[1-4]</sup>. CBZ has four anhydrous<sup>[2, 5-8]</sup> polymorphs, a dihydrate<sup>[9-11]</sup> and a number of solvates<sup>[12, 13]</sup> making it an ideal initial system to work with in this research as there are a number of possible outcomes for polymorph screen experiments.

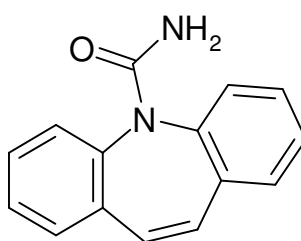


Figure 3.1 The molecular structure of carbamazepine

##### 3.1.1. Why Carbamazepine for this Research?

There are a number of reasons for the selection of this molecule; from its chemical structure to the number of polymorphs it is known to exist in. From the schematic of CBZ (Figure 3.1) it is evident that the molecule is quite rigid, with the only area

allowing for bond rotation being between the N attached to the ring and the C of the amide group (Figure 3.2).



Figure 3.2 Highlighting bond rotation in CBZ molecule

When modelling this molecule it is treated as rigid as is done in the literature,<sup>[14, 15]</sup> which is beneficial because it allows the minimum energy structure to be located more easily.<sup>[15]</sup> A more practical advantage of the rigid structure is that it may be clearer to discover molecular subtleties when modelling CBZ in different solvent force fields (explained in section 2.6.2.3)

Having more than two polymorphs is advantageous in this study also, as it may allow identification of experimental patterns. When the outcome of the polymorph screen experiments is limited to two, there are less areas of obvious difference between each form.

### 3.1.2. The Carbamazepine Polymorphs

CBZ is readily available from Sigma Aldrich in the *P*-monoclinic form (form III). From this commercial CBZ, the four characterised polymorphic forms, solvates and dihydrate can be made following the methods stated in section 2.3. All of the CBZ polymorphs have identical molecular conformations within the unit cell, but it is the way that they are arranged that creates the different polymorphic forms<sup>[2]</sup>. What is also notable in the CBZ polymorphic forms is that they all exist in hydrogen-bonded dimers<sup>[16]</sup>. This is due to the strong hydrogen-bonding between the carboxamide groups of two molecules, shown in Figure 3.3.

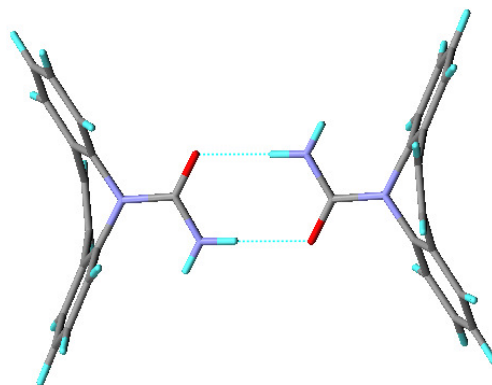


Figure 3.3 CBZ hydrogen-bonded dimer

The four anhydrous forms vary in stability, but the order follows the density rule, which states that the more dense the form, the more stable it is likely to be<sup>[17]</sup>. The densities of the anhydrous forms (in g/cm<sup>3</sup>) as determined by DSC analysis by Grzesaik et al.<sup>[2]</sup> are shown in brackets and the order of stability shown below.

Form III (1.34) > Form I (1.31) > Form IV (1.27) > Form II (1.24)<sup>[2]</sup>

#### 3.1.2.1. Form I

CBZ form I is a triclinic<sup>[6]</sup> crystal that is unique with regard to the other anhydrous forms due to the number of molecules in the unit cell. Unlike forms II-IV, form I has four molecules in the unit cell<sup>[2]</sup>, whereas the others have only one. The structure for this polymorph (Figure 3.4) may be found in the Cambridge structural database (CSD), reference CBMZPN11<sup>[2]</sup>.

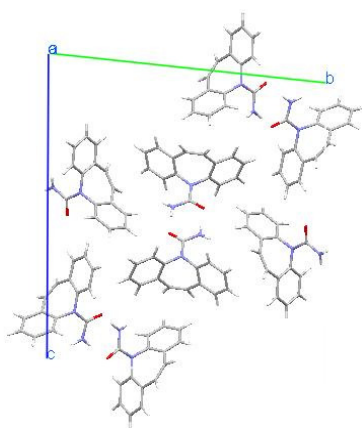


Figure 3.4 Packing diagram for form I, taken from the CSD (reference CBMZPN11<sup>[2]</sup>)



Form I and III are enantiotropically related<sup>[18]</sup>, with form III being the most stable until temperatures exceed 100°C<sup>[6]</sup>, however in other research this temperature is quoted as being 71°C<sup>[19]</sup>. All CBZ anhydrous forms transform into form I upon heating which is highlighted in thermomicroscopy work by Grzesiak et al.<sup>[2]</sup>.

#### **3.1.2.2. Form II**

The trigonal form of CBZ (CSD reference CBMZPN03<sup>[7]</sup>) is often found to crystallise when solvents with low dielectric constants are used<sup>[7, 20]</sup>. As with forms III and IV, form II melts and recrystallises to give form I. In this case, the melt is not as obvious and easily overlooked, but a small endotherm has been reported between 140-160°C<sup>[2]</sup>. This endotherm was not reported in work by Lowes et al.<sup>[7]</sup>, only the melt of form I.

Form I and II are structurally similar due to the weak hydrogen-bonding, whereby the oxygen atom will accept the hydrogens from the nearest two carbons<sup>[2]</sup>. However, unlike form I, it has been reported that if solvent is included into the structure, the stability of the form is increased<sup>[21, 22]</sup>. This research stemmed from the observation that over one hundred hypothetical polymorphs of CBZ could be computationally predicted with greater stability than form II.<sup>[21]</sup> By including solvent into the unit cell, the stability is improved<sup>[22]</sup>. The research suggested that less than 7 % of the CBZ unit cell could accommodate a solvent molecule, and therefore it is highly likely to have been missed in previous work. The solvent may also be rapidly released at room temperature providing another reason why it has not been reported earlier. The model established in work by Cruz Cabeza et al.<sup>[21]</sup>, calculated that when 3-4.5 wt.% of toluene is included in the unit cell, the stability dramatically improved. Figure 3.5 clearly shows the voids in the structure that could contain solvent molecules.

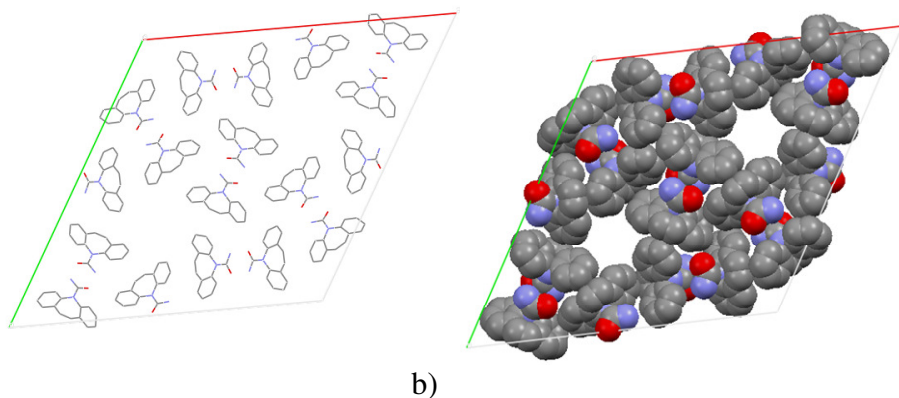


Figure 3.5 a) packing diagram of form II, taken from the CSD (reference CBMZPN03<sup>[7]</sup>) b) space filled model highlighting the possible site for solvent inclusion

### 3.1.2.3. Form III

CBZ form III is the thermodynamically stable polymorph that crystallises in a primitive monoclinic cell<sup>[1, 2, 23]</sup> and can be found in the CSD (reference CBMZPN01<sup>[23]</sup>). The packing of this crystal structure is shown in Figure 3.6.

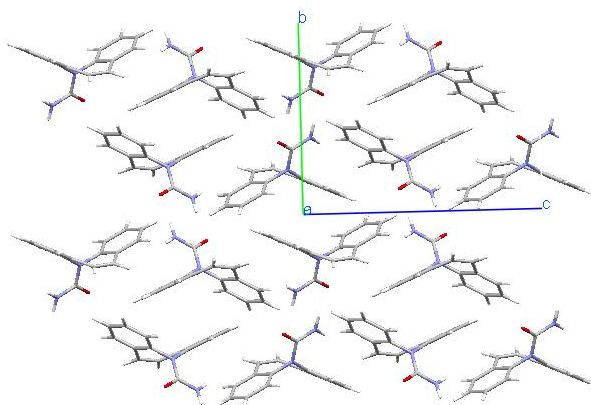


Figure 3.6 Packing diagram for CBZ form III, taken from the CSD (reference CBMZPN01<sup>[23]</sup>)

Form III is most often formed when crystallised from solvents with high dielectric constants or at slow cooling rates<sup>[7]</sup>, and is the desired output in industrial manufacture of CBZ as a drug.

### 3.1.2.4. Form IV

Form IV has a crystal structure in the *C*-monoclinic<sup>[5]</sup> form, which like the other forms is dominated by hydrogen-bonded dimers, shown in Figure 3.7 (CSD reference CBMZPN12<sup>[5]</sup>).

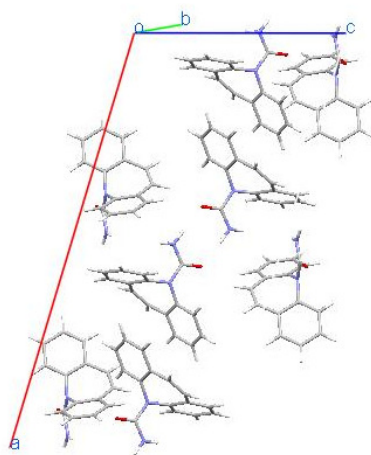


Figure 3.7 Packing diagram of form IV, taken from the CSD (reference CBMZPN12<sup>[5]</sup>)

Throughout this current research and polymorph screens documented by Florence et al.<sup>[24]</sup> form IV does not crystallise using conventional cooling methods. There have been a number of methods noted that may cause this form to crystallise and these are by dehydration of the dihydrate which acts as a precursor for form IV crystallisation, salting out from ethanol solutions and spray drying<sup>[25]</sup>.

#### 3.1.2.5. CBZ Dihydrate

It is thought that anhydrous CBZ is insoluble in water and readily converts into the CBZ dihydrate on contact with water<sup>[11, 26, 27]</sup>. Laine et al.<sup>[11]</sup> suggested that the anhydrous CBZ in water acts as “nucleation centres”<sup>[11]</sup> for the dihydrate needles to grow onto. Later work by Reck et al.<sup>[28]</sup> noted that the XRPD traces shown by Laine et al.<sup>[11]</sup> did not agree with their own theoretically calculated trace, leading them to believe there was more than one dihydrate form. This phenomenon was investigated at a later date by forming the dihydrate from forms I and III. Although initial differences were seen in the DSC analysis, they concluded on further analysis that there was in fact only one dihydrate form of CBZ<sup>[26]</sup>. Reck et al.<sup>[28]</sup> stated that there was only half a CBZ molecule and one water in the asymmetric unit. However, newer research by Harris et al.<sup>[18]</sup> and Gelbrich et al.<sup>[29]</sup> determined that there are two water molecules and one CBZ molecule in the asymmetric unit (CSD reference FEFNOT02<sup>[18]</sup>). The CBZ dihydrate is monoclinic and shares the hydrogen-bonded dimers feature seen in the polymorphs of CBZ (Figure 3.8).

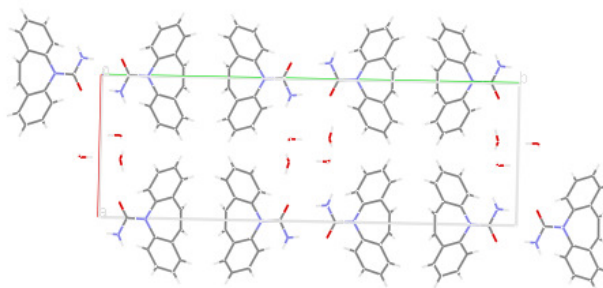


Figure 3.8 Packing of CBZ dihydrate, taken from the CSD (reference FEFNOT02<sup>[18]</sup>)

### 3.1.2.6. CBZ Solvates

CBZ is known to form a number of different solvates, with Fleischman et al.<sup>[12]</sup> identifying acetone, DMSO, methanol, ethanol, acetic acid, formic acid and butyric acid solvates, Lohani et al.<sup>[30]</sup> presented a trifluoroethanol solvate and Johnston et al.<sup>[31]</sup> presented a DMF solvate. Within this research only acetone and DMSO solvates have been crystallised. The DMSO and acetone solvates are both triclinic and both have the hydrogen-bonded dimers seen in other forms of CBZ (Figure 3.9 and Figure 3.10), their CSD references are UNEYIV<sup>[12]</sup> and CRBMZA01<sup>[12]</sup> respectively.

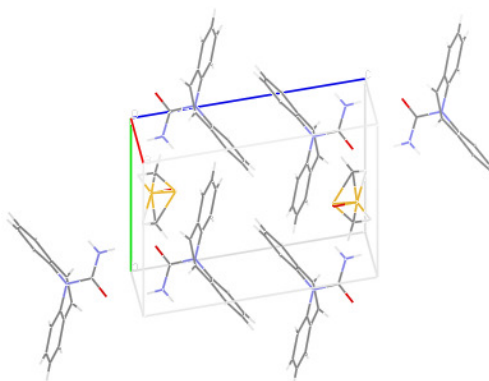


Figure 3.9 Packing of the CBZ DMSO solvate, taken from the CSD (reference UNEYIV<sup>[12]</sup>)

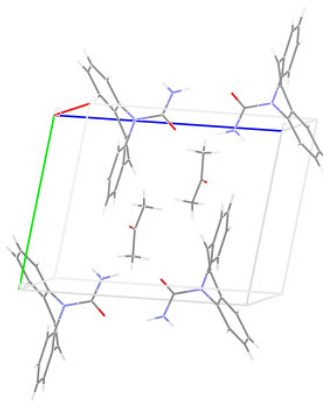


Figure 3.10 Packing of the CBZ acetone solvate, taken from the CSD (reference CRBMZA01<sup>[12]</sup>)

Table 3.1 Selected parameters of CBZ anhydrous polymorphs and dihydrate, taken from Grzesiak et al.<sup>[2]</sup> unless stated

Name	Morphology	Crystal System	Space Group [No.]	Melting point (°C) <sup>a</sup>
Form I	Needle-like	Triclinic	<i>P</i> -1 <sup>[16]</sup>	193.5
Form II	Needle-like	Trigonal	<i>R</i> -3 <sup>[7]</sup>	140-160
Form III	Prism	<i>P</i> -monoclinic	<i>P</i> 2 <sub>1</sub> / <i>n</i> <sup>[5]</sup>	174.8
Form IV	Needle-like <sup>[32]</sup>	<i>C</i> -monoclinic <sup>[5]</sup>	<i>C</i> 2/ <i>c</i> <sup>[5]</sup>	187.7
Dihydrate	Needle-like <sup>[11]</sup> / Plates <sup>[27]</sup>	<i>P</i> -monoclinic <sup>[25]</sup>	<i>P</i> 2 <sub>1</sub> / <i>c</i> <sup>[18]</sup>	48-80 <sup>b[27]</sup>

<sup>a</sup> Calculated from DSC analysis at a heating rate of 20°C/min <sup>b</sup> Calculated from DSC analysis at a heating rate of 100°C/min

### 3.2. 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile

5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile or ROY (Figure 3.11) is an intermediate in the synthesis of Olanzapine, an Eli Lilly antipsychotic drug<sup>[33, 34]</sup>. ROY stands for Red, Orange, Yellow, which are the colours of the different polymorphic forms arising from the differing levels of conjugation between the thiophene and phenyl rings in each polymorphic form<sup>[35]</sup>.

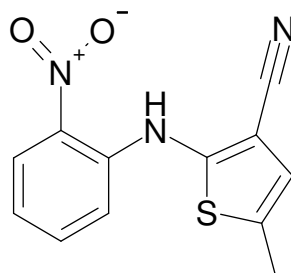


Figure 3.11 Structure of ROY

#### 3.2.1. Using ROY for Polymorphic Investigation

The ROY polymorphs are conformational, with different crystal habits and colours that may aid this investigation in a number of ways.

Being conformationally flexible allows multiple computer simulations of the different orientations of ROY. This could then enable a deeper understanding as to why different solvents interact with a particular conformation more than others, and perhaps give insight of interactions at the molecular level.

With each polymorph having a different habit and colour this would make identification of each form easier. The colour of the crystal could instantly rule out certain forms and microscopy could be used to identify the habit, making identification of the forms more efficient (see Figure 3.12). XRPD data from the CSD shows that the different forms have distinct patterns and this can and has been used in identification.

Yu et al.<sup>[36]</sup> stated that the polymorphic forms generated from different solvents are not very selective. This could be viewed in a number of ways, firstly as a means to identify subtle parameters that influence polymorph selection within the solution phase, done with a controlled polymorph screen. Alternately it could be viewed as a potential barrier in providing these crucial parameters if multiple forms were generated concomitantly.

### 3.2.2. The ROY Polymorphs

ROY currently holds the record for the highest number of characterised polymorphs in the CSD<sup>[35]</sup>. Seven of these forms have been reported<sup>[36, 37]</sup> within the (CSD) and it is known that there are currently three other forms<sup>[38]</sup> that have been crystallised. Figure 3.12 shows the ten polymorphs and clearly demonstrates the colour and morphological differences between forms. The three forms that are yet to be characterised have been found by other methods of polymorph analysis, for example melt crystallisation<sup>[38]</sup>, solid state conversion<sup>[37]</sup> and cross nucleation<sup>[38]</sup>, opposed to the conventional methods of evaporation and cooling crystallisation. This highlights the fact that different forms may be missed during a traditional solvent screen.

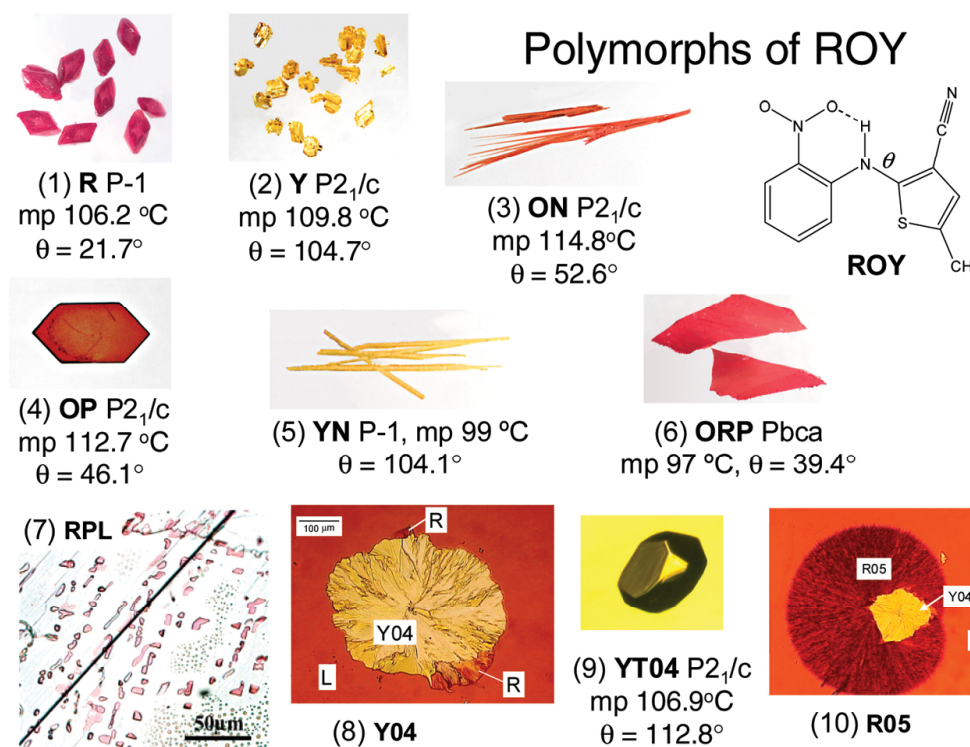


Figure 3.12 Taken from a paper by Yu, this diagram shows the different morphologies and colours of the ROY polymorphs<sup>[35]</sup>

The major differences visually between the polymorphs would allow identification of each form using microscopy. However, it would be more difficult if mixtures were crystallised, and due to this, XRPD has been used for identification purposes. Not only does each polymorph have a unique trace, but by using XRPD, quantitative analysis can be undertaken that is useful for the ANN analysis. Table 3.2 highlights key differences between the different forms, crystal information and melting points.

Table 3.2 Select parameters of the ROY polymorphs, taken from Yu<sup>[36]</sup> unless stated.

Name	Morphology and colour	Crystal System	Space Group [No.]	Melting point (°C)
R	Red Prisms	Triclinic	<i>P1</i> [2]	106.2
Y	Yellow Prisms	Monoclinic	<i>P2<sub>1</sub>/n</i> [14]	109.8
OP	Orange Plates	Monoclinic	<i>P2<sub>1</sub>/n</i> [14]	112.7
ON	Orange Needles	Monoclinic	<i>P2<sub>1</sub>/c</i> [14]	114.8
YN	Yellow Needles	Triclinic	<i>P1</i> [2]	Thermally unstable
ORP	Orange-Red Plates	Orthorhombic	<i>Pbca</i> [61]	Thermally unstable
YT04 <sup>a</sup>	Yellow Prism	Monoclinic	<i>P2<sub>1</sub>/n</i> [14]	106.9
Y04 <sup>b</sup>	Yellow Prism	Unknown	Structure not solved	Thermally unstable
RPL <sup>b</sup>	Red Plate	Unknown	Structure not solved	Thermally unstable
R05 <sup>b</sup>	Red Prism	Unknown	Structure not solved	Thermally unstable

<sup>a</sup>Details taken from Chen et al.<sup>[37]</sup> <sup>b</sup>Details taken from Chen et al.<sup>[38]</sup>

The first seven forms in Table 3.2 are known to have only one molecule in each asymmetric unit<sup>[39]</sup>, and the way in which they pack is varied (Figure 3.13, Figure 3.14, Figure 3.15). This information is not available in the literature for the other forms of ROY.

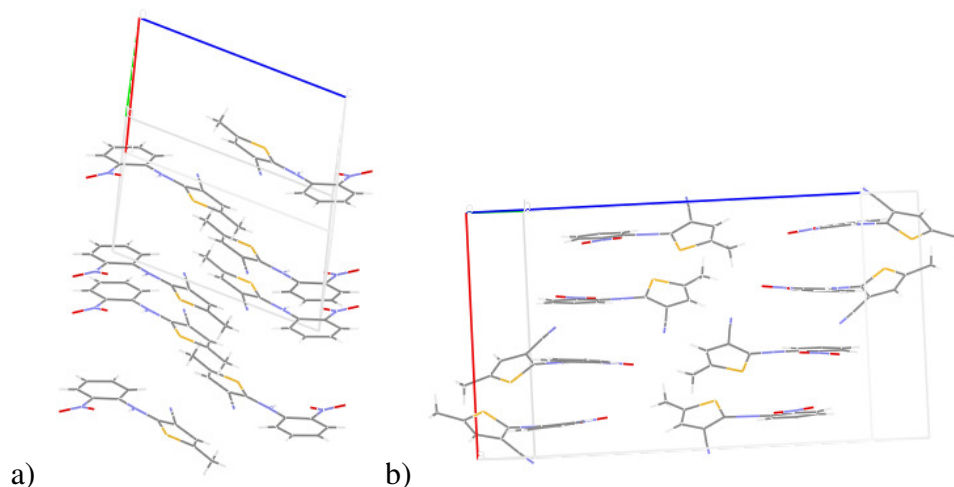


Figure 3.13 ROY crystal structures, a) form R (CSD reference QAXMEH02<sup>[36]</sup>) and b) ORP (CSD reference QAXMEH05<sup>[36]</sup>)



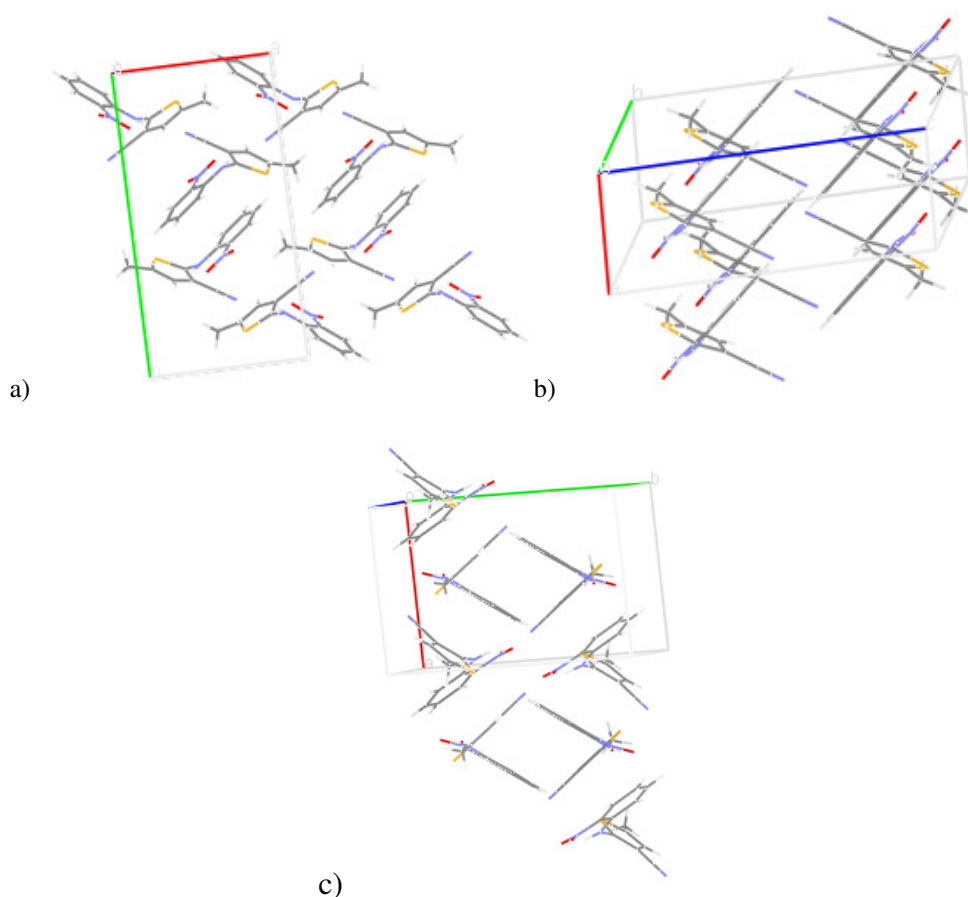


Figure 3.14 ROY crystal structures, a) form Y (CSD reference QAXMEH01<sup>[36]</sup>), b) YN (CSD reference QAXMEH04<sup>[36]</sup>) and c) YT04 (CSD reference QAXMEH12<sup>[37]</sup>)

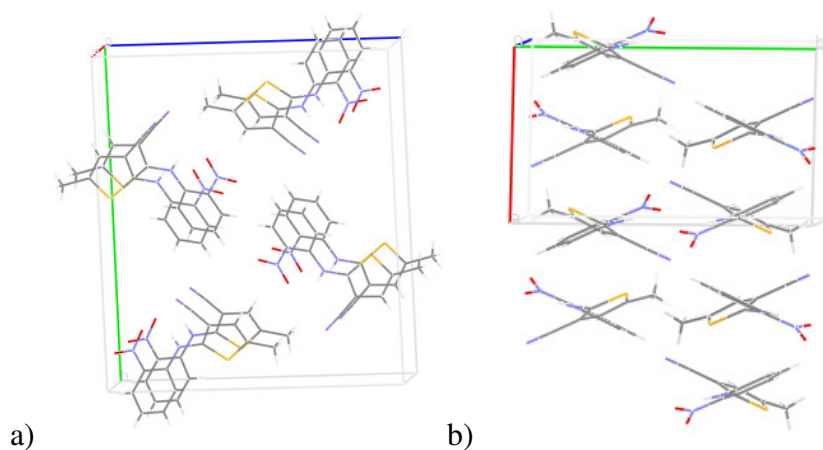


Figure 3.15 ROY crystal structures, a) form ON (CSD reference QAXMEH<sup>[36]</sup>) and b) OP (CSD reference QAXMEH03<sup>[36]</sup>)

The stability order of the forms varies with temperature and are highlighted below<sup>[38]</sup>.

Between 40-70°C:  $Y > ON \approx OP > YT04 > R > YN$

Above 70°C:  $ON > OP > Y > YT04 > R > YN$

Y is accepted as the thermodynamically stable polymorph, with Y, ON and OP being kinetically stable at room temperature.

It should also be noted that in most of the ROY polymorphs there are only intramolecular hydrogen-bonding between the amine and nitro groups.<sup>[40]</sup> However, the Y and YT04 forms also have a weak intermolecular hydrogen-bond<sup>[37]</sup> shown in Figure 3.16.

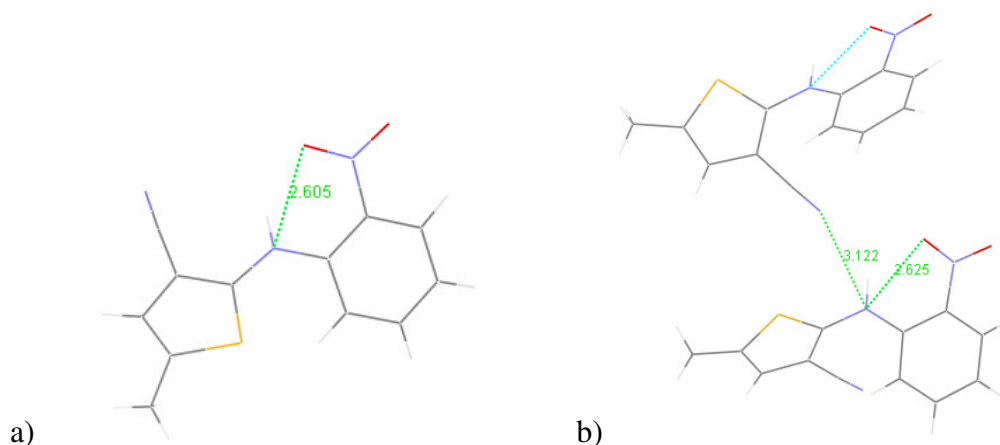


Figure 3.16 a) intramolecular H-bond in ORP b) intermolecular H-bond in Y

### 3.3. Tolbutamide

Tolbutamide (1-butyl-3-(4-methylphenylsulfonyl)urea, TBA) is a hypoglycemic agent, taken orally in the treatment of insulin-dependent diabetes sufferers<sup>[41-43]</sup>. There are five known anhydrous forms of TBA<sup>[42, 44-47]</sup> and no currently reported hydrate or solvate structures.

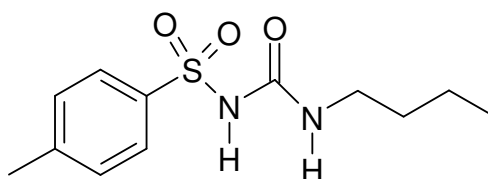


Figure 3.17 The molecular structure of tolbutamide

#### 3.3.1. Why Tolbutamide for this Research?

TBA was chosen for this research because it has a high number of different polymorphic forms and unlike CBZ it is conformationally flexible. When this research was initiated only two of the polymorphic forms crystal structures were characterised (form I and III), but now, crystal structure data is available for all five forms.

### 3.3.2. The Tolbutamide Polymorphs

TBA is available from Sigma Aldrich in the orthorhombic form (form I) and from this commercial form, the five characterised polymorphic forms can be made following the methods stated in section 2.5.

TBA is a flexible molecule, with different conformations being adopted in the different polymorphic forms<sup>[46]</sup>. A conformational search was carried out in Hyperchem<sup>TM</sup><sup>[48]</sup> and the minimum energy structure was taken forward for further geometry optimisation calculations. Based on the small amount of information known about TBA in the initial stages of this research, the method used was to take the minimum energy structure without regard for the conformation in the crystal structure. The intention behind this was to replicate the situation of a new drug molecule that has no other data associated with it other than the molecular structure. If this overall method was to perform well without any solid state knowledge of the molecule, then it would be highly valuable in drug discovery as a method for identifying potentially polymorphic molecules.

Research by Thirunahari et al.<sup>[46]</sup> described the conformations of TBA in the crystal structures of the different polymorphic forms as either U or chair like. These representations are based on whether the phenyl ring and alkyl chain are on the same side of the S-N1-C8-N2-C9 plane<sup>[46]</sup>.

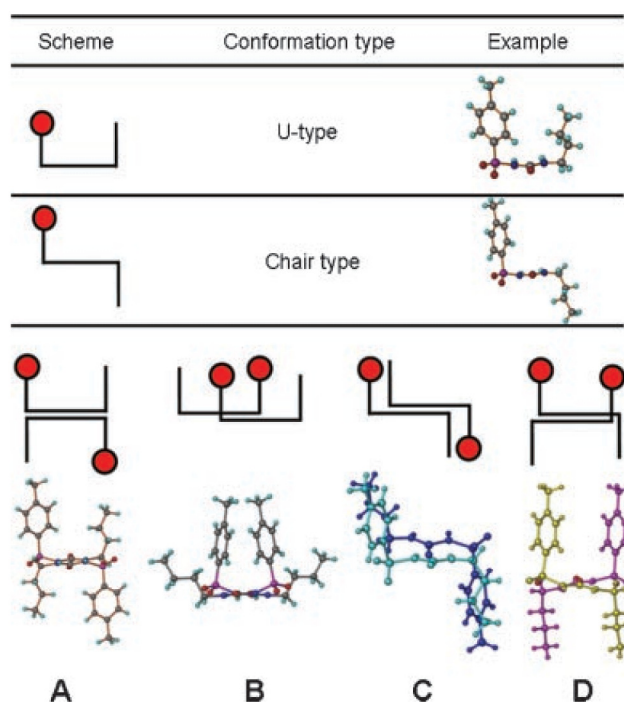


Figure 3.18 U and chair type configurations of TBA polymorphs taken from Thirunahari et al.<sup>[46]</sup>

### 3.3.2.1. Form I

The crystal structure of form I can be found in the CSD (reference ZZZPUS01<sup>[49]</sup> and ZZZPUS02<sup>[50]</sup>) and displays the U type configuration (Figure 3.19).

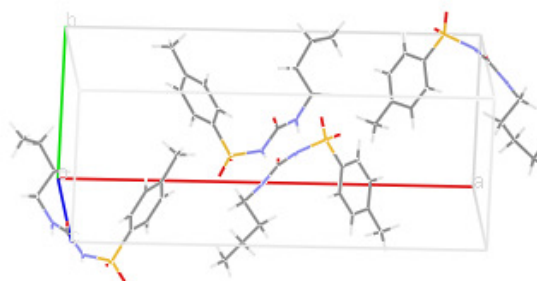


Figure 3.19 Packing diagram for form I, taken from the CSD (reference ZZZPUS02<sup>[50]</sup>)

This form is commercially available with a melting point of 127°C that is reported consistently throughout the literature. However, an unexpected peak at 39°C is present in the DSC trace (Figure 3.20).

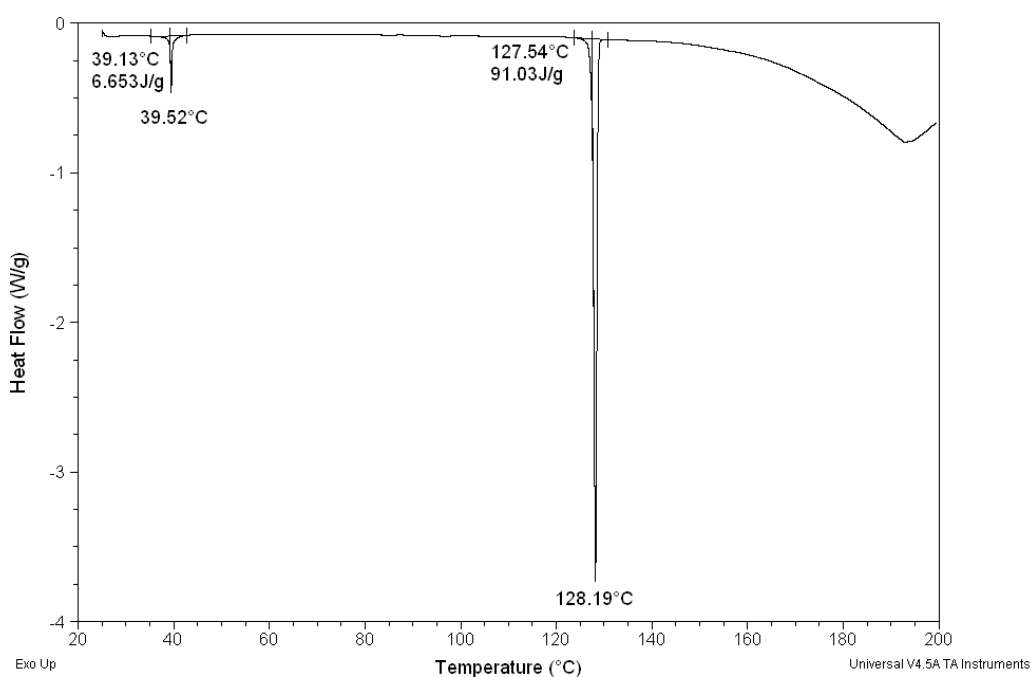


Figure 3.20 DSC of form I, run at 2°C/min

This small endotherm was commented upon in only some of the literature<sup>[43, 45, 46, 51-53]</sup> and was attributed to the “rearrangement of hydrogen bonds”<sup>[45]</sup> during a solid-solid transition<sup>[52]</sup>. This transition is also known to be kinetically reversible, as when the heating rate and direction of temperature change is altered, the position of this peak does not change<sup>[54]</sup> (Figure 3.21, Figure 3.22).

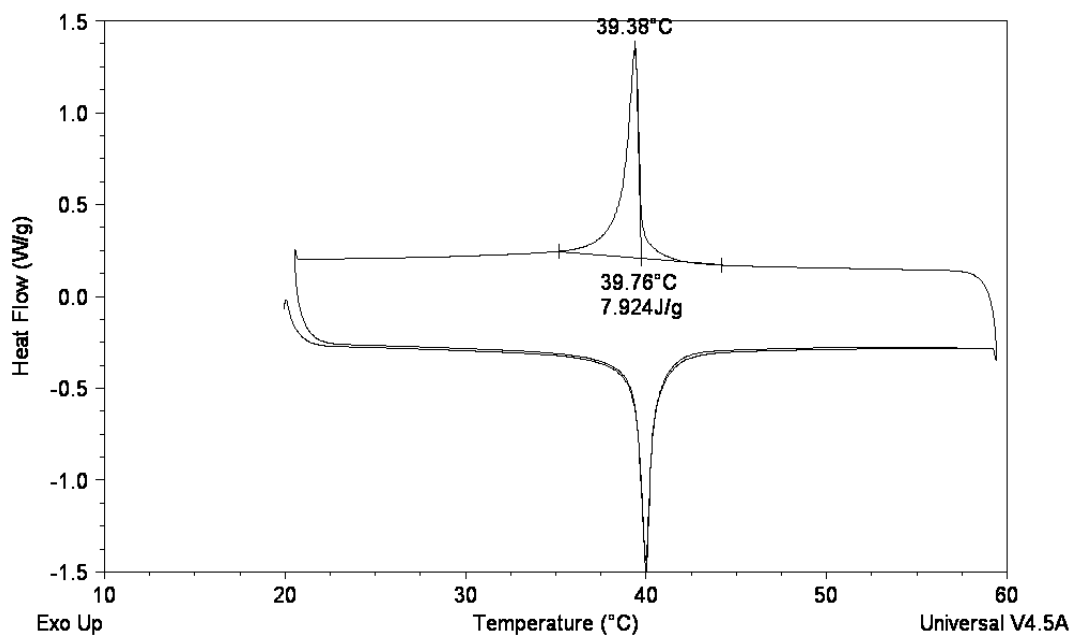


Figure 3.21 DSC of small endotherm, run at 10°C/min

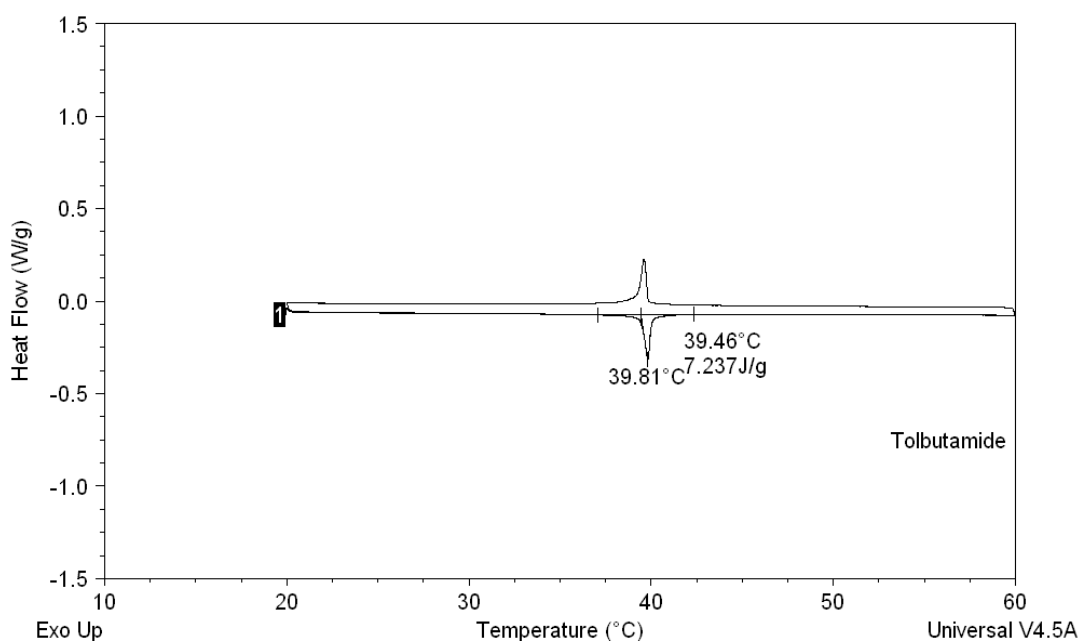


Figure 3.22 DSC of small endotherm, run at 1°C/min

Hasegawa et al.<sup>[51]</sup> analysed the form I sample above and below 39°C and observed a change in the XRPD pattern. The structure of the form I<sup>H</sup> (above the transition temp.) was solved from powder and generated a structure very similar to that of form I (denoted I<sup>L</sup> here). Figure 3.23 shows the structural difference in the low and high temperature polymorphs of form I.<sup>[51]</sup> The hot-stage XRPD analysis that was repeated in this work and is shown in Figure 3.24

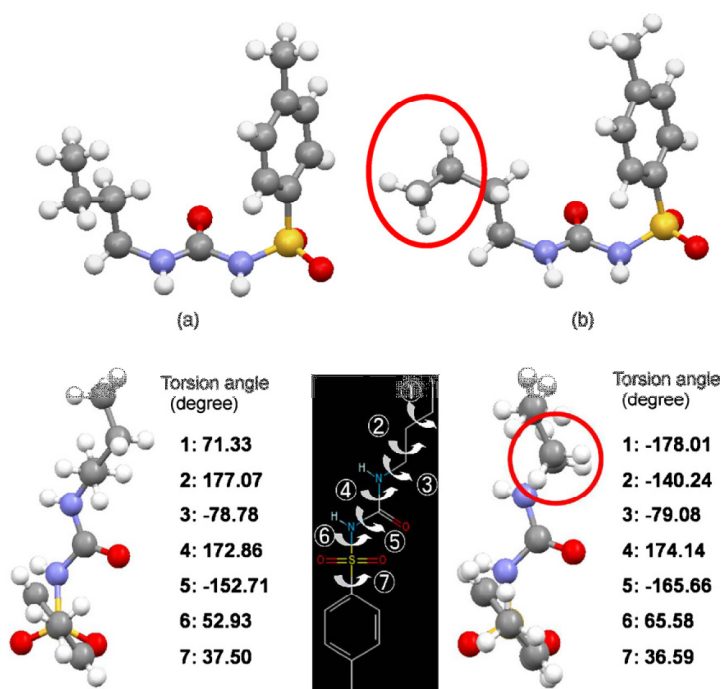


Figure 3.23 Taken from Hasegawa et al.<sup>[51]</sup> the molecular structure and torsion angles of the two forms, a) Form I<sup>L</sup> b) form I<sup>H</sup>

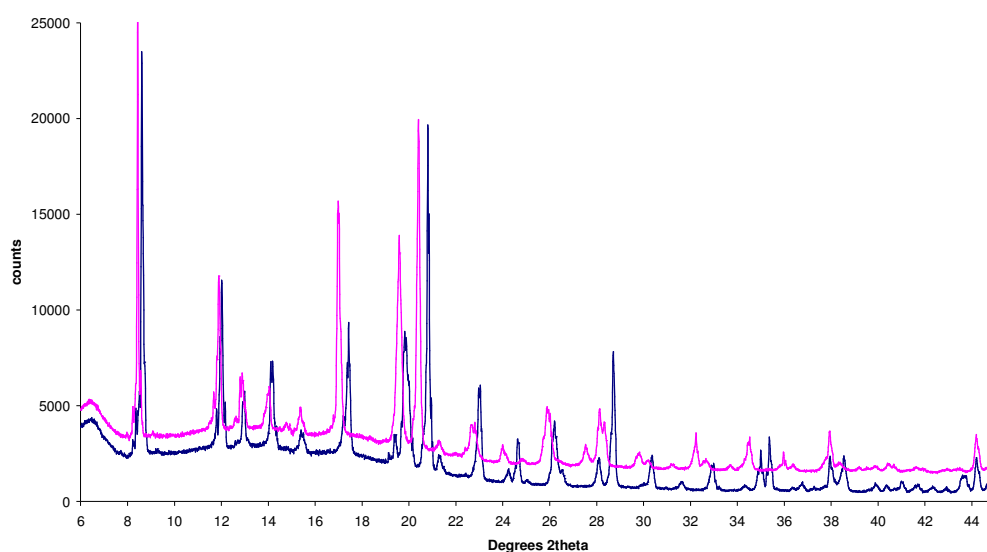


Figure 3.24 Hot-stage XRPD of TBA form I above and below small endotherm seen in DSC analysis. Blue line is the form I sample at 25°C and the pink is the form I sample at 50°C

### 3.3.2.2. Form II

In the literature there is a discrepancy in the identity of form II, with Kimura et al.<sup>[42]</sup> first publishing a crystal structure identified as form II in 1999 (CSD reference ZZZPUS03). Unfortunately there are no coordinates with this input and therefore a theoretical powder pattern could not be calculated, so data from the paper has to be

used. However, in 2010 Thirunahari et al.<sup>[46]</sup> determined that this published crystal structure is for form III and not in fact form II, and produced their own crystal structure of form II (Figure 3.25). The data generated by Thirunahari et al.<sup>[46]</sup> matches what was determined in this work and therefore the polymorphs will be named accordingly.

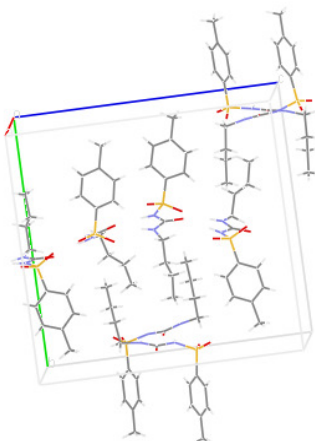


Figure 3.25 Packing diagram for form II, taken from the CSD (reference ZZZPUS03<sup>[42]</sup>)

The TBA molecules are in the chair configuration in form II, and show both C and D packing types seen in Figure 3.18 and in Figure 3.26.

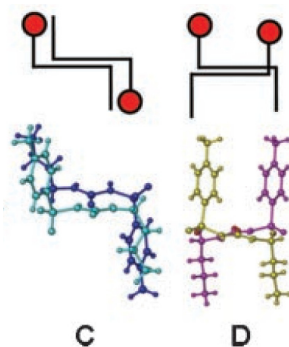


Figure 3.26 C and D chair type conformations seen in the form II crystal structure, taken from Thirunahari et al.<sup>[46]</sup>

The XRPD trace presented by Kimura et al.<sup>[42]</sup> matches that of this work and Thirunahari et al.<sup>[46]</sup>. However, the DSC value of 100°C for form II is much lower than 111-117°C stated in other literature<sup>[44, 46, 51]</sup>. The assessment made by Thirunahari et al.<sup>[46]</sup> that the data presented as form II (with the exception of the XRPD) is in fact form III seems to stand.

### 3.3.2.3. Form III

The crystal structure of form III was determined by Leary et al.<sup>[45]</sup> without presenting coordinates and does not appear in the CSD. The unit cell parameters given were  $a = 8.11$ ,  $b = 8.96$ ,  $c = 10.19$  in the space group  $P2_1$ . In more recent work by Thirunahari et al.<sup>[46]</sup> they stated that none of their polymorphs generated unit cell parameters like that of Leary, and gave a new set of unit cell parameters ( $a = 11.74$ ,  $b = 9.04$ ,  $c = 13.73$ ), which are also known by Kimura et al.<sup>[42]</sup> as form II. The XRPD pattern that was presented in their work has been used as the form III standard and does match previous patterns found in the literature<sup>[42, 51, 55, 56]</sup>. Although the crystal structure coordinates are not currently available for this monoclinic structure, the literature states that the TBA molecules are arranged in the packing arrangement A, seen in Figure 3.18 and Figure 3.27<sup>[46]</sup>.

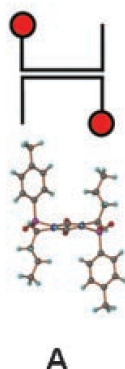


Figure 3.27 U type packing motif of TBA form III, taken from Thirunahari et al.<sup>[46]</sup>

There is again some discrepancy in the literature over the melting point of form III, Thirunahari et al.<sup>[46]</sup> gave a value of 106°C, which closely matches work by Ueda et al.<sup>[57]</sup>, Umeda et al.<sup>[58]</sup> and Hasegawa et al.<sup>[51]</sup>. Simmons et al.<sup>[59]</sup>, Burger<sup>[44]</sup>, Leary et al.<sup>[45]</sup> and Kimura et al.<sup>[42]</sup> stated values of between 113-117°C.

### 3.3.2.4. Form IV

The crystal structure of monoclinic form IV was determined by Thirunahari et al.<sup>[46]</sup> from powder data, determining that there is only one U-type molecule in the asymmetric unit, arranged in packing motif B shown in Figure 3.18 and Figure 3.28.



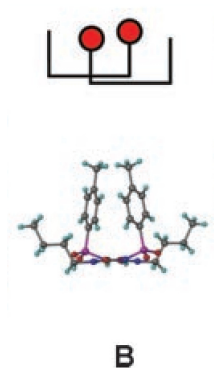


Figure 3.28 U type packing motif of form IV taken from Thirunahari et al.<sup>[46]</sup>

The unit cell parameters are  $a = 10.09$ ,  $b = 15.65$ ,  $c = 9.26$  in the  $P2_1/c$  space group which is consistent with the space group stated in earlier work by Sonoda et al.<sup>[60]</sup>.

### 3.3.2.5. Form V

A fifth form of TBA was discovered recently by adding conc.  $\text{HNO}_3$  to a TBA and methanol solution and allowing slow evaporation.<sup>[47]</sup> The crystal structure for this polymorph is not yet present in the CSD. It is a highly metastable form and readily converts to form I at RT, therefore the crystal structure was determined at 100 K to prevent moisture speeding up the polymorphic transformation. The space group was determined to be  $Pbcn$ , which is unlike any other form, with the unit cell parameters of  $a = 15.85$ ,  $b = 9.29$ ,  $c = 19.69$ . The molecules pack in a similar way to form IV which is the U type motif seen in Figure 3.28.

Table 3.3 Selected parameters of TBA anhydrous polymorphs

Name	Morphology	Crystal System	Space Group [No.]	Melting point ( $^{\circ}\text{C}$ ) <sup>a</sup>
Form I	Prismatic <sup>[59]</sup>	Orthorhombic <sup>[45]</sup>	$P2_1a n$ <sup>[45]</sup>	127 $^{\circ}\text{C}$ <sup>[59]</sup>
Form II	Needles <sup>[46]</sup>	Monoclinic <sup>[46]</sup>	$P2_1/n$ <sup>[46]</sup>	117 $^{\circ}\text{C}$ <sup>[46]</sup>
Form III	Plates <sup>[59]</sup> and needles <sup>[45]</sup>	Monoclinic <sup>[45]</sup>	$P2_1/n$ <sup>[46]</sup>	113 <sup>[59]</sup> - 114 $^{\circ}\text{C}$ <sup>[45]</sup>
Form IV	Needles <sup>[46]</sup>	Monoclinic <sup>[46]</sup>	$P2_1/c$ <sup>[46]</sup>	88 $^{\circ}\text{C}$ <sup>[46]</sup>
Form V	-	-	$Pbcn$ <sup>[47]</sup>	-

### 3.3.3. Stability

The stability order of the TBA polymorphs is an area of much conflict within the literature. Early work by Burger<sup>[44]</sup> stated that form I was the thermodynamically stable form at room temperature. This conclusion was drawn from the conversion of form II to form I upon stirring at room temperature in a pH 1.5 buffer solution. Hasegawa et al.<sup>[51]</sup> stated that form II is the most stable form, as form I converts to form II in ethanol. Similarly, Ikeda et al.<sup>[61]</sup> suspend form I in ethanol at 40 °C and saw the same conversion. The research presented in this thesis confirms these findings, with form I converting to form II in ethanol and methanol solutions at room temperature. Based on this work, it would suggest that form II is the thermodynamically stable form of TBA. Interestingly, early work by Burger<sup>[44]</sup> states that in solution form II is a solvated structure and upon drying converts to what is known to be form II. This perhaps lends to why ethanol and methanol both convert form I to form II. No further details about this solvated product are presented other than there is a change in crystal morphology, thin plates to needles upon desolvation. However, during this research it has been observed that form II often crystallises as a thin sheet of needles.

There is also evidence in the literature that form III is the most stable polymorph of TBA below 75°C. Rowe et al.<sup>[62]</sup> found form I converted to form III in water and that the aqueous solubility of form III was also lower than that of form I (13.03 mg/100 mL and 14.61 mg/100 mL at 37°C respectively) leading to the conclusion that form III is the thermodynamically stable polymorph. In the same work it was also stated that there was only a small free energy difference at room temperature between form I and III, possibly explaining the slow conversion rates<sup>[62]</sup>, something also noted by Burger<sup>[44]</sup>.

There was however no disputes that form IV was the least stable form of TBA, and readily converts to form II upon standing, but with the discovery of form V it is unknown as to which is the least stable.

Table 3.4 Summary of stability orders from literature

Literature cited	Most Stable form			Least stable form
<b>Burger</b> <sup>[44]</sup>	I	III	II	IV
<b>Rowe</b> <sup>[62]</sup>	III	I	Below 75 °C	
<b>Rowe</b> <sup>[62]</sup>	I	III	Above 75 °C	
<b>Kimura</b> <sup>[42]</sup>	I	III	II~IV	
<b>Hasegawa</b> <sup>[51]</sup>	II	I <sup>L</sup>	III	Below 80 °C
<b>Hasegawa</b> <sup>[51]</sup>	II	III	I <sup>L</sup> and I <sup>H</sup>	Above 80 °C
<b>Thirunahari</b> <sup>[46]</sup>	II	Below 80 °C		
<b>Thirunahari</b> <sup>[46]</sup>	I <sup>H</sup>	Above 80 °C		

A slurry of form I in methanol, ethanol, dichloromethane and acetone has been carried out at room temperature in order to determine the stability order. A summary of the results gathered using XRPD to identify polymorphic form is shown in Table 3.5, with XRPD traces in appendix section 12.3. The results confirm that form I does convert to form II in both methanol and ethanol, but in other solvents the conversion did not occur in the time frame assessed. Seeds of form II were placed into the slurries of dichloromethane to aid the possible conversion, but generated inconclusive results. Further analysis over a longer period of time needs to be conducted.

Table 3.5 Summary of form I slurry in different solvents

Solvent	Form identification by XRPD after					
	2 days	3 days	5 days	7 days	8 days	10 days
Methanol	II			II		
Ethanol	II					
Dichloromethane	I					
Seeded dichloromethane	I/II	I/II	I/II			
Acetone	I/II			I/II	I/II	I/II

### **3.4. Summary**

The three systems used within this research are all highly polymorphic but offer a range of different challenges to this work. CBZ is a thoroughly researched polymorphic system, with all experimental properties well known. The polymorphs are not conformational and the molecular structure is very rigid, which aids the molecular modelling. ROY offers a degree of conformational flexibility that adds complexity to the modelling. It also has conformational polymorphs that may not respond as well to the molecular descriptor analysis as molecules in which the structure is rigid. Finally TBA is a highly flexible molecule with relatively little published literature. This molecule more closely resembles a pharmaceutical molecule, which can then lead to the assessment and development of the methods presented in this thesis.

- [1] V. L. Himes, A. D. Mighell, W. H. De Camp, *Acta Crystallographica, Section B: Structural Science* **1981**, 37, 2242.
- [2] A. Grzesiak, M. Lang, K. Kim, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2003**, 92, 2260.
- [3] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tine, *Journal of Chemical Information and Modeling* **2006**, 46, 2030.
- [4] A. J. Cruz Cabeza, G. M. Day, W. D. Samuel Motherwell, W. Jones, *Journal of the American Chemical Society* **2006**, 128, 14466.
- [5] M. Lang, J. W. Kampf, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2002**, 91, 1186.
- [6] F. U. Krahn, J. B. Mielck, *International Journal of Pharmaceutics* **1989**, 53, 25.
- [7] M. M. J. Lowes, M. R. Caira, A. P. Lotter, J. G. Van Der Watt, *Journal of Pharmaceutical Sciences* **1987**, 76, 744.
- [8] C. Rustichelli, G. Gamberini, V. Ferioli, M. C. Gamberini, R. Ficarra, S. Tommasini, *Journal of Pharmaceutical and Biomedical Analysis* **2000**, 23, 41.
- [9] F. Tian, J. A. Zeitler, C. J. Strachan, D. J. Saville, K. C. Gordon, T. Rades, *Journal of Pharmaceutical and Biomedical Analysis* **2006**, 40, 271.
- [10] F. Tian, F. Zhang, N. Sandler, K. C. Gordon, C. M. McGoverin, C. J. Strachan, D. J. Saville, T. Rades, *European Journal of Pharmaceutics and Biopharmaceutics* **2007**, 66, 466.
- [11] E. Laine, V. Tuominen, P. Ilvessalo, P. Kahela, *International Journal of Pharmaceutics* **1984**, 20, 307.
- [12] S. G. Fleischman, S. S. Kuduva, J. A. McMahon, B. Moulton, R. D. Bailey Walsh, N. Rodriguez-Hornedo, M. J. Zaworotko, *Crystal Growth and Design* **2003**, 3, 909.
- [13] R. Hilfiker, J. Berghausen, F. Blatter, A. Burkhard, S. M. De Paul, B. Freiermuth, A. Geoffroy, U. Hofmeier, C. Marcolli, B. Siebenhaar, M. Szelagiewicz, A. Vit, M. Von Raumer, *Journal of Thermal analysis and Calorimetry* **2003**, 73, 429.
- [14] A. J. Cruz Cabeza, G. M. Day, W. D. Samuel Motherwell, W. Jones, *Crystal Growth and Design* **2006**, 6, 1858.
- [15] C. J. Strachan, S. L. Howell, T. Rades, K. C. Gordon, *Journal of Raman Spectroscopy* **2004**, 35, 401.
- [16] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Crystal Growth and Design* **2007**, 7, 100.
- [17] A. Burger, R. Ramberger, *Mikrochimica Acta* **1979**, II, 259.
- [18] R. K. Harris, P. Y. Ghi, H. Puschmann, D. C. Apperley, U. J. Griesser, R. B. Hammond, C. Ma, K. J. Roberts, G. J. Pearce, J. R. Yates, C. J. Pickard, *Organic Process Research and Development* **2005**, 9, 902.
- [19] R. J. Behme, D. Brooke, *Journal of Pharmaceutical Sciences* **1991**, 80, 986.
- [20] J. F. McCabe, *CrystEngComm* **2010**, 12, 1110.
- [21] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Chemical Communications* **2007**, 1600.
- [22] F. P. A. Fabbiani, L. T. Byrne, J. J. McKinnon, M. A. Spackman, *CrystEngComm* **2007**, 9, 728.
- [23] P. J. P. Reboul, B. Cristau, J. C. Soyfer, *Acta Crystallographica, Section B: Structural Science* **1981**, 37, 1844.

- [24] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, 95, 1918.
- [25] K. Kipourou, K. Kachrimanis, I. Nikolakakis, V. Tserki, S. Malamataris, *Journal of Pharmaceutical Sciences* **2006**, 95, 2419.
- [26] J. A. McMahon, P. Timmins, A. C. Williams, P. York, *Journal of Pharmaceutical Sciences* **1996**, 85, 1064.
- [27] Y. Li, P. S. Chow, R. Tan, B. H., S. N. Black, *Organic Process Research and Development* **2008**, 12, 264.
- [28] G. Reck, G. Dietz, *Crystal Research and Technology* **1986**, 21, 1463.
- [29] T. Gelbrich, M. B. Hursthouse, *CrystEngComm* **2006**, 8, 448.
- [30] S. Lohani, Y. Zhang, L. J. Chyall, P. Mougin-Andres, F. X. Muller, D. J. W. Grant, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1310.
- [31] A. Johnston, A. J. Florence, A. R. Kennedy, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1509.
- [32] P. Gosselin, F.-X. Lacasse, M. Preda, R. Thibert, S.-D. Clas, J. N. McMullen, *Pharmaceutical Development and Technology* **2003**, 8, 11.
- [33] D. O. Calligaro, J. Fairhurst, T. M. Hotten, N. A. Moore, D. E. Tupper, *Bioorganic and Medicinal Chemistry Letters* **1997**, 7, 25.
- [34] V. P. Shevchenko, I. Y. Nagaev, Y. V. Kuznetsov, E. V. Polunin, A. A. Zozulya, N. F. Myasoedov, *Russian Journal of Bioorganic Chemistry*, **2005**, 31, 378.
- [35] L. Yu, *Accounts of Chemical Research* **2010**, 43, 1257.
- [36] L. Yu, G. A. Stephenson, C. A. Mitchell, C. A. Bunnell, S. V. Snorek, J. J. Bowyer, T. B. Borchardt, J. G. Stowell, S. R. Bryn, *Journal of the American Chemical Society* **2000**, 122, 585.
- [37] S. Chen, L. A. Guzei, L. Yu, *Journal of the American Chemical Society* **2005**, 127, 9881.
- [38] S. Chen, H. Xi, L. Yu, *Journal of the American Chemical Society* **2005**, 127, 17439.
- [39] J. D. Dunitz, A. Gavezzotti, *Crystal Growth and Design* **2005**, 5, 2180.
- [40] G. A. Stephenson, T. B. Borchardt, S. R. Byrn, J. Bowyer, C. A. Bunnell, S. V. Snorek, L. Yu, *Journal of pharmaceutical sciences* **1995**, 84, 1385.
- [41] R. C. Thomas, G. J. Ikeda, *Journal of Medicinal Chemistry* **1966**, 9, 507.
- [42] K. Kimura, F. Hirayama, K. Uekama, *Journal of Pharmaceutical Sciences* **1999**, 88, 385.
- [43] P. Chakravarty, K. S. Alexander, A. T. Riga, K. Chatterjee, *International Journal of Pharmaceutics* **2005**, 288, 335.
- [44] A. Burger, *Scientia Pharmaceutica* **1975**, 43, 161.
- [45] J. R. Leary, S. D. Ross, M. J. K. Thomas, *Pharmaceutisch Weekblad Scientific Edition* **1981**, 3, 578.
- [46] S. Thirunahari, S. Aitipamula, P. S. Chow, R. B. H. Tan, *Journal of Pharmaceutical Sciences* **2010**, 99, 2975.
- [47] K. Nath, A. Nangia, *CrystEngComm* **2011**, 13, 47.
- [48] Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, pp. HyperChem(TM).
- [49] K. A. Nirmala, D. S. Gowda, *Acta Crystallographica, Section B: Structural Science* **1981**, 37, 1597.
- [50] J. D. Donaldson, J. R. Leary, S. D. Ross, M. J. K. Thomas, *Acta Crystallographica, Section B: Structural Science* **1981**, 37, 2245.

- [51] G. Hasegawa, T. Komasa, R. Bando, Y. Yoshihashi, E. Yonemochi, K. Fujii, H. Uekusa, K. Terada, *International Journal of Pharmaceutics* **2009**, 369, 12.
- [52] D. Giron, *Thermochimica Acta* **1995**, 248, 1.
- [53] P. Subra-Paternault, C. Roy, D. Vrel, A. Vega-Gonzalez, C. Domingo, *Journal of Crystal Growth* **2007**, 309, 76.
- [54] K. Kawakami, *Journal of Pharmaceutical Sciences* **2007**, 96, 982.
- [55] A. Sano, T. Kuriki, Y. Kawashima, H. Takeuchi, T. Niwa, *Chemical & Pharmaceutical Bulletin* **1989**, 37, 2183.
- [56] A. I. Olives, M. A. Martin, B. del Castillo, C. Barba, *Journal of Pharmaceutical and Biomedical Analysis* **1996**, 17, 1069.
- [57] H. Ueda, N. Nambu, T. Nagai, *Chemical & Pharmaceutical Bulletin* **1982**, 30, 2618.
- [58] T. Umeda, N. Ohnishi, T. Yokoyama, T. Kuroda, Y. Kita, K. Kuroda, E. Tatsumi, Y. Matsuda, *Chemical & Pharmaceutical Bulletin* **1985**, 33, 2073.
- [59] D. L. Simmons, R. J. Ranz, N. D. Gyanchandani, P. Picotte, *Canadian Journal of Pharmaceutical Sciences* **1972**, 7, 121.
- [60] Y. Sonoda, F. Hirayama, H. Arima, Y. Yamaguchi, W. Saenger, K. Uekama, *Crystal Growth and Design* **2006**, 6, 1181.
- [61] Y. Ikeda, Y. Ishihara, T. Moriwaki, E. Kato, K. Terada, *Chemical & Pharmaceutical Bulletin* **2010**, 58, 76.
- [62] E. L. Rowe, B. D. Anderson, *Journal of Pharmaceutical Sciences* **1984**, 73, 1673.

## 4. ANALYSIS METHODOLOGY

It is the aim of this chapter to demonstrate the procedures used within this research from the initial molecular modelling to the final artificial neural network (ANN). This will cover molecular modelling, calculation of descriptors, fuzzy logic and artificial neural network analysis.

### 4.1. Molecular Modelling

Initial molecular modelling was carried out at a low level of theory to perform a conformational search and optimise the geometry. The resulting structures were then taken to higher theory levels for further optimisation.

#### 4.1.1. Hyperchem<sup>TM</sup>

Hyperchem<sup>TM</sup><sup>[1]</sup> was used in this research to build the initial molecular structure of each target molecule and the solvents under investigation .

The solvent models taken forward to higher level calculations were initially geometry optimised in 3D at the highest level available in Hyperchem<sup>TM</sup><sup>[1]</sup>, PM3<sup>[2, 3]</sup>. The carbamazepine (CBZ) molecule was also geometry optimised in 3D and modelled using all available molecular mechanics and semi-empirical force fields available (MM+, AMBER, OPLS, BIO+, AM1, and PM3). Using the optimised structure in each force field, the molecules were compared with literature values for a number of bond lengths and angles<sup>[4]</sup> (Figure 4.1). The OPLS force field produced an optimised structure that was most like that reported experimentally<sup>[4]</sup>, with the results presented in Table 4.1, (complete results in appendix section 12.4).



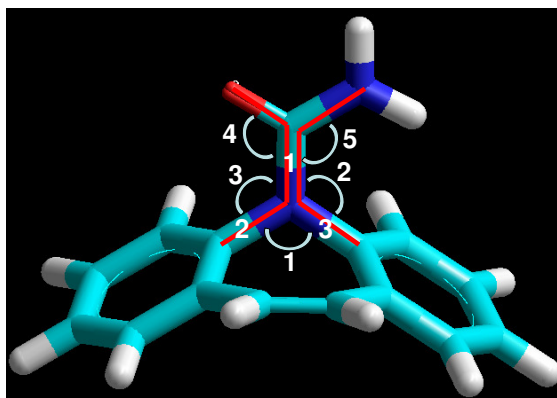


Figure 4.1 Assignments of CBZ bond lengths and angles

Table 4.1 Comparison of OPLS geometry optimised CBZ with literature values<sup>[4]</sup>

	CBZ OPLS	Literature values
<b>Bond length 1</b>	1.35 Å	1.38 Å
<b>Bond length 2</b>	1.411 Å	1.437 Å
<b>Bond length 3</b>	1.411 Å	1.434 Å
<b>Bond angle 1</b>	115.8°	116.8°
<b>Bond angle 2</b>	123.2°	121.9°
<b>Bond angle 3</b>	120.5°	120.9°
<b>Bond angle 4</b>	120.8°	121.4°
<b>Bond angle 5</b>	120.8°	116.0°
<b>Torsion angle C-O</b>	-4.1°	-9.1°
<b>Torsion angle C-N</b>	2.7°	-2.2°

Using a similar method as stated for CBZ, a 2D structure of the 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY) and tolbutamide (TBA) molecules were drawn and then optimised as 3D structures. In both cases the torsion angles within the molecules were close to zero, which deviates from experimentally determined structures<sup>[5]</sup>. A conformational search was therefore carried out upon both molecules to determine the minimum energy conformer.

In the ROY polymorphs the conformational differences between forms are well documented<sup>[5]</sup>, with the known rotatable bonds highlighted in Figure 4.2.

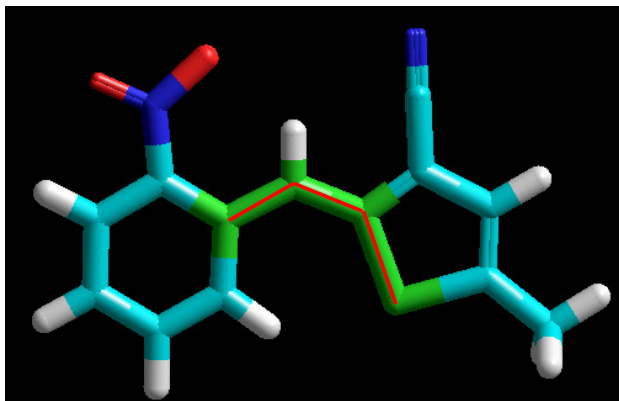


Figure 4.2 The rotatable bonds in ROY (highlighted in green with a red line)

The PM3 method found thirteen different conformations (see appendix section 12.5 for details), with a number of these closely resembling those present in known polymorphic structures. Therefore, the minimum energy PM3 structure was used in the higher level calculations.

The TBA molecule provided a further challenge to the initial geometry optimisation because it is highly flexible. For this model no experimental data was consulted in order to generate a minimum energy structure based on only basic molecular structure knowledge. Initial 3D geometry optimisation was carried out, followed by the identification of the torsion angles ( $T_x$ ) (Figure 4.3) and subsequent conformational search.

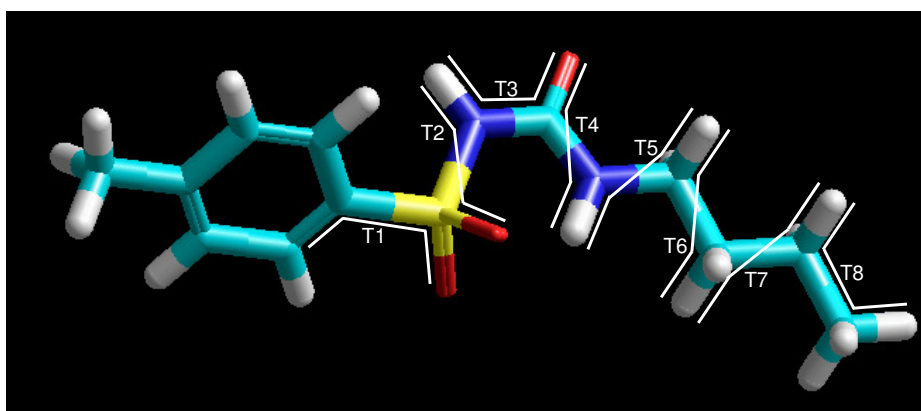


Figure 4.3 Highlighted torsion angles ( $T_x$ ) within the TBA molecule that were subjected to a conformational search

The OPLS<sup>[3]</sup> method generated 339 conformers, and the minimum energy structure of these conformers was taken forward for further optimisation (Electronic appendix, chapter 4, 4.1).

A flow chart of the Hyperchem<sup>TM</sup><sup>[1]</sup> modelling procedure is shown in Figure 4.4.

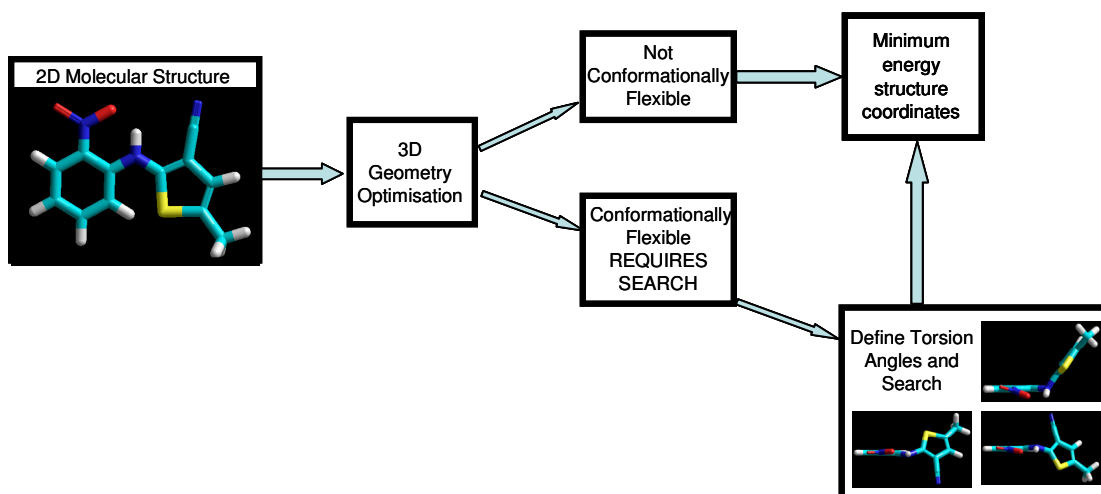


Figure 4.4 Summary of the flow of work carried out in Hyperchem<sup>TM</sup><sup>[1]</sup>

#### 4.1.2. Gaussian 03

Using the cartesian coordinates generated in Hyperchem<sup>TM</sup><sup>[1]</sup>, a gas phase optimised structure was generated for each solvent, CBZ, ROY and TBA molecule using the parameters shown in Table 4.2.

The optimised structure was then subject to polarisable continuum model (PCM) calculations that modify the geometry of the molecule in response to solvent interaction (See section 2.6.2.3). Gaussian 03<sup>[6]</sup> has predefined PCM models for many common solvents. However, it is possible for the user to define other solvents if required. In order to do this, knowledge of the static dielectric constant (EPS), the dielectric constant at infinite frequency (EPSINF), the solvent radius (RSOLV) and density (DENSITY) are required (See section 2.6.2.3). Example input parameters for both the predefined and user defined PCM methods are highlighted in Table 4.2.

Table 4.2 Input parameters for Gaussian calculations

Calculation type	Input parameters
Gas phase geometry optimisation	#p B3LYP/6-31G opt=tight freq scf=(tight,save,maxcycle=256) int=ultrafine pop=full ginput iop(6/7=3)
Predefined PCM force field optimisation (Acetone example)	#p B3LYP/6-31G* opt freq scf=(save,maxcycle=256) scrf=(pcm,read,solvent=acetone) pop=reg ginput iop(6/7=3)
User defined PCM force field optimisation (n-Butanol example)	#p B3LYP/6-31G* opt freq scf=(Tight,save,maxcycle=256) scrf=(pcm,read) pop=reg ginput iop(6/7=3) <i>Below Cartesian coordinates</i> EPS=17.84 EPSINF=1.957 RSOLV=2.669 DENSITY=0.00658

An example input file can be found in Electronic Appendix, Chapter 4, file 4.2 for the predefined and user defined PCM calculations.

When each solvent is modelled, the procedure is the same but it is only modelled within its own PCM force field. This differs to the CBZ, ROY and TBA calculations as they are modelled in each different solvent force field. By generating an optimised structure of the solvent, CBZ, ROY and TBA in solvent force fields, molecular descriptors can be calculated. A summary of the Gaussian 03<sup>[6]</sup> work is presented in Figure 4.5.

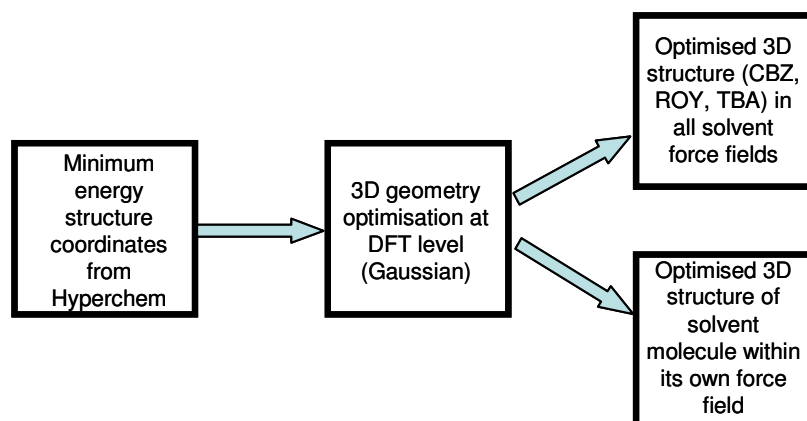


Figure 4.5 Summary of Gaussian 03<sup>[6]</sup> work flow

## 4.2. Bulk and Molecular Descriptors

The optimised structures were presented to descriptor software from the book “Molecular Descriptors in QSAR/QSPR”,<sup>[7]</sup> in .mol file format. 78 descriptors (22 of which were unique) were calculated for each optimised CBZ, ROY and TBA structure in the solvent force field. A further 78 descriptors were calculated for each solvent in their own specific force field. To increase the number of molecular descriptors available for analysis, a second program was used to calculate further values, MOE<sup>[8]</sup>. MOE<sup>[8]</sup> was used to calculate only descriptors for the polymorphic molecule in the solvent force fields due to the volume of data it generated, with the optimised structures being presented as .pdb files.

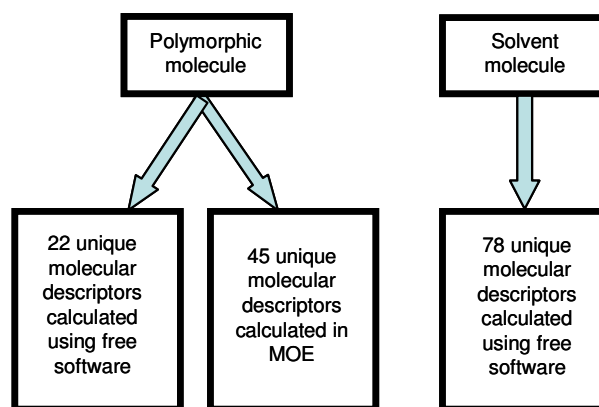


Figure 4.6 Summary of the molecular descriptor calculation process

As well as molecular descriptors, a number of bulk properties of the solvents used in the crystallisations have been included. These were primarily taken from the literature (see Electronic Appendix, Chapter 4, file 4.3) and from previous work carried out in-house at AstraZeneca.

### 4.2.1. Descriptor Reduction Methods

If all 167 descriptors were used as inputs in INForm<sup>[9]</sup> and the FormRules<sup>[10]</sup> software, not only would no specific descriptors be highlighted as useful, the network would also overtrain. When too many data are presented to a network it loses its ability to predict outside of the data seen. The network essentially learns the data it has been given and does not generalise, leading to poor predictive ability<sup>[11, 12]</sup>.

De Matas et al.<sup>[13]</sup> suggested that the input data used in the training should be at least three or four times larger than the number of descriptors presented, therefore the total number of descriptors needed to be reduced. On average, twelve descriptors were used to train the networks, which fit within the recommended strategy and also allowed meaning to be gathered from the descriptors present. A number of methods were utilised to reduce the descriptor number, which will be discussed in detail in chapters 5, 6 and 7. These were a manual search of the descriptor space, linear correlations, partial least squares (PLS) and principal component analysis (PCA).

### 4.3. Artificial Neural Network (ANN) Input File

The input file for the ANN needs to be created as a text file and contains the experimental information, the bulk and molecular descriptors that correspond to the solvent used in a particular experiment and also the experimental outcome, i.e. polymorphic form crystallised (Table 4.3). For training to occur successfully it is important to present the network with all the possible outcomes of the experiment, with values between 0 and 1. It is equally essential to inform the network that a set of parameters does not lead to one output, as the parameters that do, and based on this the experimental results were assigned a weighting. The weightings for each sample total 1 and involve all the possible CBZ forms that could be crystallised.

Table 4.3 Examples of ANN input data

INPUT				OUTPUT				
Solvent	Rate	Temp	Bulk and Molecular descriptors	Form I	Form II	Form III	Dihydrate	Solvate
EtOH	5	25	167 values	0	0	1	0	0
AcCN	15	25	167 values	0	0.5	0.5	0	0

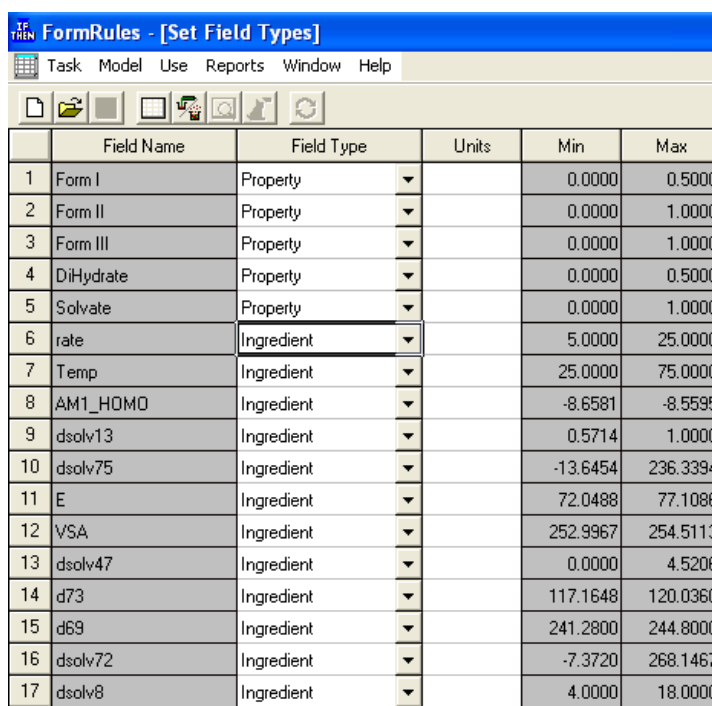
A large spreadsheet of all the experimental results combined with 167 different bulk and molecular descriptors has been created and can be found in Electronic Appendix, Chapter 4, file 4.4.

## 4.4. FormRules Analysis

FormRules<sup>[10]</sup> uses fuzzy logic to create “IF, and THEN” rules that can be used to identify which descriptors have an effect on the crystallisation of different polymorphic forms.

An example dataset using ten descriptors (AM1\_HOMO, dsolv13, dsolv75, E, VSA, dsolv47, d73, d69, dsolv72 and dsolv8 (detailed in appendix section 12.2), rate and temperature will be used to demonstrate the FormRules<sup>[10]</sup> analysis methodology.

The text file of these inputs and outputs is loaded into FormRules<sup>[10]</sup> and the user selects those columns that are the ingredients (inputs) and those that are properties (outputs) shown in Figure 4.7.



The screenshot shows the 'FormRules - [Set Field Types]' window. It has a menu bar with 'Task', 'Model', 'Use', 'Reports', 'Window', and 'Help'. Below the menu is a toolbar with icons for file operations and a refresh button. The main area is a table with 6 columns: 'Field Name', 'Field Type', 'Units', 'Min', and 'Max'. The table lists 17 fields. Fields 1-5 are 'Form I', 'Form II', 'Form III', 'DiHydrate', and 'Solvate', all with 'Property' as their field type. Fields 6-17 are 'rate', 'Temp', 'AM1\_HOMO', 'dsolv13', 'dsolv75', 'E', 'VSA', 'dsolv47', 'd73', 'd69', 'dsolv72', and 'dsolv8', all with 'Ingredient' as their field type. The 'Units' column is empty for all fields. The 'Min' and 'Max' columns contain numerical values for each field.

	Field Name	Field Type	Units	Min	Max
1	Form I	Property		0.0000	0.5000
2	Form II	Property		0.0000	1.0000
3	Form III	Property		0.0000	1.0000
4	DiHydrate	Property		0.0000	0.5000
5	Solvate	Property		0.0000	1.0000
6	rate	Ingredient		5.0000	25.0000
7	Temp	Ingredient		25.0000	75.0000
8	AM1_HOMO	Ingredient		-8.6581	-8.5595
9	dsolv13	Ingredient		0.5714	1.0000
10	dsolv75	Ingredient		-13.6454	236.3394
11	E	Ingredient		72.0488	77.1086
12	VSA	Ingredient		252.9967	254.5113
13	dsolv47	Ingredient		0.0000	4.5206
14	d73	Ingredient		117.1648	120.0360
15	d69	Ingredient		241.2800	244.8000
16	dsolv72	Ingredient		-7.3720	268.1467
17	dsolv8	Ingredient		4.0000	18.0000

Figure 4.7 Screen shot of how the inputs and outputs are identified

At this stage the data is ready to train and the user is asked to select the model to use in the training. Structure Risk Minimisation (SRM) is the default setting and has been used in this research. This method does not require any validation data to be used and therefore creates the rules based on all of the experimental information. SRM uses the bias (no. of free parameters) and variance (training data error) to look at the prediction error, and is suggested for data sets of this size<sup>[14]</sup>.

When the training is complete there are a number of tools that can be utilised to look at the results. The most straightforward method is to look at the list of rules generated in table form (Table 4.4). Each output (polymorphic outcome) has its own set of rules; in some cases a number of different models are produced. Colour coding is used in the output spreadsheet to guide the eye to the rules that have the largest positive contribution to a particular output (blue), or the smallest positive contribution to the output (red)<sup>[14]</sup>. Each rule is also given a confidence level (number in brackets after the prediction), with values between 0 and 1, with 1 representing the highest level of confidence.

Table 4.4 Example of the rules generated for analysed data in FormRules<sup>[10]</sup>

Rules generated for the example set of inputs			
--- Rules for property Form I ---			
	IF d73 is LOW	THEN Form I is	LOW (1.00)
	IF d73 is HIGH	THEN Form I is	LOW (0.96)
--- Rules for property Form II ---			
SubModel:1	IF d73 is LOW	THEN Form II is	HIGH (1.00)
	IF d73 is HIGH	THEN Form II is	LOW (1.00)
SubModel:2	IF dsolv13 is LOW	THEN Form II is	LOW (1.00)
	IF dsolv13 is HIGH	THEN Form II is	HIGH (1.00)
SubModel:3	IF rate is LOW	THEN Form II is	LOW (1.00)
	IF rate is HIGH	THEN Form II is	HIGH (0.72)
SubModel:4	IF d69 is LOW	THEN Form II is	HIGH (1.00)
	IF d69 is HIGH	THEN Form II is	LOW (1.00)
--- Rules for property Form III ---			
SubModel:1	IF dsolv47 is LOW	THEN Form III is	HIGH (1.00)
	IF dsolv47 is HIGH	THEN Form III is	LOW (1.00)
SubModel:2	IF dsolv13 is LOW	THEN Form III is	HIGH (1.00)
	IF dsolv13 is MID	THEN Form III is	LOW (1.00)
	IF dsolv13 is HIGH	THEN Form III is	LOW (1.00)
--- Rules for property Form III continued ---			
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (1.00)



Rules generated for the example set of inputs -continued			
SubModel:4	IF rate is HIGH	THEN Form III is	LOW (0.55)
	IF dsolv75 is LOW	THEN Form III is	LOW (1.00)
	IF dsolv75 is HIGH	THEN Form III is	HIGH (1.00)
--- Rules for property DiHydrate ---			
	IF d73 is LOW AND dsolv47 is LOW AND Temp is LOW	THEN DiHydrate is	LOW (1.00)
	IF d73 is LOW AND dsolv47 is LOW AND Temp is HIGH	THEN DiHydrate is	HIGH (0.55)
	IF d73 is LOW AND dsolv47 is HIGH AND Temp is LOW	THEN DiHydrate is	LOW (0.90)
	IF d73 is LOW AND dsolv47 is HIGH AND Temp is HIGH	THEN DiHydrate is	LOW (1.00)
	IF d73 is HIGH AND dsolv47 is LOW AND Temp is LOW	THEN DiHydrate is	LOW (0.57)
	IF d73 is HIGH AND dsolv47 is LOW AND Temp is HIGH	THEN DiHydrate is	LOW (1.00)
	IF d73 is HIGH AND dsolv47 is HIGH AND Temp is LOW	THEN DiHydrate is	LOW (1.00)
	IF d73 is HIGH AND dsolv47 is HIGH AND Temp is HIGH	THEN DiHydrate is	LOW (0.88)
--- Rules for property Solvate ---			
	IF E is LOW AND dsolv13 is LOW	THEN Solvate is	LOW (1.00)
	IF E is LOW AND dsolv13 is MID	THEN Solvate is	LOW (0.74)
	IF E is LOW AND dsolv13 is HIGH	THEN Solvate is	LOW (1.00)
	IF E is MID AND dsolv13 is LOW	THEN Solvate is	LOW (1.00)
	IF E is MID AND dsolv13 is MID	THEN Solvate is	LOW (1.00)
	IF E is MID AND dsolv13 is HIGH	THEN Solvate is	LOW (1.00)
	IF E is HIGH AND dsolv13 is LOW	THEN Solvate is	LOW (1.00)
	IF E is HIGH AND dsolv13 is MID	THEN Solvate is	HIGH (1.00)
	IF E is HIGH AND dsolv13 is HIGH	THEN Solvate is	LOW (1.00)

The descriptors that feature in the rules for different outputs can also be graphically represented (Figure 4.8). This functionality has not been used extensively within this research because it does not add any further understanding to the rules.

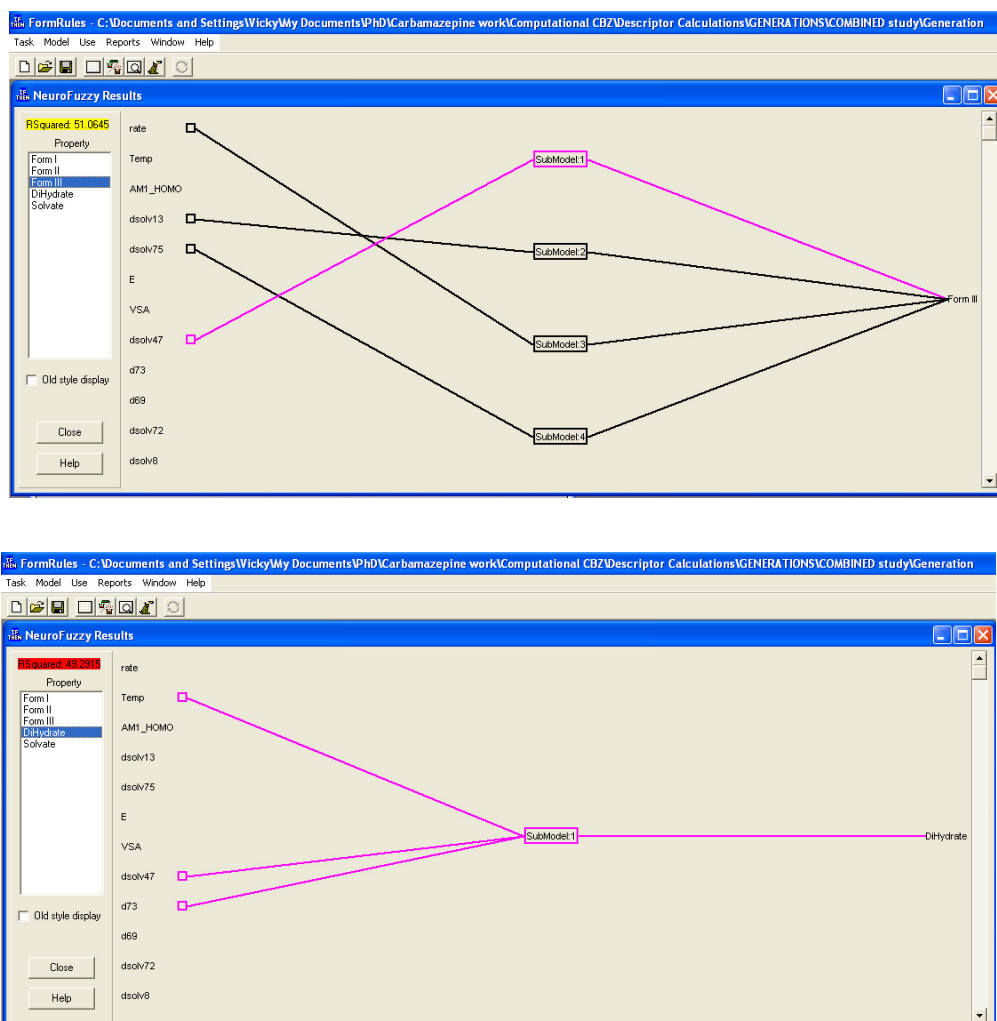


Figure 4.8 Two examples of the graphical representation of the rules for different output predictions

FormRules<sup>[10]</sup> also generates model statistics that are useful for an immediate overview of how well the model has trained. ANOVA (analysis of variance) statistics are calculated and use the following equation for the non-linear analysis of the variables, with  $\hat{y}$  representing the predicted value and  $\bar{y}$  the average of the dependent variables (Equation 4.1)<sup>[14]</sup>.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y})^2 - 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(\hat{y}_i - y_i) \quad \text{Equation 4.1}$$

This equation can also be represented in a simpler form, with  $SST$  representing the total,  $SSR$  the model and  $SSE$  is the error sum of squares<sup>[14]</sup>.

$$SST = SSR + SSE$$

Equation 4.2

The details provided on the model statistics spreadsheet have been generated in the following way (Table 4.5). The degree of freedom ( $k$ ) is equal to the number of weights and biases throughout the network. Each hidden and output layer has a bias and the hidden layers have a variable number of nodes, each with a weight.<sup>[14]</sup> Occasionally the degrees of freedom for the error can be negative, which most often occurs when there is very little data for a specific output,<sup>[14]</sup> with  $n$  representing the number of data records. The f ratio can be used to test that “the variation in the dependent variable arises from random fluctuations independent of the value of the independent (input) variable.”<sup>[14]</sup>

Table 4.5 The methods used to generate the data in the model statistics table. Taken from FormRules manual<sup>[14]</sup>

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Computed f ratio
<b>Model</b>	$SSR$	$K$ (number of weights and biases)	$SSR/k$	$(SSR/k)/(SSE/(n-k-1))$
<b>Error</b>	$SSE$	$n-k-1$	$SSE/(n-k-1)$	
<b>Total</b>	$SST$	$n-1$		

$R^2$  values are also calculated for the prediction of each form using Equation 4.3<sup>[14, 15]</sup>.

$$R^2 = \left( 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right) \times 100 \quad \text{or} \quad R^2 = \left( 1 - \frac{SSE}{SST} \right) \times 100 \quad \text{Equation 4.3}$$

All of these values are combined into another spreadsheet with a typical set of results shown in Table 4.6.

It should be noted that occasionally the  $R^2$  value is negative, highlighting the models inability to predict a given output. From Equation 4.3, if the total sum of errors ( $SSE$ ) is larger than the total variance of the data ( $SST$ ) a negative value is calculated.

Table 4.6 Model statistics produced by FormRules<sup>[10]</sup>

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	f ratio	Covariance term	Sum of Errors	Train Set R <sup>2</sup> (%)
<b>Form I</b>							<b>3</b>
Model	0.007	2	0.004	1.186			
Error	0.249	85	0.003				
Total	0.256	87			$1.93 \times 10^{-07}$	$-1.78 \times 10^{-06}$	
<b>Form II</b>							<b>54</b>
Model	8.604	5	1.721	19.236			
Error	7.336	82	0.090				
Total	15.940	87			$1.62 \times 10^{-05}$	$1.67 \times 10^{-05}$	
<b>Form III</b>							<b>51</b>
Model	8.875	6	1.479	14.087			
Error	8.505	81	0.105				
Total	17.380	87			$2.70 \times 10^{-05}$	$-4.12 \times 10^{-06}$	
<b>Dihydrate</b>							<b>49</b>
Model	0.155	8	0.019	9.599			
Error	0.159	79	0.002				
Total	0.314	87			$5.51 \times 10^{-06}$	$8.16 \times 10^{-06}$	
<b>Solvate</b>							<b>97</b>
Model	7.119	9	0.791	281.865			
Error	0.219	78	0.003				
Total	7.341	87			0.0033	$9.09 \times 10^{-05}$	

## 4.5. INForm Analysis

The input file is presented in the same way as in the FormRules<sup>[10]</sup> analysis, but in order to determine whether the network built has predictive capabilities, a test set has to be generated. This test set is used during the training of the network to assess whether the network is generalising correctly. Within the INForm<sup>[9]</sup> software there are a number of ways to select a test set, the method utilised in this research takes 15 % of the data (which is 13 rows of data) and uses what is known as the smart selection method. This method randomly selects rows for the test set, but avoids the extreme results on the edge of the experimental area<sup>[16]</sup>. A note of the test set values in every analysis was made, so the range of outputs tested could be monitored.

In this example, INForm<sup>[9]</sup> suggested a network with one hidden layer that contains 4 nodes, using the default transfer functions of the asymmetric sigmoid and a linear function. Throughout this analysis the default networks have been used to run each set of descriptors and in future, different approaches should be considered.

Table 4.7 Summary of the network architecture used in this analysis

Steps	Chosen method
Selection of the test set -	15 % of data using Smart Selection
Training parameters -	RPROP type of back propagation selected
Hidden layers and nodes -	1 hidden layer. Number of nodes is varied by the software depending on the input data. Their default settings are always used
Hidden layer transfer function -	Asymmetric sigmoid
Output transfer function -	Linear
Outputs trained -	Separately

Similarly to FormRules<sup>[10]</sup> there are a number of different ways to view the results generated in INForm<sup>[9]</sup>. The table of model statistics presented to the user is similar to that in FormRules<sup>[10]</sup>, but with  $R^2$  values for both the training and test sets. This allows the user to see immediately whether the training has been successful.

Table 4.8 Model statistics generated in INForm<sup>[9]</sup>

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F ratio	Covariance term	Sum of Errors	Test Set R <sup>2</sup> (%)	Train Set R <sup>2</sup> (%)
<b>Form I</b>							95	100
Model	0.216	57	0.004	4.807				
Error	0.013	17	0.001					
Total	0.255	74			0.026	-0.030		
<b>Form II</b>							80	59
Model	10.335	57	0.181	1.204				
Error	2.560	17	0.151					
Total	12.886	74			-0.009	0.008		
<b>Form III</b>							84	39
Model	11.955	57	0.210	1.504				
Error	2.370	17	0.139					
Total	14.331	74			0.006	-0.005		
<b>Dihydrate</b>							95	100
Model	0.267	57	0.005	4.911				
Error	0.016	17	0.001					
Total	0.313	74			0.030	-0.025		
<b>Solvate</b>							99	99
Model	6.159	57	0.108	18.520				
Error	0.099	17	0.006					
Total	6.413	74			0.154	-0.021		

The data in Table 4.8 indicate that the training has performed very well and the network has the ability to predict whether form I, dihydrate or solvates are formed successfully from the test set. However, the prediction of forms II and III is less effective from the test set and suggests that perhaps these are not the most informative descriptors.

## **4.6. Combined FormRules and INForm analysis**

A number of different methods have been determined to analyse the results and data produced in INForm<sup>[9]</sup> and FormRules<sup>[10]</sup>. One method is a more rapid method of using values generated by the software, the other is a more detailed approach.

### **4.6.1. Rapid Analysis of Combined Data**

One method of using the results from INForm<sup>[9]</sup> and FormRules<sup>[10]</sup> was to calculate an average  $R^2$  value. This generated an instant value for the success of a set of inputs at predicting the polymorphic outcome. This method has been often used within this research as an immediate identifier for successful networks and thus informative descriptors.

### **4.6.2. Detailed Analysis of Combined Data using the 3D Explorer**

A second method has also been employed to analyse the results from both pieces of software using the 3D explorer facility within INForm<sup>[9]</sup>. The 3D explorer application generates 3D plots of different descriptors versus polymorphic outputs. The plots use the predicted values from the network and can be utilised to observe general trends in the data. All descriptors in a set may be plotted against one and another, with the remaining descriptor values adjustable, for observation of the effect on the polymorphic prediction. As there are many possible combinations of descriptor values it is impractical to analyse every one, therefore the rules generated in FormRules<sup>[10]</sup> serve as a guide as to which descriptors to initially plot.

Before a plot can be generated, a set of descriptor values need to be entered as a starting point. This is done by asking the network to find a set of descriptor values that gives a high form III output (Figure 4.9). The user could ask for any of the available outputs, but form III was chosen as it occurs most frequently in the data set.

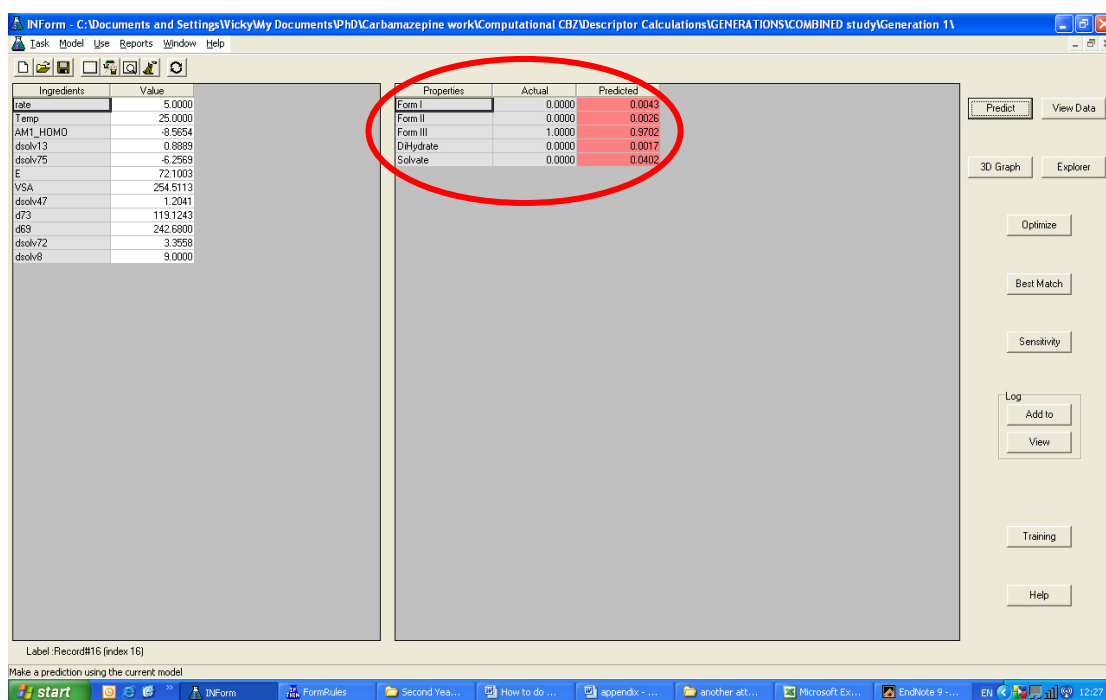


Figure 4.9 Screen shot of how the starting descriptor values are selected

Figure 4.9 shows a set of descriptor values that lead to a form III prediction. The record number (16) is shown in the bottom corner so these starting values can always be found at a later stage.

In the 3D explorer feature in INForm<sup>[9]</sup> the rules that have been previously generated are used to interrogate the networks predictive capabilities. Two examples are going to be used to highlight how the analysis is carried out. The first shows a case where the rule is very good and other descriptors have little effect on the prediction, the second shows a different result.

Table 4.9 A rule for form II prediction from FormRules<sup>[10]</sup>

Form II rule	SubModel:1	
IF d73 is LOW	THEN Form II is	HIGH (1.00)
IF d73 is HIGH	THEN Form II is	LOW (1.00)

In this example, d73 is the descriptor that was highlighted to affect the prediction of form II. The three axis of the plot need to be selected, two are known from the rule and the third needs to be decided upon. All the available descriptors are plotted against d73 and form II and the descriptor that conforms to the rule most successfully is used in the analysis. The figure below shows d73 being plot against form II and rate.



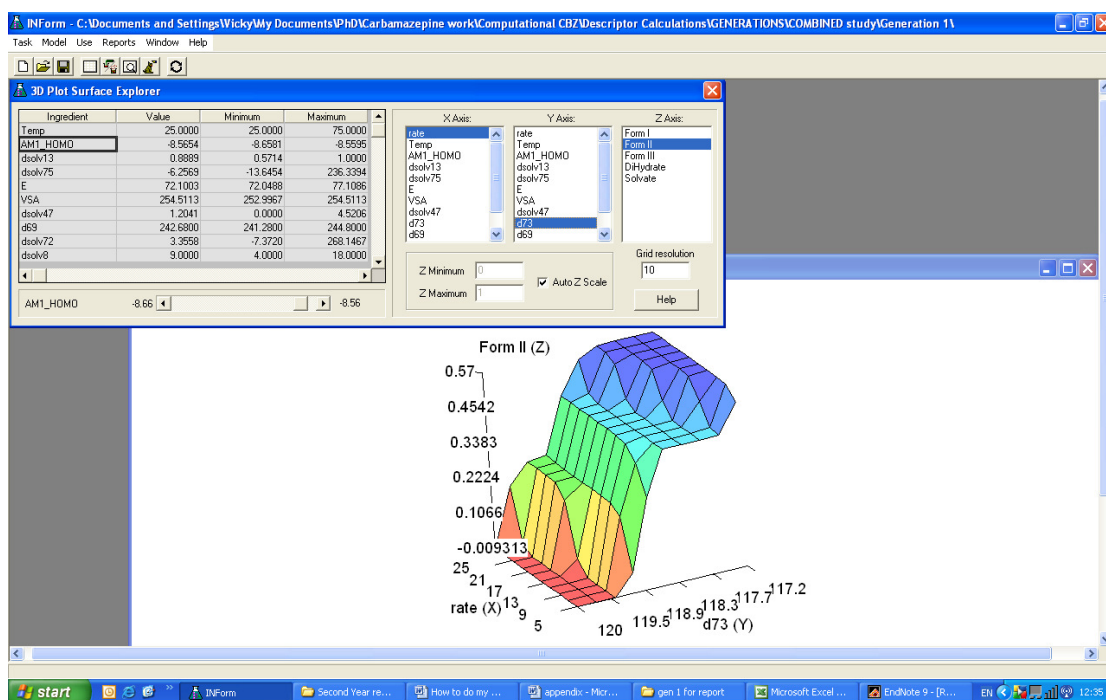


Figure 4.10 Screen shot from the 3D explorer feature showing a plot of d73 and rate against form II. The blue colouring on the plot indicates a high prediction value, and the red region is a low predictive value

Once a plot has been generated that conforms to the rule, the other descriptor values are altered one by one to see if there is any affect on the plots surface shape or prediction values. All other descriptor values are held the same, which in this case are the values from record number 16. Figure 4.11 shows the panel used to change the descriptor values that contain a list of descriptors and their maximum, minimum and current values.

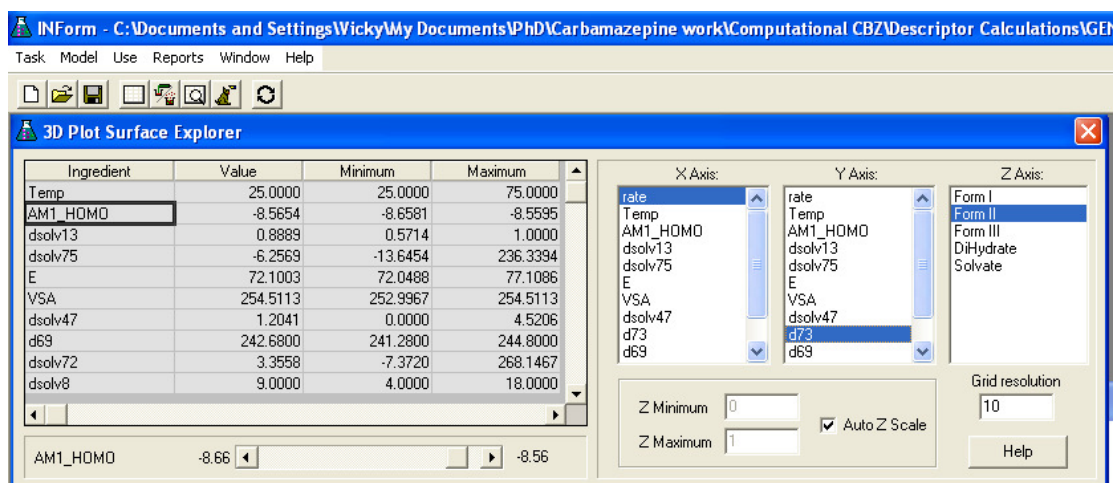


Figure 4.11 The panel used to change the axis and descriptor values in the 3D explorer

In turn, each descriptor is changed from its maximum to minimum value. Occasionally a mid range value is recorded when a large change in shape or prediction has occurred.

In Figure 4.12 the descriptor dsolv75 has been changed to highlight that in some cases other descriptors in the set do not affect the rule or prediction. When this is the case, only the axis descriptors are taken forward for further analysis.

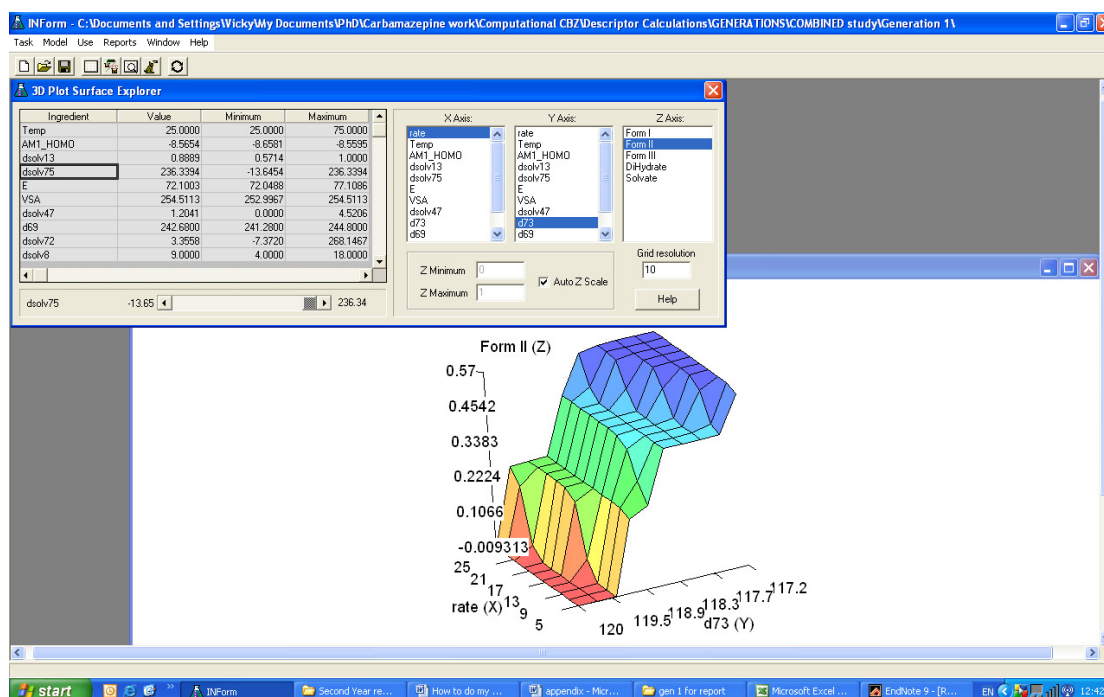


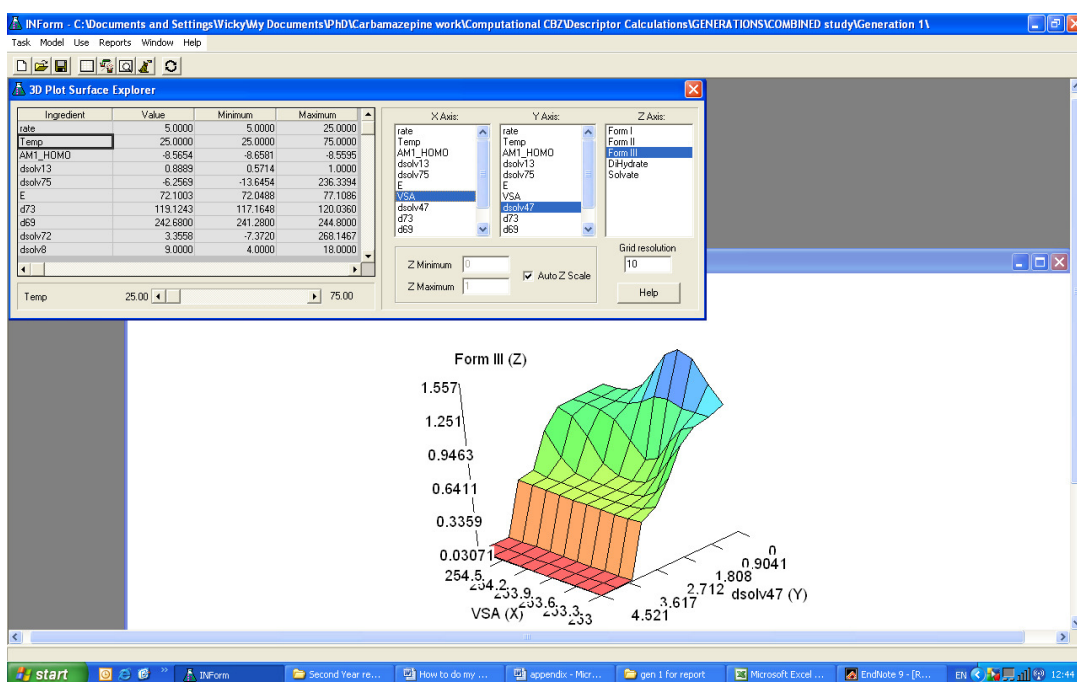
Figure 4.12 The plot of d73, rate and form II when dsolv75 is at its highest value (initially it was a very low value)

In a second example, form III is being predicted by using dsolv47 as shown in Table 4.10.

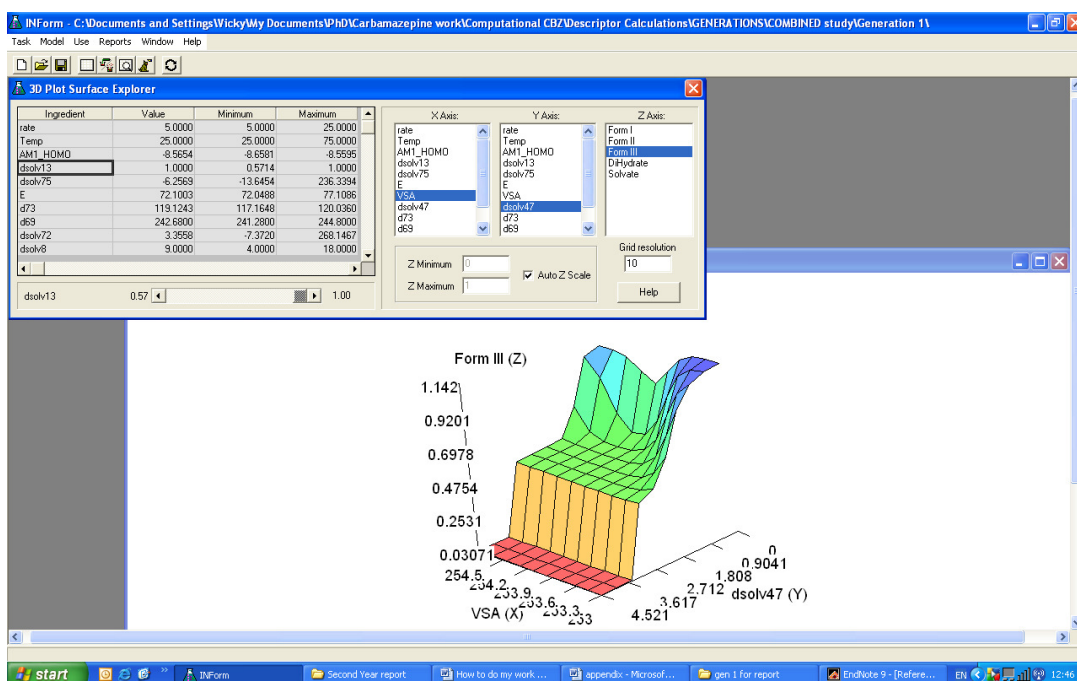
Table 4.10 Summary of the rule to predict form III

Form III rule	SubModel:1
IF dsolv47 is LOW	THEN Form III is HIGH (1.00)
IF dsolv47 is HIGH	THEN Form III is LOW (1.00)

VSA was chosen as the third axis (Figure 4.13) as it most closely follows the rule generated in FormRules<sup>[10]</sup>, with the other descriptors being altered as mentioned previously.



Within this set of descriptors there were a number of small changes in plot shape, but when dsolv13 is maximised and minimised the change is clear. The initial value of dsolv13 was mid range (its influence on the shape is demonstrated in Figure 4.13). When dsolv13 is at its maximum (Figure 4.14) and minimum (Figure 4.15) values, the plot shape changes, as does the value of the prediction of form III.



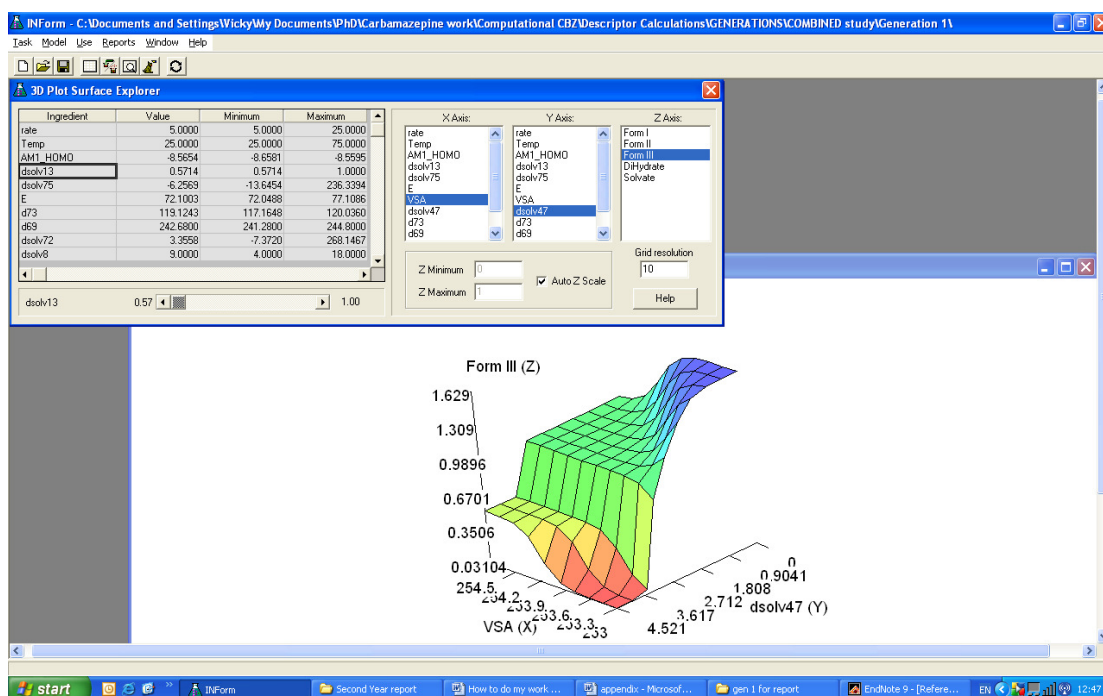


Figure 4.15 Shows the difference in plot shape when dsolv13 is set to its minimum value

The differences in plots that have occurred due to a change in the dsolv13 value are more easily compared in Figure 4.16. Although in this case the changes to the plot are not drastic, it can clearly be seen that as dsolv13 decreases, form III is the more favoured output.

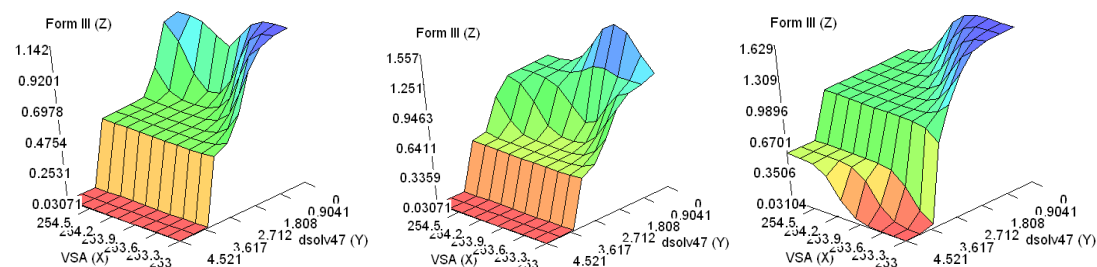


Figure 4.16 The differences in plots when dsolv13 is changed. From left to right: Maximum value, original value (mid-range) and minimum value

When results like this are observed the descriptors taken for further analysis are dsolv47, VSA and dsolv13. It is important to note that it is not only descriptors that improve the prediction of a certain form, but also those that have a negative impact that are taken forward. A negative impact on prediction would suggest it is relevant to the prediction of a different form and therefore still an important descriptor.

## 4.7. Summary of Analysis Methodology

This chapter has highlighted the molecular modelling of the polymorphic systems, from the conformational search in Hyperchem<sup>TM</sup><sup>[1]</sup> to the solvent force field optimisations in Gaussian 03<sup>[6]</sup>. A discussion of how to calculate molecular descriptors and how to reduce them in number has been made. The creation of the input file and subsequent FormRules<sup>[10]</sup> and INForm<sup>[9]</sup> analysis has been demonstrated. Figure 4.17 summarises the overall structure of the analysis.

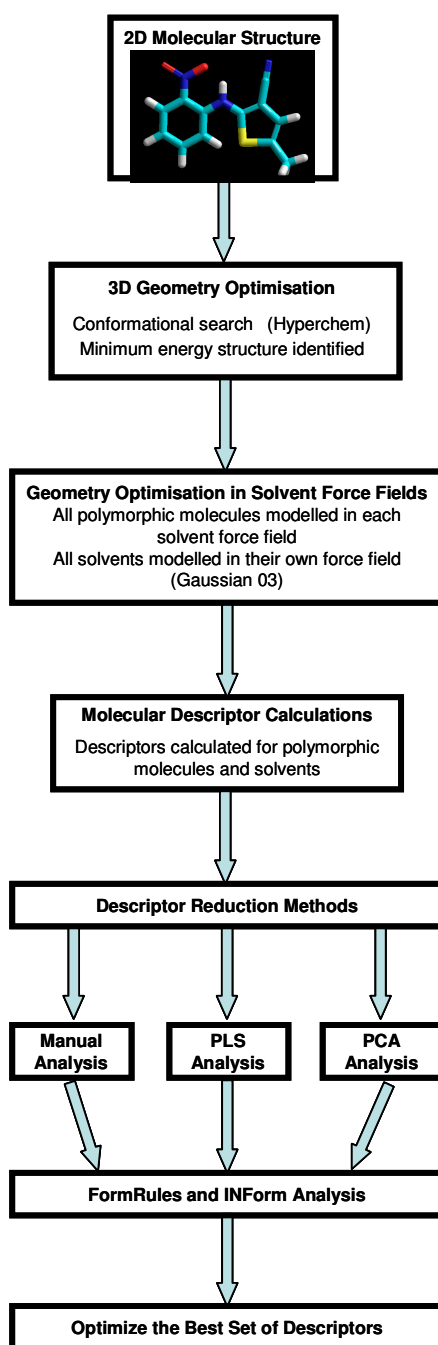


Figure 4.17 Summary of the overall analysis process

- [1] Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, pp. HyperChem(TM).
- [2] F. Jensen, *Introduction to Computational Chemistry*, 1st ed., John Wiley & Sons Ltd., Chichester, **1999**.
- [3] C. J. Cramer, *Essential of Computational Chemistry, Theories and Models*, 1st ed., John Wiley & Sons Ltd., Chichester, **2002**.
- [4] P. J. P. Reboul, B. Cristau, J. C. Soyfer, *Acta Crystallographica, Section B: Structural Science* **1981**, 37, 1844.
- [5] L. Yu, *Accounts of Chemical Research* **2010**, 43, 1257.
- [6] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople, Gaussian, Inc. Wallingford CT, **2004**.
- [7] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [8] MOE, Chemical Computing Group, p. Molecular Operating Environment.
- [9] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [10] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [11] C. M. Handley, P. L. A. Popelier, *Journal of Physical Chemistry A* **2010**, 114, 3371–3383.
- [12] D. J. Livingstone, D. W. Salt, *Bioorganic and Medicinal Chemistry Letters* **1992**, 2, 213.
- [13] M. de Matas, Q. Shao, V. L. Silkstone, H. Chrystyn, *Journal of Pharmaceutical Sciences* **2007**, 96, 3293.
- [14] *FormRules: Formulating Rules software manual v3.3*, Intelligensys Ltd., **2007**.
- [15] Q. Shao, R. C. Rowe, P. York, *European Journal of Pharmaceutical Sciences* **2006**, 28, 394.
- [16] *INForm: Intelligent Formulation manual v3.7*, Intelligensys Ltd., **2009**.

## 5. RESULTS AND DISCUSSION – MANUAL ANALYSIS

As stated previously, the number of descriptors needs to be reduced in order to create a predictive model that can generalise the data and not overtrain. It also allows meaning to be placed upon molecular descriptors that lead to successful polymorphic predictions. This chapter presents the complete data set and linear correlation data reduction methods undertaken in this research that try to determine an optimised predictive model.

### 5.1. Complete Dataset Analysis

Initially all 167 descriptors were given a random number, placed in ascending order and the first ten placed into sets. This was repeated until all descriptors were represented, generating 22 sets in which some descriptors were represented more than once. These sets of ten descriptors, combined with rate and temperature values, were then analysed using INForm<sup>[1]</sup> and FormRules<sup>[2]</sup>. The analysis technique used in this research was the detailed method (outlined in section 4.6.2) whereby the rules generated in FormRules<sup>[2]</sup> were used in conjunction with the predictions made in INForm<sup>[1]</sup>. This method was not time effective and therefore this data analysis technique did not progress.

A more rapid analysis technique was sought and found by using only the average  $R^2$  values from both INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> (discussed in section 4.6.1). Figure 5.1 summarises the method used to generate the sets of descriptors for the rapid analysis, which is discussed in the following sections.

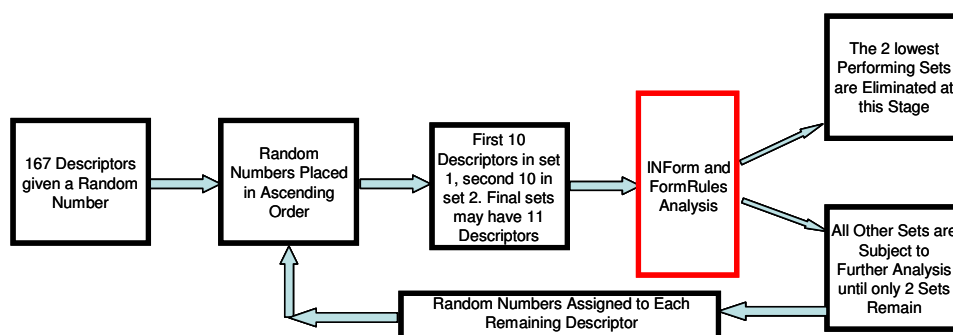


Figure 5.1 Summary of how the descriptor sets were created

16 descriptor sets were generated and were used as the starting point in both the INForm,<sup>[1]</sup> FormRules<sup>[2]</sup> and combined analysis (for details see appendix section 12.6). After the first generation of analysis the two least successful sets of descriptors at predicting the polymorphic outcome of the crystallisation experiments were discarded. The remaining descriptors were then assigned a new random number, reordered and placed into new sets. This meant that different sets of descriptors were analysed each time, but a path could be traced through each generation if certain descriptors often featured in successful sets. At each generation the two least successful sets were discarded, and the remaining descriptors randomised and put into new sets. Figure 5.2 highlights the number of sets of descriptors in each generation of analysis.

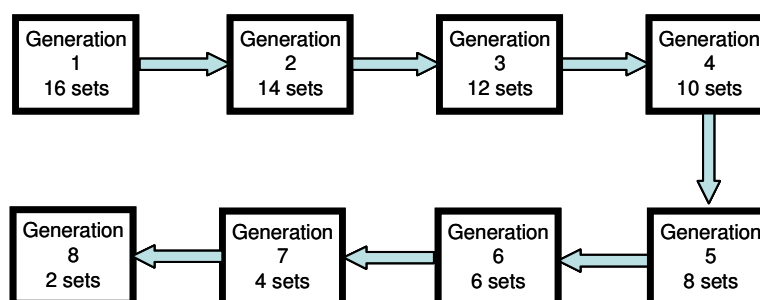


Figure 5.2 Summary of the number of descriptor sets within each generation of analysis



### 5.1.1. Complete Dataset Analysis – FormRules

Analysis of all 167 descriptors was carried out using the method outlined above using only the average  $R^2$  values generated in FormRules<sup>[2]</sup>. A full table of results can be found in Electronic Appendix, Chapter 5, file 5.1, with the final set, the first and second most successful sets (Table 5.1) presented for further analysis.

Table 5.1 The final set and first and second most successful descriptor sets in FormRules analysis

Sets	Descriptors	FormRules average $R^2$ (%)
Final Set (Gen 8, Set 2gg)	Dsolv4, dsolv18, dsolv49, dsolv50, dsolv65, dsolv71, d73, MNDO_HF, Vol, vapour density, polarisability parameter	78.31
Best Overall Set (Gen 1, Set 2)	Dsolv26, dsolv34, dsolv47, dsolv63, dsolv68, dsolv69, dsolv71, dsolv74, dsolv79, RMM,	81.26
Second Best Set (Gen 1, Set 5)	Dsolv24, dsolv32, dsolv62, d67, d71, d77, pmIZ, AM1_HOMO, PM3_LUMO, dielectric constant	80.94

The two most successful sets were both found in generation 1, which meant both sets contained unique descriptors. Therefore the linear correlations between the final, first and second most successful sets have been calculated to determine if the descriptors are linearly correlated (Table 5.2). The correlations were not only calculated between the sets, but also within the set.

Table 5.2 Linear correlations between the best set of descriptors and the final and second best set in the FormRules analysis. Number in brackets are the correlation coefficients

Best Overall Set (Gen 1, Set 2)	Total Number of Correlations with all 167 Descriptors ( $\pm 0.8 - 1$ )	Correlations within the Best Overall Set ( $\pm 0.8 - 1$ )	Correlations with Final Set (Gen 8, Set 2gg)	Correlations with Second Best Set (Gen 1, Set 5)
Dsolv26	27	Dsolv47, dsolv63, dsolv71, dsolv79	Dsolv4 (0.81) dsolv18 (0.85) dsolv71 (0.88)	Dsolv62 (0.81)
Dsolv34	14	Dsolv47, dsolv71	Dsolv71 (0.81)	Dsolv62 (0.87)
Dsolv47	25	Dsolv26, dsolv34, dsolv71	Dsolv18 (0.86) dsolv71 (0.84)	Dsolv24 (0.86) dsolv62 (0.80)
Dsolv63	12	Dsolv26	-	-
Dsolv68	11	Dsolv71, dsolv79	Dsolv50 (0.98) dsolv71 (-0.84)	Dsolv62 (-0.85)
Dsolv69	2	-	-	-
Dsolv71	31	Dsolv26, dsolv34, dsolv47, dsolv68, dsolv79	Dsolv18 (0.98) dsolv71 (1)	Dsolv62 (.096)
Dsolv74	5	-	Dsolv65 (0.96)	Dsolv32 (0.86)
Dsolv79	27	Dsolv26, dsolv68, dsolv71	Dsolv50 (-0.83) dsolv71 (0.97)	Dsolv62 (0.96)
RMM	8	-	Vapour density (1)	-
		Only 3 descriptors are not correlated with others in the best set	6 out of 11 of the final set descriptors are correlated with the best set	3 out of 10 of the second best set descriptors are correlated with the best set

Table 5.2 shows that only three descriptors in the most successful set are not correlated with others within that set. This fact may potentially allow further descriptor reduction. Six out of the eleven descriptors in the final set were correlated with descriptors in the most successful set. However, only three out of ten descriptors in the second most successful set were correlated with the most successful set.

Optimisation of the set was undertaken based on the correlations between the descriptors (Table 5.3). Dsolv71, dsolv74, RMM and d71 were chosen as the starting descriptor set. The reasons for this choice were that dsolv71 was featured in two of the three analysed sets, with a correlated descriptor in the third; therefore appearing to be an important descriptor. Dsolv74 was not correlated to any other descriptors in the best set, but was correlated to a descriptor in the final and second best sets. RMM was also not correlated to other descriptors in the best set, but highly correlated to vapour density in the final set. D71 was featured in the second best set and was the only other descriptor that was correlated between two sets that is not represented by dsolv71, dsolv74 or RMM. The addition of further descriptors was based upon which descriptors were uncorrelated in the most successful set, the final set and the second most successful set.

Table 5.3 Optimisation of the FormRules (FR) descriptor sets. X denotes the presence of the descriptor in the set

Descriptors	Best Overall Set (Gen 1, Set 2)	FR Optimised set 1	FR Optimised set 2	FR Optimised set 3	FR Optimised set 4
Dsolv26	X				
Dsolv34	X				
Dsolv47	X				
Dsolv63	X				
Dsolv68	X				
Dsolv69	X		X		
Dsolv71	X	X	X	X	X
Dsolv74	X	X	X	X	X
Dsolv79	X				
RMM	X	X	X	X	X
D71		X	X	X	X
Dsolv49				X	
MNDO_HF				X	
Vol				X	
D67					X
D77					X
AM1_HOMO					X
PM3_LUMO					X
PmiZ					X
Dielectric constant					X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>81.26</b>	<b>75.60</b>	<b>70.89</b>	<b>73.60</b>	<b>78.02</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>71.99</b>	<b>70.54</b>	<b>78.56</b>	<b>84.72</b>	<b>82.83</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>76.62</b>	<b>73.07</b>	<b>74.56</b>	<b>79.16</b>	<b>80.43</b>

No overall improvement in FormRules<sup>[2]</sup> average  $R^2$  values were observed. However, FR optimised set 4 produced a high overall average result when both FormRules<sup>[2]</sup> and INForm<sup>[1]</sup> are used (80.43 %). The descriptors within the most successful set and also FR optimised set 4 will be taken forward for further analysis.

### 5.1.2. Complete Dataset Analysis – INForm

Using the same starting sets of descriptors as in the FormRules<sup>[2]</sup> work, the analysis was carried out in the same way, but using the INForm<sup>[1]</sup> average  $R^2$  values. A full table of results can be found in Electronic Appendix, Chapter 5, file 5.2, with the final set, the first and second most successful sets presented for further analysis (Table 5.4).

Table 5.4 The final set and first and second most successful descriptor sets in FormRules analysis

Sets	Descriptors	INForm average $R^2$ (%)
<b>Final Set (Gen 8, Set 2g)</b>	Dsolv9, dsolv27, dsolv76, d67, d71, d81, d84, AM1_E, MNDO_dipole, E	<b>82.81</b>
<b>Best Overall Set (Gen 6, Set 5e)</b>	Dsolv9, dsolv10, dsolv17, dsolv45, d81, AM1_LUMO, MNDO_dipole, E, Vol, Glob, logP	<b>87.12</b>
<b>Second Best Set (Gen 1, Set 3)</b>	Dsolv1, dsolv39, dsolv60, d83, AM1_dipole, MNDO_E, E, E_Eele, dP, boiling point	<b>85.23</b>

The linear correlations between the final, first and second most successful sets have been calculated, with the results presented (Table 5.5). The correlations were again calculated between and within the sets. Table 5.5 highlights that there are no correlated descriptors within the best set, which means that a range of molecular properties is represented. Seven of the ten descriptors found in the final set are correlated to those in the best set, with four descriptors (dsolv9, d81, MNDO\_dipole and E) being present in both. Six descriptors from the second best set are also correlated and E is found in all three sets.

Table 5.5 Linear correlations between the best set of descriptors and the final and second best set in the INForm analysis. Number in brackets are the correlation coefficients

Best Overall Set (Gen 6, Set 5e)	Total Number of Correlations with all 167 Descriptors ( $\pm 0.8 - 1$ )	Correlations within the Best Overall Set ( $\pm 0.8 - 1$ )	Correlations with Final Set (Gen 8, Set 2g)	Correlations with Second Best Set (Gen 1, Set 3)
Dsolv9	12	-	Dsolv9 (1)	Dsolv1(0.89) dsolv39 (0.83)
Dsolv10	5	-	-	-
Dsolv17	8	-	Dsolv27 (0.92)	-
Dsolv45	21	-	-	-
D81	4	-	D81 (1) d84 (-1)	-
AM1_LUMO	5	-	AM1_E (-0.94)	AM1_dipole (-0.91)
MNDO_dipole	0	-	MNDO_dipole (1)	-
E	7	-	E (1)	E (1) E_ele (-0.92)
Vol	1	-	-	-
Glob	13	-	D71 (0.80)	-
LogP	12	-	D71 (-0.87)	dP (-0.85)
		None of the best set descriptors are correlated	7 out of 10 final set descriptors are correlated with the best set and 4 are the same	6 out of 10 of the second best set descriptors are correlated with the best set and 1 is the same

Optimisation was carried out using the correlation results, with no improvement in INForm average  $R^2$  values observed (Table 5.6). The initial descriptors used in this analysis were dsolv9, dsolv17, d71, d81, MNDO\_dipole, AM1\_LUMO and E. E was found in all three sets analysed here and dsolv9 was found in two sets and had a correlated descriptor in the third. D81 and MNDO\_dipole were present in two of the sets. AM1\_LUMO and d71 were found in only one set, but had correlated descriptors in the other two sets. Dsolv17 was only correlated with one other descriptor, but was the only correlated descriptor not represented in some way. The additional

descriptors used in the optimised sets were the uncorrelated descriptors in the most successful set, final set and second most successful set.

Table 5.6 Optimisation of the INForm (IN) descriptor sets. X denotes the presence of the descriptor in the set

Descriptors	Best Overall Set (Gen 1, Set 2)	IN Optimised set 1	IN Optimised set 2	IN Optimised set 3	IN Optimised set 4
Dsolv9	X	X	X	X	X
Dsolv10	X		X		
Dsolv17	X	X	X	X	X
Dsolv45	X		X		
D81	X	X	X	X	X
AM1_LUMO	X	X	X	X	X
MNDO_dipole	X	X	X	X	X
E	X	X	X	X	X
Vol	X		X		
Glob	X				
LogP	X				
D71		X	X	X	X
Dsolv27				X	
Dsolv76				X	
D67				X	
Dsolv60					X
MNDO_E					X
Boiling point					X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>41.70</b>	<b>50.40</b>	<b>50.40</b>	<b>51.25</b>	<b>50.40</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>87.12</b>	<b>81.02</b>	<b>77.50</b>	<b>44.15</b>	<b>79.36</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>64.41</b>	<b>65.71</b>	<b>63.95</b>	<b>47.70</b>	<b>64.88</b>

Overall the initial set in this analysis produced the most successful INForm result and therefore will be taken forward for further analysis. IN optimised set 1 generated the highest overall average  $R^2$  value and therefore these descriptors will also be considered. However, it should be noted that the overall performance of IN optimised set 1 is not as successful as other results and therefore less emphasis will be placed on this set of descriptors.

### 5.1.3. Complete Dataset Analysis – FormRules and INForm

Using a similar approach as discussed in 5.1.1 and 5.1.2, both the average  $R^2$  values from INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> have been used. The final, first and second most successful sets have been analysed further (Table 5.7). All results can be found in Electronic Appendix, chapter 5, file 5.3.

Table 5.7 The final set and first and second most successful descriptor sets in FormRules analysis

Sets	Descriptors	FormRules average $R^2$ (%)	INForm average $R^2$ (%)	Overall average $R^2$ (%)
<b>Final Set (Gen 8, Set 2t)</b>	Dsolv42, dsolv50, dsolv65, dsolv67, dsolv68, dsolv79, AM1_E, E_strain, ASA	<b>77.88</b>	<b>80.28</b>	<b>79.08</b>
<b>Best Overall Set (Gen 3, Set 11y)</b>	Dsolv8, dsolv22, dsolv41, dsolv49, dsolv66, MNDO_HF, Rgyr, ASA, ASA_H, RMM, logP	<b>80.17</b>	<b>86.89</b>	<b>83.53</b>
<b>Second Best Set (Gen 3, Set 5y)</b>	Dsolv32, dsolv36, dsolv39, dsolv51, d73, d83, AM1_dipole, AM1_LUMO, E, E_vdw	<b>77.90</b>	<b>86.97</b>	<b>82.44</b>

The first and second most successful sets are from the same generation, therefore containing unique descriptors. Although the descriptors are unique, there may be linear correlations between the sets, which were therefore analysed (Table 5.8).

The overall results in this combined analysis are much higher than those seen when only either FormRules<sup>[2]</sup> or INForm<sup>[1]</sup> are used; suggesting valuable information may be lost through using only one analysis method. The correlations within and between the sets have been calculated with results presented in Table 5.8.



Table 5.8 Linear correlations between the best set of descriptors and the final and second best set. Number in brackets are the correlation coefficients

Best Overall Set (Gen 3, Set 11y)	Total Number of Correlations with all 167 Descriptors ( $\pm 0.8 - 1$ )	Correlations within the Best Overall Set ( $\pm 0.8 - 1$ )	Correlations with Final Set (Gen 8, Set 2t)	Correlations with Second Best Set (Gen 3, Set 5y)
Dsolv8	28	Dsolv22	Dsolv42 (0.84) dsolv79 (0.82)	Dsolv39 (0.87)
Dsolv22	29	-	Dsolv42 (0.98) dsolv79 (0.84)	Dsolv39 (0.80) E_vdw (0.80)
Dsolv41	0	-	-	-
Dsolv49	3	-	-	Dsolv51 (0.91)
Dsolv66	4	-	Dsolv65 (0.96)	Dsolv32 (0.94)
MNDO_HF	2	-	-	-
Rgyr	14	logP	-	E_vdw (0.82)
ASA	1	ASA_H	ASA (1)	-
ASA_H	1	ASA	ASA (1)	-
RMM	8	-	-	-
logP	12	rgyr	-	D73 (-0.81) E_vdw (0.84)
		5 of the best set descriptors are correlated	4 out of 10 final set descriptors are correlated with the best set and 1 is the same	5 out of 10 of the second best set descriptors are correlated with the best set

Overall, there are fewer correlations in this analysis than have been seen in the previous work (Table 5.2 and Table 5.5). Only ASA was present in more than one of the three sets. Dsolv42 and dsolv65 have correlated descriptors in all three sets, with AM1\_E, E, E\_vdw and dsolv51 representing the remaining correlations. Optimisation was carried out based on this information (Table 5.9).

Table 5.9 Optimisation of the FormRules descriptor sets. X denotes the presence of the descriptor in the set

Descriptors	Best Overall Set (Gen 3, Set 11y)	Combined Optimised set 1	Combined Optimised set 2	Combined Optimised set 3	Combined Optimised set 4
Dsolv8	X				
Dsolv22	X				
Dsolv41	X		X		
Dsolv49	X				
Dsolv66	X				
MNDO_HF	X		X		
Rgyr	X				
ASA	X	X	X	X	X
ASA_H	X				
RMM	X		X		
logP	X				
Dsolv42		X	X	X	X
Dsolv65		X	X	X	X
AM1_E		X	X	X	X
E		X	X	X	X
E_vdw		X	X	X	X
Dsolv51		X	X	X	X
Dsolv50				X	
Dsolv67				X	
Dsolv68				X	
Dsolv69				X	
D83					X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.17</b>	<b>81.05</b>	<b>79.93</b>	<b>79.02</b>	<b>81.04</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>86.89</b>	<b>73.78</b>	<b>84.16</b>	<b>45.30</b>	<b>77.44</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>83.53</b>	<b>77.41</b>	<b>82.04</b>	<b>62.16</b>	<b>79.24</b>

The additional descriptors in the optimisation sets are the uncorrelated descriptors in the most successful, final and second most successful sets. The results (Table 5.9) show no improvement in the average  $R^2$  values, and all but one set performing well. The descriptors in the overall best set will be taken forward for further analysis.

#### 5.1.4. Optimisation of the High Performing Sets

Analysis of using FormRules<sup>[2]</sup> and INForm<sup>[1]</sup> separately and in combination has been carried out and has produced five sets of descriptors for further analysis (Table 5.10). From the average  $R^2$  values presented here, two of the sets generated percentages below 70 %, which from earlier research is known to have under performed. Therefore analysis of the three remaining sets will be carried out in order to assess whether the descriptors presented are correlated.

Table 5.10 The final set and first and second most successful descriptor sets in the high performing sets analysis

Sets	Descriptors	FormRules average $R^2$ (%)	INForm average $R^2$ (%)	Overall average $R^2$ (%)
FR Best Overall Set (Gen 1, Set 2)	Dsolv26, dsolv34, dsolv47, dsolv63, dsolv68, dsolv69, dsolv71, dsolv74, dsolv79, RMM	81.26	71.99	76.62
FR Optimised set 4	Dsolv71, dsolv74, d67, d71, d77, AM1_HOMO, dielectric constant, PmiZ, RMM, Pm3_LUMO	78.02	82.83	80.43
IN Best Overall Set (Gen 6, Set 5e)	Dsolv9, dsolv10, dsolv17, dsolv45, d81, AM1_LUMO, MNDO_dipole, E, Vol, Glob, logP	41.70	87.12	64.41
IN Optimised set 1	Dsolv9, dsolv17, d81, AM1_LUMO, d71, E, MNDO_dipole	50.40	81.02	65.71
Combined Best Overall Set (Gen 3, Set 11y)	Dsolv8, dsolv22, dsolv41, dsolv49, dsolv66, MNDO_HF, rgyr, ASA, ASA_H, RMM, logP	80.17	86.89	83.53

The correlations within and between the three sets have been calculated with results presented in Table 5.11. There are a high number of correlated descriptors within the

sets and the molecular mass of the solvent (RMM) was present in all three. This demonstrates that the different analysis methods used on the large dataset can produce similar final results.

Table 5.11 Linear correlations between the best set of descriptors and the final and second best set. Number in brackets are the correlation coefficients

Combined Best Overall Set (Gen 3, Set 11y)	Total Number of Correlations with all 167 Descriptors ( $\pm 0.8 - 1$ )	Correlations within the Best Overall Set ( $\pm 0.8 - 1$ )	Correlations with FR Best Overall Set (Gen 1, Set 2)	Correlations with FR Optimised set 4
Dsolv8	28	Dsolv22	Dsolv26 (0.90) dsolv34 (0.91) dsolv47 (0.87) dsolv63 (0.87) dsolv71 (0.85) dsolv79 (0.82)	dsolv71 (0.85)
Dsolv22	29	Dsolv8	Dsolv26 (0.96) dsolv47 (0.86) dsolv71 (0.90) dsolv79 (0.84)	dsolv71 (0.90)
Dsolv41	0	-	-	-
Dsolv49	3	-	-	-
Dsolv66	4	-	Dsolv74 (0.90)	Dsolv74 (0.90)
MNDO_HF	2	-	-	-
Rgyr	14	logP	-	D71 (-0.87) pmiZ (-0.87) Dielectric constant (-0.83)
ASA	1	ASA_H	-	-
ASA_H	1	ASA	-	-
RMM	8	-	RMM (1)	RMM (1)
logP	12	rgyr	-	D71 (-0.87) pmiZ (-0.82) Dielectric constant (-0.90)
		6 of the best set descriptors are correlated	7 out of 10 final set descriptors are correlated with the best set and 1 is the same	6 out of 10 of the second best set descriptors are correlated with the best set and 1 is the same

Optimisation was carried out based on the correlations between the three sets but no overall improvements were observed (Table 5.12).

Table 5.12 Optimisation of the top performing descriptor sets. X denotes the presence of the descriptor in the set

Descriptors	Combined Best Overall Set (Gen 3, Set 11y)	Optimised set 1	Optimised set 2	Optimised set 3	Optimised set 4
Dsolv8	X				
Dsolv22	X				
Dsolv41	X		X		
Dsolv49	X		X		
Dsolv66	X				
MNDO_HF	X		X		
Rgyr	X				
ASA	X		X		
ASA_H	X		X		
RMM	X	X	X	X	X
logP	X				
Dsolv71		X	X	X	X
Dsolv74		X	X	X	X
D71		X	X	X	X
Dsolv69				X	
D67					X
D77					X
AM1_HOMO					X
PM3_LUMO					X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.17</b>	<b>70.54</b>	<b>79.95</b>	<b>70.89</b>	<b>77.65</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>86.89</b>	<b>56.58</b>	<b>62.38</b>	<b>77.40</b>	<b>70.75</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>83.53</b>	<b>63.56</b>	<b>71.17</b>	<b>74.15</b>	<b>74.20</b>

This analysis has generated a set of eleven descriptors that can predict the polymorphic form crystallised to 83.53 %. The eleven descriptors are dsolv8, dsolv22, dsolv41, dsolv49, dsolv66, MNDO\_HF, rgyr, ASA, ASA\_H, RMM and logP. These descriptors are a mixture of bulk and molecular solvent properties and also molecular properties of the carbamazepine (CBZ) molecule.

The relative molecular mass (RMM), partition coefficient (logP) and number of bonds (dsolv8) of the solvent are represented in this set. Also molecular level solvent properties such as the Randić index (dsolv22), bonding information content (dsolv41), 3D-Kier and Hall index (dsolv49) and the moment of inertia A (dsolv66). The CBZ molecule is represented by the calculated heat of formation (MNDO\_HF), radius of gyration (rgyr) and the water accessible surface area of all atoms (ASA) and all hydrophobic atoms (ASA\_H). Details of which can be found in appendix section 12.2.

These eleven descriptors will be compared to those highlight in the linear correlation work in order to generate the most successful set from the manual analysis research. A discussion of the final descriptors will then be made relating the properties to the nucleation and growth of different polymorphic forms.

## 5.2. Linear Correlations Analysis

Linear correlations of all the 167 descriptors were calculated, with highly positively and negatively correlated descriptors grouped. This was carried out in order to reduce the dataset. Initially, all descriptors that were correlated between  $\pm 0.95$  and  $\pm 1$  were clustered together, which resulted in 69 descriptors being uncorrelated. When one descriptor from each correlated cluster was added to the uncorrelated descriptors, the dataset was reduced to 84. However, further reduction was desirable, and therefore all correlations between  $\pm 0.8$  and  $\pm 1$  were clustered. This reduced the number of uncorrelated descriptors to 18. One descriptor from each correlation cluster was selected and in the case of the largest cluster (shown in appendix section 12.7 and Electronic Appendix, Chapter 5, file 5.4, 5.5 and 5.6), 7 different descriptors, reducing the dataset to 40 values.

The 40 descriptors selected were given a random number and then numerically ordered. Four sets were created for analysis by placing the top 10 descriptors in set 1, the next 10 into set 2 and so on. The 40 selected descriptors and the initial analysis

sets they were placed in are presented (Table 5.13). This process was then repeated three times, generating four different groups of four descriptor sets (in the following tables). FormRules<sup>[2]</sup> and INForm<sup>[1]</sup> analysis was carried out upon each set with the results presented in Table 5.13, Table 5.14, Table 5.15 and Table 5.16.

Table 5.13 The 40 descriptors selected for analysis from the correlations analysis

Sets	Descriptors	FormRules average R <sup>2</sup> (%)	INForm average R <sup>2</sup> (%)	Overall average R <sup>2</sup> (%)
Correlation Set 1 Run 1	AM1_Eele, dsolv41, VSA, d72, dsolv11, MNDO_LUMO, dsolv76, ASA, d77, gutmann donor number	52.38	32.96	42.67
Correlation Set 2 Run 1	D70, MNDO_dipole, d75, PM3_E, PM3_HF, MNDO_Eele, solubility, dsolv65, rgyr, Henry's law constant	79.96	63.66	71.81
Correlation Set 3 Run 1	D67, AM1_E, d86, dsolv2, d69, dsolv57, vol, dsolv43, MNDO_E, vapour pressure	50.42	61.67	56.04
Correlation Set 4 Run 1	Dsolv78, density, dsolv71, E_nb, dsolv31, E, dipole, d82, dsolv70, E_vdw	42.82	84.38	63.60

Table 5.14 The descriptors and results of the second group of 4 sets

Sets	Descriptors	FormRules average R <sup>2</sup> (%)	INForm average R <sup>2</sup> (%)	Overall average R <sup>2</sup> (%)
Correlation Set 1 Run 2	D67, d70, dsolv31, dsolv70, MNDO_LUMO, ASA, VSA, density, solubility, Henry's law constant	60.31	49.91	55.11
Correlation Set 2 Run 2	D69, d72, d77, dsolv11, dsolv57, dsolv76, AM1_E, E, PM3_E, rgyr	42.69	66.47	54.58
Correlation Set 3 Run 2	D82, d86, dsolv41, dsolv71, MNDO_E, MNDO_Eele, PM3_HF, vapour pressure, gutmann donor number	52.82	74.13	63.47
Correlation Set 4 Run 2	D75, dsolv2, dsolv43, dsolv65, dsolv78, E_nb, E_vdw, AM1_Eele, MNDO_dipole, vol	75.84	81.75	78.80

Table 5.15 The descriptors and results of the third group of 4 sets

Sets	Descriptors	FormRules average R <sup>2</sup> (%)	INForm average R <sup>2</sup> (%)	Overall average R <sup>2</sup> (%)
Correlation Set 1 Run 3	D67, d82, dsolv65, dsolv71, ASA, MNDO_Eele, PM3_HF, VSA, rgyr, solubility	78.76	51.59	65.17
Correlation Set 2 Run 3	Dsolv2, dsolv31, dsolv70, E_nb, E, AM1_Eele, MNDO_E, MNDO_LUMO, vol, vapour pressure	43.31	66.71	55.01
Correlation Set 3 Run 3	D69, d70, d72, d75, d77, dsolv11, dsolv43, E_vdw, MNDO_dipole, density	52.12	66.29	59.20
Correlation Set 4 Run 3	D86, dipole, dsolv41, dsolv57, dsolv76, dsolv78, AM1_E, PM3_E, gutmann donor number, Henry's law constant	51.69	79.80	65.75

Table 5.16 The descriptors and results of the fourth group of 4 sets

Sets	Descriptors	FormRules average R <sup>2</sup> (%)	INForm average R <sup>2</sup> (%)	Overall average R <sup>2</sup> (%)
Correlation Set 1 Run 4	Dsolv65, dsolv71, dsolv76, AM1_E, AM1_Eele, ASA, E_vdw, PM3_E, PM3_HF, solubility	75.30	66.50	70.90
Correlation Set 2 Run 4	D70, d72, dsolv31, dsolv70, E, E_nb, MNDO_Eele, rgyr, vol, vapour pressure	44.78	81.78	63.28
Correlation Set 3 Run 4	D67, d69, d75, d82, dsolv43, dsolv57, MNDO_dipole, MNDO_LUMO, VSA, dipole	53.40	62.32	57.86
Correlation Set 4 Run 4	D77, d86, dsolv11, dsolv2, dsolv41, dsolv78, MNDO_E, gutmann donor number, Henry's law constant, density	52.08	73.22	62.65



### 5.2.1. Linear Correlation Analysis – FormRules

Using only the results acquired from the FormRules<sup>[2]</sup> analysis, the top set was taken from each group of 4 (Table 5.17). This was to assess whether any of the descriptors were featured in more than one successful set. The aim of this analysis was to highlight consistently high performing descriptors.

Table 5.17 The top set from each group based on the results of FormRules analysis

Sets	Descriptors	FormRules average R <sup>2</sup> (%)
Correlation Set 2 Run 1	D70, MNDO_dipole, d75, PM3_E, PM3_HF, MNDO_Eele, solubility, dsolv65, rgyr, Henry's law constant	79.96
Correlation Set 4 Run 2	D75, dsolv2, dsolv43, dsolv65, dsolv78, E_nb, E_vdw, AM1_Eele, MNDO_dipole, vol	75.84
Correlation Set 1 Run 3	D67, d82, dsolv65, dsolv71, ASA, MNDO_Eele, PM3_HF, VSA, rgyr, solubility	78.76
Correlation Set 1 Run 4	Dsolv65, dsolv71, dsolv76, AM1_E, AM1_Eele, ASA, E_vdw, PM3_E, PM3_HF, solubility	75.30

From the 40 descriptors in this analysis, 24 unique descriptors were present in the top four sets of the FormRules<sup>[2]</sup> analysis (Table 5.18). Twelve of these descriptors featured more than once in the four sets, with dsolv65 being present in all four sets. Dsolv65 was used in the prediction of form I and dihydrate in all cases, and also for form II on one occasion. PM3\_HF and solubility featured in three of the sets, but solubility was not present in any of the rules. PM3\_HF was used in form III rules on two occasions, but not in the third. The rules from each of these top sets can be found in Electronic Appendix, Chapter 5, file 5.7.

Table 5.18 Monitoring the repeat occurrence of descriptors in the top sets of FormRules analysis. X denotes the presence of the descriptor in the set

	Correlation Set 2 Run 1	Correlation Set 4 Run 2	Correlation Set 1 Run 3	Correlation Set 1 Run 4	Total number of occurrences
D70	X				1
D75	X	X			2
Dsolv65	X	X	X	X	4
MNDO_dipole	X	X			2
MNDO_Eele	X		X		2
PM3_E	X			X	2
PM3_HF	X		X	X	3
Rgyr	X		X		2
Solubility	X		X	X	3
Henry's law constant	X				1
Dsolv2		X			1
Dsolv43		X			1
Dsolv78		X			1
AM1_Eele		X		X	2
E_nb		X			1
E_vdw		X		X	2
Vol		X			1
D67			X		1
D82			X		1
Dsolv71			X	X	2
ASA			X	X	2
VSA			X		1
Dsolv76				X	1
AM1_E				X	1

Analysis of the 24 unique descriptors was carried out generating good results in FormRules<sup>[2]</sup>, but mediocre results in INForm<sup>[1]</sup>, 79.17 % and 66.71 % respectively (Table 5.19). Using a high number of descriptors makes it difficult to determine the most informative factors in polymorphic form prediction. Therefore, further analysis was carried out to reduce the number of descriptors used. By using only the 12 descriptors that occurred more than once in the top 4 sets, a slight reduction in the FormRules<sup>[2]</sup> average  $R^2$  value was observed (78.26 %). However, an improvement in the INForm<sup>[1]</sup> and overall average  $R^2$  values were observed, 78.08 % and 78.17 % respectively. What is noticeable about this result is that perhaps descriptors within correlation set 2, run 1 (the top performing set) that did not occur more than once were important in rule formation. Also, INForm<sup>[1]</sup> perhaps performs better when there are less descriptors used in the analysis. The rule descriptors from the correlation set 2, run 1, and also the 12 most occurring descriptors were analysed further (Table 5.19).

Table 5.19 Further analysis of FormRules results. X denotes the presence of the descriptor in the set

Descriptors	Correlation Set 2 Run 1	25 unique	12 most occurring	Correlation Set 2 Run 1 Rule descriptor only	25 unique Rule descriptor only	12 most occurring Rule descriptor only
D70	X	X		X	X	
D75	X	X	X	X		X
Dsolv65	X	X	X	X	X	X
MNDO_dipole	X	X	X	X	X	X
MNDO_Eele	X	X	X			
PM3_E	X	X	X	X		
PM3_HF	X	X	X			
Rgyr	X	X	X			
Solubility	X	X	X			
Henry's law constant	X	X		X		
Dsolv2		X				
Dsolv43		X				
Dsolv78		X			X	
AM1_Eele		X	X			
E_nb		X			X	
E_vdw		X	X		X	X
Vol		X				
D67		X				
D82		X				
Dsolv71		X	X		X	X
ASA		X	X			
VSA		X			X	
Dsolv76		X				
AM1_E		X				
<b>FormRules average R<sup>2</sup> (%)</b>	<b>79.96</b>	<b>79.17</b>	<b>78.26</b>	<b>79.96</b>	<b>79.17</b>	<b>78.26</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>63.66</b>	<b>66.71</b>	<b>78.08</b>	<b>71.55</b>	<b>72.36</b>	<b>77.75</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>71.81</b>	<b>72.94</b>	<b>78.17</b>	<b>75.76</b>	<b>75.77</b>	<b>78.01</b>

When only the rule descriptors of correlation set 2, run 1 are analysed, the FromRules<sup>[2]</sup>  $R^2$  value is at its highest (79.96 %). An improvement in INForm<sup>[1]</sup> performance from 63.66 % to 71.55 % was also observed when the none rule descriptors are removed. When the overall averages are used to determine the most successful set, the 12 most occurring descriptors generate the highest value (78.17 %). These 12 descriptors and also the rule only descriptors from correlation set 2, run will be considered in further analysis.

### 5.2.2. Linear Correlation Analysis – INForm

Similar analysis has been carried out using the same correlation sets (Table 5.13, Table 5.14, 5 and Table 5.16) but on this occasion analysing the top  $R^2$  values from INForm<sup>[1]</sup>.

Table 5.20 The top set from each group based on the results of INForm analysis

Sets	Descriptors	INForm average $R^2$ (%)
Correlation Set 4 Run 1	Dsolv78, density, dsolv71, E_nb, dsolv31, E, dipole, d82, dsolv70, E_vdw	84.38
Correlation Set 4 Run 2	D75, dsolv2, dsolv43, dsolv65, dsolv78, E_nb, E_vdw, AM1_Eele, MNDO_dipole, vol	81.75
Correlation Set 4 Run 3	D86, dipole, dsolv41, dsolv57, dsolv76, dsolv78, AM1_E, PM3_E, gutmann donor number, Henry's law constant	79.80
Correlation Set 2 Run 4	D70, d72, dsolv31, dsolv70, E, E_nb, MNDO_Eele, rgyr, vol, vapour pressure	81.78

Out of the 40 descriptors present, there are 30 unique descriptors that generated the highest INForm<sup>[1]</sup> prediction values (Table 5.21), with only 8 of these occurring more than once. An interesting result is that 7 out of 8 of these descriptors are present in the highest performing set (correlation set 4, run 1). Optimisation was carried out based on these results (Table 5.22).

In both Table 5.21 and Table 5.22 X denotes the presence of the descriptor in the set.

Table 5.21 Monitoring the repeat occurrence of descriptors in the top sets of FormRules analysis

	Correlation set 4 Run 1	Correlation set 4 Run 2	Correlation set 4 Run 3	Correlation set 2 Run 4	Number of occurrences
D82	X				1
Dsolv31	X			X	2
Dsolv70	X			X	2
Dsolv71	X				1
Dsolv78	X	X	X		3
E	X			X	2
E_nb	X	X		X	3
E_vdw	X	X			2
Density	X				1
Dipole	X		X		2
D75		X			1
Dsolv2		X			1
Dsolv43		X			1
Dsolv65		X			1
AM1_Eele		X			1
MNDO_dipole		X			1
Vol		X		X	2
D86			X		1
Dsolv41			X		1
Dsolv57			X		1
Dsolv76			X		1
AM1_E			X		1
PM3_E			X		1
Gutmann donor number			X		1
Henry's law constant			X		1
D70				X	1
D72				X	1
MNDO_Eele				X	1
Rgyr				X	1
Vapour pressure				X	1

Table 5.22 Analysis of descriptors highlighted in the INForm analysis

	Correlation set 4 Run 1	30 unique descriptors	8 most occurring	30 unique descriptors Rule only
D82	X	X		
Dsolv31	X	X	X	
Dsolv70	X	X	X	
Dsolv71	X	X		X
Dsolv78	X	X	X	X
E	X	X	X	
E_nb	X	X	X	
E_vdw	X	X	X	X
Density	X	X		
Dipole	X	X	X	
D75		X		X
Dsolv2		X		
Dsolv43		X		
Dsolv65		X		X
AM1_Eele		X		
MNDO_dipole		X		X
Vol		X	X	
D86		X		
Dsolv41		X		
Dsolv57		X		X
Dsolv76		X		
AM1_E		X		
PM3_E		X		
Gutmann donor number		X		X
Henry's law constant		X		
D70		X		
D72		X		X
MNDO_Eele		X		
Rgyr		X		
Vapour pressure		X		

	Correlation set 4 Run 1	30 unique descriptors	8 most occurring	30 unique descriptors Rule only
<b>FormRules average R<sup>2</sup> (%)</b>	<b>42.82</b>	<b>79.55</b>	<b>41.33</b>	<b>79.55</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>84.38</b>	<b>82.63</b>	<b>35.70</b>	<b>83.46</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>63.60</b>	<b>81.09</b>	<b>38.52</b>	<b>81.51</b>

Analysis of the 30 unique descriptors generated high INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> R<sup>2</sup> values, at 82.63 % and 79.55 % respectively. However, by using 30 descriptors it is difficult to know which of those are directing the correct prediction of polymorphic form. Therefore, further analysis was carried out using the descriptors present in the rules only, to reduce the number of descriptors (Table 5.22). Analysis of the 8 most occurring descriptors was also carried out, generating surprisingly poor results. INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> both performed very badly in this analysis, with average R<sup>2</sup> values of 35.70 % and 41.33 % respectively. This suggests that the descriptors in correlation set 4, run 1, that were not analysed (d82, dsolv71 and density) added a high amount of value to the set. Alternatively, the inclusion of vol may mask essential descriptors by being present.

Analysis of this hypothesis was conducted and it was established that when d82, dsolv71 and density were added both individually and in pairs the results improved (Table 5.23).

Overall the best predictive model was created by using the rule descriptors from the 30 unique descriptors analysis (overall average R<sup>2</sup> value of 81.51 %). The descriptors involved in this model were d72, d75, dsolv71, dsolv78, dsolv57, dsolv65, E\_vdw, MNDO\_dipole, gutmann donor number. The descriptors involved in correlation set 4, run 1 (d82, dsolv31, dsolv70, dsolv71, dsolv78, E, E\_nb, E\_vdw, density and dipole) will also be taken forward for further analysis as they generated the highest INForm<sup>[1]</sup> average R<sup>2</sup> value.



Table 5.23 Analysis of the effect of d82, dsolv71 and density. X denotes the presence of the descriptor in the set

	Correlation set 4 Run 1 (Best)	8 most occurring	Best + vol	8 most occurring + d82	8 most occurring + density	8 most occurring + dsolv71	8 most occurring + d82 + density	8 most occurring + d82 + dsolv71	8 most occurring + dsolv71 + density
D82	X		X	X			X	X	
Dsolv31	X	X	X	X	X	X	X	X	X
Dsolv70	X	X	X	X	X	X	X	X	X
Dsolv71	X		X			X		X	X
Dsolv78	X	X	X	X	X	X	X	X	X
E	X	X	X	X	X	X	X	X	X
E_nb	X	X	X	X	X	X	X	X	X
E_vdw	X	X	X	X	X	X	X	X	X
Density	X		X		X		X		X
Dipole	X	X	X	X	X	X	X	X	X
Vol		X	X	X	X	X	X	X	X
FormRules average R <sup>2</sup> (%)	42.82	41.33	42.82	41.04	41.33	41.33	42.82	41.02	41.33
INForm average R <sup>2</sup> (%)	84.38	35.70	62.26	67.22	73.51	84.29	79.34	77.61	70.34
Overall average R <sup>2</sup> (%)	63.60	38.52	52.54	54.12	57.42	62.81	61.08	59.31	55.83

### 5.2.3. Descriptor Overlaps in FormRules and INForm Analysis

The highest overall performing sets from the linear correlation INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> analysis have been compared (Table 5.24). Five overlapping descriptors were observed and analysed and optimised, with results in Table 5.25.

Table 5.24 Comparison of the descriptors in the most successful FormRules and INForm sets. X denotes the presence of the descriptor in the set

Descriptors	Best Set from FormRules	Best Set from INForm
D75	X	X
Dsolv65	X	X
MNDO_dipole	X	X
MNDO_Eele	X	
PM3_E	X	
PM3_HF	X	
Rgyr	X	
Solubility	X	
AM1_Eele	X	
E_vdw	X	X
Dsolv71	X	X
ASA	X	
Dsolv78		X
Dsolv57		X
Gutmann donor number		X
D72		X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>78.26</b>	<b>79.55</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>78.08</b>	<b>83.46</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>78.17</b>	<b>81.51</b>

Analysis of the overlapping descriptors resulted in high R<sup>2</sup> values. However, the average R<sup>2</sup> value (78.64 %) was a reduction of that seen in the best INForm<sup>[1]</sup> analysis. Therefore the remaining descriptors in the set were added to see what impact each of them had on prediction (Table 5.25).

Table 5.25 Analysis of overlapping descriptors with those found in INForm and FormRules best sets

Descriptors	Overlap	Set A	Set B	Set C	Set D	Set E	Set F	Set G	Set H	Set I	Set J	Set K	Set L	Set M	Set N
Dsolv65	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Dsolv71	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Dsolv78	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
MNDO_dipole	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
E_vdw	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
D72		X				X	X	X				X	X		X
D75			X			X			X	X		X		X	X
Dsolv57				X			X		X		X		X	X	X
Gutmann donor number					X			X		X	X	X	X	X	
<b>FormRules average R<sup>2</sup> (%)</b>	<b>78.44</b>	<b>78.59</b>	<b>78.38</b>	<b>79.90</b>	<b>80.12</b>	<b>77.81</b>	<b>80.06</b>	<b>80.28</b>	<b>79.90</b>	<b>80.07</b>	<b>80.12</b>	<b>79.50</b>	<b>80.27</b>	<b>80.12</b>	<b>79.33</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>78.82</b>	<b>82.38</b>	<b>72.48</b>	<b>73.55</b>	<b>42.14</b>	<b>89.64</b>	<b>84.77</b>	<b>86.88</b>	<b>65.52</b>	<b>79.04</b>	<b>85.19</b>	<b>75.04</b>	<b>84.18</b>	<b>87.43</b>	<b>84.54</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>78.64</b>	<b>80.47</b>	<b>75.43</b>	<b>76.73</b>	<b>61.13</b>	<b>83.72</b>	<b>82.42</b>	<b>83.58</b>	<b>72.71</b>	<b>79.56</b>	<b>82.66</b>	<b>77.27</b>	<b>82.23</b>	<b>83.78</b>	<b>81.94</b>

This optimisation has shown that by removing d72 from the initial best INForm<sup>[1]</sup> set of descriptors the results for both INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> are improved. The following descriptors will be taken forward for further analysis, d75, dsolv57, dsolv65, dsolv71, dsolv78, MNDO\_dipole, E\_vdw and gutmann donor number.

#### 5.2.4. Linear Correlation Analysis – FormRules and INForm

From analysis in section 5.1.3, using the combined average  $R^2$  value from INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> leads to a more successful network. Analysis has been carried out in a similar manner as the previous work in this chapter, with the most successful sets from each run being grouped together to identify any frequently occurring descriptors (Table 5.26 and Table 5.27).

Table 5.26 The top performing set from each group, based on the average result from FormRules and INForm analysis

Sets	Descriptors	FormRules average $R^2$ (%)	INForm average $R^2$ (%)	Overall average $R^2$ (%)
Correlation Set 2 Run 1	D70, MNDO_dipole, d75, PM3_E, PM3_HF, MNDO_Eele, solubility, dsolv65, rgyr, Henry's law constant	79.96	63.66	71.81
Correlation Set 4 Run 2	D75, dsolv2, dsolv43, dsolv65, dsolv78, E_nb, E_vdw, AM1_Eele, MNDO_dipole, vol	75.84	81.75	78.80
Correlation Set 4 Run 3	D86, dipole, dsolv41, dsolv57, dsolv76, dsolv78, AM1_E, PM3_E, gutmann donor number, Henry's law constant	51.69	79.80	65.75
Correlation Set 1 Run 4	Dsolv65, dsolv71, dsolv76, AM1_E, AM1_Eele, ASA, E_vdw, PM3_E, PM3_HF, solubility	75.30	66.50	70.90

Table 5.27 Comparison of the descriptors in the top performing sets using average FormRules and INForm results. X denotes the presence of the descriptor in the set

Descriptors	Correlation Set 2 Run 1	Correlation Set 4 Run 2	Correlation Set 4 Run 3	Correlation Set 1 Run 4	Number of occurrences
D70	X				1
MNDO_dipole	X	X			2
D75	X	X			2
PM3_E	X		X	X	3
PM3_HF	X			X	2
MNDO_Eele	X				1
Solubility	X			X	2
Dsolv65	X	X		X	3
Henry's law constant	X		X		2
Rgyr	X				1
E_nb		X			1
Vol		X			1
Dsolv78		X	X		2
AM1_Eele		X		X	2
Dsolv2		X			1
E_vdw		X		X	2
Dsolv43		X			1
Dsolv57			X		1
Dipole			X		1
D86			X		1
Gutmann donor number			X		1
Dsolv76			X	X	2
Dsolv41			X		1
AM1_E			X	X	2
Dsolv71				X	1
ASA				X	1

Table 5.28 Optimisation of the descriptors highlighted in the combined analysis. X denotes the presence of the descriptor in the set

Descriptors	Correlation Set 4 Run 2	26 unique descriptors	12 most occurring	26 unique descriptors Rule only
D70		X		
MNDO_dipole	X	X	X	X
D75	X	X	X	
PM3_E		X	X	
PM3_HF		X	X	
MNDO_Eele		X		
Solubility		X	X	
Dsolv65	X	X	X	X
Henry's law constant		X	X	
Rgyr		X		
E_nb	X	X		X
Vol	X	X		
Dsolv78	X	X	X	X
AM1_Eele	X	X	X	
Dsolv2	X	X		
E_vdw	X	X	X	X
Dsolv43	X	X		
Dsolv57		X		X
Dipole		X		
D86		X		X
Gutmann donor number		X		X
Dsolv76		X	X	
Dsolv41		X		
AM1_E		X	X	
Dsolv71		X		X
ASA		X		

Descriptors	Correlation Set 4 Run 2	26 unique descriptors	12 most occurring	26 unique descriptors Rule only
FormRules average $R^2$ (%)	75.84	79.99	77.97	79.99
INForm average $R^2$ (%)	81.75	79.69	76.32	82.00
Overall average $R^2$ (%)	78.80	79.84	77.15	81.00

There were 26 unique descriptors out of a possible 40 in this analysis, with 12 of these occurring more than once in the top 4 sets. In an attempt to reduce the dataset further, the rule descriptors were used from the 26 unique descriptor FormRules<sup>[2]</sup> analysis (Table 5.28).

Analysis of the 26 unique descriptors generated an improved overall average  $R^2$  value compared to the previous highest value. Further improvements were also seen when the rule descriptors from the 26 unique descriptor analysis (MNDO\_dipole, dsolv65, E\_nb, dsolv78, E\_vdw, dsolv57, d86, gutmann donor number and dsolv71) were analysed (81.00 %).

### 5.2.5. Optimisation of the Linear Correlation Best Set

Two highly successful sets of descriptors have been highlighted in the linear correlation analysis, these were the optimised set generated by comparing the best set from FormRules<sup>[2]</sup> and INForm<sup>[1]</sup> analysis (set M) and the rule only descriptors from the combined analysis of the 26 unique descriptor optimisation. The descriptors are presented (Table 5.29) and it can be observed that there are only 3 descriptors that differ between these sets, d75, d86 and E\_nb. Therefore, further analysis has been conducted to determine if further optimisation may occur.

Table 5.29 Optimisation of the best sets from the linear correlation analysis. X denotes the presence of the descriptor in the set

Descriptors	26 unique descriptors Rule only	Set M	All 10	Set 1a	Set 2a	Set 3a	Set 4a
D86	X		X		X	X	
Dsolv57	X	X	X	X	X	X	X
Dsolv65	X	X	X	X	X	X	X
Dsolv71	X	X	X	X	X	X	X
Dsolv78	X	X	X	X	X	X	X
MNDO_dipole	X	X	X	X	X	X	X
E_vdw	X	X	X	X	X	X	X
Gutmann donor number	X	X	X	X	X	X	X
E_nb	X		X	X			X
D75		X	X	X	X		
<b>FormRules average R<sup>2</sup> (%)</b>	<b>79.99</b>	<b>80.12</b>	<b>79.99</b>	<b>80.60</b>	<b>80.12</b>	<b>80.12</b>	<b>80.60</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>82.00</b>	<b>87.43</b>	<b>74.27</b>	<b>77.71</b>	<b>74.00</b>	<b>78.35</b>	<b>78.13</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>81.00</b>	<b>83.78</b>	<b>77.13</b>	<b>79.16</b>	<b>77.06</b>	<b>79.24</b>	<b>79.37</b>

The optimisation results show that there is no improvement to be made on the previously highly performing set of descriptors. What is interesting to conclude from this analysis is that by using the results from INForm<sup>[1]</sup> and FormRules<sup>[2]</sup> individually and in combination there is a large degree of descriptor overlap, which suggests that useful results can be generated using this analysis technique.

### 5.3. Overall Optimisation of Best Descriptor Set

The linear correlation analysis and analysis of all 167 descriptors each generated an optimised set (Table 5.30). These two sets were compared and further optimised in order to determine the best set of descriptors for polymorphic form prediction. From this point forward the linear correlation best set (set M) shall be referred to as Corr.



Best Set. The best set from the analysis of all 167 descriptors (Combined Best Overall Set, Gen 3, Set 11y) shall be referred to as All Best Set.

Table 5.30 The two most successful sets from linear correlation and all descriptor analysis

Sets	Descriptors	FormRules average $R^2$ (%)	INForm average $R^2$ (%)	Overall average $R^2$ (%)
<b>Corr. Best Set</b>	Dsolv57, dsolv65, dsolv71, dsolv78, MNDO_dipole, E_vdw, gutmann donor number, d75	<b>80.12</b>	<b>87.43</b>	<b>83.78</b>
<b>All Best Set</b>	Dsolv8, dsolv22, dsolv41, dsolv49, dsolv66, MNDO_HF, rgyr, ASA, ASA_H, RMM, logP	<b>80.17</b>	<b>86.89</b>	<b>83.53</b>

There are no overlapping descriptors between the two sets; therefore the linear correlations were calculated to determine if similar descriptors had been highlighted by the two different analysis techniques (Table 5.31). The correlation with All Best Set are presented, as it generated the highest average  $R^2$  value.

Table 5.31 Linear correlations between the two most successful sets from linear correlation and all descriptor analysis. Number in brackets is the correlation coefficient

All Best Set	Correlations within the All Best Set ( $\pm 0.8 - 1$ )	Correlations with Corr. Best Set ( $\pm 0.8 - 1$ )
Dsolv8	Dsolv22	Dsolv71 (0.85)
Dsolv22	Dsolv8	Dsolv71 (0.90) E_vdw (0.80)
Dsolv41	-	-
Dsolv49	-	-
Dsolv66	-	Dsolv65 (0.96)
MNDO_HF	-	-
Rgyr	LogP	E_vdw (0.82)
ASA	ASA_H	-
ASA_H	ASA	-
RMM	-	-
LogP	Rgyr	E_vdw (0.84)
	6 of the best set descriptors are correlated	3 out of 8 final set descriptors are correlated with the other set

The correlation analysis showed that 6 of the descriptors within All Best Set are positively correlated. This therefore suggests that further descriptor reduction may be possible. Table 5.32 details the descriptor reduction analysis, which highlights that further optimisation, was not possible. It should also be noted that different test sets are used in each run, which can therefore affect the success of the prediction. This method has been utilised throughout this research as a means to determine how robust the model created by a set of descriptors is. It is interesting to observe that when dsolv22 and ASA are removed from the set, the average  $R^2$  value is reduced significantly, even though a correlated descriptor is present in each case.

Table 5.32 Reduction of descriptors in All Best Set. X denotes the presence of the descriptor in the set

Descriptors	All Best Set	All Best Set Opt 1	All Best Set Opt 2	All Best Set Opt 3	All Best Set Opt 4	All Best Set Opt 5	All Best Set Opt 6	All Best Set Opt 7
Dsolv8	X		X	X	X	X	X	
Dsolv22	X	X		X	X	X	X	X
Dsolv41	X	X	X	X	X	X	X	X
Dsolv49	X	X	X	X	X	X	X	X
Dsolv66	X	X	X	X	X	X	X	X
MNDO_HF	X	X	X	X	X	X	X	X
Rgyr	X	X	X	X	X	X		
ASA	X	X	X		X	X	X	X
ASA_H	X	X	X	X		X	X	
RMM	X	X	X	X	X	X	X	X
LogP	X	X	X	X	X		X	X
<b>FormRules average <math>R^2</math> (%)</b>	<b>80.17</b>	<b>80.17</b>	<b>75.07</b>	<b>80.17</b>	<b>80.17</b>	<b>81.04</b>	<b>80.17</b>	<b>80.17</b>
<b>INForm average <math>R^2</math> (%)</b>	<b>86.89</b>	<b>85.06</b>	<b>51.19</b>	<b>56.25</b>	<b>72.02</b>	<b>72.87</b>	<b>75.91</b>	<b>74.18</b>
<b>Overall average <math>R^2</math> (%)</b>	<b>83.53</b>	<b>82.62</b>	<b>63.13</b>	<b>68.21</b>	<b>76.10</b>	<b>76.96</b>	<b>78.04</b>	<b>77.18</b>

Only 3 of the 8 descriptors in the Corr. Best Set were correlated with those in the All Best Set, therefore optimisation of all 19 descriptors presented is shown in Table 5.33. Since dsolv8, dsolv22, dsolv66, rgyr and logP are correlated with descriptors in the more successful Corr. Best Set, only the remaining six descriptors will be analysed.

Table 5.33 Linear correlations between the two most successful sets from linear correlation and all descriptor analysis

Descriptors	Corr. Best Set	Opt 1	Opt 2	Opt 3	Opt 4	Opt 5	Opt 6	Opt 7	Opt 8
Dsolv57	X	X	X	X	X	X	X	X	X
Dsolv65	X	X	X	X	X	X	X	X	X
Dsolv71	X	X	X	X	X	X	X	X	X
Dsolv78	X	X	X	X	X	X	X	X	X
MNDO_dipole	X	X	X	X	X	X	X	X	X
E_vdw	X	X	X	X	X	X	X	X	X
Gutmann donor number	X	X	X	X	X	X	X	X	X
D75	X	X	X	X	X	X	X	X	X
Dsolv41		X						X	X
Dsolv49			X					X	X
ASA				X				X	
ASA_H					X			X	
MNDO_HF						X		X	
RMM							X	X	X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>	<b>80.12</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>87.43</b>	<b>83.02</b>	<b>80.90</b>	<b>76.46</b>	<b>75.54</b>	<b>76.45</b>	<b>82.57</b>	<b>68.57</b>	<b>79.75</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>83.78</b>	<b>81.57</b>	<b>80.51</b>	<b>78.29</b>	<b>77.83</b>	<b>78.29</b>	<b>81.35</b>	<b>74.35</b>	<b>79.94</b>

Analysis of the descriptors in All Best Set and Corr. Best Set has not led to a further optimised model. This therefore suggests that the eight descriptors in Corr. Best Set are the most important in polymorphic form prediction. In this research it is interesting to observe which descriptors have been associated with the different polymorphic form prediction. This information can be obtained by using the rules generated in FormRules<sup>[2]</sup> (summarised in Table 5.34).

Table 5.34 Rules generated in FormRules when the Corr Best Set is analysed

Rules generated for Corr. Best Set			
<b>--- Rules for property Form I ---</b>			
IF dsolv65 is MID AND rate is LOW AND Temp is LOW		THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is HIGH AND Temp is HIGH		THEN Form I is	HIGH (1.00)
<b>--- Rules for property Form II ---</b>			
SubModel:1	IF E_vdw is LOW	THEN Form II is	LOW (1.00)
	IF E_vdw is HIGH	THEN Form II is	HIGH (1.00)
SubModel:2	IF MNDO_dipole is LOW	THEN Form II is	LOW (0.84)
	IF MNDO_dipole is MID	THEN Form II is	HIGH (1.00)
	IF MNDO_dipole is HIGH	THEN Form II is	LOW (1.00)
SubModel:3	IF rate is LOW	THEN Form II is	LOW (0.91)
	IF rate is HIGH	THEN Form II is	HIGH (0.78)
<b>--- Rules for property Form III ---</b>			
SubModel:1			
IF MNDO_dipole is LOW AND Gutman donor no. is LOW		THEN Form III is	HIGH (1.00)
IF MNDO_dipole is LOW AND Gutman donor no. is HIGH		THEN Form III is	LOW (1.00)
IF MNDO_dipole is HIGH AND Gutman donor no. is LOW		THEN Form III is	LOW (1.00)
IF MNDO_dipole is HIGH AND Gutman donor no. is HIGH		THEN Form III is	HIGH (1.00)

Rules generated for Corr. Best Set continued			
--- Rules for property Form III ---			
SubModel:2	IF E_vdw is LOW	THEN Form III is	HIGH (1.00)
	IF E_vdw is HIGH	THEN Form III is	LOW (1.00)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.80)
	IF rate is HIGH	THEN Form III is	LOW (1.00)
--- Rules for property Dihydrate ---			
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv71 is LOW		THEN Dihydrate is	HIGH (1.00)
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND dsolv71 is HIGH		THEN Dihydrate is	LOW (1.00)
--- Rules for property Solvate ---			
IF MNDO_dipole is LOW AND dsolv57 is LOW		THEN Solvate is	LOW (1.00)
IF MNDO_dipole is LOW AND dsolv57 is HIGH		THEN Solvate is	HIGH (0.94)
IF MNDO_dipole is MID AND dsolv57 is LOW		THEN Solvate is	LOW (0.98)
IF MNDO_dipole is MID AND dsolv57 is HIGH		THEN Solvate is	LOW (1.00)
IF MNDO_dipole is HIGH AND dsolv57 is LOW		THEN Solvate is	LOW (1.00)
IF MNDO_dipole is HIGH AND dsolv57 is HIGH		THEN Solvate is	HIGH (0.61)

When the rules are examined it becomes apparent that only one descriptor does not feature, d75. This suggested that the descriptor d75, which is the partial negative surface area of the CBZ molecule, is not important in the predictions. FormRules<sup>[2]</sup> and INForm<sup>[1]</sup> analysis were then conducted with d75 removed, to observe the impact on prediction (Table 5.35).

Table 5.35 The comparison of Corr Best Set with and without d75

	Corr. Best Set	Corr. Best Set without d75
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.12</b>	<b>80.12</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>87.43</b>	<b>87.96</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>83.78</b>	<b>84.04</b>

The removal of d75 slightly improves the INForm<sup>[1]</sup> result and overall performance of prediction, creating a further optimised model. Therefore the final set of descriptors as determined by these techniques are dsolv57, dsolv65, dsolv71, dsolv78, MNDO\_dipole, E\_vdw and gutmann donor number. These seven descriptors were taken forward for validation and analysis of their meaning.

## 5.4. Discussion of the Descriptors in the Optimised Set

The seven descriptors found in the most successful set for predicting polymorphic form, represent a range of solvent and CBZ properties, summarised in Table 5.36.

Table 5.36 Summary of the descriptors in the most successful set

Descriptor	Meaning	Type of descriptor	Calculated from Solvent or CBZ molecule
Dsolv71	Total molecular surface area	Geometrical	Solvent
Dsolv78	Difference in partial surface areas	Charge distribution	Solvent
Dsolv57	3D bonding information content (order 0)	Topological	Solvent
Dsolv65	3D bonding information content (order 2)	Topological	Solvent
MNDO_dipole	Calculated (MNDO theory) dipole moment	Quantum chemical	CBZ
E_vdw	Van der Waals contribution to the potential energy	Quantum chemical	CBZ
Gutmann donor number	Electron donating ability	Bulk	Solvent

As previously mentioned, it is very interesting to assess which descriptors have contributed to the prediction of polymorphic form. Much of this thesis discusses the reduction of the number of descriptors used in prediction, but it is of great importance to understand what the descriptors physically mean. This manual analysis method has created a predictive model comprised of seven descriptors. Each of these seven descriptors are featured within the rules generated in FormRules<sup>[2]</sup>, allowing discussion of their contribution in polymorphic form prediction.

### 5.4.1. The Prediction of Form I

The crystallisation of form I has been often noted in the literature,<sup>[3-5]</sup> but under the experimental conditions used within this work, pure form I was never crystallised. Trace amounts of form I were observed within two crystallisations (see Electronic Appendix, Chapter 4, file 4.4 for details) and therefore the prediction of form I was included in the model. However, very little training data was available and therefore the model produced for form I is likely to be unreliable. Rules were also generated to predict form I (Table 5.37), again based upon a very small amount of training data.

Table 5.37 Rules generated in FormRules for form I prediction

Rules generated for Form I prediction			
IF dsolv65 is LOW AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (0.90)	
IF dsolv65 is MID AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is HIGH AND Temp is HIGH	THEN Form I is	HIGH (1.00)	



The rules have highlighted that dsolv65, rate and temperature are important in form I prediction. However, from Table 5.37 it is clear that in most cases the prediction for form I will be low, except when all three values are high. This is in fact comparable with the outcomes of the two experiments that produced form I. Figure 5.3 displays the normalised descriptor values for the experiments producing form I and shows that in both cases at least one of the three descriptors has a mid range or lower value. Since neither experiment produced a high level of form I as their product, the experimental results obey the rules.

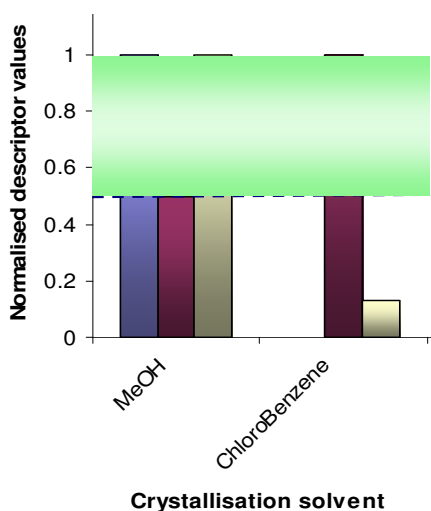


Figure 5.3 Crystallisation solvents in which form I is produced plot against the three rule descriptors, rate (blue), temperature (red) and dsolv65 (cream). The shaded area highlights the most favourable descriptor values for form I production.

Dsolv65 is a 3D bonding information content (BIC) topological descriptor for the solvent molecules. It is a measure of the structural diversity<sup>[6]</sup>, which incorporates the branching and connectivity of the molecule. It is a second order descriptor, which means that the values are based upon the connectivity of the atoms within the molecule, two bonds away from each atom in turn.<sup>[7, 8]</sup> It is calculated using Equation 5.1<sup>[7, 9]</sup>, where  $r$  represents the order of calculation,  $q$  is the “number of edges in the structural graph of the molecule”<sup>[9]</sup>,  $n$  the total number of atoms and  $n_i$  the number of atoms in the  $i$ th class in the molecule.

$$BIC_r = \frac{IC_r}{\log_2 q} \quad \text{Equation 5.1}$$

$$IC_r = -\sum_{i=1}^k \frac{n_i}{n} \log_2 \frac{n_i}{n} \quad \text{Equation 5.2}$$

The BIC descriptor, although not frequently observed in the literature, has been used in the prediction of quaternary ammonium ionic liquid melting points<sup>[10]</sup>, assessing correlations between the molecular structure and solubility<sup>[11]</sup> and in research into the solvent effects on decarboxylation<sup>[6]</sup>. Katritzky et al.<sup>[6]</sup> reported that the branching and connectivity of the solvent must impact upon solvent-solute interactions, and perhaps it is related to hydrogen bonding abilities. In other work it has been linked loosely to cavity formation during solvation<sup>[11]</sup> because it is a size related descriptor. This again suggests the link to solvent-solute interactions.<sup>[11]</sup>

Within the context of nucleation and crystallisation from solution, the interactions between the solvent molecule and the solute may play a pivotal role in which polymorphic form or solvate will be crystallised. In the case of solvate formation the solvent-solute interaction is clear, due to the presence of the solvent within the crystal structure. Perhaps the solvent-solute interactions can inhibit or promote a defined polymorph to grow based on the position and strength of the interactions. There are examples of additives that inhibit the nucleation and growth of specific polymorphs in the literature<sup>[12]</sup> and a number of discussions into the role of solvent in polymorphic crystallisation.<sup>[13-15]</sup> Sulfathiazole is a specific example of a solvent stabilising a metastable form and inhibiting conversion to a more stable polymorph.<sup>[13]</sup> This effect occurs because of a specific solvent's ability to promote or inhibit different types of intermolecular interactions.

High rates and temperatures have been commented upon in the literature as they often lead to the crystallisation of metastable forms.<sup>[16-19]</sup> A more definite picture of polymorph selection based on the descriptors in the form I rules cannot be made; this is due to the small amount of data used in the training. With an increased training set, perhaps a more obvious relationship will prevail and more firm conclusions can be made.

### 5.4.2. The Prediction of Form II

20 of the experiments carried out resulted in the crystallisation of pure form II, with a further 21 having a mixture of form II with another polymorph. These data suggest that the ability to predict this metastable form would be valuable, as it often occurs. The rules generated are presented in Table 5.38, and show the contribution to prediction of two descriptors and one experimental condition.

Table 5.38 Rules generated in FormRules for form II prediction

Rules generated for Form II prediction			
SubModel:1	IF E_vdw is LOW	THEN Form II is	LOW (1.00)
	IF E_vdw is HIGH	THEN Form II is	HIGH (1.00)
SubModel:2	IF MNDO_dipole is LOW	THEN Form II is	LOW (0.84)
	IF MNDO_dipole is MID	THEN Form II is	HIGH (1.00)
	IF MNDO_dipole is HIGH	THEN Form II is	LOW (1.00)
SubModel:3	IF rate is LOW	THEN Form II is	LOW (0.91)
	IF rate is HIGH	THEN Form II is	HIGH (0.78)

E\_vdw, MNDO\_dipole and rate have all been highlighted by FormRules<sup>[2]</sup> as important descriptors in form II prediction. Each of the experiments that crystallised pure form II were plotted with these descriptors to assess the immediate accuracy of the rules. E\_vdw was highlighted to be the most confident rule (demonstrated by the colouring in Table 5.38 and when the normalised values were plotted (Figure 5.4), it shows that the majority of the results match the rule.

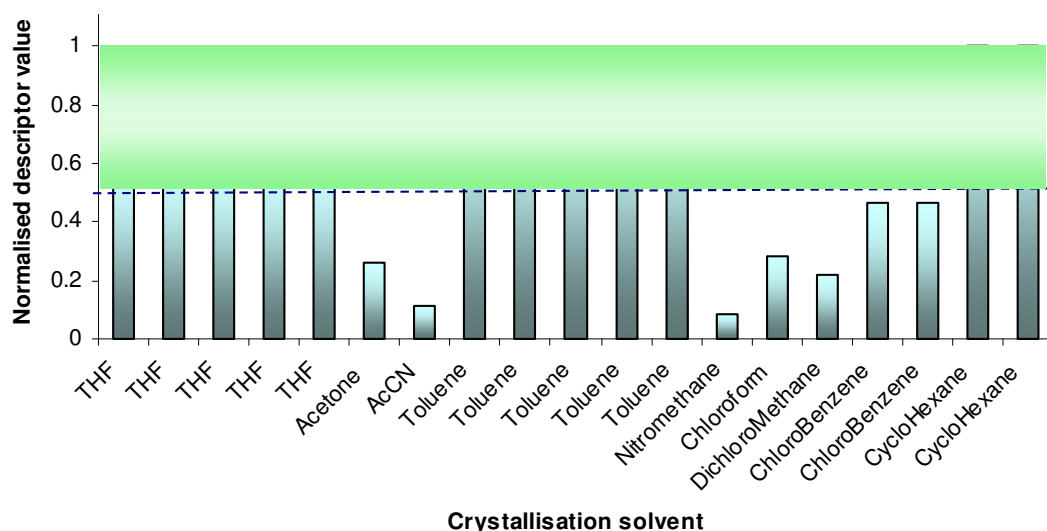


Figure 5.4 The pure form II experiments plot against the normalised  $E_{vdw}$  values. The green shaded area highlights the most favourable descriptor values for form II production.

The  $E_{vdw}$  descriptor is the van der Waals contribution to the potential energy of the CBZ molecule in a solvent forcefield. The van der Waals term can be used to describe the interactions of solvent and solute molecules<sup>[20, 21]</sup>. This descriptor was calculated for the CBZ molecule and therefore the differences between each value are very small. The slight differences in the geometry of the CBZ molecules are brought about by the modelled solvent forcefield, with these subtle changes altering the  $E_{vdw}$  value. Figure 5.5 shows the van der Waals interaction surface of the CBZ molecule (calculated in MOE<sup>[22]</sup>). If the CBZ molecular geometry was more compact, the interaction surface would be reduced, potentially affecting the molecular interactions.

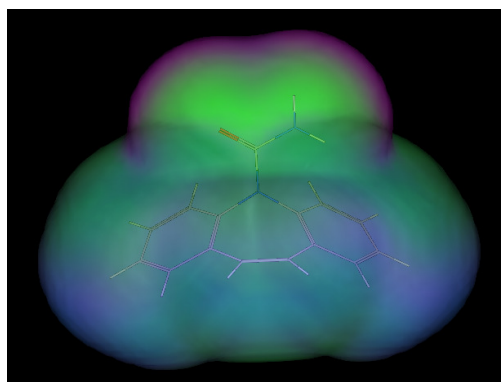


Figure 5.5 Diagram of the van der Waals interactions represented as a molecular surface. H-bonding region (pink), hydrophobic region (green) and mild polar region (blue) coloured upon the surface<sup>[22]</sup>

There are examples in the literature suggesting that the solvent-solute interactions are very important in the crystallisation of different polymorphic form.<sup>[13, 15, 23]</sup> Dunitz<sup>[23]</sup> commented specifically on the stabilising effect of coulombic interactions in the polymorphs of ROY.

The rule states that a high  $E_{vdw}$  value will lead to a high form II prediction, which perhaps indicates a greater interaction with the molecules in solution. When the crystal structure of form II is observed, there are voids within the structure, unlike other CBZ polymorphic forms. Previous research<sup>[24, 25]</sup> has suggested that solvent molecules may be included in the crystal structure of form II at very low levels, with toluene being used as an example<sup>[24]</sup>. This suggests there is a greater interaction with the solvent molecules than in other polymorphic forms. The experimental work in this thesis indicated that crystallisation in toluene often leads to form II, which supports the inclusion solvent modelling.<sup>[24]</sup>

It is interesting that a subtle solute descriptor that describes the van der Waals interactions has been highlighted. Based on the knowledge of an inclusion solvent within the form II crystal structure, perhaps this descriptor is directing our attention towards van der Waals interactions between solute and solvent molecule.

These rules hold well for pure form II experiments. However, when all of the experiments that produce any amount of form II combined with another polymorph are analysed, the majority of the  $E_{vdw}$  values are low (Figure 5.6).

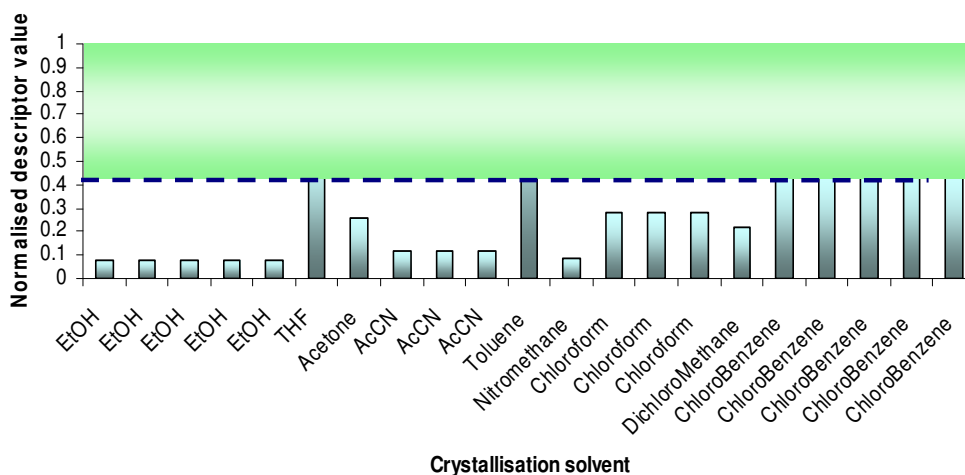


Figure 5.6 The experimental results that produced a mixture of form II and another form plot against the normalised  $E_{vdw}$  values. The shaded area highlights the most favourable descriptor values for form II production.

This could suggest that unless the solvent-solute interactions are strong, form II is unlikely to be crystallised as a pure form.

Figure 5.7 shows all the crystallisation solvents used within this research and demonstrate that from solvents with high  $E_{vdw}$  values, form II is the most likely product. There are no examples of high  $E_{vdw}$  values that do not crystallise form II.

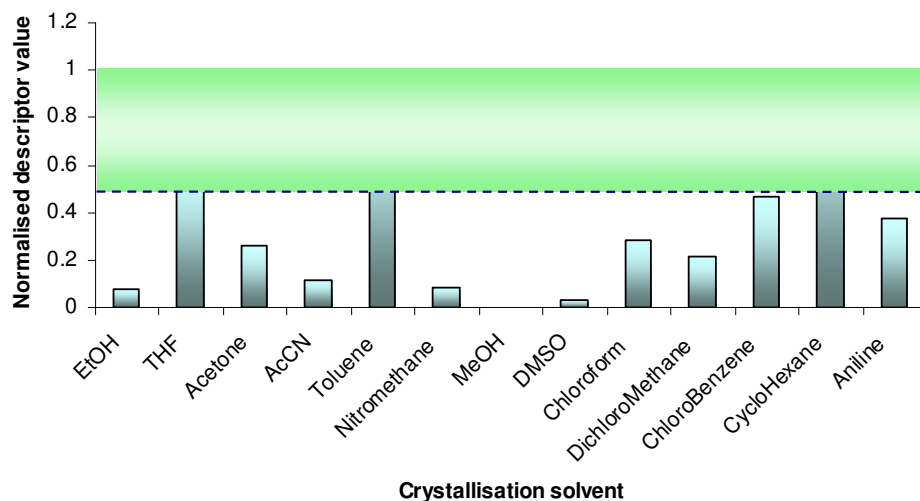


Figure 5.7 All crystallisation solvents plot against  $E_{vdw}$  normalised values. The shaded area highlights the most favourable descriptor values for form II production.

A notable correlation between the  $E_{vdw}$  descriptor values and the dielectric constant of the solvent has been observed (Figure 5.8). Dielectric constant is used within the solvent forcefield calculations and therefore plays a part in the geometry optimisation calculations. However, dielectric constant was one of the descriptors used in this analysis, but it failed to progress through the analysis individually to an optimum set of descriptors.

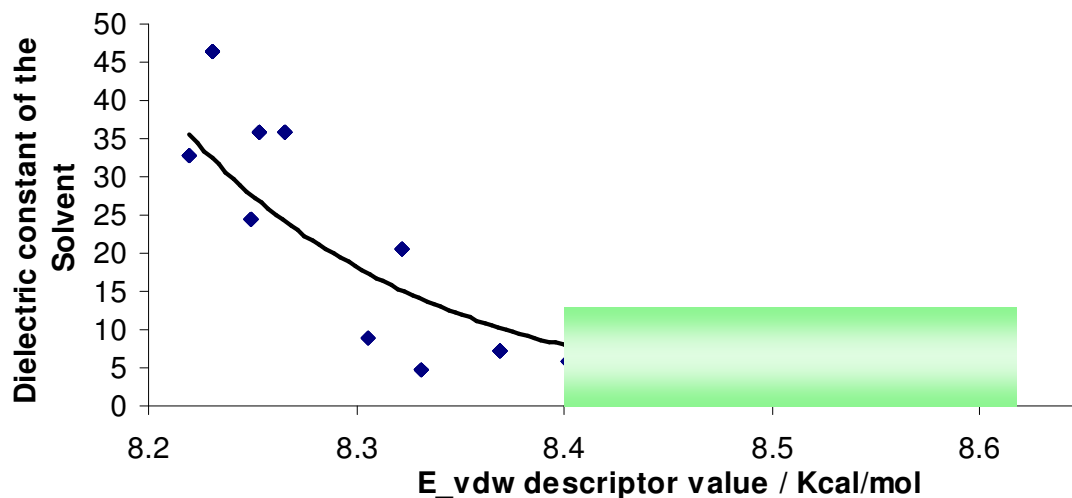


Figure 5.8 correlation of  $E_{vdw}$  and the dielectric constant of the solvents. Most favourable form II producing region is represented by the shaded area

Dielectric constants have been used in previous work<sup>[16, 26]</sup> that stated the metastable form II is most likely to be crystallised from a low dielectric solvent. The trend in Figure 5.8 agrees with this, showing that when the dielectric constant is low, the  $E_{\text{vdw}}$  value is at its highest and the rule suggests this is the most favourable region for form II crystallisation (green area on graph).

Similar analysis was carried out with the MNDO\_dipole and rate descriptors (Figure 5.9 and Figure 5.10). The rule generated for the MNDO\_dipole descriptor suggests that a medium value will most likely lead to form II being crystallised. Figure 5.9 clearly shows that most of the pure form II producing experiments fit into this range (shaded area on graph).

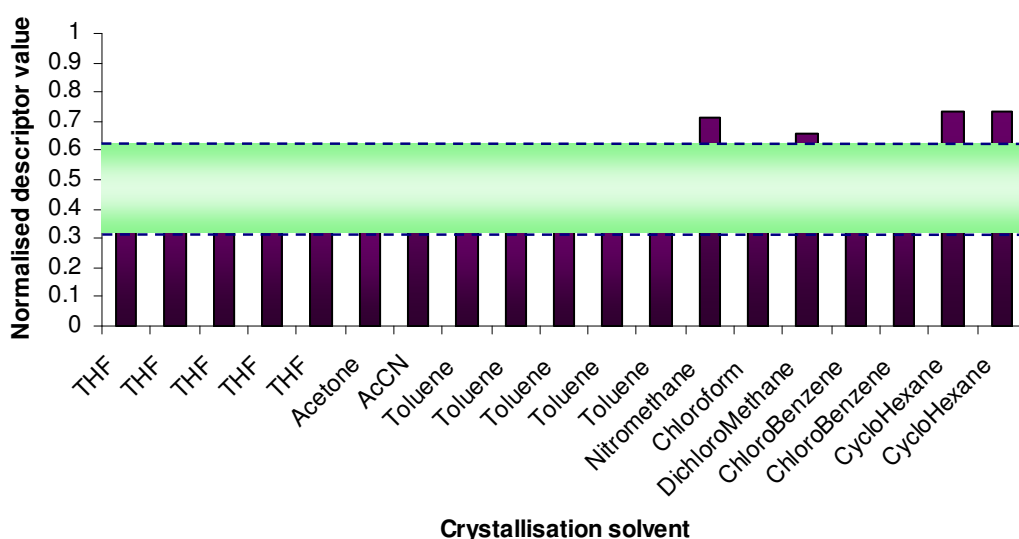


Figure 5.9 The pure form II experiments plot against the normalised MNDO\_dipole values. The shaded area highlights the most favourable descriptor values for form II production.

The MNDO\_dipole descriptor is a quantum chemical value<sup>[9, 22]</sup> that represents the calculated dipole moment of the CBZ molecule within the different solvent force fields. This descriptor gives information about the charge distribution and polarity of the molecule.<sup>[27, 28]</sup> As with the  $E_{\text{vdw}}$  descriptor, the differences between the values are very small. This is because the changes in CBZ geometry are brought about by the different interactions with the solvent force fields. With previously finding a correlation between  $E_{\text{vdw}}$  and dielectric constant, the MNDO\_dipole values were also plotted against the bulk descriptors to determine if there were any trends in the data (appendix section 12.9). No clear trends were observed between MNDO\_dipole and the bulk descriptors. MNDO\_dipole is also linearly uncorrelated to any other descriptor within the whole set, which makes its appearance in the optimised set

more interesting as it must play an important part in the predictions. Similarly to the  $E_{\text{vdw}}$  descriptor it describes the solvent-solute interactions but based upon polarity. When the evaporation rate values are plotted for pure form II producing experiments (Figure 5.10), the rule does not hold strongly. The majority of the rate values are high, but Figure 5.10 shows that there must be other influences upon the crystallisation that leads to form II production.

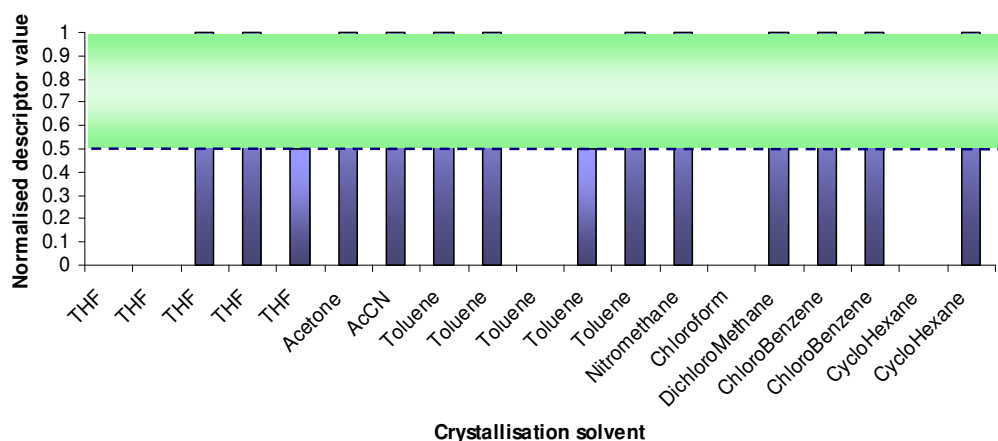


Figure 5.10 The pure form II experiments plot against the normalised rate values. The shaded area highlights the most favourable descriptor values for form II production.

The rule of a high evaporation rate producing a metastable polymorph is a sensible one, and has been commented upon in previous research<sup>[16-19, 29, 30]</sup>. By conducting a crystallisation at a high evaporation rate, high levels of supersaturation are reached more quickly. When the solution is supersaturated, nucleation can occur and if it follows Ostwald's Rule of Stages, the least stable polymorph would crystallise first<sup>[31, 32]</sup>. The least stable form of CBZ is form II. Research by Getsoian et al.<sup>[30]</sup> stated that at high supersaturations crystallisation is under kinetic control and therefore produces the metastable form.

Rules that incorporate temperature and rate are very useful in practical crystallisations, but fail to uncover molecular level properties that also impact on the formation of different polymorphic forms.

### 5.4.3. The Prediction of Form III

Form III is the thermodynamically stable form of CBZ and as such was produced as the pure product in thirty six experiments. A further twenty four experiments crystallised as a mixture of forms, including form III. To be able to predict the



experimental space in which the most stable form of a molecule would be produced, would be highly beneficial to the pharmaceutical industry. Experiments could be focused on a specific experimental region, saving time and money. Rules were generated for the prediction of form III (Table 5.39), highlighting MNDO\_dipole, Gutmann donor number, E\_vdw and rate as important descriptors and experimental conditions.

Table 5.39 Rules generated in FormRules for form III prediction

Rules generated for Form III prediction			
SubModel:1			
IF MNDO_dipole is LOW AND Gutmann donor no. is LOW	THEN Form III is	HIGH (1.00)	
IF MNDO_dipole is LOW AND Gutmann donor no. is HIGH	THEN Form III is	LOW (1.00)	
IF MNDO_dipole is HIGH AND Gutmann donor no. is LOW	THEN Form III is	LOW (1.00)	
SubModel:2	IF E_vdw is LOW	THEN Form III is	HIGH (1.00)
	IF E_vdw is HIGH	THEN Form III is	LOW (1.00)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.80)
	IF rate is HIGH	THEN Form III is	LOW (1.00)

The most significant rule as highlighted by FormRules<sup>[2]</sup> contains MNDO\_dipole and the Gutmann donor number. MNDO\_dipole featured in the rules for form II, with a medium value leading to a successful prediction. In the prediction of form III it works in tandem with the Gutmann donor number of the solvent. As mentioned previously, the MNDO\_dipole descriptor is the calculated dipole moment of the CBZ molecule in the different solvent force fields, and gives information about the charge distribution and polarity.<sup>[27, 28]</sup> The Gutmann donor number (DN) is a bulk solvent descriptor and quantifies the basicity or electron donating ability of a solvent.<sup>[33-35]</sup> It is based upon solute-solvent interactions interacting like acid-base reactions<sup>[33]</sup> and was defined by Gutmann “as the negative  $\Delta H$  value in kcal/mol for the interaction of the electron pair donor solvent with  $\text{SbCl}_5$  in a highly diluted solution of dichloroethane”.<sup>[33]</sup>

Figure 5.11 plots the normalised MNDO\_dipole and DN for each solvent that crystallised pure form III. When both descriptor values are low (the shaded area on the graph) the prediction of form III is at its highest.

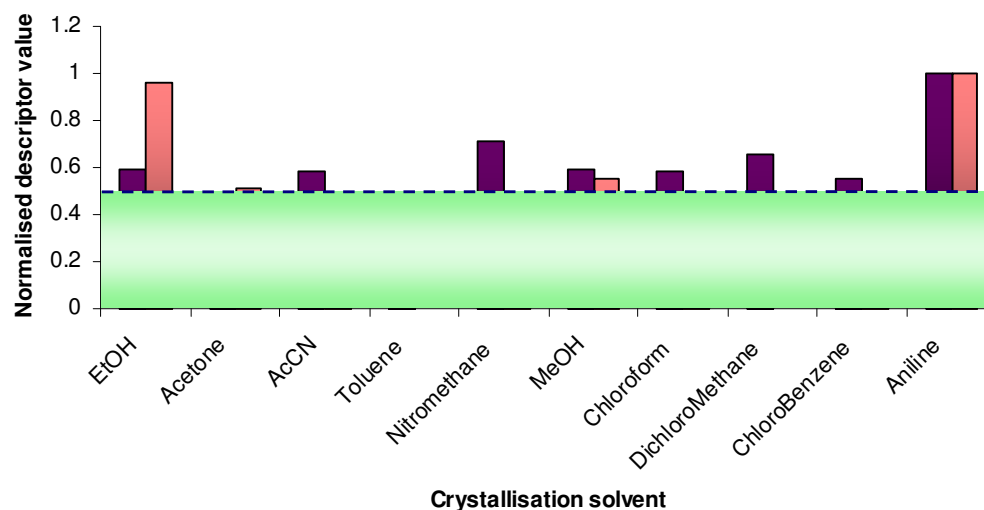


Figure 5.11 The pure form III experiments plot against the normalised MNDO\_dipole (purple) and Gutmann donor number values (pink). The shaded area highlights the most favourable descriptor values for form III production

Figure 5.11 demonstrates that this rule is generally true based on the experimental data generated in this research. This rule uses a CBZ descriptor that describes solvent-solute interactions in terms of polarity, and a bulk solvent property that quantifies the hydrogen bond donation ability. Previous research by Kelly et al.<sup>[15]</sup> discussed the importance of hydrogen bonding ability of the solvent in the preferential crystallisation of different polymorphic forms. Their research concluded that solvents with hydrogen bond acceptor capabilities preferentially crystallised CBZ form II. They also stated that a hydrogen bond donor/acceptor ratio of more than 0.5 led to concomitant crystallisation of forms II and III. No comment was made about the hydrogen bond donor ability of the solvent leading to pure form III crystallisation.

Using the Gutmann donor (DN) and acceptor numbers (AN) from this research, which are different to the values used in Kelly et al.<sup>[15]</sup>, analysis of the most favourable interactions has been carried out. DN and AN were not available for the CBZ molecule, therefore a structurally similar molecule was found. There were no DN and AN values for acetamide, so values for formamide were used.<sup>[36, 37]</sup> Figure

5.12 shows the hydrogen bonding between the formamide and dichloromethane and nitromethane.

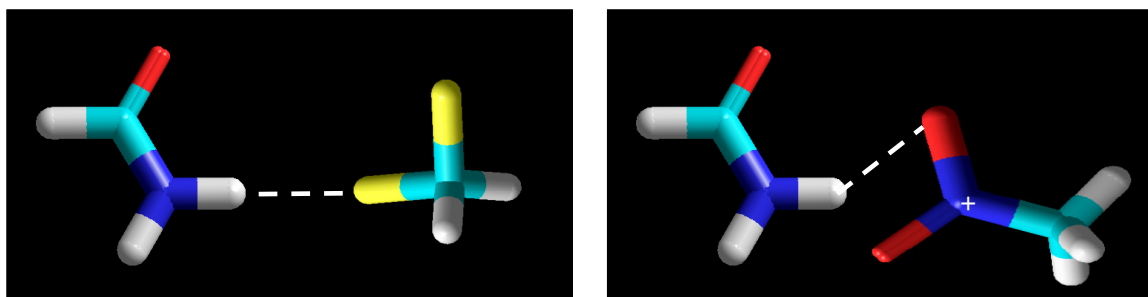


Figure 5.12 Hydrogen bonding between formamide and dichloromethane (left) and nitromethane (right)

Both dichloromethane and nitromethane produce form III and have a low DN, conforming to the rules generated. In a similar manner to the research by Kelly et al.<sup>[15]</sup> the donor acceptor ratio was calculated (Table 5.40).

Table 5.40 Solvent Gutmann donor and acceptor numbers

Solvent	Gutmann donor number (DN)	Gutmann acceptor number (AN)	DN/AN ratio
Dichloromethane	0.0	20.4	0.0
Nitromethane	2.7	20.5	0.1

Table 5.40 shows that the donor acceptor ratio is very low within these solvents, perhaps highlighting a potential reason for the formation of the stable form III. Table 5.41 shows the DN and AN values for the formamide, giving a much higher ratio. The higher ratio may suggest that the formamide is more likely to participate in solute-solute interactions, rather than a solvent-solute interaction (Figure 5.13).

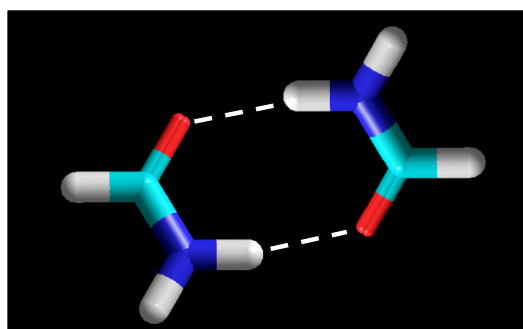


Figure 5.13 Hydrogen bonding in formamide

Table 5.41 Dimethylacetamide Gutmann donor and acceptor numbers

Example Solute	Gutmann donor number (DN)	Gutmann acceptor number (AN)	DN/AN ratio
Formamide	24	39.8	0.6

When the donor acceptor ratio is calculated based upon solvent-solute interactions and solvent-solvent interactions, perhaps it offers further information as to what is occurring in solution (Table 5.42).

Table 5.42 Solvent-solute and solvent-solvent interactions based on DN and AN ratios

Donor molecule	Acceptor molecule	Gutmann donor number (DN)	Gutmann acceptor number (AN)	DN/AN ratio
Formamide	Formamide	24	39.8	0.6
Formamide	Dichloromethane	24	20.4	1.2
Dichloromethane	Formamide	0.0	39.8	0.0
Formamide	Nitromethane	24	20.5	1.2
Nitromethane	Formamide	2.7	39.8	0.1
Formamide	Ethanol	24	37.1	0.6
Ethanol	Formamide	32	39.8	0.8
Ethanol	Ethanol	32	37.1	0.9
Formamide	Aniline	24	28.8	0.8
Aniline	Formamide	33.3	39.8	0.8
Aniline	Aniline	33.3	28.8	1.2

When only the donating abilities of the low DN value solvents are taken into account, the solute-solute interactions are more favourable. However, the low DN value solvents have a high AN and therefore still might interact with the solute. At high DN values solvent-solvent interactions are more likely to dominate in solution. Overall, no firm conclusions can be made about the role of hydrogen bonding in solution, but as a general rule, low DN solvents lead to form III crystallisation.

The DN has been used in earlier research regarding the coordination of transition metal ions.<sup>[38]</sup> Transition metal interactions have been observed with solvents that

display a range of DN values. However, the shape of the solvent molecule has a part to play in the interaction due to steric reasons, which restrict the coordination.<sup>[38]</sup> Although the solutes used in these crystallisations are not as large as the transition metal complexes discussed in Gutmann's work<sup>[38]</sup>, it may be a reason as to why there are empirical correlations between the forms produced and the crystallisation solvent used based on only the DN values.

$E_{\text{vdw}}$  is also featured in the form III rules, stating that a low value leads to a form III prediction. Figure 5.14 shows that all but one of the solvents that crystallise pure form III has a low  $E_{\text{vdw}}$  value.

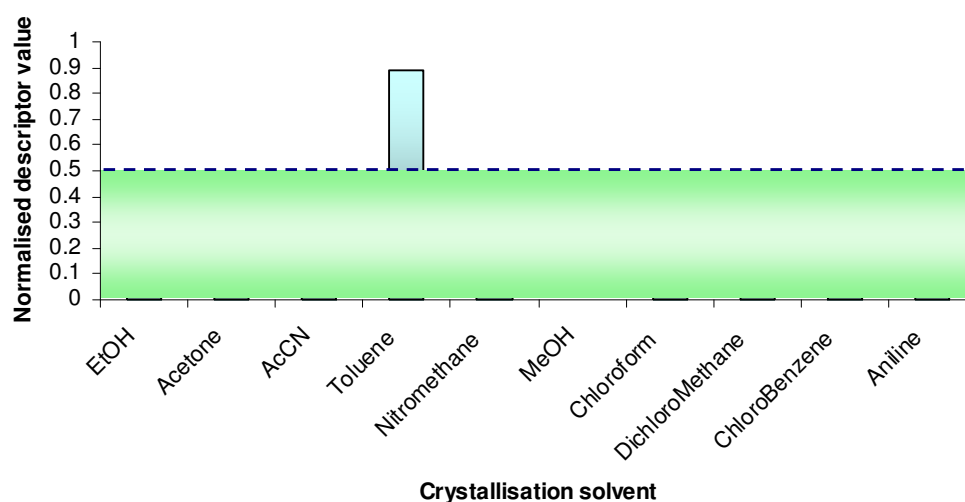


Figure 5.14 The pure form III experiments plot against the normalised  $E_{\text{vdw}}$  values. The shaded area highlights the most favourable descriptor values for form III production.

Toluene is the only solvent that produced pure form III that has a high  $E_{\text{vdw}}$  value. What is interesting about this result is that from seven experiments, five generated pure form II, one pure form III and a mixture. From the experimental results it would seem likely that toluene crystallisations would lead to form II, and that the model created by the artificial neural network would also generally suggest this.

Figure 5.8 presented the relationship between  $E_{\text{vdw}}$  and dielectric constant, displayed again in Figure 5.15, with the form III producing region highlighted in green. The relationship between high dielectric constants and the production of the thermodynamically stable form has been commented upon in the literature.<sup>[16, 26]</sup>

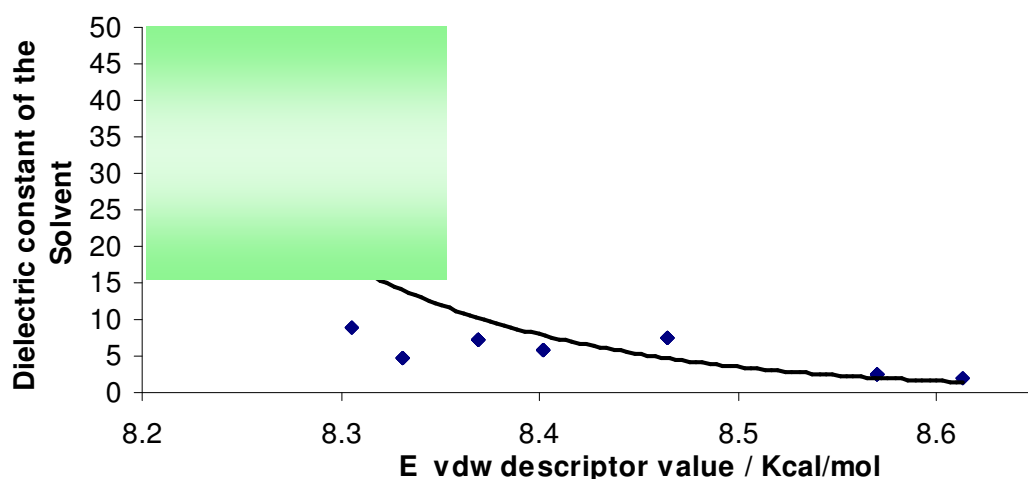


Figure 5.15 E\_vdw plot against the dielectric constants of the solvents used in the crystallisations. The shaded area on the graph represents the potentially form III producing values

Perhaps the lower E\_vdw value indicates that there are fewer interactions with the solvent, and that the hydrogen bonding between the solute molecules is the most prominent interaction. Form III is the most stable polymorph due to the number of stabilising interactions within the crystal structure.

The third rule generated for form III prediction stated that a low evaporation rate would generate high form III results. Similarly to the discussion that a high rate favours metastable form crystallisation, a low rate favours the thermodynamically stable form<sup>[16, 19]</sup>. Figure 5.16 shows the pure form III experiments normalised rate values. A general trend can be found in the data that in most pure form III forming experiments the rate is at a lower value.

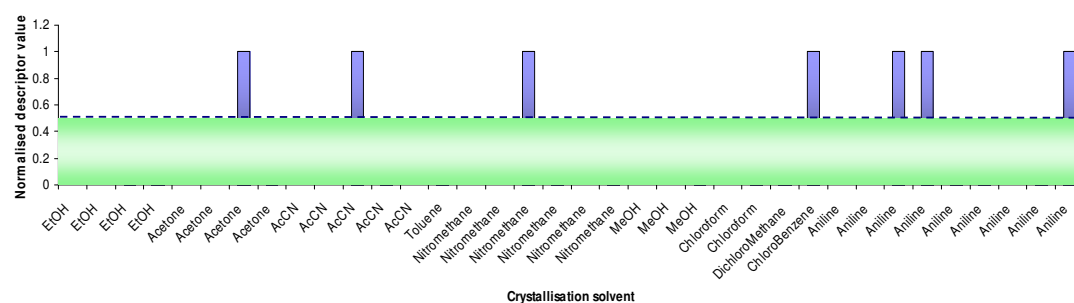


Figure 5.16 The pure form III experiments plot against the normalised rate values. The shaded area highlights the most favourable descriptor values for form III production.

#### 5.4.4. The Prediction of the Dihydrate

CBZ readily forms a dihydrate on contact with water<sup>[39-41]</sup>, but due to the scale of the crystallisations carried out within this research pure dihydrate was not formed at a large enough yield for analysis. Low levels of the dihydrate form were observed within this research, possibly suggesting the presence of water within some solvents. Similarly to form I, there were very little data available to generate a reliable prediction for the dihydrate. However, since it did feature in the results, a prediction was attempted, with rules presented in Table 5.43.

Table 5.43 Rules generated in FormRules for Dihydrate prediction

Rules generated for dihydrate prediction		
SubModel:1		
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv71 is LOW	THEN Dihydrate is	LOW (0.69)
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv71 is HIGH	THEN Dihydrate is	LOW (0.98)
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv71 is LOW	THEN Dihydrate is	LOW (0.73)
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv71 is HIGH	THEN Dihydrate is	HIGH (1.00)
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv71 is HIGH	THEN Dihydrate is	LOW (0.98)

Rules generated for dihydrate prediction continued			
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv71 is LOW	THEN Dihydrate is	HIGH (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv71 is HIGH	THEN Dihydrate is	HIGH (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv71 is LOW	THEN Dihydrate is	HIGH (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv71 is HIGH	THEN Dihydrate is	LOW (0.66)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv71 is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv71 is HIGH	THEN Dihydrate is	LOW (1.00)	
SubModel:2			
IF dsolv78 is LOW AND rate is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv78 is LOW AND rate is MID	THEN Dihydrate is	LOW (1.00)	
IF dsolv78 is LOW AND rate is HIGH	THEN Dihydrate is	LOW (0.98)	
IF dsolv78 is HIGH AND rate is LOW	THEN Dihydrate is	LOW (1.00)	
IF dsolv78 is HIGH AND rate is MID	THEN Dihydrate is	LOW (1.00)	
IF dsolv78 is HIGH AND rate is HIGH	THEN Dihydrate is	LOW (0.99)	

Three different descriptors (dsolv65, dsolv71 and dsolv78) and the two experimental conditions feature in these rules. It should be noted that sub model 2 produces a rule



that always leads to a low prediction of the dihydrate form, therefore very little can be learnt from this rule. Dsolv78 represents the difference between the positively and negatively charged surface areas on each solvent molecule. This type of charged partial surface area descriptor has been used in the literature in various areas of research.<sup>[42-46]</sup> Bodor et al.<sup>[43]</sup> commented upon the role of charge density upon the solute molecule when entering a solution, effecting the solvent-solute interactions.

The charged partial surface area descriptors have also been referred to as a measure of “weak intermolecular interactions”.<sup>[27]</sup> Unfortunately due to the limited amount of training data for the dihydrate and also the lack of a high dihydrate prediction, no further analysis into the physical meaning of dsolv78 has been carried out.

The main rule generated for the prediction of the dihydrate form involved dsolv65, dsolv71, rate and temperature. Dsolv65 also featured in the form I rules and is the 3D bonding information content (BIC) topological descriptor for the solvent molecules. The descriptor describes the branching and connectivity of the solvent molecule<sup>[6]</sup>. Dsolv71 is the total molecular surface area of the solvent molecule (TMSA). The TMSA descriptor belongs to the charged partial surface area (CPSA) group of descriptors, but represented the total geometry of the molecule. The van der Waals radii of each atom within the molecule is represented by spheres that overlap with one another (Figure 5.17), creating a molecular surface<sup>[44]</sup>. In the case of TMSA, a solvent molecule, most commonly water with a van der Waals radius of 1.5 Å<sup>[44]</sup>, is used to trace a path around the molecule, generating a solvent accessible surface area (Figure 5.17). This solvent accessible surface area is used in the charged partial surface area (CPSA) calculations and is why TMSA belongs to the CPSA set of descriptors<sup>[44]</sup>.

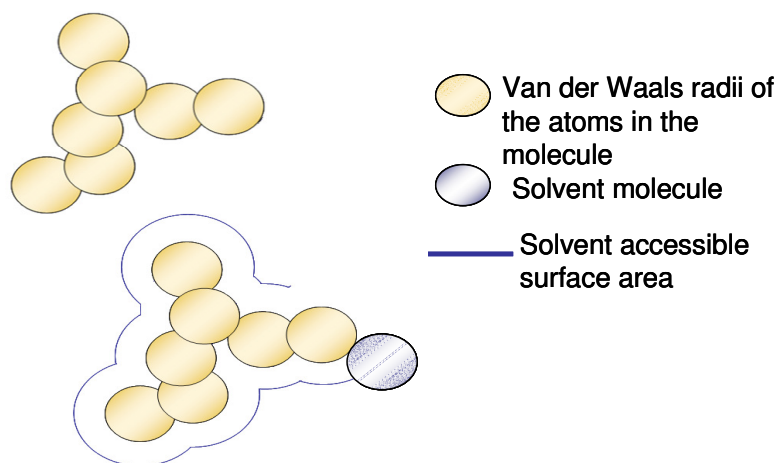


Figure 5.17 Calculation of the total molecular surface area using van der Waals radii, adapted from<sup>[44]</sup>

The solvent accessible surface area is a property that has been used in protein research to identify cavities in the structure.<sup>[47]</sup> This could also be applicable to complex or flexible small molecules. In a flexible molecule, like a protein, the solvent accessible surface area could highlight the extent of interaction certain atoms within the molecule have with the bulk solvent<sup>[47]</sup>, and therefore may be linked with the interactions in solution. If the molecule is highly flexible then larger areas may interact with the solvent or other solute molecules leading to preferential nucleation of certain polymorphs.<sup>[13, 15]</sup>

The roles of rate and temperature have been discussed previously (sections 5.4.1 and 5.4.2), with higher evaporation rates at high temperatures often leading to the metastable crystalline product.<sup>[16-19]</sup>

Based on the rules in Table 5.43 in order to generate a high dihydrate prediction, dsolv65 must be high, rate must be a mid range value and temperature and dsolv71 must be low. Figure 5.18 shows the normalised descriptor values for each experiment that generated dihydrate in the final product.

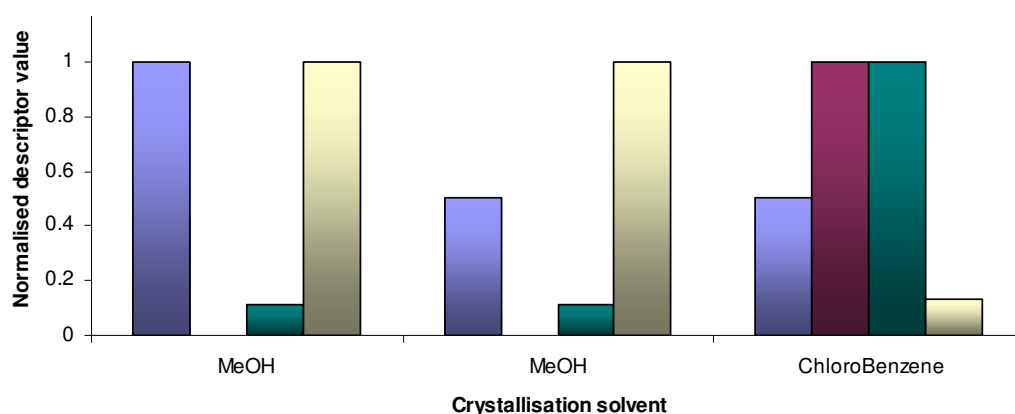


Figure 5.18 The dihydrate producing experiments plot against the normalised values of rate (blue), temperature (purple), dsolv71 (green) and dsolv65 (cream)

Due to the small amount of training data available, it is clear that the rule simply states facts based on the three occurrences of the dihydrate. In order to generate a more reliable model from which meaning can be taken from the rules, more dihydrate forming experiments need to be added to the training set.

### 5.4.5. The Prediction of Solvates

CBZ is known to produce many different solvates<sup>[25, 39, 48-52]</sup>, but only the DMSO solvate was crystallised in this research. Nine different experimental conditions were

used, generating the solvate on each occasion. Rules for solvate formation were created (Table 5.44) featuring MNDO\_dipole and dsolv57.

Table 5.44 Rules generated in FormRules for solvate prediction

Rules generated for Solvate prediction			
SubModel:1			
IF MNDO_dipole is LOW AND dsolv57 is LOW	THEN Solvate is	LOW (1.00)	
IF MNDO_dipole is LOW AND dsolv57 is HIGH	THEN Solvate is	HIGH (0.94)	
IF MNDO_dipole is MID AND dsolv57 is LOW	THEN Solvate is	LOW (0.98)	
IF MNDO_dipole is MID AND dsolv57 is HIGH	THEN Solvate is	LOW (1.00)	
IF MNDO_dipole is HIGH AND dsolv57 is LOW	THEN Solvate is	LOW (1.00)	
IF MNDO_dipole is HIGH AND dsolv57 is HIGH	THEN Solvate is	HIGH (0.61)	

The MNDO\_dipole descriptor has featured in the rules for form II and III and is the calculated dipole moment of the CBZ molecule in the different solvent force fields. Dsolv57 has not been seen previously, but like dsolv65, is a 3D bonding information content (BIC) descriptor for the solvent, but in this example it is order 0. Order 0 means that atoms within the molecule are grouped together into “equivalent classes”.<sup>[7]</sup>

When the normalised values of MNDO\_dipole and dsolv57 are plotted for the DMSO experiments Figure 5.19), it shows that the most confident rule is based upon these values.

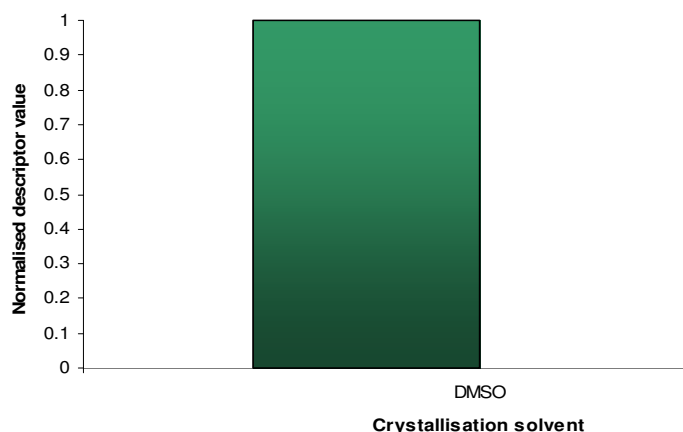


Figure 5.19 The solvate producing experiments (9 examples of DMSO solvent) plot against the normalised values of MNDO\_dipole (purple) and dsolv57 (green)

When all of the crystallisation solvents values of MNDO\_dipole and dsolv57 are plot (Figure 5.20) there are two examples of descriptor values that match the solvate forming rule. These are DMSO and acetone. Interestingly the acetone experiments never crystallised as a solvate in this research, but there are examples in the literature of its existence.<sup>[18, 26, 39, 51]</sup>

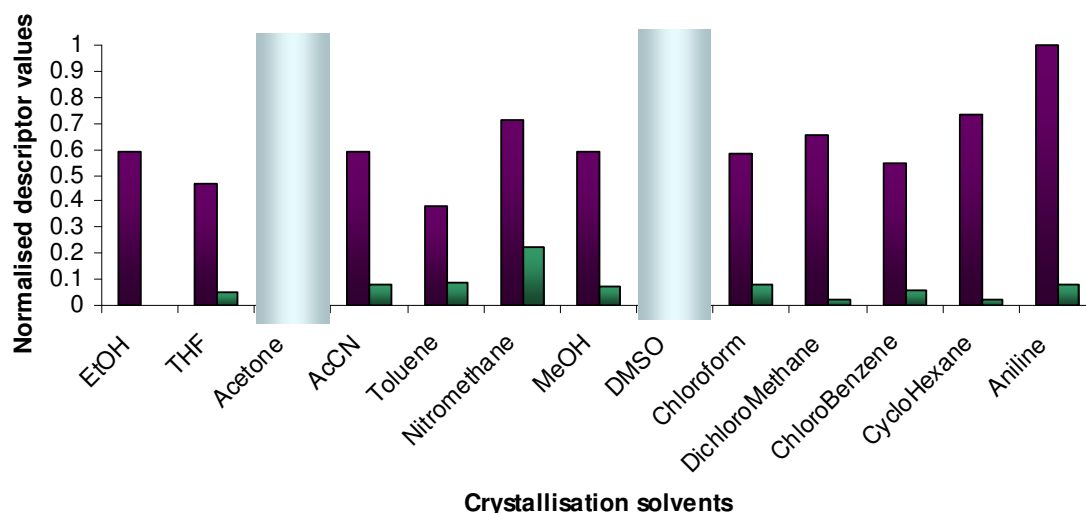


Figure 5.20 Plot of all crystallisation solvents normalised MNDO\_dipole (purple) and dsolv57 (green) values

Previous research by Johnston et al.<sup>[49]</sup> used random forest classification to predict under what conditions CBZ solvates would form. They concluded, using six parameters, the conditions required were low logP, molecular flexibility, solvent accessible surface area and molecular volume and a high dielectric constant and surface tension. All of these properties have been represented within this research, but failed to show in the final rules for solvate formation. Johnston et al.<sup>[49]</sup> did predict the formation of a nitromethane solvate, but in this research that form was not crystallised. There are many examples of CBZ solvate formation in the literature; with four solvate forming solvents were used in this research (DMSO, acetone, nitromethane and THF). To perhaps generate a more reliable rule, different solvents could be used that form solvates thus increasing the data about solvates in the training set.

#### **5.4.6. Summary of the Optimised Descriptors**

The seven descriptors present in the optimised model represent a range of properties relating to both the CBZ and solvent molecule. A summary of the descriptors involved in each rule can be found in Table 5.45. In the case of form I and the dihydrate, the lack of data generated by the crystallisations carried out reduces the reliability of these rules. With an increased number of experiments that lead to form I or dihydrate formation, the model could be rebuilt and perhaps improved. It would be unwise to draw significant conclusions from the descriptors that feature within the rules as they would be based on a very small amount of data.

The rules for form II and III involve similar descriptors that potentially describe the interactions between the solute and solvent molecules in solution. Although the differences in  $E_{\text{vdw}}$  and  $\text{MNDO\_dipole}$  values are subtle between the modelled CBZ molecules, this research has demonstrated that the effect the solvent has on the molecular geometry has an impact upon which polymorphic form is crystallised.

Similarly to the form I and dihydrate rules, the solvate rules were generated based upon one solvent crystallising as a solvate. The rule generalised that there was a possibility that acetone may also generate a solvate, something that has been observed in the literature, but not in this research. Further crystallisation work using known solvate forming solvents would lead to improved predictions of the solvate form and also generate more information, based on the descriptors used, about what is occurring at the molecular level.

Table 5.45 Summary of the descriptors involved in the CBZ predictive rules

Form predicted	Descriptor(s)	Definition(s)
I	Dsolv65	3D bonding information content (order 2) of the solvent molecule.
	Rate	Rate of nitrogen blown onto sample (L/min)
	Temperature	Temperature at which the crystallisations occurred
II	E_vdw	Van der Waals contribution to the potential energy of the CBZ molecule in a solvent force field
II	MNDO_dipole	Calculated (MNDO theory) dipole moment of the CBZ molecule in a solvent force field
II	Rate	Rate of nitrogen blown onto sample (L/min)
III	MNDO_dipole	Calculated (MNDO theory) dipole moment of the CBZ molecule in a solvent force field
	Gutmann donor number	Electron donating ability of the solvent
	E_vdw	Van der Waals contribution to the potential energy of the CBZ molecule in a solvent force field
III	Rate	Rate of nitrogen blown onto sample (L/min)
Dihydrate	Dsolv65	3D bonding information content (order 2) of the solvent molecule.
	Rate	Rate of nitrogen blown onto sample (L/min)
	Temperature	Temperature at which the crystallisations occurred
	Dsolv71	Total molecular surface area of the solvent molecule
Dihydrate	Dsolv78	Difference in partial surface areas of the solvent molecule
	Rate	Rate of nitrogen blown onto sample (L/min)
Solvate	MNDO_dipole	Calculated (MNDO theory) dipole moment of the CBZ molecule in a solvent force field
	Dsolv57	3D bonding information content (order 0) of the solvent molecule.

## **5.5. Validation of Optimised Set**

In order to assess whether the descriptors in the optimised set lead to reasonable predictions, the model needs to be validated. Two methods will be used here; firstly a cross validation method that uses 10 % of the experimental rows of data. This means that the validation set is within the experimental space of the data used in the trained network. The second method is to use data that has never been used in any training and has been generated from experiments that involve different solvents to those used in the model.

### **5.5.1. Cross Validation Results**

Nine rows of experimental results (10 %) were predicted using the model created with the remaining data. The overall average performance of the network was reduced (to 82.69 %), but this was expected as a large proportion of its training data was removed. The results are summarised in Table 5.46 and are very promising for the prediction of the major polymorphic form crystallised within the experimental space used in the training.

Table 5.46 Cross validation results

Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
					Form I	Form II	Form III	Dihydrate	Solvate
Ethanol	15	25	Form III	Form III	0.0	0.0	1.0	0.0	0.0
THF	25	25	Form II	Form II	0.0	1.2	0.6	0.1	0.0
Acetonitrile	15	50	Form III	Form III	0.0	0.3	1.1	0.0	0.0
DMSO	25	25	Solvate	Solvate	0.0	0.0	0.0	0.0	1.0
Aniline	5	50	Form III	Form III	0.0	0.0	1.0	0.0	0.0
Chlorobenzene	5	50	Form III	Form II	0.0	0.7	0.4	0.0	0.0
Toluene	15	75	Form II	Form II	0.0	0.7	0.2	0.0	0.0
Nitromethane	5	25	Form III	Form III	0.0	0.3	0.8	0.0	0.0
Chloroform	25	50	Form II / Form III	Form II	0.0	0.8	0.2	0.0	0.0



The major polymorphic form crystallised was predicted for seven out of nine of the validation set, also generating an INForm<sup>[1]</sup> average  $R^2$  value of 79.00 %. On only two occasions the major polymorphic form was incorrectly predicted. Analysis of why this occurred was conducted, and possible reasons discussed.

Firstly, the model failed to correctly predict form III as the major product for the chlorobenzene experiment in the validation set. Although the values generated in INForm<sup>[1]</sup> cannot be used to precisely determine the quantities of a mixture of forms in a product, the values can be used to assess how confident the model is that a certain form will be produced. In the case of this sample, although the major form predicted was form II, there was also a relatively high level of form III predicted (Table 5.46). The values for both these forms were lower than for other experiments, suggesting that the model could not reliably separate form II and III, and perhaps highlighting the possibility of a mixture. When the experimental data used in the training is examined, there are seven occurrences of chlorobenzene that generate a range of polymorphic outcomes. Form II is the major product in five of the experiments, but three of these are part of a polymorphic mixture. Form III is the major product in two experiments, but only pure in one instance. The model incorrectly predicted form III, but the experimental results show a range of different possible outputs. This result, although on first sight an incorrect prediction, is a valuable one. A result like this could alert the user that the solvent might produce a range of polymorphic crystals and may not be the most optimal solvent for reproducible crystallisations over a larger range of experimental conditions.

The second result incorrectly predicted was that of the crystallisation from chloroform, which produced a mixture of polymorphic forms experimentally. In this example the model predicted form II confidently for the set of experimental conditions and descriptors tested. When the training data for chloroform are analysed there are six experimental results. These chloroform crystallisations led to a combination of pure form II (1 occurrence), pure form III (2 occurrences) and mixtures of both form II and III. When this data is observed it provides a reason for the incorrect prediction of the model. However, no suggestion of form III crystallisation can be taken from the INForm<sup>[1]</sup> results. There are a number of ways in which this result can be viewed. Firstly, it would be highly beneficial for more experimental data to be obtained. This further experimentation could determine if one crystalline product is more frequently formed or if in fact a range of polymorphs

are seen in this experimental space. A second way to view these results is to note that a range of polymorphs were crystallised and the model could not generalise for this solvent well. However, from previous work<sup>[16]</sup> form II is the expected crystalline product in low dielectric solvents such as chloroform. Therefore the high occurrence of form III in this experimental work is surprising. With this in mind, it would seem that the model has generalised more successfully than initially thought.

Overall, the model created by the seven descriptors (dsolv57, dsolv65, dsolv71, dsolv78, MNDO\_dipole, E\_vdw and gutmann donor number) successfully predicted the major polymorphic form crystallised within the experimental space used in training.

### **5.5.2. External Validation Results**

In order to assess how robust the model created is, external validation has been carried out. Further crystallisations were conducted using different solvents to those used in the training of the model. This validation assesses whether the descriptors highlighted in the analysis are capable of generating a successful prediction for unknown solvents. Ethyl acetate (EtOAc) and n-butanol (n-BuOH) were used as validation solvents, with crystallisation experiments being carried out over a range of rates and temperatures. The results are presented in Table 5.47, alongside the experimental conditions used in the analysis.

Table 5.47 External validation results summary

Experiment number	Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
						Form I	Form II	Form III	Dihydrate	Solvate
1	EtOAc	5	25	Form II	Form III	0.0	0.0	0.4	0.0	0.0
2	EtOAc	5	50	Form II	Form III	0.0	0.5	0.8	0.0	0.0
3	EtOAc	25	25	Form II	Form II	0.0	0.5	0.2	0.0	0.0
4	EtOAc	25	50	Form II	Form II	0.0	1.0	0.2	0.0	0.1
5	EtOAc	15	25	Form II	Form III	0.0	0.0	0.2	0.0	0.0
6	EtOAc	15	50	Form II	Form II	0.0	0.5	0.3	0.0	0.1
7	nBuOH	5	25	Form II	Form III	0.0	0.9	1.0	0.0	0.1
8	nBuOH	5	50	Form III	Form III	0.0	0.6	1.1	0.0	0.0
9	nBuOH	25	25	Form III	Form III	0.0	0.6	0.9	0.0	0.0
10	nBuOH	25	50	Form II	Form II	0.0	1.1	0.4	0.0	0.0
11	nBuOH	15	25	Form II	Form III and form II	0.0	0.9	0.9	0.0	0.0
12	nBuOH	15	50	Form III	Form III	0.0	0.6	1.0	0.0	0.0

The model predicts the major polymorphic form crystallised in seven out of twelve of the validation experiments. With prediction of the experimental outcomes for n-butanol being more successful than the EtOAc crystallisations. However, overall the model has performed less successfully than expected.

The EtOAc crystallisations produced form II in every experiment, but half of the predictions made were for form III. When the values of prediction in INForm<sup>[1]</sup> are observed, it suggests that the model is not confident. Three of the predictions are correct, but only one confidently predicts this form II product (experiment 4). The other two form II predictions, although correct, are very low values. The situation is similar in the form III predictions made. One of these (experiment 2) is a slightly more confident prediction, but also shows the high presence of form II, the others both generated a very low value of form III only.

The n-BuOH crystallisations produce a range of pure form II and form III products, with the model successfully predicting four out of six of the experiments. The three n-butanol experiments carried out at 50°C produced confident, correct predictions. Experiment 9, although correctly predicted the major form crystallised, also presented a high value for form II.

Experiment 11 predicted form II and III at an equal value. Although form II was crystallised in that experiment, overall a mixture of the two forms have been generated.

In experiment 7, even though the major form was incorrectly predicted, high values for both form II and III were presented, possibly suggesting a mixture.

When the distribution of descriptor values for the two validation solvents are assessed, (Figure 5.21) it highlights that the descriptor space has not been adequately sampled. Only the E\_vdw descriptor values are significantly different in the EtOAc and n-butanol solvents. Perhaps to generate a more reliable validation, solvents would need to be selected to ensure the whole space was interrogated.

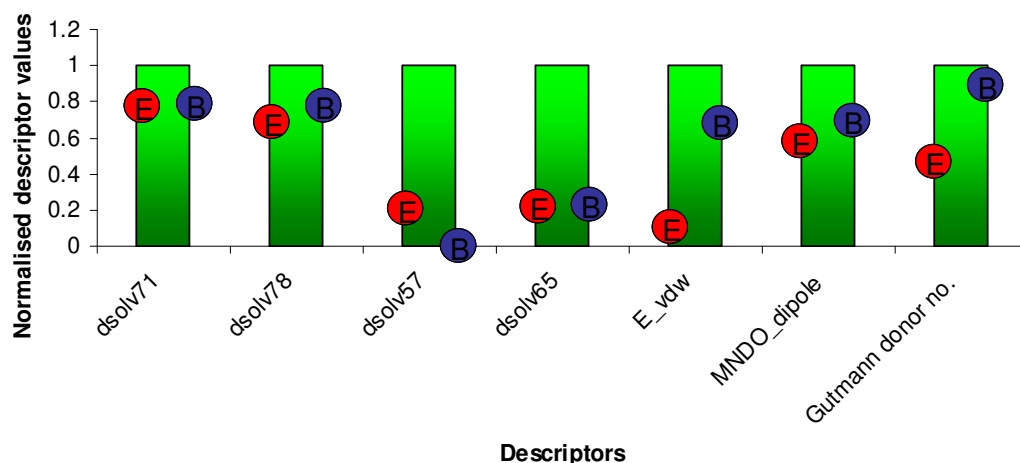


Figure 5.21 Plot of the normalised descriptor values of EtOAc (E) and n-butanol (B)

## 5.6. Conclusion of Manual Data Analysis

By analysing the linear correlations within the whole descriptor set, an optimised set of seven descriptors has been found. These seven descriptors are dsolv57, dsolv65, dsolv71, dsolv78, MNDO\_dipole, E\_vdw and gutmann donor number. In combination with the rates and temperatures used in the experimental work, this set of descriptors can successfully predict the major polymorphic form in 79 % of the cross validation experiments. When two unknown solvents were used as further test of the model, seven out of twelve of the experimental products could be predicted. To improve the model more data would be required, particularly with form I, dihydrate and solvate crystalline products. However, the model created from these seven descriptors would successfully allow polymorphic form prediction within the experimental space used in the training.

- [1] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [2] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [3] C. Rustichelli, G. Gamberini, V. Ferioli, M. C. Gamberini, R. Ficarra, S. Tommasini, *Journal of Pharmaceutical and Biomedical Analysis* **2000**, 23, 41.
- [4] A. Grzesiak, M. Lang, K. Kim, A. J. Matzger, *Journal of Pharmaceutical Sciences* **2003**, 92, 2260.
- [5] J. A. McMahon, P. Timmins, A. C. Williams, P. York, *Journal of Pharmaceutical Sciences* **1996**, 85, 1064.
- [6] A. R. Katritzky, S. Perumal, R. Petrukhim, *Journal of Organic Chemistry* **2001**, 66, 4036.
- [7] J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, Amsterdam, **1999**.
- [8] S. C. Basak, A. T. Balaban, G. D. Grunwald, B. D. Gute, *Journal of Chemical Information and Computer Sciences* **2000**, 40, 891.
- [9] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [10] D. M. Eike, J. F. Brennecke, E. J. Maginn, *Green Chemistry* **2003**, 5, 323.
- [11] A. R. Katritzky, D. B. Tatham, *Journal of Chemical Information and Computer Sciences* **2001**, 41, 358.
- [12] R. Dowling, R. J. Davey, R. A. Curtis, G. Han, S. K. Poornachary, P. S. Chow, R. Tan, B. H., *Chemical Communications* **2010**, 46, 5924.
- [13] M. M. Parmar, O. Khan, L. Seton, J. L. Ford, *Crystal Growth and Design* **2007**, 7, 1635.
- [14] D. Musumeci, C. A. Hunter, J. F. McCabe, *Crystal Growth and Design* **2010**, 10, 1661–1664.
- [15] R. C. Kelly, N. Rodriguez-Hornedo, *Organic Process Research and Development* **2009**, 13, 1291.
- [16] J. F. McCabe, *CrystEngComm* **2010**, 12, 1110.
- [17] T. Threllfall, *Organic Process Research & Development* **2003**, 7, 1017.
- [18] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, 95, 1918.
- [19] J. L. Hilden, C. E. Reyes, M. J. Kelm, J. S. Tan, J. G. Stowell, K. R. Morris, *Crystal Growth and Design* **2003**, 3, 921.
- [20] H. Hu, W. Yang, *Journal of Physical Chemistry B* **2010**, 114, 2755.
- [21] A. R. Katritzky, D. C. Fara, M. Kuanar, E. Hur, M. Karelson, *Journal of Physical Chemistry A* **2005**, 109, 10323.
- [22] M. O. Environment, Chemical Computing Group, p. Molecular Operating Environment.
- [23] J. D. Dunitz, *Chemical Communications* **2003**, 545.
- [24] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Chemical Communications* **2007**, 1600.
- [25] F. P. A. Fabbiani, L. T. Byrne, J. J. McKinnon, M. A. Spackman, *CrystEngComm* **2007**, 9, 728.
- [26] M. M. J. Lowes, M. R. Cairra, A. P. Lotter, J. G. Van Der Watt, *Journal of Pharmaceutical Sciences* **1987**, 76, 744.
- [27] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chemical Reviews* **1996**, 96, 1027.
- [28] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors, Vol. 11*, first ed., Wiley-VCH, Weinheim, **2000**.

- [29] N. Blagden, R. Davey, G. Dent, M. Song, W. I. F. David, C. R. Pulham, K. Shankland, *Crystal Growth and Design* **2005**, 5, 2218.
- [30] A. Getsoian, R. M. Lodaya, A. C. Blackburn, *International Journal of Pharmaceutics* **2008**, 348, 3.
- [31] R. Davey, J. Garside, *From Molecules to Crystallizers: An Introduction to Crystallization*, Oxford University Press, Oxford, **2000**.
- [32] N. Rodriguez-Hornedo, D. Murphy, *Journal of Pharmaceutical Sciences* **1999**, 88, 651.
- [33] V. Gutmann, *Electrochimica Acta* **1976**, 21, 661.
- [34] S. Hahn, W. M. Miller, R. N. Lichtenthaler, J. M. Prausnitz, *Journal of Solution Chemistry* **1985**, 14, 129.
- [35] L. Lavielle, J. Schultz, *Langmuir* **1991**, 7, 978.
- [36] F. L. Riddle, F. M. Fowkes, *Journal of the American Chemical Society* **1990**, 112, 3259.
- [37] Y. Marcus, *Journal of Solution Chemistry* **1984**, 13, 599.
- [38] V. Gutmann, *Coordination Chemistry Reviews* **1967**, 2, 239.
- [39] R. K. Harris, P. Y. Ghi, H. Puschmann, D. C. Apperley, U. J. Griesser, R. B. Hammond, C. Ma, K. J. Roberts, G. J. Pearce, J. R. Yates, C. J. Pickard, *Organic Process Research and Development* **2005**, 9, 902.
- [40] E. Laine, V. Tuominen, P. Ilvessalo, P. Kahela, *International Journal of Pharmaceutics* **1984**, 20, 307.
- [41] D. Murphy, F. Rodriguez-Cintron, B. Langevin, R. C. Kelly, N. Rodriguez-Hornedo, *International Journal of Pharmaceutics* **2002**, 246, 121.
- [42] D. T. Stanton, S. Dimitrov, V. Grancharov, O. G. Mekenyan, *SAR and QSAR in Environmental Research* **2002**, 13, 341.
- [43] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, 113, 9480.
- [44] D. T. Stanton, P. C. Jurs, *Analytical Chemistry* **1990**, 62, 2323.
- [45] H. Golmohammadi, *Journal of Computational Chemistry* **2009**, 30, 2455.
- [46] M. H. Fatemi, F. Karimian, *Journal of Colloid and Interface science* **2007**, 314, 665.
- [47] B. Lee, F. M. Richards, *Journal of Molecular Biology* **1971**, 55, 379.
- [48] A. Johnston, A. J. Florence, A. R. Kennedy, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1509.
- [49] A. Johnston, B. F. Johnston, A. R. Kennedy, A. J. Florence, *CrystEngComm* **2008**, 10, 23.
- [50] S. Lohani, Y. Zhang, L. J. Chyall, P. Mougin-Andres, F. X. Muller, D. J. W. Grant, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1310.
- [51] S. G. Fleischman, S. S. Kuduva, J. A. McMahon, B. Moulton, R. D. Bailey Walsh, N. Rodriguez-Hornedo, M. J. Zaworotko, *Crystal Growth and Design* **2003**, 3, 909.
- [52] A. J. Cruz Cabeza, G. M. Day, W. D. Samuel Motherwell, W. Jones, *Journal of the American Chemical Society* **2006**, 128, 14466.

## **6. RESULTS AND DISCUSSION-PLS ANALYSIS**

Partial Least Squares (PLS) analysis was carried out upon all molecular and bulk descriptors in order to reduce the dataset. PLS determines linear correlations within the data and can highlight important descriptors based on these correlations. By using the highlighted descriptors, the dataset was reduced, which led to the development of an artificial neural network (ANNs) for polymorph prediction.

### **6.1. Data Reduction using PLS**

PLS analysis was conducted as another method of descriptor selection (results in electronic appendix, chapter 6, 6.1). This data analysis technique is able to handle all of the data simultaneously, therefore making it more time effective for this type of research, rather than the manual analysis methods presented in chapter 5.

From the results generated in the PLS analysis a number of different sets of descriptors were highlighted and used in an ANN. As with principal component analysis (PCA), PLS generates score values that combine all of the descriptor data. The difference between PCA and PLS analysis is that the scores are generated based on the determination of one polymorphic outcome at a time. With form II and III being the most common outcome of the crystallisation experiments, only scores for these two polymorphic forms were generated. As well as score values, PLS determines which of the descriptors are the most important. Using these values, that once again are for form II and III separately, an ANN can be built.

Each of these outputs will be discussed separately, with the most optimal set of descriptors presented in section 6.2.

#### **6.1.1. Analysis of Score Values (Form II Model)**

Seven components were created, with each experimental row being assigned a score value. These scores are determined by different loading values being placed upon



each descriptor and summed into one value. FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> analyses were carried out and the overall results are presented in Table 6.1.

Table 6.1 Results of ANN analysis of PLS form II score values from seven components

R <sup>2</sup> values for each form (%)	FormRules		INForm Training	INForm Testing
Form I	4.40		83.10	100.00
Form II	56.91		95.70	18.71
Form III	56.65		95.31	3.16
Dihydrate	7.09		84.85	100
Solvate	97.43		98.40	96.51
<b>Average R<sup>2</sup></b>	<b>44.50 %</b>		<b>91.47 %</b>	<b>63.68 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>77.57 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>61.04 %</b>	

The average result over both FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> was 61.04 %, which is not a successful network. This result is not surprising as the components were focussed upon form II crystallisation only. However, it is more surprising that the form II prediction value in INForm<sup>[2]</sup> is very low. This indicates that this component analysis technique does not provide enough information in order to create a successful ANN model.

The PLS analysis also determines the number of significant components in each model, and in this case it highlighted only the first two. Based on the previous unsuccessful results using all seven components, it was unlikely that using only two would generate improved results, but the analysis was carried out (Table 6.2).

When only the two most significant components are used in the ANN, there is an overall reduction in the performance of FormRules<sup>[1]</sup> and INForm<sup>[2]</sup>, with an average overall result of 54.31 %.

Table 6.2 Results of ANN analysis of PLS form II score values from two components

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	1.33		75.96	100.00
Form II	56.91		80.18	-22.43
Form III	56.42		68.71	-39.37
Dihydrate	2.72		95.63	100
Solvate	96.36		99.95	100
<b>Average R<sup>2</sup></b>	<b>42.75 %</b>		<b>84.09 %</b>	<b>48.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>65.86 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>54.31 %</b>	

It should be noted that the  $R^2$  value for both the test set Form II and III results are negative. When Equation 4.3 is analysed, it becomes apparent that when the total sum of errors is larger than the total variance of the data a negative value is calculated. This highlights that the model is incapable of generating an accurate prediction based upon the training data.

The most interesting factor to observe in these analyses is that the  $R^2$  value in the form II prediction in FormRules<sup>[1]</sup> does not change. This therefore suggests that the most important information for form II prediction is contained within the first two components. However, there is not enough information in these components to successfully predict all of the different polymorphic forms.

### 6.1.2. Analysis of Score Values (Form III Model)

Similarly to above, six components that predict form III were generated. The average result when these component score values were used in FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> was 65.98 % (Table 6.3). The prediction of polymorphic form made using the form III scores is more successful than the form II scores, but does highlight that there is not enough information to predict all outcomes when using only the score values of one form.

Table 6.3 shows that the form III prediction and overall average in INForm<sup>[2]</sup> is better than in many other networks. However, when the results are combined with FormRules<sup>[1]</sup> values, which performed poorly, the average  $R^2$  result is reduced.

Table 6.3 Results of ANN analysis of PLS form III score values from six components

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	2.50		58.38	100.00
Form II	53.65		92.13	26.71
Form III	56.80		84.44	58.50
Dihydrate	43.43		99.64	100.00
Solvate	96.88		99.41	93.85
<b>Average R<sup>2</sup></b>	<b>50.65 %</b>		<b>86.80 %</b>	<b>76.00%</b>
			<b>INForm average R<sup>2</sup></b>	<b>81.31 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>65.98 %</b>	

As with the form II analysis, the first two components were highlighted as the most significant in form III prediction. The score values for these components were therefore analysed separately to see if there was any improvement (Table 6.4).

Table 6.4 Results of ANN analysis of PLS form III score values from two components

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	1.94		79.70	100.00
Form II	50.75		73.43	39.86
Form III	52.24		79.38	35.56
Dihydrate	3.88		73.38	100.00
Solvate	97.46		99.56	96.52
<b>Average R<sup>2</sup></b>	<b>41.25 %</b>		<b>81.09 %</b>	<b>74.00</b>
			<b>INForm average R<sup>2</sup></b>	<b>77.74 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>59.50 %</b>	

By training the ANN with the two most significant components the overall prediction average went down from 65.98 % to 59.50 %. Unlike in the form II analysis, the FormRules<sup>[1]</sup> R<sup>2</sup> value for form III prediction has changed, suggesting that the first

two components do not adequately provide all of the information required for a successful form III prediction.

Overall, by using the score values for both the form II and III PLS models separately, the ANN did not predict polymorphic form very successfully. This analysis also does not generate specific information about which descriptors are contributing most significantly to the prediction. This therefore means that nothing can be learned about the crystallisation of different polymorphs at the molecular level. As each score is calculated using a loading value that is placed upon each descriptor, analysis of these loading values was undertaken to determine the most influential descriptors within each component.

### **6.1.3. Analysis of the Loading Values (Form II Model)**

A method of reducing the number of descriptors analysed from the PLS analysis has been devised using the knowledge of the loading values. Each descriptor is given a loading value that contributes towards the score generated in each component. Knowing that the first two components are the most significant in form II prediction, the most highly loaded descriptors in these components must be the most influential. As a method of data reduction, the selection of the two most positively and negatively loaded descriptors from the two significant components, representing the extreme values, were run in ANN. The eight descriptors in this analysis were pmiY, E\_nb, d78 and d72 from the first component and d42, dsolv13, viscosity and d66 from the second component. A brief description of each is shown in Table 6.5 and appendix section 12.2.

Table 6.5 Brief description of the eight descriptors used in this analysis

Descriptor	Descriptor definition
PmiY	Principal moment of inertia Y, of the CBZ molecule
E_nb	Value of potential energy of the CBZ molecule when the non-bonded terms are disabled
D78	DPSA-1, the difference in partial surfaces areas upon the CBZ molecule
D72	PPSA-1, the partial positive surface area of the CBZ molecule
D42	3D-Wiener index of the CBZ molecule
Dsolv13	Relative number of single bonds of the solvent molecule
Viscosity	Of the solvent molecule
D66	Moment of inertia A of the CBZ molecule

When the eight descriptors are analysed the results are very poor (Table 6.6), with an average  $R^2$  value of 47.94 %. This highlights that there is little variation in the descriptors used and therefore no successful predictions can be made.

Table 6.6 Results of ANN analysis of PLS form II using the loading values from two most positive and negative descriptors from two components (eight descriptors)

$R^2$ values for each form (%)	FormRules		INForm Training	INForm Testing
Form I	2.71		99.60	100.00
Form II	50.30		88.23	-116.96
Form III	52.84		87.69	-90.43
Dihydrate	5.17		75.00	100.00
Solvate	96.92		99.94	99.76
<b>Average <math>R^2</math></b>	<b>41.59 %</b>		<b>90.09 %</b>	<b>18.00 %</b>
			<b>INForm average <math>R^2</math></b>	<b>54.28 %</b>
<b>FormRules and INForm average <math>R^2</math></b>			<b>47.94 %</b>	

In order to reduce the number of descriptors further, only the top and bottom descriptor in each component were selected (pmiY, d78, d42 and viscosity). The

analysis of these four descriptors actually produces a more successful prediction overall of 56.14 % (Table 6.7).

Table 6.7 Results of ANN analysis of PLS form II using the loading values from the most positive and negative descriptors from two components (four descriptors)

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	0.99		96.54	100.00
Form II	50.30		72.48	10.08
Form III	48.42		66.79	2.24
Dihydrate	1.78		79.09	100.00
Solvate	96.29		99.95	100.00
<b>Average R<sup>2</sup></b>	<b>39.56 %</b>		<b>82.97 %</b>	<b>62.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>72.72 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>56.14 %</b>	

From the FormRules<sup>[1]</sup> results it becomes apparent that when only the four descriptors are used, the form II prediction is still at the same level as with eight descriptors. When the rules generated in FormRules<sup>[1]</sup> are analysed (appendix section 12.8), there are two descriptors that are present in the form II prediction. These are d78, which is the difference in the partial positive and negative surface areas<sup>[3]</sup> and pmiY, which is the moment of inertia y<sup>[4]</sup>, both on the CBZ molecule. From the overall result generated it is clear that these descriptors alone cannot predict all the polymorphic form outcomes from the crystallisation experiments. However, these two descriptors can lead to enhanced form II prediction. Therefore these two descriptors will be taken forward for further analysis.

#### 6.1.4. Analysis of the Loading Values (Form III Model)

The same analysis was carried out with the loading values from the form III model. It should be noted that MNDO\_dipole came up twice in this analysis and therefore the sets are comprised of seven and three descriptors. The seven unique descriptors are MNDO\_dipole, d86, dsolv24 and d84 from component one and MNDO\_dipole, d77, d75 and d42 from component two, with a brief description in Table 6.8 and appendix section 12.2.

Table 6.8 Brief description of the seven descriptors used in this analysis

Descriptor	Descriptor definition
MNDO_dipole	Calculated dipole moment of the CBZ molecule
D86	FNSA-3, Fractional partial negative surface area of the CBZ molecule (PNSA-3/TMSA)
Dsolv24	Kier and Hall index (order 1) of the solvent molecule
D84	FNSA-1, Fractional partial negative surface area of the CBZ molecule (PNSA-1/TMSA)
D77	PNSA-3, Atomic charge weighted partial negative surface area of the CBZ molecule
D75	PNSA-1, Partial negative surface area of the CBZ molecule
D42	3D-Wiener index of the CBZ molecule

These descriptors were used in an ANN with the FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> results presented in Table 6.9.

Table 6.9 Results of ANN analysis of PLS form III using the loading values from the two most positive and negative descriptors from two components (seven descriptors)

R <sup>2</sup> values for each form (%)	FormRules		INForm Training	INForm Testing
Form I	2.27		98.83	100.00
Form II	48.69		81.42	23.18
Form III	56.90		83.77	46.97
Dihydrate	19.41		91.65	100.00
Solvate	96.95		99.93	99.65
<b>Average R<sup>2</sup></b>	<b>44.84 %</b>		<b>91.12 %</b>	<b>74.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>82.54 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>63.69 %</b>	

Similarly to the form II analysis, the descriptor set was reduced further, using only the most positively and negatively loaded descriptor in the first two components. Only three descriptors were used in this analysis as MNDO\_dipole occurred twice. The other descriptors were dsolv24 and d75, and were analysed in FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> (Table 6.10).

Table 6.10 Results of ANN analysis of PLS form III using the loading values from the most positive and negative descriptors from two components (three descriptors)

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	2.02		97.51	100.00
Form II	49.69		66.00	45.94
Form III	51.31		70.54	72.99
Dihydrate	19.41		97.42	100.00
Solvate	96.76		99.49	99.95
<b>Average R<sup>2</sup></b>	<b>43.84 %</b>		<b>86.19 %</b>	<b>84.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>85.10 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>64.47 %</b>	

The average results for the seven and three descriptor analyses were 63.69 % and 64.47 % respectively. There is a slight improvement in the overall average prediction value when compared to the other ANN results in the PLS analysis. It is also worth noting that the INForm<sup>[2]</sup> prediction of form III is very good when only MNDO\_dipole, dsolv24 and d75 are used. This implies that they are potentially very important descriptors. When the rules created by FormRules<sup>[1]</sup> are observed, dsolv24, which is the Kier and Hall index<sup>[5]</sup> of the solvent and d75, the partial negative surface area<sup>[3]</sup> of the CBZ molecule are used in form III prediction. These highlighted descriptors will be taken forward for further analysis.

#### 6.1.5. Analysis of the Variable Importance Values (Form II Model)

The ten most important descriptors, as determined by the PLS analysis, for form II prediction have been run in an ANN. The descriptors involved were d78, d82, d84, d81, d85, d68, pmiY, d72, d73 and E\_nb (Table 6.11), which are in order of importance.



Table 6.11 Brief description of the ten most important descriptors for form II prediction

Descriptor	Descriptor definition
D78	DPSA-1, the difference in partial surfaces areas upon the CBZ molecule
D82	FPSA-2, Fractional partial negative surface area of the CBZ molecule (PPSA-2/TMSA)
D84	FNSA-1, Fractional partial negative surface area of the CBZ molecule (PNSA-1/TMSA)
D81	FPSA-1, Fractional partial negative surface area of the CBZ molecule (PPSA-1/TMSA)
D85	FNSA-2, Fractional partial negative surface area of the CBZ molecule (PNSA-2/TMSA)
D68	Moment of inertia C, of the CBZ molecule
PmiY	Principal moment of inertia Y, of the CBZ molecule
D72	PPSA-1, the partial positive surface area of the CBZ molecule
D73	PPSA-2, the total charge weighted partial positive surface area of the CBZ molecule
E_nb	Value of potential energy of the CBZ molecule when the non-bonded terms are disabled

FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> analysis was carried out upon the ten descriptors, with the results presented in Table 6.12.

Table 6.12 Results of ANN analysis of PLS form II using variable importance values

R <sup>2</sup> values for each form (%)	FormRules		INForm Training	INForm Testing
Form I	2.71		98.65	100.00
Form II	50.30		75.48	13.10
Form III	52.92		71.39	-50.57
Dihydrate	5.17		60.41	100.00
Solvate	96.87		99.84	91.51
<b>Average R<sup>2</sup></b>	<b>41.59 %</b>		<b>81.15 %</b>	<b>51.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>66.08 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>53.84 %</b>	

The overall average  $R^2$  result of this analysis is 53.84 %, which is a poor performance. This is unsurprising as these are the ten most important descriptors in form II prediction only. It is more surprising that the INForm<sup>[2]</sup> results for form II prediction is also weak, suggesting that different properties need to be present in order to distinguish one form from another.

#### 6.1.6. Analysis of the Variable Importance Values (Form III Model)

The same analysis was carried out to determine the ten most important descriptors in form III prediction. The descriptors used were MNDO\_dipole, dsolv24, d86, dsolv25, d84, d81, d85, d82, d78 and d77 respectively. A brief description of these descriptors is presented in Table 6.13.

Table 6.13 Brief description of the ten most important descriptors for form III prediction

Descriptor	Descriptor definition
MNDO_dipole	Calculated dipole moment of the CBZ molecule
Dsolv24	Kier and Hall index (order 1) of the solvent molecule
D86	FNSA-3, Fractional partial negative surface area of the CBZ molecule (PNSA-3/TMSA)
Dsolv25	Kier and Hall index (order 2) of the solvent molecule
D84	FNSA-1, Fractional partial negative surface area of the CBZ molecule (PNSA-1/TMSA)
D81	FPSA-1, Fractional partial negative surface area of the CBZ molecule (PPSA-1/TMSA)
D85	FNSA-2, Fractional partial negative surface area of the CBZ molecule (PNSA-2/TMSA)
D82	FPSA-2, Fractional partial negative surface area of the CBZ molecule (PPSA-2/TMSA)
D78	DPSA-1, the difference in partial surfaces areas upon the CBZ molecule
D77	PNSA-3, the atomic charge weighted partial negative surfaces areas of the CBZ molecule

FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> analysis was carried out using these ten descriptors, with results presented in Table 6.14.

Table 6.14 Results of ANN analysis of PLS form III using variable importance values

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	2.27		85.82	100.00
Form II	45.84		72.41	49.01
Form III	56.12		76.69	66.48
Dihydrate	4.05		99.70	100.00
Solvate	96.78		99.92	97.75
<b>Average R<sup>2</sup></b>	<b>41.01 %</b>		<b>86.91 %</b>	<b>83.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>84.96 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>62.99 %</b>	

Overall the average result of this analysis was higher than seen in the form II analysis at 62.99 %. Notably INForm<sup>[2]</sup> has performed well, which suggests that the information contained within these ten descriptors is useful for the prediction of different polymorphic forms. However, FormRules<sup>[1]</sup> has not successfully created rules from these descriptors, suggesting that other combinations of descriptors may improve the predictions.

### 6.1.7. Analysis of the Variable Importance Descriptor Overlap

When the two sets of descriptors highlighted in the variable importance lists for both form II and form III are compared (Table 6.11 and Table 6.13), there are five descriptors that feature in both sets. The overlapping descriptors are d78, d82, d84, d81 and d85. Overall this means there are fifteen unique descriptors in the two lists. An ANN of these fifteen descriptors was run to determine if the prediction improves with the addition of further information. Also an ANN of the five overlapping descriptors was run to see whether it is more or less successful than using all fifteen descriptors.

The results (Table 6.15) of using only the 5 overlapping descriptors are very poor (39.79 %), which clearly highlights that more information is required for successful polymorphic form prediction.

Table 6.15 Results of ANN analysis of the overlapping descriptors from the top ten form II and III variable importance values

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	0.99		98.70	100.00
Form II	45.84		81.00	-57.26
Form III	30.43		83.79	-83.79
Dihydrate	1.78		99.60	100.00
Solvate	12.99		98.62	90.76
<b>Average R<sup>2</sup></b>	<b>18.41 %</b>		<b>92.34 %</b>	<b>30.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>61.17 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>39.79 %</b>	

When the fifteen unique descriptors were run in FormRules<sup>[1]</sup> and INForm<sup>[2]</sup> (Table 6.16) the overall average R<sup>2</sup> value increased to 64.07 %. This demonstrates that more information is required to predict all of the different polymorphic forms.

Table 6.16 Results of ANN analysis of the unique descriptors from the top ten form II and III variable importance values

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	2.71		98.73	100.00
Form II	50.30		82.46	18.66
Form III	56.12		79.91	13.64
Dihydrate	45.23		93.14	100.00
Solvate	96.87		99.62	90.67
<b>Average R<sup>2</sup></b>	<b>50.25 %</b>		<b>90.77 %</b>	<b>65.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>77.89 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>64.07 %</b>	

## 6.2. Optimisation of PLS Results

Overall, the PLS analysis method of data reduction does not identify the descriptors that can successfully predict the CBZ polymorphic form crystallised in a set of experiments. The analysis conducted did highlight a number of descriptors that are potentially useful in form II and III prediction. D78 and pmiY were brought forward from the form II analysis, and dsolv24, d85 and d75 from the form III analysis. By using these potentially useful descriptors, an ANN was created to assess whether both form II and form III could be successfully predicted (Table 6.17).

When the descriptors from the most successful analysis in this PLS research were observed (most positively and negatively loaded descriptors in component one and two in form III prediction), MNDO\_dipole was featured, but not brought forward as important. MNDO\_dipole, which is the calculated dipole moment<sup>[6]</sup> of the CBZ molecule, was not highlighted by the rule analysis as being an important descriptor. Therefore its effect upon addition to the five descriptors analysed in Table 6.17, has been examined by FormRules<sup>[1]</sup> and INForm<sup>[2]</sup>.

Table 6.17 Optimisation of the model using previously highlighted informative descriptors and the most successful set

Descriptors	Most Successful PLS set	Highlighted descriptors	Highlighted descriptors + MNDO_dipole
Dsolv24	X	X	X
D75	X	X	X
MNDO_dipole	X		X
D78		X	X
PmiY		X	X
D85		X	X
<b>FormRules Average R<sup>2</sup> (%)</b>	<b>43.84</b>	<b>43.97</b>	<b>43.96</b>
<b>INForm Average R<sup>2</sup> (%)</b>	<b>85.10</b>	<b>71.29</b>	<b>72.55</b>
<b>Overall Average R<sup>2</sup> (%)</b>	<b>64.47</b>	<b>57.63</b>	<b>58.26</b>

The addition of MNDO\_dipole showed a slight improvement in the overall prediction of the polymorphic forms (58.26 %), but the results are still much less successful than those seen in previous work.

### **6.3. Conclusion of PLS Work**

Overall, the use of PLS as a method of data reduction was effective in this research. However, the descriptors highlighted as important based upon their loading and variable importance values, failed to build a successful ANN for the prediction of polymorphic form. Perhaps a reason for this may be the relatively small number of experimental values involved in the predictions. It has been noted previously that PLS can be more successful with larger models, in which more importance is placed upon the information across the whole dataset and not upon individual variables<sup>[7]</sup>.

Although a successful ANN was not built, the PLS analysis did highlight a number of descriptors that shall be remembered in future analysis. The inclusion of d78 and pmiY may improve the form II predictions, similarly dsolv24, d85, d75 and MNDO\_dipole for form III predictions.

- [1] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [2] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [3] D. T. Stanton, P. C. Jurs, *Analytical Chemistry* **1990**, 62, 2323.
- [4] P. Atkins, J. de Paula, *Atkins' Physical Chemistry*, 7th ed., Oxford University Press, Oxford, **2002**.
- [5] L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research, Vol. 14*, 1st ed., Academic Press, Inc., New York, **1976**.
- [6] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chemical Reviews* **1996**, 96, 1027.
- [7] V. E. Vinzi, W. W. Chin, J. Henseler, H. Wang, *Handbook of Partial Least Squares. Concepts, Methods and Applications*, Springer-verlag Berlin Heidelberg, **2010**.

## **7. RESULTS AND DISCUSSION - PCA ANALYSIS**

All of the descriptors (molecular and bulk) were subjected to principal component analysis (PCA). Details of the PCA methods were given in section 2.9. PCA does not link the polymorphic outputs with the descriptors; it simply assesses mathematically the correlations within the information in order to reduce the dimensionality of the dataset<sup>[1]</sup>. This allows reducing the number of descriptors that need to be considered in the artificial neural network (ANN) analysis. This chapter covers the descriptor reduction analysis and a discussion of the descriptor meanings found in the most successful set.

### **7.1. Data Reduction using PCA**

Thirteen principal components (PC) were created, each encompassing features of all the descriptors. In every component each descriptor is given a loading value that places a different level of significance upon it. These values are correlations between the descriptor data and the component itself; therefore a large positive or negative loading indicates a well correlated descriptor within that variable space. The loading values are then transformed using regression-like equations to generate a score for each component<sup>[2]</sup>. Based upon this analysis, each experimental row has thirteen new descriptors, which should include all of the variation within the dataset. The PCA data can be found in Electronic Appendix, Chapter 7, file 7.1.

#### **7.1.1. Analysis of Score Values**

Using the thirteen component score values as inputs in the ANN, analysis can be carried out to determine if a successful prediction can be made. The results are highlighted in Table 7.1 and show a remarkably mediocre overall average  $R^2$  value of 60.63 %. It was expected that due to the involvement of all of the information in the descriptors, the prediction would be successful. However, this result demonstrates that the PCA has placed significance on some descriptors based on their numerical value that do not lead to a successful prediction. The rules generated



using the score values can be found in Electronic Appendix, Chapter 7, file 7.2, and refer only to the component. Therefore no specific information about which descriptors are important can be determined from these rules.

Table 7.1 FormRules and INForm results of PCA score analysis

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>	<b>INForm Training</b>	<b>INForm Testing</b>
Form I	5.77	99.39	100.00
Form II	54.66	84.90	0.12
Form III	51.66	86.04	7.88
Dihydrate	57.16	49.61	52.59
Solvate	97.43	99.92	98.77
<b>Average R<sup>2</sup></b>	<b>53.34 %</b>	<b>83.97 %</b>	<b>51.87 %</b>
		<b>INForm average R<sup>2</sup></b>	<b>67.92 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>		<b>60.63 %</b>	

A scree plot shows that as the number of PCs increases, the amount of variance in the data (represented by the R<sup>2</sup> value from the PCA) becomes less. There is always an elbow in these plots that highlights after that point there is no longer significant amounts of new data being presented<sup>[3]</sup>. The elbow in this plot was taken to be at PC5 as the change in R<sup>2</sup> value between PC5-13 is relatively small.

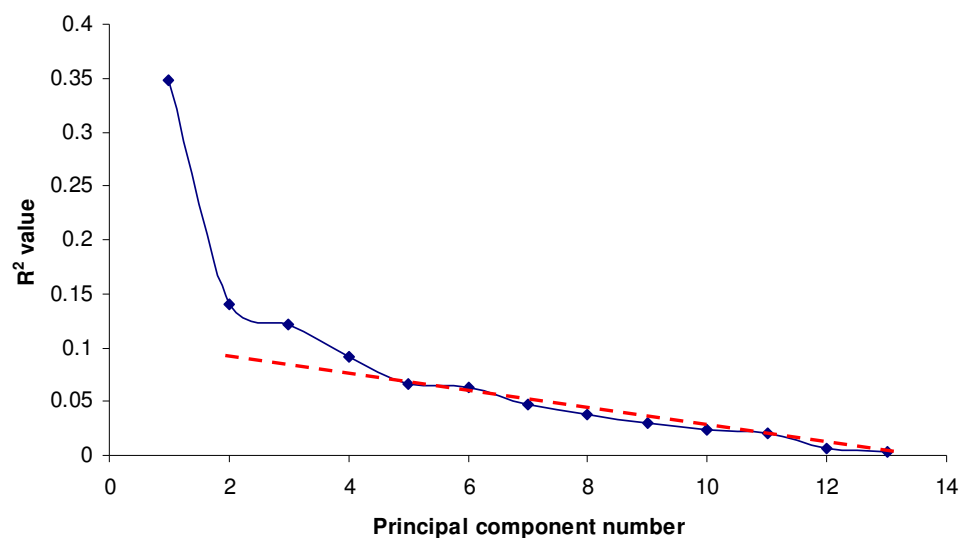


Figure 7.1 Scree plot generated from the PCA results of carbamazepine (CBZ) descriptor analysis

It was therefore examined whether omitting PCs 6-13 would affect the overall result significantly (Table 7.2).

Table 7.2 Results of PC1-5 score analysis

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	5.77		99.63	100.00
Form II	54.66		95.66	3.76
Form III	51.66		89.53	30.18
Dihydrate	57.16		98.29	100.00
Solvate	97.01		99.55	100.00
<b>Average R<sup>2</sup></b>	<b>53.25 %</b>		<b>96.53 %</b>	<b>67.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>81.66 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>67.46 %</b>	

The results do show an improvement in prediction over using all 13 PCs. However, there is still no specific information about which descriptors are most important in this predictive model.

### 7.1.2. Analysis of Loading Values

As mentioned in section 7.1, the problem associated with using the score values is that all the descriptors have been condensed into one value for each PC. Therefore, no knowledge can be acquired about which of the molecular or bulk descriptors are important in the prediction of polymorphic form. The score values are created from the loading values given to each descriptor for each experiment, and therefore by analysing the loading values, significant descriptors can be highlighted. A full list of loading values can be found in Electronic Appendix, Chapter 7, file 7.1.

PC1 is the component that represents the largest variance in the data set, with each subsequent PC containing less variation. It is the most positively and negatively loaded descriptors in each component that affect the score values most significantly. Therefore, analysis of the descriptors that have been grouped together is interesting to determine if there are any trends.

The ten most positively and negatively loaded descriptors within each PC have been analysed based on their physical meaning. For a full description of each descriptor see appendix section 12.2. This analysis has been carried out for PC1-5 only, as they are known to contain most of the variation in the data.

Table 7.3 The most significant descriptors in PC1

Most positively loaded descriptors	Most negative loaded descriptors
Dsolv43 – 3D-Randić index (order 0) of the solvent	D69 – Molecular surface area of CBZ
Dsolv20 – Randić index (order 1) of the solvent	Dielectric constant of the solvent
Dsolv22 – Randić index (order 3) of the solvent	pmiZ – Moment of inertia C of CBZ
Dsolv26 – Kier & Hall index (order 3) of the solvent	dP – Hansen solubility parameter of the solvent
Dsolv6 – Number of rings in the solvent	Dosl67 – Moment of inertia B of the solvent
Dsolv2 – Number of carbon atoms in the solvent	D71 – Total molecular surface area of the CBZ
Dsolv31 – Complementary information content (order 0) of the solvent	D79 – DPSA-2 – Difference in Charged partial surface areas of CBZ
Dolv55 – 3D-Complementary information content (order 0) of the solvent	Dsolv68 – Moment of inertia C of the solvent
Dsolv42 – 3D-Wiener index of the solvent	Glob – Globularity of CBZ
Dsolv44 – 3D-Randić index (order 1) of the solvent	Polarity Parameter (ET <sub>30</sub> ) - of the solvent

The ten most positive descriptors in PC1 (Table 7.3) are all related to the shape of the solvent molecule. Of these ten, there are four occurrences of the Randić index, one occurrence of the Wiener index and one of the Kier and Hall index. These are all classified as “second generation topological”<sup>[4]</sup> descriptors. The number of rings and carbon atoms can also be related to the shape and size of the solvent molecule. The complementary information content, although a measure of the diversity of elements within the molecule,<sup>[5]</sup> the higher order values are connected to the molecules size. Half of the negative descriptors are related to the CBZ molecule and the remainder are related to the solvent. Surface areas are featured for both the solvent and CBZ, and out of the five solvent properties, three are bulk parameters. There are also three occurrences of the moment of inertia (twice for the solvent), which like the positive descriptors is concerned with the size and shape of the molecule.

Table 7.4 Most significant descriptors in PC2

Most positively loaded descriptors	Most negative loaded descriptors
Dsolv74 – PPSA-2 – Partial positive surface area of the solvent	Dsolv69 – Molecular surface area of the solvent
Dsolv65 – 3D-Bonding information content (order 2) of the solvent	Dsolv46 – 3D-Randić index (order 3) of the solvent
Dsolv51 – 3D-Kier shape index (order 1) of the solvent	Dsolv41 – Bonding information content (order 2) of the solvent
Dsolv28 – Kier shape index (order 2) of the solvent	Dsolv60 – 3D-Structural information content (order 1) of the solvent
PM3_HOMO – Highest occupied molecular orbital energy of CBZ	Dsolv37 – Bonding Information Content (order 1) of the solvent
dH – Hansen solubility parameter of the solvent	Dsolv56 – 3D-Structural information content (order 0) of the solvent
Dsolv66 – Moment of inertia A of the solvent	Dsolv77 – PNSA-3 – Partial negative surface area of the solvent
Dsolv36 – Structural information content (order 1) of the solvent	Dsolv33 – Bonding information content (order 0) of the solvent
MNDO_IP – Ionisation potential of CBZ	Dsolv52 – 3D-Kier shape index (order 2) of the solvent
Dsolv49 – 3D-Kier & Hall index (order 2) of the solvent	PM3_IP – Ionisation potential of CBZ

In the set for PC 2 (Table 7.4), the majority of the positive descriptors are describing the solvent molecule once again. However, there is also an ionisation potential and highest occupied molecular orbital term for the CBZ molecule. These two terms are clearly related and when calculated with the same level of theory the values are the same but with one negative and one positive. The positively loaded solvent descriptors in Table 7.4 are again mostly describing the shape of the solvent molecule, but with the inclusion of a partial positive charge, the moment of inertia and bulk value, the Hansen solubility parameter.

The negative descriptor values are all associated with solvent descriptors, with the exception of PM3\_IP, which is a measure of the ionisation potential in the CBZ molecule. This descriptor was not expected to be featured here as there is an ionisation potential term in the positive results. The difference between these is the level of theory used to calculate the value, but it is still a surprising result to see a descriptor with a similar physical meaning being represented at the two ends of data

variation. The solvent descriptors are once again representing the shape and size of the molecule.

Table 7.5 Most significant descriptors in PC3

Most positively loaded descriptors	Most negative loaded descriptors
D68 – Moment of inertia C of CBZ	D84 – FNSA-1 – fractional negative surface area of CBZ
Viscosity of the solvent	D42 – 3D-Wiener index of CBZ
Dsolv7 – Number of Benzene rings in the solvent	Dsolv13 – Relative number of single bonds in the solvent
Dsolv12 – Number of aromatic bonds in the solvent	Std_dim2 – Standard dimension 2 of CBZ
Dsolv16 – Relative number of aromatic bonds in the solvent	PmiY – Moment of inertia B of CBZ
D85 – FNSA-2 – Fractional negative surface area of CBZ	D70 – Molecular volume of CBZ
D82 – FPSA-2 – Fractional positive surface area of CBZ	D75 – PNSA-1 – Partial negative surface area of CBZ
D81 – FPSA-1 – Fractional positive surface area of CBZ	E_nb – Non-bonded energy of CBZ
D78 – DPSA-1 – Difference in positive surface area of CBZ	Vapour Pressure of the solvent
D66 – Moment of inertia A of CBZ	Activity of CBZ in the solvent

For PC3 (Table 7.5), the majority of the positively loaded descriptors are representations of the CBZ molecule. There are two occurrences of moment of inertia, which highlights the subtle changes in CBZ conformation within the solvent force field, and four charged partial surface area descriptors. It is quite interesting to note that the solvent descriptors featured in this component represent very basic information with regards to the solvent molecule.

Similarly to the highly positive results the majority of the negatively loaded descriptors assess the CBZ molecular properties. The remaining descriptors are bulk solvent properties (relative number of single bonds and vapour pressure). It is noteworthy that activity features here as this is one of only two experimentally determined descriptors, therefore encompassing both solvent and CBZ information.

Table 7.6 Most significant descriptors in PC4

Most positively loaded descriptors	Most negative loaded descriptors
Dsolv76 – PNSA-2 – Partial negative surface area of the solvent	E_ang – Angle bend potential energy of CBZ
Dsolv5 – Relative number of hydrogen atoms in the solvent	E_str – Bond stretch potential energy of CBZ
E_ele – Electrostatic component of the potential energy of CBZ	E_oop – Out of plan potential energy of CBZ
Dsolv61 – 3D-Bonding information content (order 1) of the solvent	E_strain – Local strain energy of CBZ
Dsolv10 – Number of double bonds in the solvent	E – Potential energy of CBZ
Dsolv14 – Relative number of double bonds in the solvent	MNDO_IP – Ionisation potential of CBZ
MNDO_HOMO – Highest occupied molecular orbital energy of CBZ	E_tor – Torsion potential energy of CBZ
Dsolv3 – Number of hydrogen atoms in the solvent	Density – of the solvent
Dsolv57 – 3D-Bonding information content (order 0) of the solvent	MNDO_dipole – Molecular dipole of CBZ
Dsolv30 – Information content (order 0) of the solvent	LogP – of the solvent

For PC4 (Table 7.6) the positively loaded descriptors relate mostly to the bonding and atoms within the solvent, but there is also a charged partial surface area term. The MNDO\_HOMO and electrostatic component of potential energy for the CBZ molecule are also highlighted.

The majority of the negatively loaded descriptors are potential energies terms for the CBZ molecule. However, the logP and density of the solvent are also featured.

Table 7.7 Most significant descriptors in PC5

Most positively loaded descriptors	Most negative loaded descriptors
D77 – PNSA-3 – Partial negative surface area of CBZ	Vol – Van der Waals volume of CBZ
Dens – Relative molecular mass divided by the van der Waals volume of CBZ	Boiling point - of the solvent
PmiX – Moment of inertia A of CBZ	PM3_LUMO – Lowest unoccupied molecular orbital energy of CBZ
Vapour pressure – of the solvent	E_str – Bond stretch potential energy of CBZ
D86 – FNSA-3 – Fractional negative surface area of CBZ	E_tor – Torsion potential energy of CBZ
MNDO_dipole – Molecular dipole of CBZ	Glob – Globularity of CBZ molecule
AM1_LUMO – Lowest unoccupied molecular orbital energy of CBZ	E – Potential energy of CBZ
Dsolv11 – Number of triple bonds in the solvent	E_strain – Local strain energy of CBZ
Dsolv15 – Relative number of triple bonds in the solvent	pmiZ – Moment of inertia C of CBZ
Dipole – of the solvent	Std_dim3 – Standard dimension 3 of CBZ

PC5 (Table 7.7) presents a large range of descriptors. The most positively loaded descriptors are made up of six CBZ descriptors, two basic solvent molecular descriptors and two bulk solvent properties. The CBZ descriptors consist of two partial charge surface area descriptors and a calculated dipole moment, which are all related. Also featured are density, moment of inertia and LUMO energy terms. The most negatively loaded descriptors are all CBZ descriptors except boiling point. These terms are describing the energy of the molecule or its size and shape.

### 7.1.3. Selection of the Most Valuable Descriptors

Since the most positively and negatively loaded descriptors within each component are the most significant, using the descriptors from each of PC1-13, should cover all the variation in the dataset and thus generate a good predictive model. The twenty six descriptors used in this analysis are shown in Table 7.8. This set of descriptors will be known as PCA-26 from this point forward.

Table 7.8 Most positively and negatively loaded descriptors from PC1-13 (PCA-26)

	Most positively loaded descriptors	Most negative loaded descriptors
<b>PC1</b>	Dsolv43 – 3D-Randić index (order 0) of the solvent	D69 – Molecular surface area of CBZ
<b>PC2</b>	Dsolv74 - PPSA-3 – Partial positive surface area of the solvent	Dsolv69 - Molecular surface area of the solvent
<b>PC3</b>	D68 - Moment of inertia C of CBZ	D84 - FNSA-1 – Fractional negative surface area of CBZ
<b>PC4</b>	Dsolv76 - PNSA-2 – Partial negative surface area of the solvent	E_ang -Angle bend potential of CBZ
<b>PC5</b>	D77 - PNSA-3 – Partial negative surface area of CBZ	Vol – Van der Waals volume of CBZ
<b>PC6</b>	AM1_HOMO - Highest occupied molecular orbital energy of CBZ	AM1_HF – Heat of formation of CBZ
<b>PC7</b>	Activity of CBZ in the solvent	Density of the solvent
<b>PC8</b>	PM3_HF – Heat of formation of CBZ	Dsolv32 - Structural information Content (order 0) of the solvent
<b>PC9</b>	Henry's law constant of the solvent	Dsolv11 – Number of triple bonds in the solvent
<b>PC10</b>	Dsolv15 – Relative number of triple bonds in the solvent	MNDO_Eele – Electronic energy of CBZ
<b>PC11</b>	PM3_Eele – Electronic energy of CBZ	AM1_Eele – Electronic energy of CBZ
<b>PC12</b>	Vapor pressure of the solvent	Std_dim1 – Standard dimension 1 of CBZ
<b>PC13</b>	Solubility of CBZ in the solvent	Viscosity of the solvent

FormRules<sup>[6]</sup> and INForm<sup>[7]</sup> analysis of these 26 descriptors (PCA-26) was carried out, with the results summarised in Table 7.9 and rules in Electronic Appendix, Chapter 7, file 7.3. An overall average  $R^2$  value of 71.55 % was obtained, which is an improvement on the previous analyses carried out using the component score values (Table 7.1 and Table 7.2). This suggests that there is redundant information in the scores, which may mask the relevant descriptors contributions.



Table 7.9 FormRules and INForm results of PCA-26

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	90.32		94.45	100.00
Form II	51.46		81.25	-42.13
Form III	56.58		91.40	-55.61
Dihydrate	97.01		76.94	100.00
Solvate	97.02		99.95	100.00
<b>Average R<sup>2</sup></b>	<b>78.48 %</b>		<b>88.80 %</b>	<b>40.45 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>64.63 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>71.55 %</b>	

This method of selection is a good way to reduce the number of descriptors quickly. However, it would be optimal to reduce this number further. Based on the knowledge of the scree plot and that the majority of the variance of the data is contained in PC1-5, a similar selection method was employed to reduce the descriptor set.

Using only the most positively and negatively loaded descriptor from PC1-5 (d68, d69, d77, d84, dsolv43, dsolv69, dsolv74, dsolv76, E\_ang, and vol) and including rate and temperature, an ANN was run. The results are shown in Table 7.10 and rules presented in Electronic Appendix, Chapter 7, file 7.3. This set of descriptors will be referred to as PCA-10 from this point forward.

Table 7.10 FormRules and INForm results of PCA-10 analysis

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	96.60		99.61	100.00
Form II	51.46		80.61	15.68
Form III	53.83		79.30	34.52
Dihydrate	97.34		99.70	100.00
Solvate	97.02		99.94	77.61
<b>Average R<sup>2</sup></b>	<b>79.25 %</b>		<b>91.83 %</b>	<b>65.56 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>78.70 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>78.97 %</b>	

Compared to the overall average  $R^2$  value of 67.46 % generated by using the score values of PC1-5 and the 71.55 % average from PCA-26 analysis, this method shows further improvement.

Based on the improvement made to the model by using the two most significant descriptors in each of PC1-5, the question was asked as to whether the model would improve further if the two most positively and two most negatively loaded descriptors were used. This set of descriptors will be referred to as PCA-20 from this point forward and are shown in Table 7.11.

Table 7.11 Top and bottom two descriptors from PC1-5 (PCA-20)

	Most positively loaded descriptors	Most negative loaded descriptors
<b>PC1</b>	Dsolv43 – 3D-Randić index (order 0) of the solvent	D69 - Molecular surface area of CBZ
	Dsolv20 - Randić index (order 1) of the solvent	Dielectric constant - of the solvent
<b>PC2</b>	Dsolv74 - PPSA-3 – Partial positive surface area of the solvent	Dsolv69 - Molecular surface area of the solvent
	Dsolv65 -3D-Bonding information content (order 2) of the solvent	Dsolv46 - 3D-Randić index (order 3) of the solvent
<b>PC3</b>	D68 - Moment of inertia C of CBZ	D84 - FNSA-1 – Fractional negative surface area of CBZ
	Viscosity – of the solvent	D42 - 3D-Wiener index of CBZ
<b>PC4</b>	Dsolv76 - PNSA-2 – Partial negative surface area of the solvent	E_ang -Angle bend potential of CBZ
	Dsolv5 – Relative number of hydrogen atoms in the solvent	E_str - Bond stretch potential of CBZ
<b>PC5</b>	D77 - PNSA-3 – Partial negative surface area of CBZ	Vol – Van der Waals volume of CBZ
	Dens – Relative molecular mass divided by van der Waals volume of CBZ	Boiling point – of the solvent

FormRules<sup>[6]</sup> and INForm<sup>[7]</sup> analysis was carried out with the results shown in Table 7.12 and the rules presented in Electronic Appendix, Chapter 7, file 7.4. Overall, further improvement in prediction was seen, with INForm<sup>[7]</sup> improving its performance from 78.70 % to 85.68 %. However, there was only a slight improvement in FormRules<sup>[6]</sup> results, from 79.25 % to 79.31 %.

Table 7.12 FormRules and INForm results of PCA-20

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	96.66		99.64	100.00
Form II	52.90		76.02	65.98
Form III	52.02		72.32	43.20
Dihydrate	98.30		99.70	100.00
Solvate	96.67		99.94	99.99
<b>Average R<sup>2</sup></b>	<b>79.31 %</b>		<b>89.52 %</b>	<b>81.83 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>85.68 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>82.49 %</b>	

Based on the success of the analysis with reduced descriptor sets, the number of PCs used was further reduced to assess whether a successful predictive model could be built using fewer descriptors. The results are summarised in Table 7.13 and can be found in Electronic Appendix, Chapter 7, file 7.5.

Table 7.13 Summary of descriptor reduction results. The number in brackets is the number of descriptors used in the ANN

	PC1-5 (20)	PC1-5 (10)	PC1-4 (16)	PC1-4 (8)		
FormRules average R <sup>2</sup> (%)	79.31	79.25	79.38	79.25		
INForm average R <sup>2</sup> (%)	85.68	78.70	82.84	88.48		
Overall average R <sup>2</sup> (%)	82.49	78.97	81.11	83.87		
	PC1-3 (12)	PC1-3 (6)	PC1-2 (8)	PC1-2 (4)	PC1 (4)	PC1 (2)
FormRules average R <sup>2</sup> (%)	79.38	79.25	75.14	64.12	44.78	44.80
INForm average R <sup>2</sup> (%)	73.58	74.06	71.05	30.80	70.10	80.99
Overall average R <sup>2</sup> (%)	76.48	76.66	73.09	47.46	57.44	62.89

Table 7.13 shows that PC1-4 (8 descriptors, to be referred to as PCA-8) produces an average result of 83.87 %, which is the highest seen throughout this PCA analysis. A full breakdown of PCA-8 results can be found in Table 7.14, and rules presented in Electronic Appendix, Chapter 7, file 7.6.

Table 7.14 FormRules and INForm results of PCA-8 analysis

<b>R<sup>2</sup> values for each form (%)</b>	<b>FormRules</b>		<b>INForm Training</b>	<b>INForm Testing</b>
Form I	96.60		99.63	100.00
Form II	51.46		85.46	76.41
Form III	53.83		75.88	49.72
Dihydrate	97.34		99.70	100.00
Solvate	97.02		99.17	98.86
<b>Average R<sup>2</sup></b>	<b>79.25 %</b>		<b>91.97 %</b>	<b>85.00 %</b>
			<b>INForm average R<sup>2</sup></b>	<b>88.48 %</b>
<b>FormRules and INForm average R<sup>2</sup></b>			<b>83.87 %</b>	

## 7.2. Optimisation of PCA Results

From the initial analysis, the PCA-8 and PCA-20 sets performed most successfully, generating average R<sup>2</sup> values of 83.87 % and 82.49 % respectively. Further analysis of the descriptors was carried out in order to create a fully optimised set as follows. All eight descriptors in PCA-8 feature in PCA-20. However, there are twelve additional descriptors that may affect the success of the prediction. Eight of these additional descriptors have already been analysed (PCA1-4, 16 descriptors) generating a result of 81.11 %, which is less successful at predicting than PCA-20. The remaining four additional descriptors (d77, vol, dens and boiling point, detailed in appendix section 12.2) must have a positive influence on prediction to increase the average result. By using this information, further analysis was carried out (Table 7.15). As PCA-8 is the most successful set of descriptors, this will be used as the starting point. The four additional descriptors from PCA-20 (d77, vol, dens and boiling point) will then be added to PCA-8 to assess their individual and combined impact on prediction.

Table 7.15 Optimisation results

Descriptors	PCA-8	Opt 1	Opt 2	Opt 3	Opt 4	Opt 5	Opt 6	Opt 7	Opt 8	Opt 9	Opt 10	Opt 11	Opt 12	Opt 13	Opt 14	Opt 15
PCA-8	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
D77		X	X	X		X	X	X	X				X			
Vol		X	X	X	X		X			X	X			X		
Dens		X	X		X	X		X		X		X			X	
Boiling point		X		X	X	X			X		X	X				X
<b>FormRules Average R<sup>2</sup> (%)</b>	79.25	79.99	79.29	79.99	80.02	79.99	79.25	79.25	79.99	79.25	80.02	80.02	79.25	79.25	79.25	80.02
<b>INForm Average R<sup>2</sup> (%)</b>	88.48	79.78	79.22	80.29	77.05	64.70	87.38	80.90	88.62	81.67	78.68	73.75	78.87	73.67	74.49	71.45
<b>Overall Average R<sup>2</sup> (%)</b>	83.87	79.89	79.24	80.14	78.54	72.35	83.32	80.07	84.31	80.46	79.35	76.89	79.06	76.46	76.89	75.74

From Table 7.15, Opt 8, which is PCA-8 plus d77 and boiling point, has shown an improved overall average  $R^2$  value of 84.31 % from the previous 83.87 %. This set of ten descriptors is therefore determined to be the most successful set of descriptors for predicting the polymorphic form of CBZ crystallisation experiments. The ten descriptors featured in the most successful set are presented in Table 7.16 with a brief explanation of each. A more in depth discussion of these descriptors can be found in section 7.3

Table 7.16 The ten most successful descriptors for polymorphic form prediction as determined by PCA analysis

Descriptor code and Definition	
D68	Moment of inertia C of the CBZ molecule
D69	Molecular surface area of the CBZ molecule
D77	PNSA-3, atomic charge weighted partial negative surface area of the CBZ molecule
D84	FNSA-1, fractional partial negative surface area the CBZ molecule (PNSA-1/total molecular surface area)
E_ang	Angle bend potential energy of the CBZ molecule
Dsolv43	3D-Randić index (order 0) of the solvent molecule
Dsolv69	Molecular surface area of the solvent molecule
Dsolv74	PPSA-3, atomic charge weighted partial positive surface area of the solvent molecule
Dsolv76	PNSA-2, total charge weighted partial negative surface area of the solvent molecule
Boiling point	Literature value boiling point of the solvent molecule

### 7.2.1. Analysis of the Optimised Set

From work in section 5.2 it is known that the descriptor set as a whole is highly correlated. Based on this knowledge, the linear correlations between the optimised set of descriptors and the whole dataset have been calculated and represented diagrammatically (Figure 7.2). The diagram shows, using connecting lines, the descriptors that are highly correlated to the optimised set of descriptors (Table 7.16).



Figure 7.2 displays those descriptors that are highly correlated to the descriptors presented in the optimised set. D68, d69, d84, dsolv69, dsolv74, E\_ang and boiling point are all correlated to a variety of different descriptors (detailed in appendix section 12.2).

Two of the descriptors in the optimised set have no correlations with any other descriptor, these are dsolv76 and d77. Both of these descriptors describe charged partial surface areas, one of the solvent and one the CBZ molecule. See appendix section 12.2 for details

Dsolv43, which is the 3D-Randić index for the solvent, is a highly correlated descriptor. It is mainly correlated with other solvent descriptors, but is also correlated with d69, which is the molecular surface area of the CBZ molecule. The Randić index is a measure of the branching of a molecule and therefore also represents the molecules size. Initially the correlation of the Randić index of the solvent and the surface area of the CBZ appears unusual. However, the surface area of the CBZ molecule is only varied due to the solvent force field it is modelled in. If the amount of branching of the solvent molecule has an impact upon one of the parameters within a solvent forcefield, then perhaps a correlation between these two descriptors is less surprising.

A summary of these correlations is presented in Table 7.17. The total number of correlations within the whole descriptor dataset and also the most strongly correlated descriptors to the optimised set are presented.

Table 7.17 Summary of linear correlations of the optimised set

Descriptor in Optimised set	Total number of correlations ( $\pm 0.8$ to $\pm 1$ )	Most positively correlated descriptor	Most negatively correlated descriptor
<b>Dsolv43</b>	32	Dsolv20	Dsolv76
<b>Dsolv69</b>	2	Dsolv46	-
<b>Dsolv74</b>	5	Dsolv65	-
<b>Dsolv76</b>	0	-	-
<b>D68</b>	2	Viscosity	Dsolv13
<b>D69</b>	8	-	Dsolv6
<b>D77</b>	0	-	-
<b>D84</b>	4	-	D81/D82/D85
<b>E_ang</b>	5	E_str	E_ele
<b>Boiling point</b>	2	Surface tension	-



As a further test to the robustness of the PCA determination of the final descriptor set, it was examined whether one of the final descriptors could be substituted by a highly correlated descriptor. As dsolv43 is highly correlated it would not be very efficient to interrogate all possible combinations of descriptors. Therefore the most positively and negatively correlated descriptors were analysed. The results are summarised in Table 7.18.

The optimisation sets were generated by analysing all of the positively correlated descriptors, then all of the negatively correlated descriptors and then a combination of both. During the combination analysis, the descriptor was chosen based upon which was closest to either a correlation coefficient of 1 or -1. When no strongly correlated descriptor was identified, the original descriptor was used. The descriptors used within each set are highlighted with a cross in Table 7.18, with the results also presented.

Table 7.18 Summary of correlation optimisations

Descriptors	Optimised set	Corr. 1	Corr. 2	Corr. 3	Corr. 4	Corr. 5	Corr. 6
Dsolv43	X						
Dsolv69	X		X	X	X	X	
Dsolv74	X		X	X	X	X	
Dsolv76	X	X	X	X	X	X	X
D68	X						
D69	X	X					
D77	X	X	X	X	X	X	X
D84	X	X					
E_ang	X						
Boiling point	X		X	X	X	X	
Dsolv20		X					X
Dsolv46		X					X
Dsolv65		X					X
Viscosity		X					X
E_str		X					
Surface tension		X					X
Dsolv67			X	X	X	X	
Dsolv13			X	X	X	X	
Dsolv6			X	X	X	X	X
D81			X	X			
D82			X		X		X
D85			X			X	
E_ele			X				X
FormRules Average $R^2$ (%)	79.99	78.84	80.24	80.24	80.24	80.24	80.00
INForm Average $R^2$ (%)	88.62	77.03	77.47	71.60	86.84	77.40	82.59
Overall Average $R^2$ (%)	84.31	77.94	78.86	75.92	83.54	78.82	81.30

No improvement in overall average  $R^2$  value was identified by using the highly correlated descriptors in an ANN. This therefore shows that the descriptors identified by PCA are the most effective for predicting the polymorphic form of CBZ; it also indicates that the PCA optimisation technique has eliminated correlated descriptors reliably.

### 7.3. Discussion of the Descriptors in the Optimised Set

The ten descriptors featured in the optimised set from PCA data reduction techniques are a mixture of CBZ and solvent properties. Table 7.19 summarises the descriptors meanings and the type of descriptor they are.

Table 7.19 Summary of the descriptors in the most successful set

Descriptor	Meaning	Type of descriptor	Calculated from Solvent or CBZ molecule
D68	Moment of inertia C of the CBZ molecule	Geometrical	CBZ
D69	Molecular surface area of the CBZ molecule	Geometrical	CBZ
D77	PNSA-3, atomic charge weighted partial negative surface area of the CBZ molecule	Charge distribution	CBZ
D84	FNSA-1, fractional partial negative surface area of the CBZ molecule (PNSA-1/total molecular surface area)	Charge distribution	CBZ
E_ang	Angle bend potential energy of the CBZ molecule	Quantum chemical	CBZ
Dsolv43	3D-Randić index (order 0) of the solvent molecule	Topological	Solvent
Dsolv69	Molecular surface area of the solvent molecule	Geometrical	Solvent
Dsolv74	PPSA-3, atomic charge weighted partial positive surface area of the solvent molecule	Charge distribution	Solvent
Dsolv76	PNSA-2, total charge weighted partial negative surface area of the solvent molecule	Charge distribution	Solvent
Boiling point	Boiling point of the solvent molecule	Bulk	Solvent

Eight out of the ten descriptors feature within the rules generated in FormRules<sup>[6]</sup> for this set. E\_ang, which is the angle bend potential energy for the CBZ molecule in the solvent force field, and dsolv76, which is total charge weighted partial negative surface area of the solvent molecule, do not feature. When the ANN is run without these two descriptors the overall  $R^2$  value is reduced from 84.31 % to 80.72 %. This suggests that these descriptors may have an important role in the successful prediction of polymorphic form.

As outlined in section 4.6.2, a more detailed analysis method was devised to assess an individual descriptors effect upon the prediction. When this analysis was carried

out, E\_ang and dsolv76 showed some effect upon prediction of form II and III. These results will be discussed in the following corresponding sections.

### 7.3.1. The Prediction of Form I

As was previously discussed (in section 5.4.1) form I was not crystallised as a pure form in this research, and was seen on only two occasions. Because of this, there is very little training data available and therefore rules have been generated based upon only a small dataset (Table 5.37), affecting their reliability. This problem could be addressed in the future by increasing the number of form I crystallising experiments in the training set. This would allow rules to be created based upon a larger dataset, thus improving the reliability.

The rules generated (Table 5.37) contain both the rate and temperature at which the crystallisation were conducted and also dsolv74, which is the atomic charge weighted partial positive surface area of the solvent molecule.

Table 7.20 Rules generated in FormRules for form I prediction

Rules generated for Form I prediction			
IF dsolv74 is LOW AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is LOW AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is LOW AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is LOW AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is LOW AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is LOW AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is MID AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is MID AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (0.95)	
IF dsolv74 is MID AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is MID AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is MID AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is MID AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv74 is HIGH AND rate is HIGH AND Temp is HIGH	THEN Form I is	HIGH (1.00)	

The majority of the rules generated lead to a low prediction of form I. This is expected from the data within the training set. When the rule for high form I prediction are examined, rate, temperature and dsolv74 all need to be at their highest value to produce a high prediction. Figure 7.3 shows the normalised descriptor values of the two form I producing experiments. From the plot it is clear that on no occasion were the three descriptors all at a high value. This is consistent with the rule, as form I was only produced as part of a mixed product.

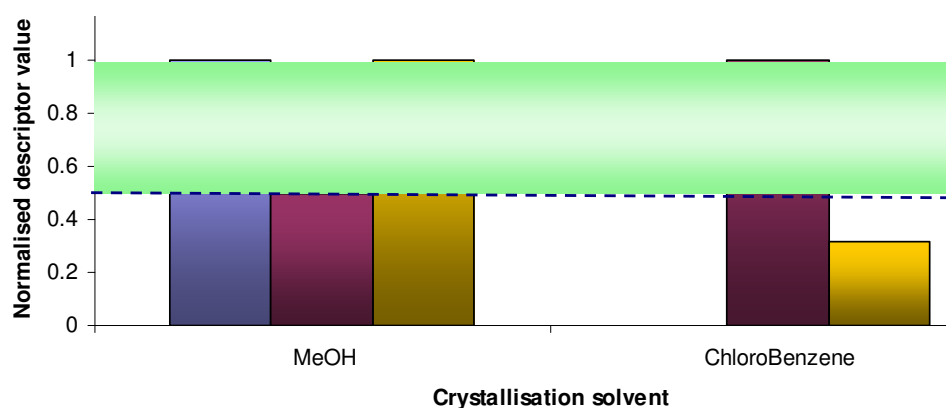


Figure 7.3 Crystallisation solvents in which form I is produced plot against the three rule descriptors, rate (blue), temperature (purple) and dsolv74 (yellow). The green section highlights the optimal value area for form I production.

As discussed in chapter 5 high rates and temperatures are often associated with the crystallisation of a metastable polymorphic form,<sup>[8-10]</sup> which is consistent with this rule. However, this research shows that these factors alone cannot predict polymorphic form; other properties of the solute and solvent have an impact.

Dsolv74 is the summation of the atomic charge weighted partial positive surface area (PPSA-3) of the solvent molecule used in the crystallisation. This type of descriptor is known as a charged partial surface area (CPSA) descriptor. It is calculated using Equation 7.1, where  $q_A$  is the atomic partial charge and  $S_A$  is the solvent accessible surface area of the molecule.

$$PPSA-3 = \sum_A q_A S_A \quad \text{Equation 7.1}$$

As previously stated in chapter 5, the CPSA descriptors feature in many different areas of research,<sup>[11-15]</sup> for example in predicting micelle-water partition

coefficients<sup>[15]</sup> and in the prediction of logP values.<sup>[14]</sup> It has been suggested that the CPSA descriptors contain information about how the molecules interact,<sup>[12, 16]</sup> which is relevant to the nucleation and crystallization of a polymorphic molecule from solution. However, since this rule was generated based upon on a very small dataset, it would be unwise to draw firm conclusions based upon these results. Further work is required in order to build up a more reliable predictive model of form I crystallization.

### 7.3.2. The Prediction of Form II

Form II was the most commonly crystallised metastable polymorph of CBZ in this research (see Electronic Appendix, Chapter 4, file 4.4). Three rules were generated (Table 7.21) that featured, rate, d68 and d69. As was mentioned in the form I discussion, high rates are often associated with the crystallisation of the metastable polymorph,<sup>[8-10]</sup> which is consistent with the rule generated. D68 and d69 are both CBZ descriptors, describing the moment of inertia C and the molecular surface area of the molecule respectively.

Table 7.21 Rules generated in FormRules for form II prediction

Rules generated for Form II prediction			
SubModel:1	IF d69 is LOW	THEN Form II is	HIGH (1.00)
	IF d69 is HIGH	THEN Form II is	LOW (1.00)
SubModel:2	IF rate is LOW	THEN Form II is	LOW (1.00)
	IF rate is HIGH	THEN Form II is	HIGH (0.64)
SubModel:3	IF d68 is LOW	THEN Form II is	HIGH (1.00)
	IF d68 is HIGH	THEN Form II is	LOW (1.00)

D69 is the molecular surface area of the CBZ molecule in the solvent force field. It is a geometrical descriptor that uses the van der Waals radii of the atoms within the molecule to give the best surface area approximation<sup>[5, 12]</sup> (Figure 7.4).

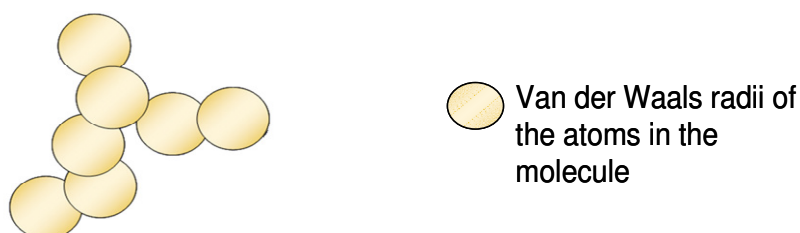


Figure 7.4 Calculation of the total molecular surface area using van der Waals radii, adapted from Stanton<sup>[13]</sup>

When the normalised descriptor values of the pure form II producing experiments are plotted (Figure 7.5), the majority of the values fit within the guidelines stated in the rules.

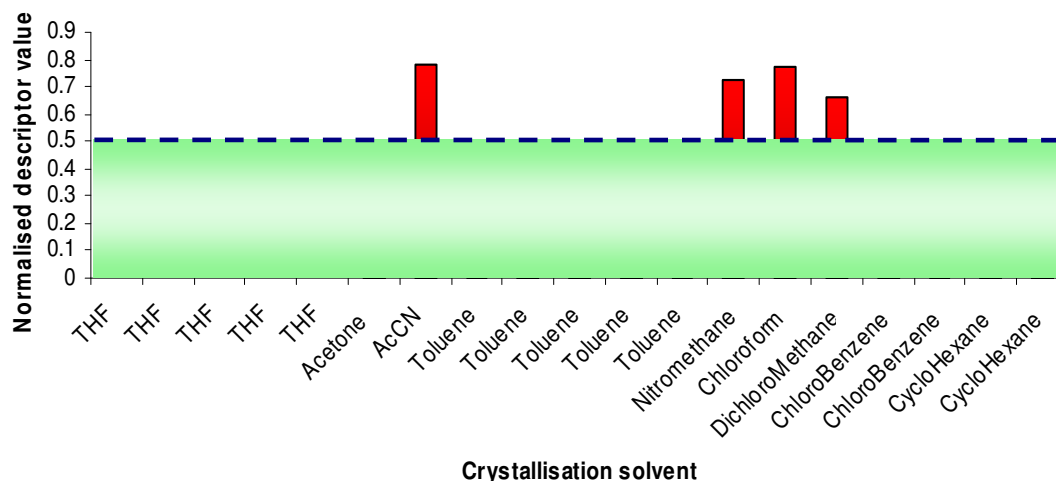


Figure 7.5 Crystallisation solvents in which form II is the pure product plot against normalised d69 descriptors values. The green section highlights the optimal value area for form II production.

The differences in molecular surface area of the CBZ molecule are very small, due to their slightly altered geometry within the solvent force field. This is similar to the discussion of  $E_{vdw}$  and  $MNDO\_dipole$  descriptors in chapter 5. These differences in geometry may affect the interactions with other solute and solvent molecules in solution. Geometric descriptors, such as molecular surface area and molecular volume, have been noted to affect solvation in research by Bodor et al.<sup>[12]</sup>, which used ANNs to predict aqueous solubility. The dipole moment also represents the solute-solvent interactions overall, but it was suggested that other descriptors, such as molecular surface area, volume and charge density are said to “refine the description of solvation”.<sup>[12]</sup>

A correlation between the molecular surface area and dielectric constant has also been made in research by Liu et al.<sup>[17]</sup>. This is significant because it represents the polarisability of the molecule, which may contribute to the intermolecular reactions in solution. These relationships were examined using the data in this research, but as the surface area was calculated for the solute and the dipole moment for the solvent, there were no significant correlations detected (appendix section 12.9).

Although not featured in the rules,  $E_{ang}$  had a slight impact upon form II prediction when the detailed analysis method (section 4.6.2) was carried out. Figure 7.6

demonstrates the reduction in form II prediction when E\_ang is at its highest value. The different E\_ang values do not affect the rule overall, but have an impact on the predicted values. This would explain why E\_ang is not featured in the rules, but has been identified in the INForm<sup>[7]</sup> analysis as important for prediction.

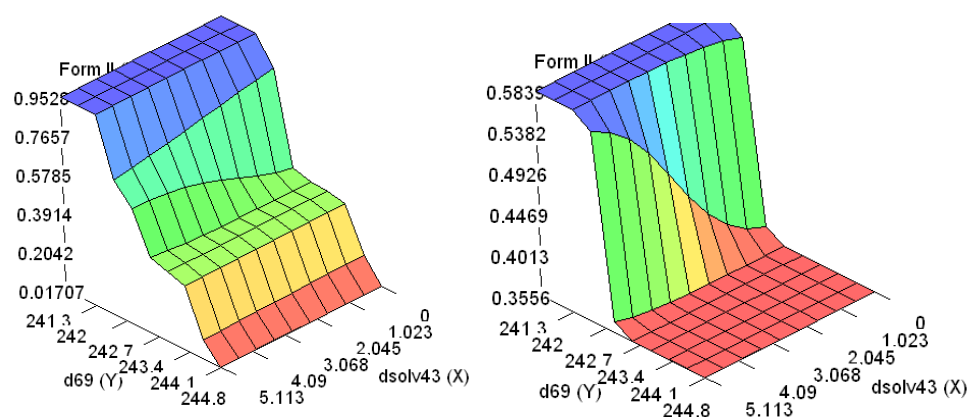


Figure 7.6 Original plot for form II prediction (left), effect on plot when a high E\_ang value is used (right)

E\_ang is the angle bend potential energy for the CBZ molecule in the solvent force fields.<sup>[18]</sup> It is a measure of deviation from the standard bond angles in the molecule. Therefore, as with many of the CBZ descriptors, the difference between each value is very small and may potentially be overlooked as a useful descriptor. This descriptor describes the geometry and flexibility of the molecule, perhaps affecting interactions with solute and solvent molecules in solution. E\_ang has been used in the literature as a descriptor in mostly biological applications,<sup>[19-21]</sup> but it is not as frequently used as other descriptors.

As stated in 7.3.1, high evaporation rates are often associated with the crystallisation of metastable forms.<sup>[8-10]</sup> Rapid crystallisation allows high supersaturation levels to be achieved more quickly, leading to nucleation in solution. If Ostwald's Rule of Stages is followed, the least stable polymorph would crystallise first,<sup>[22, 23]</sup> which in this research is form II.

When the rates of the pure form II crystallising experiments were normalised and plotted (Figure 7.7), the majority are found to follow the rule stated. This confirms the relationship between metastable form crystallisation and high evaporation rates, but does not offer any new insights into the molecular level interactions that lead to polymorphic crystallisation.



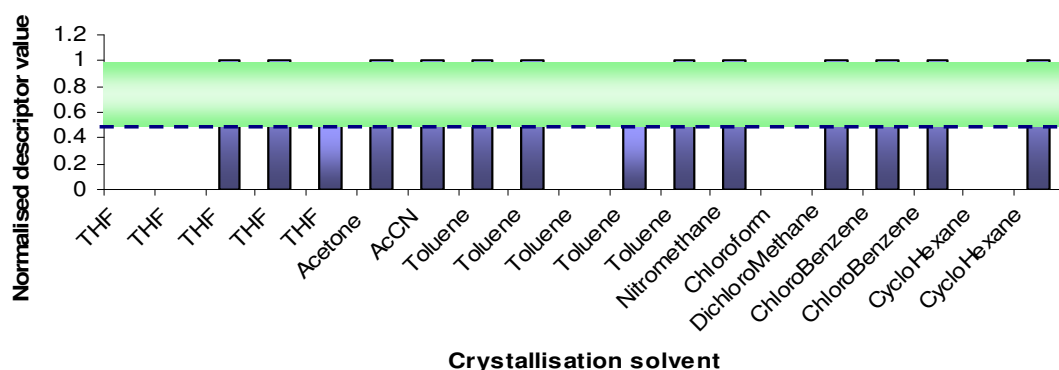


Figure 7.7 Crystallisation solvents in which form II is the pure product plot against normalised rate values. The green section highlights the optimal value area for form II production.

D68 is the moment of inertia  $C$  for the CBZ molecule in the solvent force field. These descriptors are classed as geometrical descriptors<sup>[24]</sup> and are obtained from the mass and three-dimensional coordinates of atoms in the molecule. Using the rigid rotor approximation, the moments of inertia of a single molecule  $I_A$ ,  $I_B$  and  $I_C$  are calculated using Equation 7.2, Equation 7.3 and Equation 7.4, where  $I_C > I_B > I_A$ .<sup>[25]</sup>

$$I_A = \sum_i^n m_i r_{ix}^2 \quad \text{Equation 7.2}$$

$$I_B = \sum_i^n m_i r_{iy}^2 \quad \text{Equation 7.3}$$

$$I_C = \sum_i^n m_i r_{iz}^2 \quad \text{Equation 7.4}$$

The mass of each atom is represented by  $m_i$ , with  $r_{ix/y/z}$  denoting the distance between the  $i$ th atomic nucleus and the main rotational axes,  $x$ ,  $y$  and  $z$ . The number of atoms is represented by  $n$ .<sup>[25]</sup>

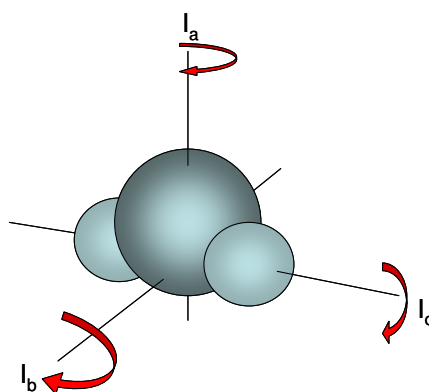


Figure 7.8 the axis of an single molecule, adapted from Atkins<sup>[25]</sup>

The moment of inertia is a measure of mass distribution in the molecule and also can determine how rotationally flexible parts of the molecule are.<sup>[4, 26]</sup> The rule states that a low d68 value leads to a high form II prediction. When the normalised d68 values for pure form II producing experiments are plotted (Figure 7.9), it is apparent that this rule holds strongly.

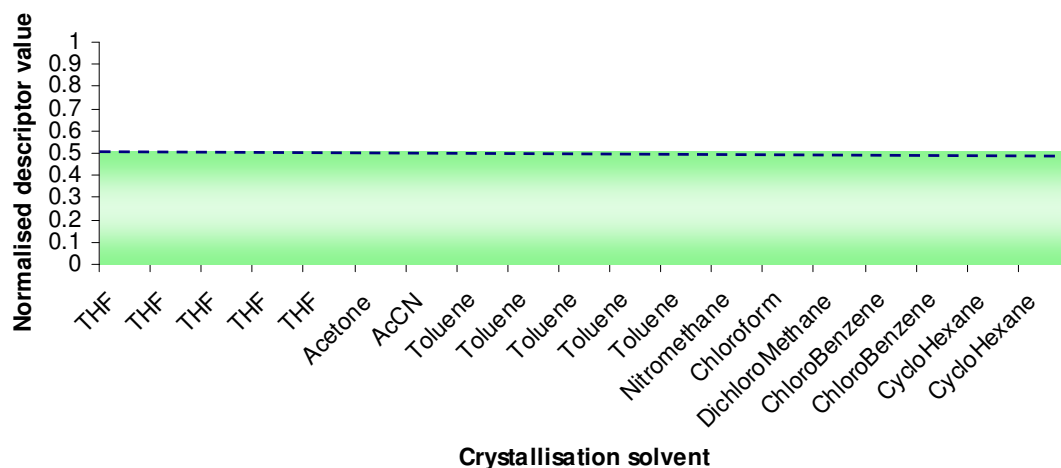


Figure 7.9 Crystallisation solvents in which form II is the pure product plot against normalised d68 values. The green section highlights the optimal value area for form II production.

Moment of inertia is a molecular descriptor that has been used in a range of research areas, from biological to chemical applications.<sup>[26-30]</sup> Moment of inertia  $x$ , or  $A$  by this notation, was used in previous research alongside other descriptors, to predict chromatographic retention times.<sup>[28, 29]</sup> The descriptor was used as a measure of molecular size, which was related to the electronic polarisability of the molecule<sup>[29]</sup>. Rohrbaugh et al.<sup>[29]</sup> concluded that by using the moment of inertia  $x$ , information was generated about the interactions between the stationary phase and the solute, which allowed retention time prediction. Collantes et al.<sup>[28]</sup> commented that retention times increased with decreasing moment of inertia  $x$ , but it was only a useful parameter for solutes of the same molecular weight. Since d68 is the moment of inertia  $c$  for the CBZ molecule in the different solvent force fields, the geometrical differences are very subtle. Perhaps because retention time increases with lower moment of inertia values, this suggests that the solute molecule interacts with the stationary phase. When these ideas are applied to solution crystallisation, perhaps a lower moment of inertia indicates more favourable interactions with solvent molecules, rather than

with other solute molecules. From the literature<sup>[31, 32]</sup> it is known that solvent molecules play an important stabilising role in form II crystallisation, and perhaps the moment of inertia is offering an indirect measure of molecular interactions. Both E\_ang and dsolv76 had an impact upon form II prediction when the detailed analysis method (section 4.6.2) was carried out based upon rules involving d68. Figure 7.10 shows the impact that mid range and high values of E\_ang have upon the prediction of form II, and Figure 7.11 shows the effect of high dsolv76 values on form II prediction.

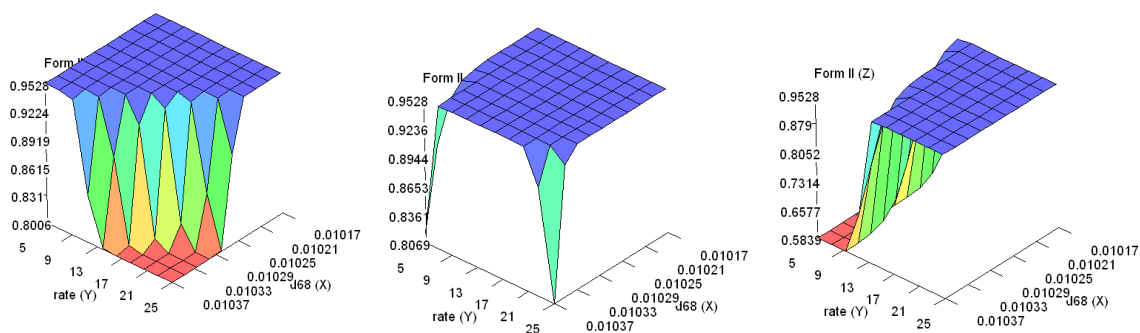


Figure 7.10 Original plot for form II prediction (left), effect on plot when a mid E\_ang value (centre) and effect on plot when high E\_ang value is used (right)

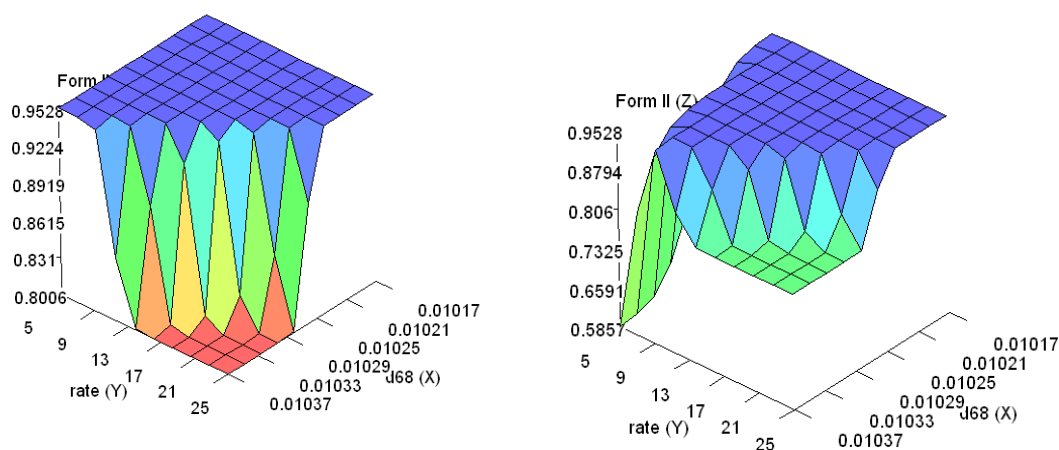


Figure 7.11 Original plot for form II prediction (left), and effect on plot when high dsolv76 value is used (right)

As E\_ang has been discussed with respect to its impact upon form II prediction, only the influence of dsolv76 will be discussed here. Dsolv76 is total charge weighted partial negative surface area (PNSA-2) of the solvent molecule and is calculated

using Equation 7.5, where  $q_A$  is the atomic partial charge and  $S_A$  is the solvent accessible surface area.<sup>[4]</sup>

$$PNSA - 2 = \sum_A q_A \sum_A S_A \quad \text{Equation 7.5}$$

Dsolv76 describes the negatively charged regions of the solvent molecule, which may therefore lead to further information about how solute and solvent molecules interact in solution. This descriptor takes into account the presence of negatively charged atoms available on the solvent accessible surface area of the molecule and also the overall solvent accessible surface area. Low values of dsolv76 are attributed to those solvents in which more electronegative atoms are featured upon the solvent accessible surface area. No direct link can be made between the individual dsolv76 values and the polymorphic form crystallised. However, this descriptor must contribute in tandem with other properties to lead to successful polymorphic predictions.

### 7.3.3. The Prediction of Form III

Form III is the thermodynamically stable form of CBZ and as such was crystallised most frequently in this research (results in Electronic Appendix, Chapter 4, file 4.4), our rules were generated for form III prediction, including the rate of evaporation, which was also seen in the manual descriptor analysis chapter. The descriptors identified as important were, d84, d77 and d68, which represent the partial negative and positive surface areas and moment of inertia of the CBZ molecule, and also the boiling point of the solvents (Table 7.22).

Table 7.22 Rules generated in FormRules for form III prediction

Rules generated for Form III prediction			
SubModel:1	IF d84 is LOW	THEN Form III is	HIGH (1.00)
	IF d84 is HIGH	THEN Form III is	LOW (1.00)
SubModel:2	IF b.p. is LOW AND d77 is LOW	THEN Form III is	HIGH (0.50)
	IF b.p. is LOW AND d77 is HIGH	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is LOW	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is HIGH	THEN Form III is	HIGH (1.00)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.72)
	IF rate is HIGH	THEN Form III is	LOW (1.00)
SubModel:4	IF d68 is LOW	THEN Form III is	HIGH (1.00)
	IF d68 is MID	THEN Form III is	HIGH (1.00)
	IF d68 is HIGH	THEN Form III is	LOW (1.00)

D68 previously featured in the form II rules, and is the moment of inertia C of the CBZ molecule in the different solvent force fields. The d68 rules for form II and III are very similar, which means that very little information can be extracted from them. A plot of the normalised d68 values has been created to assess whether the pure form III producing experiments have values in line with the guidelines in the rules (Figure 7.12). There are 36 pure form III producing experiments in the data set, therefore each solvent has been represented only once in these plots.

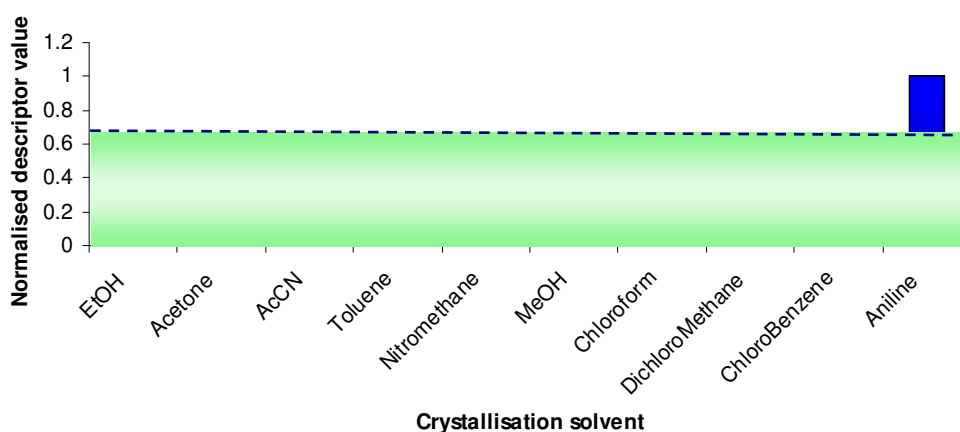


Figure 7.12 Crystallisation solvents in which form III is the pure product plot against normalised d68 values. The green section highlights the optimal value area for form III production.

Figure 7.12 demonstrates that in all but one solvent, the rule is correct. A low or medium moment of inertia value leads to a successful form III prediction. When the moment of inertia values for all of the crystallisation solvents are examined, all but

aniline are at a low value. This would perhaps suggest that very little information can be extracted from this descriptor.

The other descriptors in the form III rules have not been featured in any earlier rules. Therefore, they may present more insights into the molecular level interactions that lead to the crystallisation of form III.

D84 is a CPSA descriptor describing the fractional positive surface area (Equation 7.6) of the CBZ molecule in the solvent force field (FPSA-1). This is the ratio of the total molecular surface area (TMSA) and the partial positive surface area (PPSA), which were both discussed in section 5.4.<sup>[4]</sup> In this analysis d84 is used in form III prediction, whereas in chapter 5 similar components were found in rules for the CBZ dihydrate.

$$FPSA - 1 = \frac{PPSA - 1}{TMSA} \quad \text{Equation 7.6}$$

$$PPSA - 1 = \sum_A S_A \quad \text{Equation 7.7}$$

$$A \in \{\delta_A > 0\}$$

When the normalised d84 values are plot for the pure form III producing experiments (Figure 7.13), the majority of the values fall within the lower region (in green). The rules states that low d84 leads to high form III prediction.

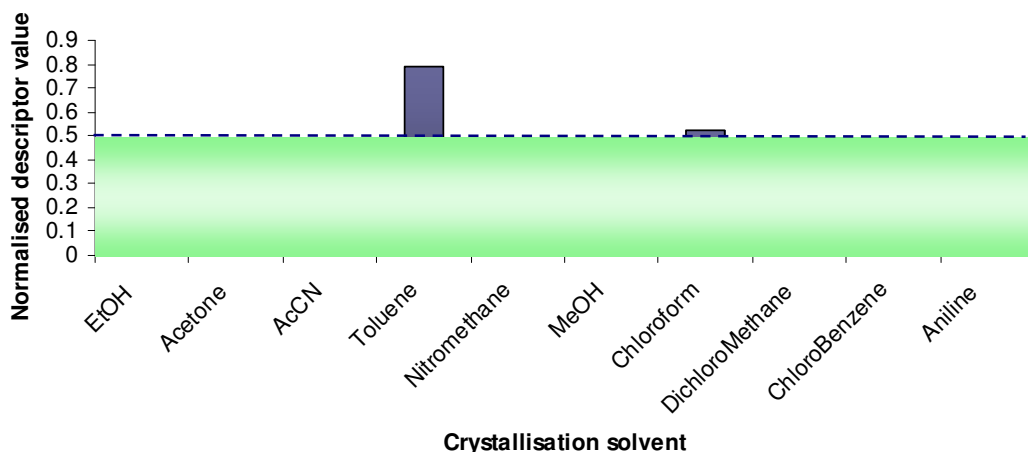


Figure 7.13 Crystallisation solvents in which form III is the pure product plot against normalised d84 values. The green section highlights the optimal value area for form III production.

Toluene and chloroform both fall outside the region for high form III prediction. This is interesting as toluene most commonly leads to the crystallisation of form II, and chloroform has lead to many mixed products in this research. Similarly to many of the other CBZ based descriptors, the differences in values are very subtle. The

TMSA value will induce the largest effect in this descriptor and based upon this assumption, the descriptor is essentially describing how compact the CBZ molecule is. How compact the molecule is may be related to its interaction surface and how it interacts with other solute and solvent molecules. CPSA descriptors have been used in the literature in biological research<sup>[33]</sup> and the prediction of physicochemical properties.<sup>[27, 34, 35]</sup> If perhaps this descriptor was describing the solvent molecule, a clearer picture of the interactions may be deduced. However, the subtlety of this descriptor makes interpretation difficult, but it is interesting that both FormRules<sup>[6]</sup> and INForm<sup>[7]</sup> have selected it as an important descriptor. This highlights how useful these machine learning techniques can be at highlighting subtle differences between molecules that may have been previously overlooked.

The rule selected as the most confident (coloured red and blue in Table 7.22) involves the boiling point of the solvent and d77, which is the partial negative surface area of the CBZ molecule (PNSA-3). Like the previous descriptor, d77 is a CPSA descriptor, which in this example examines the partial negative charges. It is calculated using Equation 7.8 where  $q_A$  is the atomic partial charge and  $S_A$  is the solvent accessible surface area<sup>[4]</sup>.

$$PNSA - 3 = \sum_A q_A S_A \quad \text{Equation 7.8}$$

The rule states that if the two descriptors are high or if they are both low, a high form III prediction will be made. When the pure form III producing experiments are plotted (Figure 7.14), a number of the experiments conform to this rule (highlighted by the green shading).

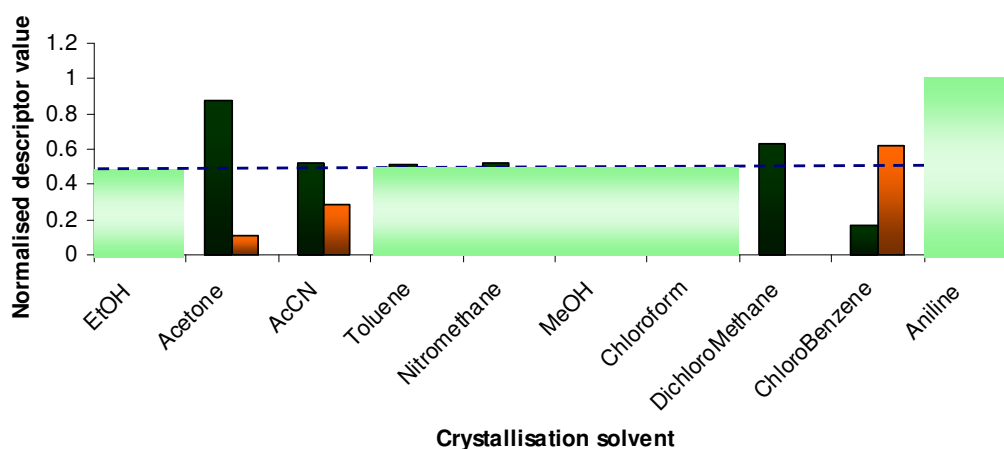


Figure 7.14 Crystallisation solvents in which form III is the pure product plot against normalised d77 and boiling point values. The green section highlights the optimal value area for form III production.

The prediction of boiling points has been the focus of other research<sup>[34, 36]</sup>, with connectivity indices, molecular energy and hydrogen bonding descriptors all being used<sup>[34]</sup>. Boiling points of solvents give an indication about how strongly solvent molecules interact, with higher boiling points indicating stronger interactions. With the rule stating that low boiling points and low d77 values can lead to the same prediction as high boiling point and high d77, it makes the rule very difficult to interpret. No correlations have been observed between either of these descriptors or with the bulk solvent properties. Also d77 is linearly uncorrelated to any other descriptor in this analysis. The identification of no obvious correlations highlights the extra information that can potentially be extracted by using these complex data analysis techniques.

Not only is the rule difficult to interpret directly, but also E\_ang and dsolv76 have been shown to have an effect on prediction of form III when detailed analysis was carried out. Both these descriptors were discussed previously because they effected form II predictions.

When the two descriptors in the rule are plotted (Figure 7.15), it is clear that this rule is not as straight forward as FormRules<sup>[6]</sup> would suggest. The general rule of both descriptors being high or both being low leading to a high form III prediction is not apparent from the ANN results in INForm.<sup>[7]</sup>

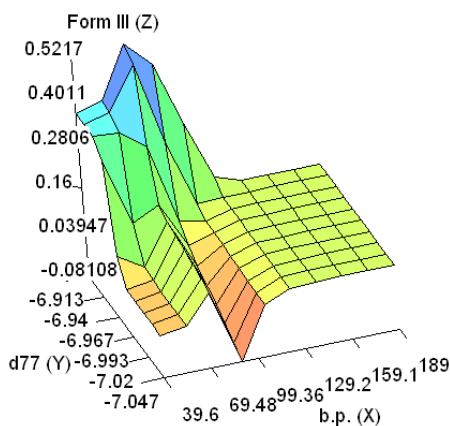


Figure 7.15 Plot of d77 and boiling point in the prediction of form III

When E\_ang is changed from its original low value to a mid range and high value, the level of form III prediction increases overall (Figure 7.16).



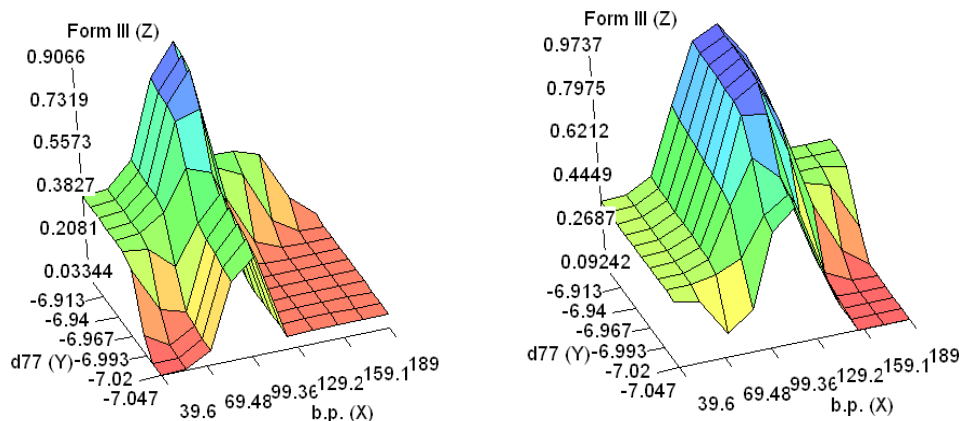


Figure 7.16 Plot of d77 and boiling point in the prediction of form III with mid range (left) and high (right) values of E\_ang

Similarly to this, the whole range of dsolv76 values has an impact upon the prediction of form III when plot against d77 and boiling point (Figure 7.17).

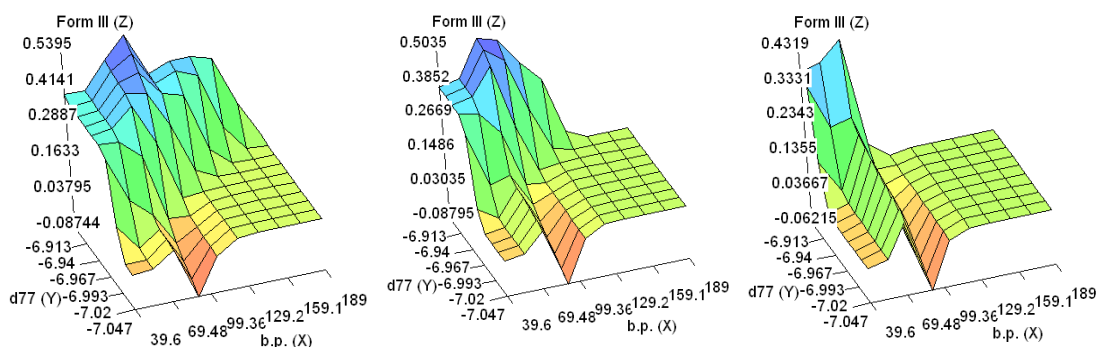


Figure 7.17 Plot of d77 and boiling point in the prediction of form III with low (left), mid range (centre) and high (right) values of dsolv76

Based on this information it is clear that an empirical rule for predicting form III crystallisation has not been found in this analysis. In fact, four different descriptors have an effect on the prediction of the form.

A final rule in which rate is featured has also been generated. As has been previously mentioned in chapter 5, low evaporation rates most commonly lead to the thermodynamically stable form being crystallised.<sup>[8-10]</sup> The normalised value for rate has been plotted against all pure form III producing solvents (Figure 7.18), and as a general rule, low rates do produce form III. When compared to other rules it is not the most confident, but it is an experimental parameter and does have an impact upon the crystallisation in many cases.

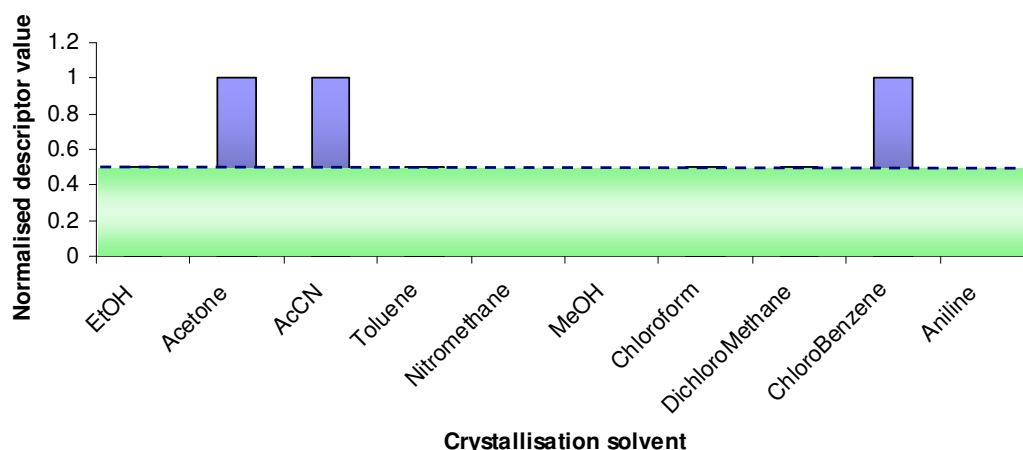


Figure 7.18 Crystallisation solvents in which form III is the pure product plot against normalised rate values. The green section highlights the optimal value area for form III production.

### 7.3.4. The Prediction of the Dihydrate

As discussed in section 5.44 the dihydrate form was never crystallised as the pure product in this research, and only featured in three crystallisation experiments (detailed in Electronic Appendix, Chapter 4, file 4.4). Similarly to form I, the data was still included in the analysis, but does call into question the reliability of the rules generated. To improve the rules, more dihydrate forming crystallisation experiments would need to be added to the training set. Rules were generated for dihydrate formation (Table 7.23) using dsolv74, dsolv69 rate and temperature in the predictions.

Table 7.23 Rules generated in FormRules for dihydrate prediction

Rules generated for dihydrate prediction		
IF dsolv74 is LOW AND Temp is LOW AND rate is LOW AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is LOW AND rate is LOW AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is LOW AND rate is MID AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is LOW AND rate is MID AND dsolv69 is HIGH	THEN DiHydrate is	LOW (0.97)
IF dsolv74 is LOW AND Temp is LOW AND rate is HIGH AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is LOW AND rate is HIGH AND dsolv69 is HIGH	THEN DiHydrate is	LOW (0.99)

Rules generated for dihydrate prediction -continued		
IF dsolv74 is LOW AND Temp is HIGH AND rate is LOW AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is HIGH AND rate is LOW AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is HIGH AND rate is MID AND dsolv69 is LOW	THEN DiHydrate is	LOW (0.50)
IF dsolv74 is LOW AND Temp is HIGH AND rate is MID AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv69 is LOW	THEN DiHydrate is	LOW (0.74)
IF dsolv74 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is LOW AND rate is LOW AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is LOW AND rate is LOW AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is LOW AND rate is MID AND dsolv69 is LOW	THEN DiHydrate is	HIGH (1.00)
IF dsolv74 is HIGH AND Temp is LOW AND rate is MID AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv69 is LOW	THEN DiHydrate is	LOW (0.50)
IF dsolv74 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv69 is HIGH	THEN DiHydrate is	LOW (0.98)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is MID AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is MID AND dsolv69 is HIGH	THEN DiHydrate is	LOW (0.66)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv69 is LOW	THEN DiHydrate is	LOW (1.00)
IF dsolv74 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv69 is HIGH	THEN DiHydrate is	LOW (1.00)

From Table 7.23 the majority of the rules lead to a low dihydrate prediction. Based upon the experimental results used in the training this is not surprising, as there were

no pure dihydrate products. The rule leading to a high dihydrate prediction states that dsolv74 must be high, rate at a mid range value and both temperature and dsolv69 to be low. When the experiments that crystallised as mixtures containing the dihydrate were plotted against the normalised descriptor values in the rule (Figure 7.19) these solvents display descriptor values that would lead to low dihydrate predictions.

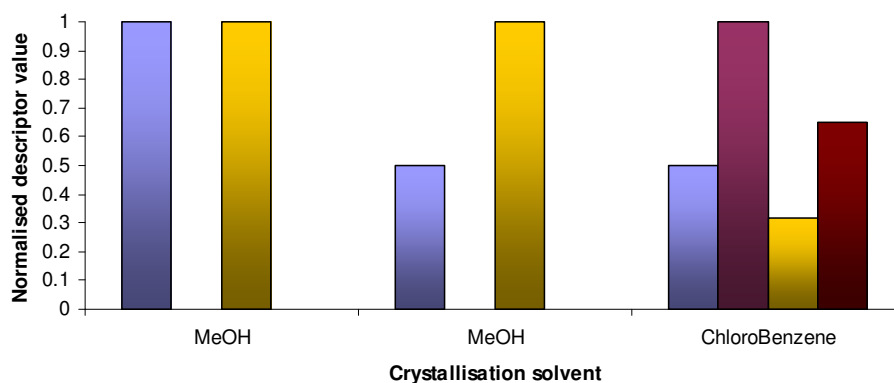


Figure 7.19 Crystallisation solvents in which dihydrate is part of the product plot against normalised rate (blue), temperature (purple), dsolv74 (orange) and dsolv69 (red) values.

Dsolv69 is the molecular surface area of the solvent molecule, with the CBZ molecular surface area (d69) featuring in the form II rules. As was previously mentioned, research by Bodor et al.<sup>[12]</sup> and Lui et al.<sup>[37]</sup> commented upon the correlations between molecular surface area and dipole moment and dielectric constants respectively. Using the data in this current research, no such empirical correlations were observed. Interestingly, the total molecular surface area descriptor (dsolv71) featured in the dihydrate rules in the manual analysis (section 5.4.4). Although a slightly different descriptor, they both describe the surface area of the solvent molecules.

Dsolv74, also observed in the form I rules, is the summation of the atomic charge weighted partial positive surface area (PPSA-3) upon the solvent molecule used in the crystallisation. Rate of evaporation and the temperature at which the crystallisations were conducted also featured in the dihydrate rules.

It is difficult to draw firm conclusions about the nucleation and crystallisation of the dihydrate based upon these descriptors due to the small amount of data available for training. Further crystallisations need to be carried out in order to generate more dihydrate forming training data. With a larger training set, perhaps more information can be extracted from the descriptors within the rules.

### 7.3.5. The Prediction of Solvates

CBZ can crystallise as many different solvates<sup>[32, 38-43]</sup>, but in this current research only the DMSO solvate was produced. Rules were generated based upon these results (Table 7.24) and featured the boiling point of the solvent and dsolv43, which is the 3D Randić index of the solvent molecule.

Table 7.24 Rules generated in FormRules for solvate prediction

Rules generated for solvate prediction			
IF b.p. is LOW AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)	
IF b.p. is LOW AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	
IF b.p. is MID AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)	
IF b.p. is MID AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	
IF b.p. is HIGH AND dsolv43 is LOW	THEN Solvate is	HIGH (1.00)	
IF b.p. is HIGH AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	

The rule for high solvate prediction matches the properties of the DMSO solvent, highlighted in green on Figure 7.20. However, no other solvents have the properties mentioned in the rule for high solvate formation. As discussed in section 5.4.5, there are examples in the literature of solvates being crystallised from acetone, THF and nitromethane,<sup>[38, 40, 44]</sup> but these solvents do not show the properties highlighted in the rules.

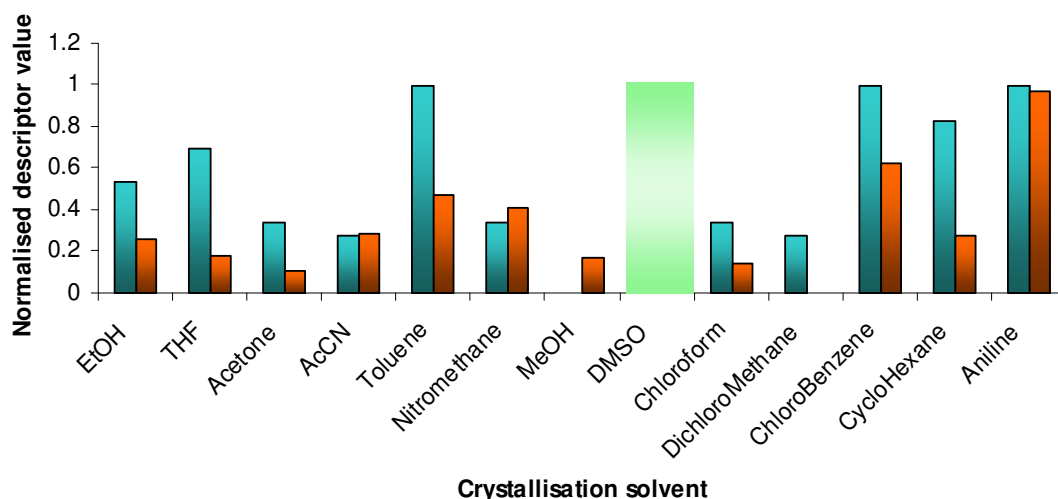


Figure 7.20 Crystallisation solvents plot against normalised dsolv43 (blue) and boiling point (orange) values.

Dsolv43, the 3D Randić index of the solvent molecule, is a topological descriptor that was created in order to uniquely identify molecules using a single parameter.<sup>[45]</sup> It places importance on molecular branching and structure.<sup>[26, 45, 46]</sup> In the molecule each non-hydrogen atom is given a number, and these atoms are then also given a value of  $1/\sqrt{\nu}$ , with  $\nu$  being “the number of non-hydrogen bonds in which the atom is involved.”<sup>[47]</sup> The numbered atoms are then used as a path.

In a more complex molecule, different weights are also placed on each bonds value depending on whether it is a single, double, triple or aromatic bond.<sup>[45, 48]</sup> The 3D structure can also be taken into account by consideration of the distances between atoms, and weighting the values appropriately<sup>[49]</sup>.

The Randić index has been used in previous research that estimates physicochemical properties of a molecule<sup>[50, 51]</sup>, to predict retention times<sup>[52]</sup> and in prediction of biological inhibitor molecules<sup>[26]</sup>. It has also been noted in research by Schweitzer and Morris<sup>[24]</sup> that in combination with CPSA descriptors, connectivity indices are essential to improve the quality of the predictive models.

It is interesting to observe that in section 5.4.5, a bonding information content descriptor was featured in the solvate rules. Perhaps descriptors that describe the branching and overall size of the solvent molecule are very important in solvate prediction. It is clear that the solvent must interact with the solute in order for a solvate to be crystallised, but perhaps simple topological properties can be used to assess whether a solvate will form.

### **7.3.6. Summary of the Optimised Descriptors**

There were ten descriptors featured in the optimised set, with eight of these featuring in the rules. E\_ang and dsolv76 did not feature in the rules, but were found to have an impact upon both form II and III prediction. The ten descriptors cover a range of CBZ and solvent properties and are summarised in Table 7.25.

Table 7.25 Summary of the descriptors involved in the CBZ predictive rules

Form predicted	Descriptor(s)	Definition(s)
I	Dsolv74	PPSA-3, atomic charge weighted partial positive surface area of the solvent molecule
	Rate	Rate of nitrogen blown onto sample (L/min)
	Temperature	Temperature at which the crystallisations occurred
II	D69	Molecular surface area of the CBZ molecule
II	Rate	Rate of nitrogen blown onto sample (L/min)
II	D68	Moment of inertia C of the CBZ molecule
III	D84	FNSA-1, fractional partial negative surface area of the CBZ molecule (PNSA-1/total molecular surface area)
	Boiling point	Literature value boiling point of the solvent molecule
	D77	PNSA-3, atomic charge weighted partial negative surface area of the CBZ molecule
	Rate	Rate of nitrogen blown onto sample (L/min)
	D68	Moment of inertia C of the CBZ molecule
Dihydrate	Dsolv74	PPSA-3, atomic charge weighted partial positive surface area of the solvent molecule
	Rate	Rate of nitrogen blown onto sample (L/min)
	Temperature	Temperature at which the crystallisations occurred
	Dsolv69	Molecular surface area of the solvent molecule
Solvate	Boiling point	Literature value boiling point of the solvent molecule
	Dsolv43	3D-Randić index (order 0) of the solvent molecule
Not in a Rule	E_ang	Angle bend potential energy of the CBZ molecule
Not in a Rule	Dsolv76	PNSA-2, total charge weighted partial negative surface area of the solvent molecule

It is important to remember that only a very limited amount of data was available for the form I, dihydrate and solvate predictions and because of this, firm conclusions cannot be made. This could be improved by training the ANN with more data that successfully crystallised as one of these forms.

## **7.4. Validation of the Optimised Set**

As discussed in chapter 5 it is important to validate the model produced, in order to determine whether the descriptors identified can lead to reasonable predictions of polymorphic form. The validation of this model was carried out in the same way as seen in section 5.5, by using a cross validation set, made up of 10 % of the experimental data, and the external validation method that uses the data of previously untested experimental work. This is comprised of two solvents that have not been used in the development of the model. It is worth referring back to the conclusions made in chapter 5 about this validation set. The two solvents used do not cover a wide range of descriptors values and are perhaps not the most effective choice as validation solvents.

### **7.4.1. Cross Validation Results**

10 % of the experimental results were predicted using the model created with the remaining experimental data. The average  $R^2$  value for INForm<sup>[7]</sup> was reduced from 88.62 % to 80.39 %. This reduction was expected as it lost a large proportion of its training data and therefore is unable to generalise as successfully. The results of the cross validation are summarised in Table 7.26 and show the model to be highly successful at predicting the major polymorphic form crystallised.



Table 7.26 Cross validation results summary

Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
					Form I	Form II	Form III	Dihydrate	Solvate
Ethanol	15	25	Form III	Form III	0.0	0.3	1.4	0.0	0.0
THF	25	25	Form II	Form II	0.0	0.9	0.2	0.2	0.0
Acetonitrile	15	50	Form III	Form III	0.0	0.2	1.1	0.0	0.0
DMSO	25	25	Solvate	Solvate	0.0	0.2	0.5	0.2	1.0
Aniline	5	50	Form III	Form III	0.0	0.0	1.1	0.0	0.0
Chlorobenzene	5	50	Form III	Form II	0.0	0.9	0.0	0.0	0.0
Toluene	15	75	Form II	Form II	0.0	0.9	0.2	0.0	0.0
Nitromethane	5	25	Form III	Form III	0.0	0.2	1.1	0.0	0.0
Chloroform	25	50	Form II / Form III	Form II	0.0	0.6	0.2	0.0	0.0

Overall the model makes predictions generally well, correctly identifying the major form 77.78 % of the time. In the two cases where the model has not predicted the major form correctly it is important examine the training data to identify possible causes. The predictions for the chlorobenzene and chloroform experiments are incorrect, for the same reasons already discussed in section 5.5.1. Both chloroform and chlorobenzene lead to crystallisation of form II and III as well as mixtures. These experimental results are reflected by the model's inability to accurately determine which polymorph is most likely to form under the experimental conditions analysed. It is encouraging to find that the model positively and correctly identifies crystallisation conditions leading to unpredictable outcomes. In an industrial setting this technique could be used to rule out such solvents, and identify solvents that consistently produce a desired form.

#### 7.4.2. External Validation Results

The same validation set as used in section 5.5.2, was applied to the model generated using PCA. The results of this analysis are presented in Table 7.27. The results clearly show the models inability to correctly predict the major form crystallised under these experimental conditions. Only three of the twelve experiments were predicted correctly. When the distribution of the descriptor values for the validation solvents were assessed (Figure 7.21), many of the descriptors for ethyl acetate (E) and n-butanol (B) values were very similar. It also showed that a number of the descriptors were outside of the range used in the training of the model.

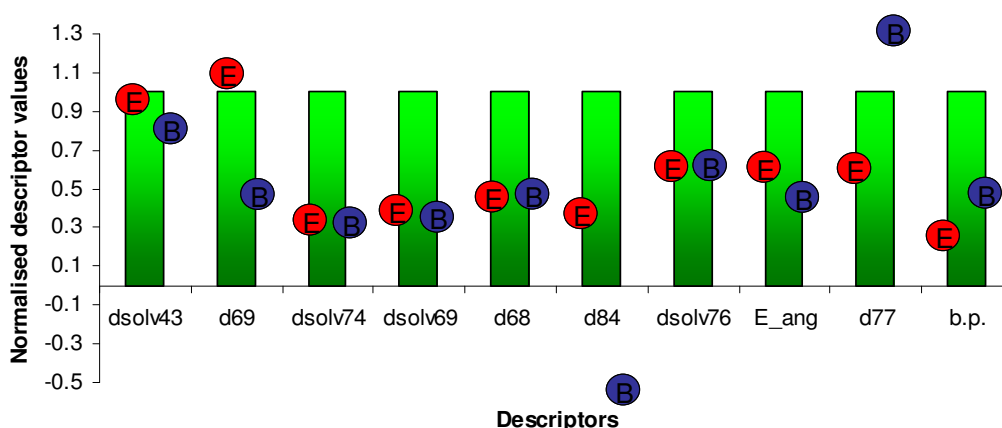


Figure 7.21 The distribution of the validation solvents descriptor values. E represents the ethyl acetate values and B the n-butanol values

This provides two possible explanations for the unsuccessful predictions. Firstly, because the descriptor values for both solvents are similar, they do not reflect the whole model and it is possible that they are in an area in which the model struggles to predict. This therefore makes the validation seem more unsuccessful than it would be if other solvents that represented a larger range of values were analysed. A second reason the predictions are unsuccessful may be due to the presence of descriptors outside of the range used in training the model. Ideally the model should be able to generalise sufficiently to make predictions outside of the range. However, the presence of these values in training may have altered the descriptor criteria used in making a prediction.

Table 7.27 External validation results summary

Experiment number	Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
						Form I	Form II	Form III	Dihydrate	Solvate
1	EtOAc	5	25	Form II	Form III	0.0	0.0	1.6	0.0	0.1
2	EtOAc	5	50	Form II	Form III	0.0	0.0	1.0	0.1	0.1
3	EtOAc	25	25	Form II	Form III	0.2	0.2	1.1	0.0	0.1
4	EtOAc	25	50	Form II	Form III	0.0	0.1	1.6	0.1	0.1
5	EtOAc	15	25	Form II	Form III	0.0	0.1	1.6	0.0	0.1
6	EtOAc	15	50	Form II	Form III	0.1	0.0	1.5	0.1	0.1
7	nBuOH	5	25	Form II	Form III	0.0	0.1	1.6	0.0	0.1
8	nBuOH	5	50	Form III	Form III	0.1	0.0	1.5	0.0	0.2
9	nBuOH	25	25	Form III	Form III	0.1	0.1	1.0	0.0	0.2
10	nBuOH	25	50	Form II	Form III	0.0	0.1	1.0	0.0	0.2
11	nBuOH	15	25	Form II	Form III	0.1	0.1	1.6	0.0	0.2
12	nBuOH	15	50	Form III	Form III	0.0	0.1	1.6	0.0	0.2

## 7.5. Conclusions

Using PCA to reduce the size of the dataset, an optimised set of descriptors for polymorphic form prediction has been created. The ten descriptors in the optimised set feature both CBZ and solvent molecule descriptors, which are dsolv43, dsolv69, dsolv74, dsolv76, d68, d69, d77, d84, E\_ang and boiling point.

In combination with the rates and temperatures used in the experimental work, this set of descriptors can successfully predict the major polymorphic form in 78 % of the cross validation experiments. When two previously untested solvents were used as a further test of the model, only three of the crystallised products were correctly predicted. This result was surprising and suggested that the model needed further optimisation. Perhaps combining the descriptors found in the manual analysis chapter with those in the optimised PCA set could lead to an improved predictive model.

- [1] M. Ringner, *Nature Biotechnology* **2008**, 26, 303.
- [2] B. G. Tabachnick, L. S. Fidell, *Using Multivariate Statistics*, 5th ed., Pearson Education, Inc., Boston, **2007**.
- [3] W. R. Dillon, M. Goldstein, *Multivariate Analysis: Methods and Applications*, 1st ed., John Wiley & Sons, Inc., New York, **1984**.
- [4] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [5] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors, Vol. 11*, first ed., Wiley-VCH, Weinheim, **2000**.
- [6] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [7] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [8] J. F. McCabe, *CrystEngComm* **2010**, 12, 1110.
- [9] T. Threllfall, *Organic Process Research & Development* **2003**, 7, 1017.
- [10] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, 95, 1918.
- [11] D. T. Stanton, S. Dimitrov, V. Grancharov, O. G. Mekenyan, *SAR and QSAR in Environmental Research* **2002**, 13, 341.
- [12] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, 113, 9480.
- [13] D. T. Stanton, P. C. Jurs, *Analytical Chemistry* **1990**, 62, 2323.
- [14] H. Golmohammadi, *Journal of Computational Chemistry* **2009**, 30, 2455.
- [15] M. H. Fatemi, F. Karimian, *Journal of Colloid and Interface science* **2007**, 314, 665.
- [16] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chemical Reviews* **1996**, 96, 1027.
- [17] J.-P. Liu, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, 55, 41.
- [18] <http://www.chemcomp.com/journal/descr.htm>, *The Chemical Computing Group Viewed on 29/04/11*.
- [19] J. Auer, J. Bajorath, *Journal of Chemical Information and Modeling* **2008**, 48, 1747–1753.
- [20] N. S. H. N. Moorthy, N. S. Cerqueira, M. J. Ramos, P. A. Fernandes, *Medical Chemistry Research* **2010**, DOI 10.1007/s00044.
- [21] H. Yuan, A. L. Parrill, *Journal of Molecular Structure (Theochem)* **2000**, 529, 273–282.
- [22] R. Davey, J. Garside, *From Molecules to Crystallizers: An Introduction to Crystallization*, Oxford University Press, Oxford, **2000**.
- [23] N. Rodriguez-Hornedo, D. Murphy, *Journal of Pharmaceutical Sciences* **1999**, 88, 651.
- [24] R. C. Schweitzer, J. B. Morris, *Analytica Chimica Acta* **1999**, 384, 285.
- [25] P. Atkins, J. de Paula, *Atkins' Physical Chemistry*, 7th ed., Oxford University Press, Oxford, **2002**.
- [26] G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, *Molecular Diversity* **2009**, 13, 301.
- [27] M. H. Fatemi, M. Ghorbanzade, *Molecular Diversity* **2009**, 13, 483.
- [28] E. R. Collantes, W. Tong, W. J. Welsh, *Analytical Chemistry* **1996**, 68, 2038.
- [29] R. H. Rohrbaugh, P. C. Jurs, *Analytical Chemistry* **1985**, 57, 2770.
- [30] T. P. Knowles, A. W. Fitzpatrick, S. Meehan, H. R. Mott, M. Vendruscolo, C. M. Dobson, M. E. Welland, *Science* **2007**, 318, 1900.

- [31] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Chemical Communications* **2007**, 1600.
- [32] F. P. A. Fabbiani, L. T. Byrne, J. J. McKinnon, M. A. Spackman, *CrystEngComm* **2007**, 9, 728.
- [33] D. T. Stanton, P. J. Madhav, L. J. Wilson, T. W. Morris, P. M. Hershberger, C. N. Parker, *Journal of Chemical Information and Computer Sciences* **2004**, 44, 221.
- [34] M. Cocchi, P. G. De Benedetti, R. Seeber, L. Tassi, A. Ulrich, *Journal of Chemical Information and Computer Sciences* **1999**, 39, 1190.
- [35] C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *Journal of Chemical Information and Modeling* **2005**, 45, 170.
- [36] A. A. Ivanova, A. A. Ivanov, A. A. Oliferenko, V. A. Palyulin, N. S. Zefirov, *SAR and QSAR in Environmental Research* **2005**, 16, 231.
- [37] J.-P. Luiu, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, 55, 41.
- [38] R. K. Harris, P. Y. Ghi, H. Puschmann, D. C. Apperley, U. J. Griesser, R. B. Hammond, C. Ma, K. J. Roberts, G. J. Pearce, J. R. Yates, C. J. Pickard, *Organic Process Research and Development* **2005**, 9, 902.
- [39] A. Johnston, A. J. Florence, A. R. Kennedy, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1509.
- [40] A. Johnston, B. F. Johnston, A. R. Kennedy, A. J. Florence, *CrystEngComm* **2008**, 10, 23.
- [41] S. Lohani, Y. Zhang, L. J. Chyall, P. Mougin-Andres, F. X. Muller, D. J. W. Grant, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1310.
- [42] S. G. Fleischman, S. S. Kuduva, J. A. McMahon, B. Moulton, R. D. Bailey Walsh, N. Rodriguez-Hornedo, M. J. Zaworotko, *Crystal Growth and Design* **2003**, 3, 909.
- [43] A. J. Cruz Cabeza, G. M. Day, W. D. Samuel Motherwell, W. Jones, *Journal of the American Chemical Society* **2006**, 128, 14466.
- [44] M. M. J. Lowes, M. R. Caira, A. P. Lotter, J. G. Van Der Watt, *Journal of Pharmaceutical Sciences* **1987**, 76, 744.
- [45] M. Randic, *Journal of Chemical Information and Computer Sciences* **1984**, 24, 164.
- [46] H. Liu, M. Lu, F. Tian, *Journal of Mathematical Chemistry* **2005**, 38, 345.
- [47] G. L. Amidon, S. T. Anik, *Journal of Pharmaceutical Sciences* **1976**, 65, 801.
- [48] M. Randic, G. M. Brissey, R. B. Spencer, C. L. Wilkins, *Computers & Chemistry* **1980**, 4, 27.
- [49] M. Randic, B. Jerman-Blazic, N. Trinajstic, *Computers & Chemistry* **1990**, 14, 237.
- [50] E. Estrada, *Journal of Physical Chemistry A* **2002**, 106, 9085.
- [51] O. Ivanciuc, T. Ivanciuc, P. A. Filip, D. Cabrol-Bass, *Journal of Chemical Information and Computer Sciences* **1999**, 39, 515.
- [52] Y. Ren, H. Liu, X. Yao, M. Liu, *Journal of Chromatography A* **2007**, 1155, 105.

## **8. FINAL OPTIMISATION AND DISCUSSION OF RESULTS**

The previous chapters have discussed different strategies to reduce the number of descriptors used in artificial neural networks (ANN) for the prediction of polymorphic forms from solution properties. The most successful sets of descriptors from the previous chapters will now be further analysed in order to determine the overall most successful set. This chapter will then conclude by presenting the final optimisation of the descriptor set alongside a discussion of the possible physical meanings of the descriptors present.

### **8.1. Introduction**

Three methods of descriptor reduction have so far been discussed; they included (i) linear correlations coefficients and manually reducing the descriptor set (chapter 5), (ii) partial least squares (PLS) analysis (chapter 6) and (iii) principal component analysis (PCA) (chapter 7). All three methods generated an optimised set of descriptors for the prediction of polymorphic form. The PLS method generated an optimised set that built a much less successful model than the other two methods. For this reason, the PLS optimised set will not be further considered in this chapter.

The linear correlation (Corr. best Set) and PCA (PCA best Set) analyses both generated a set of descriptors that allowed an ANN to predict the polymorphic form of a crystallisation experiment correctly with high probability. The descriptor sets and the resulting FormRules<sup>[6]</sup> and INForm<sup>[7]</sup>  $R^2$  values are presented in Table 8.1.



Table 8.1 The top sets of descriptors found in the manual and PCA analysis

	Corr. best Set	PCA best Set
	Dsolv57	Dsolv43
	Dsolv65	Dsolv69
	Dsolv71	Dsolv74
	Dsolv78	Dsolv76
	MNDO_dipole	D68
	E_vdw	D69
	Gutmann donor number	D77
		D84
		E_ang
		Boiling point
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.12</b>	<b>79.99</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>87.96</b>	<b>88.62</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>84.04</b>	<b>84.31</b>

It can be seen in Table 8.1 that although there are no overlapping descriptors between these two optimised sets, the overall average  $R^2$  values produced are very similar.

As in previous chapters (5 and 7) the linear correlations between the descriptors were calculated. They are presented in Table 8.2. The PCA best set has two correlated descriptors, dsolv43 and d69, which suggested that further optimisation could perhaps be performed. There are also two descriptors that are correlated across the optimised sets: these are dsolv43 and dsolv74 from the PCA best set, which correlate strongly with dsolv71 and dsolv65 from the Corr. best set respectively. All of these descriptors are solvent molecule descriptors and describe connectivity and surface areas (see appendix section 12.2 for details).

Table 8.2 Linear correlations between the two most successful sets. Number in brackets is the correlation coefficient

PCA best set	Correlations within the PCA best set ( $\pm 0.8 - 1$ )	Correlations with Corr. best set ( $\pm 0.8 - 1$ )
Dsolv43	D69	Dsolv71 (0.93)
Dsolv69	-	-
Dsolv74	-	Dsolv65 (0.96)
Dsolv76	-	-
D68	-	-
D69	Dsolv43	-
D77	-	-
D84	-	-
E_ang	-	-
Boiling point	-	-
	2 of the PCA best set descriptors are correlated	2 of the Corr. best set are correlated with the PCA best set

## 8.2. Final Optimisation of the Descriptor Sets

### 8.2.1. Optimisation Based upon Correlation Results

The correlation analysis summarised in Table 8.2 indicated that two descriptors in the PCA best set were highly correlated. The effect of removing either of these descriptors was examined, with results presented in Table 8. 3 (Opt. A and B). The correlations between the PCA and Corr. best sets were also examined, with the correlated descriptors being substituted for the corresponding descriptor in the other set (Table 8.3, Opt. C - H).

Table 8.3 Optimisation of the PCA and Corr. best sets based upon the linear correlation results. X denotes the presence of the descriptor in the set

Descriptors	PCA best set	Corr. best set	Opt. A	Opt. B	Opt. C	Opt. D	Opt. E	Opt. F	Opt. G	Opt. H
Dsolv43	X			X		X	X			X
Dsolv69	X		X	X	X		X		X	
Dsolv74	X		X	X	X			X		X
Dsolv76	X		X	X	X		X		X	
D68	X		X	X	X		X		X	
D69	X		X		X		X		X	
D77	X		X	X	X		X		X	
D84	X		X	X	X		X		X	
E_ang	X		X	X	X		X		X	
Boiling point	X		X	X	X		X		X	
Dsolv57		X				X		X	X	
Dsolv65		X				X	X			X
Dsolv71		X			X			X	X	
Dsolv78		X				X		X		X
MNDO_dipole		X				X		X		X
E_vdw		X				X		X		X
Gutmann donor number		X				X		X		X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>79.99</b>	<b>80.12</b>	<b>79.98</b>	<b>80.32</b>	<b>79.99</b>	<b>79.82</b>	<b>80.20</b>	<b>79.88</b>	<b>80.46</b>	<b>79.88</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>88.62</b>	<b>87.96</b>	<b>58.27</b>	<b>83.20</b>	<b>88.51</b>	<b>74.87</b>	<b>89.30</b>	<b>70.52</b>	<b>83.14</b>	<b>72.27</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>84.31</b>	<b>84.04</b>	<b>69.13</b>	<b>81.76</b>	<b>84.25</b>	<b>77.35</b>	<b>84.75</b>	<b>75.20</b>	<b>81.80</b>	<b>76.08</b>

The results in Table 8.3 show that a slightly improved model was built (Opt. E) when dsolv74 from the PCA best set was exchanged for dsolv65. The results also show that the presence of either dsolv43 or the correlated dsolv71 is important for the overall result generated: Opt. A clearly shows a reduced overall average R<sup>2</sup> value of 69.13 % when neither of these two descriptor was present.

When the rules generated in FormRules<sup>[6]</sup> for Opt. E are examined, only seven out of the ten descriptors were featured on the list of significant descriptors (appendix section 12.10). The effect of omitting the three remaining descriptors (dsolv69, dsolv76 and E\_ang) has been determined and the result presented in Table 8.4 under the heading ‘Opt. E Rule only’.

Table 8.4 Further optimisation of the Best sets. X denotes the presence of the descriptor in the set

Descriptors	Opt. E	Opt. E Rule only
Dsolv43	X	X
Dsolv69	X	
Dsolv76	X	
D68	X	X
D69	X	X
D77	X	X
D84	X	X
E_ang	X	
Boiling point	X	X
Dsolv65	X	X
<b>FormRules average <math>R^2</math> (%)</b>	<b>80.20</b>	<b>80.20</b>
<b>INForm average <math>R^2</math> (%)</b>	<b>89.30</b>	<b>63.87</b>
<b>Overall average <math>R^2</math> (%)</b>	<b>84.75</b>	<b>72.04</b>

It can be seen that removing the descriptors that do not feature in the rules reduces the performance of the model from an overall average  $R^2$  value of 84.75 % to one of 72.04 %. This suggests that these descriptors are subtly but significantly important in the model built by INForm.<sup>[7]</sup>

### 8.2.2. Optimisation Based upon Best Set Analysis

Due to the success of both the PCA and Corr. best sets, a further attempt to create an optimised set was made including all 17 descriptors that featured in the two best sets. They were all run in one ANN ('All Desc.'), and then the rule-only descriptors were examined (Table 8.5).

Table 8.5 Optimisation of the descriptor set. X denotes the presence of the descriptor in the set

Descriptors	PCA best set	Corr. best set	Opt. E	All Desc.	All Desc. Rule only
Dsolv43	X		X	X	
Dsolv69	X		X	X	
Dsolv74	X			X	
Dsolv76	X		X	X	X
D68	X		X	X	X
D69	X		X	X	
D77	X		X	X	
D84	X		X	X	X
E_ang	X		X	X	
Boiling point	X		X	X	
Dsolv57		X		X	X
Dsolv65		X	X	X	X
Dsolv71		X		X	X
Dsolv78		X		X	X
MNDO_dipole		X		X	X
E_vdw		X		X	X
Gutmann donor number		X		X	X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>79.99</b>	<b>80.12</b>	<b>80.20</b>	<b>81.85</b>	<b>81.85</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>88.62</b>	<b>87.96</b>	<b>89.30</b>	<b>72.76</b>	<b>81.22</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>84.31</b>	<b>84.04</b>	<b>84.75</b>	<b>77.31</b>	<b>81.54</b>

The results presented in Table 8.5 demonstrate that using a large number of descriptors does not necessarily lead to a better model. When only the rule descriptors identified by FormRules<sup>[6]</sup> are used, the model generated an improved performance compared to using all of the descriptors. It might be that information introduced by some of the less useful descriptors masks the important information required for successful prediction. Overall, none of the generated models improved on the previously determined most successful sets of descriptors.

### 8.2.3. Optimisation Based upon Descriptor Types

A number of different sets were created based upon the PCA and Corr. best sets presented in section 8.1, by taking the physical meanings of the descriptors as a guide to selection. Initially, only the solvent descriptors were used. The rationale for this choice was to create a transferable descriptor set that may be able to predict the

polymorphic outcome for different target molecules. The results obtained with this set are presented in Table 8.6 (Solvent only).

When the results in Table 8.6 are examined, there is clearly a reduction in the overall average  $R^2$  value when only the solvent descriptors are used. However, the value is still reasonably high and may produce a useful starting point in other target molecule predictions. It is interesting that all the rule-only descriptors are from the Corr. best set. To examine the impact upon prediction, the CBZ descriptors from the PCA best set have been added to these rule only descriptors. The results of this (Opt. I) show that the improvement in prediction is only small and therefore the solvent descriptors are important in the prediction.

Table 8.6 Determination of the effect of removing the target molecules descriptors. X denotes the presence of the descriptor in the set

Descriptors	PCA best set	Corr. best set	Opt. E	Solvent only	Solvent only rule only	Opt. I
Dsolv43	X		X	X		
Dsolv69	X		X	X		
Dsolv74	X			X		
Dsolv76	X		X	X		
D68	X		X			X
D69	X		X			X
D77	X		X			X
D84	X		X			X
E_ang	X		X			X
Boiling point	X		X	X	X	X
Dsolv57		X		X	X	X
Dsolv65		X	X	X	X	X
Dsolv71		X		X	X	X
Dsolv78		X		X	X	X
MNDO_dipole		X				
E_vdw		X				
Gutmann donor number		X		X	X	X
<b>FormRules average <math>R^2</math> (%)</b>	<b>79.99</b>	<b>80.12</b>	<b>80.20</b>	<b>79.79</b>	<b>79.79</b>	<b>79.91</b>
<b>INForm average <math>R^2</math> (%)</b>	<b>88.62</b>	<b>87.96</b>	<b>89.30</b>	<b>69.53</b>	<b>75.09</b>	<b>75.49</b>
<b>Overall average <math>R^2</math> (%)</b>	<b>84.31</b>	<b>84.04</b>	<b>84.75</b>	<b>74.66</b>	<b>77.44</b>	<b>77.70</b>

A second optimisation method was examined, which was based upon the class of physical property (surface area, partial charge, molecular interaction, geometry, connectivity/branching) represented by the descriptors identified by the the PCA and Corr. best sets. Figure 8.1, Figure 8.2 and Table 8.7 show the classification of the 17 descriptors.

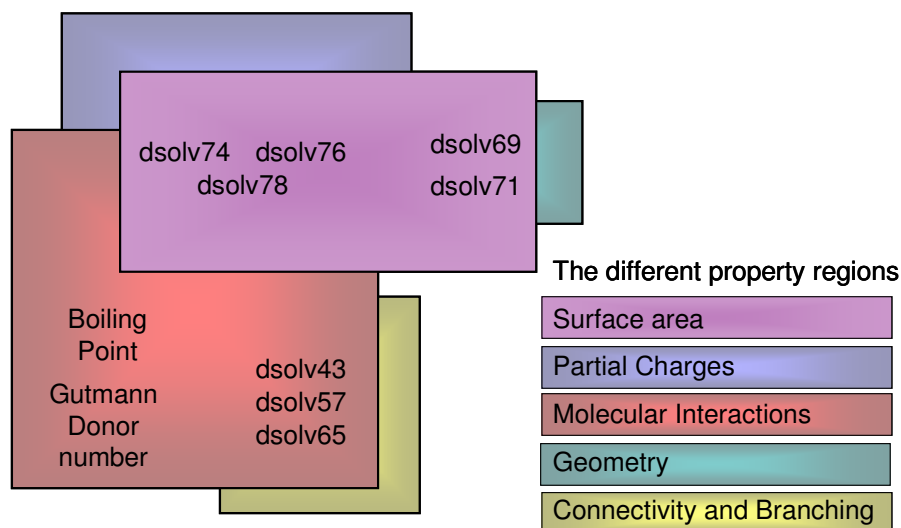


Figure 8.1 Schematic of the different property regions found for the solvent descriptors

Figure 8.1 highlights that there are four groups of solvent descriptor present in the PCA and Corr. best sets. A similar classification analysis was also carried out for the CBZ descriptors, presented in Figure 8.2.

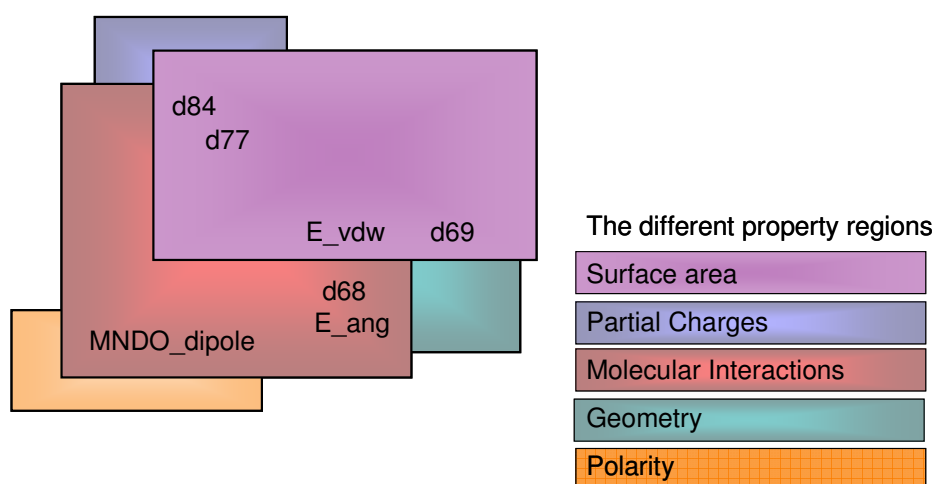


Figure 8.2 Schematic of the different property regions found for the CBZ descriptors

Figure 8.2 shows that a number of the CBZ descriptors are in the same class of physical property, but these groups are not as clearly distinguishable as those seen in the solvent descriptors. Table 8.7 summaries the classification of the descriptors.

Table 8.7 Descriptors grouped based upon their physical meaning

Physical Property	Descriptor
Solvent Descriptors: Molecular Surface Area	Dsolv69, dsolv71, Dsolv74, dsolv76, dsolv78
Solvent Descriptors: Partial Charges	Dsolv74, dsolv76, dsolv78
Solvent Descriptors: Molecular Interactions	Dsolv43, dsolv57, dsolv65, Dsolv74, dsolv76, dsolv78, Boiling point, Gutmann donor number
Solvent Descriptors: Geometry	Dsolv69, dsolv71
Solvent Descriptors: Connectivity and Branching	Dsolv43, dsolv57, dsolv65,
CBZ Descriptors: Molecular Surface Area	D69, d77, d84, E_vdw
CBZ Descriptors: Partial Charges	D77, d84
CBZ Descriptors: Molecular Interactions	D68, d77, d84, E_vdw, E_ang, MNDO_dipole
CBZ Descriptors: Geometry	D68, d69, E_vdw, E_ang
CBZ Descriptors: Polarity	MNDO_dipole

Based upon the physical properties represented by the descriptors, further optimisations were carried out to improve the set (Table 8.8). By selecting only one descriptor from each category, new sets were created and results generated in INForm<sup>[7]</sup> and FormRules.<sup>[6]</sup> If more than one descriptor was present in each category, each descriptor was tested individually, with the most successful being carried into the next set (Opt. J-R).



Table 8.8 Optimisation results based upon the physical meanings of the descriptors. X denotes the presence of the descriptor in the set

Descriptors	Opt. J	Opt. K	Opt. L	Opt. M	Opt. N	Opt. O	Opt. P	Opt. Q	Opt. R
Boiling point	X								
Gutmann donor number		X	X	X	X	X	X	X	X
Dsolv74	X	X			X	X	X	X	X
Dsolv76			X						
Dsolv78				X					
Dsolv43	X	X	X	X					
Dsolv57					X		X	X	X
Dsolv65						X			
Dsolv69	X	X	X	X	X	X			
Dsolv71							X	X	X
MNDO_dipole	X	X	X	X	X	X	X	X	X
D84	X	X	X	X	X	X	X		X
D77								X	
D69	X	X	X	X	X	X	X	X	X
E_vdw	X	X	X	X	X	X	X	X	X
D68	X	X	X	X	X	X	X	X	
E_ang									X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>79.23</b>	<b>80.71</b>	<b>54.90</b>	<b>54.29</b>	<b>80.75</b>	<b>80.80</b>	<b>80.13</b>	<b>78.58</b>	<b>80.09</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>73.82</b>	<b>74.35</b>	<b>69.36</b>	<b>88.45</b>	<b>80.94</b>	<b>78.40</b>	<b>82.99</b>	<b>73.60</b>	<b>79.14</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>76.53</b>	<b>77.53</b>	<b>62.13</b>	<b>71.37</b>	<b>80.85</b>	<b>79.60</b>	<b>81.56</b>	<b>76.09</b>	<b>79.62</b>

The results in Table 8.8 do not identify any set of descriptors that generate ANNs more successful than the PCA or best Set ANNs. In general, the sets produce high overall average R<sup>2</sup> values, with the exception of Opt. L. Opt. L contained dsolv76 instead of dsolv74, which appears to significantly affect the prediction. From the PCA and Corr. best sets presented in this chapter, dsolv74 features in the rules for form I and dihydrate prediction, which is where this set of descriptors fails. It appears necessary to include dsolv74 to generate a reasonable prediction for form I and the dihydrate.

In a way similar to the physical meaning optimisation just described, the descriptors were then grouped according to their association with predicted crystallisation outcomes, as indicated by FormRules<sup>[6]</sup>. Figure 8.3 shows a schematic of the descriptors from the two best sets and how they influence the predictions of different polymorphic forms. It can be seen that many of the descriptors contribute to the prediction of several forms.

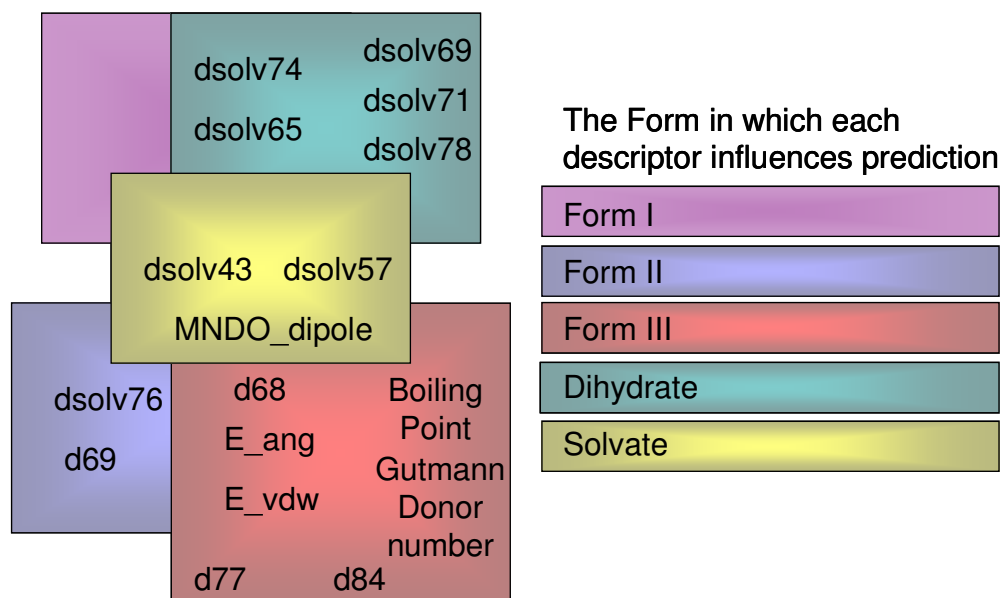


Figure 8.3 Schematic of the forms in which each descriptor influences prediction

Again, this result could be interpreted as highlighting a potential opportunity for reducing the number of descriptors in the set further. Therefore an optimisation was carried out using the groups of descriptors displayed in Figure 8.3. The results are presented in Table 8.9.

Table 8.9 Optimisation results based upon predictions made by each descriptor. X denotes the presence of the descriptor in the set

Descriptors	Opt. S	Opt. T	Opt. U	Opt. V	Opt. W	Opt. X	Opt. Y
Boiling point					X		
Gutmann donor number		X	X	X			
Dsolv74	X	X	X	X	X	X	X
Dsolv76				X	X	X	X
Dsolv78							
Dsolv43							
Dsolv57		X	X	X	X	X	X
Dsolv65							
Dsolv69							
Dsolv71		X	X	X	X	X	X
MNDO_dipole	X		X	X	X	X	X
D84							X
D77						X	
D69		X	X				
E_vdw							
D68							
E_ang							
<b>FormRules average R<sup>2</sup> (%)</b>	<b>68.49</b>	<b>68.84</b>	<b>77.81</b>	<b>77.54</b>	<b>76.92</b>	<b>74.39</b>	<b>76.25</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>68.72</b>	<b>77.69</b>	<b>78.22</b>	<b>78.98</b>	<b>71.37</b>	<b>39.49</b>	<b>80.20</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>68.61</b>	<b>73.27</b>	<b>78.02</b>	<b>78.26</b>	<b>74.15</b>	<b>56.94</b>	<b>78.23</b>

It can be seen that no improvement on the most successful set of descriptors (Opt. E with an overall average R<sup>2</sup> value of 84.75 %) was made using this technique. These results suggest that in order to generate a successful predictive model, multiple descriptors need to be used in tandem. By using only MNDO\_dipole and dsolv74, which between them have featured in the rules for all of the forms, there is not enough information to build a successful model. This once again indicates that using only descriptors presented in the rules determined by FormRules<sup>[6]</sup> represents an oversimplification. Clearly the descriptors are strongly intercorrelated and work together to create a successful ANN model. The FormRules<sup>[6]</sup> analysis may be used as a guideline to understanding what is occurring in the solutions, but the more complex intercorrelations in the ANNs are necessary for polymorph prediction.

#### 8.2.4. Optimisation Based upon Validation Results

The external validation of the PCA best set (section 7.4.2) indicated that three of the descriptors (d69, d77 and d84) were outside of the range of the descriptor values used within the training of the model. Further analysis was carried out to determine the impact of removing these descriptors from this set (Opt. E1 shown in Table 8.10). The result of this analysis was a lower overall average  $R^2$  value of 82.34%, which although lower, is still a respectably high value.

Further analysis was conducted to replace these descriptors without compromising the high average  $R^2$  value generated from Opt. E. The Opt. E2-5 sets (Table 8.10) were created based upon descriptors that featured in the Corr. best set, and were seen to influence the prediction of forms II and III. MNDO\_dipole, E\_vdw and Gutmann donor number all matched with these criteria. However, no improvements were seen in the results.

Therefore, a final strategy was devised. The linear correlations with other descriptors of d69, d77 and d84 were calculated and each was replaced in turn by a highly correlated descriptor (Opt. E6-8 in Table 8.10). D69 was replaced with dsolv6, which is the number of rings in the solvent molecule. The correlation coefficient was -0.90, which was the highest correlation determined. D84 was replaced by d81, which is the fractional partial positive surface area of the CBZ molecule. The correlation coefficient determined between these two descriptors was -1. D77 was not highly correlated with any other descriptor in this analysis and was therefore just removed from the set for testing. Descriptor details can be found in appendix section 12.2.

Table 8.10 Further optimisation of the descriptor set. X denotes the presence of the descriptor in the set

Descriptors	Opt. E	Opt. E1	Opt. E2	Opt. E3	Opt. E4	Opt. E5	Opt. E6	Opt. E7	Opt. E8
Dsolv43	X	X	X	X	X	X	X	X	X
Dsolv69	X	X	X	X	X	X	X	X	X
Dsolv76	X	X	X	X	X	X	X	X	X
D68	X	X	X	X	X	X	X	X	X
D69	X							X	X
D77	X						X	X	
D84	X						X		X
E_ang	X	X	X	X	X	X	X	X	X
Boiling point	X	X	X	X	X	X	X	X	X
Dsolv65	X	X	X	X	X	X	X	X	X
MNDO_dipole			X			X			
E_vdw				X	X	X			
Gutmann donor number					X	X			
Dsolv6							X		
D81								X	
<b>FormRules average R<sup>2</sup> (%)</b>	<b>80.20</b>	<b>79.19</b>	<b>79.16</b>	<b>79.60</b>	<b>79.60</b>	<b>79.96</b>	<b>80.66</b>	<b>80.19</b>	<b>80.22</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>89.30</b>	<b>85.48</b>	<b>73.28</b>	<b>85.38</b>	<b>70.35</b>	<b>72.94</b>	<b>85.63</b>	<b>86.60</b>	<b>73.47</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>84.75</b>	<b>82.34</b>	<b>76.22</b>	<b>82.49</b>	<b>74.98</b>	<b>76.45</b>	<b>83.15</b>	<b>83.40</b>	<b>76.85</b>

The results presented in Table 8.10 show that, as in all previous attempts to increase the predictive power of the ANNs, no improved models were generated by removal of the out-of-range descriptors.

### 8.2.5. Conclusion of the Final Optimisation Work

Across all of the final optimisation attempts, a slight and perhaps insignificant improvement could be achieved over the PCA best set results, when dsolv74 was replaced with dsolv65. (section 8.2.1).

The best descriptor set identified by the work presented here thus contains dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and the boiling point of the solvents. It generated an overall average R<sup>2</sup> value of 84.75 %.

### 8.3. Discussion of the Descriptors in the Final Set

Classifying the final set of descriptors according to their appearance in the rules for each form as determined by FormRules<sup>[6]</sup>, the ten descriptors in the final set are again presented in Table 8.11.

As expected, the rules generated for this final set have essentially remained the same as previously for the PCA best set, with dsolv74 being replaced directly with dsolv65. However, there has been one unexpected rule change in the dihydrate prediction, shown in Table 8.11. In the PCA best set rules, dsolv69 was used to predict the dihydrate. However, with the inclusion of dsolv65, the rule has now changed to replace dsolv69 with dsolv43. This suggests that dsolv65 and dsolv43 work better in tandem than dsolv65 and dsolv69 would. This was an unexpected change, which led to the removal of dsolv69 from the rules generated in FormRules<sup>[6]</sup>.

Table 8.11 Summary of the descriptors involved in the final set

Form predicted	Descriptor(s)	Definition(s)
I	Dsolv65 Rate Temperature	3D bonding information content (order 2) of the solvent molecule Rate of nitrogen blown onto sample (L/min) Temperature at which the crystallisations occurred
II	D69	Molecular surface area of the CBZ molecule
II	Rate	Rate of nitrogen blown onto sample (L/min)
II	D68	Moment of inertia C of the CBZ molecule
III	D84	FNSA-1, fractional partial negative surface area of the CBZ molecule (PNSA-1/total molecular surface area)
III	Boiling point D77	Literature value boiling point of the solvent molecule PNSA-3, atomic charge weighted partial negative surface area of the CBZ molecule
III	Rate	Rate of nitrogen blown onto sample (L/min)
III	D68	Moment of inertia C of the CBZ molecule
Dihydrate	Dsolv65 Rate Temperature Dsolv43	3D bonding information content (order 2) of the solvent molecule Rate of nitrogen blown onto sample (L/min) Temperature at which the crystallisations occurred 3D-Randić index (order 0) of the solvent molecule
Solvate	Boiling point Dsolv43	Literature value boiling point of the solvent molecule 3D-Randić index (order 0) of the solvent molecule
Not in a Rule	E_ang	Angle bend potential energy of the CBZ molecule
Not in a Rule	Dsolv76	PNSA-2, total charge weighted partial negative surface area of the solvent molecule
Not in a Rule	Dsolv69	Molecular surface area of the solvent molecule

In section 8.2.1, the rule only descriptor analysis of Opt. E (Table 8.4) found that when dsolv69, E\_ang and dsolv76 were not present, the predictive capabilities of the model were reduced. As was carried out in the analysis of PCA best set descriptors (section 7.3) detailed analysis has been conducted in order to assess the influence on the prediction of dsolv69, which will be presented in the relevant section. E\_ang and dsolv76 were found to have an impact upon forms II and III prediction (section 7.3). As the descriptor set has not changed dramatically from the PCA best set, the rules will be presented with only a brief discussion of their physical meaning in relation to nucleation and crystallisation. Full details are presented in sections 7.3.2 and 7.3.3.

### 8.3.1. The Prediction of Form I

As stated in chapters 5 and 7, the amount of data available in order to train a form I model was very small. This may have an impact upon the reliability of any model created. However, due to the presence of form I data, a model was built. The rules in Table 8.12 show that rate, temperature and dsolv65, lead to a form I prediction.

Table 8.12 Rules generated in FormRules for form I prediction

Rules generated for Form I prediction		
IF dsolv65 is LOW AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is LOW AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is LOW AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is LOW AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is LOW AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is LOW AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is MID AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is MID AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (0.90)
IF dsolv65 is MID AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is MID AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is MID AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is MID AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)
IF dsolv65 is HIGH AND rate is HIGH AND Temp is HIGH	THEN Form I is	HIGH (1.00)

The rules show that in most cases the prediction is for a low probability that form I will be observed. This reflects the training data presented to the ANN accurately, as there were no high form I yields observed under any of the experimental conditions. Nevertheless, the rules have combined to formulate one condition for high form I prediction. This rule states that when rate, temperature and dsolv65 are at their highest value, a high form I prediction will occur.

Examination of the normalised descriptor values plotted against the form I producing experiments (Figure 8.4), shows that on no occasion were conditions chosen for which the values were high for all three descriptors. This is in accordance with the levels of form I generated in the crystallisations, with neither crystallisation in chlorobenzene or in methanol generated a high yield of form I.

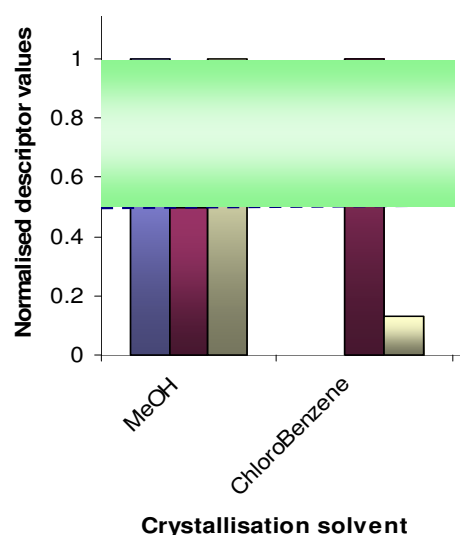


Figure 8.4 Crystallisation solvents in which form I is produced plotted against the three rule descriptors, rate (blue), temperature (purple) and dsolv65 (cream). The green shaded area highlights the most favourable descriptor values for form I production.

The physical influence of dsolv65 was already presented in section 5.1.4, and is the 3D bonding information content (BIC) for the solvent molecule. This topological descriptor may be used as a measure of structural diversity<sup>[53]</sup>, incorporating the branching and connectivity of the molecule. This descriptor has been featured in other research<sup>[53-55]</sup> and has been reported to represent the solution interactions of the solvent.

It appears therefore reasonable to assume that dsolv65 also plays a pivotal role in polymorphic crystallisation from solution. The literature provides many examples of solvents and additives inhibiting or promoting the specific nucleation of a



polymorphic form.<sup>[56-59]</sup> Perhaps the presence of this descriptor confirms the importance of how the solvent molecules interact in solution. More precisely, different levels of branching in the solvent appear to promote and inhibit interactions with the solute molecules.

The rules in Table 8.12 also feature rate and temperature. These are the conditions at which the crystallisation took place. The rule states that high values for both these conditions would lead to the formation of form I. In line with this, previous research has shown that carrying out crystallisations at high evaporation rates and temperatures often achieves metastable polymorphic forms.<sup>[8-10]</sup>

### 8.3.2. The Prediction of Form II

No changes in the rules that were presented in the PCA best set in section 7.3 were observed for the prediction of form II by this optimised set. The results presented in Table 8.13, generated by FormRules<sup>[6]</sup> contain rate, d68 and d69.

Table 8.13 Rules generated in FormRules for form II prediction

Rules generated for Form II prediction			
SubModel:1	IF d69 is LOW	THEN Form II is	HIGH (1.00)
	IF d69 is HIGH	THEN Form II is	LOW (1.00)
SubModel:2	IF rate is LOW	THEN Form II is	LOW (1.00)
	IF rate is HIGH	THEN Form II is	HIGH (0.64)
SubModel:3	IF d68 is LOW	THEN Form II is	HIGH (1.00)
	IF d68 is HIGH	THEN Form II is	LOW (1.00)

The physical meaning of d69 was already discussed in section 7.3.2; it represents the molecular surface area of the CBZ molecule in the solvent force field. It is a geometrical descriptor that uses the van der Waals radii of the atoms within the molecule to give the best surface area approximation.<sup>[5, 12]</sup> The difference between each CBZ surface area is only slight and was brought about by the solute molecules interactions within the solvent force field. The differences due to variations in the solvent force field are subtle, perhaps indicating how slight geometrical differences affect interactions with other solute or solvent molecules in solution significantly enough to modify aggregation behaviour. Previous research has indeed found that geometrical descriptors are very useful for predicting physicochemical properties of molecules.<sup>[12, 17, 37]</sup>

As presented in sections 7.3.2 and 7.3.3, E\_ang was found to have an impact upon prediction when detailed analysis was undertaken for the rule containing d69. E\_ang, which is the angle bend potential energy for the CBZ molecule in the solvent forcefield<sup>[18]</sup>, describes the flexibility and geometry of the molecule. In a similar way to the d69 descriptor, the differences between the values are subtle due to the slight changes in geometry of the CBZ in the different solvent force fields. These slight geometrical differences nevertheless appear to offer insight into significant interactions in solution. E\_ang has not commonly been used in physicochemical studies in the literature, but it has been discussed in the context of a number of biological systems.<sup>[19-21]</sup>

When the detailed analysis was carried out on this rule (section 4.6.2) dsolv69 was also found to have an impact upon the predictions of form II. Dsolv69 is the molecular surface area of the solvent model, which is calculated in the same way as mentioned in the d69 discussion in this section. When d69, which is featured in the rules, is plotted against rate, the model matches the submodel 1 rule in Table 8.13 very well. Figure 8.5 shows the effect that changing the dsolv69 values has on the prediction of form II. Although the rule is not affected, the level of form II prediction is increased at higher dsolv69 values.

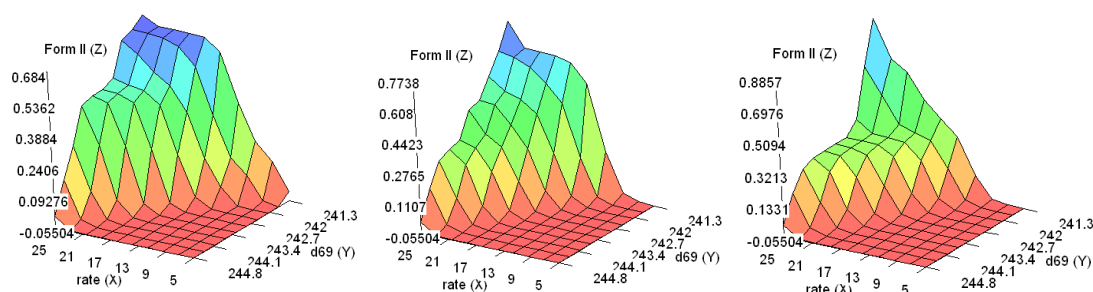


Figure 8.5 Effect of dsolv69 upon the form II predictions based upon the rules generated. Low dsolv69 values (left), mid range values (centre) and high dsolv69 values (right)

The rule in Table 8.13 that presents rate as an important property agrees with what was presented in the form I rules (Table 8.12). Form II is the least stable polymorphic form of CBZ, which coincides well with previous observations that high evaporation rates often lead to the formation of a metastable form.<sup>[8-10]</sup> High evaporation rates achieve high levels of supersaturation quickly, thereby enhancing the nucleation rate in solution. If Ostwald's Rule of Stages is followed, then the least stable polymorph, in this case CBZ form II, would crystallise first.<sup>[22, 23]</sup>

The third rule presented in Table 8.13 uses d68, which is the moment of inertia C for the CBZ molecule, to predict form II. D68 is classed as a geometrical descriptor<sup>[24]</sup> and is calculated using the mass and three-dimensional coordinates of atoms in the molecule. Moment of inertias can be used as a measure of how rotationally active parts of the molecule are,<sup>[4, 26]</sup> and as a measure of molecular size and weight.<sup>[29]</sup> This descriptor is commonly used in the physicochemical literature,<sup>[26-30]</sup> for solution systems particularly in the context of chromatographic retention time prediction, where this descriptor has generated information about molecular interactions.

When moment of inertia is considered with respect to nucleation and crystallisation, the different values likely represent an altered capacity to interact with either the solvent or other solute molecule. With the solution interactions being altered, perhaps certain forms can be inhibited or promoted. It is known that solvent is very important in form II crystallisation,<sup>[31, 32]</sup> so perhaps the moment of inertia values are offering an indirect measure of molecular interactions.

Section 7.3.2 shows the detailed analysis of this third rule with regard to the effect of both E\_ang and dsolv76 upon prediction. E\_ang has already been presented in this section, but dsolv76, which is the total charge weighted partial negative surface area (PNSA-2) of the solvent molecule, has not. This charged partial surface area descriptor provides information about the negatively charged regions on the solvent molecule, thereby once again generating potentially useful information about the interactions between solvent and solute molecules.

By using the detailed analysis method for submodel 3, it was found that high values of dsolv69 had an impact upon the prediction of form II. Figure 8.6 shows the plot of d68 and temperature when dsolv69 values are low and high.

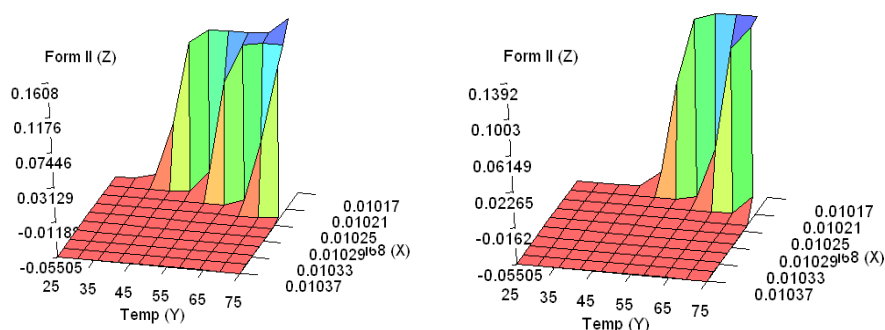


Figure 8.6 Effect of dsolv69 upon the form II predictions based upon the rules generated. Low dsolv69 values (left) and high dsolv69 values (right)

This analysis highlights once again how the rules generated in FormRules<sup>[6]</sup> can be used as a guide to what descriptor values lead to successful polymorphic predictions. However, in order to generate a more accurate picture, other descriptors that are not featured in the rules may have an influence, and can be observed from the ANN results in INForm.<sup>[7]</sup>

### 8.3.3. The Prediction of Form III

Form III is the most stable form of CBZ and as such was crystallised most frequently in this research. Table 8.14 shows the four rules generated by FormRules<sup>[6]</sup> that predict form III. These rules feature d84, d77 and d68 which are all CBZ descriptors and also the boiling point of the solvent.

Table 8.14 Rules generated in FormRules for form III prediction

Rules generated for Form III prediction			
SubModel:1	IF d84 is LOW	THEN Form III is	HIGH (1.00)
	IF d84 is HIGH	THEN Form III is	LOW (1.00)
SubModel:2	IF b.p. is LOW AND d77 is LOW	THEN Form III is	HIGH (0.50)
	IF b.p. is LOW AND d77 is HIGH	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is LOW	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is HIGH	THEN Form III is	HIGH (1.00)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.72)
	IF rate is HIGH	THEN Form III is	LOW (1.00)
SubModel:4	IF d68 is LOW	THEN Form III is	HIGH (1.00)
	IF d68 is MID	THEN Form III is	HIGH (1.00)
	IF d68 is HIGH	THEN Form III is	LOW (1.00)

Submodel 4 features d68, which is the moment of inertia C for CBZ, that also featured in the form II rules (Table 8.13). The d68 descriptor will therefore not be discussed again in this section.

D84, which is the fractional positive surface area descriptor for the CBZ molecule within a solvent force field, is the ratio of the total molecular surface area and the partial positive surface area.<sup>[4]</sup> These charged partial surface area descriptors have been used in the previous literature to predict a variety of biological and physicochemical properties.<sup>[27, 33-35]</sup> Due to the subtle differences between these descriptor values, it is difficult to interpret their effect on polymorphic crystallisation.

However, since the descriptor describes partial charges it provides information about solute and solvent interactions in solution.

The second rule incorporates a solvent and solute descriptor for the prediction of form III. D77 is the partial negative surface area of the CBZ molecule in the solvent force field, and like d84 it is a charge partial surface area descriptor. The boiling point of the solvent is also featured in this rule, and this property gives an indication into how strongly the solvent molecules interact with themselves. The boiling point is commonly used for predictions in the literature, but less so as a descriptor than a target for prediction.<sup>[34, 36]</sup>

This submodel 2 rule is interesting, as it states that the combination of high values for both d77 and the boiling point as well as a low value for both are associated with predicted high likelihood of form III crystallisation. This makes the interpretation of this rule somewhat difficult. However, closer examination reveals also that E\_ang and dsolv76 have an impact upon prediction (details can be found in 7.3.3). Therefore, perhaps this rule can be used as a simple guide for form III prediction, but there are a number of other descriptors also involved that contribute to making accurate predictions.

The third rule in Table 8.14 features the evaporation rate used in the crystallisation experiments. As mentioned in sections 8.3.1 and 8.3.2 a high evaporation rate often leads to the crystallisation of a metastable polymorphic form.<sup>[8-10]</sup> It seems therefore reasonable that a low evaporation rate should lead to the crystallisation of the thermodynamically stable form.

The evaporation rate can be used as a general rule for predicting the most likely polymorphic form to be crystallised. However, it has been found to be effected by differing values of dsolv69 in this research, shown in Figure 8.7.

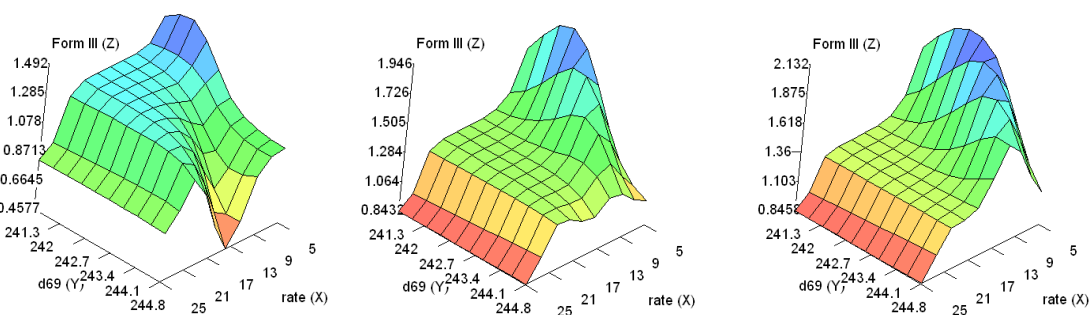


Figure 8.7 Effect of dsolv69 upon the form III predictions based upon the rules generated. Low dsolv69 values (left), mid range values (centre) and high dsolv69 values (right)

### 8.3.4. The Prediction of the Dihydrate

The rules generated by FormRules<sup>[6]</sup> for the prediction of the dihydrate were different from those presented in the PCA analysis (section 7.3.4, Table 7.23). It was expected that dsolv65 would directly replace dsolv74, which was previously used. However, it was not expected that dsolv43 would feature in this rule, as previously, dsolv69 was present. The rules presented in Table 8.15 feature, rate, temperature, dsolv65 and dsolv43 and in most instances predict a low dihydrate prediction. This general prediction of a low likelihood of hydrate crystallisation is consistent with the data used in the training, but due to the small dataset available for training, the reliability of the rules is questionable.

Table 8.15 Rules generated in FormRules for dihydrate prediction

Rules generated for dihydrate prediction			
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv43	THEN	LOW	
is LOW	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv43	THEN	LOW	
is HIGH	DiHydrate is	(0.88)	
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(0.95)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv43	THEN	LOW	
is LOW	DiHydrate is	(1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv43	THEN	LOW	
is HIGH	DiHydrate is	(0.50)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(0.99)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	

Rules generated for dihydrate prediction - continued			
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv43	THEN	HIGH	
is LOW	DiHydrate is	(0.99)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv43	THEN	LOW	
is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(0.50)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND	THEN	LOW	
dsolv43 is LOW	DiHydrate is	(1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND	THEN	LOW	
dsolv43 is HIGH	DiHydrate is	(0.90)	

Dsolv65 was present in the form I rules (section 8.3.1) and is the 3D Bonding information content of the solvent molecules. This descriptor offers topological information about the solvent molecule based upon its branching and connectivity. Rate and temperature were also discussed in the form I rules. However, the high dihydrate prediction does not state that a high rate and temperature are required. It should be noted that the dihydrate form of CBZ is not a polymorph of CBZ, and therefore Ostwald's Rule of stages<sup>[22, 23]</sup> does not hold in this instance.

Dsolv43 is the 3D Randić index of the solvent molecule and is also classed as a topological descriptor. In a similar way to dsolv65, this descriptor represents the molecular branching and structure of the solvent molecule.<sup>[26, 45, 46]</sup> The Randić index has been reported in the biological and chemical literature<sup>[26, 50-52]</sup>, and it was reported that connectivity indices were essential to improve predictive models.<sup>[24]</sup> As

was discussed previously (section 8.3.1 and 7.3.5), the branching and connectivity of a molecule may be associated with how it interacts in solution. With highly branched solvents perhaps promoting or inhibiting certain molecular interactions.

When the two experimental conditions and the two descriptors that feature in the rules in Table 8.15 are plotted against the dihydrate forming experiments (Figure 8.8), it becomes clear that on no occasion are the requirements for a high dihydrate prediction satisfied based upon the data generated in this research. However, based upon the experimental data, this is to be expected. On no occasion was a pure dihydrate crystallised.

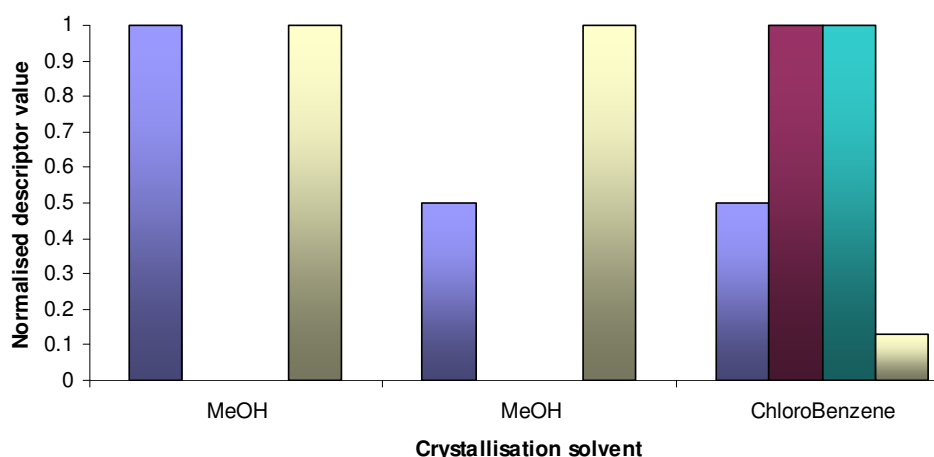


Figure 8.8 Crystallisation solvents plot against normalised rate (blue) , temperature (purple), dsolv65 (cream) and dsolv43 (green/blue) values.

### 8.3.5. The Prediction of Solvates

It has been noted in this research that only the DMSO solvate was formed. However, many other solvate forms have been reported in the literature.<sup>[32, 38-43]</sup> The rules presented in Table 8.16 are the same as those reported in section 7.3.5 (PCA best set), and contain two solvent descriptors, boiling point and dsolv43.

Table 8.16 Rules generated in FormRules for solvate prediction

Rules generated for solvate prediction		
IF b.p. is LOW AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)
IF b.p. is LOW AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)
IF b.p. is MID AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)
IF b.p. is MID AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)
IF b.p. is HIGH AND dsolv43 is LOW	THEN Solvate is	HIGH (1.00)
IF b.p. is HIGH AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)



Both boiling point and dsolv43, which is the 3D Randić index of the solvent molecule, have been presented in the previous sections (8.3.3 and 8.3.4) and will therefore not be discussed further.

### **8.3.6. Summary of the Optimised Descriptors**

Ten descriptors featured in this final set of descriptors, with seven of these highlighted in the rules generated in FormRules.<sup>[6]</sup> E\_ang, dsolv69 and dsolv76 were not mentioned in the rules, but were shown to have an impact upon form II and III prediction. This in itself demonstrates how the rules created can be used as a general guide for prediction, but that many of the descriptors are working non-linearly in tandem to create a successful prediction in INForm.<sup>[7]</sup>

## **8.4. Validation of Optimised Set**

As has been carried out in chapters 5 and 7 the model created using the optimised descriptors will now be validated. A cross validation set that is made up of 10 % of the experimental data was used and also an external validation set. This external validation set consists of experimental results generated from crystallisations in two previously untested solvents.

It is worth referring back to the conclusions made in chapter 7 about this validation set. The two solvents do not cover a wide range of descriptors values and are therefore perhaps not the most effective choice as validation solvents.

### **8.4.1. Cross Validation Results**

The 10 % of experimental data used as the validation set was removed from the training set, and the remaining data retrained. The overall average  $R^2$  value from INForm<sup>[7]</sup> was reduced from 89.30 % to 75.88 %. This was an expected reduction, as a large portion of the data has been removed.

The results of the cross validation are shown in Table 8.17 and show the model to be very successful at predicting the polymorphic outcome of these crystallisation experiments.

The major polymorphic form was correctly predicted for the first eight experimental inputs. Only the chlorobenzene input produced an output that appears to be

unconfident. This is consistent with the cross validation work in chapters 5 and 7 sections 5.5.1 and 7.4.1, where the major form was incorrectly predicted. In the experimental work carried out, crystallisation from chlorobenzene led to the formation of both forms II and III. This prediction suggests that the model is unsure as to which form would be produced. A result like this could be useful in an industrial setting that strives to consistently crystallise the thermodynamically stable form. Clearly mixed products are likely in this solvent, and therefore it should not be taken forward for further analysis.

A mixed product from crystallisation in chloroform was also tested in this validation, and it was encouraging to find that the model predicted very similar values for both forms II and III.

Table 8.17 Cross validation results summary

Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
					Form I	Form II	Form III	Dihydrate	Solvate
Ethanol	15	25	Form III	Form III	0.0	0.1	0.8	0.0	0.0
THF	25	25	Form II	Form II	0.0	1.0	0.0	0.3	0.1
Acetonitrile	15	50	Form III	Form III	0.0	0.3	1.0	0.0	0.2
DMSO	25	25	Solvate	Solvate	0.0	0.0	0.3	0.0	0.9
Aniline	5	50	Form III	Form III	0.0	0.1	1.0	0.0	0.0
Chlorobenzene	5	50	Form III	Form III	0.0	0.2	0.3	0.0	0.1
Toluene	15	75	Form II	Form II	0.0	1.0	0.0	0.0	0.0
Nitromethane	5	25	Form III	Form III	0.0	0.0	1.0	0.0	0.0
Chloroform	25	50	Form II / Form III	Form II	0.0	0.6	0.5	0.0	0.0

### 8.4.2. External Validation Results

Using the same external validation set as in chapters 5 and 7, polymorphic form predictions were made for a set of experimental results that have not been used in training the model. The results of this analysis are presented in Table 8.18 and show that only three of the 12 experiments major polymorphic forms were correctly predicted.

It has been commented upon previously (section 7.4.2) that the distribution of the descriptor values for the two external validation solvents is not broad, except for those descriptors that are outside of the range used in training. Figure 8.9 shows where the ethyl acetate (E) and n-butanol (B) descriptor values are within the range of the training descriptor values.

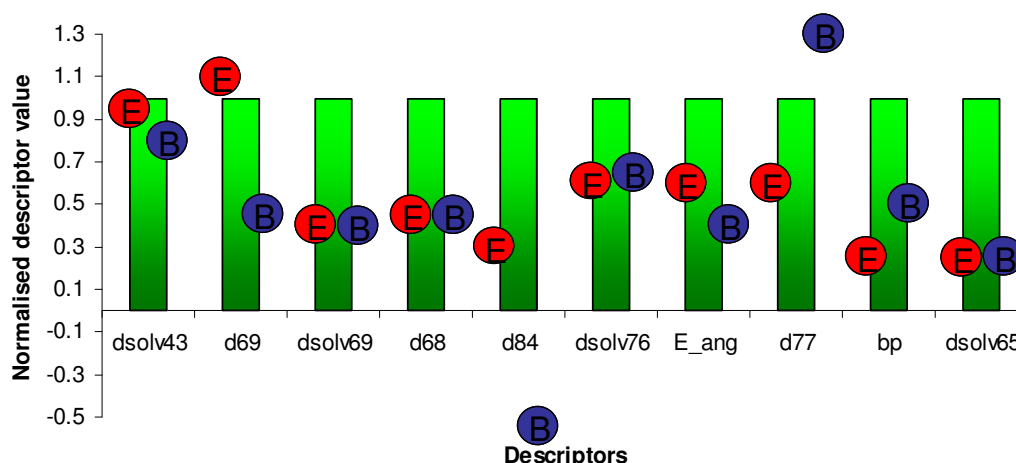


Figure 8.9 The distribution of the validation solvents descriptor values. E represents the ethyl acetate values and B the n-butanol values

As commented upon in section 7.4.2, there are two possible explanations for the poor results obtained in this external validation. As many of the descriptor values are similar in the validation set, it does not examine the whole scope of the model, and perhaps these descriptors are in an area of parameter space where the model struggles to make predictions. In the descriptors that do show differing values, at least one of the values is outside the range used in the training set. Ideally the validation solvents should have been chosen so that a large range of the descriptor space was assessed.

Table 8.18 External validation results summary

Experiment number	Solvent	Rate (L/min)	Temperature (°C)	Experimental result: Major form crystallised	Predicted result: Major form predicted	ANN predicted value				
						Form I	Form II	Form III	Dihydrate	Solvate
1	EtOAc	5	25	Form II	Form III	0.0	0.0	1.0	0.1	0.2
2	EtOAc	5	50	Form II	Form III	0.0	0.1	1.0	0.0	0.1
3	EtOAc	25	25	Form II	Form III	0.2	0.3	0.9	0.0	0.1
4	EtOAc	25	50	Form II	Form III	0.1	0.3	1.0	0.0	0.1
5	EtOAc	15	25	Form II	Form III	0.0	0.3	1.0	0.0	0.1
6	EtOAc	15	50	Form II	Form III	0.2	0.1	1.0	0.0	0.1
7	nBuOH	5	25	Form II	Form III	0.0	0.1	1.0	0.0	0.2
8	nBuOH	5	50	Form III	Form III	0.0	0.1	1.0	0.0	0.2
9	nBuOH	25	25	Form III	Form III	0.1	0.0	1.0	0.0	0.2
10	nBuOH	25	50	Form II	Form III	0.0	0.1	1.0	0.0	0.2
11	nBuOH	15	25	Form II	Form III	0.0	0.1	1.0	0.0	0.2
12	nBuOH	15	50	Form III	Form III	0.1	0.0	1.0	0.0	0.2

## 8.5. Conclusion of the Final Optimisation Analysis

Further optimisation of the Corr. and PCA best sets from chapters 5 and 7 was carried out in order to generate an improved model for polymorphic form prediction. By examining the linear correlations between the two best sets, dsolv74 was removed from the PCA best set and replaced with dsolv65 from the Corr. best set. This generated the most successful set seen in this research (overall average  $R^2$  value of 84.75 %).

It was known from the earlier PCA analysis (section 7.4.2) that when the external validation was carried out, three of the descriptors in the validation set had values outside of the range used in the training. This affected the overall success of the validation. Ideally, validation solvents should have been chosen so that the whole range of descriptor values were examined, rather than a very limit spaces (as shown in Figure 8.9).

The model produced using the optimised set of descriptors performed very well when cross validation was carried out. This therefore suggests that the model could accurately predict the major polymorphic form crystallised in a given experiment, provided that the descriptor values lie within the range used in the training. Further work could be carried out to try and extend the range of descriptor values, and determine if the same descriptors can be used.

The final set of descriptors are dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and boiling point.

- [1] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [2] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [3] A. R. Katritzky, S. Perumal, R. Petrukhim, *Journal of Organic Chemistry* **2001**, 66, 4036.
- [4] A. R. Katritzky, D. B. Tatham, *Journal of Chemical Information and Computer Sciences* **2001**, 41, 358.
- [5] D. M. Eike, J. F. Brennecke, E. J. Maginn, *Green Chemistry* **2003**, 5, 323.
- [6] R. Dowling, R. J. Davey, R. A. Curtis, G. Han, S. K. Poornachary, P. S. Chow, R. Tan, B. H., *Chemical Communications* **2010**, 46, 5924.
- [7] M. M. Parmar, O. Khan, L. Seton, J. L. Ford, *Crystal Growth and Design* **2007**, 7, 1635.
- [8] D. Musumeci, C. A. Hunter, J. F. McCabe, *Crystal Growth and Design* **2010**, 10, 1661–1664.
- [9] R. C. Kelly, N. Rodriguez-Hornedo, *Organic Process Research and Development* **2009**, 13, 1291.
- [10] J. F. McCabe, *CrystEngComm* **2010**, 12, 1110.
- [11] T. Threllfall, *Organic Process Research & Development* **2003**, 7, 1017.
- [12] A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy, N. Shankland, *Journal of Pharmaceutical Sciences* **2006**, 95, 1918.
- [13] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Vol. 11, first ed., Wiley-VCH, Weinheim, **2000**.
- [14] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, 113, 9480.
- [15] J.-P. Liu, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, 55, 41.
- [16] J.-P. Lui, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, 55, 41.
- [17] <http://www.chemcomp.com/journal/descr.htm>, *The Chemical Computing Group Viewed on 29/04/11*.
- [18] J. Auer, J. Bajorath, *Journal of Chemical Information and Modeling* **2008**, 48, 1747–1753.
- [19] N. S. H. N. Moorthy, N. S. Cerqueira, M. J. Ramos, P. A. Fernandes, *Medical Chemistry Research* **2010**, DOI 10.1007/s00044.
- [20] H. Yuan, A. L. Parrill, *Journal of Molecular Structure (Theochem)* **2000**, 529, 273–282.
- [21] R. Davey, J. Garside, *From Molecules to Crystallizers: An Introduction to Crystallization*, Oxford University Press, Oxford, **2000**.
- [22] N. Rodriguez-Hornedo, D. Murphy, *Journal of Pharmaceutical Sciences* **1999**, 88, 651.
- [23] R. C. Schweitzer, J. B. Morris, *Analytica Chimica Acta* **1999**, 384, 285.
- [24] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [25] G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, *Molecular Diversity* **2009**, 13, 301.
- [26] R. H. Rohrbaugh, P. C. Jurs, *Analytical Chemistry* **1985**, 57, 2770.
- [27] M. H. Fatemi, M. Ghorbanzade, *Molecular Diversity* **2009**, 13, 483.
- [28] E. R. Collantes, W. Tong, W. J. Welsh, *Analytical Chemistry* **1996**, 68, 2038.
- [29] T. P. Knowles, A. W. Fitzpatrick, S. Meehan, H. R. Mott, M. Vendruscolo, C. M. Dobson, M. E. Welland, *Science* **2007**, 318, 1900.

- [30] A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, *Chemical Communications* **2007**, 1600.
- [31] F. P. A. Fabbiani, L. T. Byrne, J. J. McKinnon, M. A. Spackman, *CrystEngComm* **2007**, 9, 728.
- [32] D. T. Stanton, P. J. Madhav, L. J. Wilson, T. W. Morris, P. M. Hershberger, C. N. Parker, *Journal of Chemical Information and Computer Sciences* **2004**, 44, 221.
- [33] M. Cocchi, P. G. De Benedetti, R. Seeber, L. Tassi, A. Ulrich, *Journal of Chemical Information and Computer Sciences* **1999**, 39, 1190.
- [34] C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *Journal of Chemical Information and Modeling* **2005**, 45, 170.
- [35] A. A. Ivanova, A. A. Ivanov, A. A. Oliferenko, V. A. Palyulin, N. S. Zefirov, *SAR and QSAR in Environmental Research* **2005**, 16, 231.
- [36] M. Randic, *Journal of Chemical Information and Computer Sciences* **1984**, 24, 164.
- [37] H. Liu, M. Lu, F. Tian, *Journal of Mathematical Chemistry* **2005**, 38, 345.
- [38] E. Estrada, *Journal of Physical Chemistry A* **2002**, 106, 9085.
- [39] O. Ivanciuc, T. Ivanciuc, P. A. Filip, D. Cabrol-Bass, *Journal of Chemical Information and Computer Sciences* **1999**, 39, 515.
- [40] Y. Ren, H. Liu, X. Yao, M. Liu, *Journal of Chromatography A* **2007**, 1155, 105.
- [41] R. K. Harris, P. Y. Ghi, H. Puschmann, D. C. Apperley, U. J. Griesser, R. B. Hammond, C. Ma, K. J. Roberts, G. J. Pearce, J. R. Yates, C. J. Pickard, *Organic Process Research and Development* **2005**, 9, 902.
- [42] A. Johnston, A. J. Florence, A. R. Kennedy, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1509.
- [43] A. Johnston, B. F. Johnston, A. R. Kennedy, A. J. Florence, *CrystEngComm* **2008**, 10, 23.
- [44] S. Lohani, Y. Zhang, L. J. Chyall, P. Mougin-Andres, F. X. Muller, D. J. W. Grant, *Acta Crystallographica, Section E: Structure Reports Online* **2005**, 61, 1310.
- [45] S. G. Fleischman, S. S. Kuduva, J. A. McMahon, B. Moulton, R. D. Bailey Walsh, N. Rodriguez-Hornedo, M. J. Zaworotko, *Crystal Growth and Design* **2003**, 3, 909.
- [46] A. J. Cruz Cabeza, G. M. Day, W. D. Samuel Motherwell, W. Jones, *Journal of the American Chemical Society* **2006**, 128, 14466.



## **9. RESULTS AND DISCUSSION OF ANALYSIS WITH DIFFERENT TARGET MOLECULES**

The ANN methodology based on computational and experimental descriptors was developed in the context of carbamazepine crystallisation. As a first step towards examining whether the method can be generalised to other systems, and especially to determine whether similar physical parameters influence the outcome of their crystallisation, two additional polymorphic systems were investigated. Tolbutamide (TBA) and 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY) are well known polymorphic substances. Both these systems feature a conformationally flexible molecule, setting them significantly apart from carbamazepine (CBZ), and their polymorphs have sufficiently different X-ray powder diffraction (XRPD) patterns that permit straightforward identification of polymorphic form.

This chapter will first address the performance of the artificial neural network (ANN) using only the descriptors highlighted as important in the CBZ work. Subsequently it will be assessed whether the principal component analysis (PCA) selection method can be used successfully for the different target molecules.

### **9.1. Analysis of Descriptors Highlighted in CBZ Analysis**

From the research presented in chapter 8, the most successful set of descriptors that predicted the polymorphic forms of CBZ are, dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and boiling point. These descriptors have been calculated for both TBA and ROY and will be run in an ANN as a complete set and also using the descriptors for the solvents only.

The aim of this analysis is to provide an insight into how transferable the descriptors are between different polymorphic target molecules. Validation data are not available for these systems; significant further experimental work would need to be carried out that was beyond the scope of this research. However, this analysis may potentially highlight

useful descriptors that can form part of a generic model for polymorphic form prediction.

### 9.1.1. TBA Results

The dataset available for TBA was much smaller than that used in the CBZ analysis, comprising of only 33 experimental results (Electronic Appendix, Chapter 9, file 9.1). Ideally more data are required, as 33 rows of input data are not really sufficient to reliably train an ANN. However, to obtain a rough guide as to whether the descriptors might be transferable between the different target molecules, a network was built nevertheless. The results are presented in Table 9.1.

Table 9.1 Analysis of the TBA descriptors based upon previous CBZ research. X denotes the presence of the descriptor in the set

Descriptors	Best set descriptors from CBZ analysis	Solvent descriptors only	Best set descriptors from CBZ analysis – 2 forms only	Solvent descriptors only - 2 forms only
Dsolv43	X	X	X	X
Dsolv69	X	X	X	X
Dsolv76	X	X	X	X
D68	X		X	
D69	X		X	
D77	X		X	
D84	X		X	
E_ang	X		X	
Boiling point	X	X	X	X
Dsolv65	X	X	X	X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>61.94</b>	<b>62.04</b>	<b>73.87</b>	<b>74.01</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>66.74</b>	<b>-50.21</b>	<b>71.51</b>	<b>11.69</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>64.34</b>	<b>5.92</b>	<b>72.69</b>	<b>42.85</b>

Although three different forms of TBA were present in the dataset, there was only one occurrence of the third form, prohibiting successful training of the ANN. When the data for the third form were removed from the set and the model rebuilt using the optimised CBZ descriptor set, the overall average R<sup>2</sup> values increased. The results in Table 9.1

show that by building a model that contained descriptors relating to the TBA molecule, rather than just using the solvent descriptors, the results were improved.

The rules generated for the most successful set are presented in Table 9.2 and Table 9.3 and show that rate, d68, d77 and dsolv69 are highlighted as important descriptors in TBA prediction.

Table 9.2 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting only TBA form I

Rules for best set descriptors from CBZ analysis – predicting only TBA form I			
--- Rules for property Form I ---			
SubModel:1	IF d68 is LOW AND Rate is LOW	THEN Form I is	HIGH (0.55)
	IF d68 is LOW AND Rate is HIGH	THEN Form I is	HIGH (0.58)
	IF d68 is MID AND Rate is LOW	THEN Form I is	HIGH (0.73)
	IF d68 is MID AND Rate is HIGH	THEN Form I is	LOW (1.00)
	IF d68 is HIGH AND Rate is LOW	THEN Form I is	HIGH (1.00)
	IF d68 is HIGH AND Rate is HIGH	THEN Form I is	HIGH (0.96)
SubModel:2	IF d77 is LOW	THEN Form I is	HIGH (0.74)
	IF d77 is HIGH	THEN Form I is	HIGH (1.00)

There is much dispute in the literature over the most stable form of TBA<sup>[1-5]</sup>. A summary of this controversy was already included in section 3.3.3. The experimental analysis carried out for this thesis suggests that form I is the most stable, but that it converts readily to form II in methanol and ethanol solutions (Table 3.5). If form I were the most stable form of TBA, it would be interesting to note that the descriptors found in the form I rules overlap with those present in the CBZ form III (thermodynamically most stable form) rules (section 8.3.3).

Table 9.3 presents the form II rules, and once again d68 is present. D68, which is the moment of inertia C for the target molecule, was also found in the two most frequently occurring forms of CBZ.

Table 9.3 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting only TBA form II

Rules for best set descriptors from CBZ analysis – predicting only TBA form II			
--- Rules for property Form II ---			
SubModel:1	IF d68 is LOW	THEN Form II is	LOW (1.00)
	IF d68 is MID	THEN Form II is	HIGH (1.00)
	IF d68 is HIGH	THEN Form II is	LOW (0.99)
SubModel:1	IF Rate is LOW AND dsolv69 is LOW	THEN Form II is	LOW (1.00)
	IF Rate is LOW AND dsolv69 is HIGH	THEN Form II is	LOW (0.81)
	IF Rate is HIGH AND dsolv69 is LOW	THEN Form II is	HIGH (1.00)
	IF Rate is HIGH AND dsolv69 is HIGH	THEN Form II is	LOW (0.91)

Although the results are not as successful for TBA as they were in the CBZ analysis in chapter 8, it does show that potentially a reasonable model can be made using the same descriptors. For a truly reliable model to be built, more training data for TBA are required, but based upon the relatively small dataset the set of descriptors optimised in the CBZ analysis appear to be transferable to other target molecules, suggesting that most salient properties are covered by this set. However, the results also show that it is necessary to include some target molecule-specific descriptors, as the solvent descriptors alone can not successfully predict polymorphic form.

### 9.1.2. ROY Results

In a similar way to the TBA analysis, the molecular descriptors for ROY were calculated and an ANN built based upon the optimised descriptor set in the CBZ analysis (section 8.3).

34 rows of experimental data for ROY were used in the training, with only three of these not producing purely the thermodynamically stable form Y. ROY was initially chosen for this work due to the high number of polymorphic forms that can be readily crystallised. However, under the controlled conditions of this polymorph screen (results of which may be found in Electronic Appendix, Chapter 9, file 9.2) the most stable form was crystallised on most occasions.

When the optimised descriptor set from the CBZ analysis were applied to the ROY dataset, Table 9.4 shows the results of the FormRules<sup>[6]</sup> and INForm<sup>[7]</sup> analysis.

Table 9.4 Analysis of the ROY descriptors based upon previous CBZ research. X denotes the presence of the descriptor in the set

Descriptors	Best set descriptors from CBZ analysis	Solvent descriptors only
Dsolv43	X	X
Dsolv69	X	X
Dsolv76	X	X
D68	X	
D69	X	
D77	X	
D84	X	
E_ang	X	
Boiling point	X	X
Dsolv65	X	X
<b>FormRules average R<sup>2</sup> (%)</b>	<b>99.77</b>	<b>9.44</b>
<b>INForm average R<sup>2</sup> (%)</b>	<b>66.61</b>	<b>98.13</b>
<b>Overall average R<sup>2</sup> (%)</b>	<b>83.19</b>	<b>53.79</b>

It can be seen in Table 9.4 that when both the ROY and solvent descriptors are used, a high overall average R<sup>2</sup> value was achieved. It does appear as if the INForm<sup>[7]</sup> results for the solvent-only analysis have performed extremely well. However, all of the test set used had the same experimental value, which was also the major form produced. Perhaps the results of FormRules<sup>[6]</sup> give a better insight into the ability of the descriptors used to predict polymorphic form. When the rules for the model that used both ROY and solvent descriptors are examined (Table 9.5), it is clear that it is the ROY descriptors that are of importance. The rules highlight rate, temperature, E\_ang and d69 as valuable descriptors in the predictions of the Y and R forms.

Table 9.5 Rules generated by FormRules for the best set descriptors from CBZ analysis – predicting ROY, Y and R forms

Rules for best set descriptors from CBZ analysis – predicting ROY, Y and R forms			
--- Rules for property R ---			
SubModel:1			
IF E_ang is LOW AND Temp is LOW AND rate is LOW	THEN R is	LOW (1.00)	
IF E_ang is LOW AND Temp is LOW AND rate is HIGH	THEN R is	LOW (1.00)	

Rules for best set descriptors from CBZ analysis – predicting ROY, Y and R forms – continued			
IF E_ang is LOW AND Temp is MID AND rate is LOW	THEN R is	LOW (1.00)	
IF E_ang is LOW AND Temp is MID AND rate is HIGH	THEN R is	LOW (1.00)	
IF E_ang is LOW AND Temp is HIGH AND rate is LOW	THEN R is	HIGH (0.57)	
IF E_ang is LOW AND Temp is HIGH AND rate is HIGH	THEN R is	LOW (1.00)	
IF E_ang is HIGH AND Temp is LOW AND rate is LOW	THEN R is	HIGH (1.00)	
IF E_ang is HIGH AND Temp is LOW AND rate is HIGH	THEN R is	HIGH (0.91)	
IF E_ang is HIGH AND Temp is MID AND rate is LOW	THEN R is	LOW (1.00)	
IF E_ang is HIGH AND Temp is MID AND rate is HIGH	THEN R is	LOW (1.00)	
IF E_ang is HIGH AND Temp is HIGH AND rate is LOW	THEN R is	LOW (0.92)	
IF E_ang is HIGH AND Temp is HIGH AND rate is HIGH	THEN R is	LOW (1.00)	
SubModel:2			
IF d69 is LOW AND Temp is LOW	THEN R is	HIGH (0.55)	
IF d69 is LOW AND Temp is HIGH	THEN R is	LOW (1.00)	
IF d69 is HIGH AND Temp is LOW	THEN R is	LOW (0.70)	
IF d69 is HIGH AND Temp is HIGH	THEN R is	LOW (0.80)	
--- Rules for property Y ---			
SubModel:1			
IF E_ang is LOW AND Temp is LOW AND rate is LOW	THEN Y is	HIGH (1.00)	
IF E_ang is LOW AND Temp is LOW AND rate is HIGH	THEN Y is	HIGH (0.98)	
IF E_ang is LOW AND Temp is MID AND rate is LOW	THEN Y is	HIGH (0.89)	
IF E_ang is LOW AND Temp is MID AND rate is HIGH	THEN Y is	HIGH (0.89)	
IF E_ang is LOW AND Temp is HIGH AND rate is LOW	THEN Y is	LOW (0.64)	
IF E_ang is LOW AND Temp is HIGH AND rate is HIGH	THEN Y is	LOW (1.00)	
IF E_ang is HIGH AND Temp is LOW AND rate is LOW	THEN Y is	LOW (1.00)	
IF E_ang is HIGH AND Temp is LOW AND rate is HIGH	THEN Y is	LOW (1.00)	
IF E_ang is HIGH AND Temp is MID AND rate is LOW	THEN Y is	HIGH (0.89)	
IF E_ang is HIGH AND Temp is MID AND rate is HIGH	THEN Y is	HIGH (0.89)	
IF E_ang is HIGH AND Temp is HIGH AND rate is LOW	THEN Y is	LOW (0.95)	
IF E_ang is HIGH AND Temp is HIGH AND rate is HIGH	THEN Y is	LOW (1.00)	
SubModel:2			
IF d69 is LOW AND Temp is LOW	THEN Y is	HIGH (1.00)	
IF d69 is LOW AND Temp is HIGH	THEN Y is	HIGH (1.00)	
IF d69 is HIGH AND Temp is LOW	THEN Y is	HIGH (1.00)	
IF d69 is HIGH AND Temp is HIGH	THEN Y is	HIGH (0.97)	

When the normalised descriptor values of rate, temperature and E\_ang are plotted against the R form producing experiments (Figure 9.1), it can be observed that the rules generated are based upon the descriptor values explicitly.

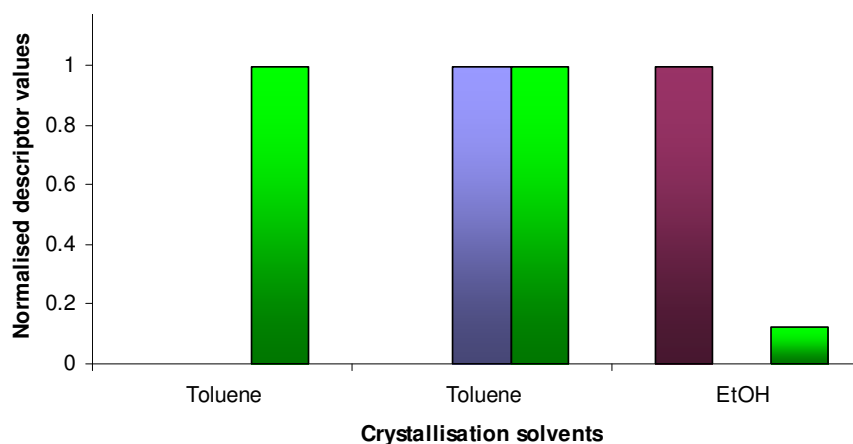


Figure 9.1 Normalised descriptor values plot for R form producing experiments. Rate (blue), temperature (purple) and E\_ang (green)

It is interesting that the highest levels of the R form are crystallised from toluene, which has the highest overall E\_ang value of all the solvents used. Perhaps with more experimental data a more reliable and potentially informative model could be generated.

### 9.1.3. Conclusion of Highlighted Descriptor Analysis

Overall, for both TBA and ROY, the models produced when the optimised descriptors from the CBZ analysis are used, perform well. Both TBA and ROY datasets require more systematic experimental work to be carried out in order to conclusively assess whether these descriptors are truly transferable. However, based upon the rough test using a small dataset, it does appear as if using a selection of both target molecule and solvent descriptors a predictive model for polymorphic form can be produced.

## 9.2. Descriptor Selection using PCA

The most efficient method of data reduction was using PCA as demonstrated in chapter 7. A scree plot for each molecule was generated, which highlights the number of components that contain most of the variation in the dataset. By selecting the descriptors

with the most significant loading values from the highlighted principal components, a reduced descriptor set was established.

This method generated an overall average  $R^2$  value of 84.31 % for CBZ after optimisation, but an initial value of 82.49 % when the scree plot analysis method was employed. It was hoped that this straightforward data reduction and descriptor selection method could be used for other polymorphic systems as a rapid method for producing a model that could predict polymorphic form.

### 9.2.1. TBA Results

Seven components were calculated for the TBA dataset, with the scree plot presented in Figure 9.2 and the results in Electronic Appendix, Chapter 9, file 9.3.

It can be seen from the scree plot that components six and seven do not contain a large amount of information from the dataset, but because of the difference between component five and six, components 1-6 were analysed using the method outlined in chapter 7.

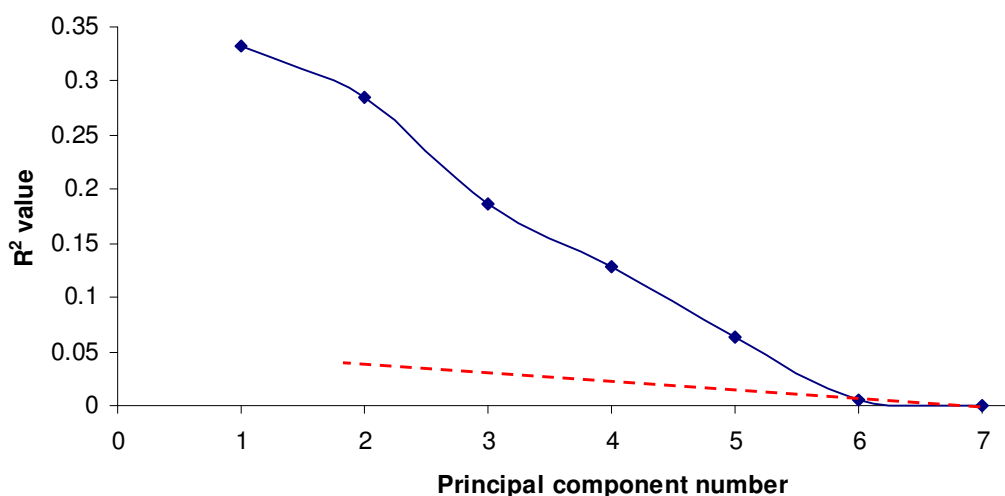


Figure 9.2 Scree plot based upon the TBA PCA data

The most positively and negatively loaded descriptors from each principal component will be used as a descriptor selection method for use in a set for ANN analysis (7.13). Table 9.6 details the descriptors to be used in this analysis.



Table 9.6 The most positively and negatively loaded descriptors taken from the TBA principal components

Principal component	Most positively loaded descriptor	Second most positively loaded descriptor	Most negatively loaded descriptor	Second most negatively loaded descriptor
1	Dsolv51	Dsolv28	Dsolv20	E_strain
2	ASA	ASA_H	D77	PM3_Eele
3	D86	Dsolv38	dP	Dsolv67
4	LogP	Henry's law constant	Dielectric constant	Dsolv3
5	Dsolv4	Dsolv76	Dsolv13	Activity
6	Solubility	Dsolv13	activity	Dsolv76

It should be noted that dsolv13, dsolv76 and activity are featured on two occasions in Table 9.6 and therefore will only be included once in the subsequent analysis. The results of which are shown in Table 9.7.

Table 9.7 Results of the PCA data reduction analysis. The number in brackets is the number of descriptors used in the ANN

	PC1-6 (21)	PC1-6 (12)	PC1-5 (20)	PC1-5 (10)	PC1-4 (16)
Form Rules average $R^2$ (%)	54.82	52.27	54.82	52.27	54.82
INForm average $R^2$ (%)	77.26	37.06	80.34	73.30	78.12
Overall average $R^2$ (%)	66.04	44.67	67.58	62.79	66.47

	PC1-4 (8)	PC1-3 (12)	PC1-3 (6)	PC1-2 (8)	PC1-2 (4)
Form Rules average $R^2$ (%)	52.27	52.27	52.27	47.44	51.91
INForm average $R^2$ (%)	77.55	71.83	68.95	-462.46	64.58
Overall average $R^2$ (%)	64.91	62.05	60.61	-207.51	58.25

It can be seen in Table 9.7 that by using the two most positively and negatively loaded descriptors from components 1-5, PC1-5 (20), the highest overall average  $R^2$  value was achieved. The analysis in section 9.1.1 indicated that the results improved when the data for the third polymorph were removed. This generated the results presented in Table 9.8.

Analysis of the descriptors that featured in the rules (rate, dsolv3 and dsolv28) was also carried out and its outcomes are summarised in Table 9.8.

Table 9.8 Results of the PCA data reduction analysis when the data for the third polymorph is removed. The number in brackets is the number of descriptors used in the ANN

	PC1-5 (20) – 2 forms only	PC1-5 (20) – Rule only descriptors
<b>Form Rules average <math>R^2</math> (%)</b>	<b>80.77</b>	<b>80.77</b>
<b>INForm average <math>R^2</math> (%)</b>	<b>94.89</b>	<b>-44.90</b>
<b>Overall average <math>R^2</math> (%)</b>	<b>87.83</b>	<b>17.94</b>

The results of the PCA data reduction technique have shown that by selecting the two most positively and negatively loaded descriptors from the first five components a good model can be produced that predicts the polymorphic outcomes of TBA crystallisations. In contrast, just as observed for CBZ in section 8.2 the results of using the rule only descriptors showed that other descriptors need to be present in order to generate a successful prediction.

The 20 descriptors in the most successful set for TBA were classified according to overall physical property in Figure 9.3. Many of the descriptors are bulk or empirical solvent properties (green box), while others relate to molecular interactions of either the TBA or the solvent molecule.

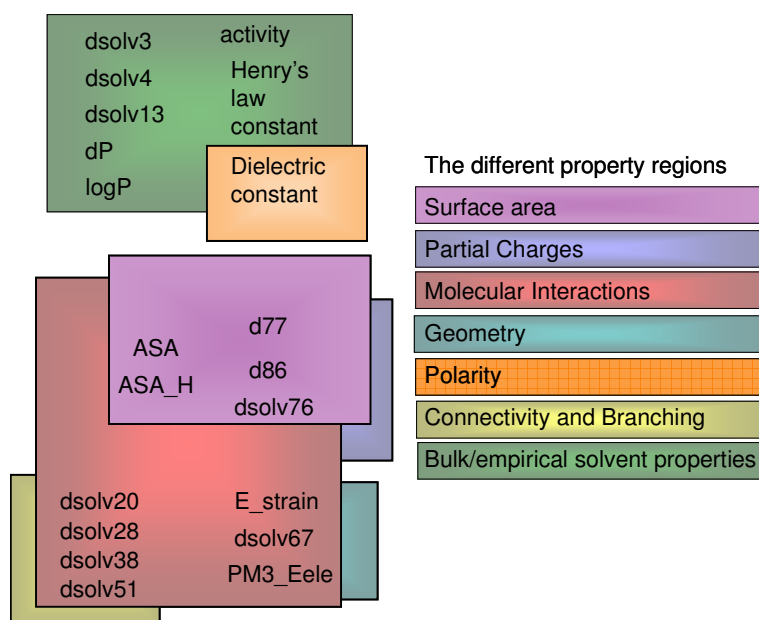


Figure 9.3 Descriptors in the most successful set grouped based upon their physical meaning

It is interesting to compare the final descriptor set identified in the CBZ analysis with this set of descriptors and observe the overlapping and similar descriptors. Table 9.9 groups the descriptors based upon their physical property class and shows both the descriptors from the PCA work and those in the final CBZ set.

Table 9.9 Descriptors grouped based upon their physical meaning

Physical Meaning of Descriptor	TBA PCA descriptor set	CBZ descriptor set
Solvent Descriptors: Molecular Surface Area	Dsolv76	Dsolv69, dsolv76
Solvent Descriptors: Partial Charges	Dsolv76	Dsolv76
		Dsolv43, dsolv65,
Solvent Descriptors: Molecular Interactions	Dsolv67, dsolv76	dsolv76, Boiling point
Solvent Descriptors: Geometry	Dsolv67	Dsolv69
Solvent Descriptors: Connectivity and Branching	Dsolv20, dsolv28, dsolv38, dsolv51	Dsolv43, dsolv65,
	Dsolv3, dsolv4,	
Solvent Descriptors: Bulk and empirical properties	dsolv13, dP, logP, activity, Henry's law constant, dielectric constant	
Solvent Descriptors: Polarity	dielectric constant	
CBZ Descriptors: Molecular Surface Area	ASA, ASA_H, d77, d84	D69, d77, d84
CBZ Descriptors: Partial Charges	D77, d84	D77, d84
	ASA, ASA_H, d77,	
CBZ Descriptors: Molecular Interactions	d84, E_strain, PM3_Eele	D68, d77, d84, E_ang
CBZ Descriptors: Geometry	E_strain, PM3_Eele	D68, d69, E_ang

There are three overlapping descriptors between the two sets, dsolv76, d77 and d84. These descriptors represent the partial negative surface area of the solvent (dsolv76) and the partial and fractional negative surface area of the TBA molecule (d77 and d84) detailed in appendix section 12.2.

Overall the PCA data reduction technique successfully reduces the descriptors and allows predictions to be made successfully. In order to improve the reliability of this

method, further experimental data would be needed, as the current model has very little training data. However, the work carried out so far already indicates that this descriptor selection technique may generally allow a predictive model to be built very rapidly.

### 9.2.2. ROY Results

Eleven components were calculated for the ROY dataset (Electronic Appendix, Chapter 9, file 9.4), with the scree plot presented in Figure 9.4.

From the plot it can be seen that the elbow is at component 6, therefore components 1-6 were analysed using the method outlined in chapter 7.

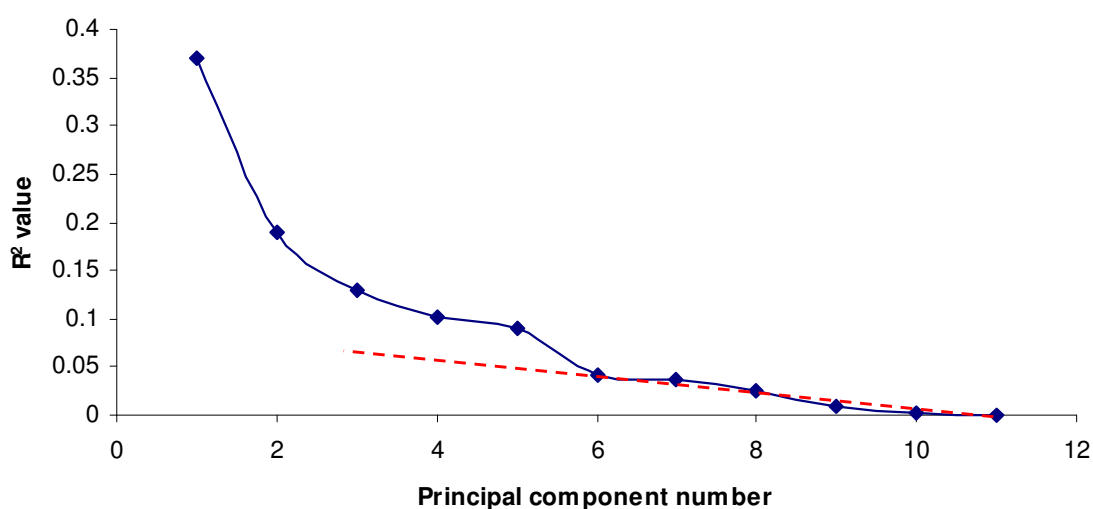


Figure 9.4 Scree plot based upon the ROY PCA data

Table 9.10 contains the two most positively and negatively loaded descriptors in components 1-6.

Table 9.10 The most positively and negatively loaded descriptors taken from the ROY principal components

Principal component	Most positively loaded descriptor	Second most positively loaded descriptor	Most negatively loaded descriptor	Second most negatively loaded descriptor
1	Dsolv47	logP	Dielectric constant	dP
2	D81	D82	D84	ASA_H
3	Dsolv57	Dsolv61	AM1_IP	Viscosity
4	PM3_LUMO	PM3_IP	PM3_dipole	PM3_Eele
5	MNDO_Eele	MNDO_HF	Std_dim1	MNDO_IP
6	Gutmann donor number	Dsolv80	Dsolv51	Freezing point

In a similar way to the TBA analysis, these descriptors were used as the basis for ANN analysis. Table 9.11 summarises the results of the PCA descriptor reduction analysis.

Table 9.11 Results of the PCA data reduction analysis. The number in brackets is the number of descriptors used in the ANN

	PC1-6 (24)	PC1-6 (12)	PC1-5 (20)	PC1-5 (10)	PC1-4 (16)
Form Rules average $R^2$ (%)	93.11	93.11	92.54	96.12	95.55
INForm average $R^2$ (%)	90.19	89.87	98.67	89.29	89.67
Overall average $R^2$ (%)	91.65	91.49	95.61	92.71	92.61

	PC1-4 (8)	PC1-3 (12)	PC1-3 (6)	PC1-2 (8)	PC1-2 (4)
Form Rules average $R^2$ (%)	96.11	95.55	96.11	95.55	96.11
INForm average $R^2$ (%)	99.42	85.03	97.99	98.67	67.11
Overall average $R^2$ (%)	97.77	90.29	97.05	97.11	81.61

It can be seen in Table 9.11 that most of the sets of descriptors effectively predict the polymorphic outcome of the ROY crystallisation experiments, with the most successful set being PC1-4 (8). This set contained only the most positively and negatively loaded descriptors from components 1-4.

When the rules from this set were examined, only temperature and d84 featured. A network was run using only these two descriptors and generated a reduced overall

average  $R^2$  value. However, even though this network didn't perform as well the PC1-4 (8) set, as it was still very successful. The results of this analysis are presented in Table 9.12.

Table 9.12 The descriptors that featured in the rules of PC1-4 (8)

Set	Descriptors	FormRules average $R^2$ (%)	INForm average $R^2$ (%)	Overall average $R^2$ (%)
<b>PC1-4 (8) rule only descriptors</b>	D84, Temp	96.11	98.04	97.08

The eight descriptors in the most successful set have been grouped in Figure 9.5 based on the classes of physical properties they represent. They are also shown in Table 9.13 next to the best sets from the CBZ and TBA analysis.

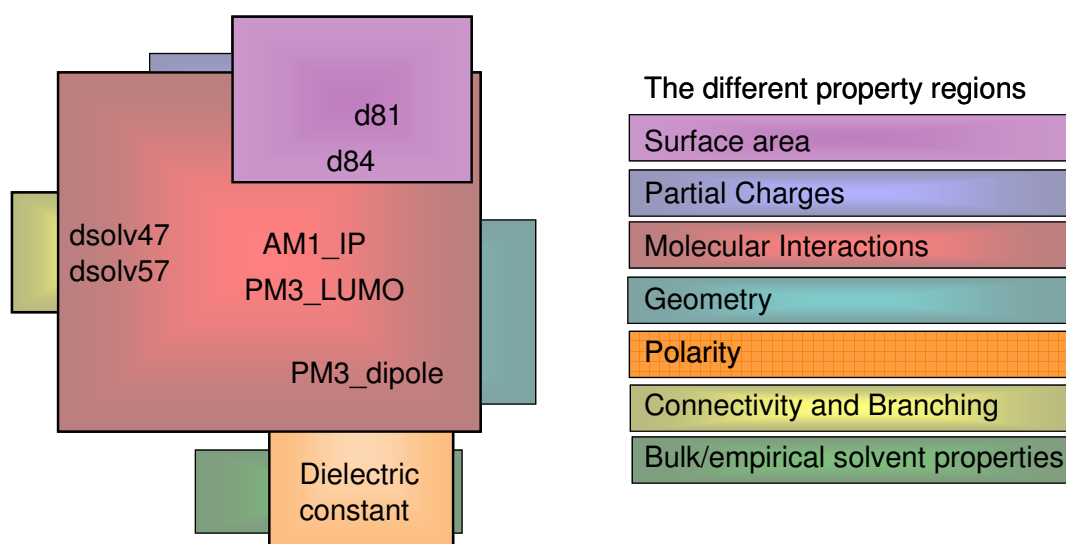


Figure 9.5 ROY PCA descriptors grouped based on their physical meaning

The majority of the descriptors in this set are ROY molecular descriptors, with only one bulk property and two branching and connectivity values for the solvent.

The most interesting feature of this set is the presence of d84, the fractional partial negative surface area of the ROY molecule (discussed in section 8.3.3). D84 features in the final sets for CBZ, TBA and ROY. This descriptor was calculated for the target molecules in the different solvent force fields, therefore the variation between each value is very subtle. This perhaps suggests that it is the slight changes in molecular geometry

in different solvents that is an important factor in the direction of which polymorphic form is crystallised.

Table 9.13 shows the descriptors that feature in the best sets for CBZ, TBA and ROY. As well as the presence of d84 in all three sets, dielectric constant is present in both the ROY and TBA sets.

Table 9.13 Descriptors grouped based upon their physical meaning

Physical Meaning of Descriptor	ROY PCA descriptor set	TBA PCA descriptor set	CBZ descriptor set
Solvent Descriptors: Molecular Surface Area		Dsolv76	Dsolv69, dsolv76
Solvent Descriptors: Partial Charges		Dsolv76	Dsolv76
Solvent Descriptors: Molecular Interactions	Dsolv47, dsolv57	Dsolv67, dsolv76	Dsolv43, dsolv65, dsolv76, Boiling point
Solvent Descriptors: Geometry		Dsolv67	Dsolv69
Solvent Descriptors: Connectivity and Branching	Dsolv47, dsolv57	Dsolv20, dsolv28, dsolv38, dsolv51 Dsolv3, dsolv4, dsolv13, dP, logP, activity, Henry's law constant, dielectric constant	Dsolv43, dsolv65,
Solvent Descriptors: Bulk and empirical properties	Dielectric constant	dielectric constant	
Solvent Descriptors: Polarity			
CBZ Descriptors: Molecular Surface Area	D81, d84	ASA, ASA_H, d77, d84	D69, d77, d84
CBZ Descriptors: Partial Charges	D81, d84	D77, d84	D77, d84
CBZ Descriptors: Molecular Interactions	D81, d84, PM3_dipole, AM1_IP, PM3_LUMO	ASA, ASA_H, d77, d84, E_strain, PM3_Eele	D68, d77, d84, E_ang
CBZ Descriptors: Geometry	PM3_dipole, AM1_IP, PM3_LUMO	E_strain, PM3_Eele	D68, d69, E_ang
CBZ Descriptors: Polarity	PM3_dipole		

Using PCA to reduce the number of descriptors and then running an ANN generates a successful predictive model for the ROY polymorphic data. Many of the crystallisations in this analysis led to the formation of the same polymorphic form, therefore in order to interrogate the model further, additional crystallisation experiments would need to be carried out, for two reasons. The first reason is to introduce other polymorphic forms to

the model, and the second to increase the dataset so that a more reliable predictive model can be created.

### **9.2.3. Summary**

The PCA data reduction technique has been applied to two different polymorphic target molecules and successfully generated a predictive model. Although there were very little training data available for these two systems, the results suggest that this descriptor selection technique is generically applicable and successful.

By generating a scree plot using the PCA data, the most important components can be determined. The descriptors can then be selected from these components based upon their loading values. Ten networks were run for each molecule in this research, in order to fully optimise the model based upon the descriptors selected.

Overall, by using PCA the most useful descriptors for predicting the polymorphic form outcome of a target molecule can be identified. When these descriptors are used in an ANN, a predictive model can be built. More experimental data would increase the reliability of the models and the conclusions drawn from them.

## **9.3. Overall Conclusions and Summary of Chapter**

This chapter has presented one set of descriptors that may be transferable across different polymorphic target molecules and one method of descriptor selection.

The descriptor set that was generated through the CBZ analysis was successfully applied to both TBA and ROY. Models were built for each molecule using these descriptors, generating overall average  $R^2$  values of 72.69 % for TBA and 83.19 % for ROY. Although not as successful as the CBZ results (84.75 %), as an immediate set of descriptors that could be used to build a predictive model for another polymorphic target molecule, the results are very good.

The PCA descriptor reduction technique demonstrates an effective method for selecting descriptors that will generate a successful predictive ANN model. Although only one common descriptor was present in the sets of each molecule (d84), the descriptors displayed similar physical meanings across each model. The overall average  $R^2$  values



generated in this analysis were higher than those observed in the analysis with the CBZ descriptors, at 87.83 % for TBA and 97.77 % for ROY.

This chapter has successfully applied a set of descriptors and a descriptor selection method to two different polymorphic target molecules. More experimental data would be required in order to improve the reliability of these models. However, the analysis has proved the concept of transferring a descriptor set and a descriptor selection method to other target molecules.

- [1] A. Burger, *Scienta Pharmaceutica* **1975**, 43, 161.
- [2] E. L. Rowe, B. D. Anderson, *Journal of Pharmaceutical Sciences* **1984**, 73, 1673.
- [3] K. Kimura, F. Hirayama, K. Uekama, *Journal of Pharmaceutical Sciences* **1999**, 88, 385.
- [4] G. Hasegawa, T. Komasa, R. Bando, Y. Yoshihashi, E. Yonemochi, K. Fujii, H. Uekusa, K. Terada, *International Journal of Pharmaceutics* **2009**, 369, 12.
- [5] S. Thirunahari, S. Aitipamula, P. S. Chow, R. B. H. Tan, *Journal of Pharmaceutical Sciences* **2010**, 99, 2975.
- [6] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [7] INForm, v3.7 ed., Intelligensys Ltd., **2009**.

## 10. CONCLUSIONS

The majority of the work described in this thesis focussed on strategies for adequately removing redundancy in the information covered by the descriptor sets. The overall aim was to generate artificial neural networks (ANNs) based on a descriptor set that is on one hand small enough to permit the derivation of plausible hypothetical relationships between physical properties and experimentally observed behaviour, yet on the other capable of predicting the polymorphic outcome of a given crystallisation experiment with high probability.

Subjective individual examination ('manual analysis') of descriptor relevance was carried out alongside partial least squares (PLS) and principal component analysis (PCA), which are more quantitative. The manual approach and the PCA methods selected the most successful sets of descriptors for polymorphic form prediction. PLS was less effective at identifying descriptors that could build a successful predictive model in an ANN.

This chapter will provide a short summary for each of the previous results chapters and then will present the overall conclusions of this work

### 10.1. Manual Analysis

An optimised model comprising of seven descriptors was determined using the manual data reduction techniques discussed in chapter 8. These seven descriptors were dsolv57, dsolv65, dsolv71, dsolv78, MNDO\_dipole, E\_vdw and gutmann donor number, which represent a mixture of solute and solvent properties (detailed in appendix section 12.2 and also in section 8.3).

These descriptors were determined by calculating the linear correlation coefficients of the whole dataset. All uncorrelated descriptors and commonly one descriptor from each correlated cluster were retained. However, as the highly correlated threshold was set at  $\pm 0.8 - 1$ , in one instance a very large cluster was generated. From this, 7 descriptors were retained. This reduced the descriptor set size to 40 parameters.

Analysis of these 40 descriptors was carried out using a variety of inspection techniques (discussed in chapter 5), eventually generating an optimised set of seven descriptors that could predict the polymorphic form of CBZ crystallisation experiments.

Cross validation analysis was carried out using 10 % of the data used in training. The model was rebuilt without these experiments and interrogated to predict the major polymorphic form obtained by crystallisation. This set of descriptors successfully predicted the major polymorphic form in 79 % of the experiments tested. When two unknown solvents were used as further validation of the model, seven out of twelve of the experimental products could be predicted.

The model created from these seven descriptors can successfully predict the major polymorphic form within the experimental space used in the training for a CBZ molecule.

## **10.2. Partial Least Squares Analysis**

The use of PLS to reduce the dataset was effective, but the identified descriptors did not build a predictive ANN model. It was noted in previous studies that PLS can be more successful with larger models, in which more importance is placed upon the information across the whole dataset and not upon individual variables<sup>[1]</sup>. Perhaps if the dataset had been much larger, PLS could have been a more effective tool.

## **10.3. Principal Component Analysis**

PCA scree plots efficiently identified tight sets of descriptors that covered most of the relevant parameter space. By using the loading values associated with each descriptor in a component, the most influential descriptors were identified. Analysis of the two most positively and negatively loaded descriptors in each component, highlighted as informative from the scree plot, were analysed in different combinations. Overall this descriptor selection method was highly effective and should plausibly be transferable to other polymorphic systems.

This PCA method of descriptor selection generated an optimised set of descriptors for CBZ polymorphic form prediction when a model was built in an ANN. The most successful model was built using ten descriptors, dsolv43, dsolv69, dsolv74, dsolv76,

d68, d69, d77, d84, E\_ang and boiling point. Similarly to the optimised set in the manual analysis work, there is a mixture of solvent and solute descriptors covering a range of different properties (discussed in chapter 7 section 7.3).

Cross validation analysis showed that this set of descriptors successfully predicted the major polymorphic form with 78 % success. However, when the external validation was carried out, the model was only able to predict the major polymorphic form crystallised on three occasions. This poor result suggested that perhaps the model needed to be optimised further, or that the external validation solvents poorly represented the range of descriptor values used in the training. A discussion of this was presented in sections 5.5.2 and 7.4.2, and clearly shows that ethyl acetate and n-butanol, the two validation solvents used, had very similar descriptor values and on a number of occasions their values were outside of the global range used in training.

The model created using PCA as a descriptor selection tool can predict the major polymorphic form of CBZ crystallisation experiments within the global range of descriptor values used within the training set.

## **10.4. Final Combined Optimisation**

As both the manual analysis technique and the PCA method both produced good predictive models, further analysis was carried out in order to optimise the descriptor set further.

Linear correlation coefficients between the two sets were calculated and by exchanging descriptors between the sets, an improved model was built.

An assessment of the predictive capabilities of all the descriptors in the two sets was made, with no improvement in prediction observed.

Classifying descriptors according to their physical meaning did not provide an avenue towards more reliable prediction of polymorphic outcomes, but suggested that many of the descriptors work in tandem to produce a successful ANN model.

Using knowledge gleaned from the poor external validation of the PCA set, those descriptors that were outside of the global range used in the training could be exchanged and removed in order to assess the effect on prediction. However, no improvements in prediction were observed. These improvements resulted in polymorphic form prediction

with 88.9 % success in cross validation experiments. External validation was less successful for similar reasons as explained in the context of the PCA strategy: the descriptor values of the solvent involved were very similar, except for those found outside the global range used in the training. This suggests that these were not the most suitable validation solvent candidates.

The most successful set of descriptors for CBZ polymorphic form prediction is dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and boiling point.

## **10.5. Overall Conclusions drawn from the investigation of the carbamazepine system**

Chapters 5, 6, 7 and 8 demonstrated the optimisation process that has established a set of CBZ and solvent descriptors that can predict the polymorphic form of CBZ in combination with an ANN. Ten descriptors have been informative, dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and boiling point, with their meanings discussed in section 8.3 and briefly in Table 10.1.

Table 10.1 Summary of the descriptors involved in the final set

Form predicted	Descriptor(s)	Definition(s)
I	Dsolv65 Rate Temperature	3D bonding information content (order 2) of the solvent molecule Rate of nitrogen blown onto sample (L/min) Temperature at which the crystallisations occurred
II	D69	Molecular surface area of the CBZ molecule
II	Rate	Rate of nitrogen blown onto sample (L/min)
II	D68	Moment of inertia C of the CBZ molecule
III	D84	FNSA-1, fractional partial negative surface area of the CBZ molecule (PNSA-1/total molecular surface area)
III	Boiling point D77	Literature value boiling point of the solvent molecule PNSA-3, atomic charge weighted partial negative surface area of the CBZ molecule
III	Rate	Rate of nitrogen blown onto sample (L/min)
III	D68	Moment of inertia C of the CBZ molecule
Dihydrate	Dsolv65 Rate Temperature Dsolv43	3D bonding information content (order 2) of the solvent molecule Rate of nitrogen blown onto sample (L/min) Temperature at which the crystallisations occurred 3D-Randić index (order 0) of the solvent molecule
Solvate	Boiling point Dsolv43	Literature value boiling point of the solvent molecule 3D-Randić index (order 0) of the solvent molecule
Not in a Rule	E_ang	Angle bend potential energy of the CBZ molecule
Not in a Rule	Dsolv76	PNSA-2, total charge weighted partial negative surface area of the solvent molecule
Not in a Rule	Dsolv69	Molecular surface area of the solvent molecule

Many of these descriptors describe the size, shape and charges on both the solvent and solute molecules, offering insight into the differing interactions between the molecules in solution. The successful optimisation of these descriptors based largely on mathematical selection processes rather than by using previously considered properties adds value to the model built. Previous literature offers many examples of how solvent-solute interactions are vital in polymorphic crystallisation<sup>[2-8]</sup>, which therefore strengthens the use of the descriptors in the final model.

From the ten descriptors featured in this final set, three of them were not mentioned in the rules created by FormRules.<sup>[9]</sup> These descriptors were E\_ang, dsolv69 and dsolv76. When detailed analysis was conducted, they were found to have an impact upon the

predictions of forms II and III. This work demonstrated how the rules generated in FormRules<sup>[9]</sup> can be used as a general guide for prediction, but that many of the descriptors work in tandem to create a successful prediction in INForm.<sup>[10]</sup>

The use of the ANN has been vital in the success of this predictive model. The relationships between the descriptors and the polymorphic outputs are not linear, and it is the highly intercorrelated nature of the descriptors that has led to the successful predictive models in the ANN.

Overall, there are a number of different conclusions that can be made from this research. An optimised set of descriptors for CBZ polymorph prediction has been generated. By using the same validation methods as in previous chapters, the major polymorphic form was predicted for 88.9 % of the cross validation experiments. The external solvent validation was again not successful. However, as has been discussed, perhaps the solvents were too similar and did not represent a variety of descriptor properties within the global range used in the training of the model.

This research has not only determined a set of descriptors for CBZ polymorph prediction, but also has discovered an effective method for rapid descriptor selection that can lead to a predictive model in an ANN.

By using PCA results, descriptors that can effectively predict the polymorphic form of crystallisation experiments when placed into an ANN can be uncovered. This was a highly effective descriptor selection technique, which was transferred to other polymorphic molecules (results presented in chapter 9).

## **10.6. Conclusions from analysis of TBA and ROY systems**

Both the set of descriptors determined in the CBZ analysis and also the PCA method of data reduction were tested on two different polymorphic systems, tolbutamide and 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (TBA and ROY respectively). This analysis was carried out in order to determine whether the descriptors that led to a successful CBZ predictive model were transferable to other polymorphic systems, or whether the descriptor selection method using PCA was transferable as a method.



The descriptor set optimised using CBZ was successfully applied to both TBA and ROY, generating overall average  $R^2$  values of 72.69 % for TBA and 83.19 % for ROY. Although these results are not as successful as the CBZ results (84.75 %), a predictive model can be created very quickly using this descriptor set. This descriptor set might create a model that is fit for purpose if a rough guide to the experimental space in which the most stable form would be crystallised was sought.

The use of PCA results to select descriptors that can lead to a successful predictive ANN model was very effective in the CBZ analysis. When all of the descriptor data for both TBA and ROY was subject to PCA, a similar selection technique was employed. The overall average  $R^2$  values generated in this analysis were higher than those observed in the analysis with the CBZ descriptors, at 87.83 % for TBA and 97.77 % for ROY. These results suggest that this is a highly effective method of descriptor selection, and when placed into an ANN, can predict the polymorphic form of a target molecule.

Only one common descriptor was observed in the sets for all three molecules, d84. However, when the descriptor meanings for these three sets were examined (Table 9.3), many similarities were displayed.

Both the set of descriptors determined in the CBZ analysis and also the PCA selection technique were successfully applied to different polymorphic target molecules. Ideally, more training data is required to make more firm conclusions. However, this analysis proved the concept of a transferable descriptor set, and also an effective selective method.

## 10.7. Summary of Conclusions

A predictive model for CBZ crystallisation experiments has been established. By reducing the number of descriptors using a variety of different techniques, an ANN model has been built that can predict the crystallisation product of CBZ experiments. Ten descriptors have been determined as the most important to this prediction, which are dsolv43, dsolv65, dsolv69, dsolv76, d68, d69, d77, d84, E\_ang and boiling point.

These descriptors appear to be transferable to other polymorphic target molecules and would allow a rapid predictive model to be built. The use of PCA to select specific descriptors for the prediction of a different polymorphic target molecule was also a

successful method. Although only carried out with two small datasets, the results show this selection technique to be very promising.

Ideally more experimental data for each polymorphic target molecule would improve the reliability of the model generated in the ANN. However, based upon the research presented in this thesis, the use of molecular and bulk descriptors related to both the solute and solvent molecules can predict the major polymorphic form of a crystallisation experiment.

- [1] V. E. Vinzi, W. W. Chin, J. Henseler, H. Wang, *Handbook of Partial Least Squares. Concepts, Methods and Applications*, Springer-verlag Berlin Heidelberg, **2010**.
- [2] S. Khoshkhoo, J. Anwar, *Journal of Physics D: Applied Physics* **1993**, 26, B90.
- [3] N. Blagden, R. J. Davey, H. F. Lieberman, L. Williams, R. Payne, R. Roberts, R. Rowe, R. Docherty, *Journal of the Chemical Society - Faraday Transactions* **1998**, 94, 1035.
- [4] P. T. Cardew, R. J. Davey, *Proceedings of the Royal Society A* **1985**, 398, 415.
- [5] R. J. Davey, J. Richards, *Journal of Crystal Growth* **1985**, 71, 597.
- [6] D. Musumeci, C. A. Hunter, J. F. McCabe, *Crystal Growth and Design* **2010**, 10, 1661–1664.
- [7] A. Spitaleri, C. A. Hunter, J. F. McCabe, M. J. Packer, S. L. Cockroft, *CrystEngComm* **2004**, 6, 489.
- [8] R. A. Chiarella, A. L. Gilon, R. C. Burton, R. J. Davey, G. Sadiq, A. Auffret, M. Cioffi, C. A. Hunter, *Faraday Discuss.* **2007**, 136, 179.
- [9] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [10] INForm, v3.7 ed., Intelligensys Ltd., **2009**.

## 11. FURTHER WORK

The presented research clearly indicates that the concept of combining calculated molecular descriptors with experimental crystallisation data is a viable avenue to the predictions of polymorphic outcomes with artificial neural network (ANN). To optimise the approach further and improve the reliability of the predictions, a number of additional steps could be taken.

Ideally, the user would want to build a model by doing as few experiments as possible and subsequently use the model to design their future experiments. In order to achieve this, a reliable set of descriptors needs to be established that is transferable between different polymorphic target molecules.

The descriptors set out by the carbamazepine (CBZ) analysis were successfully transferred to both 5-Methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY) and tolbutamide (TBA). However, a number of steps may be taken to ensure the most informative descriptors are transferred.

Obtaining more data for training this transferrable set is essential in order to gather as much information for use in the ANN as possible. The literature provides a number of examples of ANNs being trained with over 300 rows of data<sup>[1-4]</sup>, thus highlighting the potential limitations of the model presented in this thesis. However, its value should not be dismissed as there are also examples of much less data being used.<sup>[5-7]</sup>

With a larger dataset for training, the feature selection methods (PCA and PLS) would need to be reanalysed. Perhaps with a larger dataset the PLS analysis may be more successful at selecting descriptors that can lead to a predictive model. Ideally the availability of a second large dataset for a different polymorphic target molecule would allow the validation of the transferability of the model.

Generally, more datasets are needed to assess whether the ANN-based predictive method is universally transferrable.

An assessment of different theory levels used in the geometry optimisation in the initial stages of the analysis should be made. Perhaps the use of such high level calculations is not required, and would make the method more accessible to the user if the modelling process was simplified.

There are also many different components in an ANN. This research attempted to optimise the descriptors not the model architecture. Further work into the use of different transfer functions and different numbers of hidden layers would be highly valuable to determine if the model could be optimised further. The INForm<sup>[8]</sup> and FormRules<sup>[9]</sup> software does automatically determine an architecture, but refinements can be made by the user, and this process may lead to a more predictive model.

To make the overall predictive method more accessible to the user, the automation of the whole process from the initial molecular calculations to the ANN model would be highly beneficial. Integrated software for geometry optimisations, descriptor calculations, feature selection and ANN model could be very useful in drug and formulation development. With knowledge of only the molecular structure and a few experimental inputs, predictions could be made as to which solvents would crystallise the most stable form, and also generate an idea into the number of polymorphic forms that exist.

Currently the MOE<sup>[10]</sup> software would allow the geometry optimisation and descriptor calculations and the updated versions of INForm<sup>[8]</sup> has feature selection capabilities. A hybrid of these two pieces of software would allow the efficient determination of a predictive model.

The descriptors used in this research have been investigated based upon their meaning and their possible influence on the nucleation and crystallisation of polymorphic molecules. It would be beneficial to examine these descriptors in more detail and also interesting to assess whether the calculated descriptors correlate with experimental data. This would perhaps lead to more insight into the phenomenon of nucleation.

- [1] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, *113*, 9480.
- [2] R. C. Schweitzer, J. B. Morris, *Analytica Chimica Acta* **1999**, *384*, 285.
- [3] A. U. Bhat, S. S. Merchant, S. S. Bhagwat, *Industrial & Engineering Chemistry Research* **2008**, *47*, 920.
- [4] O. Engkvist, P. Wrede, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1247.
- [5] M. H. Fatemi, M. Jalali-Heravi, E. Konuze, *Analytica Chimica Acta* **2003**, *486*, 101.
- [6] M. de Matas, Q. Shao, V. L. Silkstone, H. Chrystyn, *Journal of Pharmaceutical Sciences* **2007**, *96*, 3293.
- [7] S. L. Wiskur, P. N. Floriano, E. V. Anslyn, J. T. McDevitt, *Angewandte Chemie, International Edition* **2003**, *42*, 2070.
- [8] INForm, v3.7 ed., Intelligensys Ltd., **2009**.
- [9] FormRules, v3.3 ed., Intelligensys Ltd., **2007**.
- [10] Chemical Computing Group, p. Molecular Operating Environment.

## 12. APPENDIX

### 12.1. CBZ Polymorph Screen Experimental Results

Table 12.1 CBZ polymorph screen experimental results

Crystallisation Solvent	Evaporation Rate (L/min of N <sub>2</sub> )	Crystallisation Temperature (°C)	Polymorphic Form Crystallised
Ethanol	5, 15	25, 50	Form III
Ethanol	25	25, 50	Mixture: Form II (50 %) and Form III (50 %)
Ethanol	25	75	Mixture: Form II (50 %) and Form III (50 %)
Ethanol	5	75	Mixture: Form II (0.02 %) and Form III (0.98 %)
Ethanol	15	75	Mixture: Form II (60 %) and Form III (40 %)
THF	5, 25	25, 50	Form III
THF	15	25	Form III
THF	15	50	Mixture: Form II (80 %) and Form III (20 %)
Acetone	5	25, 50	Form III
Acetone	25	25	Form III
Acetone	25	50	Form II
Acetone	15	25	Mixture: Form II (50 %) and Form III (50 %)
Acetone	15	50	Form III
Acetonitrile	5	25, 50, 75	Form III
Acetonitrile	25	25	Form III
Acetonitrile	25	50	Form II
Acetonitrile	15	25	Mixture: Form II (50 %) and Form III (50 %)
Acetonitrile	15	50	Form III
Acetonitrile	15	75	Mixture: Form II (10 %) and Form III (90 %)
Acetonitrile	25	75	Mixture: Form II (10 %) and Form III (90 %)

Crystallisation Solvent	Evaporation Rate (L/min of N <sub>2</sub> )	Crystallisation Temperature (°C)	Polymorphic Form Crystallised
Toluene	25	25, 50, 75	Form II
Toluene	5, 15	75	Form II
Toluene	5	50	Mixture: Form II (25 %) and Form III (75 %)
Toluene	15	50	Form III
Nitromethane	5	15, 50, 75	Form III
Nitromethane	25	25	Form III
Nitromethane	15	50, 75	Form III
Nitromethane	25	50	Form II
Nitromethane	25	75	Mixture: Form II (50 %) and Form III (50 %)
Methanol	5	25, 50	Form III
Methanol	25	25	Mixture: Form III (75 %) and Dihydrate (25 %)
Methanol	25	50	Mixture: Form I (50 %) and Form III (50 %)
Methanol	15	25	Mixture: Form III (50 %) and Dihydrate (50 %)
Methanol	15	50	Form III
DMSO	15,25	25, 50	Solvate
DMSO	5	50, 75	Solvate
DMSO	5	25	Mixture: Form III (50 %) and Solvate (50 %)
DMSO	15	75	Mixture: Form III (5 %) and Solvate (95 %)
DMSO	25	75	Solvate
Chloroform	5	25	Form II
Chloroform	5	50	Form III
Chloroform	25	25	Mixture: Form II (25 %) and Form III (75 %)
Chloroform	25	50	Mixture: Form II (50 %) and Form III (50 %)
Chloroform	15	25	Form III



Crystallisation Solvent	Evaporation Rate (L/min of N <sub>2</sub> )	Crystallisation Temperature (°C)	Polymorphic Form Crystallised
Chloroform	15	50	Mixture: Form II (90 %) and Form III (10 %)
Dichloromethane	5	25	Mixture: Form II (50 %) and Form III (50 %)
Dichloromethane	25	25	Form II
Dichloromethane	15	25	Form III
Chlorobenzene	25	25, 50	Form II
Chlorobenzene	5	25	Mixture: Form II (25 %) and Form III (75 %)
Chlorobenzene	5	50	Mixture: Form II (5 %) and Form III (95 %)
Chlorobenzene	15	25	Mixture: Form II (95 %) and Form III (5 %)
Chlorobenzene	5	75	Mixture: Form I (10 %) and Form II (90 %)
Chlorobenzene	15	75	Mixture: Form II (90 %) and Dihydrate (10 %)
Chlorobenzene	25	75	Form III
Cyclohexane	5, 25	75	Form II
Aniline	5, 25, 15	25, 50, 75	Form III

## 12.2. Molecular and Bulk Descriptor Meanings

**Rate** - The rate at which nitrogen was blown down onto the sample in L/min.

**Temperature** - The temperature at which the supersaturated solutions were created and the evaporations were carried out.

**D42 / dsolv42 -3D Wiener Index** - The Wiener Index (W)[1-3] is a topological descriptor, which was introduced in 1947 is defined by Equation 12.1.  $d_{ij}$  is the number of bonds between atoms  $i$  and  $j$  using the shortest path and  $N_{SA}$  represents the number of non-hydrogen atoms in the chosen molecule.[2-4] This descriptor essentially describes how compact the molecule is.

$$W = \frac{1}{2} \sum_{(i,j)}^{N_{SA}} d_{ij} \quad \text{Equation 12.1}$$

**D66, 67, 68 /dsolv 66, 67, 68 and pmiX, Y and Z** – The moment of inertia for the CBZ and solvent molecule were calculated in two different pieces of software. These descriptors are classed as geometrical descriptors<sup>[5]</sup> and are obtained from the mass and three-dimensional coordinates of atoms in the molecule. Using the rigid rotator approximation, the moments of inertia of a single molecule  $I_A$ ,  $I_B$  and  $I_C$  are calculated using Equation 12.2, Equation 12.3 and Equation 12.4, where  $I_C > I_B > I_A$ .<sup>[6]</sup>

$$I_A = \sum_i^n m_i r_{ix}^2 \quad \text{Equation 12.2}$$

$$I_B = \sum_i^n m_i r_{iy}^2 \quad \text{Equation 12.3}$$

$$I_C = \sum_i^n m_i r_{iz}^2 \quad \text{Equation 12.4}$$

The mass of each atom is represented by  $m_i$ , with  $r_{ix/y/z}$  denoting the distance between the  $i$ th atomic nucleus and the main rotational axes,  $x$ ,  $y$  and  $z$ . The number of atoms is represented by  $n$ .<sup>[6]</sup>

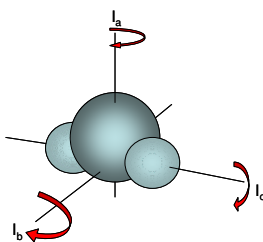


Figure 12.1 the axis of an single molecule, adapted from Atkins<sup>[6]</sup>

The moment of inertia is a measure of mass distribution in the molecule and also can determine how rotationally flexible parts of the molecule are.<sup>[7, 8]</sup>

**D69/ dsolv69** - Molecular surface area is another geometrical descriptor that uses the van der Waals radii of the atoms within the molecule to give the best surface area approximation.<sup>[1, 4, 9-11]</sup>

**D70/ dsolv70** - Molecular Volume is one of the most widely used geometrical descriptors.<sup>[12-14]</sup> Similarly to the surface area calculation, it calculates the volume of the overlapping spheres around the atoms of the molecule, based upon the van der Waals radii of the atoms.

**D71/ dsolv71 – TMSA** – The total molecular surface area is a charged partial surface area (CPSA) descriptor, but represented the total geometry of the molecule. The van der Waals radii of each atom within the molecule is represented by spheres that overlap with one another (Figure 12.2), creating a molecular surface<sup>[15]</sup>. In the case of TMSA, a solvent molecule, most commonly water with a van der Waals radius of 1.5 Å<sup>[15]</sup>, is used to trace a path around the molecule, generating a solvent accessible surface area (Figure 12.2). This solvent accessible surface area is used in the charged partial surface area (CPSA) calculations and is why TMSA belongs to the CPSA set of descriptors<sup>[15]</sup>.

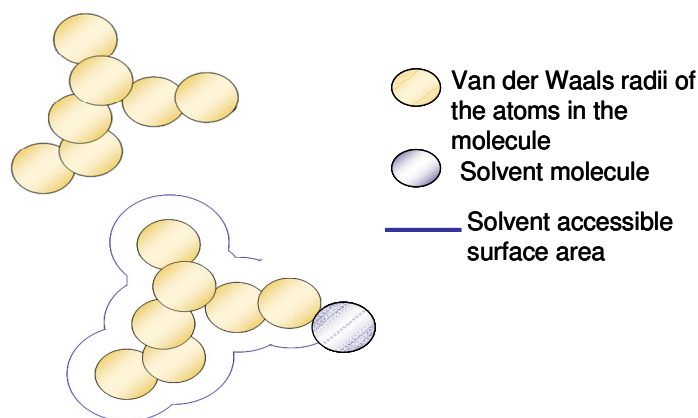


Figure 12.2 Calculation of the total molecular surface area using van der Waals radii, adapted from<sup>[15]</sup>

**D72 – PPSA-1** - Partial positive Surface is also a CPSA descriptor; it calculates the total positively charged surface area accessible by a solvent molecule ( $S_A$ ).<sup>[7, 15, 16]</sup>

$$PPSA1 = \sum_A S_A$$

$$A \in \{\delta_A > 0\}$$

Equation 12.5

**D73 – PPSA-2** – The total charge weighted partial positive surface area. This descriptor is similar to the PPSA-1 descriptor, but it includes the summation of positive atomic partial charges ( $q_A$ ) which has been calculated using a wave function, as shown in the equation below. [7, 15, 16]

$$PPSA2 = \sum_A q_A \cdot \sum_A S_A \quad \text{Equation 12.6}$$

**D74 – PPSA-3** - The atomic charge weighted partial positive surface area. Like the previous descriptor the positive atomic partial charges ( $q_A$ ) are involved in the calculation, but in this instance each individual partial charge is multiplied by the surface area ( $S_A$ ). [7, 15, 16]

$$PPSA3 = \sum_A q_A \cdot S_A \quad \text{Equation 12.7}$$

**D75 – PNSA-1** – The partial negative surface area, another CPSA descriptor, calculates the total negatively charged surface area accessible by a solvent molecule ( $S_A$ ). [7, 15, 16]

$$PNSA1 = \sum_A S_A \quad \text{Equation 12.8}$$

$$A \in \{\delta_A > 0\}$$

**D76 – PNSA-2** - The total charge weighted partial negative surface area. This descriptor is similar to the PNSA-1, but includes the summation of negative atomic partial charges ( $q_A$ ) which has been calculated using a wave function, as shown in the equation below. [7, 15, 16]

$$PNSA2 = \sum_A q_A \cdot \sum_A S_A \quad \text{Equation 12.9}$$

**D77 - PNSA-3** – The atomic charge weighted partial negative surface area. This descriptor is similar to the previous PNSAs, but in this instance the individual negative atomic partial charges are involved in the calculation are multiplied by the PNSA. [7, 15, 16]

$$PNSA3 = \sum_A q_A \cdot S_A \quad \text{Equation 12.10}$$

**D78 – DPSA-1** - The difference in charged partial surface areas. This descriptor calculates the differences between the positive and negative surface area, which may be useful to gain an idea of the polarity of the molecule.<sup>[7, 15, 16]</sup>

$$DPSA1 = PPSA1 - PNSA1 \quad \text{Equation 12.11}$$

**D79 - DPSA-2** - The difference in charged partial surface areas (PPSA2-PNSA2). PNSA-2 and PPSA-2 are total charge weighted partial surfaces areas, and it is the differences between the positive and negative surface areas.<sup>[7, 15, 16]</sup>

$$DPSA2 = PPSA2 - PNSA2 \quad \text{Equation 12.12}$$

**D80 –DPSA-3** – The difference in charged partial surface areas (PPSA3-PNSA3). PPSA-3 and PNSA-3 are the atomic charge weighted partial charged surface areas.<sup>[7, 15, 16]</sup>

$$DPSA3 = PPSA3 - PNSA3 \quad \text{Equation 12.13}$$

**D81– FPSA-1** - The fractional partial positive surface area is the ratio of the TMSA and the PPSA-1.<sup>[7, 15, 16]</sup>

$$FPSA1 = \frac{PPSA1}{TMSA} \quad \text{Equation 12.14}$$

**D82 – FPSA-2** - The fractional partial positive surface area of PPSA-2.<sup>[7, 15, 16]</sup>

$$FPSA2 = \frac{PPSA2}{TMSA} \quad \text{Equation 12.15}$$

**D83 –FPSA-3** - The fractional partial positive surface area of PPSA-3.<sup>[7, 15, 16]</sup>

$$FPSA3 = \frac{PPSA3}{TMSA} \quad \text{Equation 12.16}$$

**D84 - FNSA-1** - The fractional partial negative surface area is the ratio of the TMSA and the PNSA-1.<sup>[7, 15, 16]</sup>

$$FNSA1 = \frac{PNSA1}{TMSA} \quad \text{Equation 12.17}$$

**D85 – FNSA-2** - The fractional partial negative surface area of PNSA-2. [7, 15, 16]

$$FNSA2 = \frac{PNSA2}{TMSA} \quad \text{Equation 12.18}$$

**D86 – FNSA-3** - The fractional partial negative surface area of PNSA-3. [7, 15, 16]

$$FNSA3 = \frac{PNSA3}{TMSA} \quad \text{Equation 12.19}$$

**Dsolv2** – The number of carbon atoms in the solvent molecule

**Dsolv3** – The number of hydrogen atoms in the solvent molecule.

**Dsolv5** – The relative number of hydrogen atoms in the solvent molecule, which is the ratio between the number of hydrogen atoms and the total number of atoms in the molecule.

**Dsolv6** – The number of rings in the solvent molecule.

**Dsolv7** – The number of benzene rings in the solvent molecule.

**Dsolv8** – The total number of bonds in the solvent molecule.

**Dsolv9** – The total number of single bonds in the solvent molecule.

**Dsolv10** – The total number of double bonds in the solvent molecule.

**Dsolv13** – The relative number of single bonds in the solvent molecule, which is the ratio between the number of single bonds and the total number of bonds in the molecule.

**Dsolv14** – The relative number of double bonds in the solvent molecule, which is the ratio between the number of double bonds and the total number of bonds in the molecule.

**Dsolv16** – The relative number of aromatic bonds in the solvent molecule, which is the ratio of aromatic bonds to the total number of bonds in the solvent molecule.

**Dsolv17** – The relative molecular weight of the solvent molecule

**Dsolv20/21/22 – Randic index (order 1/2/3)** - The Randic index is a “second-generation topological”<sup>[7]</sup> descriptor, also known as the Randic molecular connectivity descriptor.<sup>[1, 17-21]</sup> It is the “sum over all pairs of edges,  $i$  and  $j$ , in the molecule”<sup>[7]</sup> “where  $D_i$  and  $D_j$  are the edge degrees”<sup>[7]</sup>.

$$\chi = \sum_{edges\,ij} (D_i D_j)^{-1/2} \quad \text{Equation 12.20}$$

To calculate the Randic indices of different orders the below equation is used.

$${}^m\chi = \sum_{path} (D_i D_j \dots D_k)^{-1/2} \quad \text{Equation 12.21}$$

**Dsolv24/25/26 - Kier & Hall index (order 1/2/3)** - This descriptor describes the valence connectivity of the molecule<sup>[20, 22, 23]</sup> and “accounts for the presence of heteroatoms and the hybridization of atoms in the molecule”.<sup>[7]</sup>  $Z_i^v$  is the number of valence electrons,  $Z_i$  is the total number of electrons in the  $i$ th atom and  $H_i$  is the number of hydrogen atoms attached to the  $i$ th atom (when  $I$  is nonhydrogen).

$$\delta_i^v = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1} \quad \text{Equation 12.22}$$

The different orders relate to the bond path, i.e.  $m=0, 1, 2$  etc. and are represented in the equation below.

$${}^m\chi^v = \sum_{i=1}^{N_s} \prod_{k=1}^{m+1} \left( \frac{1}{\delta_k^v} \right)^{1/2} \quad \text{Equation 12.23}$$

**Dsolv27/28 – Kier shape index (order 1/2)** - This descriptor is dependent on the number of atoms in the molecule and whether there are any branches. In the calculation of this descriptor, parameter  $\alpha$ , which is the ratio of atomic radius ( $r_i$ ) and the radius of the carbon in the  $sp^3$  hybridization state.<sup>[7]</sup>

$$\begin{aligned} {}^1\kappa &= (N_{SA} + \alpha)(N_{SA} + \alpha - 1)^2 ({}^1P + \alpha)^2 \\ {}^2\kappa &= (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 2)^2 ({}^2P + \alpha)^2 \\ {}^3\kappa &= (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 3)^2 ({}^3P + \alpha)^2 \dots\dots \text{if } N_{SA} \text{ is odd} \\ {}^4\kappa &= (N_{SA} + \alpha - 3)(N_{SA} + \alpha - 2)^2 ({}^3P + \alpha)^2 \dots\dots \text{if } N_{SA} \text{ is even} \end{aligned} \quad \begin{array}{l} \text{Equation} \\ 12.24 \end{array}$$

**Dsolv31/35 - Complementary Information Content (CIC) (order 0/1)** - “The  $r$ th order  $CIC_r$  measures the deviation of Information content ( $IC_r$ ) from its maximum value”.<sup>[4]</sup> In Equation 12.25,  $A$  is the atom number.<sup>[4, 24, 25]</sup>

$$CIC_r = \log_2 A - IC_r \quad \text{Equation 12.25}$$

**DSolv32/36/40 – Structural Information Content (SIC) (order 0/1/2)**- This is a topological descriptor that describes the neighbouring atoms in a molecule.<sup>[4, 7, 24, 25]</sup>

$${}^k SIC = {}^k IC / \log_2 n \quad \text{Equation 12.26}$$

**DSolv33/37/41 - Bonding Information content (BIC) (order 0/1/2)** - “The  $r$ th order  $BIC_r$  is defined in a normalized form as the  $SIC_r$  index, but taking into account the number of bonds and their multiplicity, where  $B$  is the number of bonds and  $\pi_b^*$  is the conventional bond order of the  $b$  bond.”<sup>[4, 24, 25]</sup>

$$BIC_r = \frac{Ic_r}{\log_2 \left( \sum_{b=1}^B \pi_b^* \right)} \quad \text{Equation 12.27}$$

**Dsolv34 –Information content (IC) (order 1)** - “the information content of a system having  $n$  elements is a measure of the degree of diversity of the elements in the set”

$$I_c = \sum_{g=1}^G n_g \log_2 n_g \quad \text{Equation 12.28}$$

“ $G$  is the number of different *equivalence classes* and  $n_g$  is the number of elements in the  $g$ th class”<sup>[4]</sup>

$$n = \sum_{g=1}^G n_g \quad \text{Equation 12.29}$$

**DSolv43/44/45/46 - 3D-Randic index (order 0/1/2/3)** - The Randic index is a “second-generation topological”<sup>[7]</sup> descriptor, also known as the Randic molecular connectivity descriptor. It is the “sum over all pairs of edges,  $i$  and  $j$ , in the molecule”<sup>[7]</sup> “where  $D_i$  and  $D_j$  are the edge degrees”<sup>[7]</sup>.

$$\chi = \sum_{edgesij} (D_i D_j)^{-1/2} \quad \text{Equation 12.30}$$

To calculate the Randic indices of different orders the below equation is used.

$${}^m \chi = \sum_{path} (D_i D_j \dots D_k)^{-1/2} \quad \text{Equation 12.31}$$

These two equations hold if the “continuous path of certain length,  $m > 1$ ”<sup>[7]</sup>, otherwise the below equation is used in the calculation of the Randic index.

$$\chi = \sum_{edgesij} (D_i D_j)^{-1/3} \quad \text{Equation 12.32}$$



**Dsolv47/48/49/50 – 3D-Kier and Hall index (order 0/1/2/3)** - This is a valence connectivity index (topological descriptor) that accounts “for the presence of heteroatoms and the hybridization of atoms in the molecule”<sup>[7]</sup>.

**Dsolv52 – 3D Kier Shape index (order 2)** - This descriptor ( $\kappa$ ) is dependent on the number of atoms in the molecule and whether there are any branches. In the calculation of this descriptor, parameter  $\alpha$  is the ratio of atomic radius ( $r_i$ ) and the radius of the carbon in the  $sp^3$  hybridization state.<sup>[7]</sup>

$$\begin{aligned} {}^1\kappa &= (N_{SA} + \alpha)(N_{SA} + \alpha - 1)^2 ({}^1P + \alpha)^2 \\ {}^2\kappa &= (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 2)^2 ({}^2P + \alpha)^2 \\ {}^3\kappa &= (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 3)^2 ({}^3P + \alpha)^2 \dots\dots \text{if } N_{SA} \text{ is odd} \\ {}^4\kappa &= (N_{SA} + \alpha - 3)(N_{SA} + \alpha - 2)^2 ({}^3P + \alpha)^2 \dots\dots \text{if } N_{SA} \text{ is even} \end{aligned} \quad \text{Equation 12.33}$$

**Dsolv55/59/63 - 3D-Complementary Information content (order 0/1/2)** - “The  $r$ th order  $CIC_r$  measures the deviation of  $IC_r$  from its maximum value”<sup>[4]</sup>  $A$  is the atom number. ”<sup>[4, 24-28]</sup>

$$CIC_r = \log_2 A - IC_r \quad \text{Equation 12.34}$$

**Dsolv56/60/64 -3D-Structural Information content (order 0/1/2)** - The structural information content descriptor is a topological descriptor and is an index based upon neighbouring atoms<sup>[4, 7, 24, 26, 27]</sup>

$${}^kSIC = {}^kIC / \log_2 n \quad \text{Equation 12.35}$$

This descriptor now considers the 3D structure of the solvent molecule.

**Dsolv57/61/65 - 3D-Bonding Information content (order 0/1/2)** - “The  $r$ th order  $BIC_r$  is defined in a normalized form as the  $SIC_r$  index, but taking into account the number of bonds and their multiplicity, where  $B$  is the number of bonds and  $\pi_b^*$  is the conventional bond order of the  $b$  bond.”<sup>[3, 4, 24-26, 29]</sup>

$$BIC_r = \frac{IC_r}{\log_2 \left( \sum_{b=1}^B \pi_b^* \right)} \quad \text{Equation 12.36}$$

**ASA** - Water accessible surface area calculated using a radius of 1.4 Å for the water molecule. A polyhedral representation is used for each atom in calculating the surface area.<sup>[10, 30]</sup>

**ASA\_H** - Water accessible surface area of all hydrophobic ( $|q_i| < 0.2$ ) atoms.<sup>[10, 30]</sup>

**E** - Value of the potential energy.<sup>[30]</sup>

**E\_ang** - Angle bend potential energy,<sup>[30]</sup> is a measure of deviation from the standard bond angles in the molecule.

**E\_ele** - Electrostatic component of the potential energy.<sup>[30]</sup>

**E\_vdw** - Van der Waals component of the potential energy.<sup>[30]</sup> The van der Waals term can be used to describe the interactions of solvent and solute molecules<sup>[31, 32]</sup>

**E\_nb** - Value of the potential energy with all bonded terms disabled.<sup>[30]</sup>

**E\_sol** - Solvation energy.<sup>[30, 33, 34]</sup>

**E\_strain** - Local strain energy.<sup>[30]</sup> The current energy minus the value of the energy at a near local minimum. The current energy is calculated as for the **E** descriptor. The local minimum energy is the value of the **E** descriptor after first performing an energy minimization. Current chirality is preserved and charges are left undisturbed during minimization.<sup>[30]</sup>

**Rgyr - Radius of gyration** - The radius of gyration is a “size descriptor for the distribution of atomic masses in a molecule”<sup>[4]</sup> given by Equation 12.36<sup>[4]</sup>

$$R_G = \sqrt{\frac{\sum_{i=1}^A m_i \cdot r_i^2}{MW}} \quad \text{Equation 12.37}$$

*MW* is the molecular weight, “*r<sub>i</sub>* is the distance of the *i*th atom from the centre of mass of the molecule, *m<sub>i</sub>* is the corresponding atomic mass”<sup>[4]</sup> and *A* is the atomic number.<sup>[4]</sup> This descriptor can also be calculated by using the moments of inertia for non-planar molecules as below.<sup>[4, 35]</sup>

$$R_G = \sqrt{\frac{2\pi \cdot (I_A \cdot I_B \cdot I_C)^{1/3}}{mw}} \quad \text{Equation 12.38}$$

**Pmi** - Principal moment of inertia.<sup>[36-38]</sup>

**AM1\_E** - The total energy (kcal/mol) calculated using the AM1 Hamiltonian.<sup>[30, 39, 40]</sup>

**AM1\_Eele** - The electronic energy (kcal/mol) calculated using the AM1 Hamiltonian.<sup>[30, 39, 40]</sup>

**AM1\_dipole** - The dipole moment calculated using the AM1 Hamiltonian.<sup>[30, 39, 40]</sup>

**AM1\_IP** - The ionization potential (kcal/mol) calculated using the AM1 Hamiltonian.<sup>[30, 39, 40]</sup>

**AM1\_HOMO** - The energy (eV) of the Highest Occupied Molecular Orbital calculated using the AM1 Hamiltonian.<sup>[30]</sup> This is a quantum chemical descriptor, which is the calculated energy of the highest occupies molecular orbital. When the HOMO value is high this means the molecule can donate its electrons more easily than lower HOMO values, therefore being more reactive. This descriptor is also related to ionisation potential, and also the nucleophilicity of the molecule.<sup>[4, 30, 39, 40]</sup>

**AM1\_LUMO** - The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the AM1 Hamiltonian.<sup>[30]</sup> This quantum chemical descriptor gives the energy of the lowest energy level not occupied by an electron in a molecule. This is related to the electron affinity (how electrophilic it is) of the molecule, with a lower LUMO value meaning it will more readily accept electrons.<sup>[6, 32, 40, 41]</sup>

**PM3\_IP** - The ionization potential (kcal/mol) calculated using the PM3 Hamiltonian.<sup>[30, 40, 42]</sup>

**PM3\_E** - The total energy (kcal/mol) calculated using the PM3 Hamiltonian.<sup>[30, 40, 42]</sup>

**PM3\_Eele** - The electronic energy (kcal/mol) calculated using the PM3 Hamiltonian.<sup>[30, 40, 42]</sup>

**PM3\_dipole** - The dipole moment calculated using the PM3 Hamiltonian.<sup>[30, 40, 42]</sup>

**PM3\_LUMO** - The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian.<sup>[30]</sup> This quantum chemical descriptor gives the energy of the lowest energy level not occupied by an electron in a molecule. This is related to the electron affinity (how electrophilic it is) of the molecule, with a lower LUMO value meaning it will more readily accept electrons.<sup>[6, 32, 40-42]</sup>

**PM3\_HOMO** - The energy (eV) of the Highest Occupied Molecular Orbital calculated using the PM3 Hamiltonian.<sup>[30]</sup> This is a quantum chemical descriptor, which is the calculated energy of the highest occupies molecular orbital. When the HOMO value is high this means the molecule can donate its electrons more easily than lower HOMO values, therefore being more reactive. This descriptor is also related to ionisation potential, and also the nucleophilicity of the molecule.<sup>[6, 32, 40-42]</sup>

**MNDO\_HOMO** - The energy (eV) of the Highest Occupied Molecular Orbital calculated using the MNDO Hamiltonian.<sup>[30]</sup> This is a quantum chemical descriptor, which is the calculated energy of the highest occupies molecular orbital. When the HOMO value is high this means the molecule can donate its electrons more easily

than lower HOMO values, therefore being more reactive. This descriptor is also related to ionisation potential, and also the nucleophilicity of the molecule.<sup>[4, 30, 39, 40]</sup>

**MNDO\_LUMO** - The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the MNDO Hamiltonian.<sup>[30]</sup> This quantum chemical descriptor gives the energy of the lowest energy level not occupied by an electron in a molecule. This is related to the electron affinity (how electrophilic it is) of the molecule, with a lower LUMO value meaning it will more readily accept electrons.  
[4, 30, 39, 40]

**MNDO-dipole** - The dipole moment calculated using the MNDO Hamiltonian.<sup>[30, 40]</sup>

**MNDO\_IP** - The ionization potential (kcal/mol) calculated using the MNDO Hamiltonian.<sup>[30, 40]</sup>

**MNDO\_E** - The total energy (kcal/mol) calculated using the MNDO Hamiltonian.  
[30, 40]

**MNDO\_Eele** - The electronic energy (kcal/mol) calculated using the MNDO Hamiltonian.<sup>[30, 40]</sup>

**Std\_dim1** – “Standard dimension 1: the square root of the largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.”<sup>[30]</sup>

**Std\_dim2** – “Standard dimension 2: the square root of the second largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.”<sup>[30]</sup>

**Std\_dim3** – “Standard dimension 3: the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.”<sup>[30]</sup>

**Dens** – “Mass density: molecular weight divided by van der Waals volume as calculated in the `vol` descriptor.”<sup>[30, 43]</sup>

**Glob** – “Globularity, or inverse condition number of the covariance matrix of atomic coordinates.” A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object.<sup>[30]</sup>

**Vol** - Van der Waals volume calculated using a grid approximation (spacing 0.75 Å).<sup>[11, 12, 30]</sup>

**VSA** – “The Subdivided Surface Areas are descriptors based on an approximate accessible van der Waals surface area calculation for each atom”<sup>[30]</sup>

**Dielectric Constant** - The dielectric constant is a measure of “the ability of a solvent to separate charge and to orient its dipoles.”<sup>[44]</sup> It represents the polarity of the molecule<sup>[45]</sup> and is calculated by placing the solvent between “two charged plates of a condenser” and measuring the electric field between the plates. This electric field measurement, divided by the permittivity of a vacuum gives the dielectric constant.<sup>[44]</sup>

**Dipole Moment** - When an electronically neutral molecule possesses “an unsymmetrical charge distribution”<sup>[44]</sup> it is said to have a permanent dipole moment.<sup>[8, 40]</sup>

**logP** – 1-octanol/water partition coefficient. This experimentally determined value has been correlated with biological activity<sup>[33]</sup> as it measures a substances affinity to two immiscible phases.<sup>[46]</sup>

**Boiling point (bp)** – “The temperature at which the saturated vapour pressure of a liquid equals the external atmospheric pressure”<sup>[46]</sup> Bubbles are formed within the liquid at this temperature, with the temperature remaining constant until all of the liquid has evaporated.<sup>[46, 47]</sup>

**Relative Molecular Mass (RMM)** – The summation of the molecular masses of all the atoms in a given solvent molecule<sup>[46, 48]</sup>

**Density** - Mass of a sample divided by its volume.<sup>[43, 46, 49]</sup>

**Refractive index** - The ratio of the speed of light in vacuum to that in a given medium. “The molar refractivity incorporates both the size and the polarizability of a molecule.”<sup>[50]</sup>

**Viscosity** - “A measure of the resistance to flow that a fluid offers when it is subjected to shear stress.”<sup>[46]</sup> Viscosity has been used in predictive models as a descriptor<sup>[49]</sup>, and also been predicted itself.<sup>[48]</sup> It is also very dependant on temperature.<sup>[51]</sup>

**Surface tension** – The amount of work that has to be applied to increase the solvents surface area by one unit.<sup>[51]</sup> The surface tension is the property of the solvent that makes the liquid appear to have a skin.<sup>[46]</sup> Surface tension has been used in predictive research and also in more practical crystallisation work.<sup>[23, 52]</sup>

**Hildebrand Solubility Parameter** - A parameter measuring the cohesion of a solvent (energy required to create a cavity in the solvent).<sup>[44, 53, 54]</sup>

**Freezing point** – The temperature at which a solution becomes a solid.<sup>[6]</sup>

**Vapour density** - “The density of a gas or vapour relative to hydrogen, oxygen, or air. Taking hydrogen as the reference substance, the vapour density is the ratio of the mass of a particular volume of a gas to the mass of an equal volume of hydrogen under identical conditions of pressure and temperature.”<sup>[46]</sup>

**Solubility** - “The quantity of solute that dissolves in a given quantity of solvent to form a saturated solution.”<sup>[46]</sup> Solubility features in both experimental<sup>[55-58]</sup> and predictive work.<sup>[1, 13, 27, 59]</sup> This descriptor was experimentally determined in this research for each of the target molecules in the different crystallisation solvents.

**Activity** - “A thermodynamic function used in place of concentrations (or pressures) corrected for non-ideal behaviour.”<sup>[46]</sup> This descriptor was experimentally determined in this research using the solubility values. Activity features in much of the crystallisation literature.<sup>[60-62]</sup>

**Vapour pressure** – The pressure exerted by a vapour that has been given off from the solvent in this instance.<sup>[46]</sup> Atoms or molecules have evaporated off the liquid and exert a pressure. Equilibrium is reached between the escaping and re-entering vapour molecules.<sup>[46]</sup>

**Polarity Parameter ( $E_T(30)$ )** – Dimroth and Reichardt’s polarity parameter represents the polarity of the solvent molecule.<sup>[51]</sup>

**Hansen Solubility Parameters – dD, dP, dH** - “Hansen solubility parameters (HSP) are widely used to correlate and predict the behaviour of solvents.”<sup>[63]</sup> The cohesion energy of a liquid can be divided into nonpolar atomic interactions (dispersion), dD, the permanent dipole-dipole interactions, dP, and hydrogen bonding interactions, dH.<sup>[63]</sup>

**Henry’s law constant** – “At a constant temperature the mass of gas dissolved in a liquid at equilibrium is proportional to the partial pressure of the gas”<sup>[46]</sup> outside of the solution.

**Gutmann donor number** - The Gutmann donor number (DN) is a bulk solvent descriptor and quantifies the basicity or electron donating ability of a solvent.<sup>[64-66]</sup> It is based upon solute-solvent interactions interacting like acid-base reactions<sup>[64]</sup> and was defined by Gutmann “as the negative  $\Delta H$  value in kcal/mol for the interaction of the electron pair donor solvent with  $SbCl_5$  in a highly diluted solution of dichloroethane”.<sup>[64]</sup>

**Gutmann acceptor number** - The acceptor number (AN) describes the electrophilic behaviour of the solvent.<sup>[64]</sup> Both Gutmann DN and AN have been used in the literature<sup>[65-69]</sup>

### 12.3. Tolbutamide Stability XRPD Traces

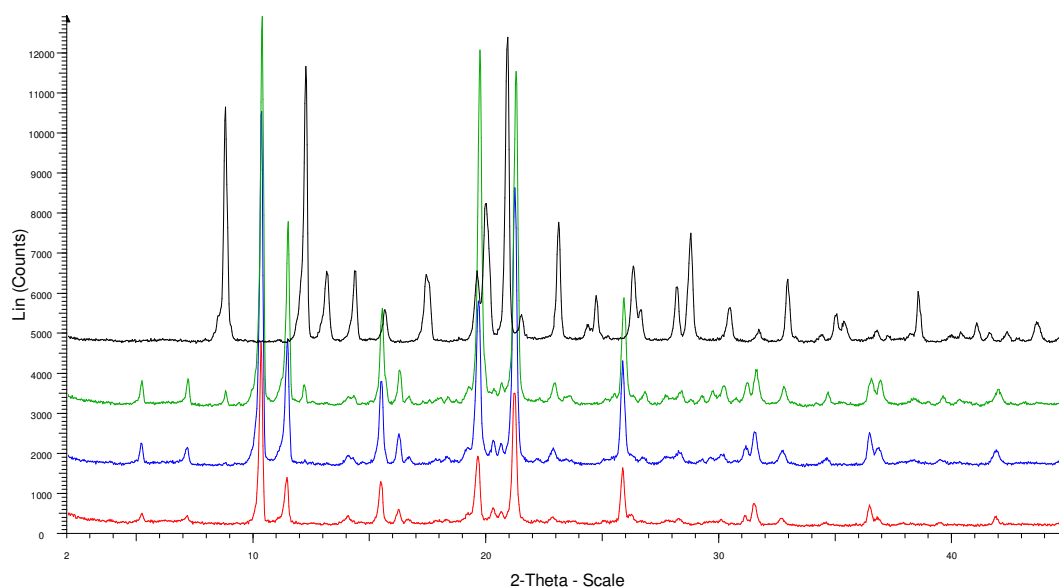


Figure 12.3 The methanol slurry to measure the stability of TBA. Commercial form I (black), two samples of the MeOH slurry after 2 days (red and green) and the MeOH slurry after 7 days (blue)

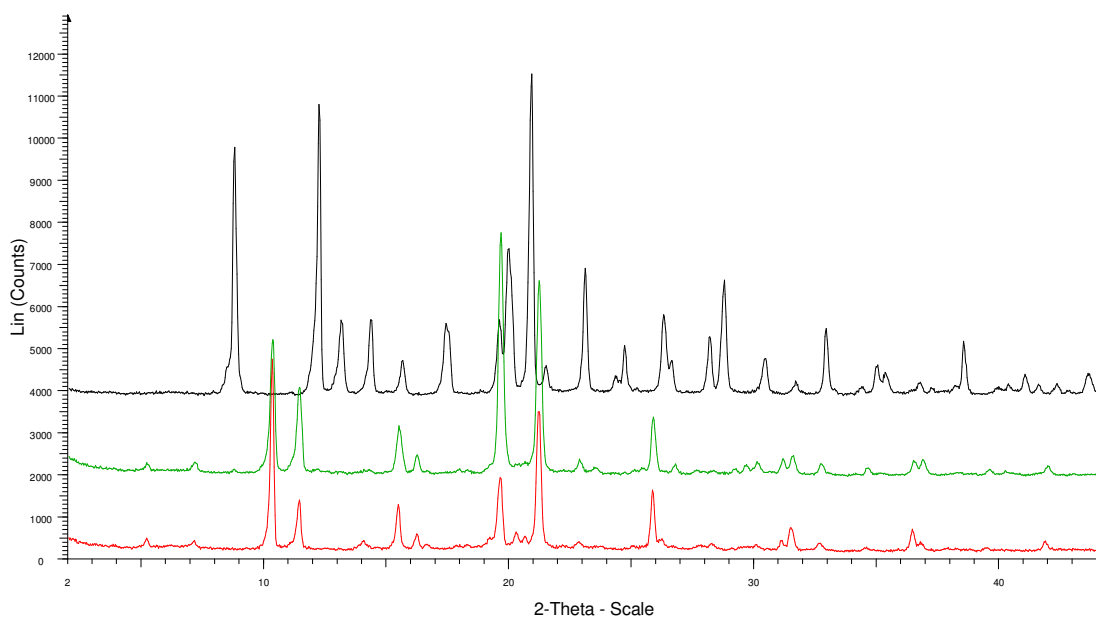


Figure 12.4 The ethanol slurry to measure the stability of TBA. Commercial form I (black), the EtOH slurry after 2 days (green) and the MeOH slurry after 2 days (red) for comparison

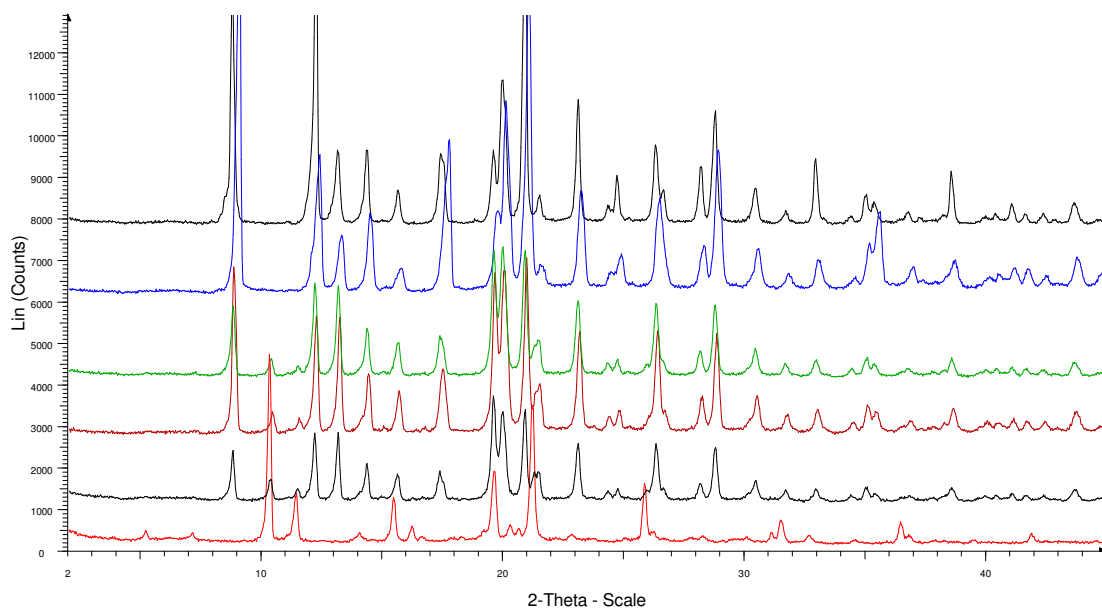


Figure 12.5 The dichloromethane slurry to measure the stability of TBA. Commercial form I (black), the DCM slurry after 2 days (blue), the DCM slurry seeded with form II after 2 days (green), the DCM slurry seeded with form II after 3 days (brown), the DCM slurry seeded with form II after 5 days (black trace above the red) and the MeOH slurry after 2 days (red) for comparison

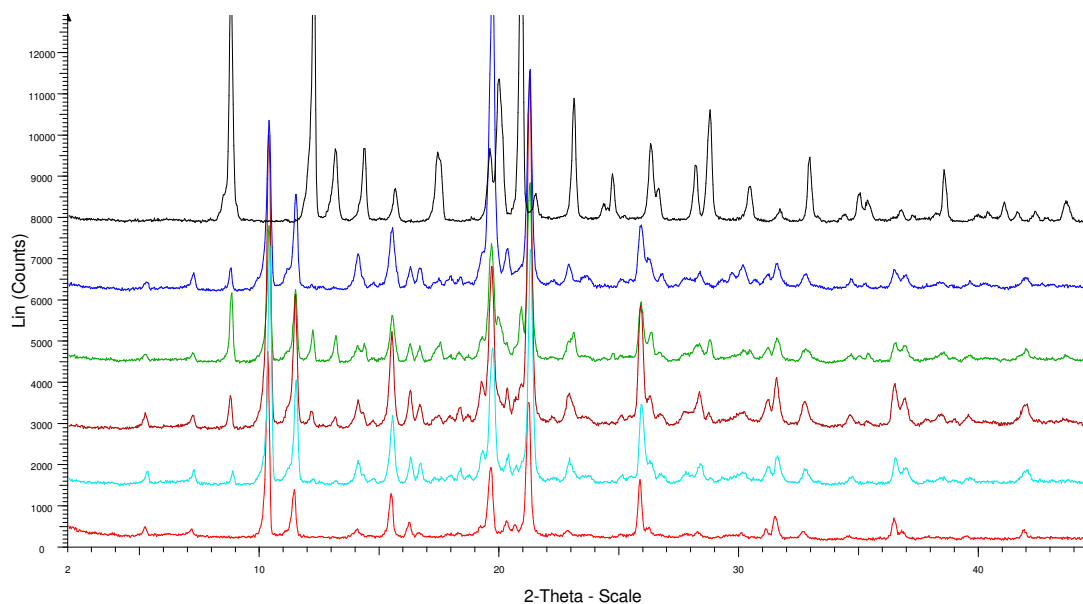


Figure 12.6 The acetone slurry to measure the stability of TBA. Commercial form I (black), the acetone slurry after 2 days (blue), the acetone slurry after 7 days (green), the acetone slurry after 8 days (brown), the acetone slurry after 10 days (light blue), and the MeOH slurry after 2 days (red) for comparison

The raw XRPD data can be found in electronic appendix, chapter 3, file 3.1.



## 12.4. CBZ OPLS Forcefield Results

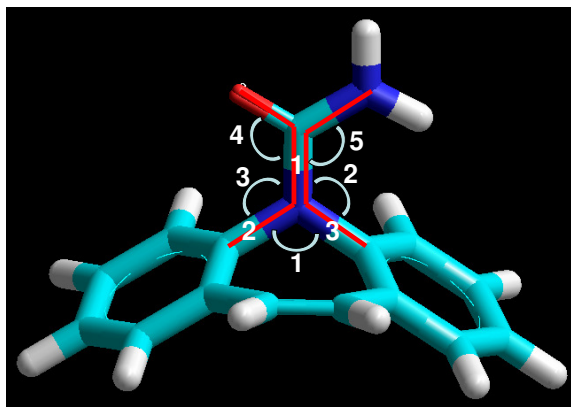


Figure 12.7 Assignments of CBZ bond lengths and angles

Table 12.2 Comparison of OPLS geometry optimised CBZ

	CBZ OPLS	Literature values
<b>Bond length 1 (Å)</b>	1.35	1.38
<b>Bond length 2 (Å)</b>	1.411	1.437
<b>Bond length 3 (Å)</b>	1.411	1.434
<b>Bond angle 1 (°)</b>	115.8	116.8
<b>Bond angle 2 (°)</b>	123.2	121.9
<b>Bond angle 3 (°)</b>	120.5	120.9
<b>Bond angle 4 (°)</b>	120.8	121.4
<b>Bond angle 5 (°)</b>	120.8	116.0
<b>Torsion angle C-O (°)</b>	-4.1	-9.1
<b>Torsion angle C-N (°)</b>	2.7	-2.2

## 12.5. ROY PM3 Conformational Search Results

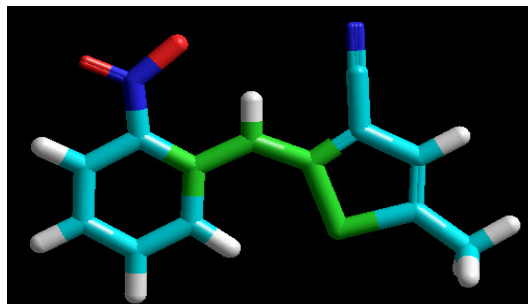


Figure 12.8 The torsion angle used in the ROY conformational analysis

Table 12.3 The Conformations of the ROY molecule and their associated energy

Torsion angle (°)	Energy (kcal/mol)
74.44586	-64934.9
-72.5912	-64934.9
-116.809	-64934.9
116.5048	-64934.9
98.11012	-64934.8
100.0112	-64934.6
-100.03	-64934.6
25.5403	-64934.1
-25.5635	-64934.1
42.18503	-64933.9
-40.3768	-64933.9
1.66E-05	-64933.8
125.668	-64929.5

## 12.6. Sets of Descriptors used in the Manual Analysis Work

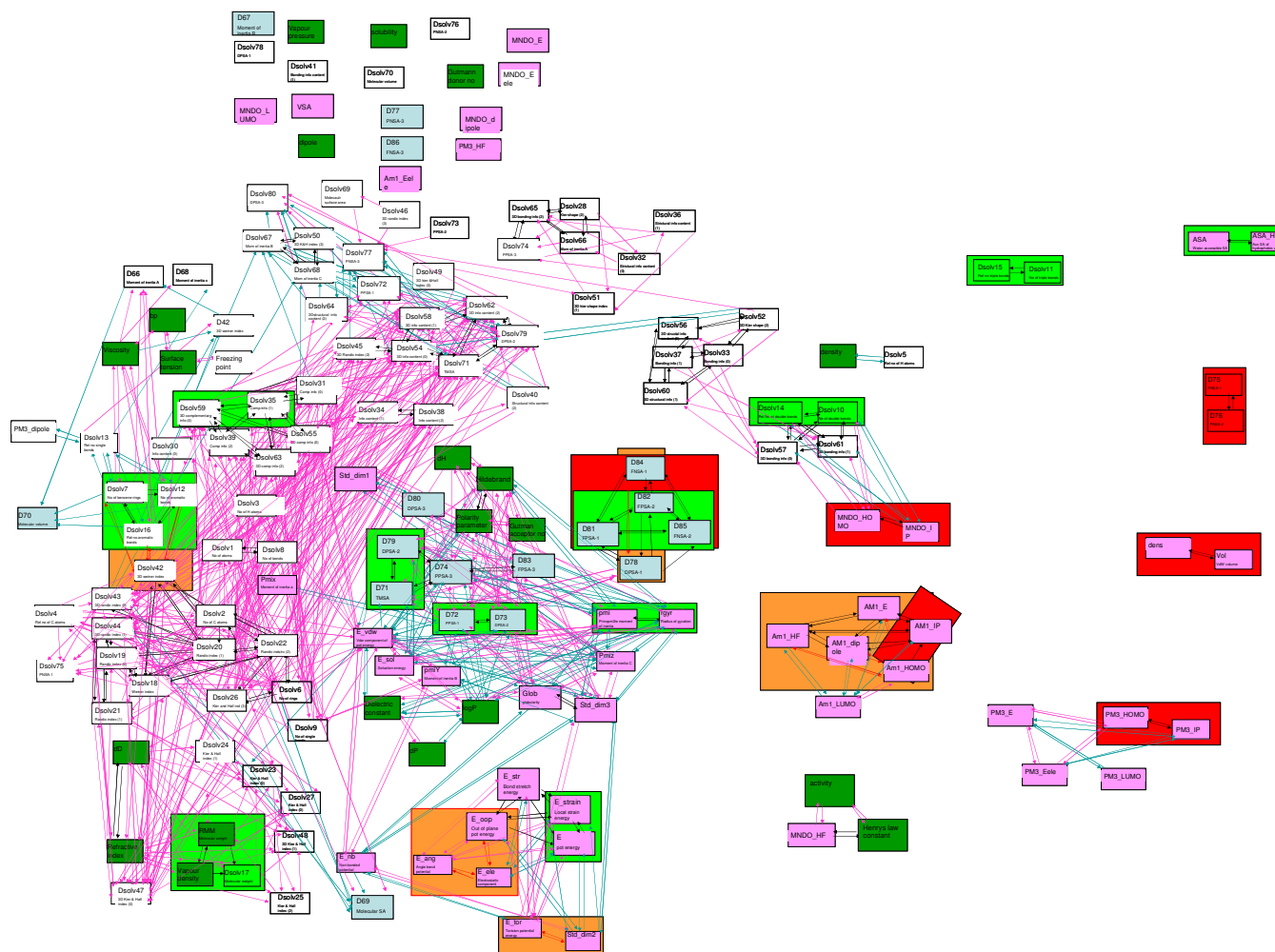
Table 12.4 The descriptor sets used in the manual analysis

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
D74	dsolv26	d83	d42	d67	d75
dsolv3	dsolv34	dsolv1	d78	d71	d79
dsolv5	dsolv47	dsolv39	d81	d77	dsolv10
dsolv40	dsolv63	dsolv60	dsolv13	dsolv24	dsolv20
dsolv42	dsolv68	MNDO_E	dsolv25	dsolv32	dsolv46
dsolv58	dsolv69	E_ele	dsolv30	dsolv62	dsolv50
pmiY	dsolv71	E	dsolv49	AM1_HOMO	dsolv75
std_dim1	dsolv74	AM1_dipole	dsolv59	PM3_LUMO	MNDO_dipole
AM1_HF	dsolv79	b.p.	dsolv76	pmiZ	Vol
ASA	RMM	dP	Polarity Parameter (ET(30))	Dielectric Constant	Refractive index

Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
d70	d68	dsolv22	d66	d72	Dsolv4
dsolv31	d69	dsolv28	d76	d84	Dsolv14
dsolv66	d85	dsolv73	d80	dsolv8	Dsolv33
dsolv80	dsolv6	E_nb	d82	dsolv16	Dsolv52
E_ang	dsolv7	E_oop	dsolv18	dsolv57	Dsolv54
E_strain	dsolv23	PM3_IP	dsolv37	dsolv64	Dsolv65
E_tor	dsolv44	rgyr	AM1_Eele	dsolv70	Dsolv77
glob	Dsolv55	Vapor Pressure (kPa)	MNDO_ HOMO	std_dim3	AM1_LUMO
ASA_H	VSA	activity	PM3_E	Freezing Point (°C)	E_str
PM3_Eele	Viscosity (cP)	LogP	Dipole (D)	Gutmann donor no.	Henry's Law Constant (atm m3/mol)
			Gutman acceptor no.	Surface tension (mN/m)	dH

Set 13	Set 14	Set 15	Set 16
d86	d73	dsolv9	dsolv15
dsolv12	dsolv2	dsolv27	dsolv38
dsolv17	dsolv11	dsolv36	dsolv48
dsolv19	dsolv35	dsolv43	dsolv72
dsolv21	dsolv41	dsolv56	MNDO_IP
dsolv45	dsolv67	AM1_E	MNDO_LU MO
dsolv51	E_vdw	AM1_IP	PM3_HF
dsolv61	MNDO_Eele	PM3_dipole	Pmi
dsolv78	MNDO_HF	std_dim2	pmiX
dens	PM3_HOMO	Hildebrand (cal <sup>1/2</sup> cm <sup>-3/2</sup> )	Vapor Density
E_sol	dD	Solubility	density (g/cm3)

## 12.7. Linear Correlation Between the Descriptors - Schematic



## 12.8. Rules of Form II Loading Value Analysis

Rules for CBZ prediction using PLS results			
--- Rules for property Form I ---			
	IF rate is LOW	THEN Form I is	LOW (1.00)
	IF rate is HIGH	THEN Form I is	LOW (0.97)
--- Rules for property Form II ---			
SubModel:1	IF d78 is LOW	THEN Form II is	HIGH (1.00)
	IF d78 is HIGH	THEN Form II is	LOW (1.00)
SubModel:2	IF rate is LOW	THEN Form II is	LOW (0.89)
	IF rate is HIGH	THEN Form II is	HIGH (0.79)
SubModel:3	IF pmiY is LOW	THEN Form II is	LOW (1.00)
	IF pmiY is HIGH	THEN Form II is	HIGH (1.00)
--- Rules for property Form III ---			
SubModel:1	IF d78 is LOW	THEN Form III is	LOW (1.00)
	IF d78 is HIGH	THEN Form III is	HIGH (1.00)
SubModel:2	IF Viscosity (cP) is LOW	THEN Form III is	HIGH (1.00)
	IF Viscosity (cP) is MID	THEN Form III is	LOW (1.00)
	IF Viscosity (cP) is HIGH	THEN Form III is	HIGH (0.85)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.83)
	IF rate is HIGH	THEN Form III is	LOW (0.97)
--- Rules for property DiHydrate ---			
	IF Temp is LOW	THEN DiHydrate is	LOW (0.96)
	IF Temp is HIGH	THEN DiHydrate is	LOW (1.00)
--- Rules for property Solvate ---			
SubModel:1	IF Viscosity (cP) is LOW AND pmiY is LOW AND d78 is LOW	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is LOW AND pmiY is LOW AND d78 is HIGH	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is LOW AND pmiY is MID AND d78 is LOW	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is LOW AND pmiY is MID AND d78 is HIGH	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is LOW AND pmiY is HIGH AND d78 is LOW	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is LOW AND pmiY is HIGH AND d78 is HIGH	THEN Solvate is	HIGH (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is LOW AND d78 is LOW	THEN Solvate is	HIGH (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is LOW AND d78 is HIGH	THEN Solvate is	HIGH (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is MID AND d78 is LOW	THEN Solvate is	HIGH (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is MID AND d78 is HIGH	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is HIGH AND d78 is LOW	THEN Solvate is	LOW (1.00)
	IF Viscosity (cP) is HIGH AND pmiY is HIGH AND d78 is HIGH	THEN Solvate is	HIGH (1.00)
SubModel:2	IF d42 is LOW	THEN Solvate is	HIGH (1.00)
	IF d42 is HIGH	THEN Solvate is	HIGH (1.00)

## 12.9. PCA Analysis – Plots of Molecular Surface Area against Bulk Properties

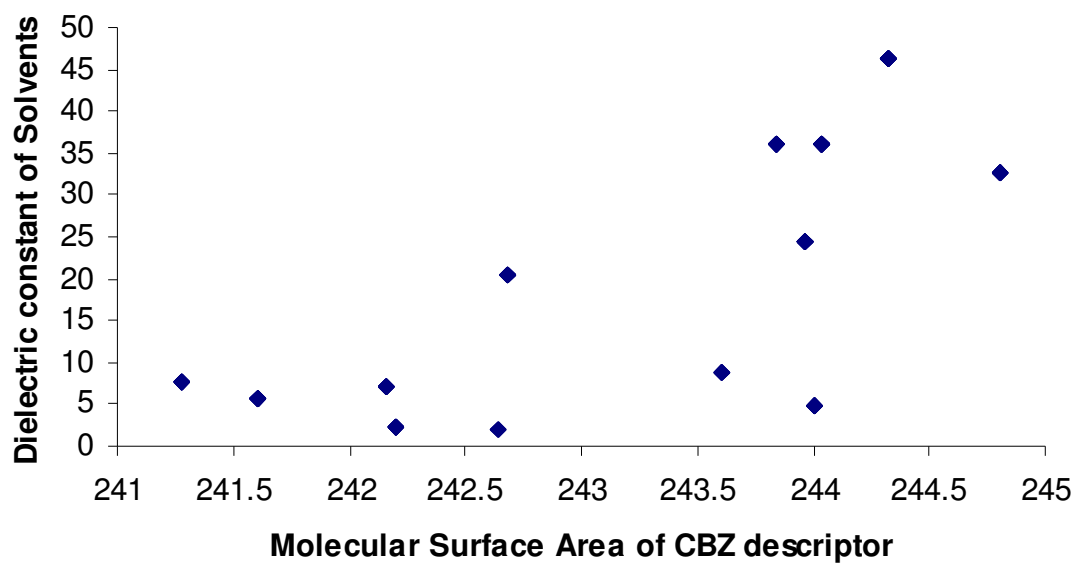


Figure 12.9 Plot of molecular surface area descriptor (d69) against the dielectric constant of the solvents

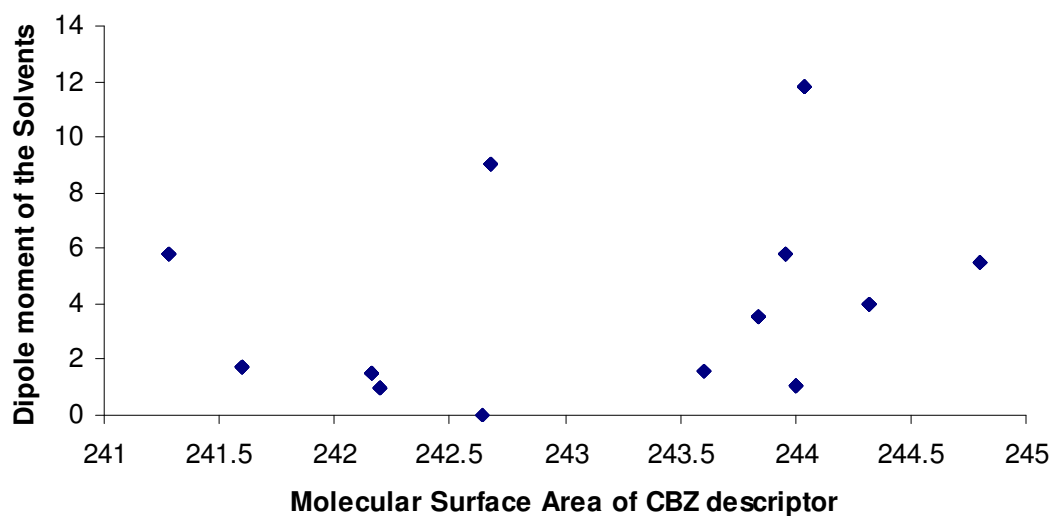


Figure 12.10 Plot of molecular surface area descriptor (d69) against the dipole moment of the solvents

## 12.10. Final CBZ Analysis – Rules of Opt. E Descriptor Set

Table 12.5 Rules from the optimised CBZ descriptor set

Rules for CBZ prediction using the final optimized descriptor set			
<b>--- Rules for property Form I ---</b>			
IF dsolv65 is LOW AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is LOW AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (0.90)	
IF dsolv65 is MID AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is MID AND rate is HIGH AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is LOW AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is LOW AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is MID AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is MID AND Temp is HIGH	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is HIGH AND Temp is LOW	THEN Form I is	LOW (1.00)	
IF dsolv65 is HIGH AND rate is HIGH AND Temp is HIGH	THEN Form I is	HIGH (1.00)	
<b>--- Rules for property Form II ---</b>			
SubModel:1	IF d69 is LOW	THEN Form II is	HIGH (1.00)
	IF d69 is HIGH	THEN Form II is	LOW (1.00)
SubModel:2	IF rate is LOW	THEN Form II is	LOW (1.00)
	IF rate is HIGH	THEN Form II is	HIGH (0.64)
SubModel:3	IF d68 is LOW	THEN Form II is	HIGH (1.00)
	IF d68 is HIGH	THEN Form II is	LOW (1.00)
<b>--- Rules for property Form III ---</b>			
SubModel:1	IF d84 is LOW	THEN Form III is	HIGH (1.00)
	IF d84 is HIGH	THEN Form III is	LOW (1.00)
SubModel:2	IF b.p. is LOW AND d77 is LOW	THEN Form III is	HIGH (0.50)
	IF b.p. is LOW AND d77 is HIGH	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is LOW	THEN Form III is	LOW (1.00)
	IF b.p. is HIGH AND d77 is HIGH	THEN Form III is	HIGH (1.00)
SubModel:3	IF rate is LOW	THEN Form III is	HIGH (0.72)
	IF rate is HIGH	THEN Form III is	LOW (1.00)
SubModel:4	IF d68 is LOW	THEN Form III is	HIGH (1.00)
	IF d68 is MID	THEN Form III is	HIGH (1.00)
	IF d68 is HIGH	THEN Form III is	LOW (1.00)
<b>--- Rules for property DiHydrate ---</b>			
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is LOW AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	

Rules for CBZ prediction using the final optimized descriptor set –cont.			
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is MID AND dsolv43 is HIGH	THEN DiHydrate is	LOW (0.88)	
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is LOW AND rate is HIGH AND dsolv43 is HIGH	THEN DiHydrate is	LOW (0.95)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is LOW AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is MID AND dsolv43 is HIGH	THEN DiHydrate is	LOW (0.50)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv43 is LOW	THEN DiHydrate is	LOW (0.99)	
IF dsolv65 is LOW AND Temp is HIGH AND rate is HIGH AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is LOW AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv43 is LOW	THEN DiHydrate is	HIGH (0.99)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is MID AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv43 is LOW	THEN DiHydrate is	LOW (0.50)	
IF dsolv65 is HIGH AND Temp is LOW AND rate is HIGH AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is LOW AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is MID AND dsolv43 is HIGH	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv43 is LOW	THEN DiHydrate is	LOW (1.00)	
IF dsolv65 is HIGH AND Temp is HIGH AND rate is HIGH AND dsolv43 is HIGH	THEN DiHydrate is	LOW (0.90)	
--- Rules for property Solvate ---			
IF b.p. is LOW AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)	
IF b.p. is LOW AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	
IF b.p. is MID AND dsolv43 is LOW	THEN Solvate is	LOW (1.00)	
IF b.p. is MID AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	
IF b.p. is HIGH AND dsolv43 is LOW	THEN Solvate is	HIGH (1.00)	
IF b.p. is HIGH AND dsolv43 is HIGH	THEN Solvate is	LOW (1.00)	



- [1] G. L. Amidon, S. T. Anik, *Journal of Pharmaceutical Sciences* **1976**, 65, 801.
- [2] B. Bogdanov, S. Nikolic, N. trinajstic, *Journal of Mathematical Chemistry* **1989**, 3, 299.
- [3] J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, Amsterdam, **1999**.
- [4] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Vol. 11, first ed., Wiley-VCH, Weinheim, **2000**.
- [5] R. C. Schweitzer, J. B. Morris, *Analytica Chimica Acta* **1999**, 384, 285.
- [6] P. Atkins, J. de Paula, *Atkins' Physical Chemistry*, 7th ed., Oxford University Press, Oxford, **2002**.
- [7] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, First ed., John Wiley & Sons, Inc., New York, **2000**.
- [8] G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, *Molecular Diversity* **2009**, 13, 301.
- [9] B. Lee, F. M. Richards, *Journal of Molecular Biology* **1971**, 55, 379.
- [10] C. Chothia, *Nature* **1974**, 248, 338.
- [11] M. Alleso, F. Van Den Berg, C. Cornett, F. S. Jorgensen, B. Halling-Sorensen, H. Lopez De Diego, L. Hovgaard, J. Aaltonen, J. Rantanen, *Journal of Pharmaceutical Sciences* **2008**, 97, 2145.
- [12] A. J. Hopfinger, *Journal of the American Chemical Society* **1980**, 102, 7196.
- [13] N. Bodor, A. Harget, M.-J. Huang, *Journal of the American Chemical Society* **1991**, 113, 9480.
- [14] H. Golmohammadi, *Journal of Computational Chemistry* **2009**, 30, 2455.
- [15] D. T. Stanton, P. C. Jurs, *Analytical Chemistry* **1990**, 62, 2323.
- [16] M. H. Fatemi, F. Karimian, *Journal of Colloid and Interface science* **2007**, 314, 665.
- [17] M. Randic, *Journal of the American Chemical Society* **1975**, 97, 6609.
- [18] M. Randic, G. M. Brissey, R. B. Spencer, C. L. Wilkins, *Computers & Chemistry* **1980**, 4, 27.
- [19] M. Randic, *Journal of Chemical Information and Computer Sciences* **1984**, 24, 164.
- [20] M. Randic, *Journal of Chemical Information and Computer Sciences* **1997**, 37, 672.
- [21] H. Liu, M. Lu, F. Tian, *Journal of Mathematical Chemistry* **2005**, 38, 345.
- [22] L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Vol. 14, 1st ed., Academic Press, Inc., New York, **1976**.
- [23] J.-P. Luiu, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, 55, 41.
- [24] S. C. Basak, A. T. Balaban, G. D. Grunwald, B. D. Gute, *Journal of Chemical Information and Computer Sciences* **2000**, 40, 891.
- [25] S. C. Basak, D. K. Harriss, V. R. Magnuson, *Journal of Pharmaceutical Sciences* **1984**, 73, 429.
- [26] R. Todeschini, R. Cazar, E. Collina, *Chemometrics and Intelligent Laboratory Systems* **1992**, 15, 51.
- [27] A. R. Katritzky, D. B. Tatham, *Journal of Chemical Information and Computer Sciences* **2001**, 41, 358.
- [28] B. Louis, V. K. Agrawal, P. V. Khadikar, *European Journal of Medicinal Chemistry* **2010**, 45, 4018.
- [29] D. M. Eike, J. F. Brennecke, E. J. Maginn, *Green Chemistry* **2003**, 5, 323.

- [30] <http://www.chemcomp.com/journal/descr.htm>, *The Chemical Computing Group Viewed on 29/04/11*.
- [31] H. Hu, W. Yang, *Journal of Physical Chemistry B* **2010**, *114*, 2755.
- [32] A. R. Katritzky, D. C. Fara, M. Kuanar, E. Hur, M. Karelson, *Journal of Physical Chemistry A* **2005**, *109*, 10323.
- [33] H. Chuman, A. Mori, H. Tanaka, *Analytical Sciences* **2002**, *18*, 1015.
- [34] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tine, *Journal of Chemical Information and Modeling* **2006**, *46*, 2030.
- [35] S. Izrailev, F. Zhu, D. K. Agrafiotis, *Journal of Computational Chemistry* **2006**, 1962.
- [36] R. H. Rohrbaugh, P. C. Jurs, *Analytical Chemistry* **1985**, *57*, 2770.
- [37] E. R. Collantes, W. Tong, W. J. Welsh, *Analytical Chemistry* **1996**, *68*, 2038.
- [38] T. P. Knowles, A. W. Fitzpatrick, S. Meehan, H. R. Mott, M. Vendruscolo, C. M. Dobson, M. E. Welland, *Science* **2007**, *318*, 1900.
- [39] M. J. S. Dewar, E. G. Zebisch, E. F. Healy, J. J. P. Stewart, *Journal of the American Chemical Society* **1985**, *107*, 3902.
- [40] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chemical Reviews* **1996**, *96*, 1027.
- [41] A. P. Harding, D. C. Wedge, P. L. A. Popelier, *Journal of Chemical Information and Modeling* **2009**, *49*, 1914.
- [42] J. J. P. Stewart, *Journal of Computational Chemistry* **1989**, *10*, 221.
- [43] T. L. Therfall, *Analyst* **1995**, *120*, 2435.
- [44] C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry*, Second ed., VCH Verlagsgesellschaft mbH, Weinheim, **1988**.
- [45] J.-P. Liu, W. V. Wilding, N. F. Giles, R. L. Rowley, *Journal of Chemical & Engineering Data* **2010**, *55*, 41.
- [46] J. Daintith, in *Oxford Dictionary of chemistry*, **2000**.
- [47] A. A. Ivanova, A. A. Ivanov, A. A. Oliferenko, V. A. Palyulin, N. S. Zefirov, *SAR and QSAR in Environmental Research* **2005**, *16*, 231.
- [48] O. Ivanciuc, T. Ivanciuc, P. A. Filip, D. Cabrol-Bass, *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 515.
- [49] M. Cocchi, P. G. De Benedetti, R. Seeber, L. Tassi, A. Ulrich, *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 1190.
- [50] C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *Journal of Chemical Information and Modeling* **2005**, *45*, 170.
- [51] Y. Marcus, *The Properties of Solvents*, Vol. 4, John Wiley & Sons Ltd., Chichester, **1998**.
- [52] A. Johnston, B. F. Johnston, A. R. Kennedy, A. J. Florence, *CrystEngComm* **2008**, *10*, 23.
- [53] A. Martin, A. N. Paruta, A. Adjei, *Journal of Pharmaceutical Sciences* **1981**, *70*, 1115.
- [54] Y. Marcus, *Chemistry Society Reviews* **1993**, 409.
- [55] D. Musumeci, C. A. Hunter, J. F. McCabe, *Crystal Growth and Design* **2010**, *10*, 1661–1664.
- [56] S. Black, F. Muller, *Organic Process Research and Development* **2010**, *14*, 661.
- [57] C. L. Leci, N. Garti, S. Sarig, *Journal of Crystal Growth* **1981**, *51*, 85.
- [58] R. Hilfiker, J. Berghausen, F. Blatter, A. Burkhard, S. M. De Paul, B. Freiermuth, A. Geoffroy, U. Hofmeier, C. Marcolli, B. Siebenhaar, M.

- Szelagiewicz, A. Vit, M. Von Raumer, *Journal of Thermal analysis and Calorimetry* **2003**, 73, 429.
- [59] J. Huuskonen, *Journal of Chemical Information and Computer Sciences* **2000**, 40, 773.
- [60] F. L. Muller, S. Black, *Organic Process Research and Development* **2009**, 13, 1315.
- [61] Y. Li, P. S. Chow, R. Tan, B. H., S. N. Black, *Organic Process Research and Development* **2008**, 12, 264.
- [62] C.-C. Chen, Y. Song, *Industrial & Engineering Chemistry Research* **2004**, 43, 8354.
- [63] C. M. Hansen, A. L. Smith, *Carbon* **2004**, 42, 1591.
- [64] V. Gutmann, *Electrochimica Acta* **1976**, 21, 661.
- [65] S. Hahn, W. M. Miller, R. N. Lichtenthaler, J. M. Prausnitz, *Journal of Solution Chemistry* **1985**, 14, 129.
- [66] L. Lavielle, J. Schultz, *Langmuir* **1991**, 7, 978.
- [67] Y. Marcus, *Journal of Solution Chemistry* **1984**, 13, 599.
- [68] K. Hiraoka, *The Chemical Society of Japan* **1986**, 39, 2571.
- [69] F. L. Riddle, F. M. Fowkes, *Journal of the American Chemical Society* **1990**, 112, 3259.

