Application of machine learning methods for the design of crystallisation processes

A thesis presented for the degree of

Doctor of Philosophy

in the Faculty of Science

of the University of Strathclyde

by

Rajesh Gurung

Strathclyde Institute of Pharmacy and Biomedical Sciences

September 2018

DECLARATION OF AUTHOR'S RIGHTS

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for the examination which has led to the award of a degree.

Abstract

There is great potential for the implementation of machine learning to aid in pharmaceutical process development. Machine learning (ML) algorithms can be applied to increase the speed with which high-value drug products are developed for the market while reducing the utilisation of material, minimising wastage and assuring the desired quality attributes are achieved. This thesis illustrates the application of ML techniques in aspects of crystallisation process design assessing the ability to predict crystallisation outcomes, including crystal habit and non-aqueous solubility of pharmaceutical drugs in a diverse range of solvents.

High throughput screening for analysing crystallisation outcomes and crystal habit of paracetamol in a diverse range of solvents was developed using *Technobis* Crystalline. Out of 94 solvents, paracetamol was observed to crystallise in 44 solvents, remain in solution at set conditions in 11 solvents, never solubilise in 36 solvents and show signs of degradation in 3 solvents. Based on these experimental data, a ML classification model was constructed for predicting the crystallisation outcomes and crystal habit of paracetamol with ~77.78 % prediction accuracy. Analysis of the ML model revealed that the physicochemical descriptors and predictive capabilities were more directed towards defining solubility of paracetamol rather than its nucleation behaviour. A rapid and efficient solvent selection tool based on relative solubility was developed using ML algorithms. The tool was not only successful in aid of rational selection of solvent but also reduce the number of screening experiments in the laboratory and thus limit material cost and usage. The regression and classification models built to predict non-aqueous solubility on 247 drug and drug-like molecules in seven commonly used solvents demonstrated that the molecular descriptors calculated using MOE were better at predicting solubility compared to structural fingerprint descriptors. Furthermore, both the regression and classification models successful predicted solubility of drugs in alcohols compared to other organic solvents.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Professor Alastair Florence who never gave up on me, kept encouraging me on every single step and without him, this thesis would not have been possible. I would also like to present my thanks to Dr Andrea Johnston who has been a fantastic mentor and my friend throughout my PhD. I would like to thank her for introducing me to the wonderful world of machine learning.

I would also like to thank my co-supervisor Dr Blair Johnston for his constant guidance and Dr Antony Vassileiou for clearly explaining even the most complex of topics. I would like to thank my parents for their constant love and support, for believing that I can and continuously quoting the phrase, '*it's not about the destination but all about the journey*'. I would like to thank the gorgeous Dr Stephanie Yerdelen for being understanding, for being there through thick and thin and making my PhD a memorable and pleasant experience. Without this PhD, we would never have met, and I am thrilled to have found you. It was a pleasure working seven feet away from you every day and seeing you was enough to bring the sunshine and joy in gloomy Glasgow. I would like to thank my other family down in Loughborough University, Dimitris Fysikopoulos and Emmanuel Kimuli for just being purely awesome and Dr Maria Briuglia for all the delicious carbonara. There is a lot of people in CMAC that I am very grateful for; particularly Dr Murray Robertson for his motivation and encouraging words. Many thanks to, Dr Thomas McGlone, Dr Cameron Brown, Dr John Robertson, Dr Martin Ward (thank you for introducing me to Tim Hortons), lovely Andrew Dunn and all the excellent CMAC team.

And finally, I would like to dedicate this thesis to my life coach, my late grandfather Bom Bahadur Gurung and my late uncle Lekh Bahadur Gurung. I am sorry I wasn't there to say goodbye. I hope we will meet someday in the future.

CONTENTS

CHAPTER 1. INTRODUCTION 1

1.1 Introduction 2

1.2 Fundamental crystallisation process parameters 4

1.2.1 Solubility and supersaturation 4

1.2.2 Nucleation 7

1.2.3 Crystal Growth 12

1.3 Machine learning (ML) in chemoinformatics 14

1.3.1 An overview of chemoinformatics 14

1.3.2 Molecular representation 16

1.3.3 Molecular descriptors 18

1.3.4 Machine learning algorithms 20

1.3.5 The performance measure of ML algorithms 34

CHAPTER 2. AIMS AND OBJECTIVES 39

2.1 Aims 40

2.2 Objectives 41

CHAPTER 3. MATERIAL AND METHODS 42

3.1 Materials 43

3.2 Experimental methodology 47

3.2.1 Crystallisation techniques 47

3.2.2 Zinsser automated platform 50

3.2.3 Optical microscopy 51

3.2.4 Malvern Morphologi G3-ID 51

3.3 Machine learning methodology 52

3.3.1 Molecular structure representation 52

3.3.2 Molecular descriptors 52

3.3.3 Molecular descriptors 56

3.3.4 ML algorithms 57

CHAPTER 4. APPLICATION OF ML ALGORITHMS FOR PREDICTING CRYSTALLISATION OUTCOMES AND CRYSTAL HABIT OF API IN A DIVERSE RANGE OF ORGANIC SOLVENTS 60

4.1 Introduction 61

4.2 Methodology 62

4.2.1 Experimental methods 62

4.3 Results and discussions 68

4.3.1 Controlled cooling experiments 68

4.3.2 ML Algorithm 76

4.4 Summary 112

CHAPTER 5. DEVELOPMENT OF A RAPID AND EFFICIENT SOLVENT SELECTION TOOL 115

5.1 Introduction 116

5.2 Methodology 118

5.3 ML workflow 122

5.4 Results and Discussions 125

5.4.1 Solvent screening of paracetamol, carbamazepine and carvedilol 125

5.4.2 Deciding optimum training set for an effective RF classification model 132

5.4.3 A rational way of selecting the solvents 141

5.4.4 RF Classification model and performance comparison 156

5.4.5 Optimisation attempts on the trained RF classification model 165

5.4.6 Assessment of the important molecular descriptors 167

5.5 Summary 169

CHAPTER 6. APPLICATION OF MACHINE LEARNING TO PREDICT SOLUBILITY OF DRUG AND DRUG-LIKE COMPOUNDS IN ORGANIC SOLVENTS 172

6.1 Introduction 173

6.2 Methodology 174

6.2.1 Solubility database 174

6.2.2 Molecular descriptors and model development *176*

6.2.3 Model development *177*

6.2.4 ML model workflow *180*

6.2.5 Model performance metrics *181*

6.3 Results and discussions 182

6.3.1 RF regression models *182*

6.3.2 RF Classification model *186*

6.4 Summary 198

CHAPTER 7. CONCLUSION AND FUTURE WORK 200

7.1 Conclusion 201

7.2 Future works 205

7.2.1 Chapter 4 .... *205*

7.2.2 Chapter 5 .... *206*

7.2.3 Chapter 6 .... *207*

CHAPTER 8. PUBLICATIONS 208

CHAPTER 9. REFERENCES 209

9.1 References 210

CHAPTER 10. APPENDICES 255

10.1 APPENDIX 1 256

10.2 APPENDIX 2 A 283

10.3 APPENDIX 2 B 290

List of Tables

Table 1.1 Examples of water-ice systems to illustrate different nucleation modes and the conditions under which they apply 8

Table 1.2 List of computational methodologies available in the spectrum of chemoinformatics (Bunin, Siesel, Morales, & Bajorath, 2006). 15

Table 1.3 Different line notations for the structure diagram of phenylalanine (Gasteiger & Engel, 2006). 17

Table 1.4 Classification of molecular descriptors based on data type (adapted from (Bunin et al., 2006)). 18

Table 1.5 Classification of molecular descriptors based on the dimensionality of the structural representation (adapted from (Bunin et al., 2006; Sliwoski, Mendenhall, & Meiler, 2016)). 19

Table 1.6 Example of a 2 x 2 confusion matrix 35

Table 1.7 Logical interpretation of Cohen's Kappa (*k*) (McHugh, 2012) 36

Table 3.1 Physical properties of the compounds selected in this thesis (Haynes, 2016; Whitesell, 1998). 43

Table 3.2 List of 94 solvents and their respective family class used for cooling crystallisation. 45

Table 3.3 List of the calculated 2D and 3D molecular descriptors with a brief description (P. Labute, 2000) 53

Table 4.1 Shown is the list of solvents categorised according to their family class and their respective crystallisation outcomes with paracetamol. The solvents are labelled numerically from 1 to 94 69

Table 4.2 Summary of the four crystallisation outcomes of paracetamol observed in 94 solvents 71

Table 4.3 Measured aspect ratio of paracetamol crystallised in 44 organic solvents by cooling crystallisation. ... 73

Table 4.4 Summary of the two crystal shape outcomes of paracetamol observed in 44 solvents. ... 76

Table 4.5 The first 5 principal components of the built PCA models for *Dataset A* containing 94 solvents displaying the eigenvalues, percent of variance and the cumulative percent of the variance ... 76

Table 4.6 The first 5 principal components of the built PCA models for *Dataset B* displaying the eigenvalues, percent of variance and the cumulative percent of the variance ... 77

Table 4.7 Partial list of top 20 variables that contributed the most to the first two PCs for *Dataset A*. The larger the value, the more the variable contributed to the component. ... 88

Table 4.8 Partial list of top 20 variables that contributed the most to the first two PCs for *Dataset B*. ... 89

Table 4.9 Breakdown of *Dataset A* in the ratio of 80:20 with 80% as training set and 20% as a test set ... 95

Table 4.10 Breakdown of updated *Dataset A* in the ratio of 80:20 with 80% as training set and 20% as a test set ... 95

Table 4.11 Confusion Matrix describing the performance of the final RF model trained on the training set after 10-fold cross validation for *Dataset A* containing the three crystallisation outcomes ... 98

Table 4.12 Confusion matrix of the predictions made on the external test set ... 100

Table 4.13 Confusion Matrix describing the performance of the stratified RF classification model on the training set for the new dataset containing the three crystallisation outcomes. ... 101

Table 4.14 Confusion matrix generated by the three RF models (RF1, RF2 and RF3) after applying OVO binarization technique on the three crystallisation multi-outcomes. 103

Table 4.15 Description of the ten most important molecular descriptors obtained from both RF2 and RF3 classification models calculated using MOE (Boyd, 2005) 105

Table 4.16 Performance measure of the 10-fold CV models calculated by iteratively removing the least important molecular descriptors for both RF2 and RF3 classification model 106

Table 4.17 Breakdown of *Dataset A* in the ratio of 80:20 with 80% as a training set and 20% as a test set 109

Table 4.18 Confusion Matrix describing the performance of the RF classification model on the training set for *Dataset B* containing the two crystal shape outcomes 110

Table 4.19 Confusion Matrix describing the predictive accuracy of the RF classification model on the test set for *Database B* containing the two crystal shape outcomes 110

Table 4.20 Confusion Matrix describing the performance of the stratified RF classification model on the training set for *Dataset B* containing the two crystal shape outcomes. 111

Table 5.1 Summary of the experimental conditions and classification for obtaining the relative solubility of the chosen APIs 119

Table 5.2 Total number of solvents found to be soluble and practically soluble in paracetamol, carbamazepine and carvedilol at two temperature points 125

Table 5.3 List of solvents with their respective qualitative solubility for paracetamol (PCM), carbamazepine (CBZ) and carvedilol (CRV) at two temperature points126

Table 5.4 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in paracetamol at low temperature points 147

Table 5.5 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in paracetamol at high temperature points 148

Table 5.6 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carbamazepine at low temperature points 149

Table 5.7 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carbamazepine at high temperature points 150

Table 5.8 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carvedilol at low-temperature points 151

Table 5.9 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carvedilol at high temperature points 152

Table 5.10 Final list of the recommended solvents for training the RF model on the three target compounds based on their frequency of observation 153

Table 5.11 Selected list of solvents using Strathclyde24. Selection criteria were that at least one solvent must be included from each solvent cluster 156

Table 5.12 Performance of the RF classification model trained using solvents selected from *method I.* 157

Table 5.13 Performance of the RF classification model trained using solvents selected from *method II.* 158

Table 5.14 Most important molecular descriptors extracted from the trained RF classification model for the three target compounds at two temperature points (Boyd, 2005) 167

Table 6.1 Shown are the number of compiled solubility data points 176

Table 6.2 Solubility definitions taken as per the Merck Index (Wolk, Agbaria, & Dahan, 2014) 178

Table 6.3 Number of solubility data points classified in each of the binary qualitative solubility regions for each organic solvent based on a solubility threshold of 33 mg/ml 179

Table 6.4 Number of solubility data points classified in each of the binary qualitative solubility regions for each organic solvent based on the solubility criteria shown in Table 6.3 180

Table 6.5 Comparison of the prediction performance of RF regression model using 10-fold cross validation on seven solvents built using MOE molecular descriptors and MACCS molecular fingerprints 183

Table 6.6 Comparison of the prediction performance of RF regression model on the 20% external validation set on seven solvents 184

Table 6.7 Breakdown of the solubility dataset in the ratio of 80:20 with 80% as a training set and 20% as an external validation set 187

Table 6.8 Average prediction accuracy after 10-fold cross validation on seven organic solvents (binary classification) 188

Table 6.9 Confusion matrix of the classification model of the alcohol models 189

Table 6.10 Confusion matrix of the classification model of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane 189

Table 6.11 Predictive performance of the RF model on the 20% external validation set for the seven organic solvents (binary classification) 190

Table 6.12 Confusion matrix on the external validation set of the alcohols models . 191

Table 6.13 Similarly, Confusion matrix on the external validation set of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane 191

Table 6.14 Breakdown of the solubility dataset on the basis of the three solubility regions in the ratio of 80:20 with 80% as a training set and 20% as an external validation set 193

Table 6.15 Average prediction accuracy after 10-fold cross validation on seven organic solvents (Three label classification) 194

Table 6.16 Confusion matrix of the classification model of the alcohol models with three outcomes 194

Table 6.17 Similarly, Confusion matrix of the classification model of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane model with three solubility outcomes 195

Table 6.18 Predictive performance on the external validation set of the RF model on seven organic solvents with three qualitative solubility outcomes 196

Table 6.19 Confusion matrix for prediction on the external validation set of alcohol models with three solubility outcomes 196

Table 6.20 Confusion matrix for prediction on an external validation set of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane models with three solubility outcomes 197

Table 10.1 Solubility data collated from literature of various drug compounds in ethanol taken at lab temperature 256

Table 10.2 Solubility data collated from literature of various drug compounds in methanol taken at lab temperature 261

Table 10.3 Solubility data collated from literature of various drug compounds in 1-butanol taken at lab temperature 265

Table 10.4 Solubility data collated from literature of various drug compounds in Acetonitrile taken at lab temperature 268

Table 10.5 Solubility data collated from literature of various drug compounds in Acetone taken at lab temperature 272

Table 10.6 Solubility data collated from literature of various drug compounds in Ethylacetate taken at lab temperature 276

Table 10.7 Solubility data collated from literature of various drug compounds in 1,4-dioxane taken at lab temperature 281

List of Figures

Figure 1.1 Graphical representation of the solubility – supersaturation phase diagram (J. M. Hughes, Aherne, & Coleman, 2012) 5

Figure 1.2 Schematic workflow of an unsupervised ML model.................................... 21

Figure 1.3 Example of a simple k-means clustering showing two distinctive clusters.22

Figure 1.4 Illustration of PCA analysis steps indicated by the purple arrow on a dataset. The data are initially represented on the x-y co-ordinates. Dimensional reduction is performed by identifying the directions in which variance is maximum. PC1 is the first principal component (longest blue, double headed arrow) which captures the maximum variance in the data points followed by the second principal component (PC2) placed orthogonally to capture the next largest variance in points. (Adapted from: Principal Component Methods in R (Kassambara, 2017).. 24

Figure 1.5 Schematic diagram of supervised ML models 25

Figure 1.6 Example of an SVM supervised classification model built on the iris dataset showing the resultant classification of the two labels: versicolour and virginica. 27

Figure 1.7 Illustration of a decision tree classification model. Here A is the parent node of B and C (child nodes). The decision node A along with the terminal nodes B and C is known as the branch of the decision tree. 30

Figure 1.8 Illustration of many decision trees forming a random forest model. 33

Figure 1.9 Illustration of reinforcement ML algorithm.. 34

Figure 3.1 Molecular structure of A. paracetamol, B. carbamazepine and C. carvedilol 45

Figure 3.2 Cooling crystallisation performed on *Technobis Crystalline* Reactor system for observing crystallisation outcomes and crystal habit of an API 48

Figure 3.3 Cooling crystallisation experiment performed on *Technobis Crystal16* for the solvent screening study. 50

Figure 3.4 Schematic workflow designed to calculate the 2D and 3D molecular descriptors of the organic solvents/compounds. The individually coloured brown and blue boxes indicate the software used for the specific parameters... 56

Figure 3.5 Schematic representation of MACCS key based fingerprints representation of the chemical structure. Each position in the vector indicated the presence (1) or absence (0) of fragments.. 57

Figure 3.6 Schematic diagram of 10-fold cross-validation. The dataset is divided into 10-folds where one fold is designated as test set while 9 folds are designated as the training set. The average accuracy of the ten test sets is the final cross-validated accuracy... 59

Figure 4.1 Schematic workflow of the implementation of the machine learning model in recommending suitable solvents for the crystallisation process. 62

Figure 4.2 Workflow presenting the implementation of machine learning models to understand the relationship between the molecular descriptors and the crystallisation outcomes and crystal habit.. 65

Figure 4.3 Workflow presenting the implementation of machine learning models to understand the relationship between the molecular descriptors and the crystallisation outcomes and crystal habit.. 67

Figure 4.4 BFDH morphology prediction of monoclinic paracetamol. 72

Figure 4.5 Optical micrographs of the crystal shape of paracetamol grown in 44 solvents by controlled cooling crystallisation arranged in the order of their decreasing mean aspect ratios: 1) 2-phenylethanol 2) Methoxyethane 3) THF 4) 2-pentanol 5) Water 6) Benzyl alcohol 7) Acetone 8) 2-butanol 9) 1,4-dioxane 10) Diethyl carbonate 11) Anisole 12) 2-butoxyethanol 13) EGME 14) Aniline 15) 1,2-propanediol 16) 2-propanol 17) Butyl acetate 18) Acetonitrile 19) Methanol

20) Ethyl acetate 21) 2-butanone 22) TFE 23) 1-butanol 24) 2-phenoxyethanol 25) Nitromethane 26) Nitrobenzene 27) Acetic Acid 28) Ethanol 29) Isoamyl alcohol 30) 1-octanol 31) Cyclohexanol 32) 1-decanol 33) 1-propanol 34) 4-methyl-2-pentanone 35) 2-methyl-1-propanol 36) N,N-DMA 37) 0E 38) Formamide 39) Formic Acid 40) N,N-DMF 41) Diiodomethane 42) IBA 43) 3-pentanone 44) THT 74

Figure 4.6 Scree plot of the extracted PC for *Dataset A* showing the fraction of total variance represented by each principal component........ 77

Figure 4.7 Scree plot of the extracted PC for *Dataset B* showing the fraction of total variance represented by each principal component........ 78

Figure 4.8 Scatter Plot based on the score values of 94 solvents in *Dataset A* projected to the first two PCs displaying the four-crystallisation outcome. Three significant clusters that were observed are outlined by the green, blue and red regions for visualisation. The solvents in the chemical space are labelled as 1 to 94 accordingly as numbered in Table 4.1. 79

Figure 4.9 3-D Scatter Plot of the score values of 94 solvents projected to the first three PCs displaying the four crystallisation outcomes. The green region indicated the solvent clusters showing the *soluble* outcomes and the red region indicated the solvent clusters showing the *practically insoluble* outcomes. 81

Figure 4.10 Projection of the scatter plot for 94 solvents based on the scores of the first two PCs. The solvents were labelled according to their respective solvent family class to identify trends within the PCA model. 82

Figure 4.11 Score plot based on the score values of 94 solvents in *Dataset A* projected to the first two PCs and displaying the solvents based on their polarity. 83

Figure 4.12 Scatter Plot based on the score values of 44 solvents in *Dataset B* projected to the first two PCs displaying the two crystal habit outcomes. The solvents in the chemical space are labelled as 1 to 44 accordingly (Table 4.3). ... 84

Figure 4.13 Variable factor map of the calculated molecular descriptors of 94 solvents from *Dataset A* 86

Figure 4.14 Variable factor map of the calculated molecular descriptors of 44 solvents from *Dataset B* 87

Figure 4.15 Biplot constructed using the first two PCs of *Dataset A* showing the overlay of the score plot and the factor map. 91

Figure 4.16 Biplot constructed using the first two PCs for *Dataset B* showing the overlay of the score plot and the factor map. 92

Figure 4.17 Average classification accuracy of the RF model with an increasing number of trees 96

Figure 4.18 Average classification accuracy of the RF model with an increasing number of *mtry* 97

Figure 4.19 Illustration of the MDS plot on the training set obtained from the RF proximity matrix for *Dataset A* with its classified three crystallisation outcomes. The shaded oval regions A and B outlined the observed two separate clusters. 99

Figure 4.20 Illustration of the One-versus-One binarization strategy applied to the new dataset with three crystallisation outcomes. 102

Figure 4.21 Graphical representation of the ten most important descriptors obtained for the RF model: i) RF2 model trained using outcomes *non-nucleated (B)* and *practically insoluble (C)* and ii) RF3 model trained using outcomes. 104

Figure 5.1 Molecular structure of the solvents selected for this study 121

Figure 5.2 Schematic workflow of the RF model developed in order to predict the qualitative solubility of three target APIs. 10,000 resampling iterations were performed. 124

Figure 5.3 Plot of qualitative transmissivity of paracetamol (PCM) in 63 organic solvents at two temperature points 129

Figure 5.4 Plot of qualitative transmissivity of carbamazepine (CBZ) in 63 organic solvents at two temperature points 130

Figure 5.5 Plot of qualitative transmissivity of carvedilol (CRV) in 63 organic solvents at two temperature points 131

Figure 5.6 10,000 RF classification models trained for paracetamol at low temperature and at varying training set sizes. 135

Figure 5.7 10,000 RF classification models trained for carbamazepine at low temperature and at varying training set sizes 136

Figure 5.8 10,000 RF classification models trained for carvedilol at low temperature and at varying training set sizes.. 137

Figure 5.9 10,000 RF classification models trained for paracetamol at high temperature and at varying training set sizes 138

Figure 5.10 10,000 RF classification models trained for carbamazepine at high temperature and at varying training set sizes 139

Figure 5.11 10,000 RF classification models trained for carvedilol at high temperature and at varying training set sizes 140

Figure 5.12 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 20 for paracetamol at low-temperature point.143

Figure 5.13 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 24 for paracetamol at low-temperature point.144

Figure 5.14 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 20 for paracetamol at high-temperature point145

Figure 5.15 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 24 for paracetamol at high-temperature point146

Figure 5.16 Visualizations of Strathclyde24 illustrating 24 individual clusters. Each cluster is an MDS plot illustrating the similarity and dissimilarity within and between the clusters of solvents (Johnston et al., 2017). 155

Figure 5.17 MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for paracetamol.. 162

Figure 5.18 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for paracetamol 162

Figure 5.19 MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for carbamazepine.. 163

Figure 5.20 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for carbamazepine 163

Figure 5.21MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for carvedilol... 164

Figure 5.22 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for carvedilol 164

Figure 5.23 The top ten important molecular descriptors for the selected RF classification model of the three target compounds taken at low and high temperatures and sorted accordingly to its average variable importance in descending order.. 166

Figure 6.1 Workflow of the RF machine learning algorithm to predict non-aqueous solubility of drug and drug like compounds at laboratory temperature. 10-fold cross validation was performed to evaluate the performance of the RF model. The trained RF model was then used to predict the non-aqueous solubility on the validation set. 181

Figure 10.1 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in ethanol. 283

Figure 10.2 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in ethanol. 283

Figure 10.3 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in methanol.. 284

Figure 10.4 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in methanol.. 284

Figure 10.5 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in 1-butanol. 285

Figure 10.6 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in 1-butanol. 285

Figure 10.7 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in acetonitrile.. 286

Figure 10.8 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in acetonitrile.. 286

Figure 10.9 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in acetone. 287

Figure 10.10 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in acetone.. 287

Figure 10.11 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in ethylacetate. 288

Figure 10.12 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in ethylacetate... 288

Figure 10.13 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the training set compounds in 1,4-dioxane. 289

Figure 10.14 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in 1,4-dioxane.. 289

Figure 10.15 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in ethanol. 290

Figure 10.16 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in ethanol. 290

Figure 10.17 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in methanol.. 291

Figure 10.18 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in methanol. 291

Figure 10.19 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in 1-butanol... 292

Figure 10.20 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using molecular descriptors on the external test compounds in 1-butanol. 292

Figure 10.21 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in acetonitrile. 293

Figure 10.22 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in acetonitrile. 293

Figure 10.23 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in acetone. 294

Figure 10.24 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in acetone. 294

Figure 10.25 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in ethylacetate. 295

Figure 10.26 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in ethylacetate. 295

Figure 10.27 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the training set compounds in 1,4-dioxane. 296

Figure 10.28 Regression plot of the experimental vs predicted solubility data obtained from the RF model trained using MACCS fingerprint on the external test compounds in 1,4-dioxane. 296

List of Abbreviations and Acronyms

ACC	Classification accuracy
ADME	Absorption, Distribution, Metabolism and Excretion
API	Active Pharmaceutical Ingredient, Active Pharmaceutical Ingredients
AR	Aspect Ratio
BFDH	Bravais-Friedel-Donnay-Harker
CoMFA	Comparative Molecular Field Analysis
CSD	Cambridge Structural Database
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FN	False Negative
FP	False Positive
ID3	Iterative Dichotomizer 3
InChl	IUPAC International Chemical Identifier
LOOCV	leave-one-out-cross-validation
MACCS	Molecular ACCess System
MAE	Mean Absolute Error
MDL	Molecular Design Limited
MDS	Multi-Dimensional Scaling
ML	Machine Learning
MOE	Molecular Operating Environment
OOB	Out-Of-Bag
OVO	One-Vs-One
PC	Principal Components

PCA	Principal Component Analysis
PP	Projection Pursuit
PPV	Positive Predictive Value
QSAR	Qualitative and Quantitative Structure-Activity
QSPR	Qualitative and Quantitative Structure Property Relationships
R^2	R-squared
RF	Random forest
RMSE	Root Mean Square Error
ROSDAL	Representation of Organic Structure Descriptions Arranged Linearly
SDU	sample dispersion unit
SHE	Safety Health and Environment
SLN	Sybyl Line Notation
SMILES	Simplified Molecular-Input Line-Entry System
SOP	Standard Operating Procedure
SVD	Singular Value Decomposition
SVM	Support Vector Machine
tBME	2-methoxy-2-methylpropane
TN	True Negative
TNR	True Negative Rate
TP	True Positive
WLN	Wiswesser Line-Formula Notation
XRPD	X-ray powder diffraction

Nomenclature

$\mu_{solution}$	chemical potential of the solute molecules in solution
μ_{solid}	chemical potential of the solute molecules in solids
R	Gas constant [8.314 J mol^{-1} K^{-1}]
T	Temperature
c	solute concentration
c_{sat}	saturated/equilibrium concentration
A	pre-exponential factor
γ	Interfacial tension
v	molecular volume
k	Boltzmann constant
S	Supersaturation ratio
$f(\varphi)$	Zeldovich factor
B_{hom}^{0}	Homogeneous nucleation
B_{het}^{0}	Heterogeneous nucleation
k_b	secondary nucleation rate constant
μ_k^j	*k*th moment of the crystal size distribution present in the crystalliser
g	overall order of the growth process
k_g	overall growth rate constant
A_T	total crystal surface area [m^2]
W	crystal weight [kg]
t	time [s]
N	Number of samples

n	number of Class A
m	number of Class B
p	ratio of n divided by N (p = n/N)
q	ratio of m divided by N
f_e	expected frequency
n_t	number of solvents in the training set size
T_s	number of solvents used for the model
ε	dielectric constant

Chapter 1.　Introduction

1.1 Introduction

Crystallisation is a widely used unit operation applied mainly as separation and purification technology in various major sectors of the chemical process industries such as pharmaceuticals, agrochemicals, microelectronics, food and petrochemicals. In the pharmaceutical industry, crystallisation mainly serves as a separation process for intermediates and often as the final step in the manufacture of active pharmaceutical ingredients (API), ideally yielding a pure, finely divided free-flowing crystalline powder (Mullin, 2001b). Operating conditions of the crystallisation process determine the properties of the crystalline product including crystal purity, particle size and shape distribution, surface properties as well as a solid-state structure including the occurrence of polymorphism or solvate formation. These essential quality attributes of the crystalline product greatly affect the efficiencies of the downstream operations such as filtration, drying, washing, flow and compaction for example. The physical properties of the APIs are also responsible for altering the bio-performance of the drug, dissolution rate, stability as well as its shelf life (Khadka et al., 2014). For pharmaceuticals that exhibit various polymorphs the crystallisation process also affects the polymorph produced. The solid-state phase and purity of the product in turn may affect drug dissolution, efficacy or potential toxicity, which are critical from a consumer safety and regulatory point of view. Thus, most pharmaceutical manufacturing processes include a series of crystallisation processes with controlled critical process parameters (e.g. cooling rate, supersaturation and impurity level) to achieve the desired crystalline product of high purity and desired final crystal form (Aamir, Nagy, Rielly, Kleinert, & Judat, 2009).

The pharmaceutical industry is continuously challenged by the need to comply with more stringent and detailed product requirements which leads to the high cost and longer development times. The design of the crystallisation process can be complex in comparison to other processes given the need to achieve multiple objectives in terms of purity, particle shape and size distribution and polymorphic form. Not only can these impacts on performance in patients but these attributes also affect the performance and ease of manufacture of the material in downstream processes such as filtration, drying, blending and compaction (Fujiwara, Nagy, Chew, & Braatz, 2005). A key challenge for the pharmaceutical industry faced with bringing a wider variety of smaller volume products to the market is to improve the efficiency of process development (Badman & Trout, 2015). This applies particularly to continuous processes, where it is essential to develop comprehensive process understanding to design and implement robust processes (Baumann & Baxendale, 2015).

There is an increasing effort to develop mechanistic models to describe key rate processes used in pharmaceutical manufacturing to inform process design and control better. While significant progress has been achieved for important unit operations such as wet granulation (e.g. Litster and Iverson's work on growth regime maps for liquid-bound granules (Iveson & Litster, 1998)), industrial crystallisation of molecular solids presents considerable challenges. Classical nucleation theory and simple power-law growth rate expression do not capture the complex interactions between nucleation, growth, attrition, agglomeration and a potentially wide range of practically important process parameters including composition, temperature, pressure, flow, undissolved solids, materials of construction or reactor geometry. Data drove approaches including machine learning (ML) are of considerable interest for their potential to model such complex interactions.

Around 90% of the world's data were generated in the last five years and the number of data is getting bigger every day (Baker, Pena, Jayamohan, & Jerusalem, 2018). ML can provide the opportunity to efficiently understand the relationships between parameters impacting complex process mechanisms and moreover, provide the basis for predicting process outcomes. This requires the utilisation of sets of data to develop and train the models over a sufficiently wide range of conditions. However, if successful, ML has the potential to minimise the traditional reliance on purely experimental efforts. Implementation of ML algorithms in the field of crystallisation process design is therefore of significant potential interest.

1.2 Fundamental crystallisation process parameters

Crystallisation involves the formation of a solid crystalline phase from solution, melt or gas. The resulting structure is one in which the atoms or molecules are arranged in a highly regular periodic array described by the crystal lattice. This ordered 'crystalline' state differs from 'amorphous' solids which lack long-range structural order (Fahlman, 2002). These well-defined internal structural features are reflected in their external morphology with clearly defined faces and sharply defined inter-facet angles and consistent physical properties including melting point and solubility. Crystallisation from solutions is extensively used in the pharmaceutical industries for the manufacture of solid bulk APIs as almost 90% of them are delivered in the crystalline state (Shekunov & York, 2000). Crystallisation confers purity and stability rendering the API in a form that can typically be handled, processed and administered to patients with relative ease. Crystallisation from solution is the most widely used industrial process and involves the dissolution of crystalline APIs in a solvent and crystallisation is induced by inducing a thermodynamic driving force e.g. by changing temperature, pressure or solvent fraction. Solution crystallisation can generally be described to be in three successive stages: supersaturation of the solution, the formation of the crystal nuclei (nucleation) and crystal growth. These are described in the sections below.

1.2.1 Solubility and supersaturation

Solubility and supersaturation are two fundamental concepts that need to be considered when characterising the dynamics of a crystallisation system. Solubility is an important physical property and is defined as the maximum concentration of a substance in a solvent at equilibrium under a given set of temperature and pressure conditions (Myerson & Ginde, 2002). In most cases, the solubility of solute rises by increasing the temperature within the bulk and therefore it is typically considered as a function of temperature (Schwartz & Myerson, 2002). Nevertheless, a few systems present an inverted behaviour, such as anhydrous sodium sulphate which shows a reduction in aqueous solubility on increasing temperature (Mullin, 2001b). A graphical representation of a typical solubility curve is shown in Figure 1.1. The solubility line defines the boundary between the stable undersaturated region of the solution phase diagram and the metastable region of the phase diagram.

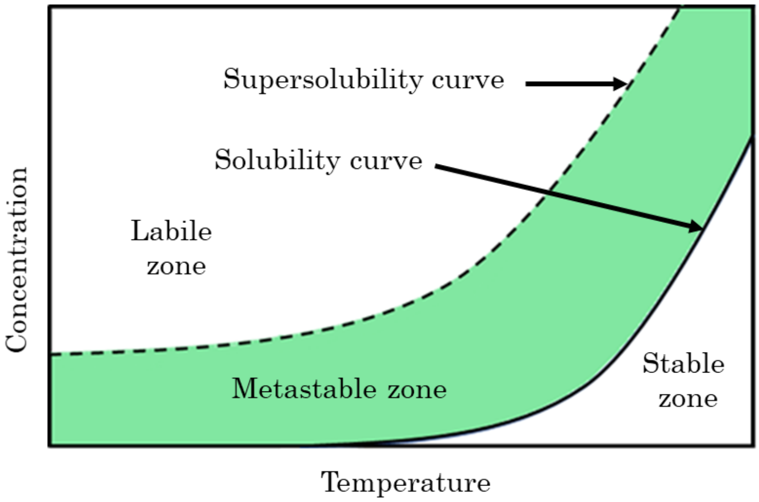


Figure 1.1 Graphical representation of the solubility – supersaturation phase diagram (J. M. Hughes, Aherne, & Coleman, 2012)

Supersaturation, describes the thermodynamic driving force of crystallisation arising from solute being dissolved in concentrations in excess of the thermodynamic equilibrium solubility (Kim & Mersmann, 2001). The extent of supersaturation dictates the kinetics of the key crystallisation rate processes and therefore impacts the achieved quality attributes including yield, size, shape, purity and polymorphic form. Supersaturation can be calculated as a function of the difference between the chemical potential of the solute molecules in solution ($\mu_{solution}$) and in the solid (μ_{solid}) state. There are a number of commonly used expressions to describe supersaturation (Mangin, Puel, & Veesler, 2009).

$$\text{Supersaturation} = \exp\left(\frac{\mu_{\text{solution}} - \mu_{\text{solid}}}{\text{RT}}\right) \qquad \text{Equation 1.1}$$

where, R is gas constant (8.314 J mol^{-1} K^{-1}) and T is temperature.

However, defining supersaturation based on solute concentration (c) and the saturated/equilibrium concentration (c_{sat}) are typically used for process engineering applications.

$$\text{Absolute concentration } (\Delta c) = c - c_{sat} \qquad \text{Equation 1.2}$$

$$\text{Supersaturation ratio (S)} = \frac{c}{c_{sat}} = \sigma + 1 \qquad \text{Equation 1.3}$$

$$\text{Relative Supersaturation } (\sigma) = \frac{c}{c_{sat}} = \text{S} - 1 \qquad \text{Equation 1.4}$$

In general, supersaturation can be practically generated in the following ways (Su, Nagy, & Rielly, 2015):

i. By chemical reaction – reactive crystallisation: High supersaturation is generated due to chemical reaction between two highly complex organic compounds mixed under stirring which produces product exhibiting a solubility several orders of magnitude lower than that of the reactants.

ii. By cooling crystallisation: By cooling a solution, the solution moves from an undersaturated to a supersaturated state whether into the metastable or labile regions. Primary nucleation can only occur in the labile region. The metastable zone is typically where seeded crystallisations are carried out to control growth rates whilst avoiding uncontrolled nucleation.

iii. By evaporative crystallisation: Due to evaporation, the solvent mass fraction can be significantly decreased driving an undersaturated or saturated solution to a metastable state.

iv. By changing the pH: Variations of the pH can also generate supersaturation by altering the ionisation state and solubility of the solvent that drives the solution to a metastable state.

v. By anti-solvent crystallisation: The addition of an anti-solvent can reduce the solubility of the API, driving the solution from a stable to a metastable region. An anti-solvent is a solvent in which the target compound is highly insoluble. This is widely used in industrial processes as can achieve the best yields although dilution by antisolvent to produce large process volumes and associated consumption of large volumes of solvent are practical considerations.

Although several approaches have been utilised for the correlation and computation of molecular solubility, none of them has proven to be of general applicability. Even thermodynamic models, such as UNIFAQ and SAFT models, fail to accurately predict the solubility of complex systems (e.g. polymorphs, multi-component systems etc.). Data generated for these models are often of poor quality or are obtained under varied experimental conditions. A small change in pH, temperature, presence of impurities etc can provide variability in solubility data. Thus, due to the lack of fully reliable characterised solubility data has let to the solubility being typically determined experimentally by using either the polythermal or the gravimetric method (Llinàs, Glen, & Goodman, 2008).

1.2.2 Nucleation

Nucleation is the process by which new crystals are formed in a crystallising environment. The first particles formed upon nucleation are referred to as nuclei/embryos. Nuclei are typically shortlived and only a few nanometres in size and occur due to the aggregation and clustering of molecules or ions in a supersaturated solution, melt or vapour. According to classical nucleation theory nuclei have to reach a critical size above which they will grow rather than re-dissolve. Nucleation can be spontaneous or induced by external stimuli e.g. mechanical shock, ultrasound, laser stimulation, electric and magnetic fields (Mullin, 2001b).

Two main types of nucleation processes are important considerations in crystallisation: primary nucleation and secondary nucleation. Primary nucleation is when a system nucleates without the presence of any crystalline material and refers to the first formation of crystallites whereas secondary nucleation is where nucleation occurs in an environment that already contains crystalline particles and is therefore influenced by their presence. Primary nucleation is further subdivided into homogeneous and heterogeneous nucleation. A homogeneous primary nucleation is a form of nucleation that occurs spontaneously without external stimuli or influence whereas heterogeneous nucleation is induced by a foreign particle whether an interface or undissolved solids particles or fibres in the supersaturated solution. Table 1.1 describes examples of the three different nucleation modes for ice and the different conditions under which they would be expected to occur (Jones, 2002).

Table 1.1 Examples of water-ice systems to illustrate different nucleation modes and the conditions under which they apply

Mode of nucleation	Example
Primary homogeneous	Crystallisation of carefully purified water (i.e. distilled and filtered). This would require cooling the water to below -30°C before ice forms.
Primary heterogeneous	Crystallisation of tap water, ice would appear at about -6°C
Secondary	Continuous crystallisation of ice in a retained bed crystalliser. Operating conditions for temperature would range between -2 and -3°C.

1.2.2.1 Primary Nucleation

Primary nucleation is the classical form of nucleation and typically occurs at very high levels of supersaturation. In unseeded crystallisation processes, primary nucleation is the prevalent process and will dictate the rate at which nuclei are formed and the growth surface is established for subsequent deposition of excess solute. The nucleation rate generated by a primary homogeneous nucleation mechanism is expressed in Equation 1.5 below.

$$B_{hom}^{0} = Aexp[-\frac{16\pi\gamma^{3}v^{2}}{3k^{3}T^{3}(lnS)^{2}}] \qquad \text{Equation 1.5}$$

where, B_{hom}^{0} is the rate of homogeneous nucleation e.g. the number of nuclei formed per unit time per unit volume, A is a pre-exponential factor, T is temperature, γ is the interfacial tension, v is the molecular volume, k is the Boltzmann constant, S is the supersaturation ratio c/c_{sat}, c is the solution concentration and c_{sat} is the equilibrium saturation concentration.

Equation 1.5 mentioned earlier can also be expressed in the form of an Arrhenius-type rate equation as shown in Equation 1.6 below,

$$B_{hom}^{0} = Aexp[-K(lnS)^{2}] \qquad \text{Equation 1.6}$$

Most primary nucleation that occurs in real processes tends to be heterogeneous rather than homogenous given the difficulty in removing all potential sources of contaminating surfaces or particulates. Heterogeneous nucleation occurs due to the presence of heteronuclei that lower the interfacial tension. The rate of primary heterogeneous nucleation is illustrated below in Equation 1.7. The presence of inert particles in the saturated solution causes nucleation to occur at a much lower level of supersaturation than in their absence.

$$B_{het}^{0} = Aexp[-\frac{16\pi\sigma^{3}v^{2}f(\varphi)}{3k^{3}T^{3}(lnS)^{2}}] \qquad \text{Equation 1.7}$$

where, the factor f (φ) also known as the Zeldovich factor, accounts for the decreased energy barrier to nucleation due to the presence of foreign solid particles (Vehkamäki, 2007).

Although expressions for both primary nucleation modes exist and are widely used, they are theoretical and whilst show the application in describing experimental process up to a point, they have very limited practical application in predicting the nucleation rate. Hence the nucleation rate and kinetic parameters must be measured and correlated empirically for each system under each specific set of experimental conditions.

1.2.2.2 Secondary Nucleation

Nucleation that takes place only in the presence of crystals of the compound being crystallised is termed as secondary nucleation (N.S. Tavare, 2013). In industrial crystallisation, seed crystals in suspension induce the formation of smaller particles and as a result enhance the rate of production of small crystals (Jones, 2002). The new crystals formed during secondary nucleation resemble crystals generated as a result of attrition but differ in the sense that they occur only in supersaturated solutions. Although a lot of work has been undertaken to uncover the mechanism by which secondary nucleation occurs (Agrawal & Paterson, 2015; Garside & Jančić, 1979), precise definitions remain unclear and indeed multiple sub-mechanisms have been described. Some of the known modes by which secondary nucleation occurs include: initial breeding; needle breeding; polycrystalline breeding; shear nucleation; and collision breeding (Jones, 2002).

The crystal surfaces at the solid-liquid interface play a significant part in driving secondary nucleation. The nucleation rate, B is expressed as the number of nuclei per mass of solvent per unit time and is represented by the semi-empirical equation below (Equation 1.8) (Narayan S. Tavare, 1995);

$$B = k_b \mu_k^j \Delta c^b \qquad \text{Equation 1.8}$$

where, k_b is the empirical secondary nucleation rate constant and a function of many variables e.g. temperature, hydrodynamics and presence of impurities, μ_k^j is the kth moment of the crystal size distribution present in the crystalliser, Δc^b is supersaturation. The kth moment of crystal size distribution is defined as:

$$\mu_k^j = \int_0^\infty L^i n(L,t)dL,\ i = 0,1,\ldots. \qquad \text{Equation 1.9}$$

where, L is the characteristic length of crystal and t is the time.

The use of a 3rd moment, i.e., magma (slurry) density which could also be defined as the volume of crystals in the solution, is found to be suitable to account for the secondary nucleation effects as secondary nucleation rate is known to increase in magma density. Therefore, when k=3, Equation 1.10, becomes the secondary nucleation rate equation (Kobari, Kubota, & Hirasawa, 2012; Narayan S. Tavare, 1995);

$$B = k_b M_\tau^j \Delta c^b \qquad \text{Equation 1.10}$$

Studies have shown that primary nucleation is more dependent on supersaturation than secondary nucleation (Jones, 2002). Denk and Bortsaris (1972) used left- and right-handed optical properties of sodium chlorate to discriminate between the crystal surface and solution properties as the causes of secondary nucleation. The study found that at higher supersaturation rates, the crystal product formed consisted of 50% of both optic forms suggesting primary nucleation was occurring. However, at lower supersaturation, all the nuclei were of the same form. Chirality helps distinguish between primary and secondary nucleation as the study assumes that secondary nucleation will give rise to a product of the same chirality as the seed crystals whereas primary nucleation will give rise to a product of both types. The overall nucleation rate in a crystalliser depends on the interaction of the secondary nucleation characteristics of the material being crystallised and the hydrodynamics of the crystal suspension and expressions to account for particle-particle; particle-vessel and particle-impeller collisions have been developed. Hence, when crystallising a given material, crystallisers of different size, agitation rate and flow pattern will tend to produce different nucleation rates.

1.2.3 Crystal Growth

A number of theoretical expressions have been developed in an attempt to represent the crystal growth process taking place either at the atomic scale or macroscopic scale. A simple empirical power law shown in Equation 1.11, can be used to represent the overall growth rate *R*, and can be sufficient for use in engineering purposes in crystal design and operation;

$$R = \frac{1}{A_\tau}\frac{dW}{dt} = k_g \Delta c^g \qquad \text{Equation 1.11}$$

where *g* is the overall order of the growth process, k_g is the overall growth rate constant which depends on the temperature, crystal size, hydrodynamics and the presence of impurities, A_T is the total crystal surface area [m^2] and *W* is the crystal weight [kg] and *t* is time [s]. The slurry density and intensity of mixing can impact the local mixture and relative crystal solution velocity in suspension hence the overall collision rate and occurrence of secondary nucleation.

The size dependence of k_g is likely due to hydrodynamic characteristics because solid molecules reach the growing surface by diffusion through the liquid phase (Jones, 2002). Once at the surface, molecules must orient and become organised into the lattice through an absorbing layer. As with nucleation, all these steps require supersaturation to occur and the extent of *S* will dictate the overall bulk kinetics. In addition, for molecular solids where low symmetry packing of molecules results in anisotropic cells, it is often the case that different crystallographic faces will have different linear growth rates. This leads to a variation in crystal shapes as different faces of the same crystal may have different growth rates, with the slowest growing face determining the crystal habit (Mullin, 2001).

For all crystal growth mechanism, the overall growth rate depends upon the following (Lewis, Seckler, Kramer, & van Rosmalen, 2015) :

i. the lateral bonds in the growing crystal face i.e. a material-specific property,

ii. interaction with the solvent and

iii. the number of growth units impinging on the crystal surfaces (related to solubility).

Crystal growth from a solution is characterised by two significant processes one of which is subdivided into three steps (Jones, 2002):

i. Mass transport from the solution to the crystal surface via diffusion, convection or a combination of both.

ii. Incorporation of material into the crystal lattice through the surface integration

    - First is the adsorption of the growth unit onto the crystal surface
    - Second is the release of its solvation shell after which the growth unit diffuses into the adsorption layer until it is either incorporated into the crystal lattice or pushed back out into the solution.
    - Finally, if the growth unit reaches a point where it can be built into the lattice, it loses the remainder of its solvation shell before it is fully incorporated into the crystal lattice.

In most crystallisation processes, more than one mechanism influences the crystal growth. However, if different mechanisms take place in parallel, the kinetics will be dictated by the faster growing mechanism.

1.3 Machine learning (ML) in chemoinformatics

1.3.1 An overview of chemoinformatics

Chemoinformatics is a broad field that utilises computer and information techniques to facilitate the collection, storage, analysis and manipulation of large quantities of chemical data like chemical formulae, molecular structures, chemical properties, chemical spectra or biochemical activities. To some extent, chemical informatics is the chemical counterpart of bioinformatics (Faulon & Bender, 2010; Gasteiger, 2003; Varnek & Baskin, 2011). According to Frank Brown, the term 'chemoinformatics'' can be generically defined as '*the mixing of information resources to transform data into information and information into knowledge, for the intended purpose of making decisions faster in the arena of drug lead identification and optimisation*' (F. K. Brown, 1998; Engel & Gasteiger, 2018). The field of chemoinformatics have been established for a long period of time and its roots can be traced back in *Annalen der Pharmacie* by Justus Liebig (1832) and in *Chemical Abstracts* (1907) (C. Smith, 2002) where it was understood simply as the application of information technology to chemistry. These documents consisted of textual and two dimensional descriptions of compounds, reaction mechanism and methods of synthesis and identification (C. Smith, 2002). Machine learning (ML) algorithms are a fundamental tool in chemoinformatics and have seen an incremental rise in usage over recent decades (Varnek & Baskin, 2012). Unlike quantum chemistry or molecular simulations that rely heavily on mathematical equations to model physical reality, chemoinformatics utilises ML algorithms to develop models that can simulate, analyse, manipulate and predict physical and chemical properties using either the two-dimensional or both two- and three-dimensional structures of a molecule (J. B. Mitchell, 2014). Furthermore, ML algorithms are much more efficient, can detect complex non-linear relationships in data and easily be scaled to big datasets without the need for extensive computational resources.

The application of chemoinformatics as defined above is heavily focused on the application of statistical/ML algorithms to a set of chemical data in order to derive predictive models and thus is broad and not limited to a specific field. The list of computational methodologies and infrastructures that describes the broadness of the field is shown in Table 1.2.

Table 1.2 List of computational methodologies available in the spectrum of chemoinformatics (Bunin, Siesel, Morales, & Bajorath, 2006).

Assemble, analyse and management of chemical data
Data management and communication
Design and organization of chemical databases
Chemical structure and property prediction (including drug-likeness)
Molecular similarity and diversity analysis
Compound or library design and optimization
Database mining
Compound classification and selection
Qualitative and Quantitative Structure-Activity (QSAR) or –Property Relationships (QSPR)
Information theory applied to chemical problems
Statistical/Machine learning models and both numerical or fingerprint descriptors in chemistry
Prediction of in vivo compound characteristics

The field of chemoinformatics is highly predominant in drug discovery. However, major applications can also be found in the field of agricultural research, food chemistry and material science. Some of these significant applications of chemoinformatics include:

i. Exploration and analysis of the chemical space (Hall, Mortenson, & Murray, 2014; Reymond, van Deursen, Blum, & Ruddigkeit, 2010)

ii. Implementing ML models in investing and predicting the relationship between the desired physicochemical properties, potency and efficacy or off-target effects of compounds (M. Liu et al., 2014; V. Svetnik et al., 2003).

iii. Similarly, develop models for predicting undesired properties of compounds relating to toxicity or Absorption, Distribution, Metabolism and Excretion (ADME), predicting plasma protein bindings and using it as an aid in improving virtual screening methods, (Maltarollo, Gertrudes, Oliveira, & Honorio, 2015; N.-N. Wang et al., 2016) and

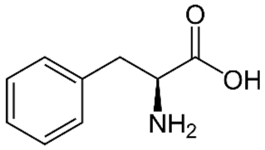
iv. Matched molecular pair analysis where every pair of molecules that differ only by a particular, well-defined, structural transformation in a database of measured properties is identified and the corresponding change in property is computed (Leach et al., 2006; Tyrchan & Evertsson, 2016).

1.3.2 Molecular representation

Numerous formats for machine-readable molecular representation have been developed for various applications ranging from web searching, text mining, and chemical identification. The most widely used method for representing the chemical structure in chemoinformatics is line notations which represent chemical compounds by encoding its connection table and stereochemistry as a linear string of symbolic characters. Linear notations are popular because they are human-readable, can be effectively and efficiently processed for characterization and identification functions as the data is structured in linear form rather than tabular form and maybe canonical (O'Boyle, 2012). Examples of line notations are

Wiswesser Line-Formula Notation (WLN), Representation of Organic Structure Descriptions Arranged Linearly (ROSDAL), Sybyl Line Notation (SLN), Simplified Molecular-Input Line-Entry System (SMILES) and IUPAC International Chemical Identifier (InChl). Examples of the different line notations for the structure diagram of paracetamol is shown in Table 1.3.

Table 1.3 Different line notations for the structure diagram of phenylalanine (Gasteiger & Engel, 2006).

Molecular Structure	
WLN	VQYZ1R
ROSDAL	1O-2-3O,2-4-5N,4-6-7=-12-7
SLN	C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH
SMILES	N[C@@H](CC1=CC=CC=C1)C(O)=O
InChl	InChI=1S/C9H11NO2/c10-8(9(11)12)6-7-4-2-1-3-5-7/h1-5,8H,6,10H2,(H,11,12)/t8-/m0/s1

Among the line notations shown in Table 1.3, SMILES is most widely used as the universal chemical nomenclature for representing the chemical structure of compounds. This is due to the fact: that SMILES are easier to read; more compact than other formats; is the simplest linear code; supports Markush, stereochemistry and reaction coding in an intuitive way. However, it presents certain drawbacks such as a lack of availability for handling aromaticity, a limited array of stereochemistry, no standard way to generate canonical representation and is not unique (Gasteiger & Engel, 2006; O'Boyle, 2012).

1.3.3 Molecular descriptors

Molecular descriptors are numerical, vectors, bit strings or quantitative representations of molecules that capture the physicochemical nature of the investigated molecule. The information content of molecular descriptors is mainly dependent upon the molecular representation of the compound and the mathematical algorithm used to calculate them. More than 5000 molecular descriptors derived from different theories and approaches have been reported in the literature (Todeschini, Consonni, & Mannhold, 2009). Furthermore, there are a plethora of available open-source and proprietary software that have been developed for calculating molecular descriptors such as DRAGON (Mauri, Consonni, Pavan, & Todeschini, 2006), RDKit (Landrum, 2006), MOE (Paul Labute, 2000), PaDEL (Yap, 2011), CDK and ChemoPy (D.-S. Cao, Q.-S. Xu, Q.-N. Hu, & Y.-Z. J. B. Liang, 2013). Molecular descriptors can be classified either by the data type of the descriptors (as shown in Table 1.4) or on the basis of the dimensionality of the structural representation (as shown in Table 1.5).

Table 1.4 Classification of molecular descriptors based on data type (adapted from (Bunin et al., 2006)).

Data Type	Examples
Boolean	The compound has at least one ring
Integer number	Number of heteroatoms, carbon atoms
Floating number	Log P, molecular weight
Vector	Dipole moment
Tensor (3 x 3 matrix)	Electronic polarizability
Scalar field	Electrostatic potential
Vector field	The gradient of the electrostatic potential, i.e., force

Table 1.5 Classification of molecular descriptors based on the dimensionality of the structural representation (adapted from (Bunin et al., 2006; Sliwoski, Mendenhall, & Meiler, 2016)).

Molecular Representations	Descriptors	Examples
0-D	Derived from the molecular Formula; Atom counts, bond counts, molecular weight, a sum of atomic properties	Molecular weight, average molecular weight, number of carbon atoms, hydrogen atoms, double bonds, aromatic bonds, a sum of atomic van der Waals volumes
1-D	captures information on bulk properties, Fragment counts	the number of H-bond donor/acceptor atoms, unsaturation index, hydrophilic factor, molar refractivity, fragment based polar surface are
2-D	Topological Descriptors	Zagreb index, Wiener index, Balaban J index, connectivity indices chi, kappa shape indices, molecular walk counts, BCUT descriptors
3-D	Geometrical descriptors	Molecular eccentricity, the radius of gyration, E-state topological parameter, WHIM descriptors
3-D Surface properties		Mean molecular electrostatic potential, hydrophobicity potential, hydrogen-bonding potential
3-D grid properties		Comparative Molecular Field Analysis (CoMFA)
4-D		It includes the conformational

		flexibility and the freedom of alignment by ensemble averaging in the conventional three dimensional descriptors (Andrade, Pasqualoto, Ferreira, & Hopfinger, 2010)

Molecular descriptors are deemed useful when they provide understandable information about the compounds whilst adding minimum noise. Thus the most useful molecular descriptors are the ones with the greatest degree of information density (amount of information utilised by the model divided by the total information) (Sliwoski et al., 2016). Descriptors containing redundant correlated information should be removed from the model as they often contribute to its poor performance (Shahlaei, 2013).

1.3.4 Machine learning algorithms

ML algorithms can be grouped into three distinctive types: unsupervised, supervised and reinforcement ML models. These three ML models are described in the following sections.

1.3.4.1 Unsupervised ML algorithms

The main goal of unsupervised machine learning is to learn the mapping of the input variables without any prior knowledge of similar outcomes and bring order to the dataset. These algorithms tend to derive insights directly from the input variables and attempt to summarise or cluster the data in order to utilise the insights learnt to make data-driven decisions. Thus, these types of algorithms are beneficial in exploratory analysis for extracting valuable insights where it is either too complicated or impractical for humans to propose trends in the data (Cios, Swiniarski, Pedrycz, & Kurgan, 2007). A schematic workflow of an unsupervised ML model is presented in Figure 1.2.

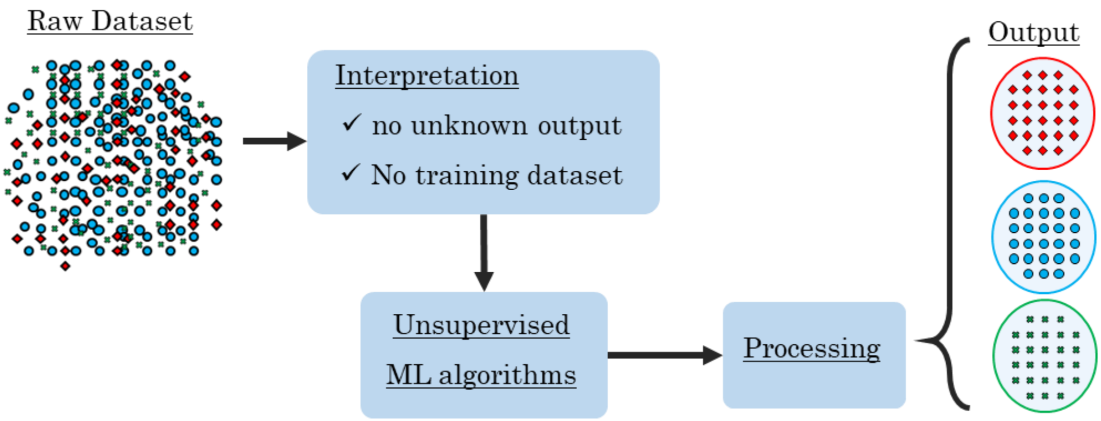


Figure 1.2 Schematic workflow of an unsupervised ML model

Unsupervised ML models apply to two main techniques i.e. clustering and dimensional reduction.

1.3.4.1.1 Clustering algorithms

The goal of this unsupervised ML is to find similarities in data points and group them into clusters based on the inherent structure within the dataset. Data points grouped in the same cluster should have similar properties or features while data points in different groups should have highly diverse properties or features. Some of the popular clustering algorithms include k-means, hierarchical or agglomerative clustering, affinity propagation, mean-shift clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to name a few. An example of a k-means clustering algorithm showing two clusters and the location of their centroids is presented in Figure 1.3.

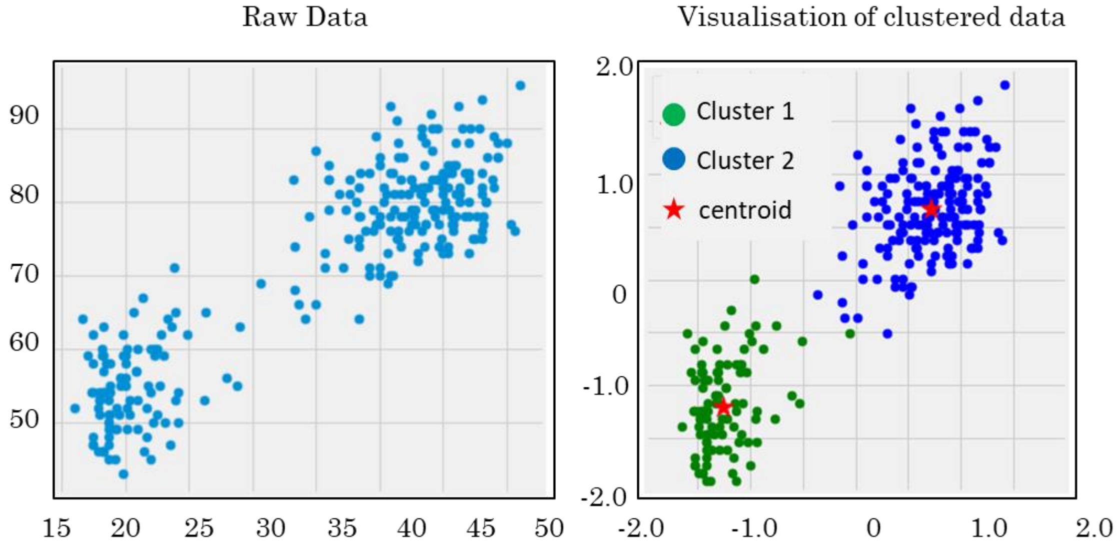


Figure 1.3 Example of a simple k-means clustering showing two distinctive clusters.

1.3.4.1.2 Dimensional reduction

Dimensional reduction aims to derive a set of new artificial variables which is smaller than the original set while still retaining most of the variance of the original data. Dimensional reduction techniques are essential as a raw dataset with high dimensional data are laced with layers of noise, becomes very sparse and thus the analysis suffers from the curse of dimensionality. Furthermore, data analysis is much less computationally intensive on a small dimensional dataset. Examples of dimensional reduction techniques include Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Singular Value Decomposition (SVD), Projection Pursuit (PP) and Sammon mapping (Dash, Liu, & Yao, 1997; Gollapudi, 2016).

1.3.4.1.2.1 Principal Component Analysis (PCA)

High-dimensional data are very common with the rapid growth of chemoinformatics, bioinformatics, healthcare and e-commerce applications. However, data analysis of high dimensional dataset suffers from the curse of dimensionality - finding meaningful similarity measures in the high dimensional space and high demand on computational and memory storage requirements (J. D. Li & Liu, 2017). Principal component analysis (PCA) is an unsupervised ML technique that helps to visualise and reduce the high dimensionality in a dataset whilst retaining the maximum amount of variance. It searches for linear combinations with the most variance from a multivariate data set and expresses this information in a set of new variables known as Principal Components (PC). The amount of variation retained by each principal component in a dataset is indicated by eigenvalue measures. The first principal component accounts for the largest variability in the data and each succeeding component, placed orthogonally, accounts for as much of the remaining variability as possible. The number of principal components is usually less than the number of original variables in the dataset. If the number of principal components is made equal to the dimensionality of the original data, then no reduction is achieved, and the original data set is simply rotated relative to the new PC space (this is not normally useful). The ultimate goal of PCA is to reduce dimensionality of a multivariate data with minimum loss of information by removing noise and redundancy in the data, identify hidden trends or patterns in the dataset, simplify the description of the dataset, extract most important information from the dataset and identify correlated variables (Husson, Le, & Pagès, 2017). Disadvantages of PCA include the lack of interpretability of the results after analysis, its sensitivity to the scale of measurement which could be fixed by standardizing the variables and the difficulty in evaluating the covariance matrix in an accurate manner. The underlying structure of the data must also be linear as PCA might miss non-linear data patterns. Overall, PCA is a fast and powerful tool for data analysis as it identifies the main axes of variance within a dataset, allows for data exploration in order to understand the key variables in the data as well as spot outliers in the dataset (Abdi & Williams, 2010; Jake Lever, Krzywinski, & Altman, 2017). The graphical representation of PCA is shown in Figure 1.4.

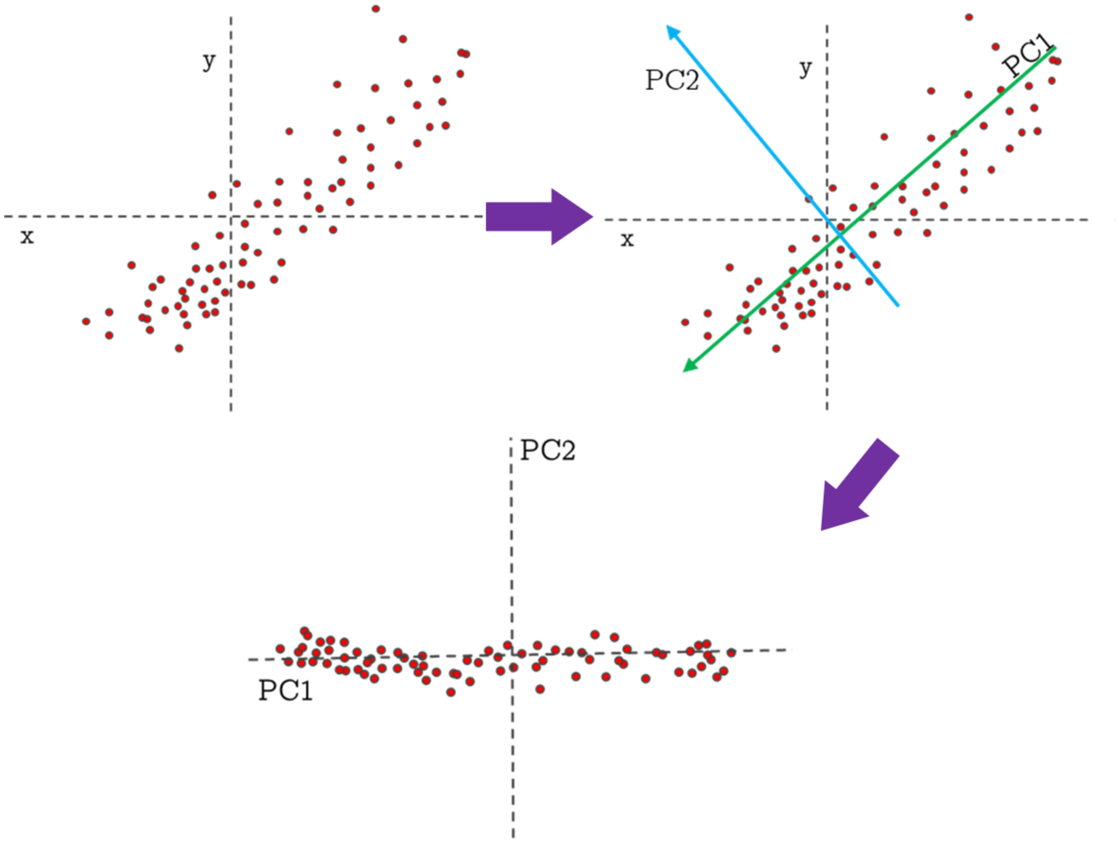


Figure 1.4 Illustration of PCA analysis steps indicated by the purple arrow on a dataset. The data are initially represented on the x-y co-ordinates. Dimensional reduction is performed by identifying the directions in which variance is maximum. PC1 is the first principal component (longest blue, double headed arrow) which captures the maximum variance in the data points followed by the second principal component (PC2) placed orthogonally to capture the next largest variance in points. (Adapted from: Principal Component Methods in R (Kassambara, 2017).

1.3.4.2 Supervised ML algorithm

Supervised ML models take known datasets consisting of both the input and associated target variables, learn the relationship between the features and the target variables and make reasonable predictions for the output to a new dataset. Prediction made by the model is usually validated by utilising an external validation set. A schematic diagram of how supervised ML models use labelled data to fit and prepare themselves to make a prediction on unseen cases is illustrated in Figure 1.5.

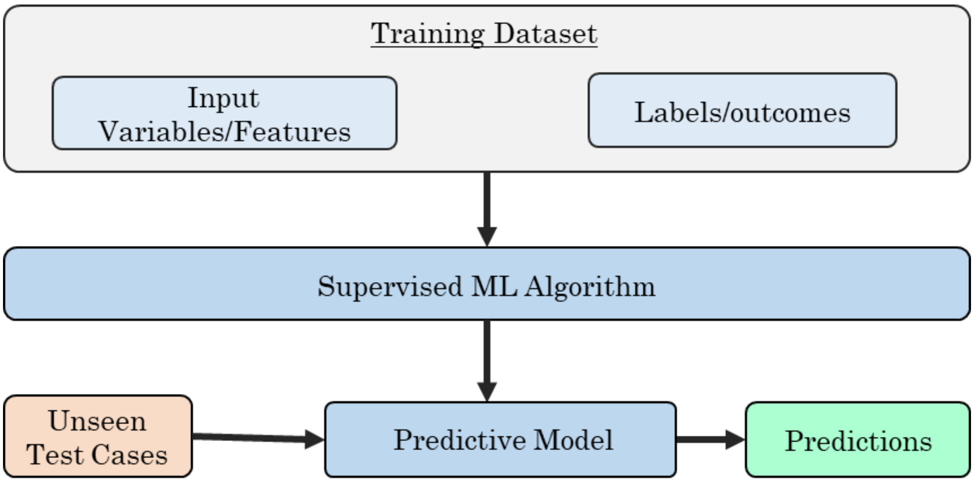


Figure 1.5 Schematic diagram of supervised ML models

Supervised ML models are often known as predictive analytic models based on their ability to predict the future based on the past. Supervised machine learning helps convert raw data into actionable insights. This in turn helps researchers to utilise the data to understand and prevent unwanted outcomes, and in some cases make decisions faster. In supervised ML, each row in the dataset is known as a training instance, the target feature is known as labels or outcomes and the overall dataset used to train the model is referred to as a training set. A simplified supervised ML algorithm can be described by an equation shown in Equation 1.12.

$$Y = f(x) \qquad \text{Equation 1.12}$$

where, x is the known input variables and Y is the output.

The two fundamental categories of supervised ML are regression and classification.

1.3.4.2.1 Regression

Regression supervised models aim to predict a continuous measurement for an observation. The predicted results are represented by a quantity or number that can be flexibly determined based on the inputs of the model rather than being confined to a set of possible discrete labels. Some of the commonly used regression supervised algorithms include: (regularised) linear regression, regression trees, support vector machines and ordinary least squares to name a few.

1.3.4.2.2 Classification

Classification supervised models aim to assign labels from a set of finite labels to an observation. In simple terms, classification models attempt to predict a categorical response, such as "blue" or "black", "disease" or "no disease" or "spam" or "no spam". Classification models with two categories or labels are known as binary classification and more than two labels are known as multiclass classification. Some of the commonly used supervised algorithms for classification include decision trees, random forest classification models, gradient boosted trees, Support Vector Machine (SVM), logistic regression, k-nearest neighbour and naïve Bayes to name a few. An example of a supervised SVM classification is shown in Figure 1.6 presenting two species (versicolour and virginica) as labels from the multivariate iris dataset and the algorithms attempt to classify the samples into these two categories (Andrews & Herzberg, 1985).

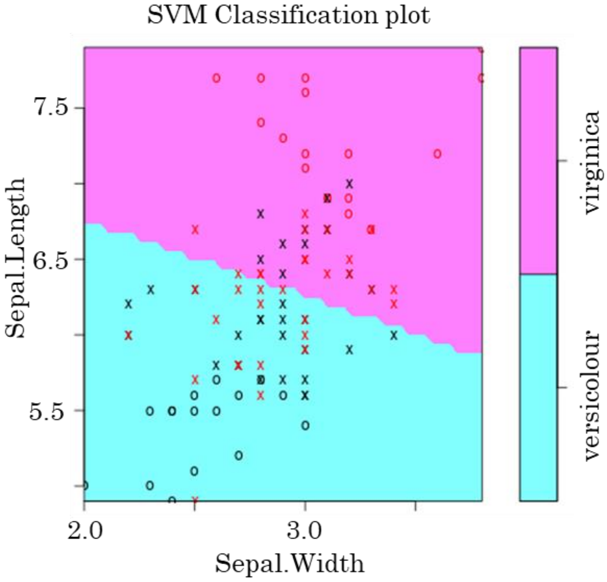


Figure 1.6 Example of an SVM supervised classification model built on the iris dataset showing the resultant classification of the two labels: versicolour and virginica.

1.3.4.2.2.1 Decision tree

Decision trees are one of the most useful and powerful supervised ML algorithms. It is an efficient non-parametric supervised approach that is mostly used for classification and regression of variables. It is a hierarchical model, encoded as a tree, for supervised learning where the local region is identified in a sequence of recursive splits in a number of smaller steps (Alpaydin, 2004). A simple illustration of decision trees is shown in Figure 1.7. From Figure 1.7, it can be observed that the decision trees consist of various internal decision nodes that represent an attribute or feature. These internal decision nodes specify all possible tests on a single attribute-value, with one branch and sub-tree for each possible outcome of the test (Chahal, 2013). These internal decision nodes were first obtained by splitting the overall data points at the root node into two homogeneous sets. The decision nodes undergo further splitting to form terminal nodes or leaf nodes which represents a classification or decision. If the outcomes are continuous, the internal decision nodes can test the value of an attribute against a threshold. The general algorithm of decision trees starts with picking the attribute which is the one that best classifies the training data. The algorithm then keeps asking relevant questions to split the training items into even smaller subsets resulting in a “tree”. The questions are asked until it has no effect on the purity of the subsets or the leaf nodes can no longer be further subdivided. The basic algorithm utilised in decision trees is the Iterative Dichotomizer 3 (ID3) algorithm by J.R. Quinlan which builds the trees using a top-down greedy approach to create the shallowest decision tree that is consistent with the dataset (Kingsford & Salzberg, 2008).

The goal while building a decision tree is to split the attributes in order to create the purest child nodes possible. To identify the best-suited attributes, some of the measures utilised are entropy, information gain, gain index and gain ratio. As the goal of decision trees is to classify the data, information gain and entropy are used by ID3 to identify the best split of the attributes and thus calculate the homogeneity of a sample.

Entropy in terms of machine learning is the measure of disorder, uncertainty or randomness. It is an indicator of how messy the data is. For a given dataset of ‘N’ number of samples, with two categories of *Class A* (n) and *Class B* (m = N - n), entropy is given by the equation shown in Equation 1.13.

$$\mathrm{E} = -\mathrm{p}\log_2(\mathrm{p}) - \mathrm{q}\log_2(\mathrm{q}) \qquad \text{Equation 1.13}$$

where, p is the ratio of n divided by N (p = n/N) and q is the ratio of m divided by N (q = m/N or, q = 1 – p).

Entropy is an absolute measure with values ranging between 0 and 1. The entropy value of 0 indicates that the sample is completely homogeneous while the value of 1 indicates an equally divided sample (Quinlan, 1986). Similarly, information gain measures how much information a feature gives us about the class or labels. Attributes with the highest information gain will split first. The equation for measuring information gain is presented in Equation 1.14. (Kotsiantis, 2013).

Inforamation gain = entropy (parent) – average entropy (child) Equation 1.14

One of the main advantages of decision trees is the easiness in the interpretability of the model's output. The graphical illustration is very intuitive and presents the visual representation of all possible outcomes, rewards and decisions in a single document. The model also assigns specific values to problems and outcomes of each decision which reduces ambiguity in decision making. Furthermore, the models also make a comprehensive analysis of the consequences of each possible decision, thus making it a very good predictive model. Decision trees are very efficient models as no normalisation of the input variables is required and they are thus resistant to outliers and missing values. The models are also capable of handling both categorical and continuous variables as well as non-linear dataset (Brijain, Patel, Kushik, & Rana, 2014; Somvanshi & Chavan, 2016).

However, decision tree models can become computationally complex if proper control and regulation measures are not taken during the growing stage. The biggest disadvantage of decision tree models is that it is prone to overfitting. Overfitting is when the model fits the training set very well but fails to make an accurate prediction on the test set. Various reasons such as having large dimensional data containing meaningless or irrelevant variables and having a small training set size can lead to overfitting. However, problems of overfitting in decision trees can be solved by either pre-pruning or post-pruning methods. The pruning process involves cutting down the trees which are done by stopping the algorithm before growing the trees to a full model. The model is usually stopped if all the attributes belong to the same class or if all the attributes' values are the same. Decision trees can also be stopped from growing if the number of instances is less than a defined threshold or growing the nodes does not improve purity. Post-pruning is more popular than pre-pruning. It involves growing the decision tree to its entirety. The data is then split into training and internal validation sets. One node is removed at a time and tested on the validation data to see if the performance improves. Nodes that increase the decrease in performance are removed. This method usually produces the smallest trees (Bramer, 2013).

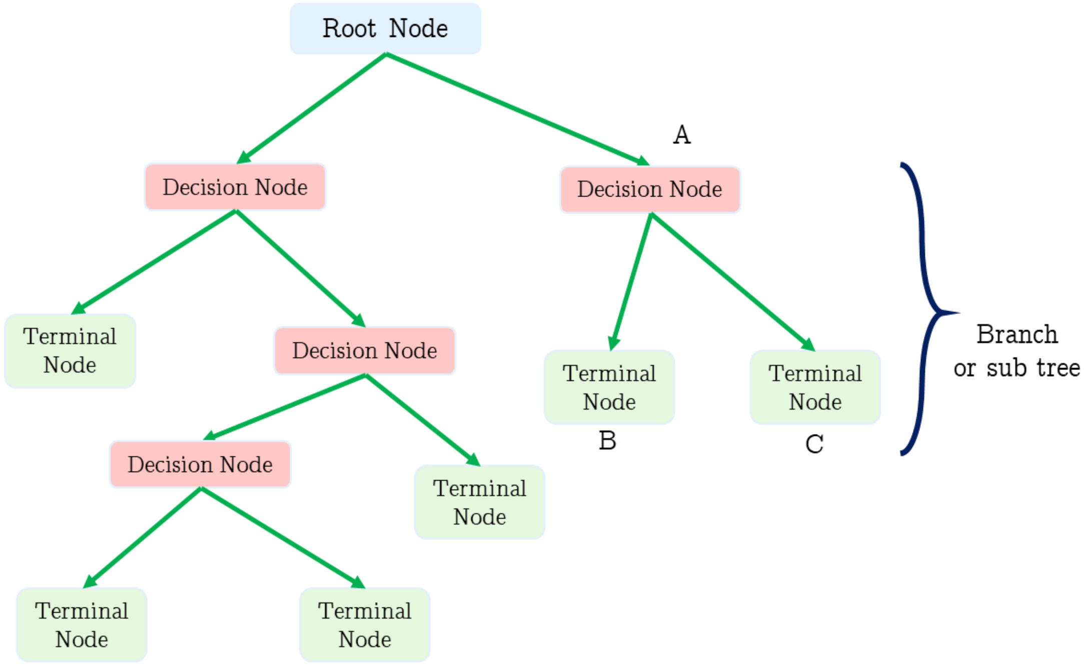


Figure 1.7 Illustration of a decision tree classification model. Here A is the parent node of B and C (child nodes). The decision node A along with the terminal nodes B and C is known as the branch of the decision tree.

1.3.4.2.2.2 Random forest (RF)

Random forest (RF) is an ensemble supervised method introduced by Leo Briemann and Adele Cutler (Breiman, Cutler, Liaw, & Wiener, 2015). Ensemble methods utilise a divide-and-conquer approach to improve performance. The main idea behind ensemble methods is the grouping of 'weak learners' together to form 'a strong learner'. As mentioned in Section 1.3.4.2.2.1, decision trees are susceptible to suffer from being a high-variance estimator i.e. making small incremental changes in the training observations can drastically alter the predictive performance of the learned tree. This problem can be mitigated by utilising bootstrap aggregation or the bagging method. Bagging ensemble method randomly samples subsets of the training dataset with replacement, fitting a decision tree to each and aggregating their result thus reducing variance. Furthermore, the bagging method utilises the entire set of variables when creating splits in the nodes which allows the decision trees to grow without pruning and thus aids in reducing tree-depth size and variance. However, it is to be noted that utilisation of the entire set of variables creates a risk of correlation between the decision trees which can increase bias in the model. This problem of correlation between the decision trees is reduced by Random Forest which selects only a subsample of the variables at each node split. The random forest algorithm is in itself primarily based on the bagging and random subspace paradigm. RF models work by taking the dataset and creating random samples with replacement to build a decision tree using each sample as the training set. Each tree is trained on roughly 2/3rd of the total training data (63.2%) while the remainder is used for the calculation of the Out-Of-Bag (OOB) error rate. The number of trees (*ntree)* grown during the building of the random forest model is raised incrementally until there is no further improvement observed on the model. At each node of the tree, some predictor variables (*mtry*) are selected at random out of all the predictor variables. The best split of *mtry* is then used to split the nodes and is held constant while the forest is being grown. The default value of *mtry* in the RF model is defined by the square root of the total number of predictor variables for classification and the total number of predictor variables divided by 3 for regression models (T. Hastie, Tibshirani, & Friedman, 2005). Using both the 2/3rd (bootstrap data) and the remaining 1/3rd (OOB error data) of the dataset, each tree in the random forest model gives a classification or regression. '*nodesize*' refers to the minimum number of terminal nodes below which leaves are no further sub-divided. The default values are different for classification and regression which are 1 and 5 respectively (Breiman, 2001b). Setting the seed in the random forest model enables reproducibility as it is the non-zero integer number that controls the random number generator (Bhardwaj, 2016). The misclassification rate or the Out-Of-Bag (OOB) error rate, is used to determine the strength of the random forest model and thus is used as a guide during the training of the model. RF models have proven very successful and are widely used in the field of chemoinformatics.

RF models have successfully been used in a diverse range of biological, physical and life science applications such as gene selection (Díaz-Uriarte & Andrés, 2006), tumor classification (Shi, Seligson, Belldegrun, Palotie, & Horvath, 2005), prediction of protein-protein interactions (Sikic, Tomic, & Vlahovicek, 2009), fault diagnosis in centrifugal pumps (Y. Wang, Lu, Liu, & Wang, 2016), quantitative structure property relationship (QSPR) (V. Svetnik et al., 2003) and predicting aqueous solubility, hydrate and solvate formations (Johnston, Johnston, Kennedy, & Florence, 2008; Palmer, O'Boyle, Glen, & Mitchell, 2007; Taldeddin, Khimyak, & Fabian, 2016). There are many advantages of implementing RF algorithms as it is highly accurate in solving both classification and regression problems and runs efficiently on large datasets. RF models can also handle datasets with high dimensions without any variable deletion. Furthermore, the model outputs variables according to their importance which is one of its main advantages. The model has the capability of generating an internal unbiased estimate of the generalization error as the forest building progresses. RF models can also be utilised in unlabelled datasets leading to unsupervised clustering. Even for labelled datasets, RF models can compute proximities between pairs of classes which can then be used in clustering and outlier detection. Even though RF models have many advantages over other ML algorithms, there are some disadvantages too. RF models are difficult to interpret and are sometimes referred to as 'black box' models. The models have also been observed to overfit for some datasets with noisy classification or regression problems. If the dataset is highly imbalanced then RF models can be biased in favour of the majority classed labels. This can result in unreliable variable importance scores. (Y. J. Qi, 2012; Sullivan, 2018; Wilson, 2017). Random forest model built on an ensemble of decision trees is illustrated in Figure 1.8.

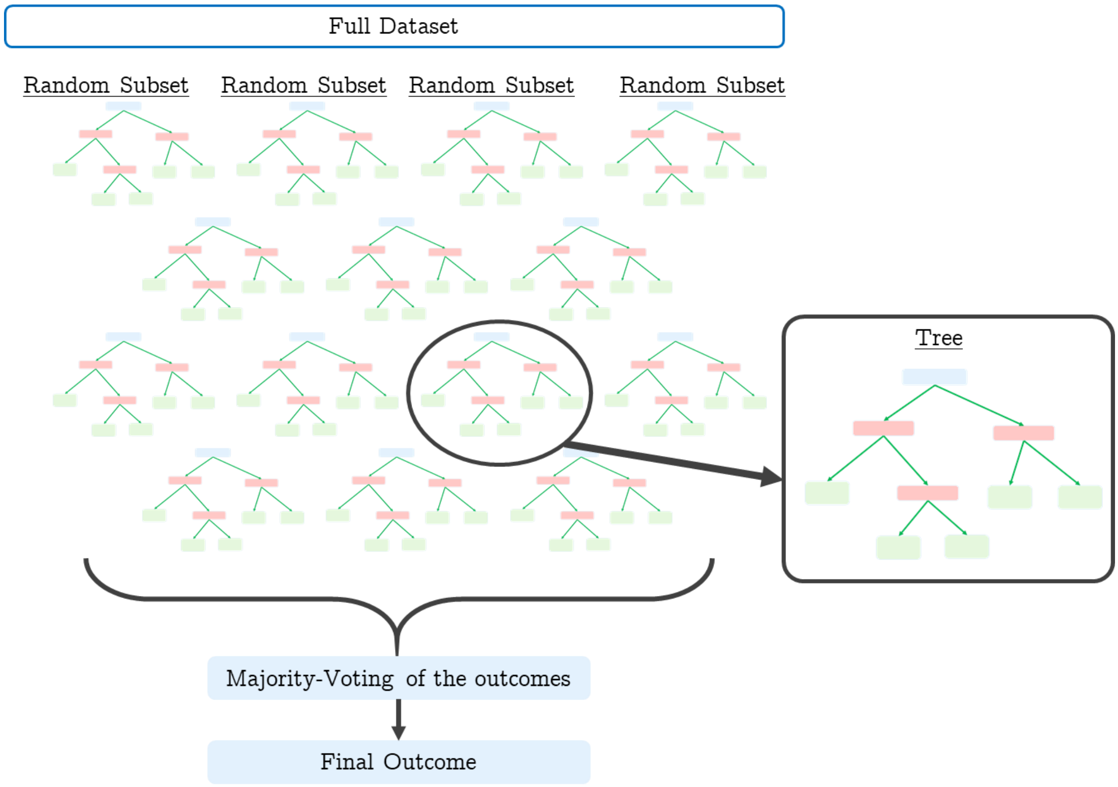


Figure 1.8 Illustration of many decision trees forming a random forest model.

1.3.4.3 Reinforcement ML algorithm

Reinforcement learning algorithms, also known as semi-supervised machine learning algorithms, are goal-oriented machine learning models that learn from the interactive environment through trial and error as feedbacks from their own actions and experiences. A systematic workflow of the reinforcement algorithm is presented in Figure 1.9. Various applications of reinforcement learning algorithms include self-navigating vacuum cleaners, driverless cars, traffic light controls etc. Reinforcement ML algorithms are not simple to build and the problems are tackled by a plethora of algorithms (Sutton & Barto, 2018; Z. P. Zhou, Li, & Zare, 2017).

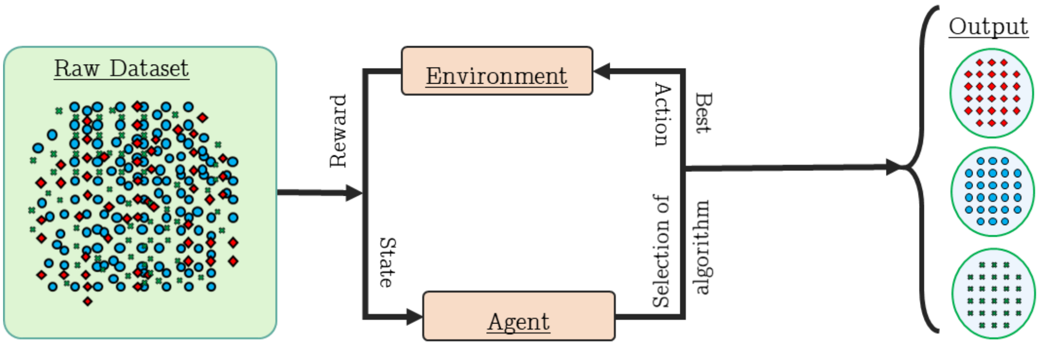


Figure 1.9 Illustration of reinforcement ML algorithm

1.3.5 The performance measure of ML algorithms

Determining an absolute measure of prediction is greatly dependent upon the selected machine learning algorithm as well as well as the dataset. Different statistical metrics can help to ensure that the trained classification and regressions models are unbiased and capable of accurate predictions.

1.3.5.1 Classification evaluation metrics

There are many metrics available for evaluating classification supervised models. Some of the popular metrics include accuracy, confusion matrix, Area under the ROC Curve (AUC), precision and log-loss to name a few. Some of the classification metrics utilised in the thesis are explained in detail in the following sections.

1.3.5.1.1 Classification accuracy (ACC)

Classification accuracy is simply the measure of the fraction of correct predictions made by the classifier. It is formally defined as the ratio between the numbers of correct predictions to the total number of predictions as shown in Equation 1.15.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \text{Equation 1.15}$$

1.3.5.1.2 Confusion matrix

The confusion matrix, n x n where n is the number of labels, is a table that helps in summarizing and visualising the predictive performance of the classification supervised model. It is also known as an error matrix. The rows of the confusion matrix present the actual classes that were observed while the columns present the model's predicted outcomes (Table 1.6).

Table 1.6 Example of a 2 x 2 confusion matrix

Predicted		Actual Class	
		Class A	Class B
	Class A	True Positive (TP)	False Negative (FN)
	Class B	False Positive (FP)	True Negative (TN)

The confusion matrix helps visualise how accurate the classification model is by exposing how frequently the model confuses or mislabels the two classes. It not only provides insight into the errors being made by the classification model but also shows the types of errors being made. For a classification model with two possible outcomes as shown in Table 1.6, a True Positive (*TP*) is the outcome where the model correctly predicts the positive class (*Class A*) while True Negative (*TN*) is the outcomes where the model correctly predicts the negative class (*Class B)*. Similarly, False Positive (*FP*) is the outcome where the model incorrectly predicts the positive class and False Negative (*FN*) is when the model incorrectly predicts the negative class (Max Kuhn & Johnson, 2013). Sensitivity or recall is the ability of the classification model to correctly identify all the positively classified values while specificity or True Negative Rate (*TNR)* is the ability of the RF classification model to correctly identify all the negatively classified values. The best value for both sensitivity and specificity is 1 while the worst value is 0. Precision or Positive Predictive Value (PPV) is the ratio of correctly detected positive instances. Higher the precision score, the higher is the confidence in the capability of the model to classify. Equations defining sensitivity, specificity and precision are presented in Equation 16, Equation 17 and Equation 18 respectively.

$$\text{Sensitivity } = \frac{TP}{TP + FN} \qquad \text{Equation 1.16}$$

$$\text{Specificity } = \frac{TN}{TN + FP} \qquad \text{Equation 1.17}$$

$$\text{Precision } = \frac{TP}{TP + FP} = \frac{True\ Positive}{Total\ Predicted\ Positive} \qquad \text{Equation 1.18}$$

1.3.5.1.3 Cohen's kappa (ƙ)

Cohen's kappa (*ƙ*) is a robust statistical measure of how well the classification model performed as compared to how well it would have performed simply by chance. Cohen's kappa can range from 0 to 1. Table 1.7 presents the interpretation of Cohen's kappa. The value of *ƙ* at 1 indicates a perfect agreement between the raters while value at 0 indicates that the agreement is equivalent to random chance. It is to be noted that there is a possibility for the value of *ƙ* to be negative but it is very unlikely to be negative in practice (Marston, 2010).

Table 1.7 Logical interpretation of Cohen's Kappa (*ƙ*) (McHugh, 2012)

Cohen's Kappa (ƙ)	Level of agreement
0	Agreement equivalent to random chance
0.1 – 0.2	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Near perfect agreement
1	Perfect agreement

Equation 1.19 presents the formula to calculate Cohen's kappa (Berry & Mielke, 1988).

$$ƙ = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \qquad \text{Equation 1.19}$$

where, p_o is the relative observed agreement among raters while p_e is the hypothetical probability of chance agreement

1.3.5.2 Regression evaluation metrics

In order to measure and assess the performance of regression models, three standard metrics i.e. Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared (R^2) are used. RMSE is the measure of the average deviation of the predicted data from the experimental observations (residuals), MAE is the absolute difference between the predicted data and the experimental observations while R-squared is the relative measure of how close the predicted data from the RF model are to the fitted regression line (Chirico & Gramatica, 2011). A picture of a good predictive model is dictated by relatively low values of RMSE and MAE whilst higher values of R-squared. The values of both RMSE and MAE range from 0 to infinity. Between the two metrics, MAE acts as a better indicator of average model performance over RMSE. This is because higher values of RMSE are influenced by the presence of the small number of high error predictions as squaring the higher prediction error will add more weight than the lower prediction errors. MAE, however are devoid of such complex parameterization and provides a straight forward determinant of prediction errors (Willmott & Matsuura, 2005). Contrary to Willmott and Matsura, Chai and Draxler (2014) stated that one cannot simply choose MAE whilst avoiding RMSE as RMSE are more appropriate when large errors are particularly undesirable (Chai & Draxler, 2014). RMSE, MAE and R^2 are mathematically defined by the following equations shown in Equation 1.20, Equation 1.21 and Equation 1.22 respectively.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}$$ Equation 1.20

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|$$ Equation 1.21

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$ Equation 1.22

where, n is the number of observations, Y_i is the experimentally observed outcome for the *i*th compound, $\widehat{Y}_i$ is the predicted outcome for the *i*th compound and $\overline{Y}$ is the average response of the training compounds

Chapter 2.　Aims and objectives

2.1 Aims

The aim of this project is to develop and implement predictive machine learning models at the earlier stages of the continuous crystallization workflows (C. J. Brown et al., 2018) in order to reduce the number of experiments necessary for full process design, understand the relationship between API properties of interest and readily-calculated physicochemical descriptors and help to automate the screening and crystallisation process development. The work also aims to widen the application of ML in the field of crystallisation by investigating the optimal training set required to build reliable predictive models, develop a rapid and efficient solvent selection tool for recommending suitable solvents for crystallisation process design and predict non-aqueous solubility of drugs and drug-like compounds in a range of diverse solvents.

i. Using ML, is it possible to reliably predict the crystallisation outcomes and crystal habit of a drug API using a set of calculated molecular descriptors?

ii. Can ML be used to identify the optimum number of experiments required to build a robust model for solvent selection? Is there a rapid and efficient way of recommending solvents for the crystallisation process design?

iii. Can ML be used to develop an in-silico method for estimating non-aqueous solubility of the drug and drug-like compounds in a diverse range of solvents?

2.2 Objectives

i. Develop experimental methodologies to investigate crystallisation behaviour and crystal habit of an API by performing small-scale cooling crystallisation experiments on a diverse range of organic solvents.

ii. Using *Technobis* Crystalline reactor system to determine solubilisation and nucleation and optical Leica microscope and Morphologi G3 to assess the obtained crystal habit.

iii. Develop a machine learning model pipeline to predict the various crystallisation outcomes and crystal habit of an API in a diverse range of solvents.

iv. Using *Technobis* Crystal16 to analyse solubilisation of paracetamol, carbamazepine and carvedilol on the diverse range of solvents at two temperature points. Qualitative solubility outcomes, i.e. *soluble* and *practically insoluble* were determined from transmissivity observations in *Crystal16*.

v. Using the experimental database to develop a rapid and efficient solvent selection tool using ML algorithms. Investigate the 'optimal' number of experimental data-points by assessment of the training set size required to build a reliable and robust predictive ML model for solvent selection.

vi. Create a *non-aqueous* solubility database of drug and drug-like compounds by curating data from various literature sources and publicly available databases. Solubility datapoints were collected for compounds on commonly used solvents at laboratory temperature.

vii. With the set of calculated physicochemical descriptors, both regression and classification models were built to predict the non-aqueous solubility of drugs and drug-like compounds in a diverse range of solvents.

Chapter 3.　Material and methods

3.1 Materials

Paracetamol (Form I, CAS ID: 103-90-2) was purchased from Molekula Ltd, UK while carbamazepine (Form III, CAS ID: 298-46-4) and carvedilol (Form I, CAS ID: 72956-09-3) were purchased from Sigma Aldrich, UK. The physical properties of the three compounds are presented in Table 3.1. X-ray powder diffraction (XRPD) was used to confirm the phase identification and purity of all the purchased three compounds used in this thesis.

Table 3.1 Physical properties of the compounds selected in this thesis (Haynes, 2016; Whitesell, 1998).

Compound	Molecular Formula	Molecular Weight (g/mol)	Melting Point (^{0}C)
Paracetamol	$C_8H_9NO_2$	151.163	168
Carbamazepine	$C_{15}H_{12}N_2O$	236.27	190.2
Carvedilol	$C_{24}H_{26}N_2O_4$	406.47	114.5

Paracetamol is a potent antipyretic and analgesic agent widely used for the relief of headaches and other minor pains. It exists in five reported polymorphic forms I, II, III, IV and V as well as a number of solvates (Heng & Williams, 2006; Lee, 2014; S. J. Smith, Bishop, Montgomery, Hamilton, & Vohra, 2014). Paracetamol Form I is used commercially and is the thermodynamically stable form. Whilst form I crystallises as a monoclinic lattice, both the metastable Form II and the highly unstable polymorphic form III exist in orthorhombic systems. Polymorphic Form IV and V are only obtained under high pressure (Espeau et al., 2005; Hiendrawan et al., 2016; S. J. Smith et al., 2014). Paracetamol has low solubility in non-polar and chlorinated hydrocarbons and high solubility in solvents with medium polarity. Its solubility in water is lower compared to other polar solvents. (Granberg & Rasmuson, 1999; Romero, Reillo, Escalera, & Bustamante, 1996). The solubility of Form II is slightly higher than that of Form I (Joiris, Di Martino, Berneron, Guyot-Hermann, & Guyot, 1998).

Carbamazepine is a first-generation anticonvulsant drug that has been used in the treatment of partial seizures, trigeminal neuralgia, manic-depressive illness, and explosive aggression (W. J. Liu, L. P. Dang, S. Black, & H. Y. Wei, 2008). Over the years, it has been a widely studied model system in the investigation of crystal polymorphism and co-crystallisation. It has been reported in the literature to crystallise in five anhydrous polymorphs and a dihydrate as well as many other solvates (Arlin, Price, Price, & Florence, 2011; Florence, 2016). Form III (monoclinic) is the commercially available and thermodynamic stable form of carbamazepine. Carbamazepine is reported to be poorly soluble in water, sparingly soluble in ethanol, isopropanol, butanol and acetone and readily soluble in methanol, dichloromethane and tetrahydrofuran (Alrashood, 2016; Kumar & Siril, 2014; W. J. Liu et al., 2008).

Carvedilol, a weak base, comes under the class of alpha and beta blockers and is widely used for the treatment of cardiovascular diseases such as hypertension, congestive heart failure, cardiac arrhythmias, myocardial infarction and angina pectoris (Feuerstein & Ruffolo, 1995; Wen, Tan, Jing, & Liu, 2004). Many polymorphic and pseudo polymorphic forms of carvedilol have been reported in literature and patents including its three polymorphic non-solvates forms I, form II, and form III. All the polymorphic non-solvates forms of carvedilol crystallise in the monoclinic system. Form I was identified as the most thermodynamically stable form of Carvedilol (Pataki, Markovits, Vajna, Nagy, & Marosi, 2012; Prado, Rocha, Resende, Ferreira, & de Figuereido Teixeira, 2014). Carvedilol was reported to be practically insoluble in water, sparingly soluble in ethanol and isopropanol and readily soluble in dimethyl sulfoxide, methanol and methylene chloride (Beattie, Phadke, & Novakovic, 2013; Brittain, 2013; Menon, Mistry, Joshi, Modi, & Shashtri, 2012; Planinsek, Kovacic, & Vrecer, 2011). These compounds were chosen in line with a greater piece of work as part of the Continuous Manufacturing and

Crystallisation Future Manufacturing research hub. Molecular structures of paracetamol, carbamazepine and carvedilol are shown in Figure 3.1.

The solvents utilised for the cooling crystallisation experiments in this thesis were of analytical grade and were all purchased from both Sigma Aldrich, UK and Fisher Scientific, UK (Table 3.2). The key driver for the choice of organic solvents was based on chemical diversity and availability in the laboratory; the choice was not limited to pharmaceutically acceptable solvents to create a larger dataset.

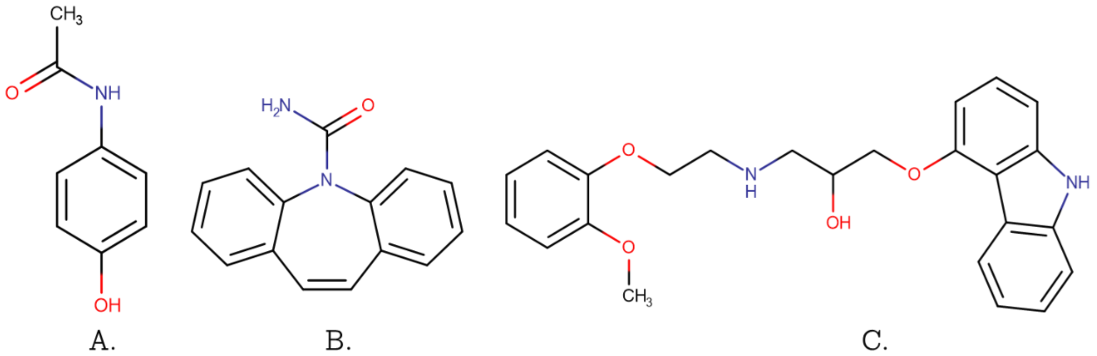


Figure 3.1 Molecular structure of A. paracetamol, B. carbamazepine and C. carvedilol

Table 3.2 List of 94 solvents and their respective family class used for cooling crystallisation.

Solvent Family	List of solvents
Alcohols	methanol, ethanol, 1-propanol, 2-propanol, 1-butanol, 2-butanol, 2-pentanol, 1-octanol, 1-decanol, 2-phenylethanol, benzyl alcohol, 2,2,2-trifluoroethanol, isoamyl alcohol, 2-methyl-1-propananol, cyclohexanol, 3-pentanol, 1-nonanol
Acids	acetic acid, formic acid, trifluoroacetic acid
Ester	ethyl acetate, butyl acetate, isobutyl acetate, diethyl carbonate, formamide, pentyl acetate, ethyl lactate
Ether	anisole, 1,4-dioxane, 2-phenoxyethanol, 2-butoxyethanol, 2-methoxyethanol, methoxyethane, 2-ethoxyethanol, 1,2-

	dimethoxyethane, 2-methoxy-2-methylpropane, cyclopentane, diisopropyl ether, dibutyl ether, diethyl ether, 2-methoxyethyl ether (Diglyme)
Ketone	acetone, 2-butanone, 3-pentanone, 4-methyl-2-pentanone
Polar Aprotic	acetonitrile, n, n-dimethylacetamide, n, n-dimethylformamide, tetrahydrofuran, dimethyl sulfoxide, n-methyl-2-pyrrolidone
Halogenated	1-bromo-2-chloroethane, diiodomethane, 1-bromobutane, 2-bromobutane, bromoform, bromobenzene, chloroform, 1-chlorobutane, chlorobenzene, 1,2-dichloroethane, dichloromethane, iodomethane, carbon tetrachloride, tetrachloroethene, trichloroethylene
Nitro	nitrobenzene, nitromethane
Aromatics	aniline, cumene, 1-methylnaphthalene, benzene, toluene, m-xylene
Diols	1,2-propanediol, 1,4-butanediol
Water	water
Amines	phenethylamine, pyridine, triethylamine, 2-amino-1-butanol
Hydrocarbons	2,2,4-trimethylpentane, 2-methyl butane, hexane, iso-hexane, methyl cyclohexane, n-dodecane, cyclohexane, heptane
Thiols	ethanethiol, 1-propanethiol
Organosulfur	3-methylthiophene, sulfolane, tetrahydrothiophene

3.2 Experimental methodology

3.2.1 Crystallisation techniques

To investigate the crystallisation outcomes, crystal morphology and to determine qualitative solubility, the cooling crystallisation technique was used. The main reasons for this was that temperature can be accurately controlled and secondly it has a wider applicability to the processes within the pharmaceutical industry. Even though evaporative crystallisation requires less apparatus and material, it presents challenges around controlling the evaporation rate and as a result was not selected.

3.2.1.1 Controlled cooling crystallisation for assessing crystallisation outcomes

XRPD was carried out at the end of the experiment to verify that only the morphology of the crystal had changed during the crystallisation process and that the API was at its thermodynamically stable form. A fixed weight of the target compound was initially measured at 5 wt/wt % and dissolved in the chosen solvent at laboratory temperature of 25^0 C. The suspension was placed inside *Technobis* Crystalline Reactor systems (Avantium, The Netherlands), a multi-reactor parallel crystalliser containing both turbidity sensors and real-time particle viewers for visualisation of the complete crystallisation process (Figure 3.2). The in-built turbidity sensors allowed the determination of clear and cloud points while the crystalline cameras provided real-time visualisation of the crystallisation process. The suspension was left to agitate at set stirring speed for an hour. The temperature was then gradually increased to 10^0 C below the boiling point of the respective solvent at a slow and constant heating rate. The temperature at which the suspension turned clear (transmissivity observed near 100% and no suspension was observed in the particle viewer) was assumed to be the saturation temperature. Once the solution was clear, the supersaturated solution was then slowly cooled to 20^0 C at a slow cooling rate of 0.1^0C/min and stored to allow maximum recovery of the crystals. Thus, the obtained crystals were gently filtered and dried in a vacuum oven. The experiment was repeated four times for reproducibility. If the suspension did not dissolve, when the temperature was increased, the concentration of the target compound was reduced to half by doubling the volume of the solvent, and the crystallisation steps were repeated. Similarly, if the suspension was found to be readily soluble at lab temperature, the weight of the target API was doubled in the respective solvent and the crystallisation steps repeated.

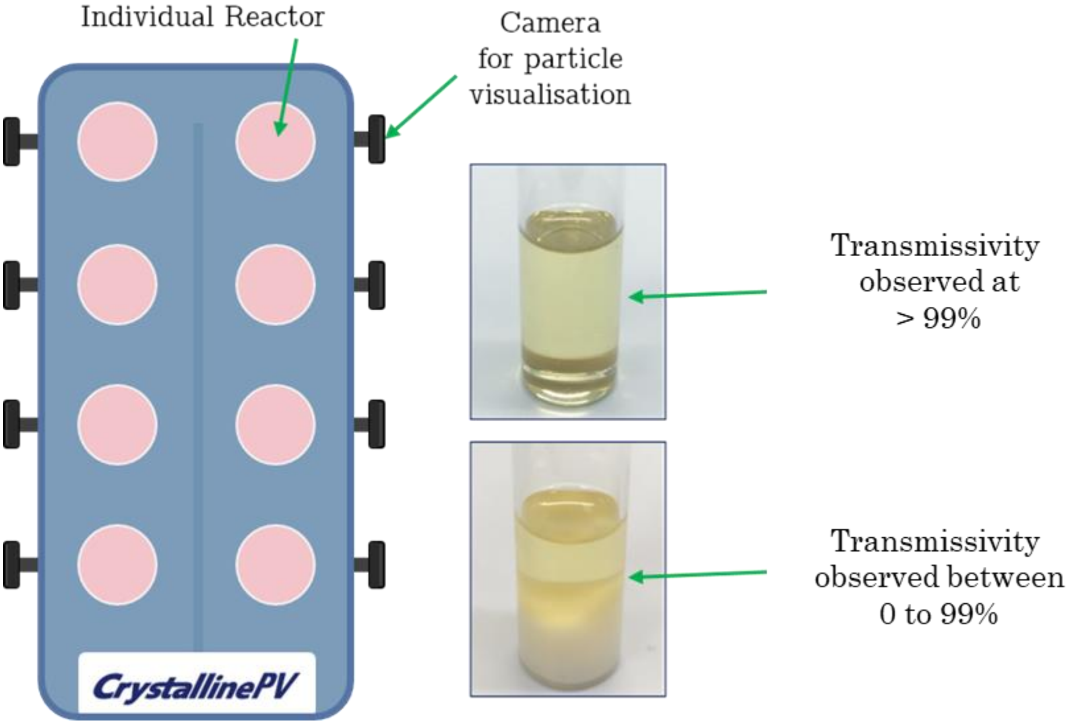

Individual Reactor
Camera for particle visualisation
CrystallinePV
Transmissivity observed at > 99%
Transmissivity observed between 0 to 99%

Figure 3.2 Cooling crystallisation performed on *Technobis Crystalline* Reactor system for observing crystallisation outcomes and crystal habit of an API

3.2.1.2 Controlled cooling crystallisation for solubility measurement

Various methods are available in the literature for measuring the solubility of solids in liquids, and the methodologies are mainly dependent upon the solvent's properties, availability of the compound and analytical instruments, the precision required or the need for additional solid phase characterisation (Mullin, 2001a). In this thesis, a simple qualitative assessment of solubility measurement was performed on the assessment of whether the compound was soluble or practically insoluble in the particular solvent at a given temperature point. Measurement of solubility was performed using Crystal16 Reactor systems (similar to *Technobis*

Crystalline), a multi-reactor benchtop parallel crystallizer which can accommodate 16 HPLC sample vials in one run (Figure 3.3). Each reactor is equipped with turbidity transmission sensors which detect the dissolution process for derivation of saturated temperatures (clear point). Calibration of the instrument was performed using vials filled with just the respective solvents before every experiment to reduce any noise or errors during the measurement. To measure solubility, a set measured volume of solvent and solid were placed in a sample vial. Measurements are performed in a closed sample vial and covered with paraffin to prevent loss of solvent by evaporation. The suspension was then stirred at a moderate rate and a set constant temperature. The temperature was initially set at laboratory temperature of 25^0 C and the suspension stirred for a set time. If the suspension at set lab temperature is found to be clear (transmissivity observed 100%), the solution is determined to be under-saturated and thus qualitatively categorised as *soluble.* If the suspension is still present after an hour of stirring in lab temperature (transmissivity observed ≥ 100%), the solution is determined to be insoluble and thus qualitatively categorised as *practically insoluble*. If the suspension was still present at low temperatures, a slow controlled stepwise heating was performed. The temperature increment was set to 10^0C below the boiling point of the solvent. The temperature point where the suspension of known concentration dissolves marks the point of the solubility line and is known as the saturation temperature. The temperature when increased was not allowed to exceed 100^0 C for health and safety reasons.

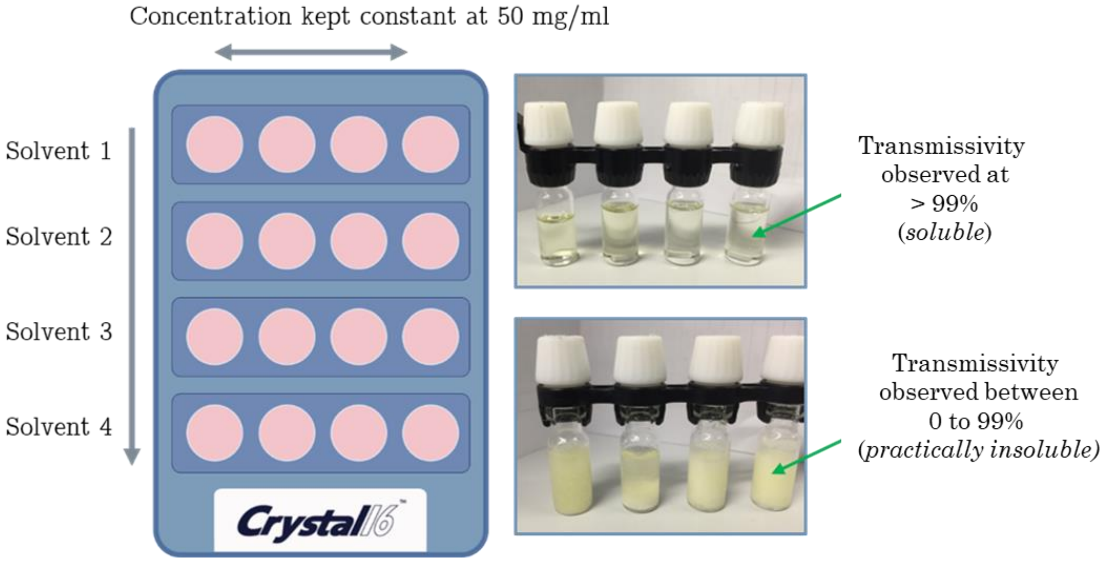

Concentration kept constant at 50 mg/ml
Solvent 1
Solvent 2
Solvent 3
Solvent 4
Crystal16
Transmissivity observed at > 99% (soluble)
Transmissivity observed between 0 to 99% (practically insoluble)

Figure 3.3 Cooling crystallisation experiment performed on *Technobis Crystal16* for the solvent screening study.

3.2.2 Zinsser automated platform

Precise solid loading and solvent dispensing steps were performed on the custom-built robotic platform acquired from Zinsser Analytics (http://www.zinsser-analytic.com). Though slightly different in the platform layout, the automated platform has been introduced and described in detail by Schuldt et al. (Schuldt & Schembecker, 2013). Powder dispensing was performed using the REDI 2002 plus, an X, Y, Z dispensing system consisting of a powder pipette and fitted with REDI VARIX©, a software controlled variable volume tip which can be adjusted for the required weight. These tips can be picked up and exchanged for each powder to prevent cross-contamination. The platform consisted of a gripper tool capable of transporting single vials of 1.5 ml and 8 ml vials as well as complete racks where up to 24 vials can be placed. The gripper tool would transport each vial to the balance where the required weight of the powder was precisely dispensed. For liquid handling, the pipetting arm was equipped with three standard pipettes and a pipette containing a filter to prevent the draw-in of solids. All operations and methods were programmed and launched via the Zinsser WinLissy Software Version 8.1.0.

3.2.3 Optical microscopy

Various crystal morphologies of paracetamol obtained from cooling crystallisation experiments were observed using the Leica DM6000M microscope (Leica, Buckinghamshire, UK). The images were visualized and captured using the LAS-AF software version 2.6.0 (Leica).

3.2.4 Malvern Morphologi G3-ID

Quantitative morphological characterization of paracetamol obtained from cooling crystallisation on various solvents (Section 3.2.2) was analysed using the Malvern Morphologi® G3 with Automated Particle Characterization System (Malvern Instruments Ltd, UK). It provides the ability to analyse, measure and characterise the size and shape of thousands of particles at a given time. The instrument comprised of a sample dispersion unit (SDU) where compressed air was supplied to disperse the sample at a set pressure, a glass plate to collect the dispersed sample, a CCD Firewire™ camera for capturing images and four optical microscopes of different magnifications (5x, 10x, 20 x and 50x) for particle analysis. Analysis of the crystal morphologies was performed using the Morphologi software version 8.12. The Morphologi software also allows the user to set up a standard operating procedure (SOP) where dispersion pressure, injection time, sample scan area and required sample characterisation can be tailored according to specific needs for analysis. The software presents the analysis in various scattergram plots which presents particle size distribution, class and filter information and particle images. There is also a comparison tab which provides the comparison of all the morphological distribution of multiple measurements.

3.3 Machine learning methodology

3.3.1 Molecular structure representation

Both the compounds and solvents in this thesis were expressed in SMILES which uniquely encodes the structure of a molecule in a single line using standard text characters.

3.3.2 Molecular descriptors

For this thesis, 340 2-D and 3-D molecular descriptors of both the solvents and APIs were calculated using the Chemical Computing Group's Molecular Operating Environment (MOE) software version 2014.09 (Boyd, 2005). The 340 descriptors consisted of a list of spatial, electronic, thermodynamic, conformational, topological, quantum mechanical and structural descriptors. Before calculating the molecular descriptors, the 3-D molecular structures of either the solvents or compounds (depending upon the study) were constructed from their canonical SMILES in Discovery Studio's Biovia Pipeline pilot 2017 software. The calculated 2-D descriptors help define the fundamental numerical properties which were calculated from the connection table representation of molecule such as formal charges and valence bonds but not atomic coordinates. These 2-D descriptors include the calculated physical properties, sub-divided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, partial charge and pharmacophore feature descriptors. Calculated 3-D descriptors, however, are dependent upon the conformation of the molecule and include the potential energy, surface area, shape and volume and the conformation-dependent charge (Vilar, Ferino, Quezada, Santana, & Friedman, 2012). Table 3.3 listed the list of the 2D and 3D molecular descriptors calculated by MOE. The workflow designed to calculate the molecular descriptors of the 63 organic solvents is shown in Figure 3.4.

Table 3.3 List of the calculated 2D and 3D molecular descriptors with a brief description (P. Labute, 2000)

2D Descriptors		
Physical Properties	AM1_dipole, AM1_E, AM1_Eele, AM1_HF, AM1_HF, AM1_IP, AM1_LUMO, AM1_HOMO, apol, bpol, fcharge, mr, smr, weight, logP(o/w), SLogP, vds_vol, density, vdw_area	Physical properties are calculated from the connection table (with no dependence on conformation) of a solvent/molecule
Subdivided surface areas	SlogP_VSA0 to SlogP_VSA9, SMR_VSA0 to SMR_VSA7	The Subdivided Surface Areas are descriptors based on an approximate accessible van der Waals surface area calculation for each atom, v_i along with some other atomic property, p_i
Atom Counts and Bond Counts	a_aro, a_count, a_heavy, a_ICM, a_IC, a_nH,a_nB, a_nC, a_nN, a_nO, a_nF, a_nP, a_nS, a_nCl, a_nBr, a_nI, b_1rotN, b_1rotR, b_ar, b_count, b_double, b_heavy, b_rotN, b_rotR, b_single, b_triple, VAdjMa, VAdjEq	The atom count and bond count descriptors are functions of the counts of atoms and bonds.
Kier&Hall Connectivity and Kappa Shape Indices	hi0, chi0_C, chi1, chi1_C, chi0v, chi0v_C, chi1v, chi1v_C, Kier1 to Kier3, KierA1 to KierA3, KierFlex,	The Kier and Hall kappa molecular shape indices compare the molecular graph with minimal and maximal molecular

	zagreb	graphs, and are intended to capture different aspects of molecular shape
Adjacency and Distance Matrix Descriptors	BalabanJ, diameter, petitjean, radius, VDistEq, VDistMa, weinerPath, weinerPol	The adjacency matrix, M, of a chemical structure is defined by the elements [Mij] where Mij is 1 if atoms i and j are bonded and zero otherwise. The distance matrix, D, of a chemical structure is defined by the elements [Dij] where Dij is the length of the shortest path from atoms i to j; zero is used if atoms i and j are not part of the same connected component
Pharmacophore Feature Descriptors	a_acc, a_acid, a_base, a_don, a_hyd, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_hyd, vsa_other, vsa_pol	The Pharmacophore Atom Type descriptors consider only the heavy atoms of a molecule and assign a type to each atom
Partial Charge Descriptors	Q_PC+, Q_PC-, Q_RPC+, Q_PRC-, Q_VSA_POS, Q_VSA_NEG, Q_VSA_PPOS, Q_VSA_PNEG, Q_VSA_HYD, Q_VSA_POL, Q_VSA_FPOS, Q_VSA_FNEG, Q_VSA_FPPOS, Q_VSA_FPNEG,	Descriptors that depend on the partial charge of each atom of a chemical structure require calculation of those partial charges.

	Q_VSA_FHYD, Q_VSA_FPOL, PEOE_VSA-6 to PEOE_VSA+6	
3D Descriptors		
Potential Energy Descriptors	E, E_ang, E_ele, E_nb, E_oop, E_sol, E_stb, E_str, E_strain, E_tor, E_vdw, E_rele, E_rsol, E_rvdw,	The energy descriptors use the MOE potential energy model to calculate energetic quantities from stored 3D conformations.
Surface Area, Volume and Shape Descriptors	ASA, dens, glob, pmi, pmiX, pmiY, pmiZ, rgyr, std_dim1 To std_dim3, vol, VSA	Descriptors depend on the structure connectivity and conformation
Conformation Dependent Charge Descriptors	ASA+, ASA-, ASA_H, ASA_P, DASA, CASA+, CASA-, DCASA, dipole, dipoleX, dipole, dipoleZ, FASA+, FASA-, FCASA+, FCASA-, FASA_H, FASA_P	Descriptors depend upon the stored partial charges of the molecules and their conformations.

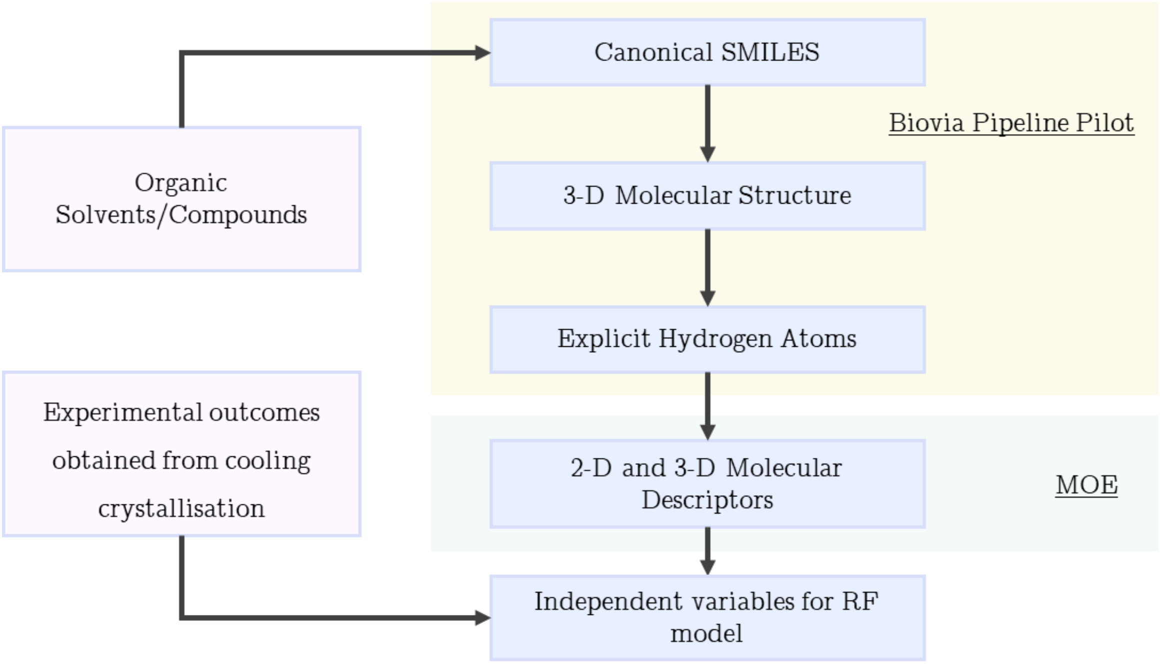

Canonical SMILES
Biovia Pipeline Pilot
Organic Solvents/Compounds
3-D Molecular Structure
Explicit Hydrogen Atoms
Experimental outcomes obtained from cooling crystallisation
2-D and 3-D Molecular Descriptors
MOE
Independent variables for RF model

Figure 3.4 Schematic workflow designed to calculate the 2D and 3D molecular descriptors of the organic solvents/compounds. The individually coloured brown and blue boxes indicate the software used for the specific parameters.

3.3.3 Molecular descriptors

Molecular fingerprints are a string representation of molecular structures that directly encodes the structure onto a series of binary bits. Molecular fingerprints are commonly used in substructure searching as the fragments of the molecule are expressed in bit sequences 0 and 1 where 1 indicates the presence of the fragment while 0 indicates the absence of the fragment as shown in Figure 3.5 (Warr, 2011) (D.-S. Cao, Q. Xu, Q. Hu, & Y.-Z. Liang, 2013). There are various types of fingerprints that are categorised broadly into four groups: binary circular fingerprints, circular fingerprints considering counts, path-based and keyed fingerprints, and pharmacophore-based fingerprints (Yuan Wang, 2009). The key type fingerprint, MACCS fingerprint, were utilised in this thesis as structural descriptors to predict the non-aqueous solubility of the drug and drug-like compounds in a diverse range of solvents alongside numerical descriptors calculated using MOE for comparison purposes. MACCS key based fingerprints were developed by Molecular Design Limited (MDL) and are based on pattern matching of molecular structure to a predefined set of 166 fragments which have been set by domain experts (Wale, Watson, & Karypis, 2008) (W. L. Chen, 2006). The MACCS key fingerprints used in this thesis were calculated using MOE following similar steps as shown in Figure 3.5.

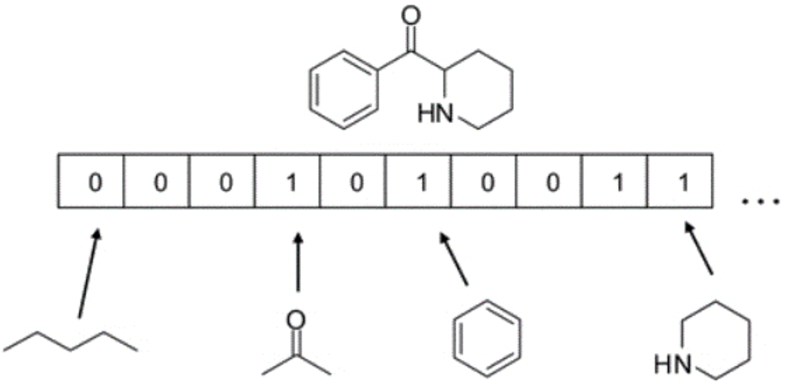


Figure 3.5 Schematic representation of MACCS key based fingerprints representation of the chemical structure. Each position in the vector indicated the presence (1) or absence (0) of fragments.

3.3.4 ML algorithms

Both regression and classification random forest models developed in this thesis were implemented using the 'randomForest' and 'caret' packages on the statistical computing environment R (version 3.3.1). PCA was constructed utilising the built-in package 'prcomp' in R and the commercially available multivariate analytical Excel add-on, XLSTAT ® (XLSTAT, 2017).

3.3.4.1 Data pre-processing

Machine learning models are dependent upon data as the algorithms learn from it. Thus, it is important to feed accurate data to develop an accurate model for prediction. Data cleaning or pre-processing is an important step in the development of machine learning models to obtain consistent and better predictive results. Pre-processing was performed on the dataset containing the 340 calculated molecular descriptors. The first stage was the identification of missing data in the dataset as it can have an impact on the accuracy. Missing data values below 5% of the sample can be omitted as the number is extremely small (Schafer, 1999). Similarly, Bennett (2001) proposed that statistical analysis is likely to be biased when 10% of the data are missing. For the research study, some of the molecular descriptors for the compounds were missing and the descriptors were omitted from the dataset. The descriptors were removed instead of the compounds in order to maintain a larger number of compounds for the model. Furthermore, random forest algorithms do not support null values. After missing data values were removed, descriptors with constant values across the dataset, *i.e.* zero variance, were removed. In this study, near-zero variance descriptors were included and the descriptors were not normalised since the random forest algorithm can handle both these kinds of descriptors. Duplicate samples in the dataset were also removed.

3.3.4.2 Cross validation

Cross-validation is an important statistical technique in evaluating the predictive capability of a model by partitioning the original dataset into a training set (to train the model) and test set (to evaluate it) (Wolpert, 1992). It is also utilised to compare various machine learning models and select the best model for a given problem. Cross-validation is easy to implement, and the cross-validated results generally have lower bias compared to other techniques (Y. L. Zhang & Yang, 2015). A few methods available for performing cross-validations are the validation set approach, leave-one-out-cross-validation (LOOCV), k-fold cross-validation, stratified k-fold cross-validation and adversarial validation. In this thesis, k-fold cross-validation was carried out on the training set where the value of k, which refers to the number of groups that the given data sample is split into, is set at 10. Choosing a lower value of k is more biased while a higher value of k suffers from large variability in overall prediction accuracy (James, Witten, Hastie, & Tibshirani, 2013). A common standard value for k is 10, with the process also known as 10-fold cross-validation. To perform 10-fold cross-validation, as mentioned earlier, the dataset was randomly split into 10 folds. For each fold in the dataset, the RF model was built on the remaining 9 folds of the dataset and tested against it. The method was repeated until each fold served as the test set once. The average error of the 10 models is the cross-validated error and is utilised as a performance metric for the RF model. Visualisation of 10-fold cross validation is shown in Figure 3.6.

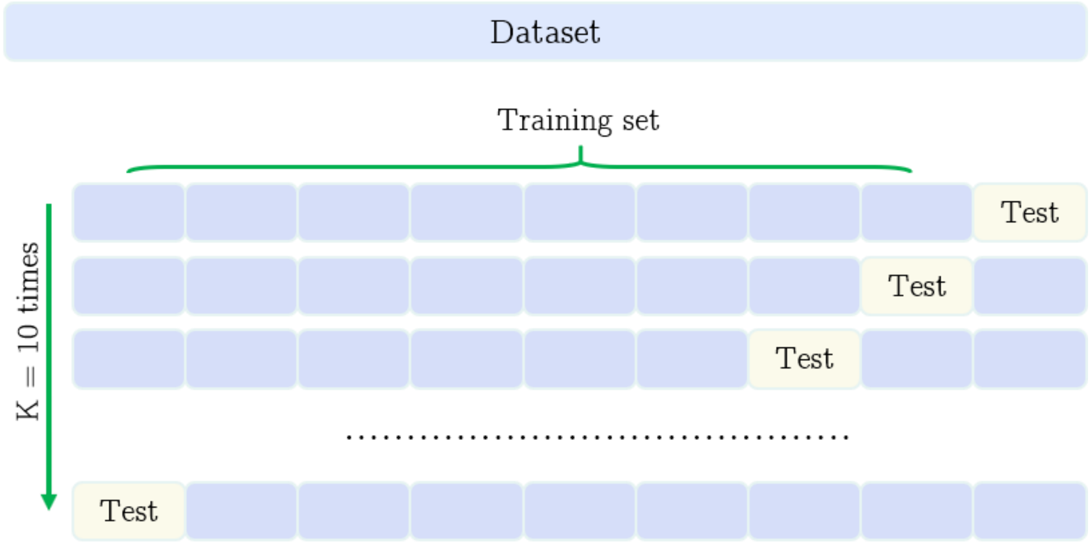


Figure 3.6 Schematic diagram of 10-fold cross-validation. The dataset is divided into 10-folds where one-fold is designated as test set while 9 folds are designated as the training set. The average accuracy of the ten test sets is the final cross-validated accuracy

3.3.4.3 RF model parameters

The value of the training parameters *mtry* and *nodesize* were kept at default values. Reducing the value of *mtry* to a small number decreased the predictive accuracy of the RF model while increasing the value of *mtry* above the default value made no change in the predictive accuracy of the model. Increasing the value of *ntree* increased the workload on the computational process. However, it made no significant change in the predictive accuracy of the RF models on the three compounds. Thus, an optimum value of *ntree* was chosen at 1,500 for the majority of RF models generated in this project, unless stated otherwise.

Chapter 4. Application of ML algorithms for predicting crystallisation outcomes and crystal habit of API in a diverse range of organic solvents

4.1 Introduction

Crystallisation is a necessary separation and purification method extensively used in the pharmaceutical industry (Jie Chen, Sarma, Evans, & Myerson, 2011). However, designing crystallisation processes is complex and challenging due to the determining molecular processes that make crystalline compounds exhibit different crystallisation behaviour. The crystallisation process is governed by the interaction between the different processes (e.g. nucleation, crystal growth) and process conditions (e.g. solution properties, and supersaturation) which determine the product quality defined by the polymorphic form, crystal size distribution, crystal shape, and purity. While extensive research has furthered the understanding of the relationship between the crystalline solid with the product quality attributes; it is still complex to predict accurately based on theoretical knowledge alone. Furthermore, with an increasing number of drug candidates and their complexity, the challenge to thoroughly understand the relationship between the crystalline solid with the required product attributes is ever more present. Limitations in predictability have been attributed to the limited level of understanding of critical processes in industrial crystallisation, e.g. nucleation (ter Horst, Schmidt, & Ulrich, 2015).

Applying systematic design strategies endeavours to develop sustainable crystallisation processes that help reduce experimental time and cost as well as limit waste generation, energy and usage of raw materials. One such method to achieve this would be through the implementation of ML algorithms (Figure 4.1). Machine learning has gained an increasing amount of interest in the field of crystallisation which includes areas such as investigating crystallisability of organic molecules, predicting various physicochemical properties (Delaney, 2005; McDonagh, van Mourik, & Mitchell, 2016; Palmer et al., 2007), protein crystallisation (Rupp & Wang, 2004; L. Y. Wei & Zou, 2016), predicting co-crystal and polymorph formation respectively (Johnston et al., 2008; Musil et al., 2018; Wicker et al., 2017). The reported advantages gained by the application of ML were an overall reduction in experimental time, cost of raw materials and process optimisation. ML has also played an important role in identifying and understanding the relationship between the features and the investigated property (Gasteiger & Engel, 2006).

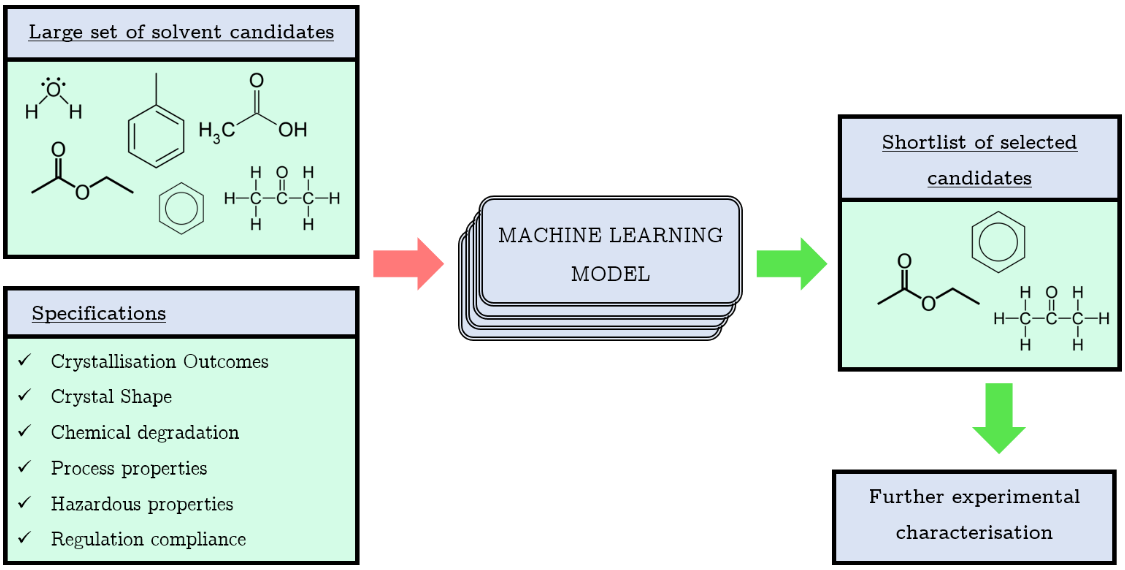


Figure 4.1 Schematic workflow of the implementation of the machine learning model in recommending suitable solvents for the crystallisation process.

The work described in this chapter aimed to gather experimental data from cooling crystallisation of an API in a diverse set of solvents which would then form the basis for the development and application of a machine learning algorithm capable of predicting various crystallisation outcomes (crystallisability, chemical degradation and crystal habit). This work has been further complemented by investigating the relationship between the calculated physicochemical molecular descriptors of the solvent with its influence on the crystallisation outcomes selected for this study.

4.2 Methodology

4.2.1 Experimental methods

4.2.1.1 Controlled cooling crystallisation

Following the methodology outined in Section 3.2.1.1, the influence of solvent on the crystallisation outcomes and shape of paracetamol was investigated on 94 solvents.The choice of organic solvents selected for this study was based on chemical diversity and availability in the laboratory and was not limited to pharmaceutically acceptable solvents. A solution containing 5 wt./wt% concentration of paracetamol was prepared by weighing a respective mass of paracetamol on a fixed volume of solvent set at 3ml. Precise, solid loading and solvent dispensing were performed using the automated Zinsser Platform. Cooling crystallisation was performed using *Technobis* Crystalline system (Figure 3.2). The solution containing paracetamol was stirred at 1000 rpm for an hour at 20^{0}C. The solution was then gradually increased to 10^{0}C below the boiling point of the solvent and left to agitate for an hour before being cooled down at a slow cooling rate of 0.1^{0}C/min. The cooled solution was then stored for a few days to allow maximum recovery of crystals. If the solution when heated remained in suspension, the concentration of paracetamol in the solution was reduced to half by doubling the volume of the respective solvent and the process was repeated. Similarly, if paracetamol was found to readily dissolve in the solvent at 20^{0}C, the concentration of paracetamol was doubled, and the process was repeated. The experiment was repeated four times for reproducibility. Crystallisation behaviour of paracetamol was observed and categorised into four outcomes. If paracetamol crystallised out in the solvent on cooling, it was categorised as '*crystallised out*'. If paracetamol remained in solution even after cooled for a set period of time, it was categorised as being '*non nucleated*'. If it remained as a suspension even after increasing the temperature, it was categorised as being *'practically insoluble'* and if change in colouration (dark brown or light pink) was observed it was categorised as '*degradation*´ in the respective solvent. These categories of crystallisation outcomes were then utilised as response for the supervised ML model.

4.2.1.2 Crystal shape analysis

For all cases of the procedure detailed in Section 4.2.1.1 where paracetamol successfully dissolved and crystallised out, the crystals were filtered and dried in a vacuum oven for 12 hours before analysis. The dried crystals were then placed under XRPD to confirm any change in polymorphic form. Crystal shapes were observed using the Leica DM6000M optical microscopy, and the mean aspect ratio of the dried crystals was measured using the Morphologi Software provided in the automated particle image analyser Malvern Morphologi G3. Crystal shapes were qualitatively divided and categorised according to the combined observations made using both the optical microscope and the measured aspect ratio (AR).

4.2.1.3 ML workflow

This study aimed to implement machine learning to predict the crystallisation outcome as well as the crystal habit of paracetamol in a diverse range of solvents. This was achieved by constructing a systematic model pipeline which involved various stages. The first stage involved the calculation of the molecular descriptors of the 94 solvents in which paracetamol was crystallised. 3D molecular structures of the solvents were built using the Biovia Pipeline Pilot 2017 software and were based on their canonical SMILES obtained from the ChemSpider database. Energy minimisation on these structures was then performed using the Clean force-field in Pipeline pilot 2017 (Hahn, 1995). The 3D molecular structures in SDF format were imported into Molecular Operating Environment. MOE was used to calculate 340 physicochemical descriptors of which 192 descriptors were 2-D, and 148 descriptors were 3-D descriptors. The list of the 2-D and 3-D calculated molecular descriptors obtained from MOE are categorised and listed in Section 3.3, as well as in the literature (P. Labute, 2000). These molecular descriptors, combined with the experimentally obtained controlled cooling crystallisation outcomes, were used to build the datasets for this project. In total, two datasets referred to as *Dataset A* and *Dataset B* were generated. *Dataset A* consisted of the 94 various solvents along with their molecular descriptors and their cooling crystallisation outcomes. The four experimental outcomes were labelled as *crystallised out*, *non-nucleated*, *practically insoluble* and *degradation. Dataset B* was a subset of *Dataset A*, including only the solvents where paracetamol was found to have crystallised out. The crystal shapes observed were qualitatively categorised as *Shape A*, and *Shape B*. The categorisation for dataset B was based on both the measured aspect ratio and the visual observation under the optical microscope. The workflow showing the generation of *Dataset A* and *Dataset B* is shown in Figure 4.2. The built dataset was then imported into the statistical computing environment R (Version 3.3.1) (R Development Core Team, 2013). The second stage involved the implementation of the machine learning classification models aimed at understanding the relationship between the physicochemical descriptors and the crystallisation outcomes and crystal habit of paracetamol. Both *Dataset A* and *Dataset B* were subjected to the machine learning workflow as outlined in Figure 4.3.

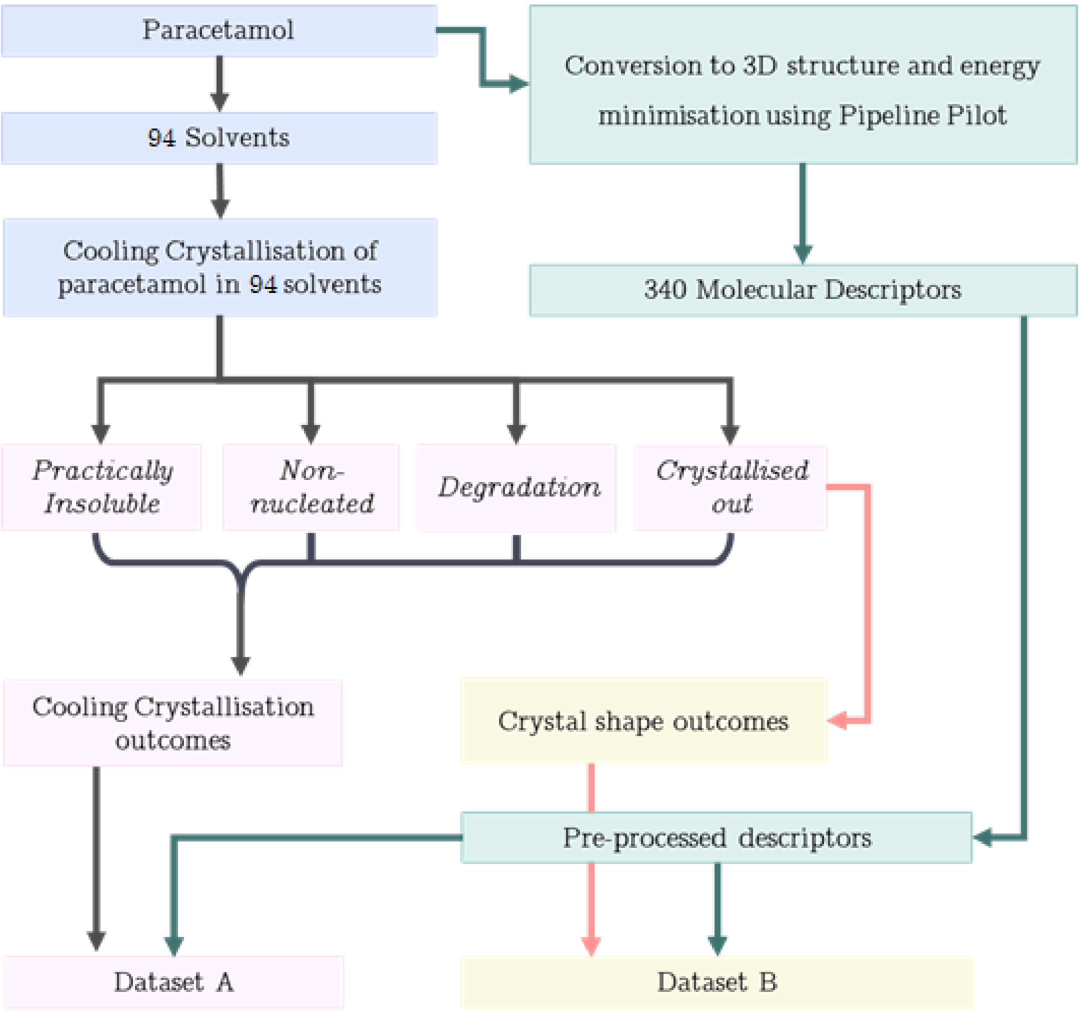

Paracetamol
94 Solvents
Cooling Crystallisation of paracetamol in 94 solvents
Conversion to 3D structure and energy minimisation using Pipeline Pilot
340 Molecular Descriptors
Practically Insoluble
Non-nucleated
Degradation
Crystallised out
Cooling Crystallisation outcomes
Crystal shape outcomes
Pre-processed descriptors
Dataset A
Dataset B

Figure 4.2 Workflow presenting the implementation of machine learning models to understand the relationship between the molecular descriptors and the crystallisation outcomes and crystal habit.

Pre-processing of the dataset was performed using the "caret" package (M. Kuhn, 2008) on the statistical computing environment R (version 3.3.1) where descriptors containing zero variance, missing values and highly correlated descriptors (>0.95) were removed. This reduced the total number of descriptors from 340 to 270 descriptors. Both unsupervised (PCA) and supervised (random forest) ML algorithms were applied independently to both the datasets. PCA was constructed utilising the commercially available multivariate analytical Excel add-on, XLSTAT (XLSTAT, 2017). PCA was performed to reduce the dimensionality of the dataset while potentially identifying the patterns or trends relating to the influence of the solvent structure on the crystallisation outcome and crystal habit of paracetamol in the respective solvents. PCA operates by reducing the number of descriptors into a set of fewer, new descriptors with minimal loss of information called principal components (J. Lever, Krzywinski, & Atman, 2017). The number of principal components to be considered is determined by the eigenvalues which measure the amount of variation retained by each principal component. Using the "factoextra" package (Le, Josse, & Husson, 2008) in R, the total contribution of each descriptor accounting for the variability of the selected principal components was determined. A biplot of the molecular descriptors and the solvents was constructed at the end of the analysis which allowed investigation of the direction of the descriptors with the position of the solvent in the factor map. Examples of these biplots can be seen in Figure 4.15 and Figure 4.16 respectively.

Similarly, the "randomForest" package (Andy Liaw & Matthew Wiener, 2002) in R, based on the original FORTAN code of Brieman and Cutler (Breiman, 2001b), was used to build the random forest classification models. RF classification models are robust to datasets containing descriptors larger than the sample size, perform internal cross-validation (i.e. using Out-Of-Bag samples) and their parameterisation only consists of few internal tuning parameters. The dataset used for the supervised classification model was randomly split in the ratio of 80:20 where 80% of the dataset was used as a training set and 20% of the dataset was used as the test set. To create an unbiased supervised classification model, 10-fold cross-validation (CV) was performed on the training set using the 'caret' function in R (M. Kuhn, 2008). RF classification models for both *Dataset A* and *Dataset B* were constructed using the following parameters: *ntree*=1500 trees and *mtry* = square root of the total number of descriptors. RF models have been proven in various literature to be quite insensitive to alteration in its internal parameters. (Cannon, Bender, Palmer, & Mitchell, 2006; L. D. Hughes, Palmer, Nigsch, & Mitchell, 2008; Palmer et al., 2007; V. Svetnik et al., 2003). The trained RF model after cross-validation was used to predict the outcomes of an external test set. The schematic workflow of the whole process from data pre-processing to the implementation of ML is presented in Figure 4.3. Both PCA and RF were discussed in detail in Chapter 1.

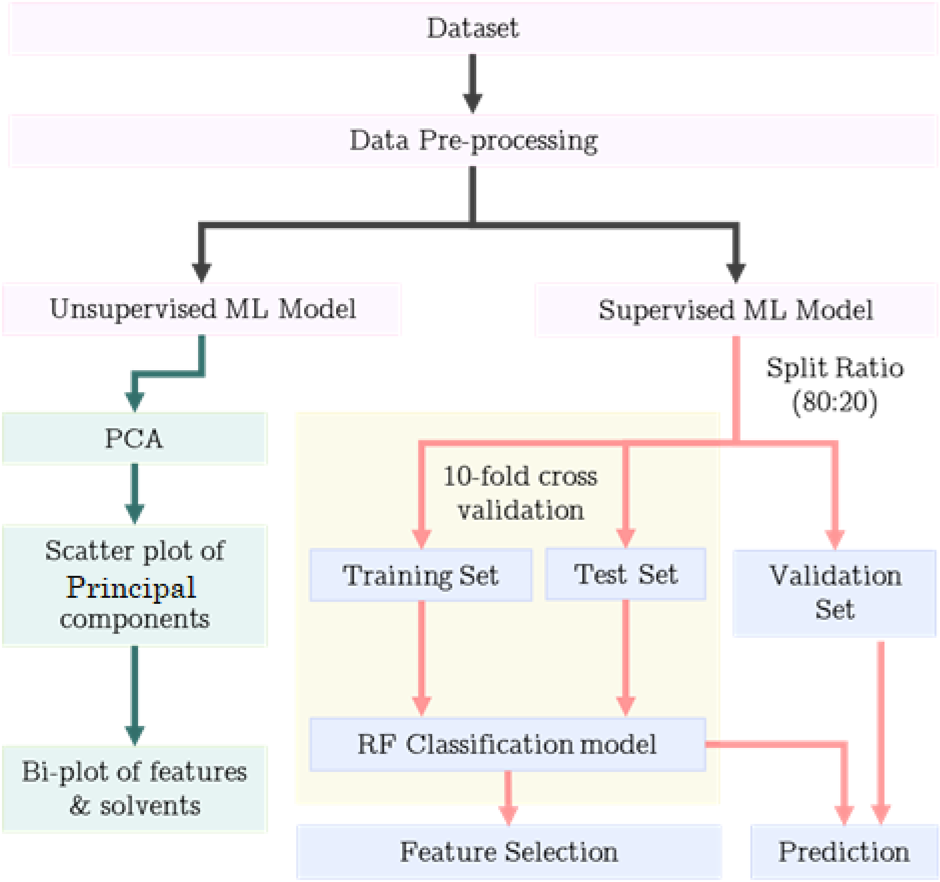


Figure 4.3 Workflow presenting the implementation of machine learning models to understand the relationship between the molecular descriptors and the crystallisation outcomes and crystal habit.

4.3 Results and discussions

4.3.1 Controlled cooling experiments

Solubility measured qualitatively from controlled cooling crystallisation experiments indicated that in a diverse range of 94 solvents, paracetamol was found to dissolve in 58 solvents, of which most were polar, and was found to be insoluble in 36 solvents, of which most were non-polar, as seen in Table 4.2. Solubility studies of paracetamol were performed in 26 solvents by Gransberg and Rasmuson (1999) and in 15 solvents by Lee et al (2006) where paracetamol was reported to be soluble in polar solvents, insoluble in non-polar and chlorinated halogenated solvents and highly soluble in dimethylsulphoxide, diethylamine and N,N dimethylformamide (Granberg & Rasmuson, 1999; Tu, Chung Shin, & Ying Hsiu, 2006). The relative solubility data obtained from the cooling crystallisation experiments were found to complement the previously referenced literature data.

Out of the 58 solvents in which it dissolved, paracetamol was found to crystallise from 44 solvents. According to the conditions set in Section 3.2.1.1 and Section 4.2.1.1 of this chapter, paracetamol failed to crystallise out from 14 solvents and remained in clear solution as shown in Table 4.2. This could be because paracetamol was highly soluble in the respective solvent and so required a lower cooling temperature than 20^{0}C, or because a longer induction time was needed to drive crystallisation. The solvents that paracetamol failed to crystallise in were also highly viscous. It has been reported in the literature that the supersaturation interval increases with increasing viscosity of the solution as it severely reduces the mobility of drug molecules in solution. Hence more time is required to align into a critical nucleus (Storm, Hazleton, & Lahti, 1970). It is to be noted that the purpose of the cooling crystallisations was to observe the appearance of crystals in the respective solvent rather than to analyse the time of the crystal's appearance. Among these 14 solvents, paracetamol was found to show degradation in the form of discolouration in 3 solvents: ethyl lactate, trifluoroacetic acid and 2-amino-1-butanol. A solution of paracetamol in ethyl lactate turned light pink while the solution of trifluoroacetic acid and 2-amino-1-butanol turned into a dark brown solution. Table 4.1 presents the list of solvents showing the various crystallisation outcomes. Overall, these four crystallisation outcomes on the 94 solvents are labelled as *Dataset A* and summarised in Table 4.2.

Table 4.1 Shown is the list of solvents categorised according to their family class and their respective crystallisation outcomes with paracetamol. The solvents are labelled numerically from 1 to 94.

Crystallisation Outcome	Solvent Family	List of solvents
Crystallised Out	Alcohols	1) methanol 2) ethanol 3) 1-propanol 4) 2-propanol 5) 1-butanol 6) 2-butanol 7) 2-pentanol 8) 1-octanol 9) 1-decanol 10) 2-phenylethanol 11) benzyl alcohol 12) 2,2,2 trifluoroethanol 13) isoamyl alcohol 14) 2-methyl-1-propananol 15) cyclohexanol
	Acids	16) acetic acid 17) formic acid
	Ester	18) ethyl acetate 19) butyl acetate 20) isobutyl acetate 21) diethyl carbonate 22) formamide
	Ether	23) anisole 24) 1,4-dioxane 25) 2-phenoxyethanol 26) 2-butoxyethanol 27) 2-methoxyethanol 28) methoxyethane
	Ketone	29) acetone 30) 2-butanone 31) 3-pentanone 32) 4-methyl-2-pentanone
	Polar Aprotic	33) acetonitrile 34) n, n-dimethylacetamide 35) n, n-dimethylformamide

		36) tetrahydrofuran
	Halogenated	37) 1-bromo-2-chloroethane 38) diiodomethane
	Nitro	39) nitrobenzene 40) nitromethane
	Organosulfur	41) tetrahydrothiophene
	Aromatics	42) aniline
	Diols	43) 1,2-propanediol
	Water	44) water
Non-Nucleated	Ether	45) 2-ethoxyethanol 46) 1,2-dimethoxyethane 47) 2-methoxyethyl ether
	Alcohols	48) 3-pentanol 49) 1-nonanol
	Amines	50) phenethylamine 51) pyridine
	Polar Aprotic	52) dimethyl sulfoxide 53) N-methyl-2-pyrrolidone
	Ester	54) pentyl acetate
	Diols	55) 1,4-butanediol
Practically Insoluble	Halogenated	56) 1-bromobutane 57) 2-bromobutane 58) bromoform 59) bromobenzene 60) chloroform 61) 1-chlorobutane 62) chlorobenzene 63) 1,2-dichloroethane 64) dichloromethane 65) iodomethane 66) carbon tetrachloride 67) tetrachloroethene 68) trichloroethylene
	Hydrocarbons	69) 2,2,4 trimethylpentane (i-octane) 70) 2-methylbutane 71) hexane 72) iso-hexane 73) methyl cyclohexane 74) n-dodecane 75) cyclohexane 76) heptane
	Aromatics	77) cumene 78) 1-methylnaphthalene

		79) benzene 80) toluene 81) m-xylene
	Ether	82) 2-methoxy-2-methylpropane 83) cyclopentane 84) diisopropyl ether 85) dibutyl ether 86) diethyl ether
	Thiols	87) ethanethiol 88) 1-propanethiol
	Organosulfur	89) 3-methylthiophene 90) sulfolane
	Amines	91) triethylamine
Degradation	Acids	92) trifluoroacetic acid
	Amines	93) 2-amino-1-butanol
	Ester	94) ethyl lactate

Table 4.2 Summary of the four crystallisation outcomes of paracetamol observed in 94 solvents

Dataset A	
Crystallisation Outcomes	**Number of solvents**
Crystallised out	44
Non-nucleated	11
Practically Insoluble	36
Degradation	3

The mean aspect ratios measured for the crystallised-out crystals in the 44 solvents are presented in Table 4.3. The predicted morphology of monoclinic paracetamol (Form I) based on the Bravais-Friedel-Donnay-Harker (BFDH) model was calculated in Mercury (Version 3.9) from its crystal structure and is shown in Figure 4.4. From the cooling crystallisation experiments, paracetamol was observed to form mainly three crystal shapes: truncated cube or blocks (e.g. methoxyethane, 2-butoxyethanol, diethyl carbonate, etc.), prismatic and parallelepiped (e.g. methanol, benzyl alcohol, nitrobenzene, etc.) and lath or rod shapes (e.g. formic acid, n, n- dimethylacetamide, 3-pentanone etc.). For the purpose of this study truncated cube and prismatic shapes were grouped under one category as aspect ratio is unable to differentiate between these two shapes. The optical micrograph of varying crystal shapes of paracetamol produced from 44 solvents in order of their decreasing mean aspect ratios is illustrated in Figure 4.5.

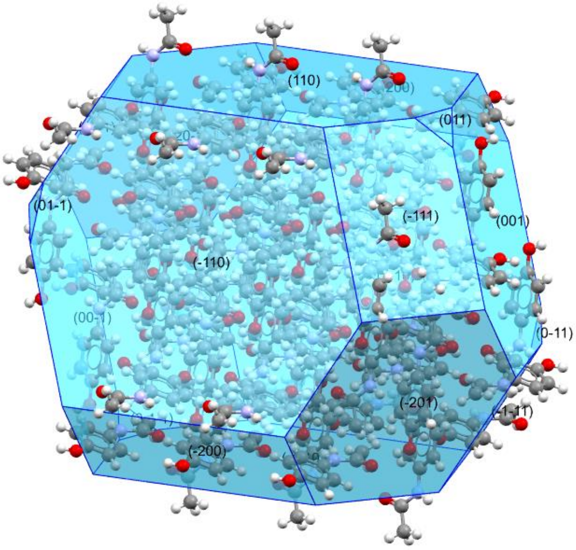


Figure 4.4 BFDH morphology prediction of monoclinic paracetamol.

Table 4.3 Measured aspect ratio of paracetamol crystallised in 44 organic solvents by cooling crystallisation.

Label	Solvent	Mean AR	Label	Solvent	Mean AR
1	2-phenylethanol	0.861	23	1-butanol	0.651
2	methoxyethane	0.854	24	2-phenoxyethanol	0.648
3	tetrahydrofuran	0.850	25	nitromethane	0.632
4	2-pentanol	0.848	26	nitrobenzene	0.632
5	water	0.827	27	acetic acid	0.631
6	benzyl alcohol	0.815	28	ethanol	0.628
7	acetone	0.801	29	isoamyl alcohol	0.622
8	2-butanol	0.800	30	1-octanol	0.618
9	1,4-dioxane	0.782	31	cyclohexanol	0.617
10	diethyl carbonate	0.775	32	1-decanol	0.595
11	anisole	0.762	33	1-propanol	0.593
12	2-butoxyethanol	0.755	34	4-methyl-2-pentanone	0.587
13	2-ethoxyethanol	0.752	35	2-methyl-1-propanol	0.586
14	aniline	0.747	36	n, n-dimethylacetamide	0.577
15	1,2-propanediol	0.706	37	1-bromo-2-chloroethane	0.498
16	2-propanol	0.703	38	formamide	0.478
17	butyl acetate	0.698	39	formic Acid	0.444
18	acetonitrile	0.681	40	n, n-dimethylformamide	0.437
19	methanol	0.673	41	diiodomethane	0.430
20	ethyl acetate	0.671	42	isobutyl acetate	0.421
21	2-butanone	0.665	43	3-pentanone	0.410
22	2,2,2-trifluoroethanol	0.659	44	tetrahydrothiophene	0.381

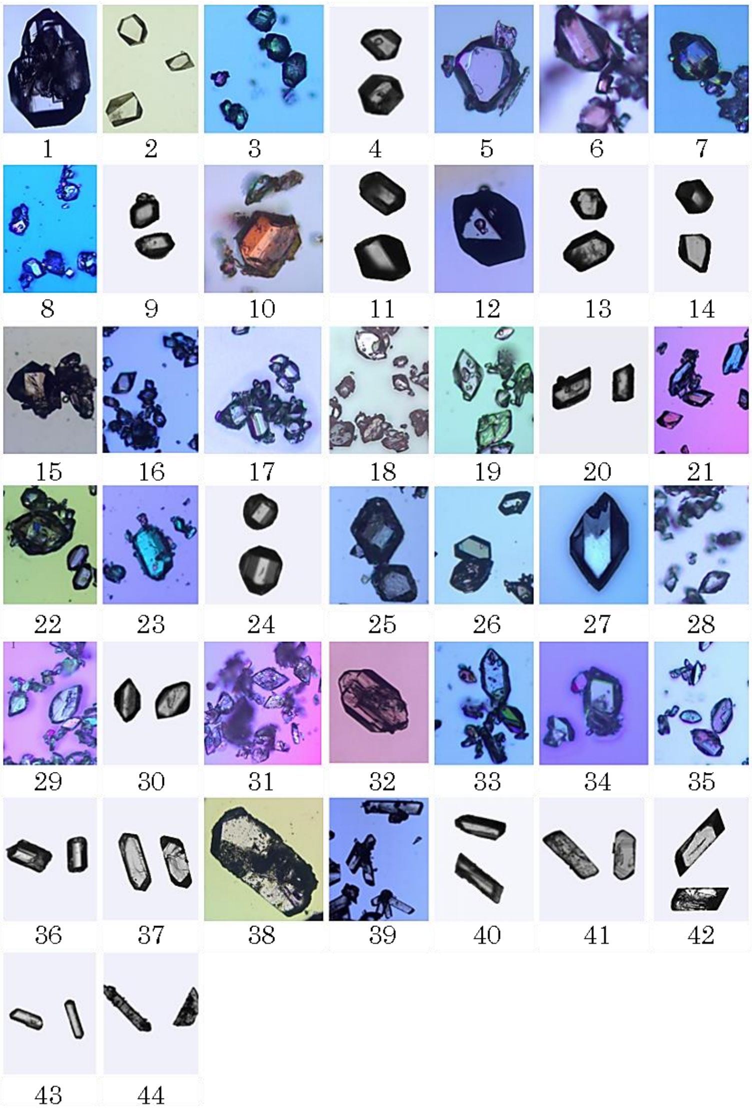


Figure 4.5 Optical micrographs of the crystal shape of paracetamol grown in 44 solvents by controlled cooling crystallisation arranged in the order of their decreasing mean aspect ratios: 1) 2-phenylethanol 2) methoxyethane 3) tetrahydrofuran 4) 2-pentanol 5) water 6) benzyl alcohol 7) acetone 8) 2-butanol 9) 1,4-dioxane 10) diethyl carbonate 11) anisole 12) 2-butoxyethanol 13) 2-methoxyethanol 14) aniline 15) 1,2-propanediol 16) 2-propanol 17) butyl acetate 18) acetonitrile 19) methanol 20) ethyl acetate 21) 2-butanone 22) 2,2,2-trifluoroethanol 23) 1-butanol 24) 2-phenoxyethanol 25) nitromethane 26) nitrobenzene 27) acetic Acid 28) ethanol 29) isoamyl alcohol 30) 1-octanol 31) cyclohexanol 32) 1-decanol 33) 1-propanol 34) 4-methyl-2-pentanone 35) 2-methyl-1-propanol 36) n, n-dimethylacetamide 37) 1-bromo-2-chloroethane 38) formamide 39) formic Acid 40) n, n-dimethylformamide 41) diiodomethane 42) isobutyl acetate 43) 3-pentanone 44) tetrahydrothiophene

It is generally assumed that for cooling crystallisation, the influence of the solvent effect on the crystal habit is mainly via the preferential adsorption of solvent molecules on the specific crystal faces and that removal of solvent molecules before the deposition of oncoming solute molecules causes retardation of crystal growth (Wells, 1946). Using the relative polarities of the various crystal faces obtained from electrostatic potential maps calculated at closest approach distances, the extent of solvation of a crystal face can be qualitatively understood (Berkovitch-Yellin, 1985).

As the purpose of this project was to implement ML models to understand the relationship between the solvents' calculated physicochemical properties and their respective crystal shapes, final crystal habit at the end of the cooling experiments was considered. The technique allowed a faster preliminary approach to crystal habit screening on a broader set of solvents. The crystal habits observed were generalised into two main distinctive shape categories, i.e. *Shape A* and *Shape* B. Crystal habits observed as being truncated cubes and prismatic were all grouped as *Shape A* while crystal shapes observed as either lath or rod were categorised as *shape B*. These two categories were made based on the visualisation of the crystals when observed through the Leica optical microscope and on their basis of the measured mean aspect ratio. As the mean aspect ratios for *Shape A* ranged from 0.861 to 0.586 while the mean aspect ratios for *Shape B* ranged from 0.577 to 0.381. Even though the crystal shapes were visually different (Figure 4.5), their mean aspect ratios were found to be in close range such as in the case of 2-methyl-1-propanol and n n-dimethylacetamide. This is because mean aspect ratios for each solvent were measured over several particles and not just a single crystal. It is possible that there was the presence of two or more shape variants which could have affected the overall measurement of mean aspect ratio. Overall, out of the 44 solvents that paracetamol was found to crystallise in, 35 of the crystal habits were categorised as being *Shape A,* and the remaining 9 crystals were categorised as being *shape B*. This dataset was labelled as *Dataset B* and summarised in Table 4.4.

Table 4.4 Summary of the two crystal shape outcomes of paracetamol observed in 44 solvents.

Dataset B	
Shapes Observed	Number of Solvents
Shape A	35
Shape B	9

4.3.2 ML Algorithm

4.3.2.1 Unsupervised ML (PCA)

PCA was performed on both *Dataset A* and *Dataset B* for understanding and explaining the total variation of the datasets. Their eigenvalues indicated the amount of variation retained by each principal component (PC). Eigenvalues are most extensive for the first PC and small for the subsequent components. Table 4.5 and Table 4.6 display the eigenvalues, percent of variance and cumulative percent of variance from the observed *Dataset A* and *Dataset B* respectively.

Table 4.5 The first 5 principal components of the built PCA models for *Dataset A* containing 94 solvents displaying the eigenvalues, percent of variance and the cumulative percent of the variance.

Principal Components (PC)	Eigenvalue	Variance (%)	Cumulative %
PC 1	73.461	26.425	26.425
PC 2	58.003	20.864	47.289
PC 3	26.163	9.411	56.701
PC 4	18.082	6.504	63.205
PC 5	12.698	4.568	67.772

Table 4.6 The first 5 principal components of the built PCA models for *Dataset B* displaying the eigenvalues, percent of variance and the cumulative percent of the variance

Principal Components (PC)	Eigenvalue	Variance (%)	Cumulative %
PC 1	85.835	31.326	31.326
PC 2	40.418	14.751	46.078
PC 3	27.610	10.077	56.154
PC 4	21.549	7.865	64.019
PC 5	15.455	5.641	69.659

A common way to decide on the number of PCs to retain is by constructing a scree plot which plots each component's percentage of explained variance against the associated component and identifies breaks or gaps between PCs of a large and smaller percentage of explained variance as shown in Figure 4.6 and Figure 4.7. Components that appear before the gap are retained as being meaningful, and those after the gap are not retained (Cattell, 1983).

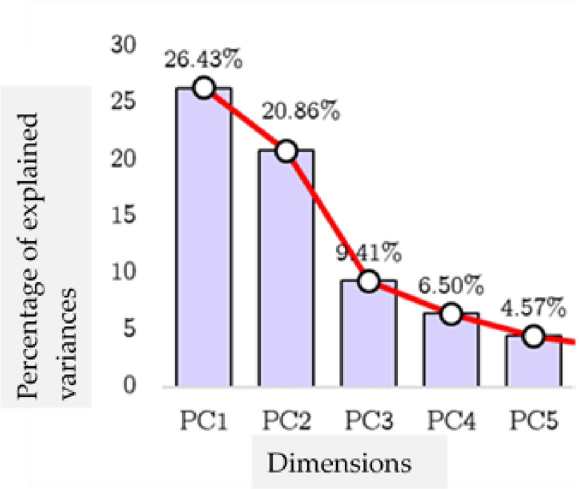


Figure 4.6 Scree plot of the extracted PC for *Dataset A* showing the fraction of total variance represented by each principal component.

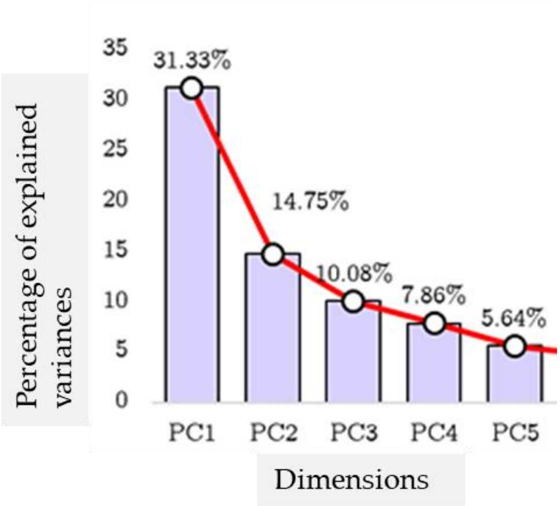


Figure 4.7 Scree plot of the extracted PC for *Dataset B* showing the fraction of total variance represented by each principal component.

The first two PCs for *Dataset A* account for 47.29% of the variation in the input. There is a significant decrease between the second and the third PCs, so the scree plot would lead us to retain only the first two components. Similarly, for *Dataset B* the first two PCs accounted for 45.73% of the variation, albeit with a much reduced fall off from PC2 to subsequent PCs.

4.3.2.1.1 Score plots for *Dataset A*

The scores of the first two PCs were plotted against each other for *Dataset A* containing 94 solvents as shown in Figure 4.8. The experimental outcomes coloured each point. Solvents, where paracetamol was found to have not dissolved, crystallised out, failed to nucleate or degraded, were outlined by the red, blue, yellow and green circles respectively.

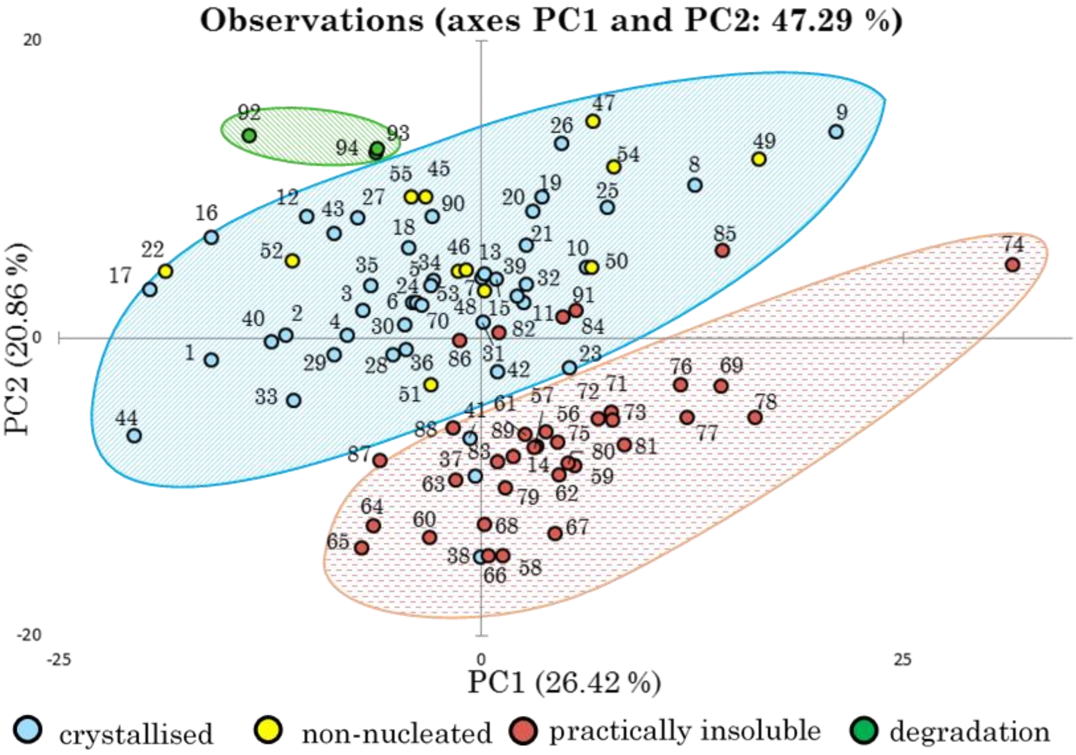


Figure 4.8 Scatter Plot based on the score values of 94 solvents in *Dataset A* projected to the first two PCs displaying the four-crystallisation outcome. Three significant clusters that were observed are outlined by the green, blue and red regions for visualisation. The solvents in the chemical space are labelled as 1 to 94 accordingly as numbered in Table 4.1.

Examining the graphical distribution of the solvents on the 2D score plot indicated three significant clusters of experimental outcomes: degradation, practically insoluble and a combined clustering between the crystallised out and non-nucleated. The solvents where paracetamol showed degradation were clustered in the top left quadrant; the solvents where paracetamol was found to have either crystallised or lacked nucleation were overlapped with each other and spread mainly among the left and right top quadrant and the solvents that paracetamol was found to be insoluble in were spread at the lower quadrant of the PCA model. Each cluster was outlined by green, blue and red regions as shown in Figure 4.8. Five of the practically insoluble labelled solvents were found to be spread in the blue clustering regions. These solvents were found to be mainly of the ether solvent family, i.e. 2-methoxy-2-methylpropane (82), diisopropyl ether (84), dibutyl ether (85), diethyl ether (86) as well as triethylamine (91).

Similarly, three solvents labelled as *crystallised out* were found to spread in the red clustering region. These solvents included 1-bromo-2-chloroethane (37), methoxyethane (38) and tetrahydrothiophene (41). Even though paracetmol crystallised out of these three solvents, its solubility these solvents was experimentally found to be very low. The lack of clustering observed between the *non-nucleated* and the *crystallised-out* solvents could be due to the possibility that besides the physicochemical descriptors of the solvents, there were other significant factors influencing nucleation of paracetamol in these solvents. Along with some of the significant factors discussed earlier in Section 4.3.1, supersaturation is one of the major driving forces behind nucleation and is not captured in the calculated physicochemical descriptors of each solvent (Myerson & Ginde, 2002).

A 3-D score plot was constructed for *Dataset A* by utilising the first three PCs obtained from Table 4.5 as adding the third PC explained 56.69% of variation compared to 47.30% of variation explained by only two PCs components. Adding a third PC also helped produce a better visualisation of the clustering than the scatter plot in Figure 4.8 as shown in Figure 4.9.

To investigate if the chosen API crystallises or not (or is degraded) in the respective solvent, the API must dissolve in the solvent. If the three crystallisation outcomes, i.e. *crystallised out*, *non-nucleated* and *degradation* were to be grouped as one outcome of *soluble*, then it can be observed from Figure 4.9 that there is a significant and near-perfect clustering between the *soluble* and *practically insoluble* outcomes. These are highlighted by the green region indicating the *soluble* and the red region indicating the *practically insoluble* regions. Thus, clustering indicates that the calculated molecular descriptors are significantly related to the relative solubility classification of paracetamol on these 94 solvents.

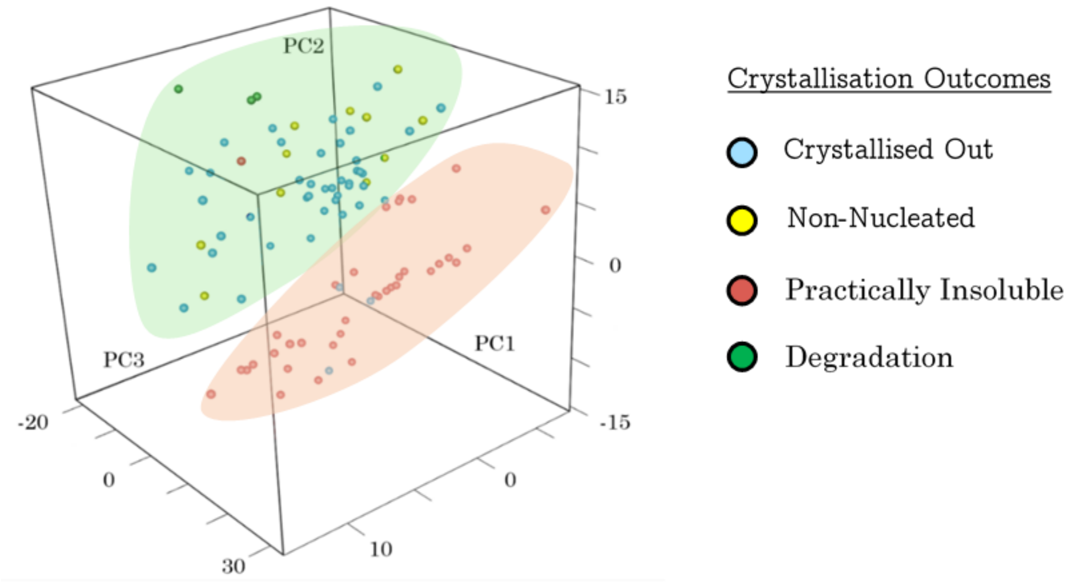


Figure 4.9 3-D Scatter Plot of the score values of 94 solvents projected to the first three PCs displaying the four crystallisation outcomes. The green region indicated the solvent clusters showing the *soluble* outcomes and the red region indicated the solvent clusters showing the *practically insoluble* outcomes.

94 solvents were categorised according to their respective functional groups (Table 4.1). Similar categories of solvents can be observed in various solvent selection guides presented in the literature. (Henderson et al., 2011; Prat et al., 2016). They were then plotted as projections on the first two PCS to visualise the diversity of the solvents in the solvent space as shown in Figure 4.10. It was observed from Figure 4.10 that solvent clusters of halogenated and thiols solvent family class were mainly observed at the lower quadrant of the scatter plot. Similarly, at the upper quadrant of the scatter plot, the clustering of diols, esters and acids was observed. Ether, polar aprotic and amine clusters occupied the middle section of the PCA scatter plot. Hydrocarbons, ketones and aromatics and primary alcohols (methanol to 1-dodecanol) formed a near-linear arrangement in the PCA score plot. Similar trends of solvent clusters observed in this chapter were reported in other studies (Diorazio, Hose, & Adlington, 2016; Johnston et al., 2017; Morten et al., 2008).

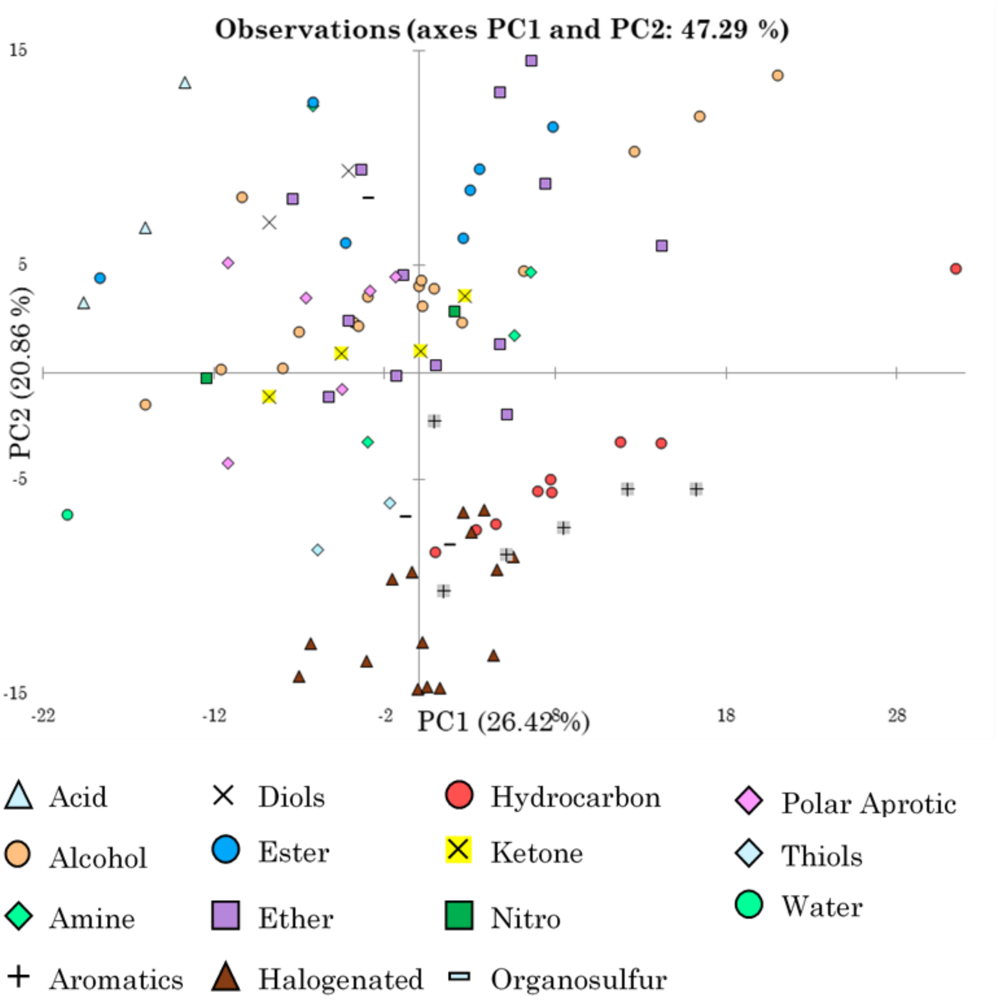


Figure 4.10 Projection of the scatter plot for 94 solvents based on the scores of the first two PCs. The solvents were labelled according to their respective solvent family class to identify trends within the PCA model.

A solvent's polarity can be estimated from its dielectric constant (ε). The higher the dielectric value, the more polar the solvent. Solvent polarity is related to the capacity of a solvent for solvating dissolved charged species. There is no unique method for quantitatively measuring polarity. Possible interaction between various solvents and solutes are too complex to be dictated by a single measurement (Alan R. Katritzky et al., 2004). Generally, solvents with dielectric constants greater than 5 are considered polar and those with dielectric constant less than 5 are regarded as being non-polar (Loudon, 2001). If the 94 solvents in this chapter were to be classified as *polar* and *non-polar* according to the mentioned criteria, the following trends were observed in the PCA scatter plot. From the PCA scatter plot, it can also be observed that solvents that are mainly polar are found on the upper half of the quadrant of the score plot, and the non-polar solvents mainly occupied the lower half of the score plot. Even though ethers are essentially non-polar, they are found to cluster mainly at the middle section of the score plot close to polar solvent family groups such as ketones, alcohols and diols. This is because even though they lack the strongly polarised O-H bond which makes hydrogen bonding possible, ethers can accept hydrogen bonds and are also able to solvate cations but not anions. (**add ref**) Furthermore, oxygen is more polar than carbon but not as polar as alcohols and thus there is a degree of polarity in ethers which explains why they are closer to polar groups (15). Paracetamol is a polar molecule due to the variation of electronegativity and the presence of lone electron pairs. Its molecular structure further lack lines of symmetry resulting in an unsymmetrical distribution of electrical charges. This prevents the charges within the structure to cancel each other out. This behaviour thus results in paracetamol to be a polar molecule. Thus, it tends to dissolve in polar solvents and is insoluble in non-polar solvents (Figure 4.11). This pattern of polarity is seen in the scatter map in Figure 4.8.

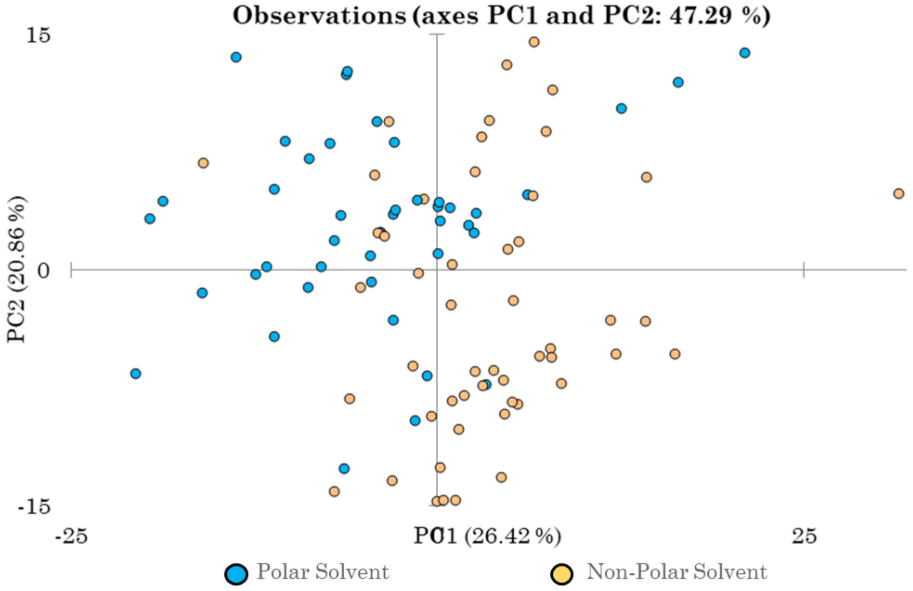


Figure 4.11 Score plot based on the score values of 94 solvents in *Dataset A* projected to the first two PCs and displaying the solvents based on their polarity.

4.3.2.1.2 Score plots for *Dataset B*

Similarly, the score plot for the PCA model of *Dataset B* for 44 solvents was plotted as shown in Figure 4.12

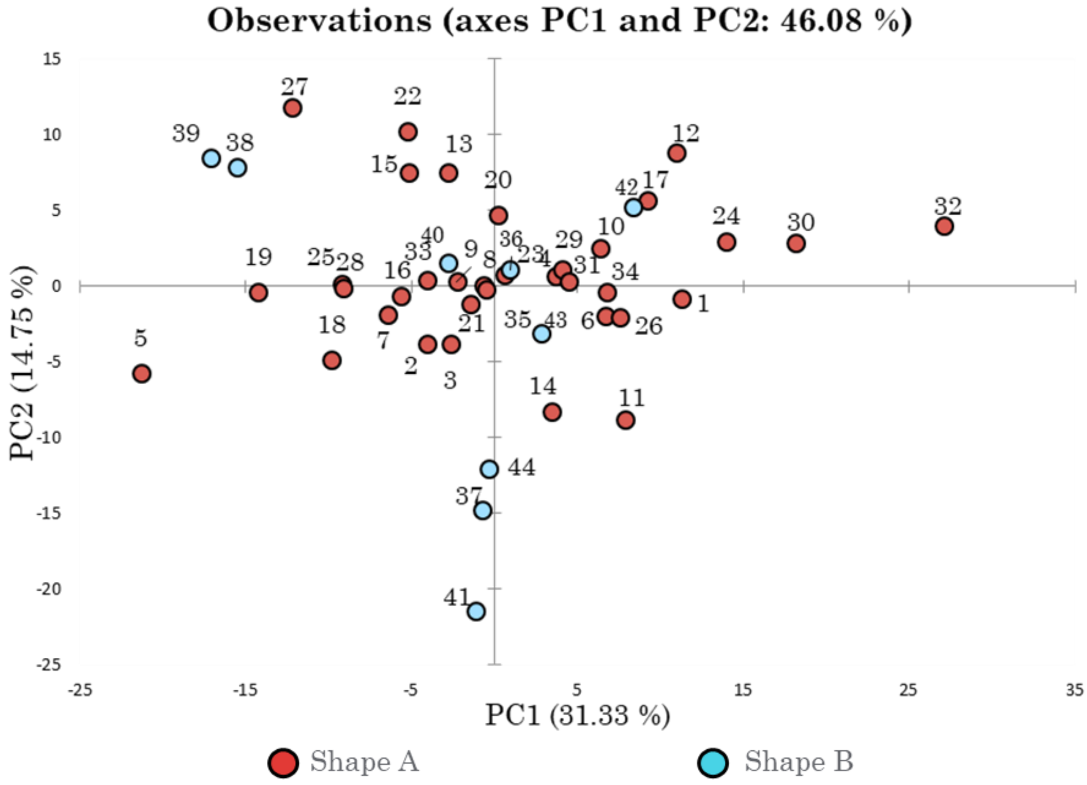


Figure 4.12 Scatter Plot based on the score values of 44 solvents in *Dataset B* projected to the first two PCs displaying the two crystal habit outcomes. The solvents in the chemical space are labelled as 1 to 44 accordingly (Table 4.3).

The score plot for the 44 solvents in *Dataset B* indicated the concentration of the classified *Shape A* crystals along the first PC. *Shape B* crystals were scattered around the chemical space. Unlike *Dataset A* (Figure 4.8), there was little evidence of an active cluster structure in Figure 4.12. The outliers observed on the far left of the PCA scatter plot labelled as (39) and (38) were formic acid and its amide derivative, formamide respectively. The solubility of paracetamol in the solvents that formed *Shape B* at the lower half of the scatter plot, i.e. (44) tetrahydrothiophene, (37) 1-bromo-2-chloroethane and (41) diiodomethane was experimentally found to be relatively poor. The solvents n, n-dimethylformamide, n, n-dimethylacetamide, isobutyl acetate and 3-pentanone labelled as (40), (36), (42) and (43) respectively also presented *Shape B* crystal habit for paracetamol. These solvents mainly clustered around the origin point of the score plot. The solubility of paracetamol in these solvents was experimentally found to be relatively higher compared to the other rod forming solvents scatter around the plot.

4.3.2.1.3 Variables factor loading map of the solvents in *Dataset A* and *Dataset B*

The descriptors' factor loading map was plotted between the first two PCs which depicts the interrelation among the calculated molecular descriptors as well as the influence of them on each principal component as shown in Figure 4.13 for *Dataset A* and Figure 4.14 for *Dataset B*. The correlation mono-plot not only represents the variance between the 270 descriptors but also indicates how much each variable contributes to each PC. The length of the vectors pointing away from the origin and its closeness to the correlation circle is proportional to its contribution to the PC while the angle between any two vectors is inversely proportional to the correlation between them. From Figure 4.13 and Figure 4.14, some of the most contributing descriptors to the PCs can be observed. A colour code indicates these with red being the most contributing variable and turquoise being the least. However, due to a large number of descriptors present (270 descriptors), it is challenging to visualise from the correlation mono-plot clearly and can only be distinguished by their correlation in the presented Figure 4.13 and Figure 4.14. To better visualise descriptors that contributed the most in explaining the variance in both principal components for *Dataset A* and *Dataset B*, a table containing the top 20 variables in order of their decreasing contribution factor was constructed as shown in Table 4.7 and Table 4.8 respectively. It is to be noted that these contributing descriptors do not relate to the experimental response labels and are attained purely based on the unsupervised PCA models.

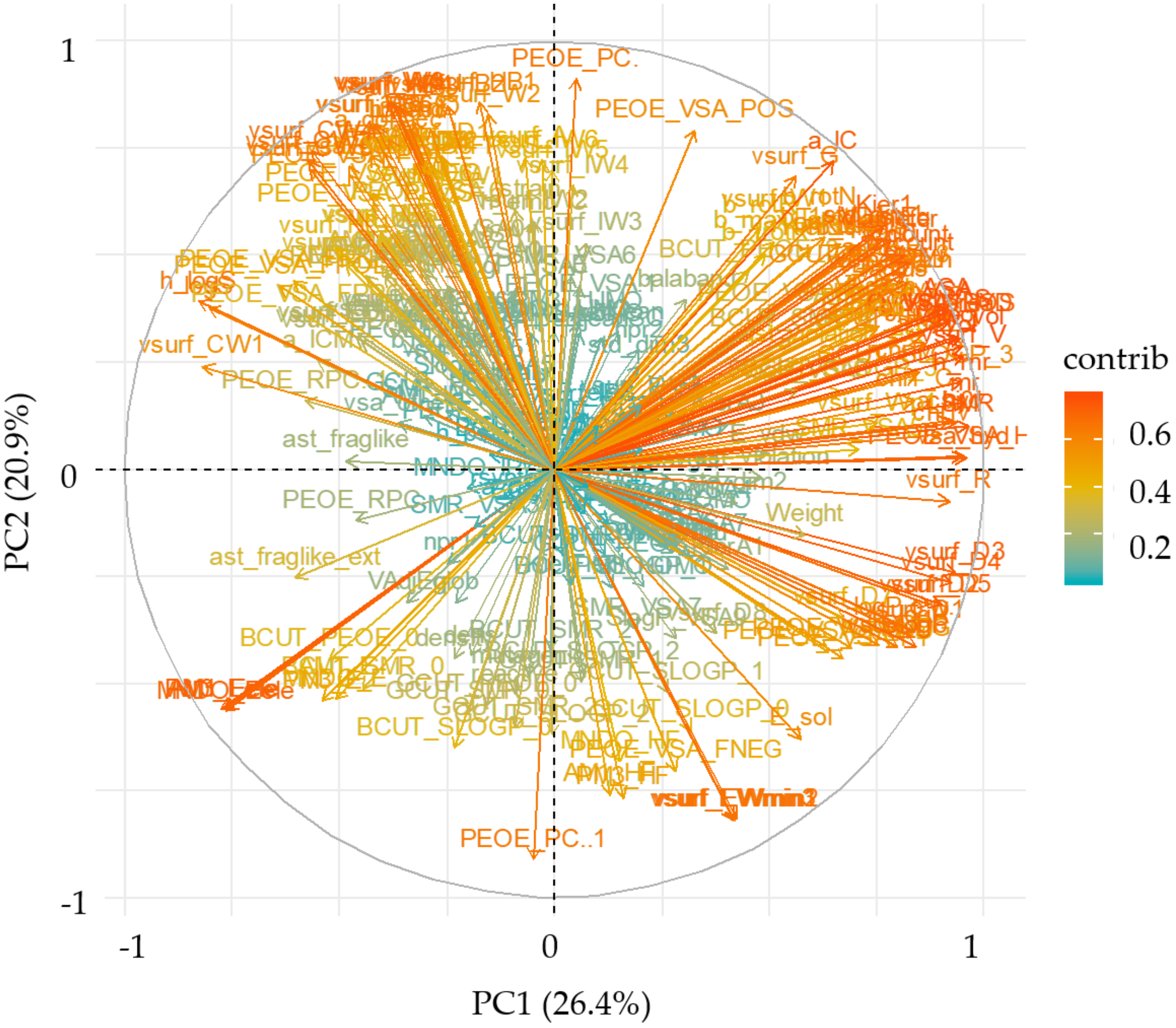


Figure 4.13 Variable factor map of the calculated molecular descriptors of 94 solvents from *Dataset A*

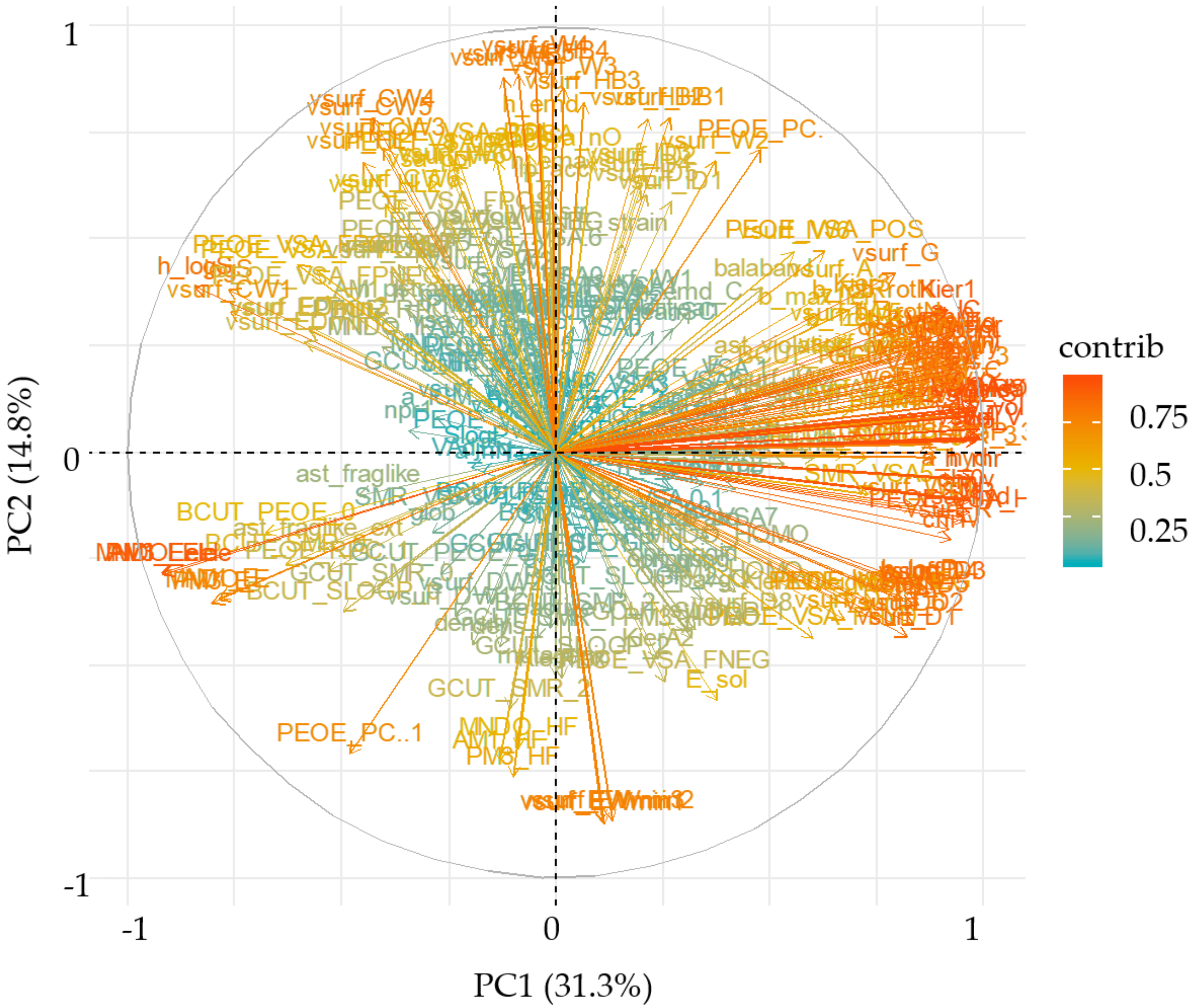


Figure 4.14 Variable factor map of the calculated molecular descriptors of 44 solvents from *Dataset B*

Table 4.7 Partial list of top 20 variables that contributed the most to the first two PCs for *Dataset A*. The larger the value, the more the variable contributed to the component.

Variables	PC1	Variables	PC2
SMR	1.260	PEOE_PC+	1.430
PEOE_VSA_HYD	1.249	PEOE_PC-	1.430
mr	1.243	vsurf_HB1	1.262
vsurf_V	1.236	vsurf_W3	1.262
vsa_hyd	1.228	vsurf_W4	1.261
vol	1.228	vsurf_HB2	1.241
h_mr	1.222	vsurf_HB4	1.239
vdw_vol	1.210	vsurf_HB3	1.220
chi0v	1.188	vsurf_W2	1.181
apol	1.188	vsurf_EWmin1	1.157
vsurf_D3	1.187	vsurf_EWmin3	1.156
vsurf_D4	1.163	vsurf_EWmin2	1.149
a_hyd	1.153	a_acc	1.129
vsurf_S	1.153	vsurf_HB5	1.111
vsurf_R	1.150	vsurf_W5	1.110
VSA	1.142	a_nO	1.086
ASA	1.133	lip_acc	1.085
vsurf_D5	1.103	h_emd	1.084
Q_VSA_HYD	1.101	PEOE_VSA_POS	1.075
Q_VSA_POS	1.101	a_don	1.022
……..		…………	

Table 4.8 Partial list of top 20 variables that contributed the most to the first two PCs for *Dataset B*.

Variables	PC1	Variables	PC2
vol	1.145	vsurf_W4	2.026
vdw_vol	1.144	vsurf_HB4	1.983
vsurf_V	1.140	vsurf_HB5	1.940
ASA	1.128	vsurf_W5	1.910
vsurf_S	1.128	vsurf_EWmin1	1.890
apol	1.126	vsurf_EWmin3	1.865
VSA	1.124	vsurf_EWmin2	1.859
h_mr	1.114	vsurf_W3	1.821
mr	1.102	vsurf_HB3	1.682
Q_VSA_HYD	1.100	vsurf_HB1	1.525
Q_VSA_POS	1.100	vsurf_HB2	1.525
vdw_area	1.100	vsurf_CW4	1.504
SMR	1.093	h_emd	1.470
chi0v	1.072	PM3_HF	1.451
PEOE_VSA_HYD	1.072	vsurf_CW5	1.416
vsa_hyd	1.064	AM1_HF	1.305
rgyr	1.037	vsurf_CW3	1.261
a_nC	1.031	PEOE_PC+	1.257
chi0_C	1.030	PEOE_PC-	1.257
chi0v_C	1.028	PEOE_VSA_POL	1.218
……..		…………	

From Table 4.7 for *Dataset A*, the first PC was found to mainly contain the descriptors defining the solvent's physical properties such as molecular refractivity (e.g. SMR, mr, h_MR), atom polarizabilities (apol), interaction field volumes (vsurf_V, vsurf_S, vsurf_R), surface area volume and shape descriptors which defined the hydrophobic parameters (vsa_hyd, vdw_vol, number of hydrophobic atoms and molecular surface rugosity. The second PC was found to mainly associate with the surface area, volume and shape descriptors which defined the hydrophilic parameters such as hydrophilic volume, hydrophilic energies, H-bond donor capacity etc and partial charge descriptors (PEOE_PC, PEOE_VSA_POS). Similarly, in Table 4.8 for *Dataset B*, the first PC was dominated by descriptors defining the solvents' van der Waals volume (vol, vdw_vol), van der Waals surface area (VSA), interaction field volume and surface area (vdw_V, vdw_S), solvent accessible surface area (ASA), sum of atomic polarizabilities (apol), molecular refractivity (h_mr, mr) etc. The second PC was mainly dominated by descriptors defining the hydrophilic properties of the solvent (vsurf_W, Vsurf_HB, vsurf_EWmin, vusurf_CW) and the H-bond donor capacity of the solvent. However, as PC1 accounted for 31.33% (Figure 4.7) of the variance compared to PC2, the most contributing variables for PC1 defined the solvent better in the chemical space plotted in Figure 4.12. Since the PCA for *Dataset B* did not significantly separate the two classes, no descriptors could be identified that correspond to paracetamol's crystal shape.

4.3.2.1.4 Biplots

Biplots were constructed to simultaneously display the scores of the solvents with the variable loading on the first two principal components. Superimposing both the score plot and the loading plot helps provide additional information about the relationship between the variables and the solvents that are absent in either of the individual plots (Jolliffe, 2002). Biplots of the first two PCs from the PCA performed on the interaction matrix of the solvents, and its calculated molecular descriptors for both *Dataset A* and *Dataset B* are presented in Figure 4.15 and Figure 4.16 respectively.

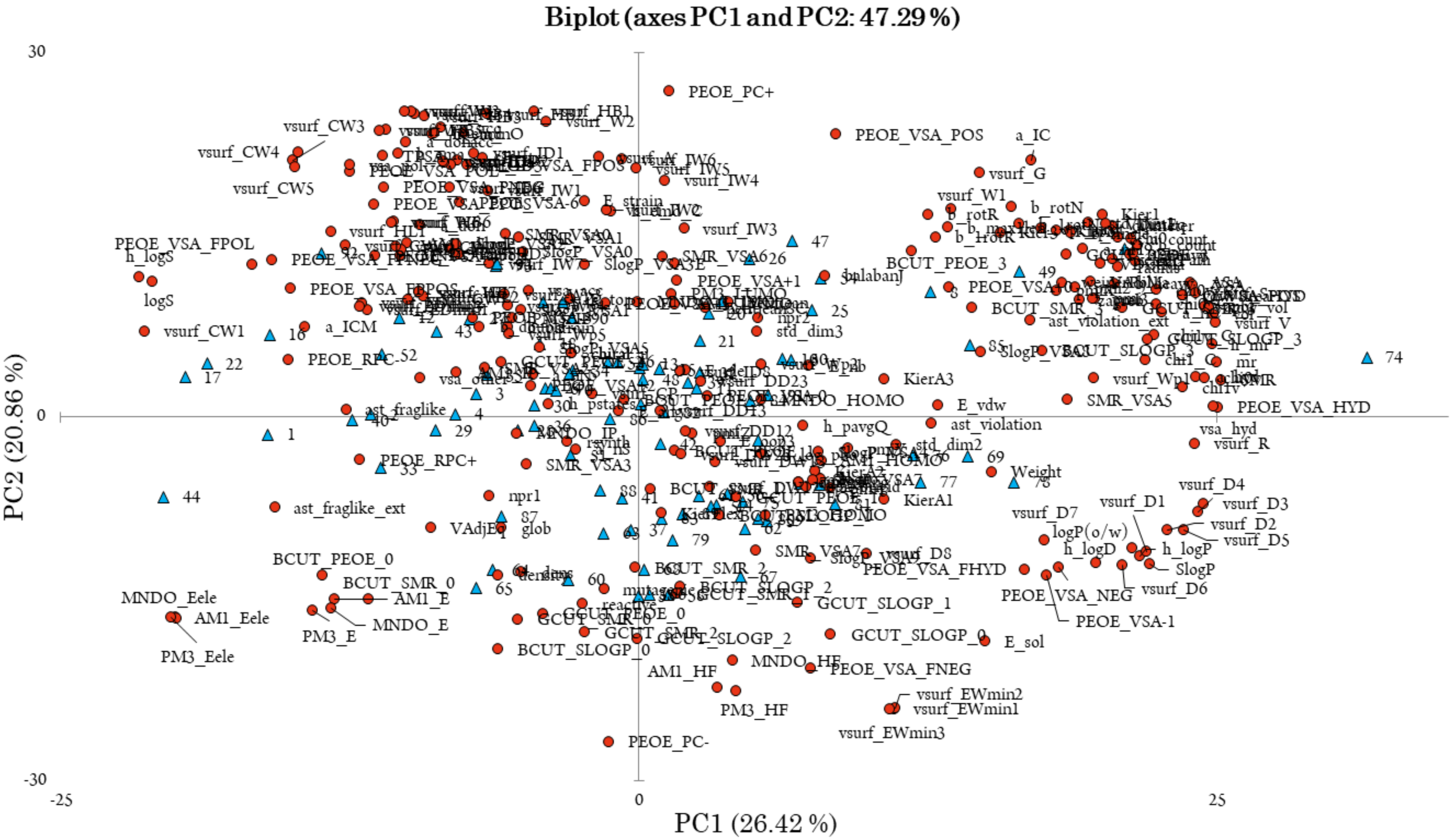


Figure 4.15 Biplot constructed using the first two PCs of *Dataset A* showing the overlay of the score plot and the factor map.

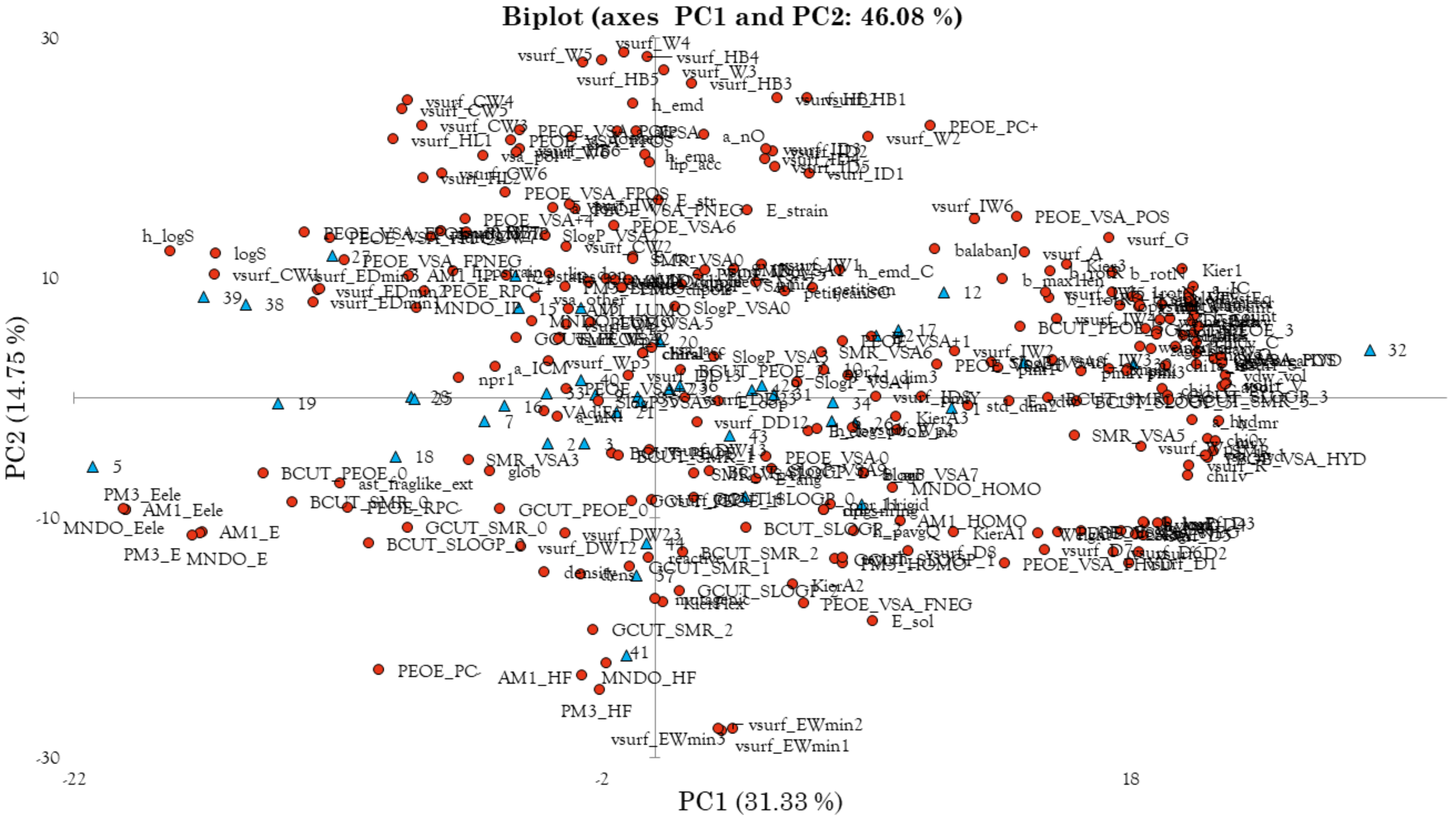


Figure 4.16 Biplot constructed using the first two PCs for *Dataset B* showing the overlay of the score plot and the factor map.

From the variables loading plot drawn in Figure 4.13 and Table 4.7, it was observed that the top contributing variables in PC1 for *Dataset A* were the various calculated descriptors defining molecular refractivity (SMR, h_mr, mr). The solvents lying on the right side of PC1 were found to complement this with higher values of molecular refractivity. Molecular refractivity and the viscosity of solvents are known to be functionally related which can also be observed from the biplot as solvents such as n-dodecane (74), 1-decanol (9), 1-methylnaphthalene (78), 1-nonanol (49), 1-octanol (8) and dibutyl ether (85) lay close to PC1 and are known to be viscous solvents (Lagemann, 1945). Most of the contributing variables explaining the first PC represented the hydrophobicity and lipophilicity parameters (vsa_hyd, vdw_vol, number of hydrophobic atoms and molecular surface rugosity). These parameters indicated the lack of hydrogen-bond accepting ability of the solvents and thus explained why non- polar solvents were mainly clustered at the lower half of the PCA score plot. Similarly, the contributing variables explaining the second PC in Figure 4.13 and Table 4.7 mainly described the hydrophilic parameters (vsurf_HB, vsurf_W, vsurf_EWmin1 etc.). Hydrophilic solvents are classified as polar solvents by IUPAC due to their ability to form intermolecular hydrogen bonds (McNaught, Wilkinson, Jenkins, International Union of, & Applied, 2006). Crystallisation outcomes were dependent upon the solubility of paracetamol in these 94 solvents. The influence of solubility and polarity can be visualised in Figure 4.8 and Figure 4.11 respectively. Most of the polar solvents present in the green region were capable of dissolving paracetamol compared to the non-polar solvents present in the red region.

Similarly, for *Dataset B*, the first component was dominated by the descriptors defining the solvent's van der Waals volume and the solvent's accessible surface area (vol, vdw_vol, VSA, ASA etc.) as shown in Table 4.8. The van der Waals volume and accessible surface areas for the 44 solvents decreased from right to left along the biplot with 1-decanol having the high values of van der Waals volume and accessible surface area compared to water. The trends observed for Dataset B were similar to those observed in Dataset A as it was a subset of Dataset A and contained the same molecular descriptors. The PCA analysis for Dataset B failed to provide any useful information about the crystal habit of paracetamol as the correlations of the variables in the dataset were either non-linear or showed no relationship with the investigated property. Thus, no further investigation was done and instead a non-linear ML algorithm was applied.

Finally, as PCA analysis is an unsupervised ML model, the patterns and trends observed are a clear indication of the ability of the calculated molecular descriptors to describe solubilisation behaviour rather than nucleation behaviour. The PCA model failed to distinctively cluster the solvents between the *non-nucleated* and *crystallised out* outcomes. Furthermore, clustering among the crystal habits indicated the lack of relative information within the calculated molecular descriptors to define these properties as well as insufficient data to build a comprehensive PCA model. A non-linear supervised ML algorithm, i.e. random forest was further applied to both datasets in the next section.

4.3.2.2 Supervised ML

4.3.2.2.1 RF Classification models trained on *Dataset A*

4.3.2.2.1.1 Classification accuracy and parameter tuning

Dataset A was randomly split into 80% as the training set and the remaining 20% as a test set. Table 4.9 presented the detailed breakdown of the dataset into the training and test sets for dataset A. To train the RF classification model; the crystallisation outcomes were labelled as A = *crystallised out*, B = *non-nucleated*, C = *practically insoluble* and D = *degradation*.

Table 4.9 Breakdown of *Dataset A* in the ratio of 80:20 with 80% as training set and 20% as a test set

	Dataset A				
	A	**B**	**C**	**D**	**Total**
Training Set	37	7	29	2	75
Test Set	7	4	7	1	19
Total	44	11	36	3	94

From Table 4.9, it can be observed that outcome D only represents a minor proportion (3%) of the overall data. Having such a small number of labels will have no effect on the classification performance of the model and hence could be omitted from the overall dataset. Thus the breakdown of the updated *Dataset A* in the ratio of 80:20 as a training set and test set is shown in Table 4.10.

Table 4.10 Breakdown of updated *Dataset A* in the ratio of 80:20 with 80% as training set and 20% as a test set

	Dataset A			
	A	B	C	Total
Training Set	37	7	29	73
Test Set	7	4	7	18
Total	44	11	36	91

The RF model was built on the dataset using 270 molecular descriptors. The internal parameters *ntree* and *mtry* were kept at default values to create a baseline for comparison, i.e. *ntree* = 500 and *mtry* = 16 (square root of the number of descriptors). 10-fold cross-validation was performed on the training set using the ‘caret’ function in R. RF classification model trained after performing 10-fold CV gave an average accuracy across the holdout predictions of 76.65% with an average Cohen's kappa statistics of 0.592. The obtained value of kappa, 0.592 fell within the range 0.4 to 0.6 which indicated that when validated the predicted outcomes from the trained model were in moderate agreement with the experimental outcomes (Altman, 1999).

After implementing the CV process, *ntree* and *mtry* were tuned to see if the performance of the RF model could be further improved. Increasing the number of trees (*ntree*) from its default value can result in better accuracy. However, it also leads to an increase in computational cost. After growing a certain number of trees, the improvement was found to be negligible as seen in Figure 4.17. According to Oshiro et al. (2012), after 128 trees, there was no significant improvement in the accuracy of the model tested on 29 data sets (Oshiro, Perez, & Baranauskas, 2012). Overall, the required number of trees depends upon the task and the total number of variables that are used. For this chapter, the value of *ntree* was set at 1500 trees.

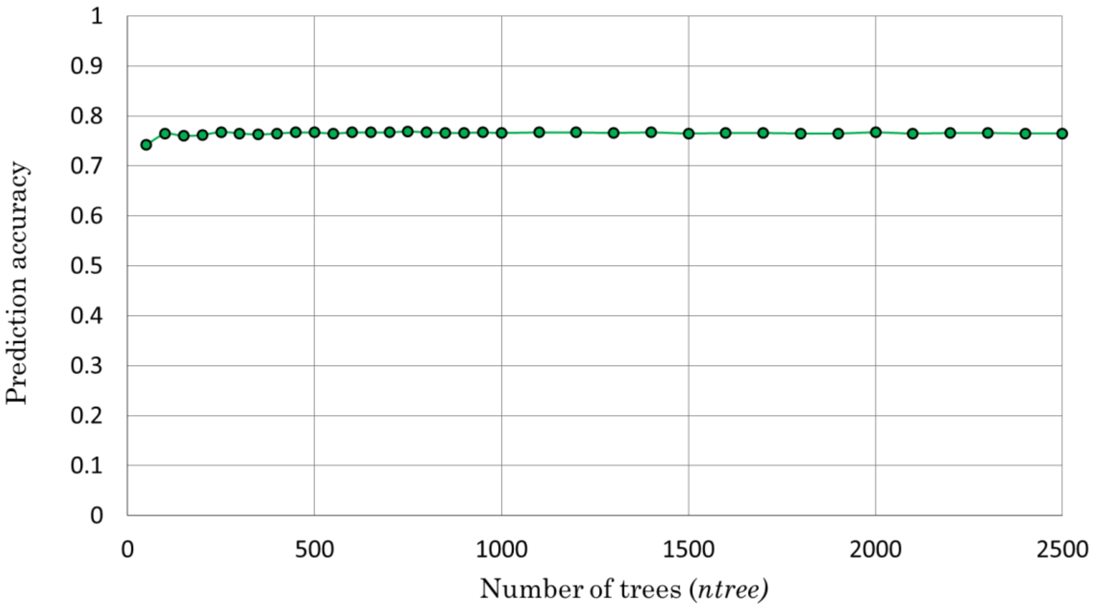


Figure 4.17 Average classification accuracy of the trained RF model with an increasing number of trees

Grid search approach on the *mtry* parameter was performed using the caret package in R which provided an excellent tool for tuning the *mtry* of the RF model. The best parameter value was then chosen based on the cross-validation results.

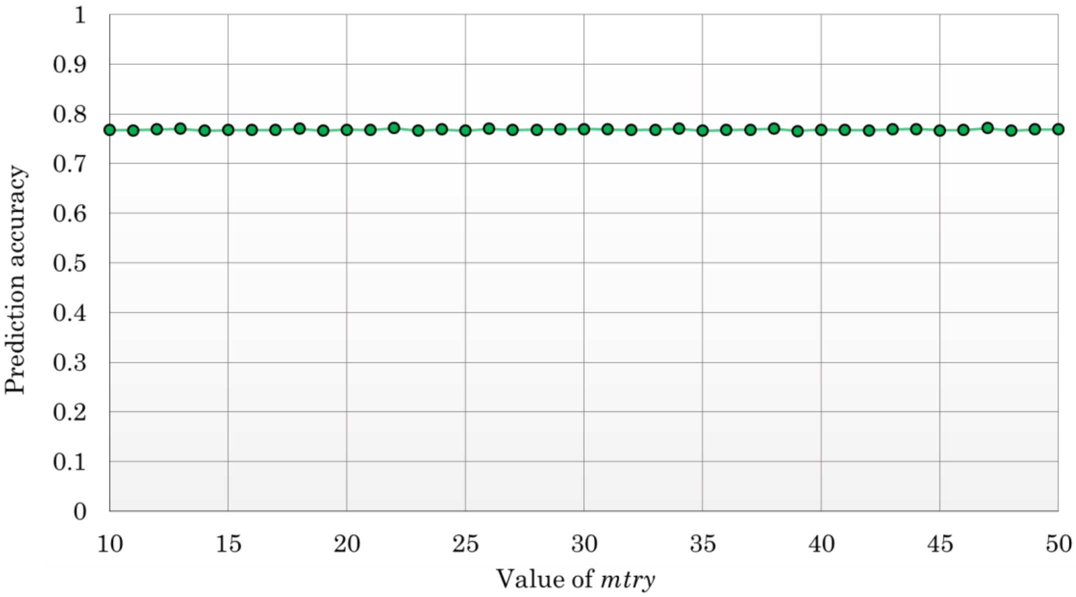


Figure 4.18 Average classification accuracy of the RF model with an increasing number of *mtry*

From Figure 4.18, it can be observed that the classification accuracy of the RF model did not vary significantly between the various values of *mtry*. The most accurate value for *mtry* was 22 with a classification accuracy of 77.17%. However, when compared with the default value of *mtry* at 16, the RF model gave a classification accuracy of 76.65%, which was not far off from the optimum value. As the varying value of *mtry* did not really make a drastic effect on the classification accuracy of the RF model, the best option was to keep the value of *mtry* for classification at its default value. Based on these analyses, the parameters for the RF classification models were selected with a value of *mtry* = 16, node size = 1 and *ntree* = 1500.

The confusion matrix in Table 4.11 showed the summary of the classification accuracy of the final RF model after 10-fold cross-validation with three crystallisation outcomes.

Table 4.11 Confusion Matrix describing the performance of the final RF model trained on the training set after 10-fold cross validation for *Dataset A* containing the three crystallisation outcomes

Predicted Data	**Reference Data**				
		A	**B**	**C**	**Class Error (%)**
	Crystallised Out (A)	32	1	4	13.514
	Non-Nucleated (B)	6	1	0	85.714
	Practically Insoluble (C)	4	0	25	13.793

The confusion matrix in Table 4.11 represented an OOB accuracy of 79.45 % for the three crystallisation outcomes. It was observed that the labels *crystallised out* and *practically insoluble* were correctly classified with an accuracy of 86.49% and 86.21% respectively by the RF model. However, the model failed to accurately classify the solvents where paracetamol was found to be *non-nucleated* (B) with 85.71% error.

In growing the RF model, a proximity matrix was computed between each pair of outcomes for the training dataset which can be useful in constructing Multi-Dimensional (MDS) plot. Like PCA, MDS is a dimension reduction technique aimed at projecting high dimensional data down to 2D or 3D dimensions while still preserving relative distances between the observations. Based on the prediction outputs from the trained RF model, MDS plots help in visualising clustering among the observations and identifying which ones are effectively close to one another based on their outcomes and how dissimilar each of the outcomes is with each other (Trevor Hastie, Tibshirani, & Friedman, 2009). MDS plot of the RF proximities was computed on the 75-training set based on the classification performance as shown in Figure 4.19.

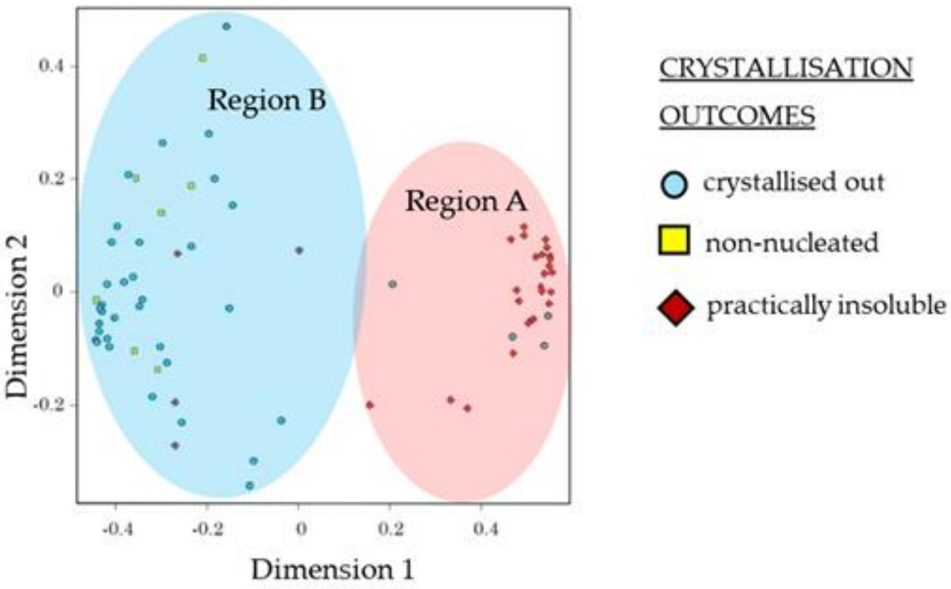


Figure 4.19 Illustration of the MDS plot on the training set obtained from the RF proximity matrix for *Dataset A* with its classified three crystallisation outcomes. The shaded oval regions A and B outlined the observed two separate clusters

The MDS plots in Figure 4.19 revealed two distinctive clusters in region A (practically insoluble solvents) and region B (crystallised out and non-nucleated solvents). The clustering observed from the MDS plot were similar to ones in the PCA score plot in Figure 4.8. The lack of clustering between the non-nucleated and crystallised out solvents is because both the RF model and the PCA models were trained using the same calculated molecular descriptors. Both the ML models failed to cluster as the descriptors used did not carry enough information to relate the properties of the solvent to these two outcomes. However, as RF is a supervised ML technique and is trained with the knowledge of the set outcomes, the separation between red coloured Region A and blue coloured Region B was larger compared to the unsupervised PCA.

The RF model was then applied to predict the crystallisation outcomes on the randomly selected 20% external test set which is shown by the confusion matrix in Table 4.12.

Table 4.12 Confusion matrix of the predictions made on the external test set

Predicted Data	Reference Data				
		A	B	C	Class Error (%)
	Crystallised Out (A)	7	0	0	0
	Non-Nucleated (B)	4	0	0	100
	Practically Insoluble (C)	0	0	7	0

The confusion matrix in Table 4.12 represented the prediction made on the 20% external test set. The overall prediction accuracy on the unknown test set was found to be 77.78%. Among the three crystallisation outcomes, both *crystallised out* (A) and *practically insoluble* (C) outcomes were predicted with 100% accuracy whist the label *Non-Nucleated(B)* was found to be predicted with 0% accuracy. Instead of similar to the classification in the training set, these labels were misclassified as being under the crystallised out category. These findings were discussed in more detail later in the section.

There are various strategies available that can be applied to solve a highly imbalanced dataset. These can be divided into three broad categories: data stratification processing based on sampling method such as 'stratified random sampling', 'random oversampling', 'random undersampling' and 'cluster-based oversampling', data stratification by modifying the existing machine learning algorithm and by application of cost-sensitive methods (Anyfamis, Karagiannopoulos, Kotsiantis, & Pintelas, 2007; Chawla, Bowyer, & Hall, 2002; Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). When faced with an imbalanced dataset, there is no standard method for improving the accuracy of the prediction model. The effectiveness of these techniques is dependent upon the characteristics of the imbalanced dataset.

For the imbalanced new dataset with three crystallisation outcomes, the 'stratified random sampling' method was applied. This stratification method aimed to balance the outcome distribution by dividing the sample population into smaller groups called 'strata' which is based on the samples shared molecular descriptors. Random sampling is then applied within each stratum. One of the significant advantages of this method is that it can capture key sample characteristics (Ye, Wu, Huang, Ng, & Li, 2013). The confusion matrix of the RF model after performing stratified sampling was presented in Table 4.13.

Table 4.13 Confusion Matrix describing the performance of the stratified RF classification model on the training set for the new dataset containing the three crystallisation outcomes.

	Reference Data				
		A	B	C	Class Error (%)
Predicted Data	Crystallised Out (A)	25	8	4	0.324
	Non-Nucleated (B)	5	2	0	0.714
	Practically Insoluble (C)	2	1	26	0.103

The stratified random sampling method was deemed ineffective as it did not improve the classification accuracy of outcome B by much as observed in Table 4.13. As outcome B kept being misclassified even after performing stratified sampling on the dataset, the focus turned to develop binary classifiers for each possible pair of outcomes and built an ensemble. Instead of training a single RF model using all the three crystallisation outcomes, One-Vs-One (OVO) binarization strategy was implemented. As the name suggests, OVO technique involved picking a pair of outcomes from a set of outcomes and training the RF model on every possible pair of outcomes (Furnkranz, 2002; Rifkin & Klautau, 2004). For the new dataset containing three crystallisation outcomes A, B and C, 3 pairs of binary outcomes were composed i.e. A against B, A against C and B against C as shown in Figure 4.20. Three RF models (RF1, RF2 and RF3) were then constructed using these binary combinations. OVO binarization strategy could help understand whether outcome B was misclassified due to class imbalance in the dataset or due to the choice of molecular descriptors used to train the model. The confusion matrix generated by the three RF models (RF1, RF2, RF3) after utilising OVO binarization technique is presented in Table 4.14.

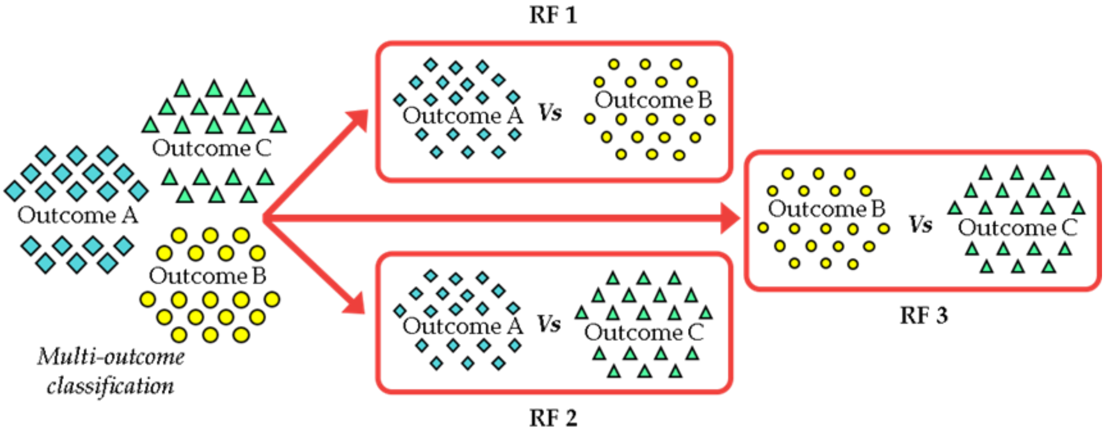


Figure 4.20 Illustration of the One-versus-One binarization strategy applied to the new dataset with three crystallisation outcomes.

Table 4.14 Confusion matrix generated by the three RF models (RF1, RF2 and RF3) after applying OVO binarization technique on the three crystallisation multi-outcomes.

	A Vs B [RF1]			**B Vs C [RF2]**			**A Vs C [RF3]**		
	A	**B**	**Class Error**	**B**	**C**	**Class Error**	**A**	**C**	**Class Error**
Crystallised Out (A)	37	0	0				33	4	0.108
Non-Nucleated (B)	6	1	0.857	5	2	0.286			
Practically Insoluble (C)				1	28	0.034	5	24	0.172
OOB Accuracy	0.86			0.916			0.863		
Cohen Kappa for training set	0.2189			0.722			0.739		

It can be generalised from the confusion matrices presented in Table 4.14 that the RF model was most accurate in classifying the solvents on binary outcomes A vs C and on the outcomes B vs C. The model however was least accurate in classifying the solvents on outcomes A and B as most of outcome B was misclassified as A. The conclusion that can be generalised from the confusion matrix is that the RF model with the calculated descriptors was able to accurately classify the solvents that crystallised and those that never nucleated against the practically insoluble solvents. In other words, the RF models were capable of predicting solubility of the compounds in the respective solvents rather than its nucleating behaviour. The confusion matrix also indicates that Cohen's kappa is a better classification metric for accurately pointing out the poor performance of the RF models rather than the OOB accuracy as all the three RF models had high OOB accuracy. The lower kappa value of RF1 compared to RF2 and RF3 indicates that the classification made in RF1 model was in poor agreement with the experimental outcomes.

4.3.2.2.1.2 Variable importance

OVO outcome binarisation strategy was also adopted for investigating outcome-specific variable importance to understand the relationship between the calculated molecular descriptors with the respective crystallisation property. Variable importance was assessed on only the two RF classification models: RF2 and RF3 as shown in Figure 4.21. The RF classifier RF1 which was trained to classify the outcomes *crystallised out (A)* and *non-nucleated (B)* was omitted for variable importance selection due to having a high misclassification rate for outcome B.

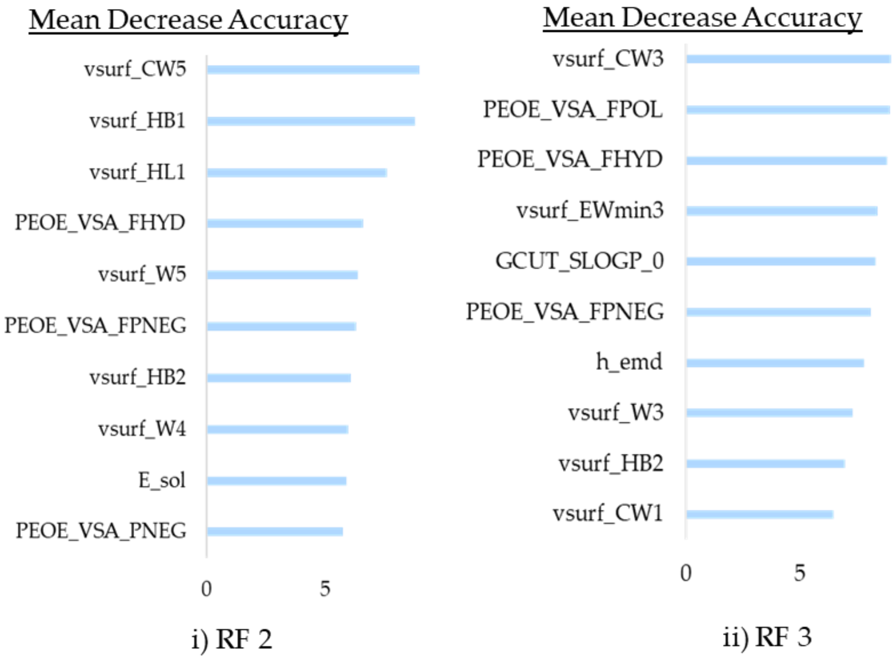


Figure 4.21 Graphical representation of the ten most important descriptors obtained for the RF model: i) RF2 model trained using outcomes *non-nucleated (B)* and *practically insoluble (C)* and ii) RF3 model trained using outcomes.

Table 4.15 Description of the ten most important molecular descriptors obtained from both RF2 and RF3 classification models calculated using MOE (Boyd, 2005)

Descriptor Category	Molecular Descriptor	Descriptor Description
Partial Charge Descriptors that combines the shape and electronic information to characterise the solvents	PEOE_VSA_FHYD	Fractional hydrophobic van der Waals surface area
	PEOE_VSA_FPNEG	Factional negative polar van der Waals surface area.
	PEOE_VSA_FPOL	Fractional polar van der Waals surface area
	PEOE_VSA_POL	Total polar van der Waals surface area
Adjacency and Distance Matrix Descriptors	GCUT_SLOGP_0	Descriptors calculated using the atomic contribution to logP and quantify lipophilicity
Semi-Empirical Quantum descriptors	h_emd	the sum of Extended Huckel Theory (EHT) donor strengths
Potential Energy Descriptors	E_sol	Solvation energy
Surface Area, Volume, and shape descriptors	vsurf_CW3, vsurf_CW5	The capacity factor of the molecules calculated at different energies. It provides information on the number of hydrophilic regions per unit surface area
	vsurf_EWmin2, vsurf_EWmin3	Lowest hydrophilic energy of the solvent
	vsurf_HB1	H-bond donor capacity
	vsurf_HL1	Hydrophilic-lipophilic ratio

	vsurf_IW1, vsurf_IW2	Hydrophilic integy moment
	vsurf_W4	Hydrophilic volume

Both the RF models were retrained by iteratively removing a portion of the least important molecular descriptors. The average predictive accuracy and its Cohen's Kappa after 10-fold cross validation on the training set with the reduced descriptors were evaluated. This process was continued until only the top ten molecular descriptors remained. Table 4.16 presented the average predicted accuracy obtained from 10-fold cross validation of both the RF models RF2 and RF3 respectively along with their calculated Cohen Kappa.

Table 4.16 Performance measure of the 10-fold CV models calculated by iteratively removing the least important molecular descriptors for both RF2 and RF3 classification model

	RF2 model		RF3 model	
Number of Variables in Dataset	Average CV accuracy (%)	Kappa	Average CV Accuracy (%)	Kappa
278	94.44	0.823	86.36	0.739
139	94.44	0.823	86.36	0.739
70	94.44	0.823	86.36	0.739
25	94.44	0.823	87.88	0.754
15	94.44	0.823	87.88	0.754
10	94.44	0.823	87.88	0.754

It was observed that removal of the least important molecular descriptors did not affect the predictive accuracy of the RF classification accuracy for both the RF models RF2 and RF3. The model trained using only the top 10 molecular descriptors provided the same average predictive accuracy as the ones trained using all the calculated molecular descriptors. This further proves the predictive capability of the ensemble algorithm and its ability to handle relatively high dimensional data and small sample sizes. This has also been proved and tested in various studies where RF built classification models have shown good predictive accuracy when trained on small sample size and high dimensions (C. Li, 2016; Schwarz, König, & Ziegler, 2010; Svetnik et al., 2003; B. X. Xu, Huang, Williams, Wang, & Ye, 2012).

Even though reducing the variables showed no change in the average predictive accuracy of the CV RF model, it is rather unwise to simply remove the molecular descriptors with less importance. Reduction in the molecular descriptors does not really make a big reduction in computation time (Banfield, Hall, Bowyer, & Kegelmeyer, 2007; Breiman, 2001a; Palmer et al., 2007). Furthermore, there exists intercorrelation between some of the calculated molecular descriptors which have an influence in the order of importance. For example, if two molecular descriptors were highly correlated, either one of the two can be selected to make a highly similar split in a decision tree. Since, RF model indiscriminately uses either one of the descriptors to yield the optimum split regardless of their intercorrelation, the order of the calculated variable importance for the ensemble model can be distorted. It is therefore difficult to pinpoint the exact order of the important molecular descriptors contributing to the prediction accuracy of the model.

4.3.2.2.1.3 Assessment of important molecular descriptors

The top ten descriptors for the classification model RF2 built on outcomes *non-nucleated (B)* against *practically insoluble (C)* mainly consisted of vsurf_ descriptors (vsurf_CW5, vsurf_HB1, vsurf_HL1, vsurf_W5, vsurf_HB2 and vsurf_W4). As mentioned in Table 4.16, the vsurf_ descriptors define the surface area, volume and shape properties depending upon the structural connectivity and the conformation of the drug molecules. Similar to the volsurf descriptors, the vsurf_ descriptors help describe the interaction of the drug molecule with the hydrophobic and hydrophilic part of the solvent through surface properties such as shape, electrostatic, hydrogen-bonding and hydrophobicity (Cruciani, Crivori, Carrupt, & Testa, 2000; Omkvist et al., 2010). Solvents possessing higher values of these vsurf_ descriptors and lower values of the partial charge descriptors (PEOE_VSA_FHYD, PEOE_VSA_FPNEG and PEOE_VSA_PNEG) were found to favour outcome *B* i.e. *non-nucleated*. vsurf_HB descriptor represents the hydrogen bonding capabilities of the solvents.

Finally, the solvation energy descriptor E_sol contributed negatively to the outcome B. Like the RF2 model, the top ten descriptors for the classification model RF3 built on outcomes '*crystallised out*' against '*practically insoluble*' mainly consisted of the vsurf_ and partial charge descriptors (Figure 4.21). Both the RF2 and RF3 models showed similar top molecular descriptors with the addition of adjacency and distance matrix descriptors (GCUT_SLOGP_0).

Since solvents with outcomes *crystallised out* and *non-nucleated* must have initially solubilised paracetamol, they can be merged and grouped as soluble (i.e. the difference is due to the nucleation (or lack of) and not dissolution). The similarity in the variable importance between the two RF models (RF2 and RF3) indicates this statement as these descriptors describe solubilisation of paracetamol better than nucleation. Thus, this explains why there were persisting misclassification and a lack of separation between the outcomes *crystallised out* against *non-nucleated* in both supervised and unsupervised ML models.

Even though these listed important molecular descriptors obtained from RF2 and RF3 classification models help understand the relationship between the descriptors and the crystallisation outcomes A, B and C, it is to be noted that these descriptors are mainly representative of the paracetamol crystallisation model and will vary depending upon the choice of drug molecule selected for the study. However, these descriptors still act as an important factor in understanding the solute-solvent interaction and provides vital information on predicting the crystallisation outcomes.

4.3.2.2.2 RF Classification models trained on *Dataset B*

Similar methods as mentioned in section 4.3.2.2.1 were followed for dataset B. As dataset B is a subset of *Dataset A* i.e. only the solvents that paracetamol crystallised out from were chosen, the total sample size is comparatively small. The crystal shape outcomes for *Dataset B* were labelled as *Shape A* and *Shape B* to train the RF classification model. A random stratified split of the dataset into 80% as a training set and the remaining 20% as test set was performed as the dataset was imbalanced.

Table 4.17 Breakdown of *Dataset A* in the ratio of 80:20 with 80% as a training set and 20% as a test set

Dataset B			
	Shape A	Shape B	Total
Training Set	30	5	35
Test Set	5	4	9
Total	35	9	44

The RF classification model after performing ten repeats of 10-fold cross-validation on the training set gave the performances across holdout predictions to have an average OOB accuracy of 91.43 % with an average Cohen's Kappa statistic of 0.4. The average classification accuracy of the 10-fold CV RF model here was high but misleading. This is because the value of Cohen's Kappa indicated a poor level of agreement between the model and the experimental outcomes which can be further investigated using the confusion matrix of the RF model. The RF classification model was trained using the default value of m_{try} and 1500 trees. The confusion matrix in Table 4.18 and Table 4.19 showed the classification accuracy of the RF model on the training set and the prediction accuracy on the unknown test set respectively.

Table 4.18 Confusion Matrix describing the performance of the RF classification model on the training set for *Dataset B* containing the two crystal shape outcomes

Predicted Data	Reference Data			
		Shape A	Shape B	Class Error (%)
	Shape A	30	0	0
	Shape B	3	2	60

Table 4.19 Confusion Matrix describing the predictive accuracy of the RF classification model on the test set for *Database B* containing the two crystal shape outcomes

Predicted Data	Reference Data			
		Shape A	Shape B	Class Error (%)
	Shape A	24	6	20
	Shape B	2	3	40

The confusion matrix presented in Table 4.18 showed an OOB classification accuracy of the RF model as 91.43 % for the two crystal shape outcomes. The Cohen Kappa value was calculated as 0.4. *Shape B* crystal shapes were accurately classified while only 2 out of the 5 *Shape A* crystals were correctly classified by the RF classifier. Similarly, when predicted on the unknown test set, the prediction accuracy was obtained as 55.56% with a kappa value of 0 (Table 4.19). From the confusion matrices, it can be confirmed that the RF model simply failed to correctly predict the crystal habit of paracetamol on organic solvents. This could be due to the small size of the dataset (44 samples). The smaller dataset doesn't have the capability to provide meaningful performance estimates and ML models trained on small datasets are more prone towards overfitting (especially on the validation set).

Furthermore, due to the small dataset size for crystal shapes, the molecular descriptors utilised could potentially not provide relevant information for prediction.

'Stratified random sampling' method as applied in 4.3.2.2.1.1, of the chapter for *Dataset A* was applied to improve the RF model on the imbalanced *Dataset B*. It can be observed from Table 4.20 that applying stratification did improve the classification accuracy but not by much. This indicates that, along with the very small sample size, the RF model could not find trends between the calculated variables and the crystal habit of paracetamol in the respective solvent. Thus, when the model was trained for *Dataset B*, the predictions were simply biased toward the majority outcome i.e. *shape A*. No further statistical analysis was done on *Dataset B* as no relationship between the variables and crystal habit could be observed.

Table 4.20 Confusion Matrix describing the performance of the stratified RF classification model on the training set for *Dataset B* containing the two crystal shape outcomes.

	Reference Data			
		Shape A	Shape B	Class Error (%)
Predicted Data	Shape A	24	6	20
	Shape B	2	3	40

4.4 Summary

Machine learning algorithms were applied to predict the crystallisation outcomes of paracetamol on 94 solvents and crystal habit on 44 organic solvents where it has crystallised. Controlled cooling crystallisation was performed to generate the dataset for the ML algorithm. Four crystallisation outcomes: *crystallised out*, *non-nucleated, practically insoluble* and *degradation* were observed from the controlled cooling crystallisation on 94 solvents. Out of the 94 solvents, paracetamol was crystallised in 44 solvents. These observed crystals were categorised as *Shape A* and *Shape B a*ccording to their observed crystal habit on the respective solvent. Both unsupervised (PCA) and supervised (RF) machine learning algorithms were utilised to aid in the prediction of both the crystallisation outcome and crystal habit.

The score plot obtained from the PCA on the crystallisation outcomes of 94 solvents presented two distinctive clusters of degradation and practically insoluble and a third cluster comprising the crystallised out and non-nucleated solvents (Figure 4.8). The MDS plot obtained from the RF classification model showed similar clustering to PCA with two distinctive cluster regions: Region A (practically insoluble) and Region B (crystallised out and non-nucleated) (Figure 4.19). Solvents where paracetamol showed degradation were removed from the RF model as this response represented a small minority (3%) in the overall dataset. The trained RF model classified the three crystallisation outcomes of paracetamol in 91 solvents with an average classification accuracy of 79.45% and kappa value of 0.63. The trained model was capable of predicting the outcomes on the 20% external test set with accuracy of 77.78% and kappa value of 0.64.

Even though, both the classification accuracy and predictive accuracy of the trained RF model were relatively high, it was observed from the confusion matrices that the RF model had accurately predicted the crystallisation outcomes: *crystallised out* and *practically insoluble* but failed to accurately classify and predict the response: *non-nucleated*. The accuracy of the model came mostly from the two well represented classes: *crystallised out* and *practically insoluble.*

As the dataset was imbalanced with the response *Non-Nucleated* a minority in the overall dataset, stratified sampling and one-vs-one binarisation techniques were further applied on *Dataset A* to deal with the imbalanced dataset and understand the precise distinctions between classes the model succeeded and failed to capture. It was observed that performing stratification did not improve the classification nor prediction on the minority response (*Non-Nucleated).* Performing the OVO binarisation techniques further concluded that the trained RF model was indeed capable of accurately classifying and predicting the outcomes: '*crystallised out*' and '*practically insoluble*' but the model simply failed to predict the non-nucleating behaviour of paracetamol in the respective solvents.

As only 12% of the outcomes was *non-nucleated*, it was originally assumed that the model failed to simply represent the minority outcomes. However, upon performing stratification and the OVO binarisation techniques concluded that small sample size was not the only problem but rather the choice of the molecular descriptors utilised to train the model could have failed to clearly classify the outcome *non-nucleated.* It is well known from literature that external factors such as supersaturation, impurities, process time and operating conditions are known to have more dominant roles in inducing nucleation (Myerson & Ginde, 2002). It has also been reported that physical properties such as viscosity, molecular weight and density of solvents were found to directly affect the onset of nucleation while surface tension, dielectric constant and concentration were reported to indirectly affect the onset of nucleation (Storm et al., 1970). These properties were not included in the model and thus could have contributed to the error in misclassification.

Both unsupervised and supervised ML algorithms failed to predict the nucleating behaviour of paracetamol in 91 solvents. However, an observation was made whilst training the ML algorithms. The outcome *non-nucleated* were constantly misclassified as being 'crystallised out' rather than '*practically insoluble*' by the algorithms. Both the PCA score plot and the MDS plot from the RF model also clustered these two outcomes together. According to the predictions made by the algorithms, had the solution been left to crystallise out for an unspecific infinite amount of time, paracetamol may have eventually crystallised. It is not impossible for this to happen as the set induction time for the experiments were controlled. However, in reality, it is impracticable to leave the solution to crystallise for an infinite amount of time.

Since the outcome '*non-nucleated*' was constantly misclassified as '*crystallised out*', it can simply be combined and classified as one outcome *'soluble'*. This category is true as in order for the compounds to crystallise out, it must first dissolve in the respective solvent. Thus, it can be concluded that the ML algorithms were actually capable of accurately predicting the solubilisation behaviour of paracetamol in the respective solvent rather than its the nucleating behaviour. This is an important observation as it can also confirm that the molecular descriptors utilised for the project were defining the solubility of paracetamol in 91 solvents and thus along with the ML algorithms can be utilised to develop a rapid and efficient solvent selection tool.

For the model trained to predict the crystal habit of paracetamol, the score plot obtained from PCA on the crystal habit outcomes of 44 solvents showed no significant clustering between the two outcomes, *Shape A* and *Shape B*. The *Shape A* outcomes were mostly clustered along PC1 while the *shape B* outcome was spread around the scatter plot (Figure 4.12). The RF model was not successful in accurately classifying the outcomes *Shape B* as the size of *Dataset B* was very small and highly imbalanced. Collating crystal habit data on a particular API can prove challenging as only the number of solvents are quite limited compared to the abundance in a number of drug APIs. It would be wiser to perform this study orthogonally, i.e. predicting a dataset containing crystal habit observations of various drug compounds on a single solvent.

Chapter 5.　Development of a rapid and efficient solvent selection tool

5.1 Introduction

The selection of solvent is a fundamental step in the design of an efficient and effective solvent-based crystallisation process that can have a significant influence on the product quality and the manufacturing process. Physical properties such as solubility, polymorphism, crystal shape, size and habit of a crystalline product are all influenced by the choice of a solvent selected. These properties, in turn, will influence the downstream processing of the drug product and its quality (Maghsoodi, 2015; Variankaval, Cote, & Doherty, 2008). The use of solvents also accounts for greater than 70% of the total waste from the pharmaceutical process as reported from a study carried out by Jimenez-Gonzalez and co-authors at GlaxoSmithKline (Jimenez-Gonzalez, Ponder, Broxterman, & Manley, 2011). Thus, the solvent selection is not only crucial for the manufacture of high-value drug products but also fundamental in the design of an efficient and sustainable chemical process.

Solvent selection is normally performed at the early stages in the design for a crystallisation process development. There are various systematically designed tools available to aid in solvent selection for a crystallisation process. Most of them are based either on thermodynamic criteria or upon the solvent's impact on Safety, Health and the Environment (SHE). Major pharmaceutical companies such as GlaxoSmithKline, AstraZeneca, Pfizer and Sanofi-Aventis have published their own solvent selection guides where the solvents are ranked based on their impact on SHE, economic criteria, regulatory concerns and on some physical properties of the solvent such as polarity, melting point and boiling points (Curzons, Constable, & Cunningham, 1999; Henderson et al., 2011; Prat et al., 2016). However, a recent publication on solvent selection by AstraZeneca has moved away from its original SHE-only focus to more of a comprehensive process design solvent selection tool that covers chemical reactivity, tautomerization and molecular conformation (Diorazio et al., 2016). There are many publications highlighting and comparing the various available predictive methods based on thermodynamic models for solvent selection. Gani et al (2006) published an article comparing the various solvent selection approaches, detailing tools and software databases used in both laboratory and industry environments (Gani et al., 2006).

Solvent selection for any given crystallisation process is not a straight forward process as most pharmaceutical drug products are multifunctional, polarizable and can form specific interactions with the solvent (Kolar et al., 2002). The change from using one solvent to another remains a major challenge as the new solvent brings significant changes in several of the chemical properties of the solution. Furthermore, there are hundreds of solvents to choose from and new solvents being introduced which makes the selection process tedious and computationally challenging. The choice of solvent selected is also determined by many other factors such as literature precedent, pharmaceutically acceptable solvents, solubility, trial and error or simply a favourite solvent of the chemist and the solvent's availability in the laboratory. Thus, there is a need for the development of an efficient and effective solvent selection tool.

The model developed for predicting the crystallisation behaviour of paracetamol in various solvents indicated that the trained models were capable of predicting solubility rather than nucleation behaviour (Chapter 4). Thus, this finding led to the implementation of ML algorithms in the development of a rapid and efficient solvent selection tool for continuous crystallisation process. The chapter also investigates on the minimum number of data points required to train an effective ML model, attempts on filtering specific solvents from the solvent database which are involved in the training set of a highly accurate predictive model and accesses the important molecular descriptors which played an important role in predicting relative solubility.

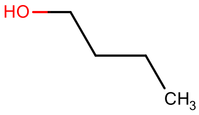
5.2 Methodology

A diverse range of 63 solvents selected for the screening study and their molecular structures are presented in Figure 5.1. Solvents were selected based on their availability in the laboratory and were not limited to pharmaceutically accepted solvents. Paracetamol, carbamazepine and carvedilol were the chosen compounds on which the solvent screening was studied. These compounds and solvents were chosen in line with a greater piece of work as part of the Continuous Manufacturing and Crystallisation Future Manufacturing research hub (C. J. Brown et al., 2018).

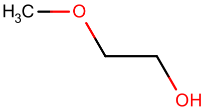
All the cooling crystallizations performed for this study were conducted in uniform conditions which included a constant weight for the chosen compounds, constant volume for the solvents and the utilization of the same instrument. Standardization was performed in order to collect comparable data. For each target compound, 50 mg was weighed precisely and added to 1ml of the respective solvent in an HPLC vial using the Zinsser Automated platform. Cooling crystallisation was then performed on *Crystal16* Reactor systems (Technobis) following the steps mentioned in Section 3.2.3 (Material and Methods). Operating conditions such as heating rate, cooling rate and stirring speed were kept constant throughout the experiment for all three APIs. Transmissivity observed at the end of each temperature cycle was noted for each solvent. Compounds with transmissivity observed at 100% were labelled as being *soluble* in the respective solvent as no floating particles were visible in the solution at the end of the cycle while compounds with transmissivity observed below 100 % were labelled as being *practically insoluble*. These two experimental responses were used as the class labels for the random forest model. The experiment conditions for this study are summarized in Table 5.1.

H_2O

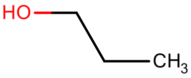
water



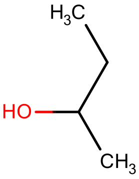
1-butanol



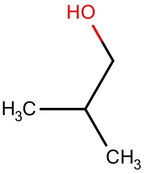
2-methoxyethanol



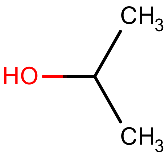
1-propanol



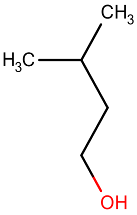
2-butanol



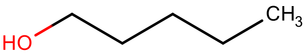
2-methyl-1-propanol



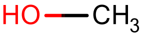
2-propanol



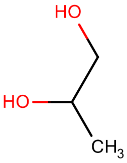
Isoamyl alcohol



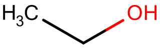
1-pentanol



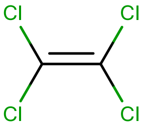
methanol



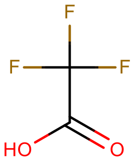
1,2-propanediol



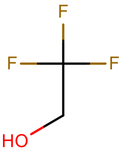
ethanol



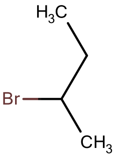
tetrachloroethene



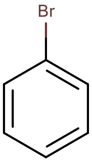
trifluoroacetic acid



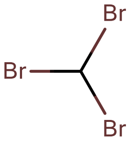
2,2,2-trifluoroethanol



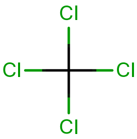
2-bromobutane



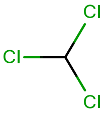
bromobenzene



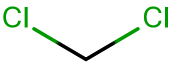
bromoform



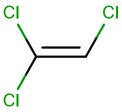
carbon tetrachloride



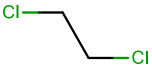
chloroform



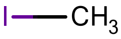
dichloromethane



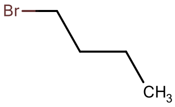
trichloroethylene



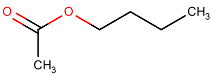
1,2-dichloroethane



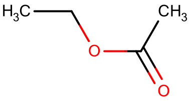
iodomethane



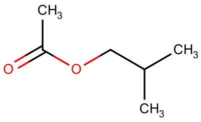
1-bromobutane



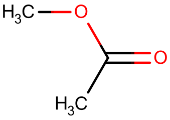
butyl Acetate



ethyl acetate

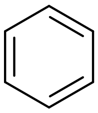


isobutyl

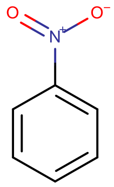


methyl

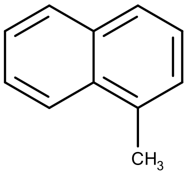
acetate Acetate



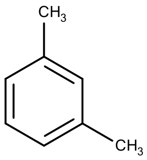
benzene



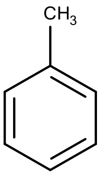
nitrobenzene



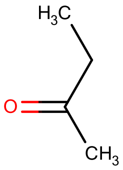
1-methylnapthalene



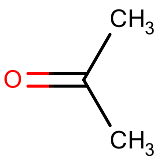
m- xylene



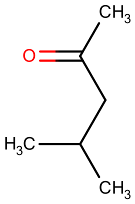
toluene



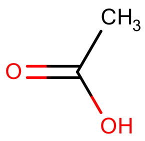
2-butanone



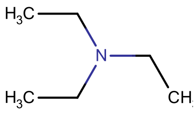
acetone



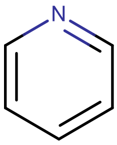
4-methyl-2-pentanone



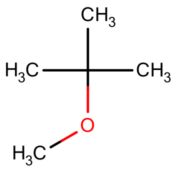
acetic Acid



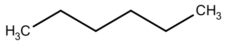
triethylamine



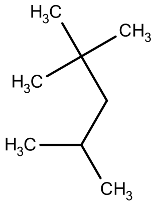
pyridine



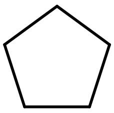
2-methoxy-2-methylpropane



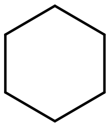
hexane



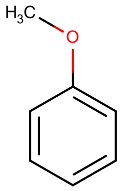
isooctane



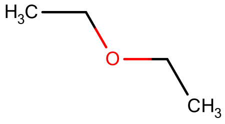
cyclopentane



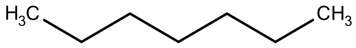
cyclohexane



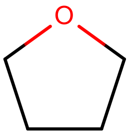
anisole



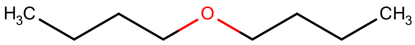
diethyl ether



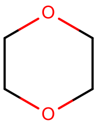
heptane



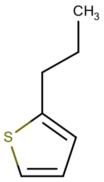
Tetrahydrofuran



dibutyl ether



1,4-dioxane



2-propane thiol

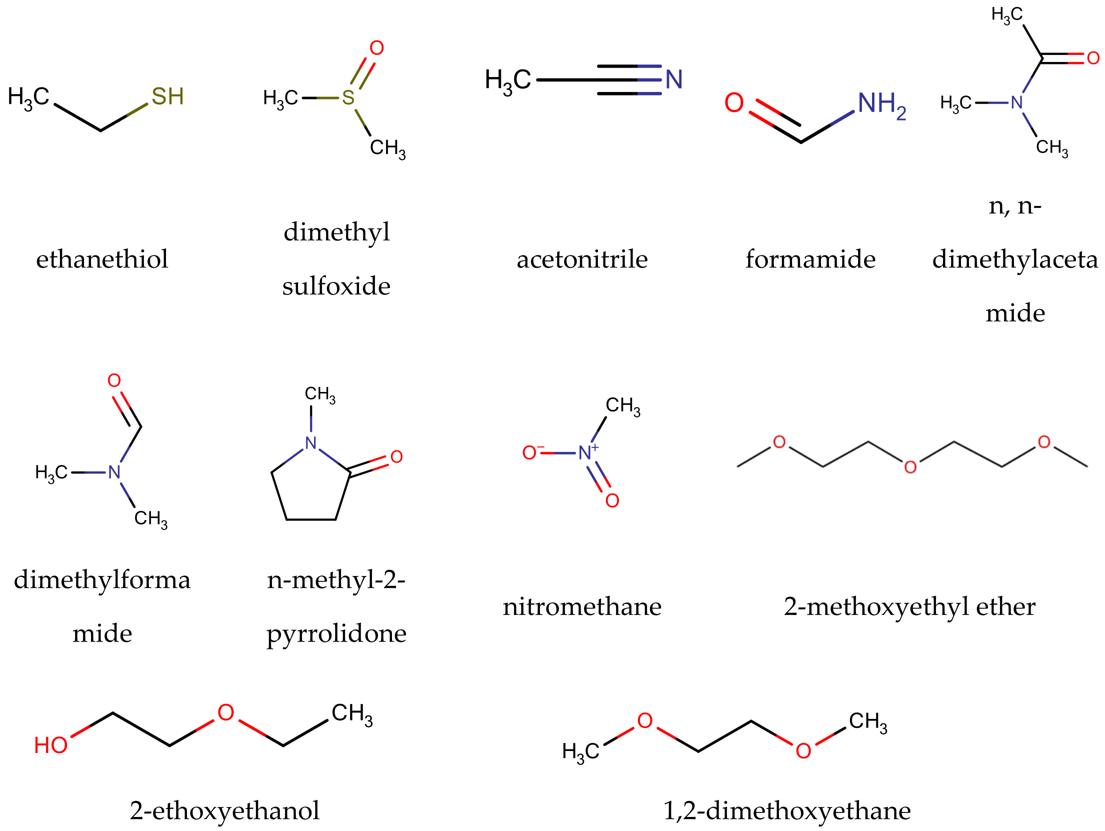


Figure 5.1 Molecular structure of the solvents selected for this study

Table 5.1 Summary of the experimental conditions and classification for obtaining the relative solubility of the chosen APIs

Weight of the target compound	50mg
The volume took for each solvent	1 ml
Stirring Speed	1000 rpm
Low-temperature point	20^{0}C
Elevated temperature point	10^{0} C below the boiling point of the solvent and capped at 100^{0} C for solvents with higher boiling points
Heating and Cooling rate	0.1^{0}C per minute
Transmissivity at 100%	soluble
Transmissivity below 100%	Practically insoluble

5.3 ML workflow

The first stage involved the calculation of the molecular descriptors of the 63 solvents. 2D structure of the solvents based on their canonical SMILES were constructed using the Biovia Pipeline pilot 2017 software. Energy minimisations on these structures were than performed using the Clean force-field in Pipeline pilot 2017 (Hahn, 1995). Molecular Operating Environment (MOE), an all-in-one molecular modelling and visualization tool was used to calculate 340 physicochemical 2-D and 3-D molecular descriptors. The list of the 2-D and 3-D calculated molecular descriptors obtained from MOE are categorised and listed in Chapter 3: *Material and Methodology* (P. Labute, 2000). Pre-processing was performed on the dataset containing the calculated molecular descriptors and the solubility. Pre-processing involved the removal of the descriptors with zero variance and highly correlated descriptors (>0.95) using the statistical computing environment R (version 3.3.1) (Andy Liaw & Matthew Wiener, 2002). The final reduced number of molecular descriptors after pre-processing was 170 molecular descriptors. The

Random forest model was built using the randomForest package in R. The overall dataset for the three APIs contained the 250 calculated molecular descriptors and 63 solubility outcomes.

A good solvent selection tool promotes the need for fewer screening experiments to reduce time and expense in the laboratory. Similarly, the success of a machine learning model is highly dependent upon both the quantity and quality of the data available. The larger the number of data points, the better the performance of the machine learning algorithm and thus delivery of accurate results. The quality of the dataset is dependent upon the accuracy of the experimental observations, choice of observations in the training dataset and how balanced the dataset was for the classification model (Witten, Frank, & Hall, 2011).

In order to balance the desire for fewer experiments with the need for an effective machine learning predictive model, the first objective was to identify the optimum number of training sample size (*ntrain)* required to train an effective random forest classification model. Even though any good algorithm available struggles with a small dataset, one of the many unique advantages of random forest is its ability to deal with small sample sizes, high-dimensional feature spaces and complex data structures which makes it very suitable for this process (GÃŠrard Biau, 2012; G. Biau & Scornet, 2016; Y. Qi, 2012). Thus, to find the optimum training set size (*ntrain)*, the RF model is trained using default parameters of *mtry* (number of variables available for splitting at each tree node) and *ntree* (number of trees grown) with varying training set sizes between 8 to 32 samples while keeping the hold-out test set fixed at 28 samples. Resampling – taking random samples from the dataset, with replacement – is performed for each training set to evaluate the performance of the predictive model. For this study, 10,000 resampling iterations were performed per value of training set as outlined in the workflow in Figure 5.2. The consistency of the classification models at each size of training sets was evaluated, revealing the level of accuracy that was expected for a given value of the training set. Each model was tested on a randomized test set of 28 solvents. The mean value of the prediction error rate at each size of the training set was plotted along with quartile ranges of the error rate. The actual prediction (AP) error rate is the total number of correct classifications predicted over the total number of samples in the dataset as compared to the out-of-bag (OOB) error rate which is the fraction of the number of incorrect classifications over the total number of out of bag samples. Between the two, the OOB error rates are mainly useful for determining hyper-parameters while AP is preferred for estimation of the actual performance of the trained RF model. The data from the training set sizes with low mean prediction error rates were then mined using an in-house python code to determine the optimum number of training set size and a method for recommending suitable solvents in the training set to build a robust and accurate RF model.

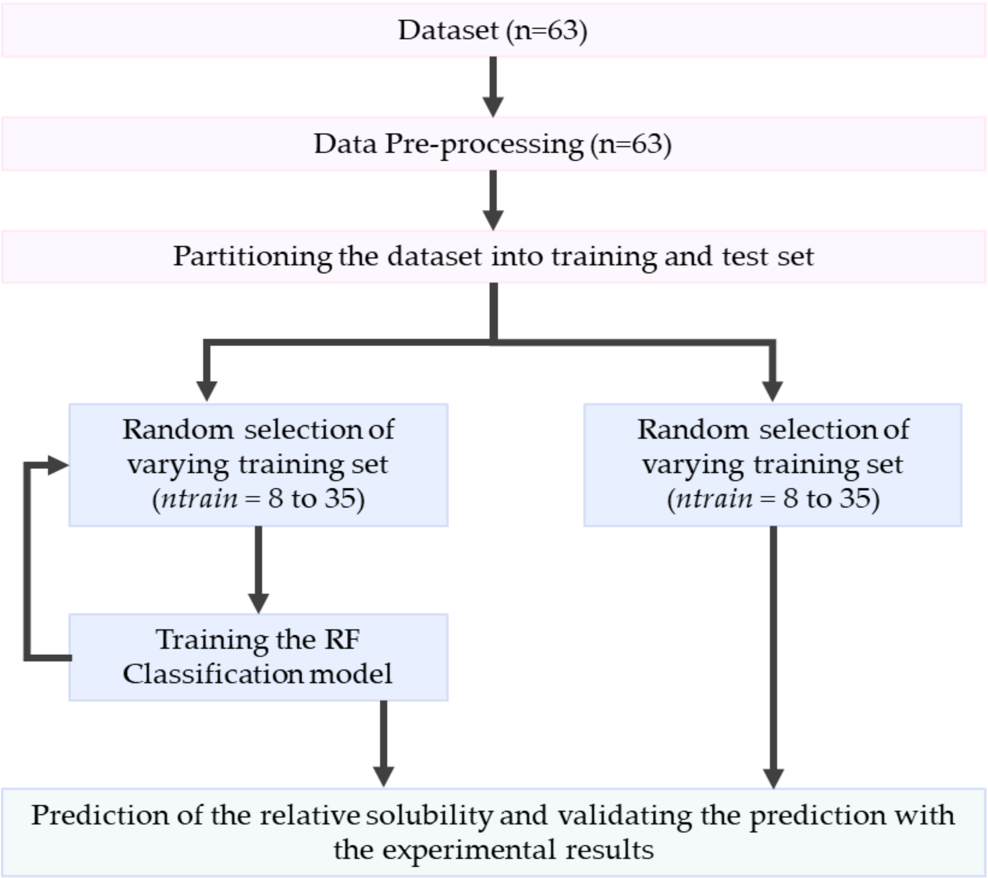

Dataset (n=63)
Data Pre-processing (n=63)
Partitioning the dataset into training and test set
Random selection of varying training set (*ntrain* = 8 to 35)
Random selection of varying training set (*ntrain* = 8 to 35)
Training the RF Classification model
Prediction of the relative solubility and validating the prediction with the experimental results

Figure 5.2 Schematic workflow of the RF model developed in order to predict the qualitative solubility of three target APIs. 10,000 resampling iterations were performed.

5.4 Results and Discussions

5.4.1 Solvent screening of paracetamol, carbamazepine and carvedilol

Following the method outlined in Section 3.2.1.2 and 5.2, solvent screening was performed on paracetamol, carbamazepine and carvedilol on 63 diverse organic solvents at two respective temperature points (low and high). Cooling crystallisation for each target compound on each solvent was repeated four times for validation and reproducibility. Table 5.2 presents the total number of solvents in which the compounds were found to be *soluble* and *practically insoluble* at two set temperature points. A list of the solvents detailing the observed experimental responses at the two set temperature points are presented in Table 5.3 for lab temperature and high temperature point respectively (see methodology 3.2.1.2). The soluble outcomes are indicated by the '*S*' while the practically insoluble outcomes are indicated by the *PI* for all the three compounds at two temperature points (Table 5.3). The qualitative data obtained from the cooling crystallisation were also visualized in a histogram where transmissivity of the target compounds in 63 solvents was plotted at two temperature points as shown in Figure 5.3, Figure 5.4 and Figure 5.5 respectively.

Table 5.2 Total number of solvents found to be soluble and practically soluble in paracetamol, carbamazepine and carvedilol at two temperature points

	Low temperature		High temperature	
Compounds	Soluble	Practically Insoluble	Soluble	Practically Insoluble
Paracetamol	24	39	33	30
Carbamazepine	15	48	36	27
Carvedilol	15	48	42	21

Table 5.3 List of solvents with their respective qualitative solubility for paracetamol (PCM), carbamazepine (CBZ) and carvedilol (CRV) at two temperature points

Solvent Name	Paracetamol		Carbamazepine		Carvedilol	
	Low Temp	High Temp	Low Temp	High Temp	Low Temp	High Temp
water	PI	S	PI	PI	PI	PI
nitromethane	PI	S	PI	S	PI	S
methanol	S	S	S	S	PI	S
formamide	S	S	PI	S	PI	S
dichloromethane	PI	PI	S	S	PI	PI
iodomethane	PI	PI	PI	PI	PI	PI
1,2-dichloroethane	PI	PI	PI	S	PI	S
acetonitrile	PI	S	PI	S	PI	S
acetone	S	S	PI	S	PI	PI
trifluoroethanol	S	S	S	S	PI	S
n-methyl-2-pyrrolidone	S	S	S	S	S	S
propanediol	S	S	S	S	PI	S
pyridine	S	S	S	S	S	S
1,4-dioxane	PI	S	PI	S	S	S
2-methoxyethanol	S	S	S	S	S	S
ethylacetate	PI	PI	PI	PI	PI	S
tetrachloroethene	PI	PI	PI	PI	S	S
2-bromobutane	PI	PI	PI	PI	PI	PI
bromobenzene	PI	PI	PI	S	PI	PI
cyclopentane	PI	PI	PI	PI	PI	PI
dibutyl ether	PI	PI	PI	PI	S	S

1-butanol	S	S	PI	S	PI	S
1-methylnapthalene	PI	PI	PI	S	PI	S
2,2,4-trimethylpentane	PI	PI	PI	PI	PI	PI
acetic Acid	S	S	PI	S	S	S
bromoform	PI	PI	S	S	PI	S
ethane thiol	PI	PI	PI	PI	PI	PI
dimethylsulphoxide	S	S	PI	PI	S	S
trifluoroacetic acid	S	S	S	S	S	S
n, n-dimethylformamide	S	S	S	S	S	S
ethanol	S	S	PI	PI	PI	S
nitrobenzene	PI	S	S	S	PI	PI
tetrahydrofuran	S	S	PI	PI	S	S
2-methoxyethylether	S	S	PI	S	S	S
4-methyl-2-pentanone	PI	S	PI	PI	PI	S
trichloroethylene	PI	PI	PI	PI	PI	PI
1-bromobutane	PI	PI	PI	PI	PI	PI
xylene	PI	PI	PI	PI	PI	S
cyclohexane	PI	PI	PI	PI	PI	PI
2-methoxy-2-methylpropane	PI	PI	PI	PI	PI	PI
triethylamine	PI	PI	PI	PI	PI	PI
hexane	PI	PI	PI	PI	PI	PI
1,2-dimethoxyethane	S	S	PI	S	S	S
1-pentanol	PI	S	PI	S	PI	S
1-propanol	S	S	PI	S	PI	S
2-butanol	S	S	PI	S	PI	S

2-butanone	S	S	PI	PI	PI	S
2-ethoxyethanol	S	S	S	S	S	S
2-methyl-1-propanol	PI	S	PI	S	PI	S
2-propane thiol	PI	PI	PI	PI	PI	PI
2-propanol	S	S	S	S	PI	S
3-methyl-1-butanol	S	S	PI	S	PI	S
anisole	PI	PI	PI	S	PI	S
benzene	PI	PI	PI	PI	PI	S
butyl acetate	PI	PI	PI	S	PI	S
carbon tetrachloride	PI	PI	PI	PI	PI	PI
chloroform	PI	PI	S	S	PI	S
diethyl ether	PI	PI	PI	PI	PI	PI
heptane	PI	PI	PI	PI	PI	PI
isobutyl acetate	PI	PI	PI	PI	PI	S
methyl acetate	PI	S	PI	S	PI	PI
n, n-dimethylacetamide	S	S	S	S	S	S
toluene	PI	PI	PI	S	PI	S

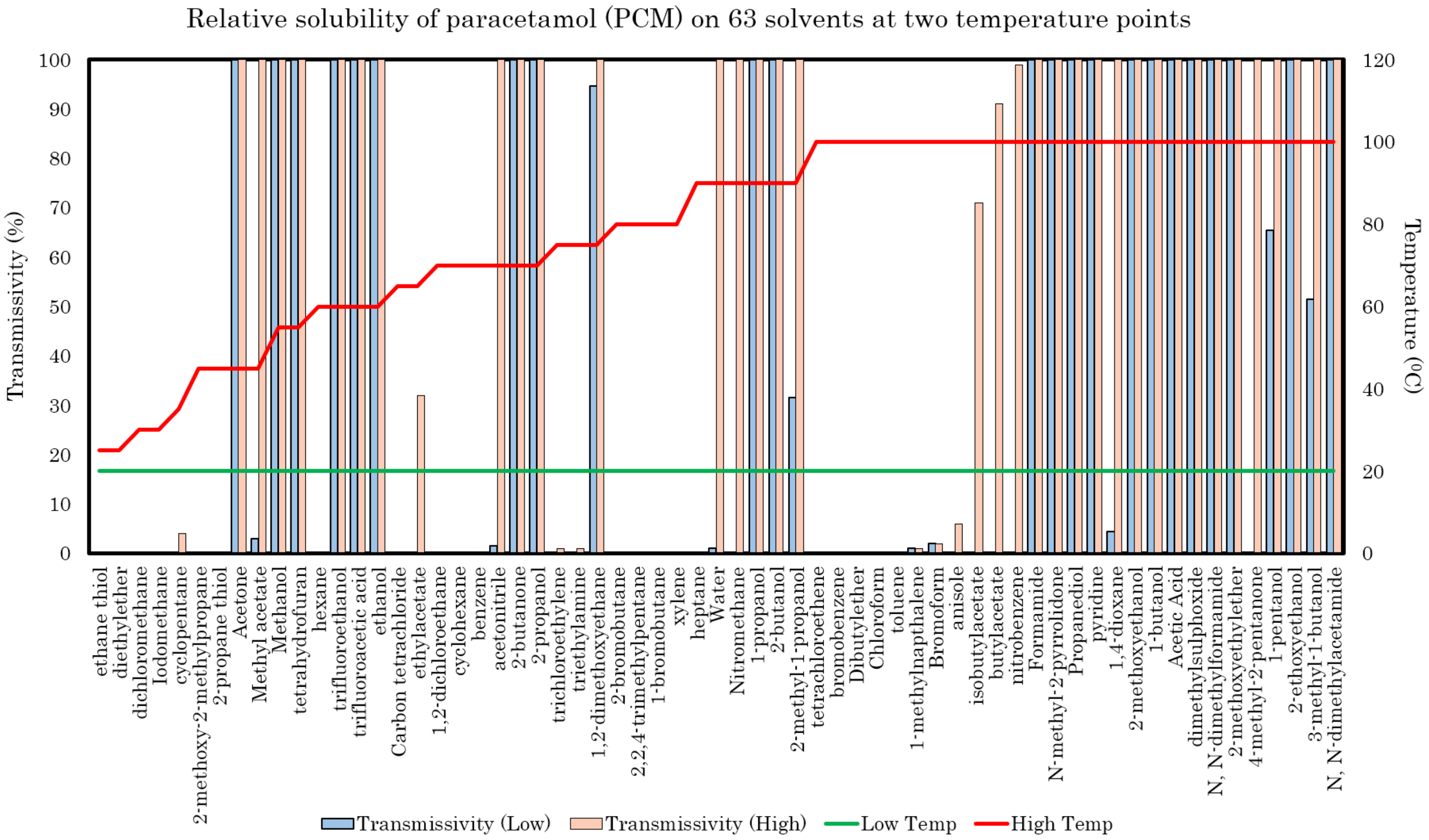


Figure 5.3 Plot of qualitative transmissivity of paracetamol (PCM) in 63 organic solvents at two temperature points

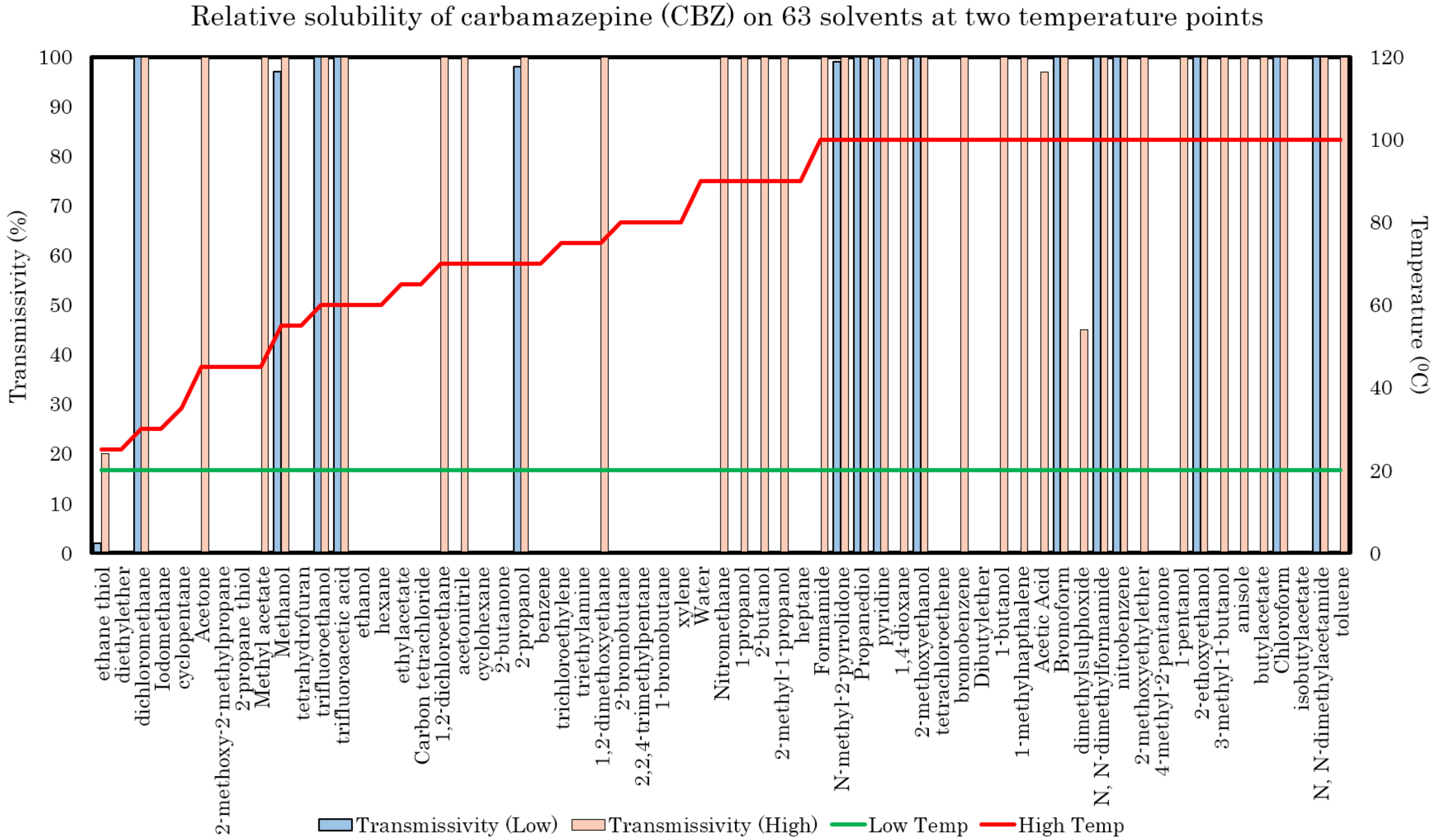


Figure 5.4 Plot of qualitative transmissivity of carbamazepine (CBZ) in 63 organic solvents at two temperature points

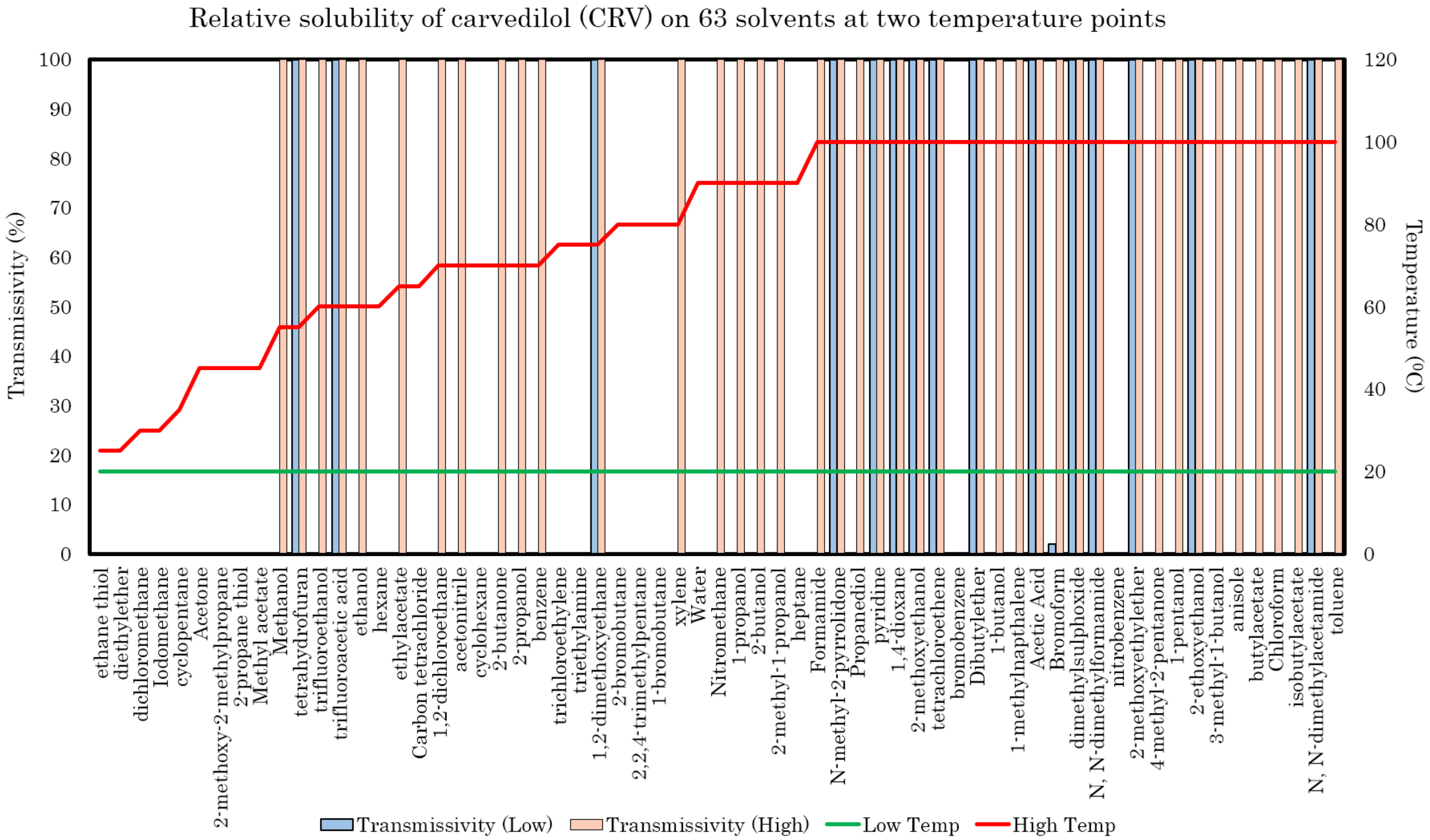


Figure 5.5 Plot of quality ative transmissivity of carvedilol (CRV) in 63 organic solvents at two temperature points

5.4.2 Deciding optimum training set for an effective RF classification model

For developing a successful predictive model with this solubility data, it was of interest to know, a) how many solubility data points were required to train an accurate model to predict the rest, b) whether the choice of solvents for the training set had a major effect on the results, and c) if this were the case, which solvents were important to include in the training set to maximise accuracy. To answer these questions, a large battery of RF models was generated, varying both training set size and composition.

While it was not computationally feasible to investigate every possible combination of training set at every possible training set size, a reasonable compromise was settled upon. 10,000 RF classification models were constructed for each compound using the following parameters: *ntree* = 1,500, *mtry* = 15 (default value), randomly selecting the training set from the pool of available data for each repeat. The size of the training set (*ntrain*) was then varied, generating 10,000 models in the same manner at each value. This procedure was performed for each compound, at both low and high temperature points. The results are summarised as a series of box plots: the three compounds at low temperatures are visualized in Figure 5.6, Figure 5.7 and Figure 5.8 while the box plots for the three compounds at high temperatures are visualized in Figure 5.9, Figure 5.10 and Figure 5.11.

The box plots provide a useful way of understanding the sensitivity of the accuracy of the RF predictions both to the solvents sampled for inclusion in the training set and the training set size. The mean and median prediction error rate (AP) for the three compounds was reasonably low at around 30% even at a small training sample size at both temperature points. It was observed from the box plots that both the mean and median error rate gradually decreases with the increasing number in training sets which indicates that the bigger the training set size, the lesser the prediction error rate. At higher temperatures, the mean prediction error rate was observed to be even lower than compared to low temperature for paracetamol as it tends to dissolve in most solvents when the temperature is increased. However, for carvedilol and carbamazepine, the error rates were observed to be slightly higher than those at low temperatures. This could be due to the ability of these compounds to form various polymorphic forms and solvates in different solvents and thus the transformation could have happened during the initial stage where the compound was stirred in the solvent for an hour at lab temperature or when the temperature was increased. This has an adverse effect on the solubility of the compound. As the models do not take these polymorphic form changes into consideration, this could have contributed to the high error rates at high temperatures.

From the series of box plots shown in Figure 5.6 to Figure 5.11, the conclusion drawn was that even by randomly selecting the solvents without any prior knowledge on its suitability for the model, there was a significant chance of constructing a RF classification model with an estimated error rate below 30% for binary classes at training sample size as small as 8 solvents for all the three compounds. As the training set size increases, it can be noticed that the size of the box plots was relatively shorter as well as their whiskers. The reduction in the whiskers indicates the decrease in the variability of the prediction error rate. The reduction in the box plot size as *ntrain* increases indicates that the data points are more clustered around the median value. This is a good indication as with smaller box sizes the majority of the data points indicate a lower prediction error rate. For paracetamol at low temperature point in Figure 5.6, it can be observed that the median of the prediction error rate is consistent after *ntrain* =16. However, box sizes are the smallest between *ntrain* = 21 to 28. This indicates that the majority of the data points after *ntrain*=21 were in agreement with the median and mean prediction error rate below 25%. This also indicates that the chances of getting a lower prediction error rate were higher when randomly selecting 21 solvents in the training set than compared to *ntrain* less than 21. For carbamazepine at low temperature point, the smallest box sizes are observed at *ntrain* = 13 however the median value was consistent until after *ntrain* = 28 where it reduced to 0.2 (Figure 5.7). Similarly, for carvedilol at low temperature, the low median value was observed to be consistent after *ntrain* = 20 (Figure 5.8) while the trend of a shorter condensed box plot was observed at *ntrain* = 20. Similarly, at high temperature points, for paracetamol and carvedilol, the condensed small box plots were observed at *ntrain*=20 and 23 respectively while the median value was consistent after *ntrain* = 20 for paracetamol and *ntrain*=17 for carvedilol. Similarly, for carbamazepine at high temperature point, small condensed box plots were observed a bit later at *ntrain*=25 and the median value was consistent at *ntrain*=20. A trend can be observed between all these six box plots (Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9, Figure 5.10 and Figure 5.11) that if solvents were to be randomly selected to build a classification model, the smallest *ntrain* size with a relatively low prediction error rate (i.e. for all the three compounds at two temperature points, the box plots were the most condensed around the region indicating that the majority of the data points with low prediction error rate lied close to the median and mean value) would be estimated between 20 to 24. Thus, this estimate was chosen as the optimum number of training set sizes for building future RF classification models.

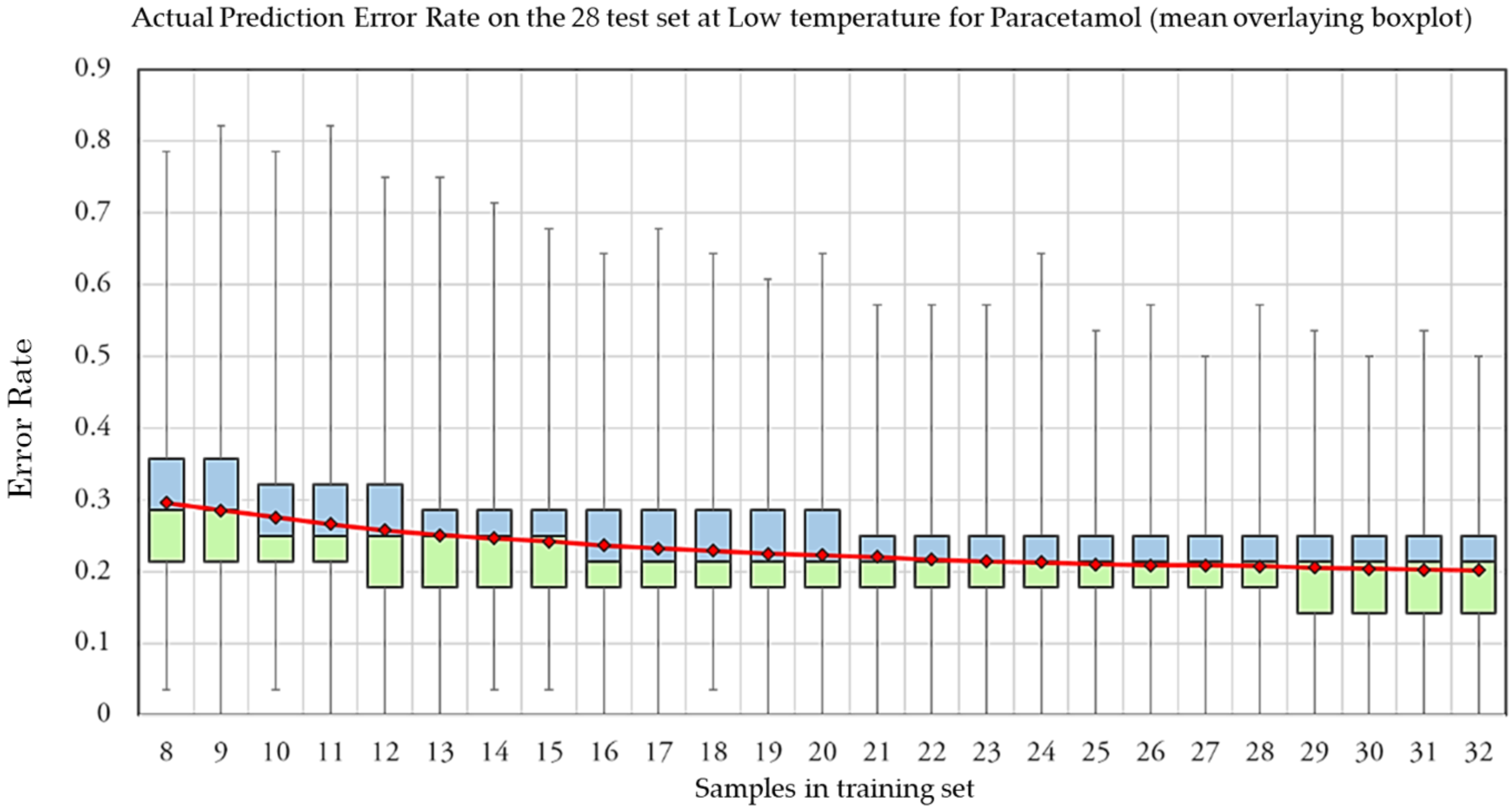


Figure 5.6 10,000 RF classification models trained for paracetamol at low temperature and at varying training set sizes.

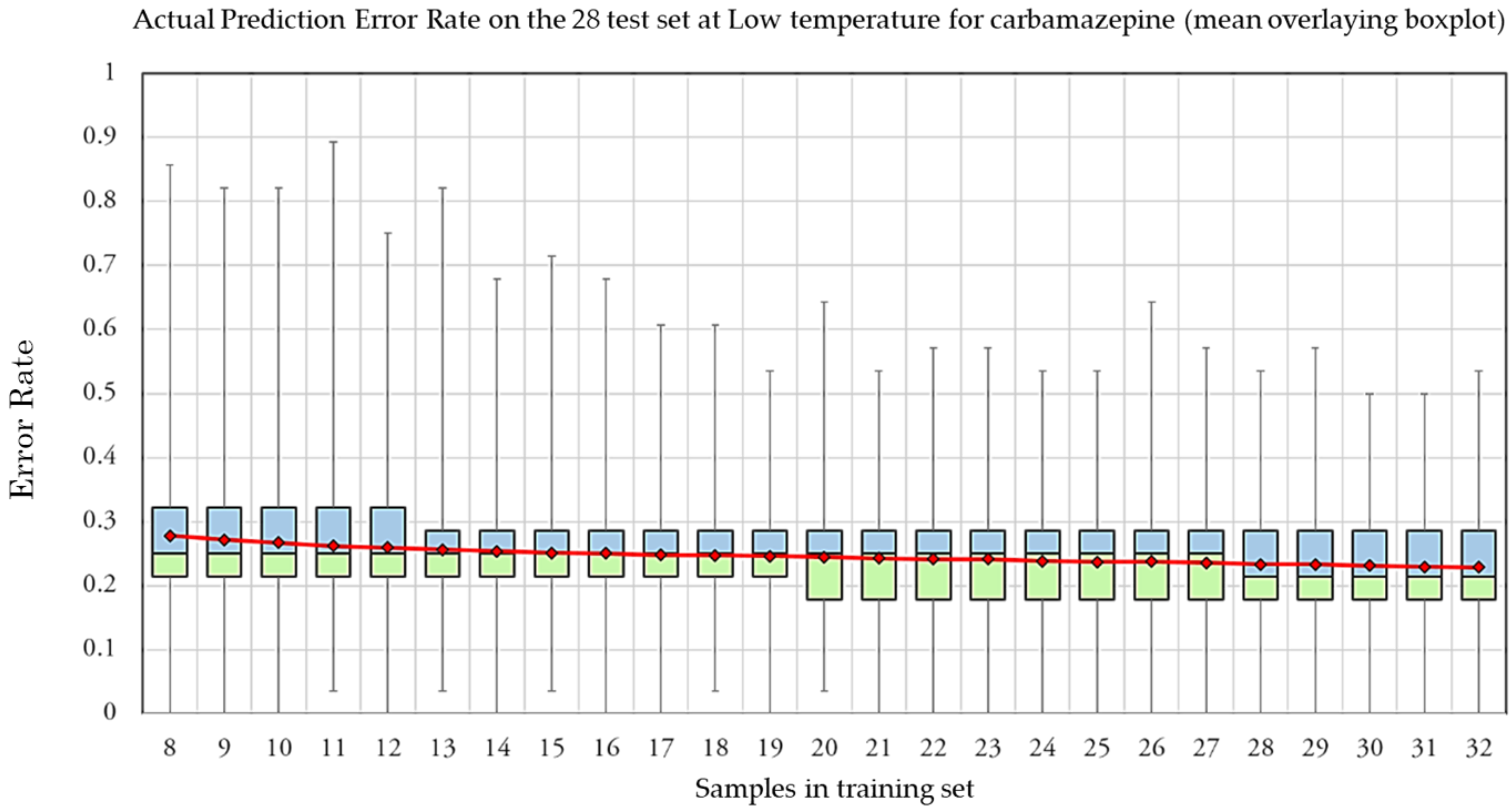


Figure 5.7 10,000 RF classification models trained for carbamazepine at low temperature and at varying training set sizes

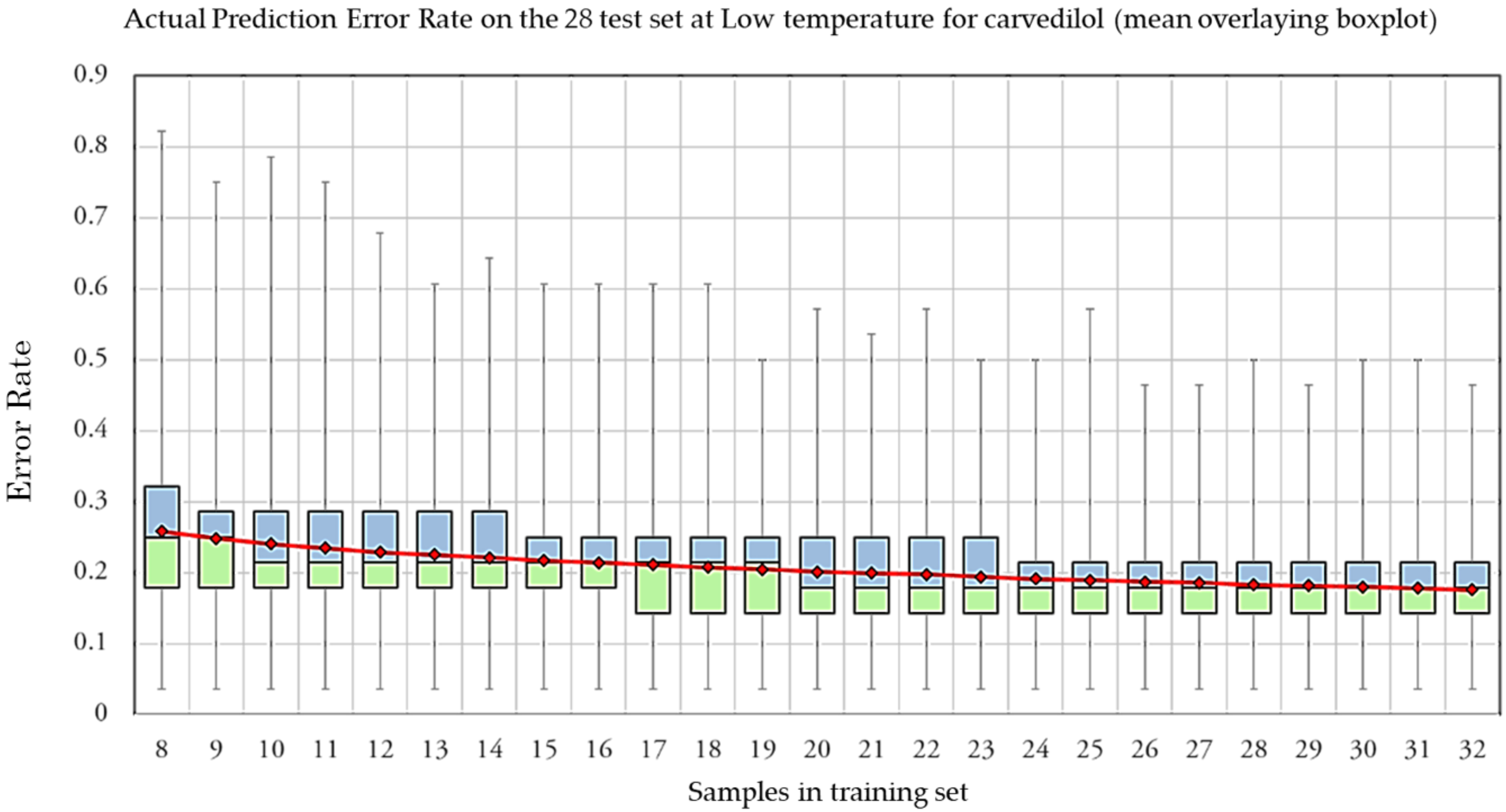


Figure 5.8 10,000 RF classification models trained for carvedilol at low temperature and at varying training set sizes

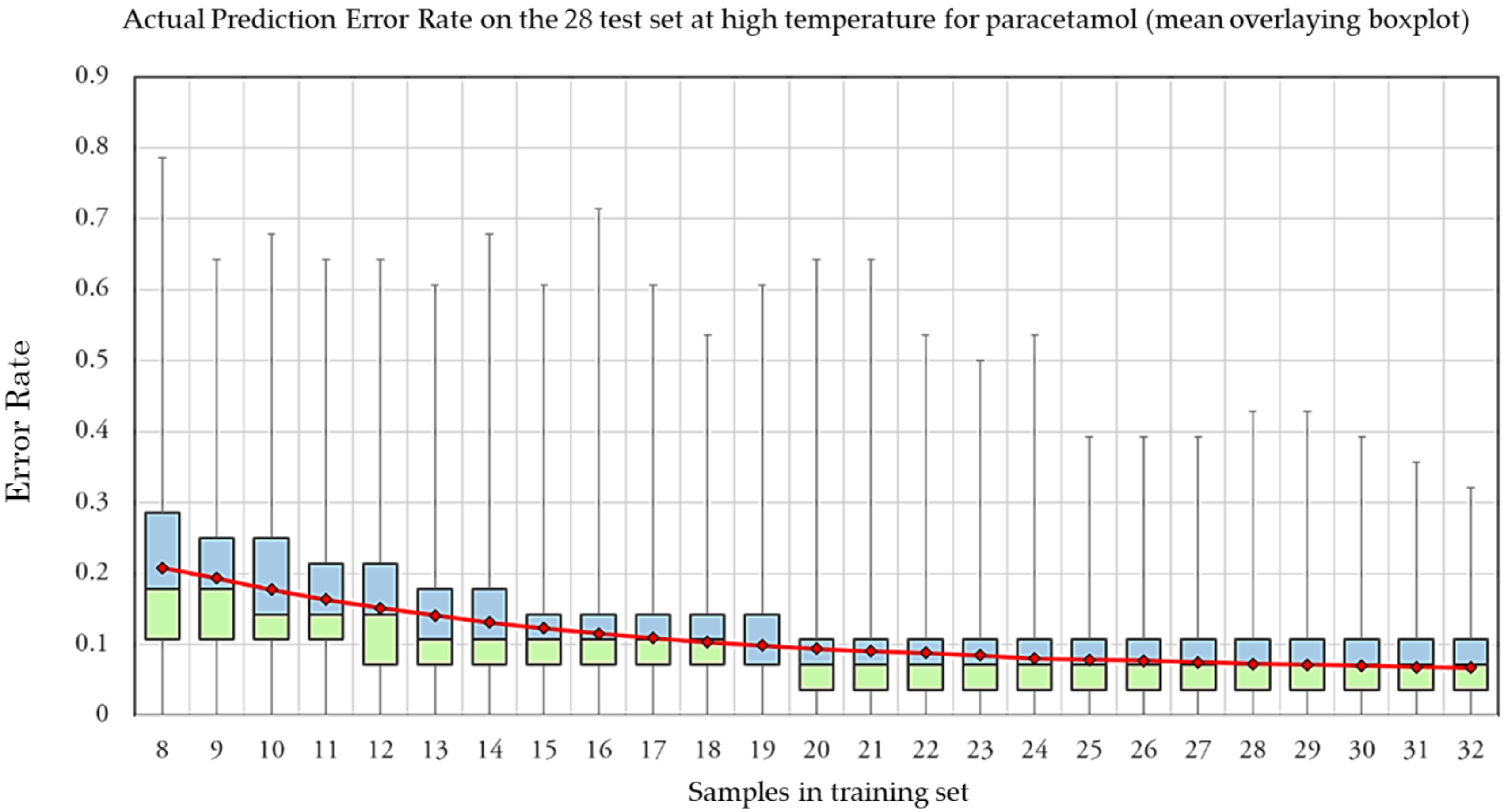


Figure 5.9 10,000 RF classification models trained for paracetamol at high temperature and at varying training set sizes

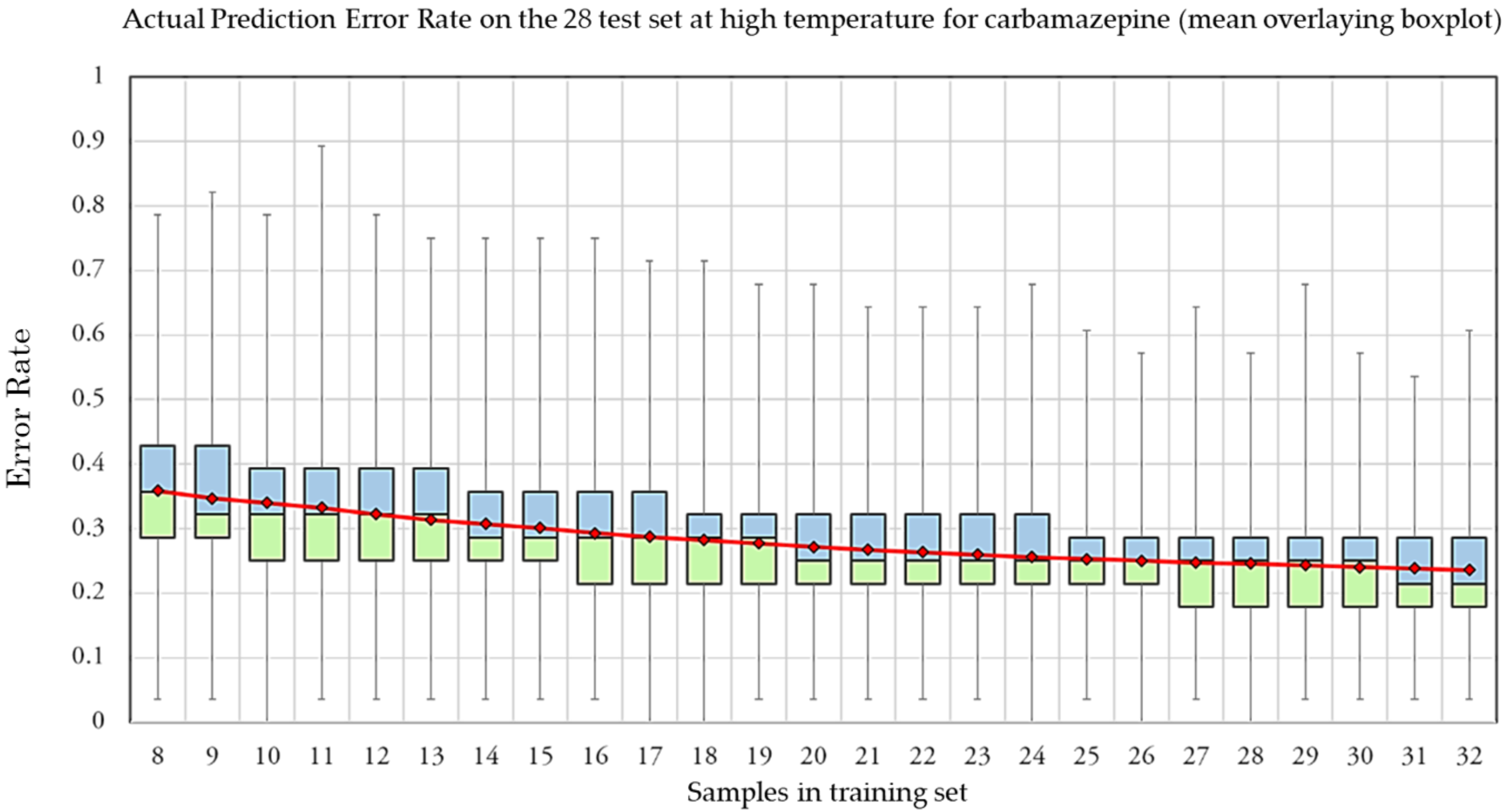


Figure 5.10 10,000 RF classification models trained for carbamazepine at high temperature and at varying training set sizes

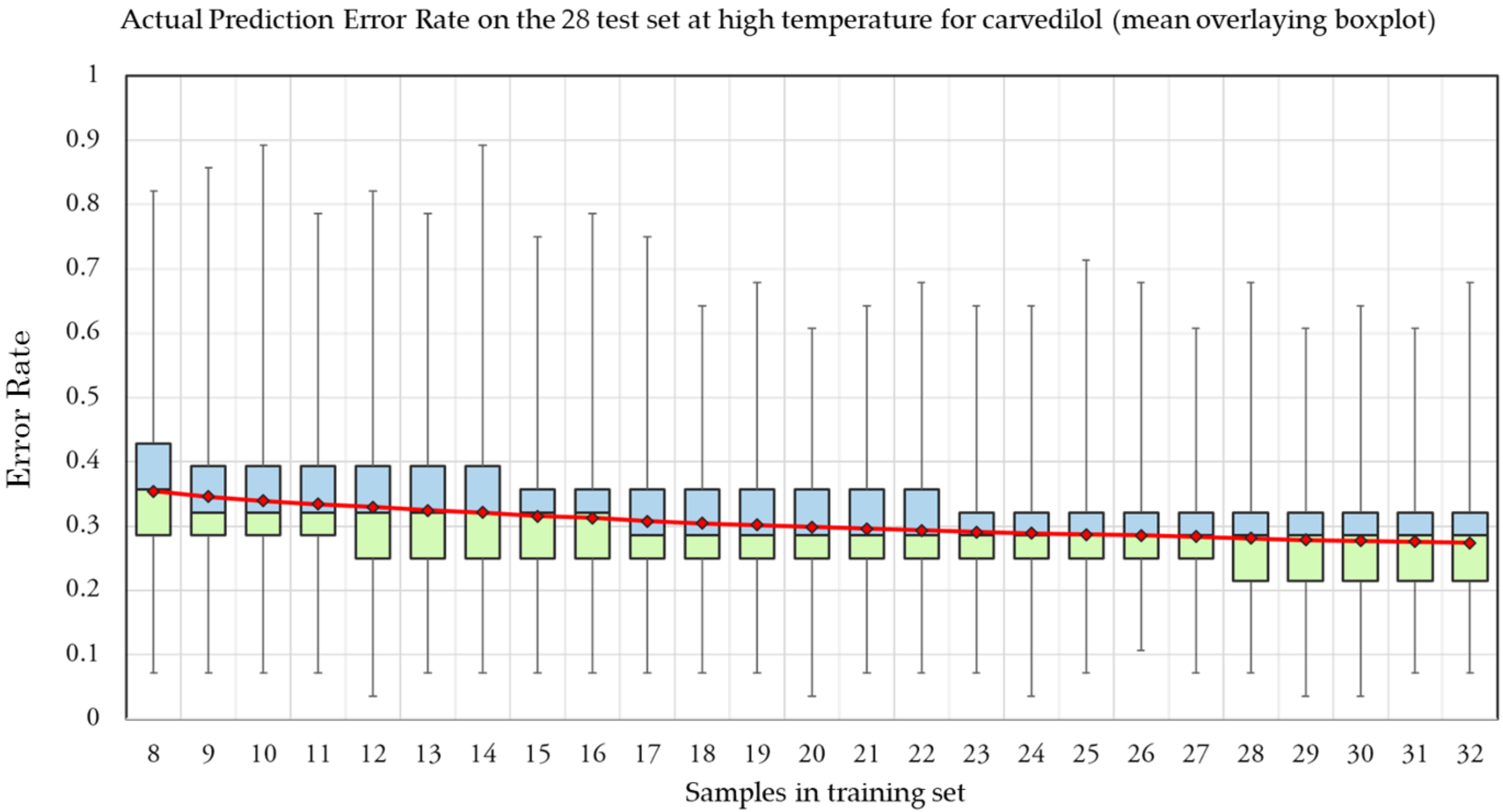


Figure 5.11 10,000 RF classification models trained for carvedilol at high temperature and at varying training set sizes

5.4.3 A rational way of selecting the solvents

As observed from the analysis of the box plots in Section 5.4.2, the optimum value of the training set size was estimated to be between 20 to 24 solvents for this study. However, the figures between Figure 5.6 to Figure 5.11 also demonstrated that there is a wide variance in model accuracy for any given training set size which was dependent on the choice of the solvents on the training set. The question arises as to which 20 to 24 solvents can be chosen from a database of 63 solvents for consistently attaining a low prediction error rate. Thus, there was a need for a rational method for selecting solvents in the training set. Two methods were applied in this chapter for rational selection of solvents in the database follows.

5.4.3.1 *Method I*: Extracting the most frequent solvents observed in high-performing models

The aim here was to identify the list of solvents that, when selected in the training set, contributed to a relatively low error rate of prediction from the trained 10,000 RF models. For each model in the 4th quartile region from the preceding analysis (i.e. the lowest error rates), the list of solvents forming its training set was extracted. This was performed for all training set sizes (8 to 32), compounds and temperatures. The next step was to calculate the expected frequency of each solvent being randomly sampled at each training set size using Equation 5.1.

The expected frequency is calculated using the probability theory of random selection and is given by (Corcoran, 2006).

$$f_e = \frac{n_t}{T_s} \qquad \text{Equation 5.1}$$

where, f_e is the expected frequency, n_t is the number of solvents in the training set size and T_s is the total number of solvents used for the model

The actual (observed) frequency was then obtained by counting the number of times each solvent actually appeared in the extracted lists for each training set size. The difference between the observed expected frequencies gave the deviation from the expected frequency of selection for each solvent. Appearing more often than chance could indicate that the solvent was important for a high-performing predictive model.

Since the optimum number for the training set size to build an effective RF model was estimated between *ntrain*=20 to 24, the most frequent solvents for these training set sizes were determined for each compound at the two temperature points. Figure 5.12 and Figure 5.13 presented the plot of the deviation from the expected frequency of the solvents for paracetamol at the training set size 20 and 24 respectively for low temperature. Similarly, Figure 5.14 and 5.15 present the plot relative change in the frequency of the solvents for paracetamol at the training set size 20 and 24 respectively for high temperature. The higher the positive deviation of the solvent in each diagram, the more important its role was in the training set and thus more accurate the predictive capability of the trained RF model.

The process of mining the most frequently attained solvents was repeated for both carbamazepine and carvedilol between training set sizes 20 to 24 at two temperature points. The results from the extraction of the solvents for all the three compounds between training set sizes 20 to 24 at two temperature points were analysed and the top ten frequent solvents in order of their frequency from each training sets were taken which are summarized in Table 5.4, Table 5.5, Table 5.6, Table 5.7, Table 5.8 and Table 5.9 respectively. The frequency of solvents where the difference between the observed and expected frequencies was calculated to near 0% or less than 0% (negative) was omitted from the selection process. This is because the low or negative value reflects the impact of the solvent as being low or irrelevant to the lower predicted error rate.

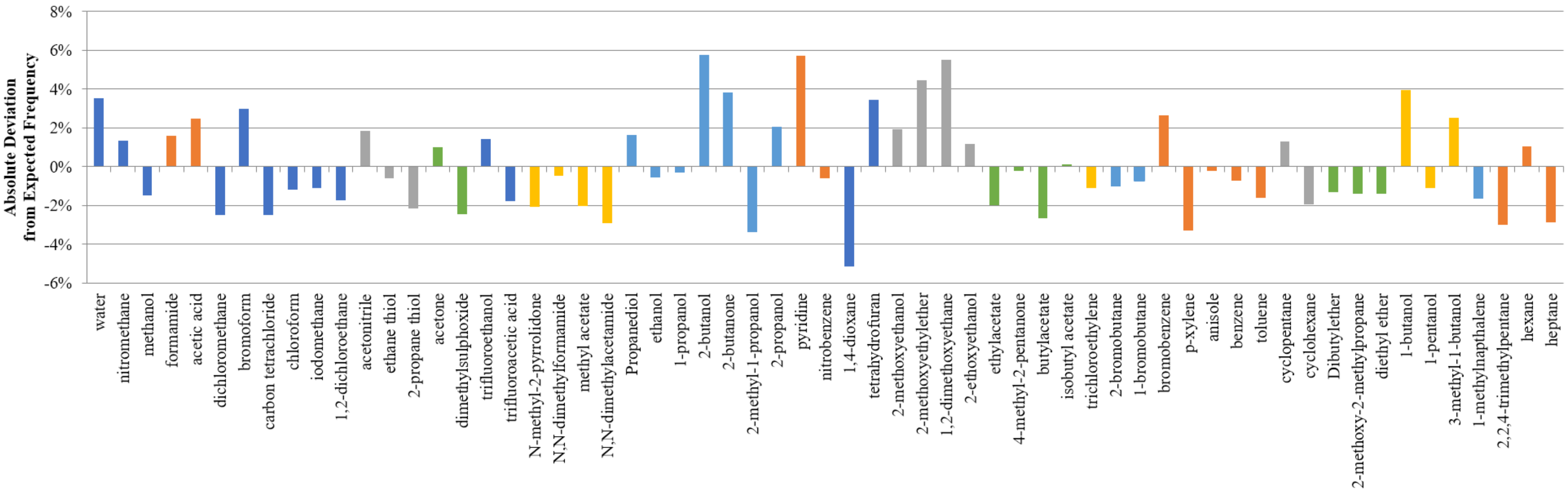


Figure 5.12 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 20 for paracetamol at low-temperature point.

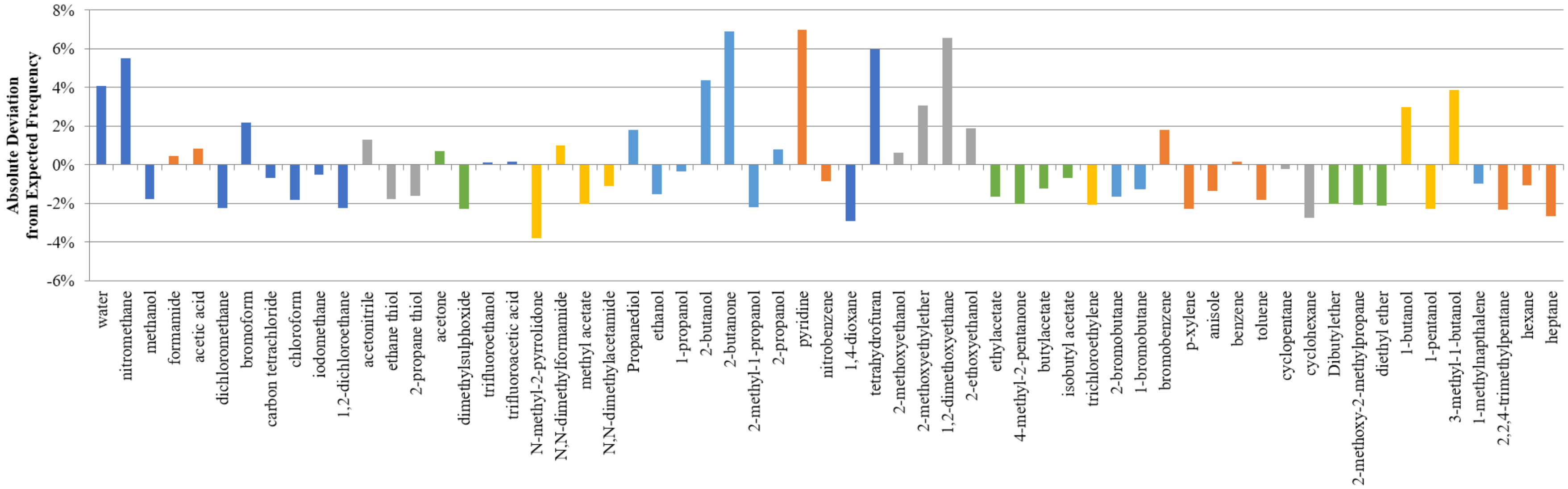


Figure 5.13 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 24 for paracetamol at low-temperature point.

Figure 5.14 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 20 for paracetamol at high-temperature point

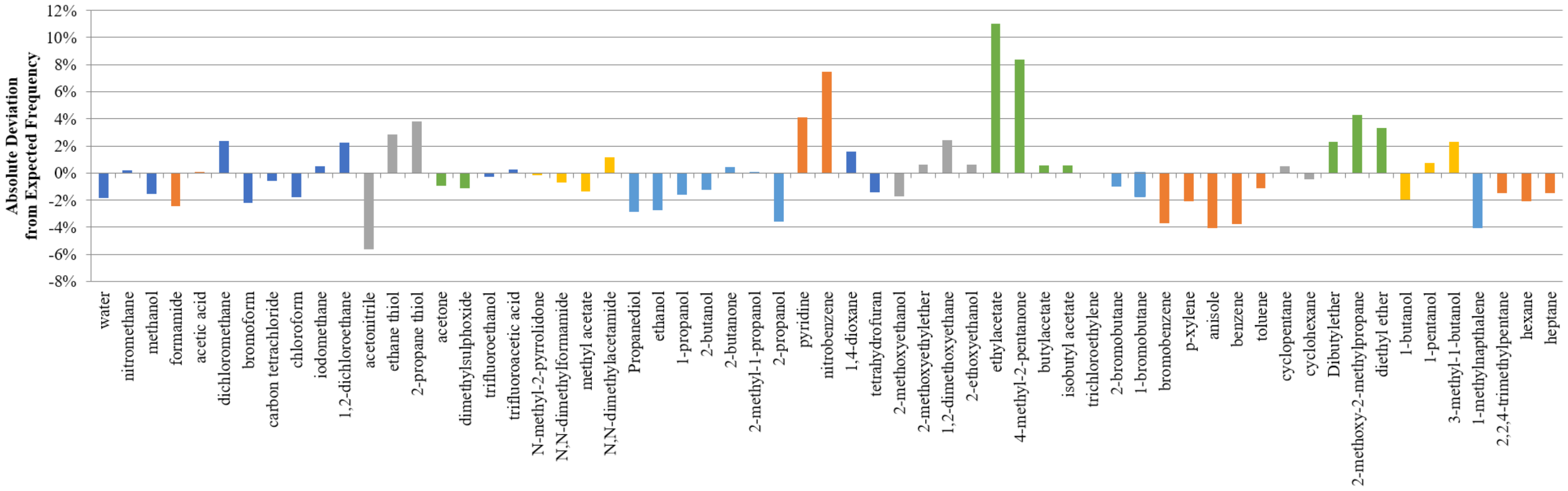


Figure 5.15 Graphical plot showing the absolute deviation from expected frequency of the solvents in training set size 24 for paracetamol at high-temperature point

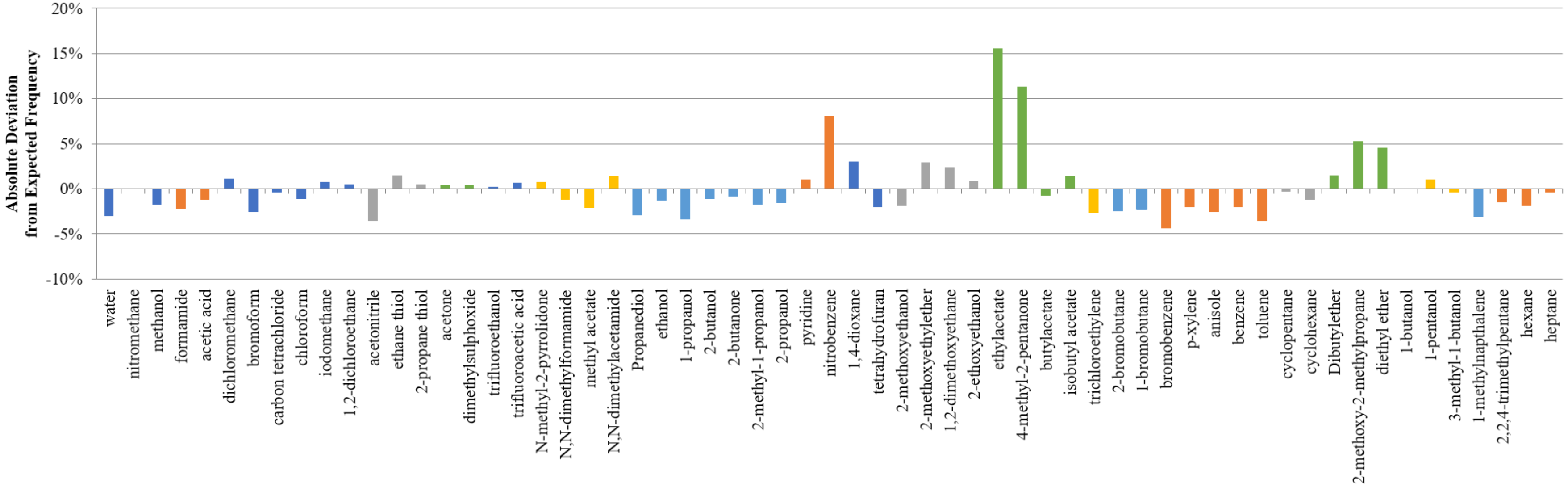


Table 5.4 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in paracetamol at low temperature points

ntrain = 20	ntrain = 21	ntrain = 22	ntrain = 23	ntrain = 24
2-butanol	2-butanone	2-butanol	tetrahydrofuran	pyridine
pyridine	2-butanol	2-butanone	1,2-dimethoxyethane	2-butanone
1, 2- dichloroethane	pyridine	1,2-dimethoxyethane	pyridine	1,2-dimethoxyethane
diglyme	tetrahydrofuran	tetrahydrofuran	2-butanone	tetrahydrofuran
1-butanol	water	pyridine	water	nitromethane
2-butanone	1,2-dimethoxyethane	water	2-butanol	2-butanol
water	isoamyl alcohol	diglyme	isoamyl alcohol	water
tetrahydrofuran	diglyme	isoamyl alcohol	diglyme	isoamyl alcohol
bromoform	1-butanol	nitromethane	bromoform	diglyme
bromobenzene	nitromethane	acetic acid	bromobenzene	1-butanol

Table 5.5 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in paracetamol at high temperature points

ntrain = 20	ntrain = 21	ntrain = 22	ntrain = 23	ntrain = 24
ethyl acetate	ethylacetate	ethylacetate	ethylacetate	ethylacetate
methyl isobutyl ketone	methyl isobutyl ketone	nitrobenzene	nitrobenzene	methyl isobutyl ketone
nitrobenzene	nitrobenzene	methyl isobutyl ketone	methyl isobutyl ketone	nitrobenzene
methyl isobutyl ketone	methyl isobutyl ketone	diethyl ether	diethyl ether	methyl isobutyl ketone
pyridine	diethyl ether	methyl isobutyl ketone	dibutyl ether	diethyl ether
2-propanethiol	1,4-dioxane	1,2-dimethoxyethane	1,2-dimethoxyethane	1,4-dioxane
diethyl ether	dibutyl ether	ethanethiol	methyl isobutyl ketone	diglyme
ethane thiol	diglyme	1,4-dioxane	methyl isobutyl ketone	1,2-dimethoxyethane
1,2-dimethoxyethane	2-propanethiol	dibutyl ether	diglyme	dibutyl ether
dichloromethane	methyl isobutyl ketone	2-propane thiol	1, 2- dichloroethane	ethane thiol

Table 5.6 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carbamazepine at low temperature points

ntrain = 20	ntrain = 21	ntrain = 22
ethanethiol	Ethanethiol	ethanethiol
n, n-dimethylacetamide	n, n-dimethylacetamide	n, n-dimethylacetamide
ethanol	Bromobenzene	2-butanol
Bromobenzene	ethanol	ethanol
methyl tert-butyl ether	methyl tert-butyl ether	Bromobenzene
heptane	acetone	propanediol
1-propanol	dibutyl ether	dibutyl ether
acetone	propanediol	methyl tert-butyl ether
dibutyl ether	2-butanol	n-methyl-2-pyrrolidone
propanediol	heptane	1-propanol
ntrain = 23	**ntrain = 24**	
ethanethiol	ethanethiol	
n, n-dimethylacetamide	ethanol	
ethanol	n, n-dimethylacetamide	
methyl tert-butyl ether	bromobenzene	
bromobenzene	acetone	
2-butanol	2-butanol	
Heptane	heptane	
dibutyl ether	dibutyl ether	
acetone	propanediol	
1-butanol	methyl tert-butyl ether	

Table 5.7 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carbamazepine at high temperature points

ntrain = 20	ntrain = 21	ntrain = 22
formamide	formamide	formamide
ethylacetate	tetrahydrofuran	2-butanone
tetrahydrofuran	2-propanethiol	tetrahydrofuran
heptane	2-butanone	ethylacetate
2-butanone	heptane	heptane
2-propanethiol	ethyl acetate	iodomethane
methyl tert-butyl ether	toluene	1-methylnapthalene
ethanethiol	1, 2- dichloroethane	ethanol
1-methylnapthalene	trichloroethylene	toluene
iodomethane	cyclohexane	ethanethiol
ntrain = 23	ntrain = 24	
tetrahydrofuran	2-butanone	
formamide	formamide	
ethanol	tetrahydrofuran	
2-butanone	heptane	
heptane	Iodomethane	
cyclohexane	ethanol	
2-propanethiol	toluene	
ethyl acetate	1, 2- dichloroethane	
cyclopentane	2-propanethiol	
methyl isobutyl ketone	1-methylnapthalene	

Table 5.8 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carvedilol at low-temperature points

ntrain = 20	ntrain = 21	ntrain = 22
tetrahydrofuran	tetrahydrofuran	tetrahydrofuran
formamide	ethanethiol	acetic acid
acetone	formamide	formamide
ethanethiol	acetic acid	acetone
acetic acid	2-ethoxyethanol	ethanethiol
2-ethoxyethanol	acetone	2-ethoxyethanol
2-methoxyethanol	1,2-dimethoxyethane	acetonitrile
1,2-dimethoxyethane	acetonitrile	1,4-dioxane
1,4-dioxane	2-methoxyethanol	1,2-dimethoxyethane
acetonitrile	isobutyl acetate	benzene
ntrain = 23	ntrain = 24	
tetrahydrofuran	tetrahydrofuran	
formamide	formamide	
ethanethiol	ethanethiol	
isobutyl acetate	acetic acid	
acetic acid	isobutyl acetate	
acetone	pyridine	
acetonitrile	1,2-dimethoxyethane	
2-ethoxyethanol	acetone	
1,2-dimethoxyethane	2-ethoxyethanol	
1,4-dioxane	2-methoxyethanol	

Table 5.9 Top ten most frequently observed solvents in the training set sizes between 20 and 24 which contributed to low error rate in carvedilol at high temperature points

ntrain = 20	ntrain = 21	ntrain = 22
ethyl acetate	ethyl acetate	ethyl acetate
methyl tert-butyl ether	methyl tert-butyl ether	methyl tert-butyl ether
heptane	2-propanethiol	2-propanethiol
2-propanethiol	heptane	heptane
toluene	toluene	toluene
m-xylene	carbon Tetrachloride	water
methanol	1-methylnapthalene	carbon Tetrachloride
Carbon Tetrachloride	1-bromobutane	iso-octane
1-methylnapthalene	methyl acetate	2-bromobutane
water	pyridine	m-xylene
ntrain = 23	ntrain = 24	
ethyl acetate	ethyl acetate	
methyl tert-butyl ether	methyl tert-butyl ether	
2-propanethiol	2-propanethiol	
heptane	heptane	
toluene	toluene	
carbon Tetrachloride	water	
water	chloroform	
1-methylnapthalene	carbon Tetrachloride	
ethanol	m-xylene	
iso-octane	trichloroethylene	

It can be observed from all the summarised tables (Table 5.4, Table 5.5, Table 5.6, Table 5.7, Table 5.8 and Table 5.9), that there is a recurring trend among the solvents' frequencies in the varying training set sizes for the three compounds. The most frequently appearing classes of solvents for all the three compounds in these selected training set sizes were found to be ethers and alcohols. This suggests that these classes of solvents were most important for producing a high-performing RF model. 52 solvents appeared in the top ten list of solvents for all the three target compounds within training set sizes 20 to 24. The solvents appearing with the frequency of 5 or less were removed as their presence in the training set would not have a large impact on the prediction accuracy compared to solvents with higher frequency. This resulted in a final list of 18 solvents recommended for inclusion in a training set in order to build a RF model with high prediction accuracy for the three target compounds; these are shown in Table 5.10.

Table 5.10 Final list of the recommended solvents for training the RF model on the three target compounds based on their frequency of observation

SN	Solvent	SN	Solvent
1	1,4-dioxane	10	ethanol
2	1-methylnapthalene	11	ethyl acetate
3	2-butanol	12	formamide
4	2-butanone	13	heptane
5	2-propane thiol	14	isoamyl alcohol
6	acetone	15	2-methoxy-2-methylpropane
7	dibutyl ether	16	tetrahydrofuran
8	2-methoxyethyl ether	17	toluene
9	ethanethiol	18	water

5.4.3.2 *Method II*: Selecting solvents using a cluster map (Strathclyde24)

Analysing and mining the data for finding a trend in the frequency of solvents in the training set for the 10,000 RF models was found to be computationally demanding and time consuming. The list of frequently observed solvents was also found to vary with the changing temperature points and the choice of a compound. This makes it challenging to recommend solvents for a new compound. Thus, an alternative way of recommending suitable solvents for building a RF model with high prediction accuracy was performed using a cluster map. Various reports of solvent clustering based on calculated molecular descriptors can be found in the literature (Chastrette, Rajzmann, Chanon, & Purcell, 1985; Cleophas, 2012; Gu, Li, Gandhi, & Raghavan, 2004; D. Xu & Redman-Furey, 2007). Most of these cluster maps share a common theme i.e. an unsupervised learning algorithm combined with a clustering algorithm to form clusters. A novel in-house solvent clustering method was attempted where the solvents were clustered based on 250 calculated physicochemical molecular descriptors and clustering was performed using the ClusterSim package in R and multidimensional scaling. In this approach, 94 solvents were clustered into 24 individual clusters based on their similarity in thermodynamic, electronic, topological, spatial and feature-count molecular descriptors and distance criterion. The cluster map is known as Strathclyde24 (Johnston et al., 2017) and is visualised in Figure 5.16. For solvent screening, one solvent from each individual cluster can be selected for the training dataset thus resulting in the value of *ntrain* = 24. The list of solvents randomly selected using the Strathclyde24 cluster map with the criteria of at least one per solvent cluster is shown in Table 5.11.

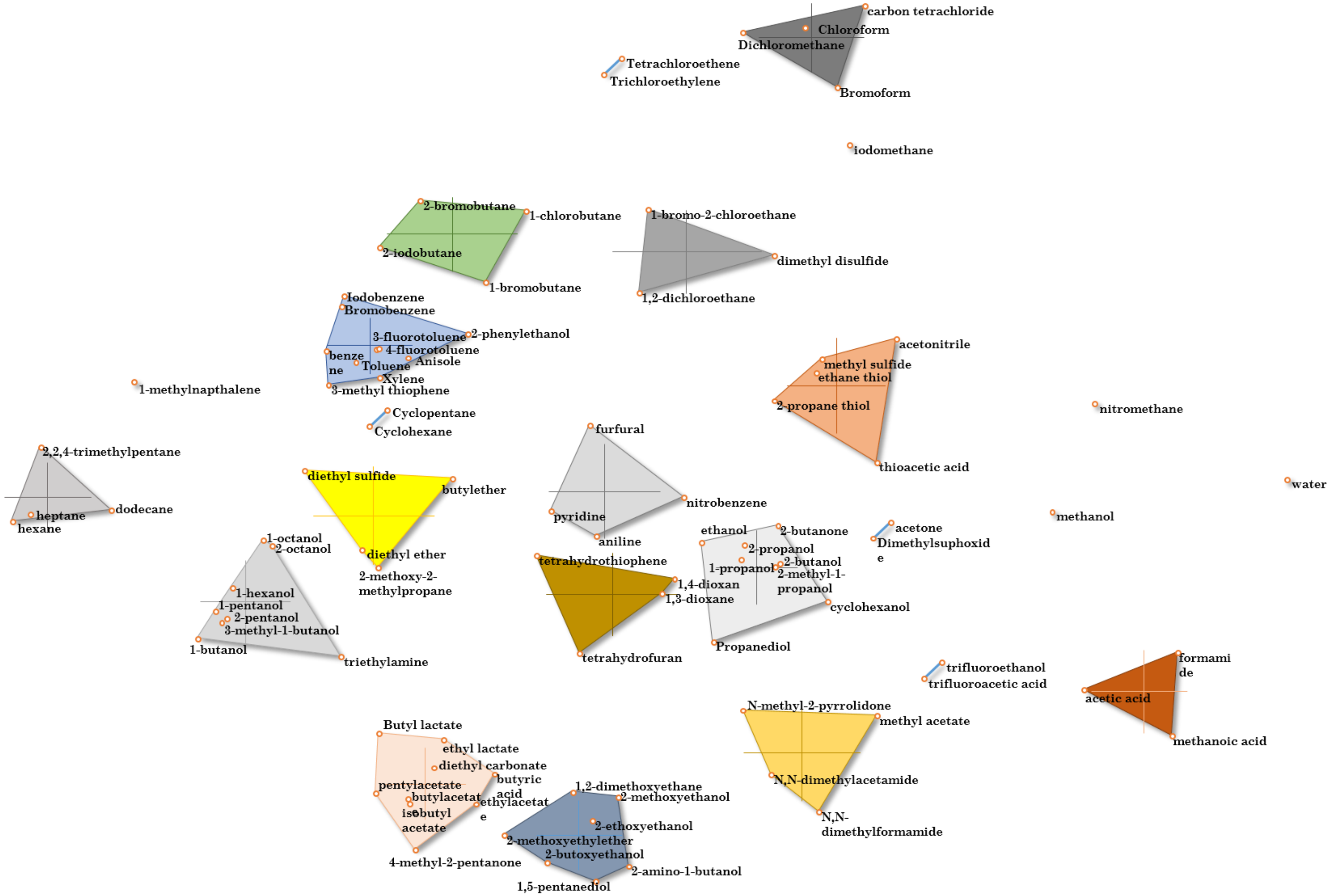


Figure 5.16 Visualizations of Strathclyde24 illustrating 24 individual clusters. Each cluster is an MDS plot illustrating the similarity and dissimilarity within and between the clusters of solvents (Johnston et al., 2017).

Table 5.11 Selected list of solvents using Strathclyde24. Selection criteria were that at least one solvent must be included from each solvent cluster

Clusters	Solvent Selected	Clusters	Solvent Selected
1	Water	13	pyridine
2	Nitromethane	14	1,4-dioxane
3	Methanol	15	2-methoxyethanol
4	Formamide	16	Ethyl acetate
5	dichloromethane	17	Tetrachloroethene
6	Iodomethane	18	2-bromobutane
7	1,2-dichloroethane	19	Bromobenzene
8	Acetonitrile	20	Cyclopentane
9	Acetone	21	Dibutylether
10	224-trifluoroethanol	22	1-butanol
11	N-methyl-2-pyrrolidone	23	1-methylnapthalene
12	1,2 propanediol	24	2,2,4-trimethylpentane

5.4.4 RF Classification model and performance comparison

The random forest models for all the three compounds were trained using all the 250 calculated molecular descriptors with the recommended solvents obtained from *method I* (Table 5.10) and *method II* (Table 5.11) as training dataset and remaining solvents as the test set with two qualitative solubility outcomes. The trained RF model was then used to predict the qualitative solubility in solvents in the test dataset and thus obtained predicted responses were validated using the experimental dataset. The predictive performance of the trained binary classification RF models was then evaluated for models obtained from both the methods. From the confusion matrix obtained from the classification models, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) were extracted and used to measure sensitivity (Sn), specificity (Sp), classification accuracy (ACC) and Cohen's kappa (κ) parameters which help analyses the performance of the classification model. Sn is the ability of the classification model to correctly identify the solvents that the target compounds are *soluble* in while Sp is the measurement of the model to correctly recognize the solvents that the target compounds are *practically insoluble* in. For a classification model, high values of Sn and Sp are desirable but usually there is a tradeoff. The predictive performance of the classification model trained using both methods **I** and **II** are presented in Table 5.12 and Table 5.13 respectively.

Table 5.12 Performance of the RF classification model trained using solvents selected from *method I*.

Low Temperature						
	κ	Sn (%)	Sp (%)	ACC (%)	OOB (%)	Mean accuracy at ntrain =18
Paracetamol	0.633	79.31	87.5	82.2	61.11	78 %
Carbamazepine	invalid	invalid	invalid	invalid	invalid	75 %
Carvedilol	0.362	27.27	100	82.22	83.33	80%
High Temperature						
	κ	Sn (%)	Sp (%)	ACC (%)	OOB (%)	Mean accuracy at ntrain =18
Paracetamol	0.911	91.30435	100	95.556	72.22	90%
Carbamazepine	0.419	67.9	76.5	71.11	55.56	70%
Carvedilol	0.5	83.33	66.7	77.8	61.11	70%

Table 5.13 Performance of the RF classification model trained using solvents selected from *method II*.

Low Temperature						
	κ	Sn (%)	Sp (%)	ACC (%)	OOB (%)	Mean accuracy at ntrain =24
Paracetamol	0.67	80	87.5	84.6	66.67	80 %
Carbamazepine	0.614	50	100	89.74	62.5	75 %
Carvedilol	0.658	55.56	100	89.74	66.67	80 %
High Temperature						
	κ	Sn (%)	Sp (%)	ACC (%)	OOB (%)	Mean accuracy at ntrain =24
Paracetamol	0.796	80	100	89.74	95.83	92 %
Carbamazepine	0.43	90	47.37	71.8	79.17	75%
Carvedilol	0.25	84.62	38.47	69.23	66.67	70%

It can be seen from both the tables (Table 5.12 and table 5.13) that the RF model trained for paracetamol had the best predictive performance among the three compounds with ACC value of 82.2 % and 95.56% at low and high temperatures respectively using solvents selected from *method I* and 84.6% and 89.74% at both temperature points using solvents selected from *method II*. Carbamazepine and carvedilol had ACC values of above 65% from both methods. Even though the overall ACC value obtained from both methods is above 65%, using only ACC as an appropriate metric for performance measure is not recommended as it can be misleading especially when there is an imbalance of the classes in the dataset. It can be observed from the table that the OOB accuracy does not correspond to the ACC accuracy. In *method I*, the OOB accuracy is underestimated than the value of ACC in all the three target compounds at both temperature points. Even though the OOB error rate is the unbiased estimate of the mean prediction error rate, it is reported to be overestimated when the value of n<<p (M. W. Mitchell, 2011).

The analysis presented in the method I was carried out assuming that over the 10,000 models there was a good correlation between the composition of the training set and the model's performance i.e. the choice of solvents that appeared in the training set of the high performing model contributed to its high performance. However, if the cause of the high performance was due to the composition of the solvents in the test set and not the training set, there would be little or no trend in the expected model's performance as selecting the final training set based on such analysis would not improve the model. Looking at Table 5.12, the overall prediction accuracy of the model trained with the final selected solvents were found to be better than the mean average accuracy at low temperature points and significantly close to the mean value for carbamazepine and carvedilol at high temperature. Although, the study has not ruled anything out in terms of the test set composition, it has shown that choosing the training set in this way led to an above-average performance. However, the value of Sn was found to be 0% and the value of Sp as 100% for carbamazepine at low temperature. On analysis of the responses of the solvents in the training set, it was found that the training set consisted of only one class label in the training set and thus the model was invalid and unsuitable. For carvedilol at lab temperature, the labels in the training set obtained from the method I was highly unbalanced with 14 solvents labeled as 'Practically Insoluble' and only 4 solvents labelled as 'soluble'. This resulted in the low value of Sn. It can be clearly seen that class imbalanced in the training set can have an impact on the performance of the classification model. Class imbalance and its impact on model performance have been well researched in literature and the solution of tacking this problem is recommended either by sampling techniques such as down-sampling by reducing the majority or up sampling by increasing the minority or by cost sensitive learning techniques (C. Chen & Breiman, 2004). However, even with the availability of these techniques to tune for better model performance, it is to be noted that the performance of the best model is not based on the performance evaluation of the training set but rather of the test set. For a solvent selection model, it is difficult to know the true misclassification of the dataset at the learning stage without performing all the experiments in the lab. This is further problematic when the target compound is only soluble or practically soluble in most or all of the solvents in the training set which was the case for carbamazepine at low temperature in *method I*. Sample size also has an effect on the class imbalance. As for solvent selection, a minimum number of experiments is desired. This leads to a smaller sample size and the possibility of less information on the minority class. As the solvents selected using *method I* is purely based on the frequency of the solvent's appearance on the best models, it can be hard to generalize a stable set of solvents for selection as well as a set number. This is because the frequency is proportional to the choice of target compound selected for the screen.

Issues related to solvents selected by *method I* was readily solved by *method II* as the method of the solvent selected in *method II* was one solvent per cluster. The solvents in the cluster map were clustered according to the similarity in the calculated molecular descriptors, covered a diverse range of solvents and the distance between each solvent was set according to their dissimilarity in molecular properties (Johnston et al., 2017).

The Cohen's kappa parameter measured for the target compounds at two temperature points for both *method I* and *method II* helps evaluate the performance of the model by comparing the observed accuracy with the expected accuracy (random chance) (McHugh, 2012). The value of κ obtained for paracetamol from *method I* was very high which indicates a good agreement between the model's prediction and the actual labels. However, the value of κ was 0 for carbamazepine due to an invalid model at low temperature point (Table 5.12). The value of κ was found to be quite consistent above 0.6 for all target compounds at low temperatures for *method II* (Table 5.13). However, at high temperature, the value of κ for carbamazepine and especially for carvedilol were low which indicates a poor agreement between the actual experimental and model's classification labels.

The high misclassification for these two target compounds could potentially be due to the nature of these compounds to form various enantiotropic polymorphs and pseudo-polymorphs (Defossemont, Randzio, & Legendre, 2007; Hiendrawan, Widjojokusumo, Veriansyah, & Tjandrawinata, 2017). Paracetamol on the other hand has not been reported to change in polymorphism under the experimental conditions done on this study. Oiling out is found to occur on carvedilol in certain solvents when the temperature is increased. This affects the accuracy of the turbidity measurements and provides unreliable experimental data. Fouling and degradation were also observed at the end of the experiments in both carbamazepine and carvedilol at high temperature which could have contributed to experimental errors.

Once the RF classification models were trained for the target compounds, a proximity matrix was constructed which quantifies the similarity between the solvents. Only the proximity of the solvents selected using *method II* was drawn as there were greater class imbalance and poor or invalid model performance using method I in carbamazepine and carvedilol. The proximity measures between two solvents is the measurement of the frequency of the placement of the two solvents at the same terminal node of the same RF tree divided by the number of trees built. Treating the proximity matrix as a set of point to point distances, multidimensional scaling (MDS) can be performed to allow the visualization of the data in two dimensions, with spatially-proximal data points having high proximity in the model. Ideally, well-separated clusters of points corresponding to each class should be observed. The proximities of solvents for the three target compounds are presented as MDS plots at low temperature points in Figure 5.17, Figure 5.19 and Figure 5.21 and at high temperature points in Figure 5.18, Figure 5.20 and Figure 5.22. The MDS plots show a general clustering of the solvent belonging to each solubility outcome of *soluble* and *practically insoluble.*

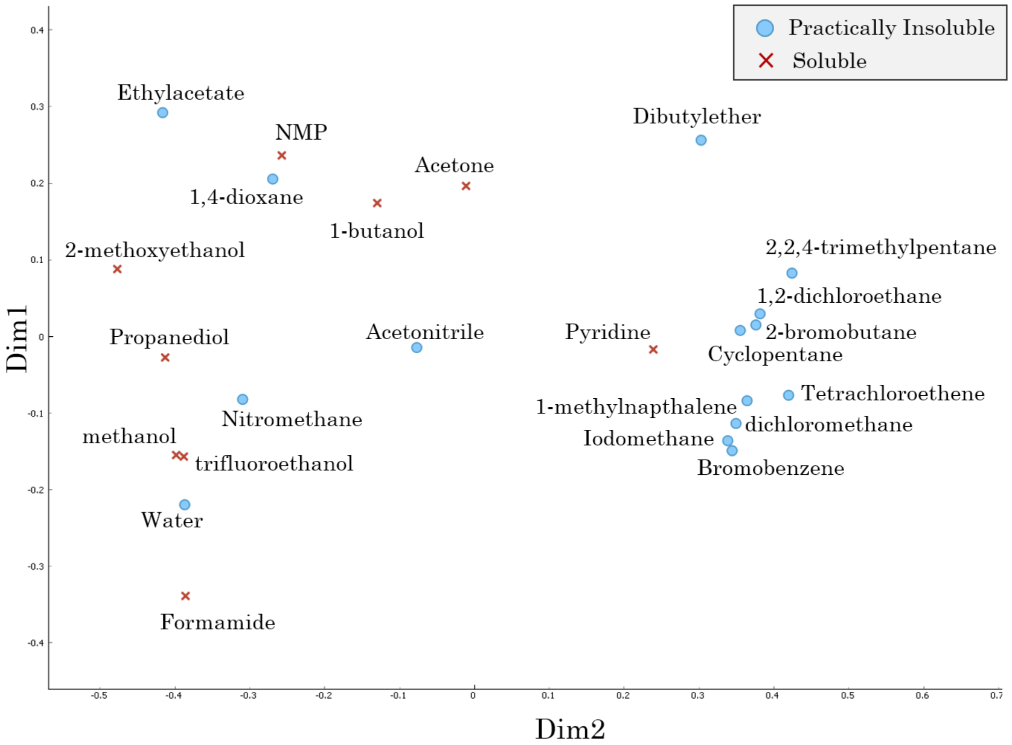


Figure 5.17 MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for paracetamol in 24 solvents (*Method II*).

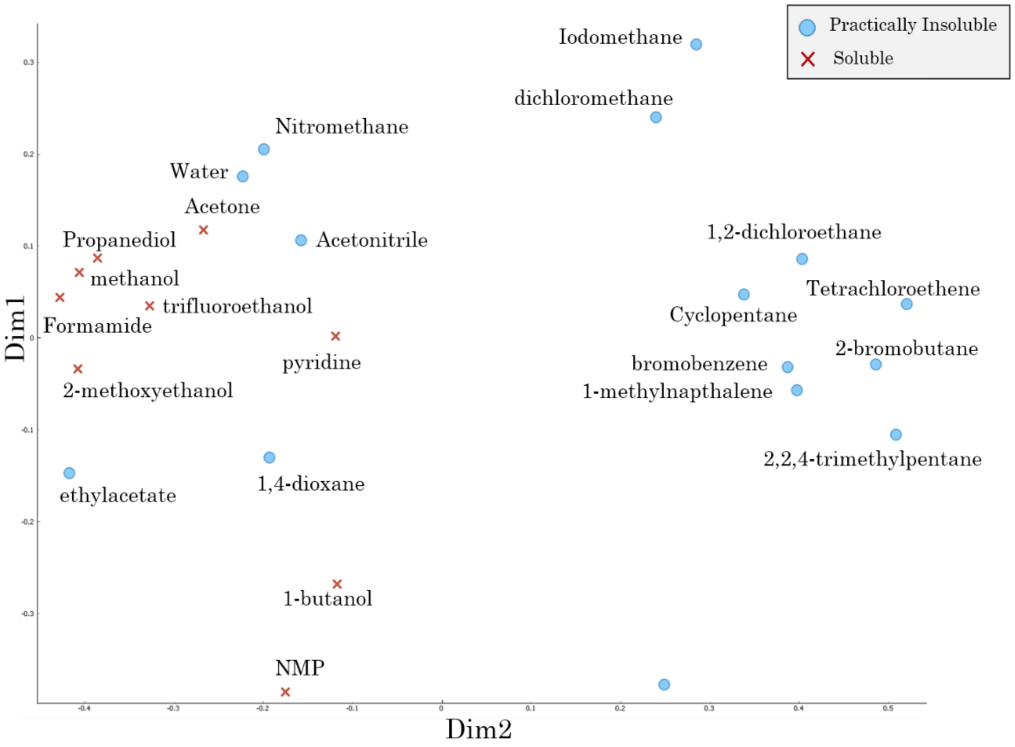


Figure 5.18 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for paracetamol in 24 solvents (*Method II*).

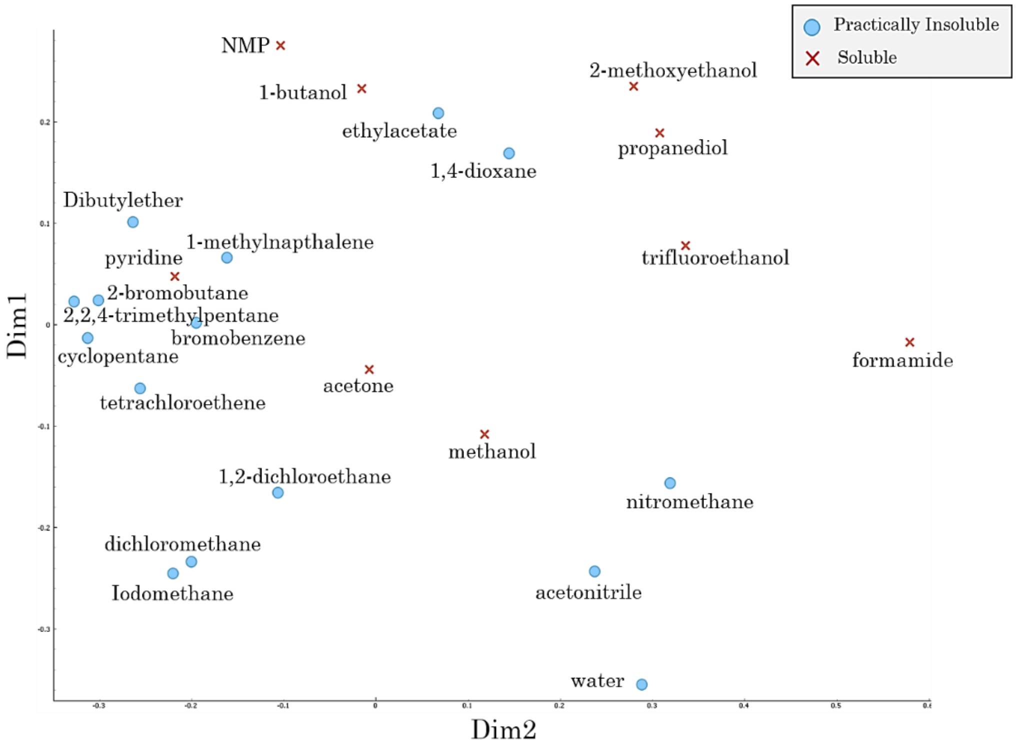


Figure 5.19 MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for carbamazepine in 24 solvents (*Method II*)..

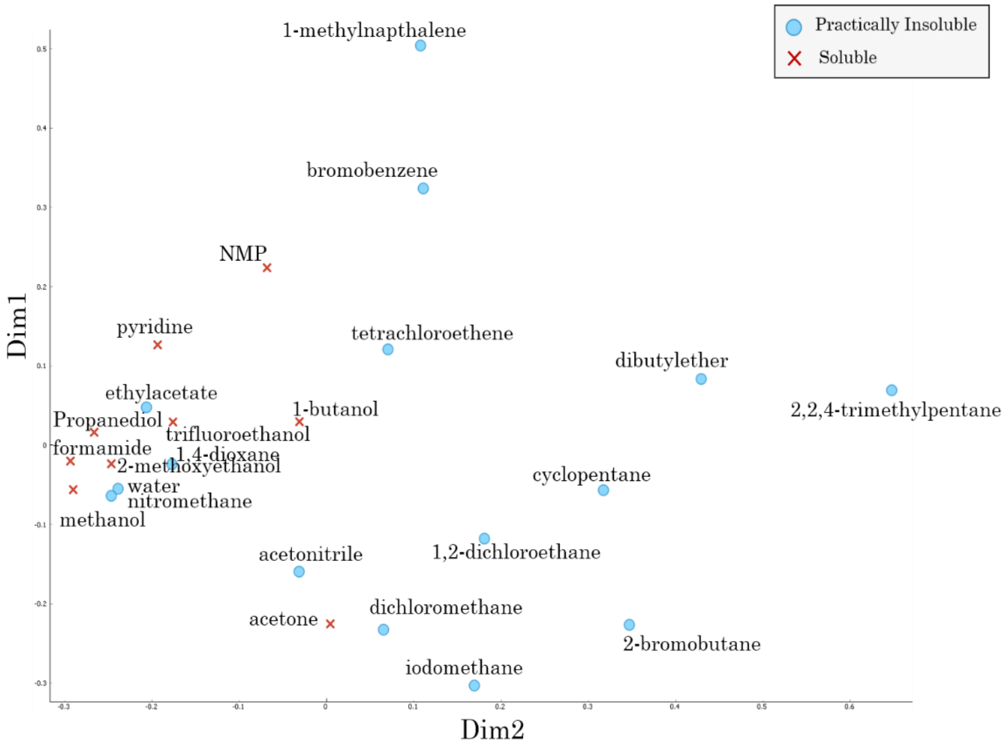


Figure 5.20 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for carbamazepine in 24 solvents (*Method II*).

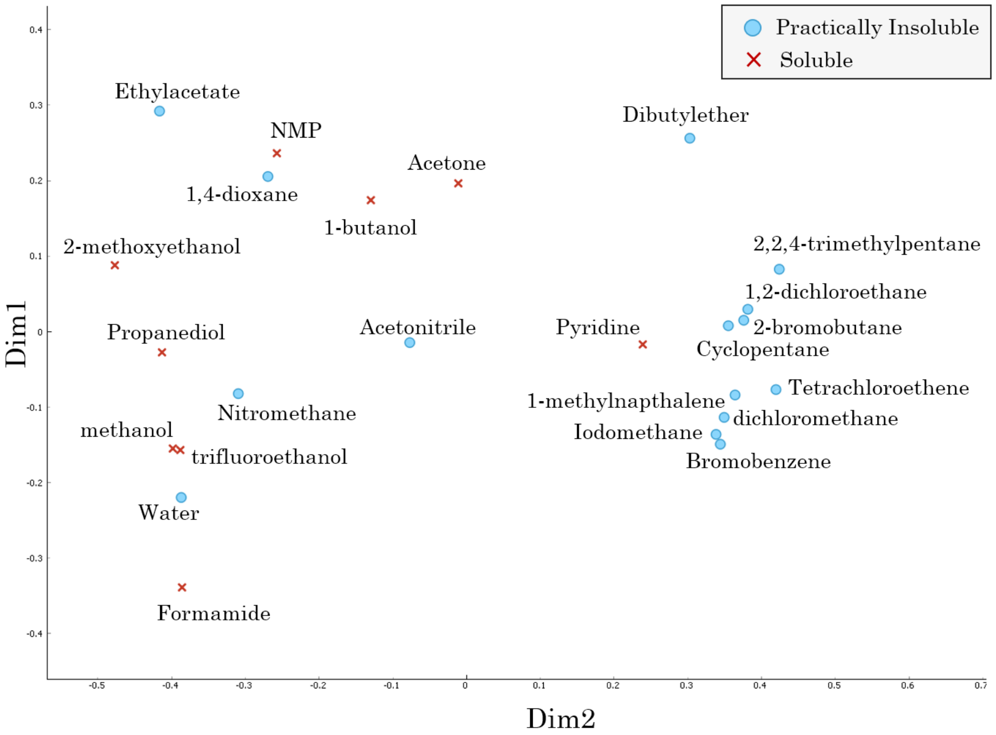


Figure 5.21MDS plot obtained from the RF classification proximity matrix showing clustering at low temperature for carvedilol in 24 solvents (*Method II*).

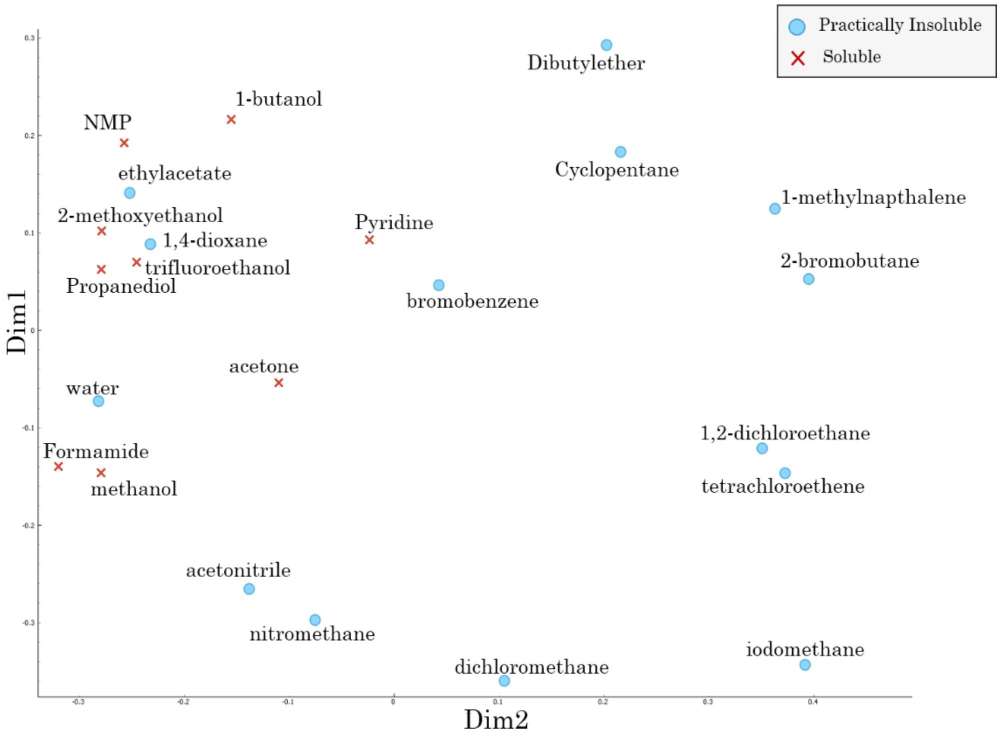
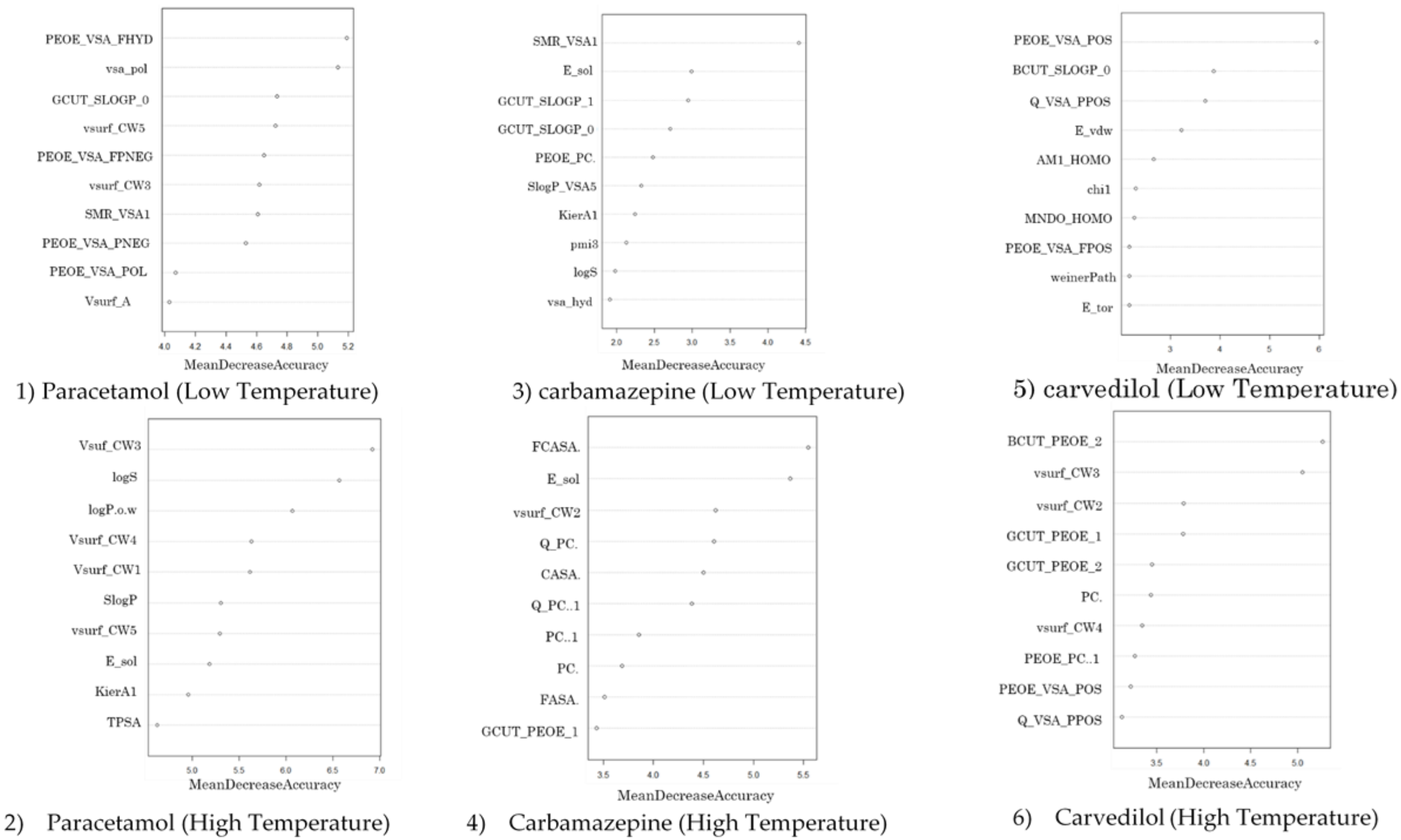


Figure 5.22 MDS plot obtained from the RF classification proximity matrix showing clustering at high temperature for carvedilol in 24 solvents (*Method II*).

5.4.5 Optimisation attempts on the trained RF classification model

An important feature of the RF model is its ability to directly measure the impact of each variable on the prediction accuracy of the model. Consistency in the selection of the molecular descriptors by the RF model for each run is of great importance which indicates the stability of the trained classification model. The analysis and interpretation of these important molecular descriptors can help extract relative information on the molecular structure of the organic solvents and its effect on the relative solubility of paracetamol, carvedilol and carbamazepine in the organic solvents. The top ten molecular descriptors obtained from the variable importance of the RF classification models trained using solvents selected by method II for all the three target compounds at two temperature points are shown in Figure 5.23.



1) Paracetamol (Low Temperature)

3) carbamazepine (Low Temperature)

5) carvedilol (Low Temperature)

2) Paracetamol (High Temperature)

4) Carbamazepine (High Temperature)

6) Carvedilol (High Temperature)

Figure 5.23 The top ten important molecular descriptors for the selected RF classification model of the three target compounds taken at low and high temperatures and sorted accordingly to its average variable importance in descending order

The RF model was retrained by selecting only one quarter of the 250 molecular descriptors (62 descriptors) while removing the least important descriptors in the order of their mean decrease accuracy. The ACC, Sn, Sp and κ were then evaluated and compared with the model trained with 250 variables. The value of ACC, Sn, Sp and κ obtained with 62 variables were exactly the same as the original trained model with 250 variables. This process was repeated with the top 30 variables and the results again did not vary. This indicates that the remaining variables had little to no influence in the prediction of relative solubility and that only 30 variables were enough to develop a robust classification model.

5.4.6 Assessment of the important molecular descriptors

Table 5.14 Most important molecular descriptors extracted from the trained RF classification model for the three target compounds at two temperature points (Boyd, 2005)

Descriptors	Category
BCUT_PEOE_2, BCUT_SLOGP_0, GCUT_PEOE_1, GCUT_PEOE_2, CUT_SLOGP_0, GCUT_SLOGP_1, weinerPath	Adjacency and distance matrix descriptors
CASA, FASA, FCASA, chi1	Conformation Dependent Charge Descriptors
Chi1, KierA1	Kier & Hall Connectivity and Kappa Shape Indices
AM1_HOMO, MNDO_HOMO	MOPAC descriptors
PC, PC+, PC-, PEOE_PC+, PEOE_PC-, PEOE_VSA_FHYD, PEOE_VSA_FPNEG, PEOE_VSA_FPOL, PEOE_VSA_FPOS, PEOE_VSA_PNEG, PEOE_VSA_POS, PEOE_VSA4, Q_PC, Q_PC-, Q_VSA_PPOS	Partial Charge Descriptors
vsa_hyd	Pharmacophore Feature

	Descriptors
logP.o.w, logS, slogP, TPSA	Physical Properties
E_sol, E_tor, E_vdw	Potential Energy Descriptors
SlogP_VSA5, SMR_VSA1	Subdivided Surface Areas
Pmi3, Vsurf_A, vsurf_cw1, vsurf_CW2, Vsurf_CW3, Vsurf_CW4, Vsurf_CW5	Surface Area, Volume and Shape Descriptors

The important variables obtained from Figure 5.23 were categorized in Table 5.14. For paracetamol at low and high temperature points, the top descriptors responsible for solubilisation were observed to be capacity factors (Vsurf_CW) which describes the amount of hydrophilic regions per unit surface area, vsa_pol which shows the polar groups such as hydrogen bond donor and acceptor groups in the van der Waals surface of the molecule, the partial charge descriptors which calculates partial charge on the van der Waals surface area (PEOE_VSA_FHYD) and the log of aqueous solubility (Baurin et al., 2004; Hari Narayana Moorthy, Ramos, & Fernandes, 2011). Similarly, for carbamazepine, the top important descriptors included the partial charge descriptors, conformation dependent charge descriptors, the adjacency and distance matrix descriptors, log of aqueous solubility, third largest principle moment of inertia (pmi3), first alpha modified shape index (KierA3), capacity factors, subdivided surface area descriptors and approximation to the sum of the van der Waals surface areas of hydrophobic atoms. Similarly, for carvedilol at two temperature points, the important variables included the MOPAC descriptors, the adjacency and distance matrix descriptors, atomic connectivity index, torsion and van der Waals component of the potential energy, partial charge descriptors, capacity factors and the Wiener path number. It is difficult to gain a deeper chemical understanding relating to qualitative solubility simply by looking at the importance of molecular descriptors; however, some general trends can be drawn. There is precedent for these descriptors contributing to solubility prediction such as the Kier and Hall connectivity and kappa shape indices, partial charge descriptors and the pharmacophore descriptors (A. R. Katritzky, Fara, Kuanar, Hur, & Karelson, 2005; A. R. Katritzky et al., 2003). The top ten descriptors calculated for paracetamol at two temperature points obtained in this chapter was found to be similar and shared some of the descriptors predicting crystallisation outcome in previous studies (Section 4.3.2.2.1.2).

5.5 Summary

There are many examples in the literature where RF is shown to outperform other machine learning algorithms (Caruana & Niculescu-Mizil, 2006; Martin et al., 2013; Ogutu, Piepho, & Schulz-Streeck, 2011; Palmer et al., 2007), which led to choosing RF for this study. Furthermore, from Chapter 4, models trained using RF algorithms showed success in predicting solubility of paracetamol in diverse organic solvents. The aim was to develop a rapid and efficient solvent selection tool. The trained RF model was able to achieve that goal by predicting qualitative solubility of paracetamol, carvedilol and carbamazepine with an accuracy of ~85%, ~71% and ~69%, respectively. The purpose of the model was to reduce the number of screening experiments in the laboratory, and so reduce material cost and usage. Thus, in order to balance the desire for fewer experiments with the need for an effective machine learning predictive model, 10,000 RF classification models with randomly-sampled training sets were trained at varying training set sizes for each compound and temperature, and the spread results analysed. Doing so helped set the optimum training set size between 20 and 24 solvents. However, selecting the optimum training set size was not sufficient as the accuracy of the RF models depended upon the selection of solvents for the training sets. Two methods were suggested in order to rationally select these solvents. Method I, where data mining was performed on the built 10,000 RF models with low error rate in order to observe the trend of the solvent's frequency in the good models and Method II, where the solvents were clustered accordingly to the similarity in the physicochemical properties (Johnston et al., 2017). Method II was preferred over method I. The RF models trained using the selected solvents from method I not only suffered from class-imbalance but a major problem was that experimental knowledge was still required to identify the frequently occurring solvents. Moreover, data mining was computationally demanding and complex. However, method II was comparatively trivial, allowing the researcher to simply select the solvents randomly from each cluster and develop RF models with a relatively low prediction error rate.

Even though predictions made using Method II gave relatively high prediction accuracy, further improvements could still be made to make the model even more accurate. Prediction accuracy could have been influenced by human error during data recording in the laboratory. Since the model was based on calculated molecular descriptors, it ignored the effect of impurities, change in compound's solid forms and the reaction of the API with the solvent. As mentioned previously, it is difficult to gain a deeper chemical understanding from examining only the molecular descriptors. However, analysis of descriptor importance highlighted the partial charge descriptors, Kier and Hall connectivity and kappa shape indices, and the pharmacophore descriptors as the most important for characterising intermolecular interactions in solution. Similar descriptors (partial charge descriptors and Surface Area, Volume, and shape descriptors) were also observed as being the most important molecular descriptors describing the solubility in paracetamol in Chapter 4 (Figure 4.21). This leads to the conclusion that these top descriptors describe the solubility behaviour of drug and drug like compounds in organic solvents and their relationship with solubility can be further investigated.

Finally, it can be concluded that RF was successfully utilised as a rapid and efficient predictive tool for recommending solvents for the initial stages of crystallisation process design which helped reduce computational cost and experimental time. This method helps in quickly eliminating solvents not suitable for the crystallisation process and recommends solvents for further analysis of precise solubility measurements. Addition of Safety, Health and Environment (SHE) information of the solvents to the modelling tool adds further value to the screening platform in recommending a suitable solvent for a crystallisation process design.

Chapter 6.　Application of machine learning to predict solubility of drug and drug-like compounds in organic solvents

6.1 Introduction

Numerous chemoinformatics methods have been applied for estimating the aqueous solubility of various drug and drug-like compounds in water (S. J. Ali & Rajini, 2012; Delaney, 2005; Hewitt et al., 2009; Lusci, Pollastri, & Baldi, 2013; McDonagh, Nath, De Ferrari, van Mourik, & Mitchell, 2014; Palmer et al., 2007; Shayanfar, Fakhree, & Jouyban, 2010). The application of chemoinformatics has also been widely reported in the estimation of various other physicochemical properties (melting points, boiling points, logP) which aid in the prediction of aqueous solubility (Bhat, Merchant, & Bhagwat, 2008; Tetko, Lowe, & Williams, 2016). However, when compared to aqueous solubility studies, far fewer studies reporting the application of ML for solubility estimation of drug compounds in organic solvents can be found. This is likely to be in large part due to the lack of sufficient publicly available data for training ML algorithms, as well as the reliability of what solubility data is available. Even though companies do tend to screen thousands of molecules, these data are rarely made public. In 2015, Buanaiuto and Lang implemented a RF regression model to predict the solubility of 261 compounds in 1-octanol with an out-of-bag (OOB) mean squared error of 0.34 M (Buonaiuto & Lang, 2015). Similarly, in 2013 Tetko et al. compared various classification ML algorithms to estimate the solubility of 163000 molecules obtained from UCB and Enamine Ltd in DMSO (Balakin, Savchuk, & Tetko, 2006); however, the data used for the analysis was not made available.

The aim of the chapter was to curate solubility data from the literature to develop and implement an ML-based model capable of predicting the solubility of drug-like compounds in common organic solvents. This approach, collecting solubility data for a large number of drug-like compounds in a small selection of solvents, is "orthogonal" to the approach presented in Chapter 5, which focussed on the development of models on an all-solvent, per-solute basis. Both classification and regression models were constructed for comparison purposes and to allow a deeper understanding of the most accurate way of predicting non-aqueous solubility.

6.2 Methodology

6.2.1 Solubility database

A solubility database of drug and drug-like molecules was compiled to train and test the RF model. The data were chiefly gathered from two sources: the Handbook of solubility data for pharmaceuticals (Jouyban, 2009) and text mining of the American Chemistry Society and ScienceDirect journal archives (this was performed with tools available in KNIME). The Handbook of solubility data for pharmaceuticals consisted of approximately 5000 solubility data points of pharmaceutical compounds in pure solvents, measured mainly in mole fraction or g/L at temperatures in the range of -5 to 70°C. Compilation of the solubility data from the journal archives was on the basis that solubility measurements determined for the pharmaceutical compounds were consistent, compounds were structurally diverse and present as solid-phase at room temperature. For this study, solubility measurements on seven common organic solvents at laboratory temperature (25°C) were collated for the solubility database. These solvents were chosen for the study because they were the frequently occurring solvents used in the crystallisation of pharmaceuticals in the Cambridge Structural Database (CSD). During the curation process, more than one solubility datum for the same compound was observed depending upon the number of its polymorphic forms, the number of studies done on the same compound by different researchers and the choice of measurement technique applied. It was observed that most of the thermodynamic solubility measurements in the literature were determined using synthetic methods. In synthetic methods, the composition of the saturated solution is determined by overall weight or measuring the individual components and the solubility is determined by the state in which the solid phase just disappears. The disappearance of the solid phase is induced by the change in temperature or by the addition of a known amount of solvent. High definition camera is utilised to visualise the disappearance of the solid phase in solution at a given temperature. Synthetic methods overcome the limitations of analytical methods such as HPLC and UV methods which are tedious and time consuming (Alves, Condotta, & Giulietti, 2001; N. Tang, Shi, & Yan, 2016; Y. N. Wang, Fu, Jia, Qian, & Chen, 2013). Thus, solubility measurements performed using these methods were preferred for the curated database. Similarly, for compounds existing with various polymorphs, the solubility of only the most stable polymorph was considered. For example, three polymorphic forms of Lamivudine (Form I, II and III) are reported in the literature, with Form II being the most thermodynamically stable form (Chadha, Arora, & Bhandari, 2012); only form II was only considered in the solubility dataset. Furthermore, salts, hydrates and solvates were omitted from the selection and a Lipinski filter was performed (Gimenez, Santos, Ferrarini, & Fernandes, 2010; Lipinski, Lombardo, Dominy, & Feeney, 1997; Pickett, 2007). A list of the number of solubility measurements for each solvent at laboratory temperature is presented in Table 6.1. Among the seven solvents, ethanol was observed to be the most commonly used solvent (no. of occurrences:30,462) according to a CSD data search performed in March 2008 (Brittain, 2009). Similar findings were reported by Hosakawa et al. in 2005 where, from a list of 6397 compounds in the Cambridge Structural Database (CSD), 1328 compounds were crystallised in ethanol (Hosokawa, Goto, & Hirayama, 2005). Overall, the final solubility database featured solubility measurements of 247 unique compounds in seven solvents, expressed in mole fraction at 25°C. These measurements of compounds are listed for each solvents in Appendix 1.

Table 6.1 Shown are the number of compiled solubility data points

Solvent name	Number of solubility data points
Ethanol	181
Methanol	148
1-butanol	113
Acetonitrile	104
Acetone	102
Ethyl acetate	125
1,4-dioxane	47

There was no upper bound to the ideal dataset size for this task. However, the prescribed constraints on solutes (drug-like) solvents (one of seven specific organic solvents), temperature (25°C) and experimental protocol (synthetic method) were necessary to ensure that the dataset was fit for purpose. There are also only a sparse number of non-aqueous solubility databases available for public access beyond the *Handbook of solubility data for pharmaceuticals*, such as the Open Notebook Science Solubility Database (Bradley *et al.*, 2010), NIST solubility database (Acree, 2014) and the OCHEM chemical database (Sushko et al., 2011). These datasets were considered for inclusion in this project, but there were too many instances of incomplete data, incorrectly specified parameters (e.g. temperature) or lack of reference to the original study. Moreover, there were many overlapping data points with the *Handbook of solubility data for pharmaceuticals*. Thus, the compiled data shown in Table 6.1 are still an adequate and reasonable starting point for developing and testing a predictive ML model.

6.2.2 Molecular descriptors and model development

340 physicochemical molecular descriptors (both 2D and 3D descriptors) were calculated for the 247 unique compounds using the Molecular Operating

Environment (MOE, 2014.09 release). These numerical physicochemical descriptors were pre-processed by removing any missing values and descriptors with zero-variance which reduced the variables from 340 to 270. MOE software was also used to calculate Molecular ACCess System (MACCS) structural fingerprints, a classical fingerprint developed mainly for substructure and similarity screening studies in chemoinformatics. The 166 public MACCS fingerprints containing 1500 binary vectors were used as raw descriptors for *in silico* prediction of solubility in this chapter (Fernandez-de Gortari, Garcia-Jacas, Martinez-Mayorga, & Medina-Franco, 2017). Both the numerical and complex fingerprint descriptors were calculated from their respective 3D molecular structures of the compounds based on canonical SMILES and converted using Pipeline pilot 2017 (Hahn, 1995). (Section 3.3.2).

6.2.3 Model development

Both RF regression and classification models were developed in this study to predict the non-aqueous solubility of compounds. The RF regression model was developed to predict non-aqueous solubility as a continuous, numerical value. Solubility was expressed as the logarithmic transformation of mole fractions at laboratory temperature. The mole fractions were expressed in the logarithmic form to handle the non-linear relationship between the independent and dependent descriptors (Benoit, 2011). On the other hand, RF classification was used to predict the range of non-aqueous solubility as a categorical value. For this, solubility was qualitatively divided into regions in accordance with the solubility ranges specified on the Merck Solubility Index (Whitesell, 1998). The Merck Index provides a qualitative description of drug solubility rather than a specific value of solubility as shown in Table 6.2. Similar solubility ranges were also utilised to correlate the solubility of common organic compounds (Qiu & Albrecht, 2018). The unit of solubility for the Merck Index is mg/ml; the collected data, expressed in mole fractions, were therefore converted to mg/ml using each solvent's density at 25°C.

Table 6.2 Solubility definitions taken as per the Merck Index (Wolk, Agbaria, & Dahan, 2014)

Descriptive term (solubility definition)	Solubility range (mg/ml)
Very soluble	≥ 1,000
Freely soluble	100 to 1000
Soluble	33 to 100
Sparingly soluble	10 to 33
Slightly soluble	1 to 10
Very slightly soluble	0.1 to 1
Practically insoluble	< 0.1

For this study, two different sets of solubility regions were defined. Initially, only two solubility regions, i.e. *soluble* region and *insoluble* region were defined using the Merck Solubility Index. Compounds with solubility > 33 mg/ml were regarded as being in the *soluble* region while compounds with solubility ≤ 33 mg/ml were regarded as being in the *insoluble* region. This threshold was chosen because, as well as being an intuitive place to start, it produced a reasonably balanced split between the two classes, as presented in Table 6.3.

Table 6.3 Number of solubility data points classified in each of the binary qualitative solubility regions for each organic solvent based on a solubility threshold of 33 mg/ml

List of solvents	'*Soluble*' region (> 33mg/ml)	'*Insoluble*' region (≤ 33 mg/ml)
Ethanol	88	93
Methanol	60	88
1-butanol	61	52
Acetone	46	56
Acetonitrile	38	66
Ethyl acetate	73	52
1,4-dioxane	15	32

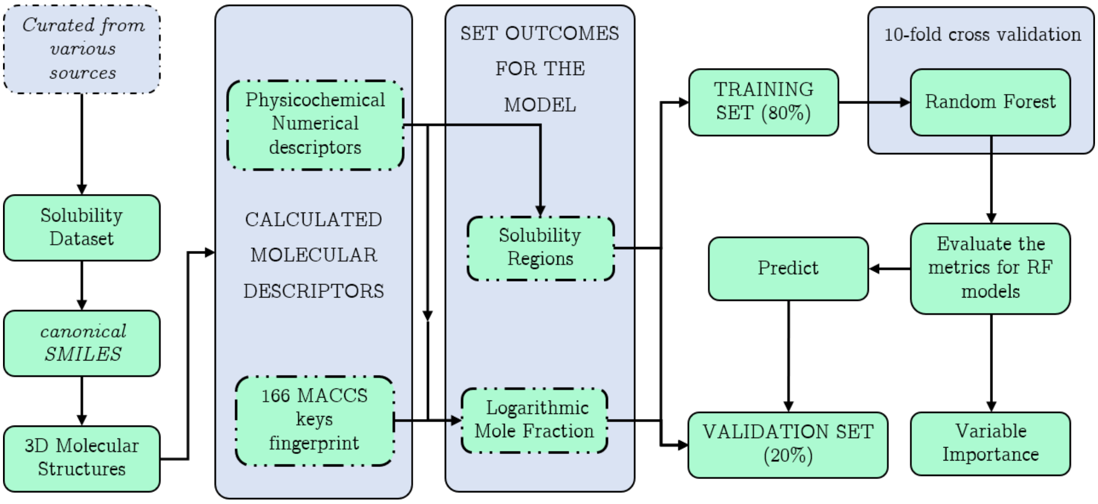
Nevertheless, class imbalance varied across solvents. It can be observed from Table 6.3, solubility dataset for methanol, acetonitrile, ethyl acetate and especially 1, 4-dioxane were more imbalanced compared to the solubility dataset for ethanol, 1-butanol and acetone. Beside these binary solubility regions, a further attempt was made to classify the solubility data by creating three solubility regions: *practically insoluble*, *soluble* and *highly soluble,* defined, respectively, as solubility less than 10 mg/ml; solubility between 10 mg/ml and 100 mg/ml, and solubility above 100 mg/ml. Table 6.4, showed the number of solubility data points set at these three solubility regions for each solvent.

Table 6.4 Number of solubility data points classified in each of the binary qualitative solubility regions for each organic solvent based on the solubility criteria shown in Table 6.3.

List of solvents	Practically Insoluble	Soluble	Highly Soluble
Ethanol	57	71	53
Methanol	39	48	61
1-butanol	37	51	25
Acetone	27	39	36
Acetonitrile	25	33	46
Ethyl acetate	50	47	28
1,4-dioxane	10	13	24

6.2.4 ML model workflow

Solubility dataset containing the compounds and their calculated numerical and fingerprint descriptors were randomly divided into 80:20 ratios, from which 80% were used as the training set and the remaining 20% were used as the external validation set. 10-fold cross validation was performed on the training set with the calculated descriptors in order to evaluate the performance of the RF models. It evaluates performance and stability by randomly splitting the dataset into 10 groups, training a model with 9/10 groups and testing it on the remaining 1/10. This procedure is performed 10 times, sequentially leaving out each group for testing and aggregating the results (Bengio & Grandvalet, 2004). 10-fold cross validation on the training set was performed using the 'caret' package in R (Version 3.5.1) (A. Liaw & M. Wiener, 2002). After evaluating model performance by 10-fold cross validation, a final model was trained on the full training set to predict non-aqueous solubility on the remaining 20% external validation set. For the RF classification model, the default value of *mtry* (number of randomly selected descriptors) was used (square root of the total number of predictors). The default value of *mtry* was also used for the RF regression model, where it is a third of the number of predictors. The number of trees generated for the RF classification model was set at 1500 trees while that for the RF regression model was set at 500 trees. Both regression and classification RF models were trained using the 'randomForest' package in R (Version 3.5.1) (Ihaka & Gentleman, 1996). A schematic workflow of the model development showing the various stages is shown in Figure 6.1.



Curated from various sources
Solubility Dataset
canonical SMILES
3D Molecular Structures
Physicochemical Numerical descriptors
CALCULATED MOLECULAR DESCRIPTORS
166 MACCS keys fingerprint
SET OUTCOMES FOR THE MODEL
Solubility Regions
Logarithmic Mole Fraction
TRAINING SET (80%)
VALIDATION SET (20%)
10-fold cross validation
Random Forest
Evaluate the metrics for RF models
Predict
Variable Importance

Figure 6.1 Workflow of the RF machine learning algorithm to predict non-aqueous solubility of drug and drug like compounds at laboratory temperature. 10-fold cross validation was performed to evaluate the performance of the RF model. The trained RF model was then used to predict the non-aqueous solubility on the validation set.

6.2.5 Model performance metrics

In order to measure and assess the performance of the RF regression model, three standard metrics i.e. root mean square error (RMSE), mean absolute error (MAE) and the square of the Pearson correlation coefficient or R-squared (R^2) were used. RMSE is the measure of the average deviation of the predicted data from the experimental observations (residuals), MAE is the absolute difference between the predicted data and the experimental observations while R-squared is the relative measure of how close the predicted data from the RF model are to the fitted regression line (Chirico & Gramatica, 2011). A picture of a good predictive model is dictated by relatively low values of RMSE and MAE and higher values of R-squared. The values of both RMSE and MAE range from 0 to infinity. There are no universal threshold values of RMSE and MAE to reflect a model's predictive ability. Similarly, the performance of the RF classification models was evaluated using the accuracy of the statistical measure (ACC), *sensitivity* (Sn), *specificity* (Sp), *precision* or *positive predictive value* (PPV), *Cohen's kappa* (κ) and summarised by a confusion matrix. These performance metrics for both the RF regression and classification models are further explained in detail, along with their respective equations, in Section 1.3.5.

6.3 Results and discussions

6.3.1 RF regression models

Results obtained from the RF regression model constructed by performing 10-fold cross validation using both numerical and MACCS fingerprints are summarised in Table 6.5 and their regression plots are visualised in Appendix 2A. It can be observed that, overall, the predictive ability of the regression models built using MOE descriptors was superior compared to the models built using MACCS fingerprints. The RMSE and MAE values of acetonitrile, 1-butanol and ethanol were found to be relatively lower compared to the values found for methanol, acetone, ethyl acetate and 1,4-dioxane. The RF regression model built on acetonitrile with MOE physicochemical descriptors had the best average prediction accuracy according to the cross-validation results with the lowest RMSE value of 0.67 LogS and MAE value of 0.506 LogS. This indicates a good fit of the RF model to the acetonitrile data compared to other solvents i.e. the observed data was close to the model's predicted values. The random forest regression model predicted the non-aqueous solubility of 247 compounds in the following order of their average predictive performance: acetonitrile > 1-butanol > ethanol > methanol > acetone > ethyl acetate > 1, 4-dioxane. RF models trained using both the numerical descriptors and fingerprints were then selected for validation on the 20% external test set. The predictive performance of the trained model using both descriptors on the validation set is summarised in Table 6.6 and their regression plots are visualised in Appendix 2B.

Table 6.5 Comparison of the prediction performance of RF regression model using 10-fold cross validation on seven solvents built using MOE molecular descriptors and MACCS molecular fingerprints

	MOE descriptors			MACCS descriptors		
Solvents	RMSE	R^2	MAE	RMSE	R^2	MAE
Ethanol	0.721	0.578	0.541	0.765	0.516	0.584
Methanol	0.734	0.562	0.565	0.799	0.469	0.636
1-butanol	0.674	0.582	0.513	0.571	0.669	0.441
Acetonitrile	0.670	0.628	0.506	0.722	0.584	0.525
Acetone	0.757	0.363	0.585	0.773	0.337	0.591
Ethyl acetate	0.964	0.400	0.732	0.939	0.428	0.721
1,4-dioxane	0.959	0.008	0.685	0.998	-0.209	0.725

Table 6.6 Comparison of the prediction performance of RF regression model on the 20% external validation set on seven solvents

	MOE descriptors			MACCS descriptors		
Solvents	RMSE	R^2	MAE	RMSE	R^2	MAE
Ethanol	0.650	0.522	0.479	0.704	0.412	0.551
Methanol	0.657	0.497	0.505	0.618	0.538	0.468
1-butanol	0.524	0.609	0.375	0.781	0.471	0.65
Acetonitrile	0.805	0.374	0.570	0.563	0.516	0.438
Acetone	0.794	0.614	0.608	0.898	0.365	0.688
Ethyl acetate	0.689	0.602	0.545	0.744	0.508	0.607
1,4-dioxane	0.839	0.692	0.559	0.849	0.445	0.577

If the evaluation metric values were significantly better for the training set (i.e. lower value of RMSE and MAE while higher value of R^2) than that of the validation set, this would indicate one of two things; either the model has highly over-fitted where the model learns patterns highly specific to the training set and then struggles to maintain performance on independent validation sets or the validation set size is too small and not enough for the model to identify trends (Domingos, 2012). Even though similar predicted values between the training and validation set indicates a stable and valid model, the predictive capability of the model is determined by low values of both RMSE and MAE and higher values of R^2.

Results summarised in Table 6.6 suggested that the regression model built using MOE descriptors was moderately good at predicting solubility in alcohols whilst the model struggled to predict accurately in acetone, acetonitrile and 1,4 dioxane. This was indicated by higher values of RMSE and MAE. However, models built using MACCS fingerprints were good at predicting solubility in acetonitrile, methanol and ethanol while failed with high errors for ethyl acetate, 1-butanol and acetone.

The prediction metrics of acetonitrile were the best among the seven solvents in the cross-validation model using MOE descriptors, however the metrics were poor for the external validation set. As suggested by Hewitt et al (2009), this is a clear sign that the trained model underwent overfitting. The training set for acetonitrile thus did not contain enough information to generalise the trends to predict non-aqueous solubility for compounds in the test set (Hewitt et al., 2009). It can also be observed from Table 6.5 that model for 1, 4-dioxane, acetone and ethylacetate had high values of RMSE and lower values of R^2 in the training set. Lower values of R^2 indicates that the model explains none of the variability of the solubility data around its mean and that the predicted data points fall further from the regression line. Thus a poor fit of the model. This can also be observed from their regression plots in Appendix 2B. This could either be due to the extremely small dataset size for 1,4- dioxane (47 data points) or the inability of the models to fully capture the relationship between the molecular descriptors and the target property for ethylacetate, acetone and 1,4-dioxane.

The RF regression model built to predict non-aqueous solubility of the drug and drug-like compounds on seven organic solvents was compared with RF solubility predictions available in the literature. Buonaiuto and Lang (2015) created a similar random forest regression model using 86 physicochemical descriptors to predict the solubility of 259 compounds in 1-octanol with an R^2 value of 0.63 and RMSE of 0.616 LogS (Buonaiuto & Lang, 2015). These values were comparable in performance to the cross-validation results of the presented alcohol datasets (ethanol, methanol and 1-butanol) in Table 6.5. As mentioned earlier in the chapter, RF regression models have been extensively researched in the prediction of aqueous solubility rather than non-aqueous solubility in literature. Aqueous solubility data are found in abundance in public databases and are thus larger in number. For example, Palmer *et al.* developed a RF model to predict aqueous solubility on 988 compounds with an RMSE value of 0.685 which was comparable to the RMSE results obtained on non-aqueous solvents in this study (Palmer et al., 2007). However, R^2 value from this study was lower compared to the value reported by Palmer *et al* (0.896). This could be due to the significantly smaller solubility dataset size used in this study which prevented the models to fully capture the underlying trends between the molecular descriptors and the dependent variable property of interest, potentially be due to the failure in the algorithms themselves or due to the incomplete set of descriptors.

Overall, the built RF regression models on the external test set produced moderately good predictions of non-aqueous solubility for drug and drug-like compounds in alcohols (methanol, ethanol and 1-butanol) but failed for ethyl acetate, acetone, acetonitrile and 1,4-dioxane. This remains a significant finding given the comparatively small sample size compared to aqueous solubility databases; the regression models were nevertheless able to correlate some information obtained from the molecular descriptors with the experimental solubility data for some solvents. Rather than aiming to predict continuous and quantitative values of solubility of compounds on respective solvents, classification modelling with discretized labels was applied.

6.3.2 RF Classification model

The same non-aqueous solubility prediction problem was treated as a classification task. RF classification models (in classification mode) was were evaluated to map out where the solubility of drug and drug like compounds would map out under a finite set of possible outcomes. RF Classification model was initially built on the criteria set in section 6.2.3, of the chapter i.e. with two classification outcomes, where compounds with solubility > 33 mg/ml were regarded as being in the Soluble region and compounds with solubility < 33 mg/ml were regarded as being in the Insoluble region (Table 6.3). The dataset was randomly split into 80% as a training set and 20% as the external validation set. 10-fold cross validation was performed on the training set as shown in Table 6.7. Performance measures of the classification model after 10-fold cross validation are summarised in the following Table 6.8.

Table 6.7 Breakdown of the solubility dataset in the ratio of 80:20 with 80% as a training set and 20% as an external validation set

Solvent		Soluble Region	Insoluble Region
Ethanol	Training Set	69	75
	Validation Set	19	18
Methanol	Training Set	49	69
	Validation Set	11	19
1-butanol	Training Set	52	38
	Validation Set	11	12
Acetone	Training Set	38	43
	Validation Set	8	13
Acetonitrile	Training Set	29	54
	Validation Set	7	14
Ethyl acetate	Training Set	59	41
	Validation Set	14	11
1,4-dioxane	Training Set	13	24
	Validation Set	2	8

Table 6.8 Average prediction accuracy after 10-fold cross validation on seven organic solvents (binary classification)

Solvent	Sensitivity	Specificity	Cohen kappa	Precision	Accuracy
Ethanol	0.8000	0.7848	0.5816	0.7536	79.17
Methanol	0.7500	0.7838	0.5218	0.6735	77.20
1-butanol	0.7391	0.6364	0.3762	0.68	68.89
Acetonitrile	0.6786	0.8182	0.4924	0.6552	77.11
Acetone	0.6316	0.7674	0.4015	0.7059	70.37
Ethyl acetate	0.8182	0.6889	0.5112	0.7627	76.00
1,4-dioxane	0.5455	0.6953	0.2204	0.4286	64.86

It can be observed from Table 6.8 that, on the whole, the RF binary classifiers after 10-fold cross validation on the seven solvents performed reasonably well, with classification accuracy above 64%. The average predictive accuracy obtained from performing cross validation was highest in the ethanol model and lowest for the 1, 4-dioxane model. The Cohen's kappa values indicate that the predictions made for the ethanol model were in moderate agreement with the experimental outcomes while those for 1, 4-dioxane had an only slight agreement with the experimental outcomes. Sensitivity and specificity indicate how well the model identities the two outcomes, with higher values indicate the ability of the model to correctly identity the compounds with the correct outcome. The training sets for acetonitrile, methanol, ethyl acetate and 1,4-dioxane were imbalanced (Table 6.7), though for acetonitrile, methanol and ethyl acetate models were fairly accurate in classification the respective outcome with high accuracy of above 76%. Sensitivity and specificity for 1, 4-dioxane were low, indicating that the RF model was not capable of accurately classifying the compounds according to their respective outcomes. Confusion matrices drawn on Table 6.9 and Table 6.10 helps to understand the classification of the RF trained models for which the performance metrics are based on.

Table 6.9 Confusion matrix of the classification model of the alcohol models

		Reference Data					
		Ethanol		Methanol		1-Butanol	
		Soluble	Insoluble	Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	52	17	33	16	34	16
	Insoluble	13	62	11	58	12	28

Table 6.10 Confusion matrix of the classification model of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane

		Reference Data			
		Acetone		Acetonitrile	
		Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	26	12	19	10
	Insoluble	12	31	9	45
		Reference Data			
		Ethyl acetate		1,4-dioxane	
		Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	45	14	6	8
	Insoluble	10	31	5	18

From the confusion matrices (Table 6.9 and Table 6.10), RF model for ethanol had shown the least misclassification when compared to the other solvents and thus high specificity and sensitivity and specificity. The 1, 4- dioxane model struggled to correctly classify the soluble outcome with an error rate of 57.14%. This misclassification is due to the small sample size as well as the imbalance in the outcomes; the much better result for the insoluble outcome is equally an artefact of the class imbalance. Aside from 1, 4-dioxane, it was observed that the RF model the models were able to classify the outcomes in the six other solvents with less misclassification and fairly high accuracy. The prediction was then performed in the remaining 20% external validation set. The performance of the RF classification model when tested on the external validation set is shown in Table 6.11. Confusion matrices of the predicted model on the external validation set are presented to interpret the classification of the seven organic solvents (Table 6.12 and Table 6.13).

Table 6.11 Predictive performance of the RF model on the 20% external validation set for the seven organic solvents (binary classification)

Solvent	Sensitivity	Specificity	Cohen kappa	Precision	Accuracy
Ethanol	0.790	0.833	0.622	0.833	0.811
Methanol	0.727	0.842	0.569	0.727	0.800
1-butanol	0.636	0.583	0.219	0.583	0.609
Acetonitrile	0.333	1.000	0.364	1.000	0.714
Acetone	0.750	0.846	0.596	0.750	0.810
Ethyl acetate	0.714	0.546	0.262	0.667	0.640
1,4-dioxane	0.500	1.000	0.615	1.000	0.900

Table 6.12 Confusion matrix on the external validation set of the alcohols models

		Reference Data					
		Ethanol		Methanol		1-Butanol	
		Soluble	Insoluble	Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	14	5	5	6	7	4
	Insoluble	3	15	7	12	5	7

Table 6.13 Similarly, Confusion matrix on the external validation set of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane

		Reference Data			
		Acetone		Acetonitrile	
		Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	6	2	3	4
	Insoluble	2	11	5	9
		Reference Data			
		Ethyl acetate		1,4-dioxane	
		Soluble	Insoluble	Soluble	Insoluble
Predicted Data	Soluble	10	4	1	1
	Insoluble	5	6	0	8

Overall, it was observed that, when tested on the validation set, the models classified the outcomes on all the seven solvents with prediction accuracy of above 60%. The ethanol model had the most accurate prediction of the binary outcomes with the overall prediction accuracy of 81.08% and Cohen's kappa at 0.6219 interpreting the model as having a substantial agreement with the experimental outcomes (Table 6.11 and Table 6.12). The confusion matrix showed a good prediction between the two outcomes with fewer errors. RF model failed to accurately predict the soluble outcome for compounds in acetonitrile and 1, 4-dioxane. Predictions on 1-butanol and ethyl acetate models were also not strong, with significant error rates and low Cohen's kappa values. Soluble outcomes were predicted in the acetonitrile validation set with an accuracy of 42.86% while that for 1,4-dioxane was 50%. The validation set for both acetonitrile and 1, 4-dioxane were highly imbalanced which led to the RF to be more biased toward the majority cases. However, for 1-butanol and ethyl acetate, the validation sets were balanced yet a lot of compounds were misclassified. This confirms that the training data set for these two compounds do not represent the problem space. Similar high errors were observed in the regression models for these two solvents described earlier in this chapter, especially ethyl acetate (Table 6.5 and Table 6.6). In summary, the trained RF classification models were accurate in classifying drug-like compounds into a binary *soluble/insoluble* outcome for ethanol, methanol and acetone, whilst proving inaccurate for ethyl acetate, 1,4-dioxane, 1-butanol and acetonitrile models.

The solubility dataset was reclassified into three outcomes instead of two in order to investigate whether the extreme ends of the solubility spectrum (termed *highly soluble* and *practically insoluble*) could be well separated when considered apart from the central region. As mentioned in section 6.2.3, the dataset was divided into three solubility regions *i.e.* compounds with solubility less than 10 mg/ml were classified as *practically insoluble (PI)*. Compounds with solubility between 10 mg/ml to 100 mg/ml were classified as *soluble (S)* and compounds with solubility above 100 mg/ml were classified as *highly soluble (HS)*. The same workflow was followed as with the binary outcomes: the dataset was randomly split into 80% as a training set and 20% as the external validation set and 10-fold cross validation was performed on the training set, as shown in Table 6.14. Table 6.15 and Table 6.16 present the confusion matrices of the cross-validation classification model.

Table 6.14 Breakdown of the solubility dataset on the basis of the three solubility regions in the ratio of 80:20 with 80% as a training set and 20% as an external validation set

Solvent		Highly Soluble (HS)	Soluble (S)	Practically Insoluble (PI)
Ethanol	Training Set	19	27	37
	Validation Set	6	6	9
Methanol	Training Set	31	37	50
	Validation Set	8	11	11
1-butanol	Training Set	32	41	17
	Validation Set	5	10	8
Acetone	Training Set	23	26	32
	Validation Set	4	13	4
Acetonitrile	Training Set	19	27	37
	Validation Set	6	6	9
Ethyl acetate	Training Set	38	39	23
	Validation Set	11	9	5
1,4-dioxane	Training Set	8	8	21
	Validation Set	2	5	3

Table 6.15 Average prediction accuracy after 10-fold cross validation on seven organic solvents (Three label classification)

			Sensitivity			Specificity		
Solvent	ACC (%)	Kappa	HS	S	PI	HS	S	PI
Ethanol	63.19	0.443	0.781	0.540	0.625	0.874	0.753	0.808
Methanol	58.47	0.370	0.606	0.460	0.667	0.847	0.741	0.786
1-butanol	72.22	0.538	0.815	0.667	0.778	0.841	0.861	0.877
Acetonitrile	61.45	0.407	0.600	0.462	0.730	0.841	0.754	0.826
Acetone	43.21	0.129	0.429	0.200	0.550	0.767	0.639	0.751
Ethyl acetate	56	0.324	0.667	0.500	0.500	0.813	0.714	0.8
1,4-dioxane	56.76	0.1801	0.500	0	0.7037	0.8182	0.3710	0.752

Table 6.16 Confusion matrix of the classification model of the alcohol models with three outcomes

		Reference Data								
		Ethanol			Methanol			1-Butanol		
		HS	S	PI	HS	S	PI	HS	S	PI
Predicted Data	H. S	15	3	1	19	8	4	26	5	1
	Soluble	5	15	8	9	17	11	7	27	7
	PI	2	12	23	4	13	33	0	2	15

Table 6.17 Similarly, Confusion matrix of the classification model of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane model with three solubility outcomes

		Reference Data					
		Acetone			Acetonitrile		
		HS	S	PI	HS	S	PI
Predicted Data	H. S	11	11	3	11	6	2
	Soluble	10	6	10	8	12	6
	PI	5	13	14	2	8	27

		Reference Data					
		Ethyl acetate			1,4-dioxane		
		HS	S	PI	HS	S	PI
Predicted Data	H. S	25	10	3	2	4	2
	Soluble	8	20	12	2	0	6
	PI	3	8	12	0	2	19

From Table 6.15, it can be observed from the performance metrics of the cross validated model that 1-butanol, ethanol and acetonitrile had the least misclassified outcomes out of the seven compounds whilst acetone, ethyl acetate and especially 1,4-dioxane gave the worst classification of the compounds. This overall trend is similar to the regression models and binary classification models. From the confusion matrices (Table 6.16 and Table 6.17), it can be observed that acetone and 1, 4-dioxane models failed to accurately classify the outcome Soluble. Instead, most of the soluble compounds in acetone were classified as either being highly soluble or practically insoluble. Similarly, the majority of the compounds labelled as soluble were misclassified as being practically insoluble in 1, 4-dioxane model.

Following the 10-fold cross-validation, a final model for each solvent was then trained on the training set and tested on the remaining 20% external validation set. The results are shown in Table 6.18. The confusion matrices are also shown to help interpret the classification of the seven organic solvents (Table 6.19 and Table 6.20).

Table 6.18 Predictive performance on the external validation set of the RF model on seven organic solvents with three qualitative solubility outcomes

			Sensitivity			Specificity		
Solvent	ACC (%)	Kappa	HS	S	PI	HS	S	PI
Ethanol	64.86	0.460	0.667	0.647	0.625	0.840	0.800	0.808
Methanol	66.67	0.460	0.500	0.700	0.714	1.000	0.700	0.786
1-butanol	82.61	0.720	0.800	1.000	0.625	1.000	0.692	0.877
Acetonitrile	80.95	0.702	1.000	0.714	0.818	0.833	0.929	0.826
Acetone	47.62	0.260	0.750	0.308	0.750	0.824	0.875	0.756
Ethyl acetate	60.00	0.343	0.539	0.600	1.000	0.833	0.600	0.8
1,4-dioxane	70.00	0.167	0.000	0.000	1.000	0.889	1.000	0.769

Table 6.19 Confusion matrix for prediction on the external validation set of alcohol models with three solubility outcomes

		Reference Data								
		Ethanol			Methanol			1-Butanol		
		HS	S	PI	HS	S	PI	HS	S	PI
Predicted Data	H. S	4	1	1	8	0	0	5	0	0
	Soluble	8	29	3	2	6	3	1	7	2
	PI	1	4	5	1	2	8	0	0	8

Table 6.20 Confusion matrix for prediction on an external validation set of acetone, acetonitrile, ethyl acetate and 1, 4-dioxane models with three solubility outcomes

		Reference Data					
		Acetone			Acetonitrile		
		HS	S	PI	HS	S	PI
Predicted Data	H. S	2	1	1	3	2	1
	Soluble	3	10	0	0	5	1
	PI	0	3	1	0	0	9
		Reference Data					
		Ethyl acetate			1,4-dioxane		
		HS	S	PI	HS	S	PI
Predicted Data	H. S	9	2	0	1	0	1
	Soluble	5	5	0	2	2	2
	PI	0	3	3	0	2	1

For the external validation set, the classification models predicted the outcomes accurately for 1-butanol, ethanol and methanol. The acetonitrile model performed slightly better than on the cross-validation set. Compounds labelled as *practically insoluble* were largely misclassified as *soluble* in the acetone model. Similarly, for ethyl acetate and 1, 4-dioxane models, the majority of the compounds labelled as *soluble* were misclassified as *highly soluble* and *practically insoluble,* respectively. This was not surprising given that both the models had a poor performance in regression and in binary classification models too. It was interesting to observe that, with the three defined solubility regions, the compounds labelled at the extreme ends of the solubility spectrum *i.e. highly soluble* and *practically insoluble* were generally predicted correctly or misclassified as *soluble.* In other words, the two ends of the solubility spectrum were well-separated by the classification model. If the 'middle' class of *soluble* were to be treated as an unknown category that required experimental confirmation, this methodology could feasibly be used to reduce the number of solubility tests that must be performed, using the high accuracy of the *highly soluble* and *practically insoluble* classes to filter out unnecessary combinations and only testing anything predicted as *soluble.*

6.4 Summary

This study reports the application of random forest to predict non-aqueous solubility of 247 unique drug and drug-like molecules in seven commonly used solvents. The maximum number of solubility data points curated for this study for any one solvent was 181 (for ethanol) which is a relatively small number when compared to aqueous solubility datasets, some of which have over 1000 data points. Furthermore, solubility data was taken at a specific temperature which limited the number of data points that could be utilised. However, in spite of this, both regression and classification models built using physicochemical molecular descriptors were successfully implemented to predict non-aqueous solubility. As indicated by the results, the RF regression models were able to predict solubility with relatively low values of RMSE and MAE whilst maintaining high R^2 values in ethanol, 1-butanol and methanol. The regression models however failed to predict accurately in ethyl acetate, acetone, acetonitrile and 1, 4-dioxane. Equivalent regression models built using MACCS fingerprints showed relatively poor performance compared to the ones built using numerical descriptors using MOE. Similarly, RF was successfully implemented in developing classification models capable of accurately classifying which compounds lie in the respective solubility regions. A binary classification response with set outcomes *soluble* and *insoluble* based on the Merck Index was initially used. The binary classification models worked well in ethanol, methanol and acetone whilst failing to accurately classify compounds in 1-butanol, ethyl acetate, acetonitrile and 1,4-dioxane. The binary response was then replaced by three solubility regions to be able to identify compounds on the far extremes of the solubility spectrum i.e. *highly soluble* and *practically insoluble* regions. RF classification models trained using three outcomes were accurate for the alcohol models (ethanol, methanol and 1-butanol) and acetonitrile models but failed to accurately predict on ethyl acetate, acetone and 1,4-dioxane. Predicting the solubility of compounds in ethyl acetate failed for both regression and classification models which could indicate that the calculated descriptors were incomplete and thus the model failed to learn the underlying patterns between the descriptors and the outcomes within the training set in order to make accurate predictions in the external validation set or that the algorithm implemented was unable to identify the underlying trends. Similarly, 1, 4-dioxane model was relatively poor due to its small sample size (47 samples) and the descriptors could not fully explain the targeted outcomes.

It is worth noting that, if all cases predicted as the central class (*soluble*) in the three-way classification setup were to be considered as unknowns requiring further testing, the remaining classes on either side of the solubility spectrum were rarely misclassified as one another. This means that this method could be used to filter out many solvent-solute combinations that are very high or very low, reducing the number required to be taken forward to the experimental screening.

This study covers a wider range of industrially acceptable solvents than what is reported in the literature and demonstrates that machine learning can be used as a predictive tool in recommending solvents for solubility studies during crystallisation process design. Perhaps most importantly, this method is purely based on literature data and as such it can help reduce experimental cost and time needed in the laboratory.

Chapter 7.　Conclusion and Future work

7.1 Conclusion

The work performed in this thesis demonstrates the ability of machine learning algorithms in predicting various essential parameters required at the initial stages of crystallisation process design such as crystallisation outcomes and crystal habit which helps reduce experimental time, cost and wastage. The research not only successfully implemented predictive machine learning models but also provided an understanding of the relationship between the calculated 2D and 3D numerical descriptors and the investigated properties. A solvent selection tool was developed to help recommend solvents for rapid and efficient screening at early stage of a crystallisation process with an average prediction accuracy of ~75%. An in-house solubility database consisting of 247 unique drug and drug-like compounds was utilised to predict non-aqueous solubility in seven common organic solvents (methanol, ethanol, 1-butanol, acetone, acetonitrile, ethyl acetone, 1, 4-dioxane) with an average predictive accuracy of ~74 %.

Cooling crystallisation was utilised throughout the research to collect experimental data required for training the machine learning algorithm. Overall, the experimental and predictive computational methods have provided an understanding of the various influences of organic solvents in the crystallisation process.

The conclusions for each of the study done in the thesis are summarised as follows:

Chapter4: Implementing the RF algorithm for predicting crystallisation outcomes and crystal habit of paracetamol in a diverse range of organic solvents

The studies investigated the implementation of machine learning algorithms in the prediction of crystallisation outcomes of paracetamol on 94 organic solvents and the crystal habit on 44 solvents on which paracetamol crystallised out. The crystallisation outcomes obtained from controlled cooling crystallisation experiments were qualitatively categorised as *crystallised out*, *non-nucleated*, *practically insoluble* and *degradation*. The models developed in this study can help provide valuable assistance in making quick decisions about suitable solvents for the selected drug at the initial stage of a crystallisation process by providing initial information about the drug's crystallisability and crystal habit.

The results obtained from the unsupervised machine learning technique, PCA, provided an efficient visualisation of dominant patterns in the solvent chemical space. The experimental outcomes when painted over the first two principal components provided two distinctive clustering: a cluster of the solvents showing outcomes *practically insoluble* and combined clustering of *crystallised out, non-nucleated and degradation*. The cluster pattern observed from the unsupervised ML model demonstrated solubility behaviour as the solvents in combined clusters of *crystallised out, non-nucleated and degradation* all solubilises paracetamol while the solvents in the other cluster did not. The PCA model also provided further information on the distribution of the solvents and its clusters according to its polarity and functional groups in the chemical space. Similarly, the RF classification model successfully predicted the solubility outcomes but failed to correctly classify and predict the behaviour of paracetamol in solvents where it failed to nucleate i.e. *non-nucleated* outcome and on those that it degraded in i.e. *degradation*. Furthermore, the number of solvents that paracetamol degraded were very few in the dataset and had no influence in the prediction of the model. Thus, solvents showing degradation were omitted from the dataset. Overall, the classification accuracy of the trained model was observed to be 77.33% with kappa as 0.594 and prediction accuracy on the validation sets of 73.68% with kappa as 0.583. The classification and prediction accuracy came mostly from the two well-represented classes i.e. *crystallised out* and *practically insoluble*. The Cohen's kappa indicated the presence of misclassification in the model with a moderate agreement value of between 0.55 and 0.6. Stratified sampling and One-vs-one binarisation techniques when applied further provided a better understanding that the trained models were capable of only predicting whether paracetamol is soluble in the solvent or not. The results presented the limitation of the selected molecular descriptors to provide enough information to the RF models in order to accurately classify and predict solvents where paracetamol failed to nucleate and degrade. This is because nucleation is not only dependent upon the physicochemical properties but is also influenced by supersaturation, temperature and process control parameters. However, as the chapter presented that the ML algorithms built using these same molecular descriptors were able to accurately classify the solvents into two distinct clusters, i.e. solvent clusters where paracetamol solubilised in and the clusters containing solvents where paracetamol was found to be insoluble in. This indicated that the trained classification models could provide a valuable tool for rapidly predicting solubility of compounds and thus aid in solvent selection.

The trained ML models were also unable to predict the crystal habit of paracetamol in 44 solvents. This could be due to the small and highly unbalanced dataset size of the crystallised paracetamol. Small dataset sizes decrease statistical power as it is incapable of fully detecting the relationship between the outcomes and the descriptors.

Overall, as mentioned earlier the built machine learning models provided great potential as a rational solvent selection tool for predicting the solubilisation behaviour of the drug (i.e. whether the drug will dissolve in the respective solvent and which solvents the drug was insoluble) with a good degree of confidence. The model, however, failed to provide relevant information on the drug's nucleation and degradation as well as crystal habit.

Chapter 5: Development of a rapid and efficient solvent selection tool

The work in this chapter focused on utilising the ability of the ML algorithm and the calculated molecular descriptors of the solvents to predict solubility in order to develop a rapid and efficient solvent selection tool for crystallisation process design. The method explored reducing the number of screening experiments in the laboratory and the number of materials used for screening studies by investigating the optimum number of the sample sizes required for building a robust random forest model. Data mining approaches utilised in the study also help extract the frequent and important solvents which when selected on the training set would provide an accurate predictive model. The RF solvent selection approach, when combined with the Strathclyde24 Solvent cluster map, proved an even more effective solvent selection tool for screening with an average predictive accuracy above 80% (Johnston et al., 2017). The developed approach for solvent selection provided an efficient way of eliminating unsuitable solvents at early stage of the crystallisation process. Furthermore, solvents in which the compounds were predicted as insoluble by the model could be recommended as wash solvents or as an antisolvent in the crystallisation process design.

Chapter 6: RF models to predict non-aqueous solubility of the drug and drug-like compounds in organic solvents

The work performed in the chapter focused on the development of random forest models purely based on collated literature data to predict the non-aqueous solubility of the drug and drug-like compounds on common organic solvents. The trained model was the first of its kind as most of the predictions in literature is done on aqueous solubility. A dataset was curated consisting of solubility data of 247 unique compounds in seven organic solvents. Both the regression and classification model built on calculated 2D and 3D molecular properties provided good predictions of non-aqueous solubility on alcohols except 1,4-dioxane and ethyl acetate. The size of dataset curated for 1, 4-dioxane was limited in number, and thus the RF model struggled to predict solubility accurately. However, in the case of ethyl acetate, even though the number of data set size was similar to other solvents, the model failed to make predictions and find trends between the descriptors and the solubility outcomes. The solubility predictions made by the trained random forest model were found to be comparable with similar studies done in literature

7.2 Future works

7.2.1 Chapter 4

The research on predicting crystallisation outcomes and crystal habit were performed on 94 solvents for only one drug compound. The chosen API, Paracetamol, was a good starting point as a lot of research has been heavily performed on the compound. Furthermore, paracetamol is a stable API and does not change polymorphic form in the set experimental conditions. It is also ideal for crystal habit studies as it is reported its crystal habit is dependent upon the solvent used. However, more crystallisation experiments performed on a number of diverse drug compounds could help better understanding of the crystallisation behavior on the drug compounds in a diverse range of solvents. This could also in turn provide more data for the machine learning algorithms and a thus better understanding of the relationship between the molecular descriptors and the targeted outcomes. The choice of calculated molecular descriptors failed to sufficiently correlate with the nucleation and degradation outcomes of the drug compound. Investigating and including descriptors that define these outcomes better would help improve the quality of the machine learning model. The number of solvents in which paracetamol was crystallised in was few in number thus resulting in a small dataset size for the ML model. This could potentially be solved by crystallising various drug compounds on one solvent using constant cooling crystallisation conditions rather than crystallising one drug compound on a diverse range of solvents. As the number of drug compounds is in the range of hundreds of thousands compared to the number of solvents, this could help solve the issues with dataset size. Also, utilizing databases such as Cambridge Structural Database (CSD) in the future to data mine crystal habits to develop predictive machine learning models would be the better route to build reliable models as the CSD database in 2016 consists of over 800,000 crystal structure (Groom, Bruno, Lightfoot, & Ward, 2016). Cooling crystallisation was the preferred crystallisation technique utilised in this study. The addition of other techniques such as evaporation, antisolvent crystallisation, solvent diffusion etc would help generate more data and thus further understand the crystallisation outcomes and habit of drug compounds in various solvents.

Implementing robotic platforms such as the Zinsser automated platform for high throughput screening studies could help generate a large number of experimental data points on a number of diverse organic solvents in a more accurate and efficient way which in turn will help increase the number of data points for the ML model.

7.2.2 Chapter 5

The solvent selection tool was tested on paracetamol, carbamazepine and carvedilol. The next step would be to investigate the efficiency of the solvent selection tool on salts and co-crystals. This would add cover a diverse range of compounds and add more value to the solvent selection tool. The solvent cluster map consisted of only 63 solvents. Adding more solvents to the cluster map would also help cover a wider solvent chemical space. The result from the current study presented the random forest model capable of accurately predicting whether the drug was soluble or insoluble in the respective solvent at two temperature points which helped filter out the unsuitable solvents for the crystallisation process. Safety, toxicology/health and environment criteria utilised by various industries in their solvent selection tool can be added either as descriptors during model training or later to further filter out the solvents. Doing so can add more value to the solvent selection tool (Alder et al., 2016; Prat et al., 2013). Similarly, properties such as polymorphic change of the drug in the respective solvent as well as a chemical reaction, oiling out, degradation etc were not included in the solvent selection tool and if included could help recommend suitable solvents with high accuracy. The solvent selection tool was developed for pure solvents. However, this approach can also be utilised on solvent mixtures by incorporating experimental outcomes and relevant molecular descriptors of solvent mixtures in the training set.

7.2.3 Chapter 6

Random forest models were successfully developed to predict the solubility of drug and drug like compounds. Implementing other supervised and unsupervised machine learning algorithms or even combining with RF models would help investigate the predictive accuracy. A comparison of machine learning algorithms has been done previously in the literature (Ahmad, Mourshed, & Rezgui, 2017; Banfield et al., 2007). Even though the random forest was successful in predicting non-aqueous solubility as well as the RF model being easy to train, flexible and highly accurate, it is to be noted that there is no perfect algorithm and that all algorithms are dependent upon the quality and quantity of the data. Similarly, besides MACCS fingerprint, various other fingerprints such as circular Morgan fingerprints, RDKit fingerprints and e-state count fingerprints and numerical descriptors from RDKit could also be implements. Furthermore, clustering of the drug structures based on their molecular structural fingerprint or numerical descriptors using clustering algorithms could aid in selecting suitable drug compounds for developing accurate RF models. 247 unique compounds were present in the dataset curated in this chapter. More data under the set parameters such as stable polymorphic form, same temperature and same solubility measurement techniques could be mined from literature and electronic laboratory notebooks. The same method could be applied to the limited number of solubility databases available in the literature and compared with the developed RF model in this study.

Chapter 8. Publications

i. Enabling precision manufacturing of active pharmaceutical ingredients: workflow for seeded cooling continuous crystallisation

C. J. Brown, T. McGlone, S. Yerdelen, V. Srirambhatla, F. Mabbott, R. Gurung, M. L. Briuglia, B. Ahmed, H. Polyzois, J. McGinty, F. Perciballi, D. Fysikopoulos, P. MacFhionnghaile, H. Siddique, V. Raval, T. S. Harrington, A. D. Vassileiou, M. Robertson, E. Prasad, A. Johnston, B. Johnston, A. Nordon, J. S. Srai, G. Halbert, J. H. ter Horst, C. J. Price, C. D. Rielly, J. Sefcik and A. J. Florence, Molecular Systems Design & Engineering 2018, 3, 518-549.

ii. Combined Chemoinformatics Approach to Solvent Library Design Using clusterSim and Multidimensional Scaling

A. Johnston, R. Bhardwaj-Miglani, R. Gurung, A. D. Vassileiou, A. J. Florence and B. F. Johnston, J Chem Inf Model 2017, 57, 1807-1815.

Chapter 9.　References

9.1 References

1. Aamir, E., Nagy, Z. K., Rielly, C. D., Kleinert, T., & Judat, B. (2009). Combined Quadrature Method of Moments and Method of Characteristics Approach for Efficient Solution of Population Balance Models for Dynamic Modeling and Crystal Size Distribution Control of Crystallization Processes. *Industrial & Engineering Chemistry Research, 48*(18), 8575-8584. doi:10.1021/ie900430t

2. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459. doi:doi:10.1002/wics.101

3. Abdul Mudalip, S. K., Abu Bakar, M. R., Jamal, P., & Adam, F. (2013). Solubility and Dissolution Thermodynamic Data of Mefenamic Acid Crystals in Different Classes of Organic Solvents. *Journal of Chemical & Engineering Data, 58*(12), 3447-3452. doi:10.1021/je400714f

4. Acree, W. E. (2014). IUPAC-NIST Solubility Data Series. 102. Solubility of Nonsteroidal Anti-inflammatory Drugs (NSAIDs) in Neat Organic Solvents and Organic Solvent Mixtures. *Journal of Physical and Chemical Reference Data, 43*(2), 023102. doi:10.1063/1.4869683

5. Agrawal, S., & Paterson, A. (2015). *Secondary Nucleation: Mechanisms and Models* (Vol. 202).

6. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77-89. doi:https://doi.org/10.1016/j.enbuild.2017.04.038

7. Alder, C. M., Hayler, J. D., Henderson, R. K., Redman, A. M., Shukla, L., Shuster, L. E., & Sneddon, H. F. (2016). Updating and further expanding GSK's solvent sustainability guide. *Green Chemistry, 18*(13), 3879-3890. doi:10.1039/C6GC00611F

8. Ali, H. S. M., York, P., Blagden, N., Soltanpour, S., Acree, W. E., & Jouyban, A. (2010). Solubility of Budesonide, Hydrocortisone, and Prednisolone in Ethanol + Water Mixtures at 298.2 K. *Journal of Chemical & Engineering Data, 55*(1), 578-582. doi:10.1021/je900376r

9. Ali, S. J., & Rajini, P. S. (2012). Elicitation of Dopaminergic Features of Parkinson's Disease in C. elegans by Monocrotophos, an Organophosphorous Insecticide. *Cns & Neurological Disorders-Drug Targets, 11*(8), 993-1000.

10. Alpaydin, E. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*: The MIT Press.

11. Alrashood, S. T. (2016). Chapter Three - Carbamazepine. In H. G. Brittain (Ed.), *Profiles of Drug Substances, Excipients and Related Methodology* (Vol. 41, pp. 133-321): Academic Press.

12. Alshehri, S., & Shakeel, F. (2017). Solubility measurement, thermodynamics and molecular interactions of flufenamic acid in different neat solvents. *Journal of Molecular Liquids, 240*, 447-453. doi:https://doi.org/10.1016/j.molliq.2017.05.105

13. Altman, D. G. (1999). *Practical statistics for medical research*. Boca Raton, Fla.: Chapman & Hall/CRC.

14. Alves, K. C. M., Condotta, R., & Giulietti, M. (2001). Solubility of Docosane in Heptane. *Journal of Chemical & Engineering Data, 46*(6), 1516-1519. doi:10.1021/je010032h

15. Andrade, C. H., Pasqualoto, K. F. M., Ferreira, E. I., & Hopfinger, A. J. (2010). 4D-QSAR: perspectives in drug design. *Molecules (Basel, Switzerland), 15*(5), 3281-3294. doi:10.3390/molecules15053281

16. Andrews, D. F., & Herzberg, A. M. (1985). Iris Data *Data: A Collection of Problems from Many Fields for the Student and Research Worker* (pp. 5-8). New York, NY: Springer New York.

17. Aniya, V., De, D., Mohammed, A. M., Thella, P. K., & Satyavathi, B. (2017). Measurement and Modeling of Solubility of para-tert-Butylbenzoic Acid in Pure and Mixed Organic Solvents at Different Temperatures. *Journal of Chemical & Engineering Data, 62*(4), 1411-1421. doi:10.1021/acs.jced.6b00965

18. Anyfamis, D., Karagiannopoulos, M., Kotsiantis, S., & Pintelas, P. (2007). Robustness of learning techniques in handling class noise in imbalanced datasets. *Artificial Intelligence and Innovations 2007: From Theory to Applications*, 21-+.

19. Arlin, J. B., Price, L. S., Price, S. L., & Florence, A. J. (2011). A strategy for producing predicted polymorphs: catemeric carbamazepine form V. *Chemical Communications, 47*(25), 7074-7076. doi:10.1039/c1cc11634g

20. Badman, C., & Trout, B. L. (2015). Achieving Continuous Manufacturing May 20–21 2014 Continuous Manufacturing Symposium. *Journal of Pharmaceutical Sciences, 104*(3), 779-780. doi:https://doi.org/10.1002/jps.24246

21. Baker, R. E., Pena, J. M., Jayamohan, J., & Jerusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol Lett, 14*(5). doi:10.1098/rsbl.2017.0660

22. Balakin, K. V., Savchuk, N. P., & Tetko, I. V. (2006). In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Current medicinal chemistry, 13*(2), 223-241.

23. Baluja, S., Bhalodia, R., Gajera, R., Vekariya, N., & Bhatt, M. (2009). Solubility of Difloxacin in Acetone, Methanol, and Ethanol from (293.15 to 313.15) K. *Journal of Chemical & Engineering Data, 54*(3), 1091-1093. doi:10.1021/je800742d

24. Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2007). A comparison of decision tree ensemble creation techniques. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 29*(1), 173-180. doi:Doi 10.1109/Tpami.2007.250609

25. Baumann, M., & Baxendale, I. R. (2015). The synthesis of active pharmaceutical ingredients (APIs) using continuous flow chemistry. *Beilstein journal of organic chemistry, 11*, 1194-1219. Retrieved from http://europepmc.org/abstract/MED/26425178

26. http://europepmc.org/articles/PMC4578405?pdf=render

27. http://europepmc.org/articles/PMC4578405

28. https://doi.org/10.3762/bjoc.11.134 doi:10.3762/bjoc.11.134

29. Baurin, N., Mozziconacci, J.-C., Arnoult, E., Chavatte, P., Marot, C., & Morin-Allory, L. (2004). 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *Journal of Chemical Information and Computer Sciences, 44*(1), 276-285.

30. Beattie, K., Phadke, G., & Novakovic, J. (2013). Chapter Four - Carvedilol. In H. G. Brittain (Ed.), *Profiles of Drug Substances, Excipients and Related Methodology* (Vol. 38, pp. 113-157): Academic Press.

31. Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research, 5*, 1089-1105.

32. Benoit, K. (2011). *Linear Regression Models with Logarithmic Transformations*.

33. Berkovitch-Yellin, Z. (1985). Toward an ab initio derivation of crystal morphology. *Journal of the American Chemical Society, 107*(26), 8239-8253. doi:10.1021/ja00312a070

34. Berry, K. J., & Mielke, P. W. (1988). A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters.

Educational and Psychological Measurement, 48(4), 921-933. doi:10.1177/0013164488484007

35. Bhardwaj, R. M. (2016). *Control and Prediction of Solid-State of Pharmaceuticals: Experimental and Computational Approaches*: Springer International Publishing.
36. Bhat, A. U., Merchant, S. S., & Bhagwat, S. S. (2008). Prediction of melting points of organic compounds using extreme learning machines. *Industrial & Engineering Chemistry Research, 47*(3), 920-925.
37. Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research, 13*(Apr), 1063-1095.
38. Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197-227. doi:10.1007/s11749-016-0481-7
39. Boyd, S. (2005). Molecular operating environment. *Chemistry World, 2*(9), 66-66.
40. Bramer, M. (2013). Avoiding Overfitting of Decision Trees. In M. Bramer (Ed.), *Principles of Data Mining* (pp. 121-136). London: Springer London.
41. Breiman, L. (2001a). Random forests. *Machine Learning, 45*(1), 5-32. doi:Doi 10.1023/A:1010933404324
42. Breiman, L. (2001b). Random Forests. *Machine Learning, 45*. doi:10.1023/a:1010933404324
43. Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2015). *Breiman and Cutler's random forests for classification and regression*.
44. Brijain, M., Patel, R., Kushik, M., & Rana, K. (2014). A survey on decision tree algorithm for classification.
45. Brittain, H. G. (2009). *Polymorphism in pharmaceutical solids* (2nd ed.). New York: Informa Healthcare.

46. Brittain, H. G. (2013). Profiles of Drug Substances, Excipients, and Related Methodology Preface. *Profiles of Drug Substances, Excipients, and Related Methodology, Vol 38, 38*, Xi-Xi.

47. Brown, C. J., McGlone, T., Yerdelen, S., Srirambhatla, V., Mabbott, F., Gurung, R., . . . Florence, A. J. (2018). Enabling precision manufacturing of active pharmaceutical ingredients: workflow for seeded cooling continuous crystallisations. *Molecular Systems Design & Engineering, 3*(3), 518-549. doi:10.1039/C7ME00096K

48. Brown, F. K. (1998). Chemoinformatics, what it is and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry, 33*, 375-384.

49. Bunin, B. A., Siesel, B., Morales, G., & Bajorath, J. (2006). *Chemoinformatics: Theory, Practice, & Products*: Springer Netherlands.

50. Buonaiuto, M. A., & Lang, A. S. (2015). Prediction of 1-octanol solubilities using data from the Open Notebook Science Challenge. *Chemistry Central Journal, 9*(1), 50.

51. Cannon, E. O., Bender, A., Palmer, D. S., & Mitchell, J. B. O. (2006). Chemoinformatics-Based Classification of Prohibited Substances Employed for Doping in Sport. *Journal of Chemical Information and Modeling, 46*(6), 2369-2380. doi:10.1021/ci0601160

52. Cao, D.-S., Xu, Q.-S., Hu, Q.-N., & Liang, Y.-Z. J. B. (2013). ChemoPy: freely available python package for computational biology and chemoinformatics. *29*(8), 1092-1094.

53. Cao, D.-S., Xu, Q., Hu, Q., & Liang, Y.-Z. (2013). *manual for chemopy*.

54. Cao, X.-X., Tong, R.-J., Zhao, Y., Lv, T.-T., Song, Y., & Yao, J.-C. (2012). Determination and Correlation of Pyridazin-3-amine Solubility in Eight Organic Solvents at Temperatures Ranging from (288.05 to 333.35) K. *Journal of Chemical & Engineering Data, 57*(8), 2360-2366. doi:10.1021/je300517q

55. Caruana, R., & Niculescu-Mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. Paper presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA.

56. Cattell, R. B. (1983). Citation Classic - the Scree Test for the Number of Factors. *Current Contents/Social & Behavioral Sciences*(5), 16-16.

57. Chadha, R., Arora, P., & Bhandari, S. (2012). Polymorphic Forms of Lamivudine: Characterization, Estimation of Transition Temperature, and Stability Studies by Thermodynamic and Spectroscopic Studies. *ISRN Thermodynamics, 2012*, 8. doi:10.5402/2012/671027

58. Chahal, H. (2013). *ID3 Modification and Implementation in Data Mining* (Vol. 80).

59. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250. doi:10.5194/gmd-7-1247-2014

60. Chang, Q.-L., Li, Q.-S., Wang, S., & Tian, Y.-M. (2007). Solubility of Phenacetinum in Methanol, Ethanol, 1-Propanol, 1-Butanol, 1-Pentanol, Tetrahydrofuran, Ethyl Acetate, and Benzene between 282.65 K and 333.70 K. *Journal of Chemical & Engineering Data, 52*(5), 1894-1896. doi:10.1021/je700209v

61. Chastrette, M., Rajzmann, M., Chanon, M., & Purcell, K. F. (1985). Approach to a General Classification of Solvents Using a Multivariate Statistical Treatment of Quantitative Solvent Parameters. *Journal of the American Chemical Society, 107*(1), 1-11. doi:DOI 10.1021/ja00287a001

62. Chawla, N. V., Bowyer, K. W., & Hall, L. O. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res, 16*.

63. Chen, C., & Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*.

64. Chen, G., Chen, J., Cheng, C., Cong, Y., Du, C., & Zhao, H. (2017). Solubility determination and thermodynamic modelling of 2-amino-5-methylthiazole in eleven organic solvents from T=(278.15 to 313.15)K and mixing properties of solutions. *Journal of Molecular Liquids, 232*, 226-235. doi:https://doi.org/10.1016/j.molliq.2017.02.084

65. Chen, G., Chen, J., Cheng, C., Cong, Y., Jian, P., & Zhao, H. (2017). Solubility modelling and dissolution properties of 5-phenyltetrazole in thirteen mono-solvents and liquid mixtures of (methanol+ethyl acetate) at elevated temperatures. *The Journal of Chemical Thermodynamics, 112*, 114-121. doi:https://doi.org/10.1016/j.jct.2017.04.019

66. Chen, G., Chen, J., Jian, P., & Zhao, H. (2017). Solubility modelling and mixing properties of biologically active 5-amino-3-methyl-1-phenylpyrazole in ten neat solvents from 283.15K to 318.15K. *Journal of Molecular Liquids, 240*, 532-541. doi:https://doi.org/10.1016/j.molliq.2017.05.124

67. Chen, J., Chen, G., Cong, Y., Du, C., & Zhao, H. (2017). Solubility of 2-isopropylimidazole in nine pure organic solvents and liquid mixture of (methanol+ethyl acetate) from T=(278.15 to 313.15)K: Experimental measurement and thermodynamic modelling. *The Journal of Chemical Thermodynamics, 107*, 133-140. doi:https://doi.org/10.1016/j.jct.2016.12.028

68. Chen, J., Sarma, B., Evans, J. M. B., & Myerson, A. S. (2011). Pharmaceutical Crystallization. *Crystal Growth & Design, 11*(4), 887-895. doi:10.1021/cg101556s

69. Chen, W. L. (2006). Chemoinformatics: Past, Present, and Future†. *Journal of Chemical Information and Modeling, 46*(6), 2230-2255. doi:10.1021/ci060016u

70. Cheng, Y., Wang, D., Zhang, Z., & Wang, Z. (2015). Solubility and solution thermodynamics of rhein in eight pure solvents from (288.15 to 313.15) K. *RSC Advances, 5*(98), 80548-80552. doi:10.1039/C5RA17881A

71. Chirico, N., & Gramatica, P. (2011). Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *Journal of Chemical Information and Modeling, 51*(9), 2320-2335. doi:10.1021/ci200211n

72. Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). Unsupervised Learning: Association Rules. In K. J. Cios, R. W. Swiniarski, W. Pedrycz, & L. A. Kurgan (Eds.), *Data Mining: A Knowledge Discovery Approach* (pp. 289-306). Boston, MA: Springer US.

73. Cleophas, T. J. (2012). Machine Learning in Pharmaceutical research: Data clustering, Why so and how so. *Journal of Pharmaceutical Sciences, Volume 4*(11), 1964-1969.

74. Corcoran, K. (2006). Statistics for evidence-based practice and evaluation. *Research on Social Work Practice, 16*(5), 546-547. doi:10.1177/1049731506291442

75. Cruciani, G., Crivori, P., Carrupt, P. A., & Testa, B. (2000). Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *Journal of Molecular Structure: THEOCHEM, 503*(1), 17-30. doi:https://doi.org/10.1016/S0166-1280(99)00360-7

76. Cruz-Monteagudo, M., Borges, F., Cordeiro, M., Helguera, A. M., Tejera, E., Paz, Y. M. C., . . . Perez-Castillo, Y. (2017). Chemoinformatics Profiling of the Chromone Nucleus as a MAO-B/A2AAR Dual Binding Scaffold. *Curr Neuropharmacol, 15*(8), 1117-1135. doi:10.2174/1570159X15666170116145316

77. Curzons, A. D., Constable, D. C., & Cunningham, V. L. (1999). Solvent selection guide: a guide to the integration of environmental, health and safety criteria into the selection of solvents. *Clean Products and Processes, 1*(2), 82-90. doi:10.1007/s100980050014

78. Dash, M., Liu, H., & Yao, L. (1997). Dimensionality reduction of unsupervised data. *Ninth Ieee International Conference on Tools with Artificial Intelligence, Proceedings*, 532-539. doi:Doi 10.1109/Tai.1997.632300

79. Defossemont, G., Randzio, S., & Legendre, B. (2007). Identification of an enantiotropic system with hindered multiphase transitions. *Journal of Thermal Analysis and Calorimetry, 89*(3), 751-755. doi:10.1007/s10973-007-8397-9

80. Delaney, J. S. (2005). Predicting aqueous solubility from structure. *Drug Discovery Today, 10*(4), 289-295. doi:Pii S1359-6446(04)03365-3

81. Doi 10.1016/S1359-6446(04)03365-3

82. DELGADO, D. R., R. HOLGUIN, A., & MARTÍNEZ, F. (2012). SOLUTION THERMODYNAMICS OF TRICLOSAN AND TRICLOCARBAN IN SOME VOLATILE ORGANIC SOLVENTS. *Vitae, 19*, 79-92.

83. Díaz-Uriarte, R., & Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinforma, 7*. doi:10.1186/1471-2105-7-3

84. Ding, L., Wang, B., Wang, F., Dong, J., Zhou, G., & Li, H. (2017). Measurement and correlation of the solubility of dipyrone in ten mono and water+ethanol mixed solvents at temperatures from (293.15 to 332.85) K. *Journal of Molecular Liquids, 241*, 742-750. doi:https://doi.org/10.1016/j.molliq.2017.06.072

85. Diorazio, L. J., Hose, D. R. J., & Adlington, N. K. (2016). Toward a More Holistic Framework for Solvent Selection. *Organic Process Research & Development, 20*(4), 760-773. doi:10.1021/acs.oprd.6b00015

86. Domańska, U., Pobudkowska, A., & Pelczarska, A. (2011). Solubility of Sparingly Soluble Drug Derivatives of Anthranilic Acid. *The Journal of Physical Chemistry B, 115*(11), 2547-2554. doi:10.1021/jp109905r

87. Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the Acm, 55*(10), 78-87. doi:10.1145/2347736.2347755

88. Dong, X., Cao, Y., Lin, H., Yao, Y., Guo, Y., Wang, T., . . . Wu, Z. (2017). Solubilities of formononetin and daidzein in organic solvents: Effect of molecular structure and interaction on solvation process. *Journal of Molecular Liquids, 231*, 542-554. doi:https://doi.org/10.1016/j.molliq.2017.02.051

89. El-Badry, M., Haq, N., Fetih, G., & Shakeel, F. (2014). Measurement and Correlation of Tadalafil Solubility in Five Pure Solvents at (298.15 to 333.15) K. *Journal of Chemical & Engineering Data, 59*(3), 839-843. doi:10.1021/je400982r

90. Engel, T., & Gasteiger, J. (2018). *Chemoinformatics- Basic Concepts and Methods*: Wiley.

91. Espeau, P., Ceolin, R., Tamarit, J. L., Perrin, M. A., Gauchi, J. P., & Leveiller, F. (2005). Polymorphism of paracetamol: relative stabilities of the monoclinic and orthorhombic phases inferred from topological pressure-temperature and temperature-volume phase diagrams. *J Pharm Sci, 94*(3), 524-539. doi:10.1002/jps.20261

92. Fahlman, B. D. (2002). Chemical vapor deposition of carbon nanotubes - An experiment in materials chemistry. *Journal of Chemical Education, 79*(2), 203-206. doi:DOI 10.1021/ed079p203

93. Fang, J., Zhang, M., Zhu, P., Ouyang, J., Gong, J., Chen, W., & Xu, F. (2015). Solubility and solution thermodynamics of sorbic acid in eight pure organic solvents. *The Journal of Chemical Thermodynamics, 85*, 202-209. doi:https://doi.org/10.1016/j.jct.2015.02.004

94. Fathi-Azarbayjani, A., Abbasi, M., Vaez-Gharamaleki, J., & Jouyban, A. (2016). Measurement and correlation of deferiprone solubility: Investigation of solubility parameter and application of van't Hoff equation and Jouyban–Acree model. *Journal of Molecular Liquids, 215*, 339-344. doi:https://doi.org/10.1016/j.molliq.2015.12.005

95. Faulon, J.-L., & Bender, A. (2010). *Handbook of chemoinformatics algorithms*. Boca Raton, FL: Chapman & Hall/CRC.

96. Fernandez-de Gortari, E., Garcia-Jacas, C. R., Martinez-Mayorga, K., & Medina-Franco, J. L. (2017). Database fingerprint (DFP): an approach to represent molecular databases. *Journal of Cheminformatics, 9*. doi:ARTN 9

97. 10.1186/s13321-017-0195-1

98. Feuerstein, G. Z., & Ruffolo, R. R. (1995). Carvedilol, a Novel Multiple Action Antihypertensive Agent with Antioxidant Activity and the Potential for Myocardial and Vascular Protection. *European Heart Journal, 16*, 38-42.

99. Florence, A. J. (2016). Control and Prediction of Solid-State of Pharmaceuticals Experimental and Computational Approaches Conclusions and Further Work. *Control and Prediction of Solid-State of Pharmaceuticals: Experimental and Computational Approaches*, 195-205. doi:10.1007/978-3-319-27555-0_8

100. Fujiwara, M., Nagy, Z. K., Chew, J. W., & Braatz, R. D. (2005). First-principles and direct design approaches for the control of pharmaceutical crystallization. *Journal of Process Control, 15*(5), 493-504. doi:https://doi.org/10.1016/j.jprocont.2004.08.003

101. Furnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research, 2*(4), 721-747. doi:Doi 10.1162/153244302320884605

102. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *Ieee Transactions on Systems Man and Cybernetics Part C-Applications and Reviews, 42*(4), 463-484. doi:10.1109/Tsmcc.2011.2161285

103. Gani, R., Jimenez-Gonzalez, C., ten Kate, A., Crafts, P. A., Jones, M., Powell, L., . . . Cordiner, J. L. (2006). A modern approach to solvent selection. *Chemical Engineering, 113*(3), 30-43.

104. Garside, J., & Jančić, S. J. (1979). Measurement and scale-up of secondary nucleation kinetics for the potash alum-water system. *25*(6), 948-958. doi:10.1002/aic.690250605

105. Gasteiger, J. (2003). *Handbook of chemoinformatics : from data to knowledge*. Weinheim: Wiley-VCH.

106. Gasteiger, J., & Engel, T. (2006). *Chemoinformatics: A Textbook*: Wiley.

107. Gimenez, B. G., Santos, M. S., Ferrarini, M., & Fernandes, J. P. S. (2010). Evaluation of blockbuster drugs under the Rule-of-five. *Pharmazie, 65*(2), 148-152. doi:10.1691/ph.2010.9733

108. Gollapudi, S. (2016). *Practical Machine Learning*: Packt Publishing.

109. Gracin, S., & Rasmuson, Å. C. (2002). Solubility of Phenylacetic Acid, p-Hydroxyphenylacetic Acid, p-Aminophenylacetic Acid, p-Hydroxybenzoic Acid, and Ibuprofen in Pure Solvents. *Journal of Chemical & Engineering Data, 47*(6), 1379-1383. doi:10.1021/je0255170

110. Granberg, R. A., & Rasmuson, A. C. (1999). Solubility of paracetamol in pure solvents. *Journal of Chemical and Engineering Data, 44*(6), 1391-1395. doi:DOI 10.1021/je990124v

111. Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B, Structural Science, Crystal Engineering and Materials, 72*(Pt 2), 171-179. doi:10.1107/S2052520616003954

112. Gu, C. H., Li, H., Gandhi, R. B., & Raghavan, K. (2004). Grouping solvents by statistical analysis of solvent property parameters: implication to

polymorph screening. *International Journal of Pharmaceutics, 283*(1-2), 117-125. doi:10.1016/j.ijpharm.2004.06.021

113. Guo, H.-j., Cao, D.-l., Liu, Y., Dang, X., Yang, F., Li, Y.-x., . . . Li, Z.-h. (2017). Determination and correlation of solubility of N-methyl-3,4,5-trinitropyrazole (MTNP) in ten pure solvents from 283.15 K to 323.15 K. *Fluid Phase Equilibria, 444*, 13-20. doi:https://doi.org/10.1016/j.fluid.2017.04.008

114. Hahn, M. (1995). Receptor surface models. 1. Definition and construction. *J Med Chem, 38*(12), 2080-2090.

115. Hall, R. J., Mortenson, P. N., & Murray, C. W. (2014). Efficient exploration of chemical space by fragment-based screening. *Progress in Biophysics and Molecular Biology, 116*(2), 82-91. doi:https://doi.org/10.1016/j.pbiomolbio.2014.09.007

116. Hao, J., Yang, W., Li, H., Fan, S., Zhao, W., & Hu, Y. (2017). Solubility of 4-Methylsulfonylacetophenone in Nine Pure Solvents and (Ethyl Acetate + n-Hexane) Binary Solvent Mixtures from 278.15 K to 328.15 K. *Journal of Chemical & Engineering Data, 62*(1), 236-242. doi:10.1021/acs.jced.6b00617

117. Hari Narayana Moorthy, N., Ramos, M. J., & Fernandes, P. A. (2011). Topological, hydrophobicity, and other descriptors on α-glucosidase inhibition: a QSAR study on xanthone derivatives. *Journal of enzyme inhibition and medicinal chemistry, 26*(6), 755-766.

118. Hastie, T., Tibshirani, R., & Friedman, J. (2005). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

119. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

120. Haynes, W. M. (2016). *CRC Handbook of Chemistry and Physics, 94th Edition*: CRC Press.

121. Henderson, R. K., Jimenez-Gonzalez, C., Constable, D. J. C., Alston, S. R., Inglis, G. G. A., Fisher, G., . . . Curzons, A. D. (2011). Expanding GSK's solvent selection guide - embedding sustainability into solvent selection starting at medicinal chemistry. *Green Chemistry, 13*(4), 854-862. doi:10.1039/c0gc00918k

122. Heng, J. Y. Y., & Williams, D. R. (2006). Wettability of Paracetamol Polymorphic Forms I and II. *Langmuir, 22*(16), 6905-6909. doi:10.1021/la060596p

123. Hewitt, M., Cronin, M. T. D., Enoch, S. J., Madden, J. C., Roberts, D. W., & Dearden, J. C. (2009). In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *Journal of Chemical Information and Modeling, 49*(11), 2572-2587. doi:10.1021/ci900286s

124. Hiendrawan, S., Veriansyah, B., Widjojokusumo, E., Soewandhi, S. N., Wikarsa, S., & Tjandrawinata, R. R. (2016). Physicochemical and mechanical properties of paracetamol cocrystal with 5-nitroisophthalic acid. *International Journal of Pharmaceutics, 497*(1-2), 106-113. doi:10.1016/j.ijpharm.2015.12.001

125. Hiendrawan, S., Widjojokusumo, E., Veriansyah, B., & Tjandrawinata, R. R. (2017). Pharmaceutical Salts of Carvedilol: Polymorphism and Physicochemical Properties. *Aaps Pharmscitech, 18*(4), 1417-1425. doi:10.1208/s12249-016-0616-x

126. Hong, M., Wu, S., Qi, M., & Ren, G. (2016). Solubility correlation and thermodynamic analysis of two forms of Metaxalone in different pure solvents. *Fluid Phase Equilibria, 409*, 1-6. doi:https://doi.org/10.1016/j.fluid.2015.09.013

127. Hong, M., Xu, L., Ren, G., Chen, J., & Qi, M. (2012). Solubility of Lansoprazole in different solvents. *Fluid Phase Equilibria, 331*, 18-25. doi:https://doi.org/10.1016/j.fluid.2012.06.011

128. Hosokawa, K., Goto, J., & Hirayama, N. (2005). Prediction of Solvents Suitable for Crystallization of Small Organic Molecules. *Chemical and Pharmaceutical Bulletin, 53*(10), 1296-1299. doi:10.1248/cpb.53.1296

129. Hu, X., Wang, Y., Xie, C., Wang, G., & Hao, H. (2013). Determination and Correlation of Solubility of Cefradine Form I in Five Pure Solvents from (283.15 to 308.15) K. *Journal of Chemical & Engineering Data, 58*(7), 2028-2034. doi:10.1021/je400218d

130. Hu, Y., Zhang, Q., Shi, Y., Yang, Y., Cheng, L., Cao, C., & Yang, W. (2014). Thermodynamic Models for Determination of the Solubility of Sulfanilic Acid in Different Solvents at Temperatures from (278.15 to 328.15) K. *Journal of Chemical & Engineering Data, 59*(11), 3938-3943. doi:10.1021/je500867u

131. Huang, C., Xie, Z., Xu, J., Qin, Y., Du, Y., Du, S., & Gong, J. (2015). Experimental and Modeling Studies on the Solubility of d-Pantolactone in Four Pure Solvents and Ethanol–Water Mixtures. *Journal of Chemical & Engineering Data, 60*(3), 870-875. doi:10.1021/je500996n

132. Huang, X., Wang, J., Hao, H., Ouyang, J., Gao, Y., Bao, Y., . . . Yin, Q. (2015). Determination and correlation of solubility and solution thermodynamics of coumarin in different pure solvents. *Fluid Phase Equilibria, 394*, 148-155. doi:https://doi.org/10.1016/j.fluid.2015.03.022

133. Hughes, J. M., Aherne, D., & Coleman, J. N. (2012). Generalizing solubility parameter theory to apply to one- and two-dimensional solutes and to incorporate dipolar interactions. *Journal of Applied Polymer Science, 127*(6), 4483-4491. doi:10.1002/app.38051

134. Hughes, L. D., Palmer, D. S., Nigsch, F., & Mitchell, J. B. O. (2008). Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *Journal of Chemical Information and Modeling, 48*(1), 220-232. doi:10.1021/ci700307p

135. Husson, F., Le, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R, Second Edition*: CRC Press.

136. Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314. doi:10.2307/1390807

137. Iveson, S. M., & Litster, J. D. (1998). Growth regime map for liquid-bound granules. *44*(7), 1510-1518. doi:doi:10.1002/aic.690440705

138. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. New York: Springer.

139. Ji, W., Meng, Q., Ding, L., Wang, F., Dong, J., Zhou, G., & Wang, B. (2016). Measurement and correlation of the solubility of caffeic acid in eight mono and water+ethanol mixed solvents at temperatures from (293.15 to 333.15) K. *Journal of Molecular Liquids, 224*, 1275-1281. doi:https://doi.org/10.1016/j.molliq.2016.10.110

140. Ji, W., Meng, Q., Li, P., Yang, B., Wang, F., Ding, L., & Wang, B. (2016). Measurement and Correlation of the Solubility of p-Coumaric Acid in Nine Pure and Water + Ethanol Mixed Solvents at Temperatures from 293.15 to 333.15 K. *Journal of Chemical & Engineering Data, 61*(10), 3457-3465. doi:10.1021/acs.jced.6b00361

141. Jia, D., Li, Y., Li, Y., & Li, C. (2013). Measurement and Correlation of Solubility of 2-Chloro-4-ethylamino-6-isopropylamino-1,3,5-triazine in Different Organic Solvents. *Journal of Chemical & Engineering Data, 58*(11), 3183-3189. doi:10.1021/je400639m

142. Jimenez-Gonzalez, C., Ponder, C. S., Broxterman, Q. B., & Manley, J. B. (2011). Using the Right Green Yardstick: Why Process Mass intensity Is Used in the Pharmaceutical Industry To Drive More Sustainable Processes. *Organic Process Research & Development, 15*(4), 912-917. doi:10.1021/op200097d

143. Jing, D., Wang, J., & Wang, Y. (2010). Solubility of Penicillin Sulfoxide in Different Solvents. *Journal of Chemical & Engineering Data, 55*(1), 508-509. doi:10.1021/je900326e

144. Johnston, A., Bhardwaj-Miglani, R., Gurung, R., Vassileiou, A. D., Florence, A. J., & Johnston, B. F. (2017). Combined Chemoinformatics Approach to Solvent Library Design Using clusterSim and Multidimensional Scaling. *Journal of Chemical Information and Modeling, 57*(8), 1807-1815. doi:10.1021/acs.jcim.71300038

145. Johnston, A., Johnston, B. F., Kennedy, A. R., & Florence, A. J. (2008). Targeted crystallisation of novel carbamazepine solvates based on a retrospective Random Forest classification. *CrystEngComm, 10*(1), 23-25. doi:10.1039/b713373a

146. Joiris, E., Di Martino, P., Berneron, C., Guyot-Hermann, A. M., & Guyot, J. C. (1998). Compression behavior of orthorhombic paracetamol. *Pharmaceutical Research, 15*(7), 1122-1130. doi:Doi 10.1023/A:1011954800246

147. Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

148. Jones, A. G. (2002). *Crystallization Process Systems*: Elsevier Science.

149. Jouyban, A. (2009). *Handbook of Solubility Data for Pharmaceuticals*: CRC Press.

150. Kai, Y., Hu, Y., Cao, Z., Liu, X., Liu, Y., & Yang, W. (2013). Measurement and correlation solubility and mixing properties of l-malic acid in pure and mixed organic solvents. *Fluid Phase Equilibria, 360*, 466-471. doi:https://doi.org/10.1016/j.fluid.2013.10.011

151. Kassambara, A. (2017). *Practical Guide To Principal Component Methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra*: CreateSpace Independent Publishing Platform.

152. Katritzky, A. R., Fara, D. C., Kuanar, M., Hur, E., & Karelson, M. (2005). The classification of solvents by combining classical QSPR methodology with principal component analysis. *J Phys Chem A, 109*(45), 10323-10341. doi:10.1021/jp050395e

153. Katritzky, A. R., Fara, D. C., Yang, H., Tämm, K., Tamm, T., & Karelson, M. (2004). Quantitative Measures of Solvent Polarity. *Chemical Reviews, 104*(1), 175-198. doi:10.1021/cr020750m

154. Katritzky, A. R., Oliferenko, A. A., Oliferenko, P. V., Petrukhin, R., Tatham, D. B., Maran, U., . . . Acree, W. E., Jr. (2003). A general treatment of solubility. 2. QSPR prediction of free energies of solvation of specified solutes in ranges of solvents. *J Chem Inf Comput Sci, 43*(6), 1806-1814. doi:10.1021/ci034122x

155. Khadka, P., Ro, J., Kim, H., Kim, I., Kim, J. T., Kim, H., . . . Lee, J. (2014). Pharmaceutical particle technologies: An approach to improve drug solubility, dissolution and bioavailability. *Asian Journal of Pharmaceutical Sciences, 9*(6), 304-316. doi:https://doi.org/10.1016/j.ajps.2014.05.005

156. Kim, K.-J., & Mersmann, A. (2001). Estimation of metastable zone width in different nucleation processes. *Chemical Engineering Science, 56*(7), 2315-2324.

157. Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology, 26*(9), 1011-1013. doi:10.1038/nbt0908-1011

158. Kobari, M., Kubota, N., & Hirasawa, I. (2012). Secondary nucleation-mediated effects of stirrer speed and growth rate on induction time for unseeded solution. *CrystEngComm, 14*(16), 5255-5261. doi:10.1039/C2CE25248A

159. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5), 1-26.

160. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

161. Kumar, R., & Siril, P. F. (2014). Ultrafine carbamazepine nanoparticles with enhanced water solubility and rate of dissolution. *Rsc Advances, 4*(89), 48101-48108. doi:10.1039/c4ra08495k

162. Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics Modelling, 18*(4-5), 464-477.

163. Labute, P. (2000). A widely applicable set of descriptors. *J Mol Graph Model, 18*(4-5), 464-477.

164. Lagemann, R. T. (1945). A Relation between Viscosity and Refractive Index. *Journal of the American Chemical Society, 67*(3), 498-499. doi:10.1021/ja01219a509

165. Landrum, G. (2006). RDKit: Open-source cheminformatics.

166. Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1-18.

167. Leach, A. G., Jones, H. D., Cosgrove, D. A., Kenny, P. W., Ruston, L., MacFaul, P., . . . Law, B. (2006). Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *Journal of Medicinal Chemistry, 49*(23), 6672-6682. doi:10.1021/jm0605233

168. Lee, E. H. (2014). A practical guide to pharmaceutical polymorph screening & selection. *Asian Journal of Pharmaceutical Sciences, 9*(4), 163-175. doi:https://doi.org/10.1016/j.ajps.2014.05.002

169. Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods, 14*, 641. doi:10.1038/nmeth.4346

170. Lever, J., Krzywinski, M., & Atman, N. (2017). POINTS OF SIGNIFICANCE Principal component analysis. *Nature Methods, 14*(7), 641-642. doi:DOI 10.1038/nmeth.4346

171. Lewis, A., Seckler, M., Kramer, H., & van Rosmalen, G. (2015). *Industrial Crystallization: Fundamentals and Applications*: Cambridge University Press.

172. Li, B., Wu, Y., Zhu, J., Chen, K., Wu, B., & Ji, L. (2016). Determination and correlation of solubility and mixing properties of isonicotinamide (form II) in some pure solvents. *Thermochimica Acta, 627-629*, 55-60. doi:https://doi.org/10.1016/j.tca.2016.01.012

173. Li, C. (2016). The Application of high-dimensional Data Classification by Random Forest based on Hadoop Cloud Computing Platform. *3rd International Conference on Applied Engineering, 51*, 385-390. doi:10.3303/Cet1651065

174. Li, H., Yang, W., Hao, J., Fan, S., Yang, S., & Guo, Q. (2016). Experimental measurement and thermodynamic models for solid–liquid equilibrium of 3-amino-1-adamantanol in different pure solvents and in (H2O+ethanol) binary solvent mixtures. *Journal of Molecular Liquids, 215*, 127-134. doi:https://doi.org/10.1016/j.molliq.2015.12.052

175. Li, H., Zhu, J., Hu, G., Jiang, P., Zhao, L., & Zhang, Y. (2010). Measurement and Correlation of Solubility of Pimelic Acid in Ether, Tetrahydrofuran, Ethanol, and Methanol. *Journal of Chemical & Engineering Data, 55*(3), 1443-1445. doi:10.1021/je900629v

176. Li, J. D., & Liu, H. (2017). Challenges of Feature Selection for Big Data Analytics. *Ieee Intelligent Systems, 32*(2), 9-15.

177. Li, Q.-S., Li, Z., & Wang, S. (2008). Solubility of Trimethoprim (TMP) in Different Organic Solvents from (278 to 333) K. *Journal of Chemical & Engineering Data, 53*(1), 286-287. doi:10.1021/je700497h

178. Li, W., Chen, J., Han, S., Du, C., & Zhao, H. (2016). Thermodynamic solubility of tetraethyl ranelate in ten organic solvents at different temperatures. *Journal of Molecular Liquids, 216*, 771-780. doi:https://doi.org/10.1016/j.molliq.2016.02.002

179. Li, X., Cong, Y., Cunbin, D., & Hongkun, Z. (2016). Solubility and solution thermodynamics of 2-methyl-4-nitroaniline in eleven organic solvents at elevated temperatures. *The Journal of Chemical Thermodynamics, 105*. doi:10.1016/j.jct.2016.10.037

180. Li, X., Cong, Y., Du, C., & Zhao, H. (2017). Solubility and solution thermodynamics of 2-methyl-4-nitroaniline in eleven organic solvents at elevated temperatures. *The Journal of Chemical Thermodynamics, 105*, 276-288. doi:https://doi.org/10.1016/j.jct.2016.10.037

181. Li, X., Du, C., Cong, Y., & Zhao, H. (2017). Solubility determination and thermodynamic modelling of 3-amino-1,2,4-triazole in ten organic solvents from T=283.15K to T=318.15K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 104*, 189-200. doi:https://doi.org/10.1016/j.jct.2016.09.033

182. Li, X., Wang, M., Cong, Y., Du, C., & Zhao, H. (2017). Solubility of 4-methyl-2-nitroaniline in fourteen organic solvents from T=(278.15 to 313.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 110*, 33-40. doi:https://doi.org/10.1016/j.jct.2017.02.008

183. Li, Y., Li, C., Cong, Y., Du, C., & Zhao, H. (2017). Solubility and thermodynamic functions of tebuconazole in nine organic solvents from T=(278.15 to 313.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 106*, 243-255. doi:https://doi.org/10.1016/j.jct.2016.11.032

184. Li, Z., Zhang, T., Huang, C., Wang, H., Yu, B., & Gong, J. (2016). Measurement and Correlation of the Solubility of Maltitol in Different Pure Solvents, Methanol–Water Mixtures, and Ethanol–Water Mixtures. *Journal of Chemical & Engineering Data, 61*(3), 1065-1070. doi:10.1021/acs.jced.5b00565

185. Liang, A., Wang, S., & Qu, Y. (2017). Determination and Correlation of Solubility of Phenylbutazone in Monosolvents and Binary Solvent Mixtures.

Journal of Chemical & Engineering Data, 62(2), 864-871. doi:10.1021/acs.jced.6b00911

186. Liang, S., Li, H., Shen, L., Li, H., Mao, Z., & Li, H. (2016). Measurement and correlation of the solubility of (1-benzyl-1H-1,2,3-triazole-4-yl)methanol in water and alcohols at temperatures from 292.15K to 310.15K. *Thermochimica Acta, 630*, 1-10. doi:https://doi.org/10.1016/j.tca.2016.01.009

187. Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*.

188. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.

189. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 23*(1-3), 3-25. doi:Doi 10.1016/S0169-409x(96)00423-1

190. Liu, J.-Q., Chen, S.-Y., & Ji, B. (2014). Solubility and Thermodynamic Functions of Isatin in Pure Solvents. *Journal of Chemical & Engineering Data, 59*(11), 3407-3414. doi:10.1021/je500396b

191. Liu, M., Bienfait, B., Sacher, O., Gasteiger, J., Siezen, R. J., Nauta, A., & Geurts, J. M. (2014). Combining chemoinformatics with bioinformatics: in silico prediction of bacterial flavor-forming pathways by a chemical systems biology approach "reverse pathway engineering". *PLoS One, 9*(1), e84769. doi:10.1371/journal.pone.0084769

192. Liu, W., Dang, L., Black, S., & Wei, H. (2008). Solubility of Carbamazepine (Form III) in Different Solvents from (275 to 343) K. *Journal of Chemical & Engineering Data, 53*(9), 2204-2206. doi:10.1021/je8002157

193. Liu, W. J., Dang, L. P., Black, S., & Wei, H. Y. (2008). Solubility of Carbamazepine (Form III) in Different Solvents from (275 to 343) K (vol 53,

pg 2204, 2008). *Journal of Chemical and Engineering Data, 53*(12), 2918-2918. doi:10.1021/je800762j

194. Liu, Y., Gao, H., Ren, F., & Ren, G. (2015). Solubility of Agomelatine Crystal Form I and Form II in Pure Solvents and (Isopropanol + Water) Mixtures. *Journal of Chemical & Engineering Data, 60*(11), 3347-3352. doi:10.1021/acs.jced.5b00586

195. Llinàs, A., Glen, R. C., & Goodman, J. M. (2008). Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *Journal of Chemical Information and Modeling, 48*(7), 1289-1303. doi:10.1021/ci800058v

196. Loudon, G. M. (2001). Organic chemistry *Organic chemistry*. New York: Oxford University Press.

197. Lusci, A., Pollastri, G., & Baldi, P. (2013). Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *Journal of Chemical Information and Modeling, 53*(7), 1563-1575. doi:10.1021/ci400187y

198. Maghsoodi, M. (2015). Role of Solvents in Improvement of Dissolution Rate of Drugs: Crystal Habit and Crystal Agglomeration. *Advanced Pharmaceutical Bulletin, 5*(1), 13-18. doi:10.5681/apb.2015.002

199. Maia, G. D., & Giulietti, M. (2008). Solubility of Acetylsalicylic Acid in Ethanol, Acetone, Propylene Glycol, and 2-Propanol. *Journal of Chemical & Engineering Data, 53*(1), 256-258. doi:10.1021/je7005693

200. Maltarollo, V. G., Gertrudes, J. C., Oliveira, P. R., & Honorio, K. M. (2015). Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opinion on Drug Metabolism & Toxicology, 11*(2), 259-271. doi:10.1517/17425255.2015.980814

201. Mangin, D., Puel, F., & Veesler, S. (2009). Polymorphism in processes of crystallization in solution: a practical review. *Organic Process Research & Development, 13*(6), 1241-1253.

202. Marston, L. (2010). *Introductory statistics for health and nursing using SPSS*. Los Angeles: SAGE.

203. Martin, T. M., Grulke, C. M., Young, D. M., Russom, C. L., Wang, N. Y., Jackson, C. R., & Barron, M. G. (2013). Prediction of Aquatic Toxicity Mode of Action Using Linear Discriminant and Random Forest Models. *Journal of Chemical Information and Modeling, 53*(9), 2229-2239. doi:10.1021/ci400267h

204. Matsuda, H., Kaburagi, K., Matsumoto, S., Kurihara, K., Tochigi, K., & Tomono, K. (2009). Solubilities of Salicylic Acid in Pure Solvents and Binary Mixtures Containing Cosolvent. *Journal of Chemical & Engineering Data, 54*(2), 480-484. doi:10.1021/je800475d

205. Mauger, J. W., Paruta, A. N., & Gerraughty, R. J. (1972). Solubilities of sulfadiazine, sulfisomidine, and sulfadimethoxine in several normal alcohols. *Journal of Pharmaceutical Sciences, 61*(1), 94-97. doi:10.1002/jps.2600610117

206. Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. J. M. (2006). Dragon software: An easy approach to molecular descriptor calculations. *56*(2), 237-248.

207. McDonagh, J. L., Nath, N., De Ferrari, L., van Mourik, T., & Mitchell, J. B. O. (2014). Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *Journal of Chemical Information and Modeling, 54*(3), 844-856. doi:10.1021/ci4005805

208. McDonagh, J. L., van Mourik, T., & Mitchell, J. B. O. (2016). Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation (vol 34, pg 715, 2015). *Molecular Informatics, 35*(10), 538-538. doi:10.1002/minf.201681041

209. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282.

210. McNaught, A. D., Wilkinson, A., Jenkins, A. D., International Union of, P., & Applied, C. (2006). *IUPAC compendium of chemical terminology : the gold book*. [Research Triangle Park, N.C.]: International Union of Pure and Applied Chemistry.

211. Mealey, D., Svärd, M., & Rasmuson, Å. C. (2014). Thermodynamics of risperidone and solubility in pure organic solvents. *Fluid Phase Equilibria, 375*, 73-79. doi:https://doi.org/10.1016/j.fluid.2014.04.028

212. Meng, Z., Hu, Y., Kai, Y., Yang, W., Cao, Z., & Shen, F. (2013). Thermodynamics of solubility of thiomalic acid in different organic solvents from 278.15K to 333.15K. *Fluid Phase Equilibria, 352*, 1-6. doi:https://doi.org/10.1016/j.fluid.2013.05.002

213. Menon, S. K., Mistry, B. R., Joshi, K. V., Modi, N. R., & Shashtri, D. (2012). Evaluation and solubility improvement of Carvedilol: PSC[n]arene inclusion complexes with acute oral toxicity studies. *Journal of Inclusion Phenomena and Macrocyclic Chemistry, 73*(1-4), 295-303. doi:10.1007/s10847-011-0056-x

214. Mitchell, J. B. (2014). Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci, 4*(5), 468-481. doi:10.1002/wcms.1183

215. Mitchell, M. W. (2011). Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. *Open Journal of Statistics, Vol.01No.03*, 7. doi:10.4236/ojs.2011.13024

216. MonÁrrez, C. I., Stovall, D. M., Woo, J. H., Taylor, P., & Acree, W. E. (2002). Solubility of Xanthene in Organic Nonelectrolyte Solvents: Comparison of Observed Versus Predicted Values Based Upon Mobile Order Theory. *Physics and Chemistry of Liquids, 40*(6), 703-714. doi:10.1080/0031910021000018581

217. Morten, A., Frans, v. d. B., Claus, C., Steen, J. F., Bent, H. S., Lopez, d. D. H., . . . Jukka, R. (2008). Solvent diversity in polymorph screening. *Journal of Pharmaceutical Sciences, 97*(6), 2145-2159. doi:doi:10.1002/jps.21153

218. Mullin, J. W. (2001a). *Crystallisation, 4th Edition*: Butterworth-Heinemann.

219. Mullin, J. W. (2001b). *Crystallization* (4th ed.). Oxford ; Boston: Butterworth-Heinemann.

220. Musil, F., De, S., Yang, J., Campbell, J. E., Day, G. M., & Ceriotti, M. (2018). Machine learning for the structure–energy–property landscapes of molecular crystals. *Chemical Science, 9*(5), 1289-1300. doi:10.1039/C7SC04665K

221. Myerson, A. S., & Ginde, R. (2002). 2 - Crystals, crystal growth, and nucleation *Handbook of Industrial Crystallization (Second Edition)* (pp. 33-65). Woburn: Butterworth-Heinemann.

222. Nam, K., Ha, E.-S., Kim, J.-S., Kuk, D.-H., Ha, D.-H., Kim, M.-S., . . . Hwang, S.-J. (2017). Solubility of oxcarbazepine in eight solvents within the temperature range T=(288.15–308.15)K. *The Journal of Chemical Thermodynamics, 104*, 45-49. doi:https://doi.org/10.1016/j.jct.2016.09.011

223. Nordström, F. L., & Rasmuson, Å. C. (2006). Solubility and Melting Properties of Salicylamide. *Journal of Chemical & Engineering Data, 51*(5), 1775-1777. doi:10.1021/je060178m

224. O'Boyle, N. M. (2012). Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics, 4*(1), 22-22. doi:10.1186/1758-2946-4-22

225. Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings, 5*(3), S11. doi:10.1186/1753-6561-5-s3-s11

226. Omkvist, D. H., Larsen, S. B., Nielsen, C. U., Steffansen, B., Olsen, L., Jørgensen, F. S., & Brodin, B. (2010). A Quantitative Structure–Activity

Relationship for Translocation of Tripeptides via the Human Proton-Coupled Peptide Transporter, hPEPT1 (SLC15A1). *The AAPS Journal, 12*(3), 385-396. doi:10.1208/s12248-010-9195-z

227. Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings* (pp. 154-168). Berlin, Heidelberg: Springer Berlin Heidelberg.

228. Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. O. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling, 47*(1), 150-158. doi:10.1021/ci060164k

229. Pataki, H., Markovits, I., Vajna, B., Nagy, Z. K., & Marosi, G. (2012). In-Line Monitoring of Carvedilol Crystallization Using Raman Spectroscopy. *Crystal Growth & Design, 12*(11), 5621-5628. doi:10.1021/cg301135z

230. Perlovich, G. L., & Bauer-Brandl, A. (2003). Thermodynamics of Solutions I: Benzoic Acid and Acetylsalicylic Acid as Models for Drug Substances and the Prediction of Solubility. *Pharmaceutical Research, 20*(3), 471-478. doi:10.1023/A:1022624725495

231. Pickett, S. D. (2007). 4.15 - Library Design: Reactant and Product-Based Approaches. In J. B. Taylor & D. J. Triggle (Eds.), *Comprehensive Medicinal Chemistry II* (pp. 337-378). Oxford: Elsevier.

232. Planinsek, O., Kovacic, B., & Vrecer, F. (2011). Carvedilol dissolution improvement by preparation of solid dispersions with porous silica. *International Journal of Pharmaceutics, 406*(1-2), 41-48. doi:10.1016/j.ijpharm.2010.12.035

233. Prado, L. D., Rocha, H. V. A., Resende, J. A. L. C., Ferreira, G. B., & de Figuereido Teixeira, A. M. R. (2014). An insight into carvedilol solid forms:

effect of supramolecular interactions on the dissolution profiles. *Crystengcomm, 16*(15), 3168-3179. doi:10.1039/C3CE42403K

234. Prat, D., Pardigon, O., Flemming, H. W., Letestu, S., Ducandas, V., Isnard, P., . . . Hosek, P. (2013). Sanofi's Solvent Selection Guide: A Step Toward More Sustainable Processes. *Organic Process Research & Development, 17*(12), 1517-1525. doi:10.1021/op4002565

235. Prat, D., Wells, A., Hayler, J., Sneddon, H., McElroy, C. R., Abou-Shehada, S., & Dunn, P. J. (2016). CHEM21 selection guide of classical- and less classical-solvents. *Green Chemistry, 18*(1), 288-296. doi:10.1039/c5gc01008j

236. Qi, Y. (2012). Random forest for bioinformatics *Ensemble machine learning* (pp. 307-323): Springer.

237. Qi, Y. J. (2012). Random Forest for Bioinformatics. *Ensemble Machine Learning: Methods and Applications*, 307-323. doi:10.1007/978-1-4419-9326-7_11

238. Qin, Y., Wang, H., Yang, P., Du, S., Huang, C., Du, Y., . . . Yin, Q. (2015). Measurement and correlation of solubility and dissolution properties of flunixin meglumine in pure and binary solvents. *Fluid Phase Equilibria, 403*, 145-152. doi:https://doi.org/10.1016/j.fluid.2015.06.026

239. Qiu, J., & Albrecht, J. (2018). Solubility Correlations of Common Organic Solvents. *Organic Process Research & Development, 22*(7), 829-835. doi:10.1021/acs.oprd.8b00117

240. R Development Core Team. (2013). R: A language and environment for statistical computing. 3.3.1. Retrieved from http://www.R-project.org/

241. Ren, Y., Duan, X., & Yang, J. (2014). Experimental measurement and correlation of the solubility of 2-Cyanoguanidine in different pure solvents. *Journal of Molecular Liquids, 191*, 53-58. doi:https://doi.org/10.1016/j.molliq.2013.11.029

242. Reymond, J.-L., van Deursen, R., Blum, L. C., & Ruddigkeit, L. (2010). Chemical space as a source for new drugs. *MedChemComm, 1*(1), 30-38. doi:10.1039/C0MD00020E

243. Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research, 5*, 101-141.

244. Romero, S., Reillo, A., Escalera, B., & Bustamante, P. (1996). The behavior of paracetamol in mixtures of amphiprotic and amphiprotic-aprotic solvents. Relationship of solubility curves to specific and nonspecific interactions. *Chemical & Pharmaceutical Bulletin, 44*(5), 1061-1064.

245. Rupp, B., & Wang, J. (2004). Predictive models for protein crystallization. *Methods, 34*(3), 390-407. doi:https://doi.org/10.1016/j.ymeth.2004.03.031

246. Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8*(1), 3-15. doi:Doi 10.1191/096228099671525676

247. Schuldt, S., & Schembecker, G. (2013). A Fully Automated Ad- and Desorption Method for Resin and Solvent Screening. *Chemical Engineering & Technology, 36*(7), 1157-1164. doi:10.1002/ceat.201200725

248. Schwartz, A. M., & Myerson, A. S. (2002). 1 - Solutions and solution properties *Handbook of Industrial Crystallization (Second Edition)* (pp. 1-31). Woburn: Butterworth-Heinemann.

249. Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics, 26*(14), 1752-1758. doi:10.1093/bioinformatics/btq257

250. Shahlaei, M. J. C. r. (2013). Descriptor selection methods in quantitative structure–activity relationship studies: a review study. *113*(10), 8093-8103.

251. Shakeel, F., Haq, N., Alanazi, F., & Alsarra, I. (2015). Solubility of anti-inflammatory drug lornoxicam in ten different green solvents at different

temperatures. *Journal of Molecular Liquids, 209*, 280-283. doi:10.1016/j.molliq.2015.05.035

252. Shakeel, F., Haq, N., Shazly, G. A., Alanazi, F. K., & Alsarra, I. A. (2015). Solubility and Thermodynamic Analysis of Tenoxicam in Different Pure Solvents at Different Temperatures. *Journal of Chemical & Engineering Data, 60*(8), 2510-2514. doi:10.1021/acs.jced.5b00382

253. Shakeel, F., Salem-Bekhit, M. M., Haq, N., & Siddiqui, N. A. (2017). Solubility and thermodynamics of ferulic acid in different neat solvents: Measurement, correlation and molecular interactions. *Journal of Molecular Liquids, 236*, 144-150. doi:https://doi.org/10.1016/j.molliq.2017.04.014

254. Shalmashi, A., & Eliassi, A. (2008). Solubility of l-(+)-Ascorbic Acid in Water, Ethanol, Methanol, Propan-2-ol, Acetone, Acetonitrile, Ethyl Acetate, and Tetrahydrofuran from (293 to 323) K. *Journal of Chemical & Engineering Data, 53*(6), 1332-1334. doi:10.1021/je800056h

255. Shalmashi, A., & Golmohammad, F. (2010). Solubility of caffeine in water, ethyl acetate, ethanol, carbon tetrachloride, methanol, chloroform, dichloromethane, and acetone between 298 and 323 K. *Latin American applied research, 40*, 283-285.

256. Shan, Y., Fu, M., & Yan, W. (2017). Solubilities of 4′,5,7-Triacetoxyflavanone in Fourteen Organic Solvents at Different Temperatures. *Journal of Chemical & Engineering Data, 62*(1), 568-574. doi:10.1021/acs.jced.6b00929

257. Shayanfar, A., Fakhree, M. A. A., & Jouyban, A. (2010). A simple QSPR model to predict aqueous solubility of drugs. *Journal of Drug Delivery Science and Technology, 20*(6), 467-476. doi:Doi 10.1016/S1773-2247(10)50080-7

258. Shekunov, B. Y., & York, P. (2000). Crystallization processes in pharmaceutical technology and drug delivery design. *Journal of Crystal Growth, 211*(1-4), 122-136. doi:Doi 10.1016/S0022-0248(99)00819-2

259. Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology, 18*(4), 547-557. doi:10.1038/modpathol.3800322

260. Sikic, M., Tomic, S., & Vlahovicek, K. (2009). Prediction of Protein-Protein Interaction Sites in Sequences and 3D Structures by Random Forests. *Plos Computational Biology, 5*(1). doi:ARTN e1000278

261. 10.1371/journal.pcbi.1000278

262. Sliwoski, G., Mendenhall, J., & Meiler, J. (2016). Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign. *Journal of computer-aided molecular design, 30*(3), 209-217. doi:10.1007/s10822-015-9893-9

263. Smith, C. (2002). *Cheminformatics: Redefining the crucible* (Vol. 16).

264. Smith, S. J., Bishop, M. M., Montgomery, J. M., Hamilton, T. P., & Vohra, Y. K. (2014). Polymorphism in paracetamol: evidence of additional forms IV and V at high pressure. *J Phys Chem A, 118*(31), 6068-6077. doi:10.1021/jp411810y

265. Somvanshi, M., & Chavan, P. (2016, 12-13 Aug. 2016). *A review of machine learning techniques using decision tree and support vector machine.* Paper presented at the 2016 International Conference on Computing Communication Control and automation (ICCUBEA).

266. Song, L., Li, M., & Gong, J. (2010). Solubility of Clopidogrel Hydrogen Sulfate (Form II) in Different Solvents. *Journal of Chemical & Engineering Data, 55*(9), 4016-4018. doi:10.1021/je100022w

267. Storm, T. D., Hazleton, R. A., & Lahti, L. E. (1970). Some effects of solvent properties on nucleation. *Journal of Crystal Growth, 7*(1), 55-60. doi:https://doi.org/10.1016/0022-0248(70)90114-4

268. Su, Q., Nagy, Z. K., & Rielly, C. D. (2015). Pharmaceutical crystallisation processes from batch to continuous operation using MSMPR stages: Modelling, design, and control. *Chemical Engineering and Processing: Process Intensification, 89*, 41-53.

269. Sullivan, W. (2018). *Decision Tree and Random Forest: Machine Learning and Algorithms: The Future Is Here!* : CreateSpace Independent Publishing Platform.

270. Sun, D., Ren, R., Dun, W., Zhang, H., Zhao, L., Zhang, L., . . . Gong, J. (2014). Measurement and Correlation of the Solubility of l-Carnitine in Different Pure Solvents and Ethanol–Acetone Solvent Mixture. *Journal of Chemical & Engineering Data, 59*(6), 1984-1990. doi:10.1021/je500078n

271. Sun, F., Kang, H., Zhang, K., Liu, B., & Zhang, B. (2012). Solubility of chlocyphos in different solvents. *Fluid Phase Equilibria, 330*, 12-16. doi:https://doi.org/10.1016/j.fluid.2012.06.010

272. Sun, H., Gong, J.-b., & Wang, J.-k. (2005). Solubility of Lovastatin in Acetone, Methanol, Ethanol, Ethyl Acetate, and Butyl Acetate between 283 K and 323 K. *Journal of Chemical & Engineering Data, 50*(4), 1389-1391. doi:10.1021/je0500781

273. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., . . . Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design, 25*(6), 533-554. doi:10.1007/s10822-011-9440-2

274. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning : an introduction* (Second edition. ed.). Cambridge, MA: The MIT Press.

275. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for

compound classification and QSAR modeling. *J Chem Inf Comput Sci, 43*. doi:10.1021/ci034160g

276. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947-1958. doi:10.1021/ci034160g

277. Taldeddin, K., Khimyak, Y. Z., & Fabian, L. (2016). Prediction of Hydrate and Solvate Formation Using Statistical Models. *Crystal Growth & Design, 16*(1), 70-81. doi:10.1021/acs.cgd.5b00966

278. Tang, N., Shi, W., & Yan, W. (2016). Modified Method for Measuring the Solubility of Pharmaceutical Compounds in Organic Solvents by Visual Camera. *Journal of Chemical & Engineering Data, 61*(1), 35-40. doi:10.1021/acs.jced.5b00122

279. Tang, W., Dai, H., Feng, Y., Wu, S., Bao, Y., Wang, J., & Gong, J. (2015). Solubility of tridecanedioic acid in pure solvent systems: An experimental and computational study. *The Journal of Chemical Thermodynamics, 90*, 28-38. doi:https://doi.org/10.1016/j.jct.2015.05.026

280. Tavare, N. S. (1995). *Industrial crystallization : process simulation analysis and design*. New York: Plenum Press.

281. Tavare, N. S. (2013). *Industrial Crystallization: Process Simulation Analysis and Design*: Springer US.

282. ter Horst, J. H., Schmidt, C., & Ulrich, J. (2015). 32 - Fundamentals of Industrial Crystallization. In P. Rudolph (Ed.), *Handbook of Crystal Growth (Second Edition)* (pp. 1317-1349). Boston: Elsevier.

283. Tetko, I. V., Lowe, D. M., & Williams, A. J. (2016). The development of models to predict melting and pyrolysis point data associated with several

hundred thousand compounds mined from PATENTS. *Journal of Cheminformatics, 8*(1), 2.

284. Thati, J., Nordström, F. L., & Rasmuson, Å. C. (2010). Solubility of Benzoic Acid in Pure Solvents and Binary Mixtures. *Journal of Chemical & Engineering Data, 55*(11), 5124-5127. doi:10.1021/je100675r

285. Todeschini, R., Consonni, V., & Mannhold, R. (2009). *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*: Wiley-VCH.

286. Tu, L., Chung Shin, K., & Ying Hsiu, C. (2006). Solubility, Polymorphism, Crystallinity, and Crystal Habit of Acetaminophen and Ibuprofen by Initial Solvent Screening. *Pharmaceutical Technology, 30*(10), 72-80,82,84,86,88,90,92.

287. Tully, G., Hou, G., & Glennon, B. (2016). Solubility of Benzoic Acid and Aspirin in Pure Solvents Using Focused Beam Reflective Measurement. *Journal of Chemical & Engineering Data, 61*(1), 594-601. doi:10.1021/acs.jced.5b00746

288. Tyrchan, C., & Evertsson, E. (2016). Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Computational and structural biotechnology journal, 15*, 86-90. doi:10.1016/j.csbj.2016.12.003

289. Variankaval, N., Cote, A. S., & Doherty, M. F. (2008). From form to function: Crystallization of active pharmaceutical ingredients. *AIChE Journal, 54*(7), 1682-1688. doi:10.1002/aic.11555

290. Varnek, A., & Baskin, II. (2011). Chemoinformatics as a Theoretical Chemistry Discipline. *Mol Inform, 30*(1), 20-32. doi:10.1002/minf.201000100

291. Varnek, A., & Baskin, I. (2012). Machine learning methods for property prediction in chemoinformatics: Quo Vadis? *J Chem Inf Model, 52*(6), 1413-1437. doi:10.1021/ci200409x

292. Vehkamäki, H., Määttänen, A., Lauri, A., Napari, I., and Kulmala, M. (2007). Technical Note: The heterogeneous Zeldovich factor. *Atmospheric Chemistry and Physics, 7*, 309-313.

293. Vilar, S., Ferino, G., Quezada, E., Santana, L., & Friedman, C. (2012). Predicting Monoamine Oxidase Inhibitory Activity Through Ligand-Based Models. *Current Topics in Medicinal Chemistry, 12*(20), 2258-2274. doi:Doi 10.2174/156802612805219987

294. Vilas Boas, S. A. M. (2017). *Studies on the solubility of phenolic compounds.* (Chemical Engineering), Instituto Politécnico de Bragança. Retrieved from http://hdl.handle.net/10198/14601

295. Wale, N., Watson, I. A., & Karypis, G. (2008). Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems, 14*(3), 347-375. doi:10.1007/s10115-007-0103-5

296. Wang, H., & Zhang, W. (2009). Solubility of 3,5-Dimethoxybenzoic Acid, 4-Cyanobenzoic Acid, 4-Acetoxybenzoic Acid, 3,5-Diaminobenzoic Acid, and 2,4-Dichlorobenzoic Acid in Ethanol. *Journal of Chemical & Engineering Data, 54*(6), 1942-1944. doi:10.1021/je900190n

297. Wang, J., Xu, A., & Xu, R. (2017a). Solubility and solution thermodynamics of 4-hydroxybenzaldehyde in twelve organic solvents from T=(278.15 to 318.15)K. *Journal of Molecular Liquids, 237*, 226-235. doi:https://doi.org/10.1016/j.molliq.2017.04.082

298. Wang, J., Xu, A., & Xu, R. (2017b). Solubility of 2-nitro-p-phenylenediamine in nine pure solvents and mixture of (methanol+N-methyl-2-pyrrolidone) from T=(283.15 to 318.15)K: Determination and modelling. *The Journal of Chemical Thermodynamics, 108*, 45-58. doi:https://doi.org/10.1016/j.jct.2017.01.006

299. Wang, J., Xu, R., & Xu, A. (2017c). Solubility determination and thermodynamic functions of 2-chlorophenothiazine in nine organic solvents

from T=283.15K to T=318.15K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 106*, 132-144. doi:https://doi.org/10.1016/j.jct.2016.11.029

300. Wang, L.-H., & Cheng, Y.-Y. (2005). Solubility of Puerarin in Water, Ethanol, and Acetone from (288.2 to 328.2) K. *Journal of Chemical & Engineering Data, 50*(4), 1375-1376. doi:10.1021/je050076g

301. Wang, L., Wang, J., Bao, Y., & Li, T. (2007). Solubility of Irbesartan (Form A) in Different Solvents between 278 K and 323 K. *Journal of Chemical & Engineering Data, 52*(5), 2016-2017. doi:10.1021/je700296x

302. Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., . . . Cao, D.-S. (2016). ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *Journal of Chemical Information and Modeling, 56*(4), 763-773. doi:10.1021/acs.jcim.5b00642

303. Wang, S., Li, Q.-S., Li, Z., & Su, M.-G. (2007). Solubility of Xylitol in Ethanol, Acetone, N,N-Dimethylformamide, 1-Butanol, 1-Pentanol, Toluene, 2-Propanol, and Water. *Journal of Chemical & Engineering Data, 52*(1), 186-188. doi:10.1021/je060348v

304. Wang, S., Qin, L., Zhou, Z., & Wang, J. (2012). Solubility and Solution Thermodynamics of Betaine in Different Pure Solvents and Binary Mixtures. *Journal of Chemical & Engineering Data, 57*(8), 2128-2135. doi:10.1021/je2011659

305. Wang, Y. (2009). Molecular Complexity Effects and Fingerprint-Based Similarity Search Strategies.

306. Wang, Y., Lu, C., Liu, H. M., & Wang, Y. J. (2016). Fault Diagnosis for Centrifugal Pumps Based on Complementary Ensemble Empirical Mode Decomposition, Sample Entropy and Random Forest. *Proceedings of the 2016 12th World Congress on Intelligent Control and Automation (Wcica)*, 1317-1320.

307. Wang, Y. N., Fu, S. X., Jia, Y. X., Qian, C., & Chen, X. Z. (2013). Solubility of Benzyl Disulfide in Five Organic Solvents between (283.45 and 333.15) K. *Journal of Chemical and Engineering Data, 58*(9), 2483-2486. doi:10.1021/je400326r

308. Warr, W. A. (2011). Representation of chemical structures. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 1*(4), 557-579. doi:doi:10.1002/wcms.36

309. Watterson, S., Hudson, S., Svärd, M., & Rasmuson, Å. C. (2014). Thermodynamics of fenofibrate and solubility in pure organic solvents. *Fluid Phase Equilibria, 367*, 143-150. doi:https://doi.org/10.1016/j.fluid.2014.01.029

310. Wei, D., & Pei, Y. (2008). Measurement and Correlation of Solubility of Diphenyl Carbonate in Alkanols. *Industrial & Engineering Chemistry Research, 47*(22), 8953-8956. doi:10.1021/ie801263x

311. Wei, D., Pei, Y., & Yan, F. (2009). Solubility of Salol in Pure Alcohols from (283.15 to 308.15) K. *Journal of Chemical & Engineering Data, 54*(1), 142-143. doi:10.1021/je800668j

312. Wei, D., Wang, L., Liu, C., & Wang, B. (2010). β-Sitosterol Solubility in Selected Organic Solvents. *Journal of Chemical & Engineering Data, 55*(8), 2917-2919. doi:10.1021/je9009909

313. Wei, L. Y., & Zou, Q. (2016). Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition. *International Journal of Molecular Sciences, 17*(12). doi:ARTN 2118

314. 10.3390/ijms17122118

315. Wells, A. F. (1946). XXI. Crystal habit and internal structure.—I. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 37*(266), 184-199. doi:10.1080/14786444608561072

316. Wen, X. H., Tan, F., Jing, Z. J., & Liu, Z. Y. (2004). Preparation and study the 1 : 2 inclusion complex of carvedilol with beta-cyclodextrin. *Journal of Pharmaceutical and Biomedical Analysis, 34*(3), 517-523. doi:10.1016/S0731-7085(03)00576-4

317. Whitesell, J. K. (1998). The Merck Index, 12th Edition, CD-ROM (Macintosh): An encyclopedia of chemicals, drugs & biologicals. *Journal of the American Chemical Society, 120*(9), 2209-2209. doi:DOI 10.1021/ja975911w

318. Wicker, J. G. P., Crowley, L. M., Robshaw, O., Little, E. J., Stokes, S. P., Cooper, R. I., & Lawrence, S. E. (2017). Will they co-crystallize? *CrystEngComm, 19*(36), 5336-5340. doi:10.1039/c7ce00587c

319. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82. doi:DOI 10.3354/cr030079

320. Wilson, R. (2017). *Machine Learning: A Visual Beginners Guide to Machine Learning With Python, Data Science, Tensorflow, Artificial Intelligence, Random Forests and Decision Trees*: CreateSpace Independent Publishing Platform.

321. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining Practical Machine Learning Tools and Techniques Third Edition Preface. *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition*, Xxi-+.

322. Wolk, O., Agbaria, R., & Dahan, A. (2014). Provisional in-silico biopharmaceutics classification (BCS) to guide oral drug product development. *Drug Design Development and Therapy, 8*, 1563-1575. doi:10.2147/Dddt.S68909

323. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259. doi:https://doi.org/10.1016/S0893-6080(05)80023-1

324. Wu, Y., Di, Y., Zhang, X., & Zhang, Y. (2016). Solubility determination and thermodynamic modeling of 3-methyl-4-nitrobenzoic acid in twelve organic

solvents from T=(283.15–318.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 102*, 257-269. doi:https://doi.org/10.1016/j.jct.2016.07.023

325. XLSTAT. (2017). Data Analysis and Statistical Solution for Microsoft Excel (Version 2017). Paris, France: Addinsoft.

326. Xu, B. X., Huang, J. Z., Williams, G., Wang, Q., & Ye, Y. M. (2012). Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces. *International Journal of Data Warehousing and Mining, 8*(2), 44-63. doi:10.4018/jdwm.2012040103

327. Xu, D., & Redman-Furey, N. (2007). Statistical cluster analysis of pharmaceutical solvents. *International Journal of Pharmaceutics, 339*(1-2), 175-188. doi:10.1016/j.ijpharm.2007.03.002

328. Xu, R., Wang, J., Du, C., Han, S., Meng, L., & Zhao, H. (2016). Solubility determination and thermodynamic dissolution functions of 1,3-diphenylguanidine in nine organic solvents at evaluated temperatures. *The Journal of Chemical Thermodynamics, 99*, 86-95. doi:https://doi.org/10.1016/j.jct.2016.03.011

329. Xu, R., Wang, J., Han, S., Du, C., Meng, L., & Zhao, H. (2016). Solubility modelling and thermodynamic dissolution functions of phthalimide in ten organic solvents. *The Journal of Chemical Thermodynamics, 94*, 160-168. doi:https://doi.org/10.1016/j.jct.2015.10.024

330. Xu, R., Xu, A., Du, C., Cong, Y., & Wang, J. (2016). Solubility determination and thermodynamic modeling of 2,4-dinitroaniline in nine organic solvents from T=(278.15 to 318.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 102*, 178-187. doi:https://doi.org/10.1016/j.jct.2016.07.009

331. Yang, H., & Rasmuson, Å. C. (2010). Solubility of Butyl Paraben in Methanol, Ethanol, Propanol, Ethyl Acetate, Acetone, and Acetonitrile.

Journal of Chemical & Engineering Data, 55(11), 5091-5093. doi:10.1021/je1006289

332. Yang, J., Qian, G., Wu, Y., Zhang, X., Liu, Y., Li, Z., & Zhou, X. (2017). Thermodynamic analysis of the solubility of polymorphic cytarabine in a variety of pure solvents. *Fluid Phase Equilibria, 445*, 1-6. doi:https://doi.org/10.1016/j.fluid.2017.04.021

333. Yang, W., Chen, Z., Jiang, X., Hu, Y., Li, Y., Shi, Y., & Wang, J. (2013). Solubility of succinic anhydride in different pure solvents and binary solvent mixtures with the temperature range from 278.15 to 333.15K. *Journal of Molecular Liquids, 180*, 7-11. doi:https://doi.org/10.1016/j.molliq.2012.12.027

334. Yang, W., Fan, S., Guo, Q., Hao, J., Li, H., Yang, S., . . . Hu, Y. (2016). Thermodynamic models for determination of the solubility of 4-(4-aminophenyl)-3-morpholinone in different pure solvents and (1,4-dioxane+ethyl acetate) binary mixtures with temperatures from (278.15 to 333.15)K. *The Journal of Chemical Thermodynamics, 97*, 214-220. doi:https://doi.org/10.1016/j.jct.2016.01.022

335. Yang, W., Hu, Y., Chen, Z., Jiang, X., Wang, J., & Wang, R. (2012). Solubility of itaconic acid in different organic solvents: Experimental measurement and thermodynamic modeling. *Fluid Phase Equilibria, 314*, 180-184. doi:https://doi.org/10.1016/j.fluid.2011.09.027

336. Yang, Y., Hu, Y., Zhang, Q., Cheng, L., Cao, C., Yang, W., & Shen, F. (2015). Experimental measurement and thermodynamic models for solid–liquid equilibrium of hyodeoxycholic acid in different organic solvents. *Journal of Molecular Liquids, 202*, 17-22. doi:10.1016/j.molliq.2014.12.007

337. Yang, Y., Tang, W., Li, X., Han, D., Liu, Y., Du, S., . . . Gong, J. (2017). Solubility of Benzoin in Six Monosolvents and in Some Binary Solvent Mixtures at Various Temperatures. *Journal of Chemical & Engineering Data, 62*(10), 3071-3083. doi:10.1021/acs.jced.7b00238

338. Yao, G., Li, Z., Xia, Z., & Yao, Q. (2016). Solubility of N-phenylanthranilic acid in nine organic solvents from T=(283.15 to 318.15)K: Determination and modelling. *The Journal of Chemical Thermodynamics, 103*, 218-227. doi:https://doi.org/10.1016/j.jct.2016.08.017

339. Yao, G., Yao, Q., Xia, Z., & Li, Z. (2017). Solubility determination and thermodynamic modelling of 3,5-dimethylpyrazole in nine organic solvents from T=(283.15 to 313.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 110*, 99-109. doi:https://doi.org/10.1016/j.jct.2017.02.011

340. Yap, C. W. J. J. o. c. c. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *32*(7), 1466-1474.

341. Ye, Y. M., Wu, Q. Y., Huang, J. Z. X., Ng, M. K., & Li, X. T. (2013). Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition, 46*(3), 769-787. doi:10.1016/j.patcog.2012.09.005

342. Yin, D.-p., Liu, M.-x., Fu, H.-l., Shu, G., Zhou, J.-y., Qing, X.-y., & Wu, W.-b. (2016). Solubility of Trimethoprim in Selected Pure Solvents and (Water + Ethanol/2-Propanol) Mixed-Solvent Systems. *Journal of Chemical & Engineering Data, 61*(1), 404-411. doi:10.1021/acs.jced.5b00616

343. Yuan, F., Wang, Y., Xiao, L., Huang, Q., Xu, J., Jiang, C., & Hao, H. (2016). Solubility of cefoxitin acid in different solvent systems. *The Journal of Chemical Thermodynamics, 103*, 125-133. doi:https://doi.org/10.1016/j.jct.2016.07.049

344. Zhang, C.-L., Li, B.-Y., & Wang, Y. (2010). Solubilities of norfloxacin in ethanol, 1-propanol, acetone, and chloroform from 294.15 to 318.15 K. *The Canadian Journal of Chemical Engineering, 88*(1), 63-66. doi:10.1002/cjce.20247

345. Zhang, C.-L., Wang, F.-A., & Wang, Y. (2007). Solubilities of Sulfadiazine, Sulfamethazine, Sulfadimethoxine, Sulfamethoxydiazine,

Sulfamonomethoxine, Sulfamethoxazole, and Sulfachloropyrazine in Water from (298.15 to 333.15) K. *Journal of Chemical and Engineering Data - J CHEM ENG DATA, 52*. doi:10.1021/je0603978

346. Zhang, F., Tang, Y., Wang, L., Xu, L., & Liu, G. (2015). Solubility Measurement and Correlation for 2-Naphthaldehyde in Pure Organic Solvents and Methanol + Ethanol Mixtures. *Journal of Chemical & Engineering Data, 60*(8), 2502-2509. doi:10.1021/acs.jced.5b00377

347. Zhang, H., Yin, Q., Liu, Z., Gong, J., Bao, Y., Zhang, M., . . . Xie, C. (2014). Measurement and correlation of solubility of dodecanedioic acid in different pure solvents from T=(288.15 to 323.15)K. *The Journal of Chemical Thermodynamics, 68*, 270-274. doi:https://doi.org/10.1016/j.jct.2013.09.012

348. Zhang, J., Wang, Y., Wang, G., Hao, H., Wang, H., Luan, Q., & Jiang, C. (2014). Determination and correlation of solubility of spironolactone form II in pure solvents and binary solvent mixtures. *The Journal of Chemical Thermodynamics, 79*, 61-68. doi:https://doi.org/10.1016/j.jct.2014.07.011

349. Zhang, J., Yang, X., Han, Y., Li, W., & Wang, J. (2012). Measurement and correlation for solubility of levofloxacin in six solvents at temperatures from 288.15 to 328.15K. *Fluid Phase Equilibria, 335*, 1-7. doi:https://doi.org/10.1016/j.fluid.2012.05.027

350. Zhang, L., Huang, Z., Wan, X., Li, J., & Liu, J. (2012). Measurement and Correlation of the Solubility of Febuxostat in Four Organic Solvents at Various Temperatures. *Journal of Chemical & Engineering Data, 57*(11), 3149-3152. doi:10.1021/je300647k

351. Zhang, Q., Hu, Y., Shi, Y., Yang, Y., Cheng, L., Cao, C., & Yang, W. (2014). Thermodynamic Models for Determination of the Solubility of Dibenzothiophene in Different Solvents at Temperatures from (278.15 to 328.15) K. *Journal of Chemical & Engineering Data, 59*(9), 2799-2804. doi:10.1021/je500437m

352. Zhang, Y., Guo, F., Cui, Q., Lu, M., Song, X., Tang, H., & Li, Q. (2016). Measurement and Correlation of the Solubility of Vanillic Acid in Eight Pure and Water + Ethanol Mixed Solvents at Temperatures from (293.15 to 323.15) K. *Journal of Chemical & Engineering Data, 61*(1), 420-429. doi:10.1021/acs.jced.5b00619

353. Zhang, Y. L., & Yang, Y. H. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics, 187*(1), 95-112. doi:10.1016/j.jeconom.2015.02.006

354. Zhao, Y., Jiang, X., & Hou, B. (2010). Measurement and Correlation of Solubility of Cefuroxime Acid in Pure and Binary Solvents at Various Temperatures. *Journal of Chemical & Engineering Data, 55*(9), 3369-3372. doi:10.1021/je100397q

355. Zheng, S., Han, Y., Zhang, J., & Li, W. (2017). Determination and correlation of solubility of linezolid form II in different pure and binary solvents. *Fluid Phase Equilibria, 432*, 18-27. doi:https://doi.org/10.1016/j.fluid.2016.10.021

356. Zhong, J., Tang, N., Asadzadeh, B., & Yan, W. (2017). Measurement and Correlation of Solubility of Theobromine, Theophylline, and Caffeine in Water and Organic Solvents at Various Temperatures. *Journal of Chemical & Engineering Data, 62*(9), 2570-2577. doi:10.1021/acs.jced.7b00065

357. Zhou, G., Du, C., Han, S., Meng, L., Wang, J., Li, R., & Zhao, H. (2015). Solubility measurement and modelling of 1,8-dinitronaphthalene in nine organic solvents from T=(273.15 to 308.15)K and mixing properties of solutions. *The Journal of Chemical Thermodynamics, 90*, 259-269. doi:https://doi.org/10.1016/j.jct.2015.07.010

358. Zhou, Z. P., Li, X. C., & Zare, R. N. (2017). Optimizing Chemical Reactions with Deep Reinforcement Learning. *Acs Central Science, 3*(12), 1337-1344. doi:10.1021/acscentsci.7b00492

359. Zhu, P., Chen, Y., Fang, J., Wang, Z., Xie, C., Hou, B., . . . Xu, F. (2016). Solubility and solution thermodynamics of thymol in six pure organic solvents. *The Journal of Chemical Thermodynamics, 92*, 198-206. doi:https://doi.org/10.1016/j.jct.2015.09.010

360. Zi, J., Peng, B., & Yan, W. (2007). Solubilities of rutin in eight solvents at T=283.15, 298.15, 313.15, 323.15, and 333.15K. *Fluid Phase Equilibria, 261*(1), 111-114. doi:https://doi.org/10.1016/j.fluid.2007.07.030

Chapter 10. APPENDICES

10.1 APPENDIX 1

Table 10.1 Solubility data collated from literature of various drug compounds in ethanol taken at lab temperature

Compound	Mol. F	Reference
1,5-dinitronaphthalene	0.0078	(G. Zhou et al., 2015)
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0294	(Jouyban, 2009)
1,3-Dimethylurea	0.3960	(Jouyban, 2009)
2-amino-4-chloro-6-methoxypyrimidine	0.0040	(Jouyban, 2009)
2-Amino-4-chlorobenzoic acid	0.0191	(Jouyban, 2009)
2-Chlorophenothiazine	0.0055	(J. Wang, R. Xu, & A. Xu, 2017c)
2-Isopropylimidazole	0.2528	(Jiao Chen, Chen, Cong, Du, & Zhao, 2017)
2-methyl-4-nitroaniline	0.0152	(X. Li, Cong, Du, & Zhao, 2017)
2-methyl-6-nitroaniline	0.0233	(Jouyban, 2009)
2-Naphthaldehyde	0.0653	(F. Zhang, Tang, Wang, Xu, & Liu, 2015)
2-nitro-p-phenylenediamine	0.0089	(J. Wang, A. Xu, & R. Xu, 2017b)
3-methyl-4-nitrobenzoic acid	0.0102	(Wu, Di, Zhang, & Zhang, 2016)
3-nitro-o-toluic acid	0.0286	(Jouyban, 2009)
4-Amino-3,6-Dichloropyridazine	0.0154	(Jouyban, 2009)
4-methyl-2-nitroaniline	0.0194	(X. Li, Y. Cong, et al., 2017)
4-nitrobenzaldehyde	0.0310	(Jouyban, 2009)
5,5-Diethylbarbituric acid (Barbital)	0.0292	(Jouyban, 2009)
5-Ethyl-5-(1-methylpropyl)-barbituric acid (Butabarbital)	0.0215	(Jouyban, 2009)
5-Ethyl-5-(2-methylbutyl)-barbituric acid (Pentobarbital)	0.0572	(Jouyban, 2009)
5-Ethyl-5-(3-methylbutyl)-barbituric acid (Amobarbital)	0.0526	(Jouyban, 2009)
5-Ethyl-5-isopropylbarbituric acid (Probarbital)	0.0188	(Jouyban, 2009)
5-Ethyl-5-pentylbarbituric acid	0.1028	(Jouyban, 2009)
5-Ethyl-5-phenylbarbituric acid (Phenobarbital)	0.0309	(Jouyban, 2009)
Acetylsalicylic acid	0.0855	(Tully, Hou, & Glennon, 2016)
Aloe-emodin	0.0001	(Jouyban, 2009)
Anhydrous citric acid	0.1436	(Jouyban, 2009)
Apigenin	0.0003	(Jouyban, 2009)

apremilast	0.0001	(Jouyban, 2009)
Artemisnin	0.0048	(Jouyban, 2009)
Aspartame	0.0001	(Jouyban, 2009)
Atractylenolide III	0.0081	(Jouyban, 2009)
Atrazine	0.0034	(Jia, Li, Li, & Li, 2013)
Benzoic acid	0.1789	(Perlovich & Bauer-Brandl, 2003)
Betulin	0.0009	(Jouyban, 2009)
Butyl paraben	0.3600	(H. Yang & Rasmuson, 2010)
Caffeic acid	0.0184	(Ji, Meng, Ding, et al., 2016)
Caffeine (form 1)	0.0017	(A. Shalmashi & Golmohammad, 2010)
Carbamazepine	0.0059	(W. Liu, L. Dang, S. Black, & H. Wei, 2008)
Cefradine Form I	0.0003	(X. Hu, Wang, Xie, Wang, & Hao, 2013)
Cefuroxime Acid	0.0034	(Zhao, Jiang, & Hou, 2010)
Coumarin	0.0533	(X. Huang et al., 2015)
Daidzein	0.0004	(Jouyban, 2009)
Danazol	0.0039	(Fathi-Azarbayjani, Abbasi, Vaez-Gharamaleki, & Jouyban, 2016)
Dehydroepiandrosterone acetate	0.0057	(Jouyban, 2009)
Diclofenac	0.0087	(Jouyban, 2009)
Diflunisal	0.0191	(Jouyban, 2009)
Diphenyl Carbonate	0.0134	(D. Wei & Pei, 2008)
Eflucimibe (form A)	0.0007	(Jouyban, 2009)
eszopiclone	0.0004	(Jouyban, 2009)
ferulic acid Form I	0.0241	(Shakeel, Salem-Bekhit, Haq, & Siddiqui, 2017)
Flurbiprofen	0.0612	(Jouyban, 2009)
Gallic acid	0.0604	(Vilas Boas, 2017)
Haloperidol	0.0057	(Jouyban, 2009)
Ibuprofen	0.1422	(Jouyban, 2009)
isatin	0.0041	(J.-Q. Liu, Chen, & Ji, 2014)
Isonicotinamide (form II)	0.0316	(B. Li et al., 2016)
Ketoprofen	0.0640	(Jouyban, 2009)
l-(+)-Ascorbic acid	0.0027	(Anvar Shalmashi & Eliassi, 2008)
Lactose	0.0001	(Jouyban, 2009)
Lamivudine (form 2)	0.0029	(Jouyban, 2009)
linezolid form II	0.0017	(Zheng, Han, Zhang, & Li, 2017)
Loratadine	0.0232	(Jouyban, 2009)
Lovastatin	0.0035	(H. Sun, Gong, & Wang, 2005)
Luteolin	0.0019	(Jouyban, 2009)

Maltitol	0.0001	(Z. Li et al., 2016)
Mannitol	0.0001	(Jouyban, 2009)
Mefenamic acid	0.0019	(Abdul Mudalip, Abu Bakar, Jamal, & Adam, 2013)
Meloxicam	0.0001	(Jouyban, 2009)
Methyl D-(-)-4-hydroxy-phenylglycinate	0.0032	(Jouyban, 2009)
Methyl paraben	0.1470	(Jouyban, 2009)
Naproxen	0.0201	(Jouyban, 2009)
Nifedipine	0.0042	(Jouyban, 2009)
Niflumic acid	0.0110	NIIST database (Domańska, Pobudkowska, & Pelczarska, 2011)
Nimesulide	0.0006	(Jouyban, 2009)
Norfloxacin	0.0001	(C.-L. Zhang, Li, & Wang, 2010)
N-phenylanthranilic acid	0.0153	(Yao, Li, Xia, & Yao, 2016)
Oleanolic acid	0.0008	(Jouyban, 2009)
Oxcarbazepine	0.0004	(Nam et al., 2017)
paclobutrazol	0.0149	(Jouyban, 2009)
Paracetamol	0.0601	(Jouyban, 2009)
para-tert Butylbenzoic acid	0.0882	(Aniya, De, Mohammed, Thella, & Satyavathi, 2017)
Phenacetinum	0.0184	(Chang, Li, Wang, & Tian, 2007)
Phenothiazine	0.0089	(Gracin & Rasmuson, 2002)
p-Hydroxybenzoic acid	0.1261	(Jouyban, 2009)
p-Hydroxyphenylacetic acid	0.3074	(Jouyban, 2009)
Pimetic Acid	0.1208	(H. Li et al., 2010)
Pimozide	0.0012	(Jouyban, 2009)
Piroxicam	0.0001	(Jouyban, 2009)
pronamide	0.0131	(Jouyban, 2009)
pyraclostrobin	0.0117	(Jouyban, 2009)
risperidone form I	0.0020	(Mealey, Svärd, & Rasmuson, 2014)
Salicylamide	0.0333	(Nordström & Rasmuson, 2006)
Salicylic acid	0.1100	(Matsuda et al., 2009)
sorbic acid	0.0643	(Fang et al., 2015)
Spironolactone Form II	0.0039	(J. Zhang et al., 2014)
Sulfadiazine	0.0001	(Mauger, Paruta, & Gerraughty, 1972)
Sulfadimethoxine	0.0007	(C.-L. Zhang, Wang, & Wang, 2007)
Sulfamethoxypyridazine	0.0013	(Y. Hu et al., 2014)
Sulfisomidine	0.0006	(El-Badry, Haq, Fetih, & Shakeel, 2014)
tebuconazole	0.0299	(Shakeel, Haq, Shazly, Alanazi, & Alsarra, 2015)
Temazepam	0.0030	(W. Li, Chen, Han, Du, & Zhao, 2016)

Tetraethyl ranelate	0.0018	(Jouyban, 2009)
Triclosan	0.4490	(W. Tang et al., 2015)
tridecanedioic acid	0.0044	(Q.-S. Li, Li, & Wang, 2008)
Trimethoprim	0.0009	NIIST database
Vanillic Acid	0.0345	(Y. Zhang et al., 2016)
Xanthene	0.0062	(MonÁrrez, Stovall, Woo, Taylor, & Acree, 2002)
Dodecanedioic acid	0.0145	(H. Zhang et al., 2014)
3,5-dimethylpyrazole	0.1626	(Yao, Yao, Xia, & Li, 2017)
5-amino-3-methyl-1-phenylpyrazole	0.1068	(G. Chen, Chen, Jian, & Zhao, 2017)
5-phenyltetrazole	0.0200	(G. Chen, J. Chen, C. Cheng, Y. Cong, P. Jian, et al., 2017)
Dibenzothiophene	0.0075	(Q. Zhang et al., 2014)
4-hydroxybenzaldehyde	0.1146	(J. Wang, A. Xu, & R. Xu, 2017a)
Betaine	0.0372	(S. Wang, Qin, Zhou, & Wang, 2012)
Ampelopsin	0.0001	(Jouyban, 2009)
2-Cyanoguanidine	0.0059	(Ren, Duan, & Yang, 2014)
l-malic acid	0.1872	(Kai et al., 2013)
thiomalic acid	0.2046	(DELGADO, R. HOLGUIN, & MARTÍNEZ, 2012)
chlocyphos	0.0069	(F. Sun, Kang, Zhang, Liu, & Zhang, 2012)
itaconic acid	0.0782	(W. Yang et al., 2012)
N-methyl-3,4,5-trinitropyrazole	0.0180	(Guo et al., 2017)
Flufenamic acid	0.0681	(Alshehri & Shakeel, 2017)
Lornoxicam	0.0001	(Shakeel, Haq, Alanazi, & Alsarra, 2015)
4-Cyanobenzoic acid	0.0214	(H. Wang & Zhang, 2009)
2,4-Dichlorobenzoic acid	0.0581	(H. Wang & Zhang, 2009)
3,4-Dimethoxybenzoic acid	0.0072	(Jouyban, 2009)
Lansoprazole	0.0221	(Hong, Xu, Ren, Chen, & Qi, 2012)
cytarabine	0.0003	(J. Yang et al., 2017)
pyrazinamide	0.0028	(Jouyban, 2009)
Syringic Acid	0.0169	(Jouyban, 2009)
Tenoxicam	0.0001	(Jouyban, 2009)
Benzoin	0.0025	(Y. Yang et al., 2017)
4-Nitrophthalimide	0.0020	(Jouyban, 2009)
p-Toluenesulfonamide	0.0005	(Jouyban, 2009)
o-Toluenesulfonamide	0.0003	(Jouyban, 2009)
Theobromine	0.0001	(Meng et al., 2013)
Theophylline	0.0012	(Zhu et al., 2016)

Protocatechuic Acid	0.1425	(Vilas Boas, 2017)
Gentisic Acid	0.0376	(Vilas Boas, 2017)
S-Hesperetin	0.0044	(Jouyban, 2009)
Ethyl Paraben	0.9098	(Jouyban, 2009)
Phenylacetic acid	0.4112	(A. Liang, Wang, & Qu, 2017)
p-Aminophenylacetic acid	0.0005	(Gracin & Rasmuson, 2002)
Triclocarban	0.0015	(W. Tang et al., 2015)
4-(4-aminophenyl)-3-morpholinone	0.0032	(W. Yang et al., 2016)
hyodeoxycholic acid	0.0472	(H. Li et al., 2016)
4-Aminobenzoic acid	0.0506	(Jouyban, 2009)
Vanillin	0.0794	(Jouyban, 2009)
4-Acetoxybenzoic Acid	0.0249	(H. Wang & Zhang, 2009)
3,5-Diaminobenzoic Acid	0.0033	(Jouyban, 2009)
3,5-Dimethoxybenzoic Acid	0.0049	(H. Wang & Zhang, 2009)
(1-benzyl-1H-1,2,3-triazole-4-yl)methanol	0.0833	(S. Liang et al., 2016)
1,3-diphenylguanidine	0.0315	(R. Xu, Wang, Du, et al., 2016)
2,4-dinitroaniline	0.0028	(R. Xu, Xu, Du, Cong, & Wang, 2016)
Clopidogrel	0.0107	(Song, Li, & Gong, 2010)
Dipyrone	0.0037	(Ding et al., 2017)
Metaxalone (Form B)	0.0062	(Hong, Wu, Qi, & Ren, 2016)
p-Coumaric Acid	0.0456	(Ji, Meng, Li, et al., 2016)
phthalimide	0.0024	(R. Xu, Wang, Han, et al., 2016)
Propylparaben	0.1970	(Jouyban, 2009)
Pyridazine-3-amine	0.0237	(Cao et al., 2012)
Salol	0.1272	(D. Wei, Pei, & Yan, 2009)
thymol	0.6361	(DELGADO et al., 2012)
3-amino-1,2,4-triazole	0.0240	(X. Li, Du, Cong, & Zhao, 2017)
2-amino-5-methylthiazole	0.0950	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Difloxacin	0.0507	(Baluja, Bhalodia, Gajera, Vekariya, & Bhatt, 2009)
Flunixin meglumine	0.0058	(Qin et al., 2015)
Formononetin	0.0003	(Dong et al., 2017)
Irbesartan (form A)	0.0011	(L. Wang, Wang, Bao, & Li, 2007)
L-Carnitine	0.0483	(D. Sun et al., 2014)
Piracetam (Form III)	0.0550	(Jouyban, 2009)
Sulfanilic Acid	0.0003	(Mauger et al., 1972)
Xylitol	0.0024	(S. Wang, Li, Li, & Su, 2007)
4',5,7-Triacetoxyflavanone	0.0013	(Shan, Fu, & Yan, 2017)
4-Methylsulfonylacetophenone	0.0197	(Hao et al., 2017)
Cefoxitin acid	0.0009	(Yuan et al., 2016)

Febuxostat	0.0018	(L. Zhang, Huang, Wan, Li, & Liu, 2012)
levofloxacin	0.0011	(J. Zhang, Yang, Han, Li, & Wang, 2012)
D-Pantolactone	0.5062	(C. Huang et al., 2015)
Rhein	0.0001	(Cheng, Wang, Zhang, & Wang, 2015)

Table 10.2 Solubility data collated from literature of various drug compounds in methanol taken at lab temperature

Compound	Mol. F	Reference
1,5-dinitronaphthalene	0.0008	(G. Zhou et al., 2015)
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0312	(Jouyban, 2009)
1,3-Dimethylurea	0.4480	(Jouyban, 2009)
2-amino-4-chloro-6-methoxypyrimidine	0.0037	(Jouyban, 2009)
2-Chlorophenothiazine	0.0059	(J. Wang et al., 2017c)
2-Isopropylimidazole	0.0826	(J. Chen et al., 2017)
2-methyl-4-nitroaniline	0.0168	(X. Li, Cong, Cunbin, & Hongkun, 2016)
2-methyl-6-nitroaniline	0.0172	(Jouyban, 2009)
2-Naphthaldehyde	0.0274	(F. Zhang et al., 2015)
2-nitro-p-phenylenediamine	0.0104	(J. Wang et al., 2017b)
3-methyl-4-nitrobenzoic acid	0.0104	(Wu et al., 2016)
3-nitro-o-toluic acid	0.0319	(Jouyban, 2009)
4-Amino-3,6-Dichloropyridazine	0.0153	(Jouyban, 2009)
4-methyl-2-nitroaniline	0.0169	(X. Li, Wang, Cong, Du, & Zhao, 2017)
4-nitrobenzaldehyde	0.0731	(Jouyban, 2009)
5,5-Diethylbarbituric acid (Barbital)	0.0352	(Jouyban, 2009)
5-Ethyl-5-(1-methylpropyl)-barbituric acid (Butabarbital)	0.0246	(Jouyban, 2009)
5-Ethyl-5-(2-methylbutyl)-barbituric acid (Pentobarbital)	0.0531	(Jouyban, 2009)
5-Ethyl-5-(3-methylbutyl)-barbituric acid (Amobarbital)	0.0498	(Jouyban, 2009)
5-Ethyl-5-isopropylbarbituric acid (Probarbital)	0.0257	(Jouyban, 2009)
5-Ethyl-5-pentylbarbituric acid	0.0901	(Jouyban, 2009)
5-Ethyl-5-phenylbarbituric acid (Phenobarbital)	0.0421	(Jouyban, 2009)
Acetylsalicylic acid	0.0719	(Maia & Giulietti, 2008)
Aloe-emodin	0.0001	(Jouyban, 2009)

Anhydrous citric acid	0.1575	(Jouyban, 2009)
Apigenin	0.0001	(Jouyban, 2009)
apremilast	0.0001	(Jouyban, 2009)
Artemisnin	0.0014	(Jouyban, 2009)
Aspartame	0.0008	(Jouyban, 2009)
Atrazine	0.0033	(Jia et al., 2013)
Benzoic acid	0.1632	(Thati, Nordström, & Rasmuson, 2010)
Betulin	0.0004	(Jouyban, 2009)
Butyl paraben	0.3360	(H. Yang & Rasmuson, 2010)
Caffeic acid	0.0178	(Ji, Meng, Ding, et al., 2016)
Caffeine (form 1)	0.0020	(A. Shalmashi & Golmohammad, 2010)
Carbamazepine	0.0129	(W. Liu et al., 2008)
Cefradine Form I	0.0008	(X. Hu et al., 2013)
Coumarin	0.0802	(X. Huang et al., 2015)
Daidzein	0.0003	(Jouyban, 2009)
Dehydroepiandrosterone acetate	0.0036	(Jouyban, 2009)
Diclofenac	0.0059	(Jouyban, 2009)
Diflunisal	0.0151	(Jouyban, 2009)
Diphenyl Carbonate	0.0153	(D. Wei & Pei, 2008)
ferulic acid Form I	0.0249	(Shakeel et al., 2017)
Flurbiprofen	0.0478	(Jouyban, 2009)
Gallic acid	0.0678	(Vilas Boas, 2017)
Haloperidol	0.0015	(Jouyban, 2009)
Ibuprofen	0.0247	(Jouyban, 2009)
isatin	0.0057	(J.-Q. Liu et al., 2014)
Ketoprofen	0.0428	(Jouyban, 2009)
l-(+)-Ascorbic acid	0.0108	(Anvar Shalmashi & Eliassi, 2008)
Lactose	0.0001	(Jouyban, 2009)
linezolid form II	0.0031	(Zheng et al., 2017)
Loratadine	0.0339	(Jouyban, 2009)
Lovastatin	0.0031	(H. Sun et al., 2005)
Luteolin	0.0005	(Jouyban, 2009)
Maltitol	0.0003	(Z. Li et al., 2016)
Mannitol	0.0004	(Jouyban, 2009)
Meloxicam	0.0001	(Jouyban, 2009)
Methyl D-(-)-4-hydroxy-phenylglycinate	0.0042	(Jouyban, 2009)
Methyl paraben	0.1210	(Jouyban, 2009)
Naproxen	0.0126	(Jouyban, 2009)

Niflumic acid	0.0076	NIIST database
Nimesulide	0.0011	(Jouyban, 2009)
N-phenylanthranilic acid	0.0083	(Yao et al., 2016)
Oxcarbazepine	0.0008	(Nam et al., 2017)
Paracetamol	0.0658	(Jouyban, 2009)
para-tert Butylbenzoic acid	0.0622	(Aniya et al., 2017)
Phenacetinum	0.0203	(Chang et al., 2007)
Phenothiazine	0.0051	(Gracin & Rasmuson, 2002)
p-Hydroxybenzoic acid	0.1142	(Jouyban, 2009)
p-Hydroxyphenylacetic acid	0.3589	(Jouyban, 2009)
Pimozide	0.0008	(Jouyban, 2009)
Piroxicam	0.0001	(Jouyban, 2009)
pronamide	0.0124	(Jouyban, 2009)
risperidone form I	0.0040	(Mealey et al., 2014)
Salicylamide	0.0406	(Nordström & Rasmuson, 2006)
Salicylic acid	0.1280	(Matsuda et al., 2009)
sorbic acid	0.0553	(Fang et al., 2015)
Spironolactone Form II	0.0006	(J. Zhang et al., 2014)
Sulfadiazine	0.0002	(C.-L. Zhang et al., 2007)
Sulfadimethoxine	0.0012	(C.-L. Zhang et al., 2007)
Sulfamethoxypyridazine	0.0029	(Y. Hu et al., 2014)
Sulfisomidine	0.0011	(Mauger et al., 1972)
tebuconazole	0.0400	(Y. Hu et al., 2014)
Temazepam	0.0055	(W. Li et al., 2016)
Tetraethyl ranelate	0.0039	(Jouyban, 2009)
Triclosan	0.4265	(W. Tang et al., 2015)
tridecanedioic acid	0.0045	(Q.-S. Li et al., 2008)
Trimethoprim	0.0017	(Yin et al., 2016)
Xanthene	0.0045	(MonÁrrez et al., 2002)
3,5-dimethylpyrazole	0.1434	(Yao et al., 2017)
5-amino-3-methyl-1-phenylpyrazole	0.1150	(G. Chen, J. Chen, P. Jian, et al., 2017)
5-phenyltetrazole	0.0209	(G. Chen, J. Chen, C. Cheng, Y. Cong, P. Jian, et al., 2017)
Dibenzothiophene	0.0029	(Q. Zhang et al., 2014)
4-hydroxybenzaldehyde	0.0969	(J. Wang et al., 2017a)
Betaine	0.1308	(S. Wang et al., 2012)
2-Cyanoguanidine	0.0231	(Ren et al., 2014)
thiomalic acid	0.2032	(DELGADO et al., 2012)
chlocyphos	0.0071	(F. Sun et al., 2012)
itaconic acid	0.1093	(W. Yang et al., 2012)

N-methyl-3,4,5-trinitropyrazole	0.0326	(Guo et al., 2017)
Flufenamic acid	0.0439	(Alshehri & Shakeel, 2017)
Lornoxicam	0.0001	(Shakeel, Haq, Alanazi, et al., 2015)
cytarabine	0.0011	(J. Yang et al., 2017)
pyrazinamide	0.0047	(Jouyban, 2009)
Syringic Acid	0.0190	(Jouyban, 2009)
Tenoxicam	0.0001	(Jouyban, 2009)
Benzoin	0.0031	(J. Yang et al., 2017)
4-Nitrophthalimide	0.0032	(Jouyban, 2009)
p-Toluenesulfonamide	0.0006	(Jouyban, 2009)
o-Toluenesulfonamide	0.0004	(Jouyban, 2009)
Theobromine	0.0001	(Zhong, Tang, Asadzadeh, & Yan, 2017)
Theophylline	0.0137	(Zhu et al., 2016)
Genistein	0.0015	(Jouyban, 2009)
Protocatechuic Acid	0.1414	(Vilas Boas, 2017)b
Gentisic Acid	0.0827	(Vilas Boas, 2017)
S-Hesperetin	0.0038	(Jouyban, 2009)
Ethyl Paraben	0.9256	(Jouyban, 2009)
Phenylacetic acid	0.4069	(Gracin & Rasmuson, 2002)
p-Aminophenylacetic acid	0.0010	(Gracin & Rasmuson, 2002)
Triclocarban	0.0006	(DELGADO et al., 2012)
4-(4-aminophenyl)-3-morpholinone	0.0046	(W. Yang et al., 2016)
hyodeoxycholic acid	0.0518	(H. Li et al., 2016)
4-Aminobenzoic acid	0.0539	(Jouyban, 2009)
(1-benzyl-1H-1,2,3-triazole-4-yl)methanol	0.1715	(S. Liang et al., 2016)
2,4-dinitroaniline	0.0028	(R. Xu, Xu, et al., 2016)
Dipyrone	0.0318	(Ding et al., 2017)
p-Coumaric Acid	0.0395	(Ji, Meng, Li, et al., 2016)
phthalimide	0.0026	(R. Xu, Wang, Han, et al., 2016)
Propylparaben	0.1720	(Jouyban, 2009)
Pyridazine-3-amine	0.0455	(Cao et al., 2012)
Salol	0.0946	(D. Wei et al., 2009)
thymol	0.6708	(Zhu et al., 2016)
3-amino-1,2,4-triazole	0.0471	(X. Li, C. Du, et al., 2017)
2-amino-5-methylthiazole	0.1075	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Difloxacin	0.0073	(Baluja et al., 2009)
Dimethyl 1,4-Cyclohexanedione-2,5-dicarboxylate	0.0004	(Jouyban, 2009)

Flunixin meglumine	0.0349	(Qin et al., 2015)
Formononetin	0.0002	(Dong et al., 2017)
L-Carnitine	0.1190	(D. Sun et al., 2014)
Sulfanilic Acid	0.0003	(Y. Hu et al., 2014)v
4',5,7-Triacetoxyflavanone	0.0022	(Shan et al., 2017)
4-Methylsulfonylacetophenone	0.0246	(Hao et al., 2017)
Cefoxitin acid	0.0110	(Yuan et al., 2016)
Febuxostat	0.0009	(L. Zhang et al., 2012)
D-Pantolactone	0.3553	(C. Huang et al., 2015)
Rhein	0.0001	(Cheng et al., 2015)

Table 10.3 Solubility data collated from literature of various drug compounds in 1-butanol taken at lab temperature

Compound	Mol. F	Reference
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0378	(Jouyban, 2009)
1,3-Dimethylurea	0.3690	(Jouyban, 2009)
2-Amino-4-chlorobenzoic acid	0.0157	(Jouyban, 2009)
2-Chlorophenothiazine	0.0044	(J. Wang et al., 2017c)
2-methyl-4-nitroaniline	0.0118	(X. Li et al., 2016)
2-Naphthaldehyde	0.0561	(F. Zhang et al., 2015)
3-methyl-4-nitrobenzoic acid	0.0108	(Wu et al., 2016)
4-methyl-2-nitroaniline	0.0253	(X. Li, M. Wang, et al., 2017)
4-nitrobenzaldehyde	0.0200	(Jouyban, 2009)
5,5-Diethylbarbituric acid (Barbital)	0.0200	(Jouyban, 2009)
5-Ethyl-5-(1-methylpropyl)-barbituric acid (Butabarbital)	0.0119	(Jouyban, 2009)
5-Ethyl-5-(2-methylbutyl)-barbituric acid (Pentobarbital)	0.0576	(Jouyban, 2009)
5-Ethyl-5-(3-methylbutyl)-barbituric acid (Amobarbital)	0.0514	(Jouyban, 2009)
5-Ethyl-5-isopropylbarbituric acid (Probarbital)	0.0060	(Jouyban, 2009)
5-Ethyl-5-pentylbarbituric acid	0.1159	(Jouyban, 2009)
5-Ethyl-5-phenylbarbituric acid (Phenobarbital)	0.0208	(Jouyban, 2009)
Acetylsalicylic acid	0.0453	(Maia & Giulietti, 2008)
Aloe-emodin	0.0001	(Jouyban, 2009)

Apigenin	0.0006	(Jouyban, 2009)
apremilast	0.0002	(Jouyban, 2009)
Artemisnin	0.0025	(Jouyban, 2009)
Atractylenolide III	0.0602	(Jouyban, 2009)
Atrazine	0.0056	(Jia et al., 2013)
Benzoic acid	0.2016	(Thati et al., 2010)
Betulin	0.0024	(Jouyban, 2009)
Butyl paraben	0.3640	(H. Yang & Rasmuson, 2010)
Caffeic acid	0.0103	(Ji, Meng, Ding, et al., 2016)
Carbamazepine	0.0005	(W. Liu et al., 2008)
Cefradine Form I	0.0005	(X. Hu et al., 2013)
Dehydroepiandrosterone acetate	0.0107	(Jouyban, 2009)
Diflunisal	0.0266	(Jouyban, 2009)
Diphenyl Carbonate	0.0155	(D. Wei & Pei, 2008)
ferulic acid Form I	0.0161	(Shakeel et al., 2017)
Flurbiprofen	0.0667	(Jouyban, 2009)
Haloperidol	0.0078	(Jouyban, 2009)
isatin	0.0049	(J.-Q. Liu et al., 2014)
Isonicotinamide (form II)	0.0421	(B. Li et al., 2016)
Ketoprofen	0.0868	(Jouyban, 2009)
Lactose	0.0514	(Jouyban, 2009)
Lamivudine (form 2)	0.0022	(Jouyban, 2009)
Loratadine	0.0416	(Jouyban, 2009)
Luteolin	0.0018	(Jouyban, 2009)
Meloxicam	0.0001	(Jouyban, 2009)
Methyl D-(-)-4-hydroxy-phenylglycinate	0.0164	(Jouyban, 2009)
Methyl paraben	0.1460	(Jouyban, 2009)
Naproxen	0.0142	(Jouyban, 2009)
Nimesulide	0.0006	(Jouyban, 2009)
N-phenylanthranilic acid	0.0187	(Yao et al., 2016)
Oleanolic acid	0.0035	(Jouyban, 2009)
Oxcarbazepine	0.0004	(Nam et al., 2017)
paclobutrazol	0.0224	(Jouyban, 2009)
Paracetamol	0.0392	(Jouyban, 2009)
Phenacetinum	0.0289	(Chang et al., 2007)
Phenothiazine	0.0110	(Chang et al., 2007)
Pimozide	0.0039	(Jouyban, 2009)
Piroxicam	0.0514	(Jouyban, 2009)
pyraclostrobin	0.0132	(Jouyban, 2009)
risperidone form I	0.0035	(Mealey et al., 2014)

sorbic acid	0.0973	(Fang et al., 2015)
Spironolactone Form II	0.0047	(J. Zhang et al., 2014)
Sulfadiazine	0.0001	(C.-L. Zhang et al., 2007)
Sulfadimethoxine	0.0004	(C.-L. Zhang et al., 2007)
Sulfamethoxypyridazine	0.0007	(C.-L. Zhang et al., 2007)
Sulfisomidine	0.0003	(Mauger et al., 1972)
Temazepam	0.0049	(Jouyban, 2009)
Tetraethyl ranelate	0.0014	(Jouyban, 2009)
Trimethoprim	0.0006	(Q.-S. Li et al., 2008)
Vanillic Acid	0.0263	(Y. Zhang et al., 2016)
Xanthene	0.0176	(MonÁrrez et al., 2002)
3,5-dimethylpyrazole	0.1869	(Yao et al., 2017)
Dibenzothiophene	0.0132	(Q. Zhang et al., 2014)
4-hydroxybenzaldehyde	0.1303	(J. Wang et al., 2017a)
Betaine	0.0125	(S. Wang et al., 2012)
l-malic acid	0.1397	(Kai et al., 2013)
thiomalic acid	0.1526	(Meng et al., 2013)
N-methyl-3,4,5-trinitropyrazole	0.0087	(Guo et al., 2017)
Flufenamic acid	0.1190	(Alshehri & Shakeel, 2017)
Lornoxicam	0.0001	(Shakeel, Haq, Alanazi, et al., 2015)
Lansoprazole	0.0027	(Hong et al., 2012)
pyrazinamide	0.0024	(Jouyban, 2009)
Syringic Acid	0.0052	(Jouyban, 2009)
Tenoxicam	0.0001	(Jouyban, 2009)
Benzoin	0.0020	(Y. Yang et al., 2017)
p-Toluenesulfonamide	0.0004	(Jouyban, 2009)
o-Toluenesulfonamide	0.0003	(Jouyban, 2009)
Theophylline	0.0013	(Zhong et al., 2017)
Agomelatin (Form II)	0.0451	(Y. Liu, Gao, Ren, & Ren, 2015)
S-Hesperetin	0.0001	(Jouyban, 2009)
Ethyl Paraben	0.9082	(Jouyban, 2009)
Triclocarban	0.0029	(DELGADO et al., 2012)
hyodeoxycholic acid	0.0351	(H. Li et al., 2016)
4-Aminobenzoic acid	0.0314	(Jouyban, 2009)
Vanillin	0.0647	(Jouyban, 2009)
(1-benzyl-1H-1,2,3-triazole-4-yl)methanol	0.0481	(S. Liang et al., 2016)
1,3-diphenylguanidine	0.0378	(R. Xu, Wang, Du, et al., 2016)
2,4-dinitroaniline	0.0033	(R. Xu, Xu, et al., 2016)
Clonazepam	0.0008	(Jouyban, 2009)
Clopidogrel	0.0016	(Song et al., 2010)

Dipyrone	0.0004	(Ding et al., 2017)
p-Coumaric Acid	0.0371	(Ji, Meng, Li, et al., 2016)
phthalimide	0.0029	(R. Xu, Wang, Han, et al., 2016)
Propylparaben	0.2060	(Jouyban, 2009)
Pyridazine-3-amine	0.0170	(Cao et al., 2012)
Salol	0.1817	(D. Wei et al., 2009)
Sulfanilamide	0.0189	(Jouyban, 2009)
thymol	0.5471	(Zhu et al., 2016)
4',5,7-Triacetoxyflavanone	0.0005	(Shan et al., 2017)
4-Methylsulfonylacetophenone	0.0140	(Hao et al., 2017)
levofloxacin	0.0022	(J. Zhang et al., 2012)
Rhein	0.0001	(Cheng et al., 2015)
Rutin	0.0003	(Zi, Peng, & Yan, 2007)
Propyl *p*-hydroxybenzoate	0.2110	(Jouyban, 2009)
Sulfisoxazole	0.0015	(Y. Li, Li, Cong, Du, & Zhao, 2017)

Table 10.4 Solubility data collated from literature of various drug compounds in Acetonitrile taken at lab temperature

Compound	Mol. F	Reference
1,5-dinitronaphthalene	0.0102	(G. Zhou et al., 2015)
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0128	(Jouyban, 2009)
2-amino-4-chloro-6-methoxypyrimidine	0.0043	(Jouyban, 2009)
2-Amino-4-chlorobenzoic acid	0.0030	(Jouyban, 2009)
2-Chlorophenothiazine	0.0053	(J. Wang et al., 2017c)
2-methyl-4-nitroaniline	0.0566	(X. Li et al., 2016)
2-nitro-p-phenylenediamine	0.0286	(J. Wang et al., 2017b)
3-methyl-4-nitrobenzoic acid	0.0041	(Wu et al., 2016)
4-nitrobenzaldehyde	0.1025	(Jouyban, 2009)
Acetylsalicylic acid	0.0185	(Maia & Giulietti, 2008)
Artemisnin	0.0018	(Jouyban, 2009)
Benzoic acid	0.0539	(Thati et al., 2010)
Butyl paraben	0.1899	(H. Yang & Rasmuson, 2010)
Deferiprone	0.0005	(Fathi-Azarbayjani et al., 2016)

Dehydroepiandrosterone acetate	0.0083	(Jouyban, 2009)
Diflunisal	0.0015	(Jouyban, 2009)
Gallic acid	0.0012	(Vilas Boas, 2017)
isatin	0.0067	(J.-Q. Liu et al., 2014)
l-(+)-Ascorbic acid	0.0002	(Anvar Shalmashi & Eliassi, 2008)
Lamivudine (form 2)	0.0002	(Jouyban, 2009)
Loratadine	0.0058	(Jouyban, 2009)
N-phenylanthranilic acid	0.0037	(Yao et al., 2016)
Oxcarbazepine	0.0008	(Nam et al., 2017)
paclobutrazol	0.0054	(Jouyban, 2009)
Paracetamol	0.0074	(Jouyban, 2009)
Phenothiazine	0.0117	(J. Wang et al., 2017c)
Salicylamide	0.0333	(Nordström & Rasmuson, 2006)
Salicylic acid	0.0294	(Maia & Giulietti, 2008)
sorbic acid	0.0112	(Fang et al., 2015)
tebuconazole	0.0134	(Y. Li et al., 2017)
Temazepam	0.0110	(W. Li et al., 2016)
Tetraethyl ranelate	0.0334	(Jouyban, 2009)
Triclosan	0.4840	(DELGADO et al., 2012)
Trimethoprim	0.0004	(Yin et al., 2016)
Xanthene	0.0197	(MonÁrrez et al., 2002)
3,5-dimethylpyrazole	0.0568	(Yao et al., 2017)
5-amino-3-methyl-1-phenylpyrazole	0.1642	(G. Chen, J. Chen, P. Jian, et al., 2017)
Dibenzothiophene	0.0157	(Q. Zhang et al., 2014)
4-hydroxybenzaldehyde	0.0782	(J. Wang et al., 2017a)
l-malic acid	0.0608	(Kai et al., 2013)
thiomalic acid	0.0392	(Meng et al., 2013)
itaconic acid	0.0055	(W. Yang et al., 2012)

pyrazinamide	0.0029	(Jouyban, 2009)
4-Nitrophthalimide	0.0062	(Jouyban, 2009)
p-Toluenesulfonamide	0.0011	(Jouyban, 2009)
o-Toluenesulfonamide	0.0005	(Jouyban, 2009)
Protocatechuic Acid	0.0155	(Vilas Boas, 2017)
Gentisic Acid	0.0086	(Vilas Boas, 2017)
S-Hesperetin	0.0021	(Jouyban, 2009)
Ethyl Paraben	0.0634	(Jouyban, 2009)
Triclocarban	0.0004	(DELGADO et al., 2012)
hyodeoxycholic acid	0.0006	(Y. Yang et al., 2015)
1,3-diphenylguanidine	0.0214	(R. Xu, Wang, Du, et al., 2016)
2,4-dinitroaniline	0.0149	(R. Xu, Xu, et al., 2016)
phthalimide	0.0030	(R. Xu, Wang, Han, et al., 2016)
thymol	0.7535	(Zhu et al., 2016)
3-amino-1,2,4-triazole	0.0019	(X. Li, C. Du, et al., 2017)
2-amino-5-methylthiazole	0.0586	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Dimethyl 1,4-Cyclohexanedione-2,5-dicarboxylate	0.0023	(Jouyban, 2009)
Flunixin meglumine	0.0001	https://www.sciencedirect.com/science/article/pii/S0378381215003453#sec0030
Succinic Anhydride	0.1462	(Jouyban, 2009)
Sulfanilic Acid	0.0002	(Y. Hu et al., 2014)
Cefoxitin acid	0.0006	(Yuan et al., 2016)
2,3-Dichlorophenol	0.5400	(Jouyban, 2009)
2,3-Dimethylphenol	0.3795	(Jouyban, 2009)
2,4,5-Trichlorophenol	0.5124	(Jouyban, 2009)
2,4,6-Trichlorophenol	0.5095	(Jouyban, 2009)
2,5-Dimethylphenol	0.3355	(Jouyban, 2009)
2,6-Dichlorophenol	0.4723	(Jouyban, 2009)
2-Iodophenol	0.7533	(Jouyban, 2009)

2-Nitrophenol	0.5819	(Jouyban, 2009)
3,4-Dichlorophenol	0.6362	(Jouyban, 2009)
3,5-Dimethylphenol	0.5024	(Jouyban, 2009)
3-Cyanophenol	0.3827	(Jouyban, 2009)
3-Nitroaniline	0.1080	(Jouyban, 2009)
3-Nitrophenol	0.4009	(Jouyban, 2009)
4-Aminoacetophenone	0.0215	(Jouyban, 2009)
4-Bromophenol	0.5294	(Jouyban, 2009)
4-Fluorophenol	0.6106	(Jouyban, 2009)
4-Isopropylphenol	0.3842	(Jouyban, 2009)
4-Methoxyphenol	0.5672	(Jouyban, 2009)
4-Nitroaniline	0.0674	(Jouyban, 2009)
4-Nitrophenol	0.3795	(Jouyban, 2009)
4-Nitrotoluene	0.4677	(Jouyban, 2009)
4-Phenylphenol	0.0507	(Jouyban, 2009)
4-tert-Butylphenol	0.2931	(Jouyban, 2009)
Benzamide	0.0266	(Jouyban, 2009)
Carbamazepine	0.0086	(W. Liu et al., 2008)
Cortexolone	0.0009	(Jouyban, 2009)
Cortisone	0.0011	(Jouyban, 2009)
Estradiol	0.0008	(Jouyban, 2009)
Estriol	0.8264	(Jouyban, 2009)
Estrone	0.0007	(Jouyban, 2009)
Flubiprofen	0.0308	(Jouyban, 2009)
Hydrocortisone	0.0005	(H. S. M. Ali et al., 2010)
Lidocaine	0.8252	(Jouyban, 2009)
Methyl 4-aminobenzoate	0.0978	(Jouyban, 2009)
n-Butyl 4-aminobenzoate	0.2618	(Jouyban, 2009)
Pentachlorophenol	0.0236	(Jing, Wang, & Wang, 2010)
Phenol	0.7777	(Jouyban, 2009)

Phenyl benzoate	0.1637	(Jouyban, 2009)
Prednisone	0.0005	(Jouyban, 2009)
Propyl paraben	0.0674	(Jouyban, 2009)

Table 10.5 Solubility data collated from literature of various drug compounds in Acetone taken at lab temperature

Compound	Mol. F	Reference
1,5-dinitronaphthalene	0.0223	(G. Zhou et al., 2015)
2-Amino-4-chlorobenzoic acid	0.0307	(Jouyban, 2009)
2-Chlorophenothiazine	0.0378	(J. Wang et al., 2017c)v
2-methyl-4-nitroaniline	0.1266	(X. Li et al., 2016)
3-methyl-4-nitrobenzoic acid	0.0260	(Wu et al., 2016)
3-nitro-o-toluic acid	0.0684	(Jouyban, 2009)
4-Amino-3,6-Dichloropyridazine	0.0258	(Jouyban, 2009)
4-nitrobenzaldehyde	0.1344	(Jouyban, 2009)
5-Ethyl-5-phenylbarbituric acid (Phenobarbital)	0.0138	(Jouyban, 2009)
Acetylsalicylic acid	0.0828	(Maia & Giulietti, 2008)
Artemisnin	0.0085	(Jouyban, 2009)
Benzoic acid	0.1857	(Thati et al., 2010)
Betulin	0.0016	(Jouyban, 2009)
Caffeic acid	0.0380	(Ji, Meng, Ding, et al., 2016)
Caffeine (form 1)	0.0045	(A. Shalmashi & Golmohammad, 2010)
Carbamazepine	0.0037	(W. Liu et al., 2008)
Daidzein	0.0009	(Jouyban, 2009)
Dehydroepiandrosterone acetate	0.0282	(Jouyban, 2009)
Diclofenac	0.0302	(Jouyban, 2009)
eszopiclone	0.0019	(Jouyban, 2009)

Haloperidol	0.0031	(Jouyban, 2009)
Ibuprofen	0.3508	(Jouyban, 2009)
isatin	0.0140	(J.-Q. Liu et al., 2014)
l-(+)-Ascorbic acid	0.0003	(Anvar Shalmashi & Eliassi, 2008)
Lactose	0.0001	(Jouyban, 2009)
Lamivudine (form 2)	0.0003	(Jouyban, 2009)
Loratadine	0.0229	(Jouyban, 2009)
Lovastatin	0.0123	(H. Sun et al., 2005)
Luteolin	0.0016	(Jouyban, 2009)
Naproxen	0.0692	(Jouyban, 2009)
Niflumic acid	0.0010	(Jouyban, 2009)
Norfloxacin	0.0001	(C.-L. Zhang et al., 2010)
N-phenylanthranilic acid	0.0335	(Yao et al., 2016)
Oleanolic acid	0.0010	(Jouyban, 2009)
Oxcarbazepine	0.0013	(Nam et al., 2017)
paclobutrazol	0.0239	(Jouyban, 2009)
Paracetamol	0.0368	(Jouyban, 2009)
p-Hydroxybenzoic acid	0.1194	(Jouyban, 2009)
p-Hydroxyphenylacetic acid	0.2213	(Jouyban, 2009)
Piroxicam	0.0028	(Jouyban, 2009)
risperidone form I	0.0018	(Mealey et al., 2014)
Salicylamide	0.1294	(Nordström & Rasmuson, 2006)
Salicylic acid	0.1792	(Matsuda et al., 2009)
sorbic acid	0.0701	(Fang et al., 2015)
Sulfadiazine	0.0010	(C.-L. Zhang et al., 2007)
Sulfamethoxypyridazine	0.0092	(C.-L. Zhang et al., 2007)
tebuconazole	0.0668	(Y. Li et al., 2017)

Temazepam	0.0207	(W. Li et al., 2016)
Tetraethyl ranelate	0.0582	(Jouyban, 2009)
Triclosan	0.5780	(DELGADO et al., 2012)
tridecanedioic acid	0.0041	(W. Tang et al., 2015)
Trimethoprim	0.0009	(Q.-S. Li et al., 2008)
Vanillic Acid	0.0661	(Y. Zhang et al., 2016)
Dodecanedioic acid	0.0050	(H. Zhang et al., 2014)
3,5-dimethylpyrazole	0.1164	(Yao et al., 2017)
5-phenyltetrazole	0.0217	(G. Chen, J. Chen, C. Cheng, Y. Cong, P. Jian, et al., 2017)
4-hydroxybenzaldehyde	0.1591	(J. Wang et al., 2017a)
2-Cyanoguanidine	0.0110	(Ren et al., 2014)
l-malic acid	0.1703	(Meng et al., 2013)
thiomalic acid	0.1681	(Meng et al., 2013)
itaconic acid	0.0514	(W. Yang et al., 2012)
Lansoprazole	0.0039	(Hong et al., 2012)
pyrazinamide	0.0049	(Jouyban, 2009)
Benzoin	0.0159	(Y. Yang et al., 2017)
4-Nitrophthalimide	0.0271	(Jouyban, 2009)
Theobromine	0.0001	(Zhong et al., 2017)
Theophylline	0.0009	(Zhong et al., 2017)
Agomelatin (Form II)	0.0429	(Y. Liu et al., 2015)
S-Hesperetin	0.0237	(Jouyban, 2009)
Phenylacetic acid	0.4232	(Gracin & Rasmuson, 2002)
p-Aminophenylacetic acid	0.0036	(Gracin & Rasmuson, 2002)
Triclocarban	0.0077	(DELGADO et al., 2012)
hyodeoxycholic acid	0.0022	(Y. Yang et al., 2015)
4-Aminobenzoic acid	0.0527	(Jouyban, 2009)

1,3-diphenylguanidine	0.0928	(R. Xu, Wang, Du, et al., 2016)
2,4-dinitroaniline	0.0497	(R. Xu, Xu, et al., 2016)
Clopidogrel	0.0007	(Song et al., 2010)
Dipyrone	0.0001	(Ding et al., 2017)
p-Coumaric Acid	0.0428	(Ji, Meng, Li, et al., 2016)
phthalimide	0.0128	(R. Xu, Wang, Han, et al., 2016)
Sulfanilamide	0.0944	(Jouyban, 2009)
3-amino-1,2,4-triazole	0.0079	(X. Li, C. Du, et al., 2017)
2-amino-5-methylthiazole	0.0979	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Difloxacin	0.0067	(Baluja et al., 2009)
Dimethyl 1,4-Cyclohexanedione-2,5-dicarboxylate	0.0046	(Jouyban, 2009)
Fenofibrate (formI I)	0.1922	(Watterson, Hudson, Svärd, & Rasmuson, 2014)
Flunixin meglumine	0.0016	(Qin et al., 2015)
Formononetin	0.0006	(Dong et al., 2017)
Irbesartan (form A)	0.0007	(L. Wang et al., 2007)
L-Carnitine	0.0001	(D. Sun et al., 2014)
Succinic Anhydride	0.1426	(Jouyban, 2009)
Sulfanilic Acid	0.0003	(Y. Hu et al., 2014)
Sulfatiazole	0.0724	(Y. Hu et al., 2014)
Xylitol	0.0002	(S. Wang et al., 2007)
4',5,7-Triacetoxyflavanone	0.1100	(Shan et al., 2017)
Febuxostat	0.0074	(L. Zhang et al., 2012)
Flubiprofen	0.1240	(Jouyban, 2009)
p-Aminobenzoic acid	0.0527	(Jouyban, 2009)
Puerarin	0.0280	(L.-H. Wang & Cheng, 2005)

Rutin	0.0003	(Zi et al., 2007)
Stearic acid	0.0077	(W. Yang et al., 2013)
Sulfaguanidine	0.0054	(Jouyban, 2009)

Table 10.6 Solubility data collated from literature of various drug compounds in Ethylacetate taken at lab temperature

Compound	Mol. F	Reference
1,5-dinitronaphthalene	0.0088	(G. Zhou et al., 2015)
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0349	(Jouyban, 2009)
2-amino-4-chloro-6-methoxypyrimidine	0.0155	(Jouyban, 2009)
2-Amino-4-chlorobenzoic acid	0.0241	(Jouyban, 2009)
2-Chlorophenothiazine	0.0291	(J. Wang et al., 2017c)
2-Isopropylimidazole	0.3102	(J. Chen et al., 2017)
2-methyl-4-nitroaniline	0.1100	(X. Li et al., 2016)
2-methyl-6-nitroaniline	0.1675	(Jouyban, 2009)
2-nitro-p-phenylenediamine	0.0342	(J. Wang et al., 2017b)
3-methyl-4-nitrobenzoic acid	0.0140	(Wu et al., 2016)
3-nitro-o-toluic acid	0.0406	(Jouyban, 2009)
4-Amino-3,6-Dichloropyridazine	0.0013	(Jouyban, 2009)
4-methyl-2-nitroaniline	0.1280	(X. Li et al., 2016)
4-nitrobenzaldehyde	0.0870	(Jouyban, 2009)
Acetylsalicylic acid	0.0448	(Maia & Giulietti, 2008)
Anhydrous citric acid	0.0062	(Jouyban, 2009)
Apigenin	0.0002	(Jouyban, 2009)
apremilast	0.0001	(Jouyban, 2009)
Artemisnin	0.0142	(Jouyban, 2009)
Atractylenolide III	0.0242	(Jouyban, 2009)
Atrazine	0.0125	(Jia et al., 2013)
Benzoic acid	0.1649	(Thati et al., 2010)

Betulin	0.0003	(Jouyban, 2009)
Caffeic acid	0.0015	(Ji, Meng, Ding, et al., 2016)
Caffeine (form 1)	0.0040	(Cruz-Monteagudo et al., 2017)
Carbamazepine	0.5110	(W. Liu et al., 2008)
Daidzein	0.0001	(Jouyban, 2009)
Deferiprone	0.0001	(Fathi-Azarbayjani et al., 2016)
Dehydroepiandrosterone acetate	0.0449	(Jouyban, 2009)
Diclofenac	0.0229	(Jouyban, 2009)
ferulic acid Form I	0.0207	(Shakeel et al., 2017)
Flurbiprofen	0.1110	(Jouyban, 2009)
Gallic acid	0.0051	(Vilas Boas, 2017)
Haloperidol	0.0101	(Jouyban, 2009)
Ibuprofen	0.3348	(Jouyban, 2009)
isatin	0.0057	(J.-Q. Liu et al., 2014)
Isonicotinamide (form II)	0.0082	(B. Li et al., 2016)
Ketoprofen	0.1530	(Jouyban, 2009)
l-(+)-Ascorbic acid	0.0001	(Anvar Shalmashi & Eliassi, 2008)
Lactose	0.0001	(Jouyban, 2009)
Lamivudine (form 2)	0.0001	(Jouyban, 2009)
Loratadine	0.0293	(Jouyban, 2009)
Lovastatin	0.0057	(H. Sun et al., 2005)
Luteolin	0.0026	(Jouyban, 2009)
Mannitol	0.0001	(Jouyban, 2009)
Mefenamic acid	0.0039	(Abdul Mudalip et al., 2013)
Meloxicam	0.0001	(Jouyban, 2009)

Methyl D-(-)-4-hydroxy-phenylglycinate	0.0008	(Jouyban, 2009)
Naproxen	0.0355	(Jouyban, 2009)
Niflumic acid	0.0001	(Domańska et al., 2011)
N-phenylanthranilic acid	0.0267	(Yao et al., 2016)
paclobutrazol	0.0228	(Jouyban, 2009)
Paracetamol	0.0001	(Jouyban, 2009)
Phenacetinum	0.0115	(Chang et al., 2007)
p-Hydroxybenzoic acid	0.0698	(Jouyban, 2009)
p-Hydroxyphenylacetic acid	0.0961	(Jouyban, 2009)
Pimozide	0.0021	(Jouyban, 2009)
Piroxicam	0.0024	(Jouyban, 2009)
risperidone form I	0.0022	(Mealey et al., 2014)
Salicylamide	0.0755	(Nordström & Rasmuson, 2006)
sorbic acid	0.0045	(Fang et al., 2015)
Spironolactone Form II	0.0281	(J. Zhang et al., 2014)
Sulfadiazine	0.0001	(C.-L. Zhang et al., 2007)
Sulfamethoxypyridazine	0.0021	(C.-L. Zhang et al., 2007)
tebuconazole	0.0564	(Y. Li et al., 2017)
Temazepam	0.0145	(W. Li et al., 2016)
Tetraethyl ranelate	0.0509	(Jouyban, 2009)
Triclosan	0.5650	(DELGADO et al., 2012)
tridecanedioic acid	0.0015	(W. Tang et al., 2015)
Trimethoprim	0.0003	(Q.-S. Li et al., 2008)
Dodecanedioic acid	0.0024	(H. Zhang et al., 2014)
3,5-dimethylpyrazole	0.1125	(Yao et al., 2017)
5-phenyltetrazole	0.0077	(G. Chen, J. Chen, C. Cheng, Y. Cong, P. Jian, et al., 2017)
4-hydroxybenzaldehyde	0.1549	(J. Wang et al., 2017a)

l-malic acid	0.0134	(Kai et al., 2013)
thiomalic acid	0.0595	(Meng et al., 2013)
itaconic acid	0.0110	(W. Yang et al., 2012)
Lornoxicam	0.0001	(Shakeel, Haq, Alanazi, et al., 2015)
Lansoprazole	0.0025	(Hong et al., 2012)
pyrazinamide	0.0035	(Jouyban, 2009)
Syringic Acid	0.0013	(Jouyban, 2009)
Tenoxicam	0.0007	(Jouyban, 2009)
Benzoin	0.0129	(Y. Yang et al., 2017)
4-Nitrophthalimide	0.0163	(Jouyban, 2009)
p-Toluenesulfonamide	0.0009	(Jouyban, 2009)
o-Toluenesulfonamide	0.0006	(Jouyban, 2009)
Theobromine	0.0001	(Zhong et al., 2017)
Theophylline	0.0007	(Zhong et al., 2017)
Genistein	0.0024	(Jouyban, 2009)
Protocatechuic Acid	0.0432	(Vilas Boas, 2017)
Gentisic Acid	0.0186	(Vilas Boas, 2017)
S-Hesperetin	0.0069	(Jouyban, 2009)
Phenylacetic acid	0.3482	(Gracin & Rasmuson, 2002)
p-Aminophenylacetic acid	0.0104	(Gracin & Rasmuson, 2002)
Triclocarban	0.0033	(DELGADO et al., 2012)
4-(4-aminophenyl)-3-morpholinone	0.0020	(W. Yang et al., 2016)
hyodeoxycholic acid	0.0016	(Y. Yang et al., 2015)
4-Aminobenzoic acid	0.0576	(Jouyban, 2009)
Vanillin	0.1230	(Jouyban, 2009)
1,3-diphenylguanidine	0.0566	(R. Xu, Wang, Du, et al., 2016)

2,4-dinitroaniline	0.0221	(R. Xu, Xu, et al., 2016)
Dipyrone	0.0001	(Ding et al., 2017)
Metaxalone (Form B)	0.0085	(Hong et al., 2016)
p-Coumaric Acid	0.0128	(Ji, Meng, Li, et al., 2016)
phthalimide	0.0075	(R. Xu, Wang, Han, et al., 2016)
Pyridazine-3-amine	0.0032	(Cao et al., 2012)
3-amino-1,2,4-triazole	0.0029	(S. Liang et al., 2016)
2-amino-5-methylthiazole	0.1008	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Dimethyl 1,4-Cyclohexanedione-2,5-dicarboxylate	0.0075	(Jouyban, 2009)
Succinic Anhydride	0.0771	(Jouyban, 2009)
4',5,7-Triacetoxyflavanone	0.0450	(Shan et al., 2017)
4-Methylsulfonylacetophenone	0.0361	(Hao et al., 2017)
Cefoxitin acid	0.0003	(Yuan et al., 2016)
Febuxostat	0.0046	(L. Zhang et al., 2012)
levofloxacin	0.0022	(J. Zhang et al., 2012)
D-Pantolactone	0.3854	(C. Huang et al., 2015)
Hydrocortisone	0.0008	(H. S. M. Ali et al., 2010)
p-Aminobenzoic acid	0.0576	(Jouyban, 2009)
Rutin	0.0011	(Zi et al., 2007)
1H-1,2,4-Triazole	0.0410	(S. Liang et al., 2016)
2-Hydroxybenzoic acid	0.1425	(Gracin & Rasmuson, 2002)
β-Sitosteryl maleate	0.0059	(D. Wei, Wang, Liu, & Wang, 2010)
Methyl *p*-hydroxybenzoate	0.1270	(Jouyban, 2009)
Ricobendazole	0.0001	(Jouyban, 2009)
Saccharose	0.0001	(Jouyban, 2009)

Table 10.7 Solubility data collated from literature of various drug compounds in 1,4-dioxane taken at lab temperature

Compound	Mol. F	Reference
2,4-dihydro-5-methyl-2-(4-methylphenyl)-3H-pyrazol-3-one	0.0549	(Jouyban, 2009)
2-amino-4-chloro-6-methoxypyrimidine	0.0367	(Jouyban, 2009)
2-Amino-4-chlorobenzoic acid	0.0271	(Jouyban, 2009)
2-nitro-p-phenylenediamine	0.0593	(J. Wang et al., 2017b)
3-methyl-4-nitrobenzoic acid	0.0545	(Wu et al., 2016)
3-nitro-o-toluic acid	0.0779	(Jouyban, 2009)
4-Amino-3,6-Dichloropyridazine	0.0043	(Jouyban, 2009)
4-methyl-2-nitroaniline	0.0794	(R. Xu, Xu, et al., 2016)
Acetylsalicylic acid	0.0516	(Perlovich & Bauer-Brandl, 2003)
Anhydrous citric acid	0.1413	(Jouyban, 2009)
Benzoic acid	0.2853	(Thati et al., 2010)
Deferiprone	0.0002	(Fathi-Azarbayjani et al., 2016)
Diclofenac	0.1055	(Jouyban, 2009)
Flurbiprofen	0.1750	(Jouyban, 2009)
Haloperidol	0.0056	(Jouyban, 2009)
Ibuprofen	0.0372	(Jouyban, 2009)
isatin	0.0188	(J.-Q. Liu et al., 2014)
Ketoprofen	0.1530	(Jouyban, 2009)
Lactose	0.0001	(Jouyban, 2009)
Mannitol	0.0001	(Jouyban, 2009)
Meloxicam	0.0018	(Jouyban, 2009)
Naproxen	0.1040	(Jouyban, 2009)
Niflumic acid	0.0483	(Jouyban, 2009)
paclobutrazol	0.0324	(Jouyban, 2009)
Paracetamol	0.0314	(Jouyban, 2009)
Phenothiazine	0.1026	(J. Wang et al., 2017c)
p-Hydroxybenzoic acid	0.0844	(Jouyban, 2009)
Pimozide	0.0113	(Jouyban, 2009)
Piroxicam	0.0049	(Jouyban, 2009)
Salicylamide	0.1373	(Nordström & Rasmuson, 2006)
Salicylic acid	0.2610	(Matsuda et al., 2009)
Sulfadiazine	0.0005	(Mauger et al., 1972)
Sulfamethoxypyridazine	0.0239	(Mauger et al., 1972)

5-amino-3-methyl-1-phenylpyrazole	0.2104	(G. Chen, J. Chen, P. Jian, et al., 2017)
5-phenyltetrazole	0.0125	(G. Chen, J. Chen, C. Cheng, Y. Cong, P. Jian, et al., 2017)
4-hydroxybenzaldehyde	0.1227	(J. Wang et al., 2017a)
Lornoxicam	0.0002	(Shakeel, Haq, Alanazi, et al., 2015)
4-Nitrophthalimide	0.0295	(Jouyban, 2009)
Nabumetone	0.0053	(Jouyban, 2009)
4-(4-aminophenyl)-3-morpholinone	0.0102	(W. Yang et al., 2016)
4-Aminobenzoic acid	0.0700	(Jouyban, 2009)
3-amino-1,2,4-triazole	0.0051	(S. Liang et al., 2016)
2-amino-5-methylthiazole	0.0806	(G. Chen, J. Chen, C. Cheng, Y. Cong, C. Du, et al., 2017)
Irbesartan (form A)	0.0008	(L. Wang et al., 2007)
p-Aminobenzoic acid	0.0632	(Jouyban, 2009)
2-Hydroxybenzoic acid	0.2945	(Gracin & Rasmuson, 2002)
Saccharose	0.0001	(Jouyban, 2009)

10.2 APPENDIX 2 A

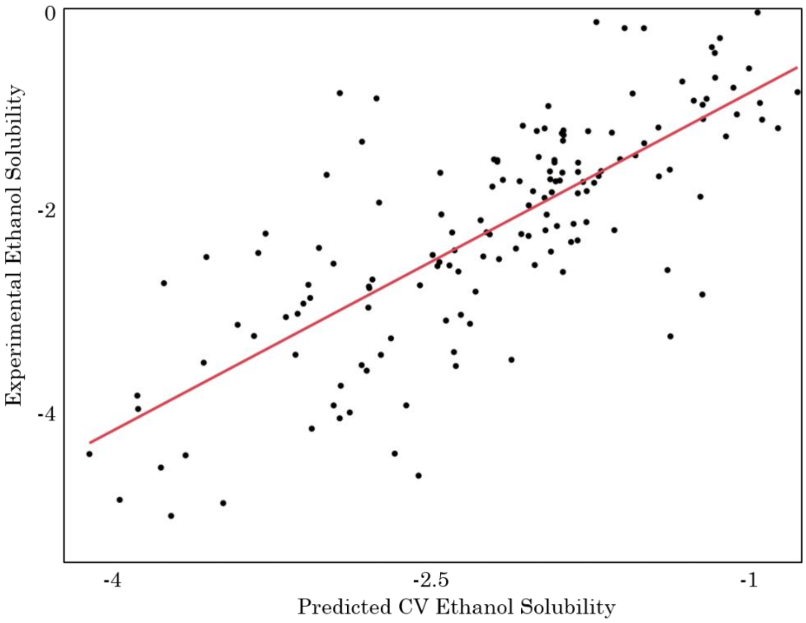


Figure 10.1 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in ethanol.

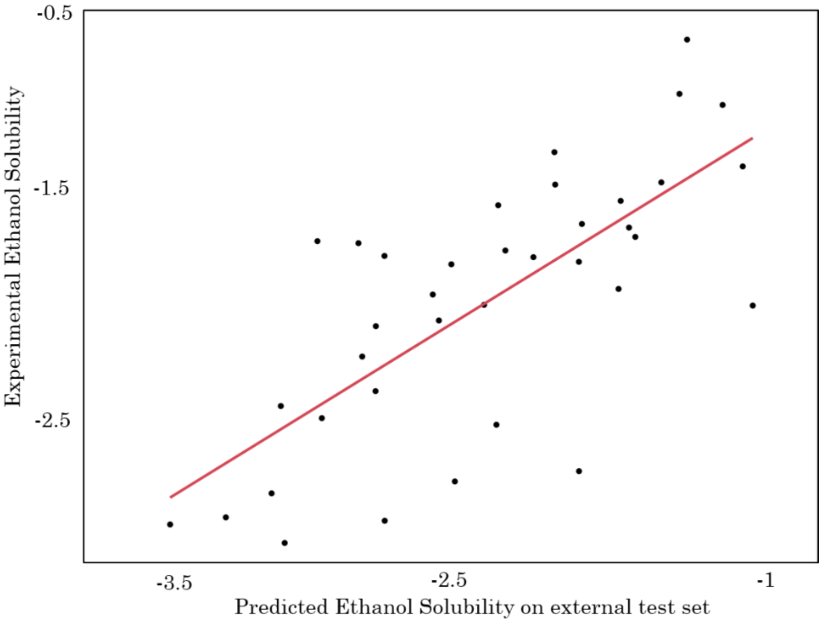


Figure 10.2 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in ethanol.

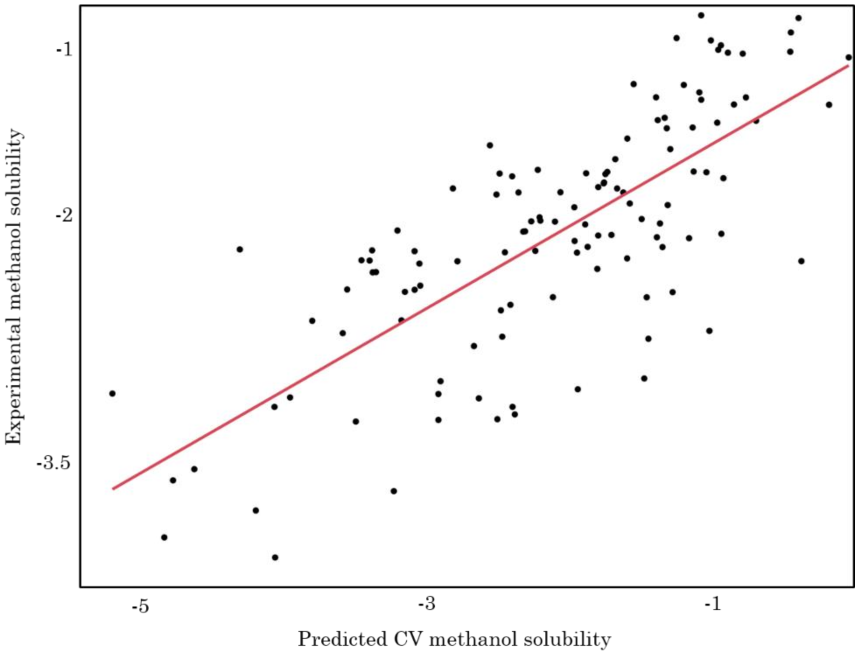


Figure 10.3 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in methanol.

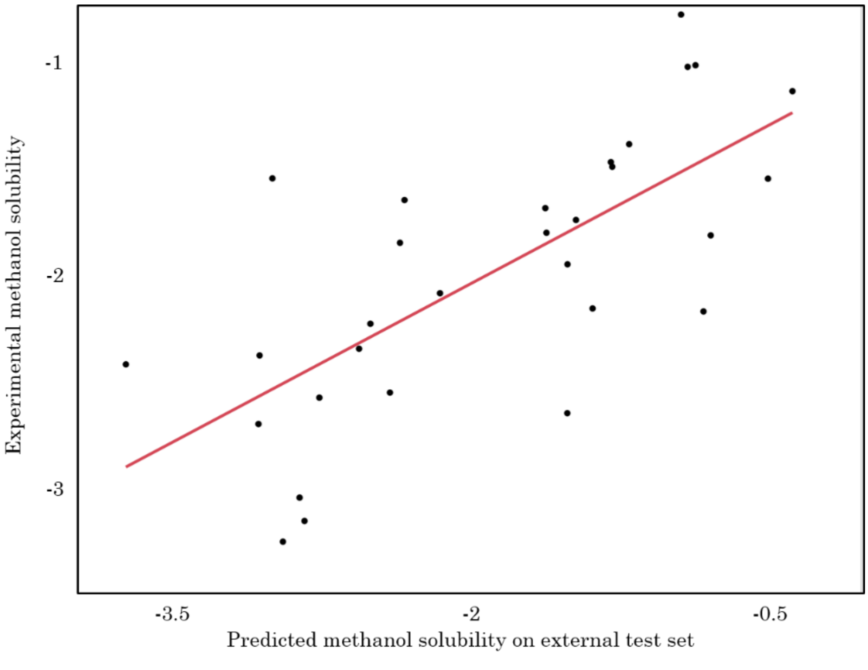


Figure 10.4 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in methanol.

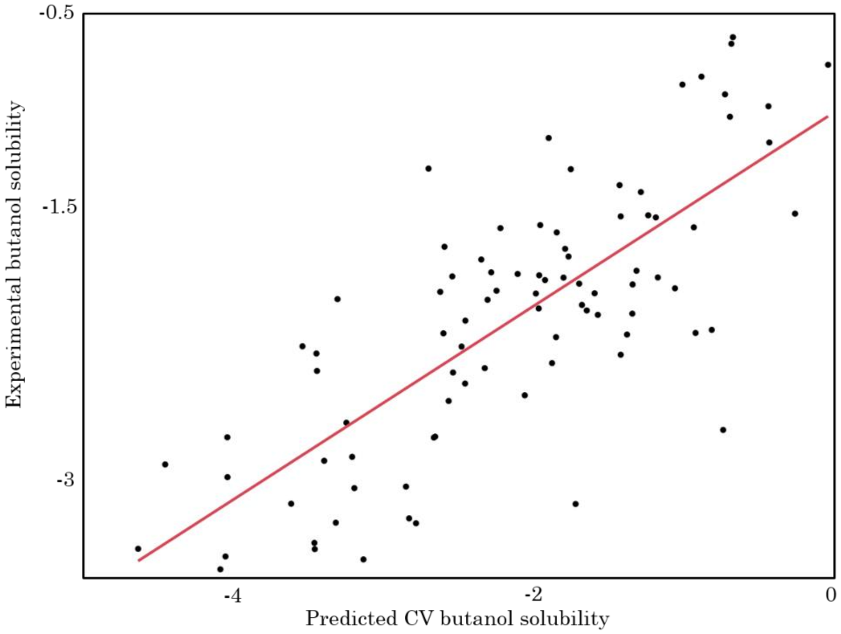


Figure 10.5 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in 1-butanol.

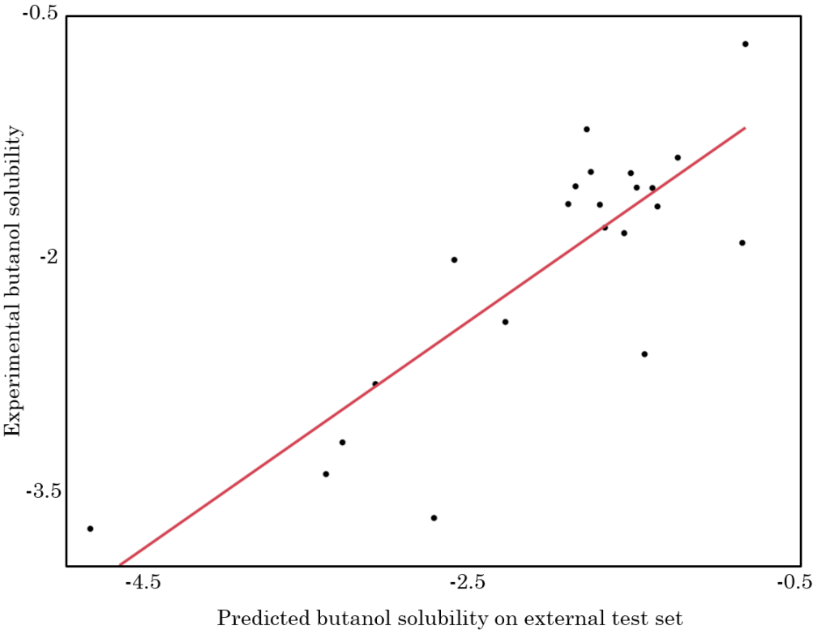


Figure 10.6 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in 1-butanol.

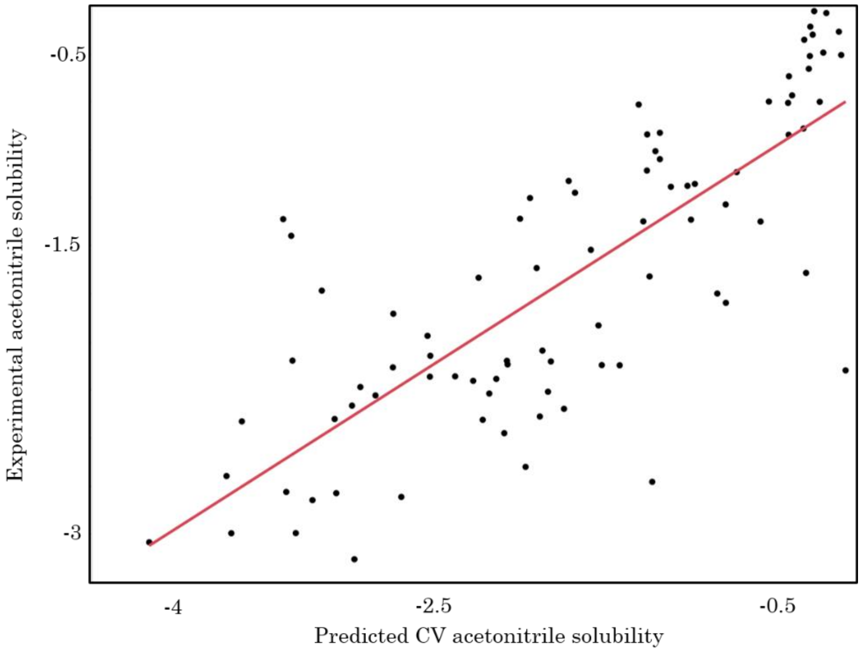


Figure 10.7 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in acetonitrile.

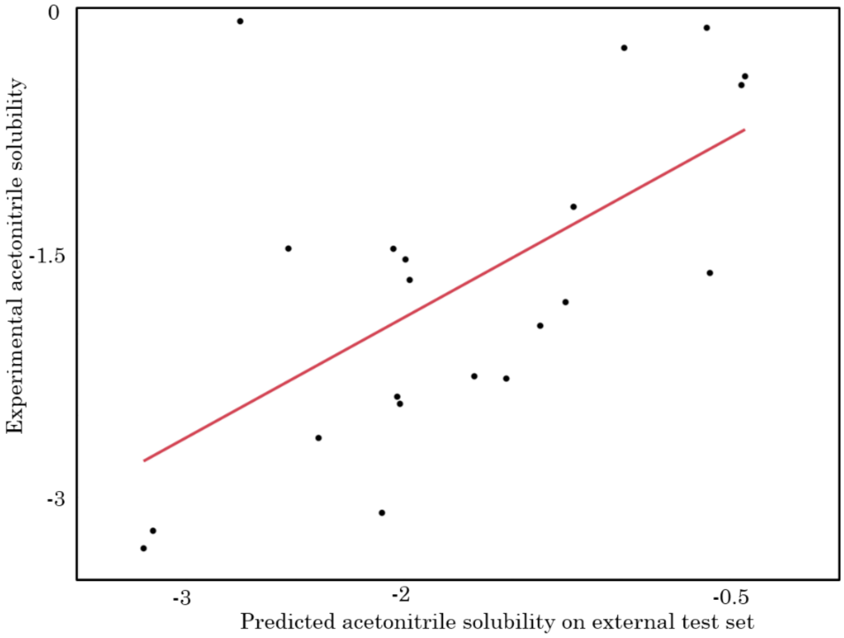


Figure 10.8 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in acetonitrile.

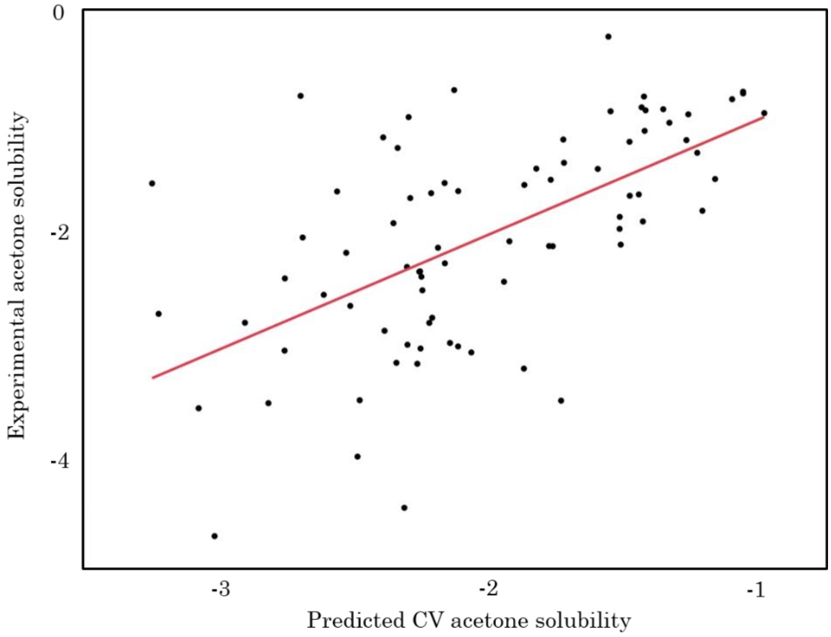


Figure 10.9 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in acetone.

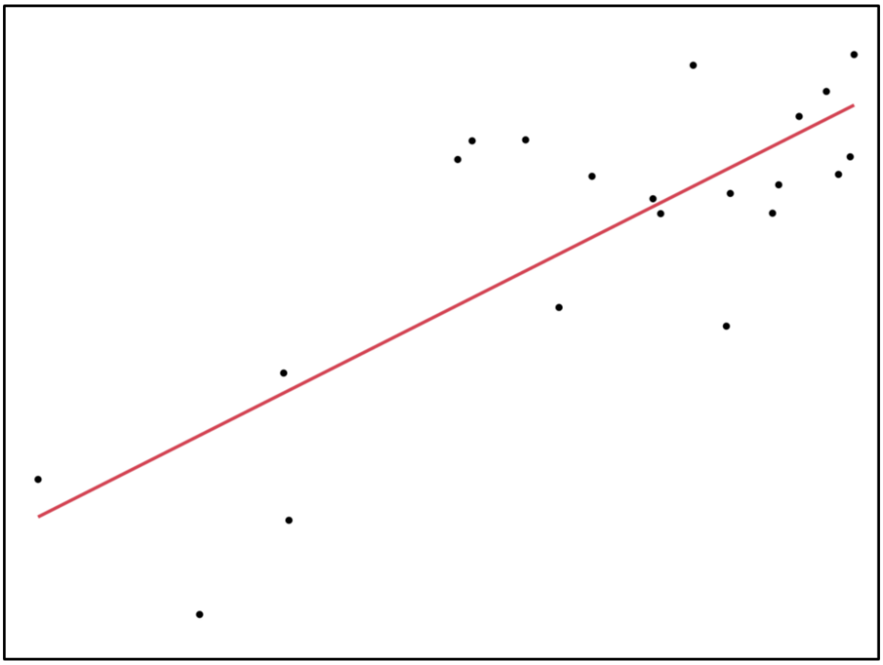


Figure 10.10 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in acetone.

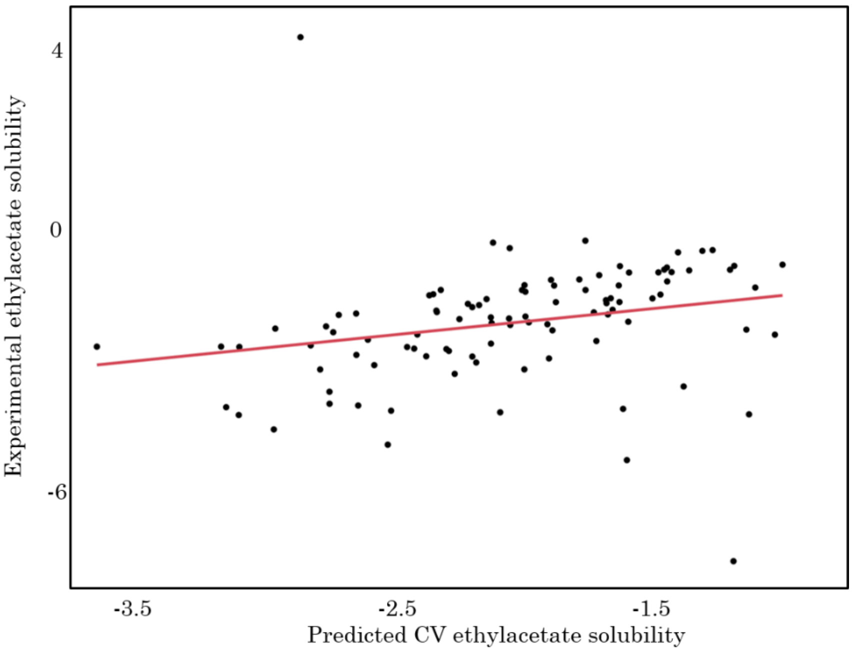


Figure 10.11 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in ethylacetate.

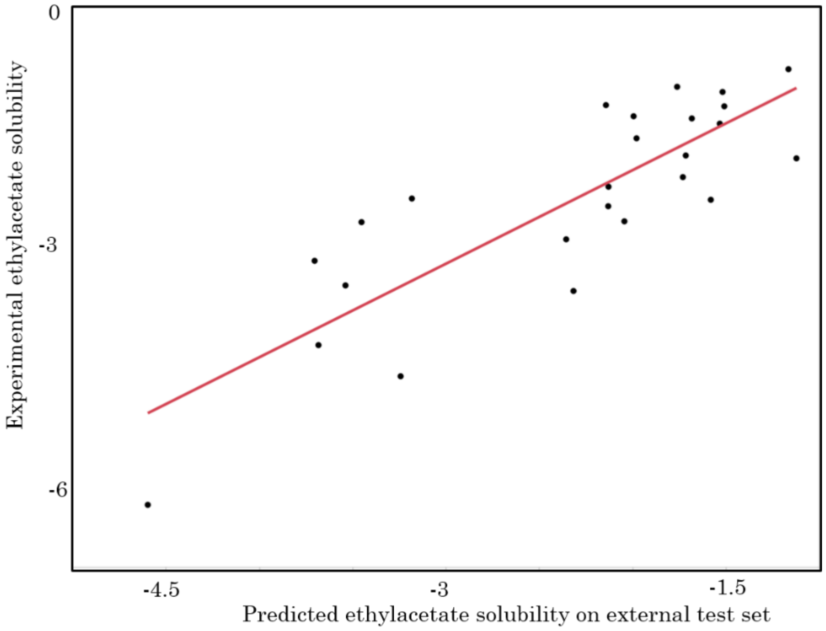


Figure 10.12 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in ethylacetate.

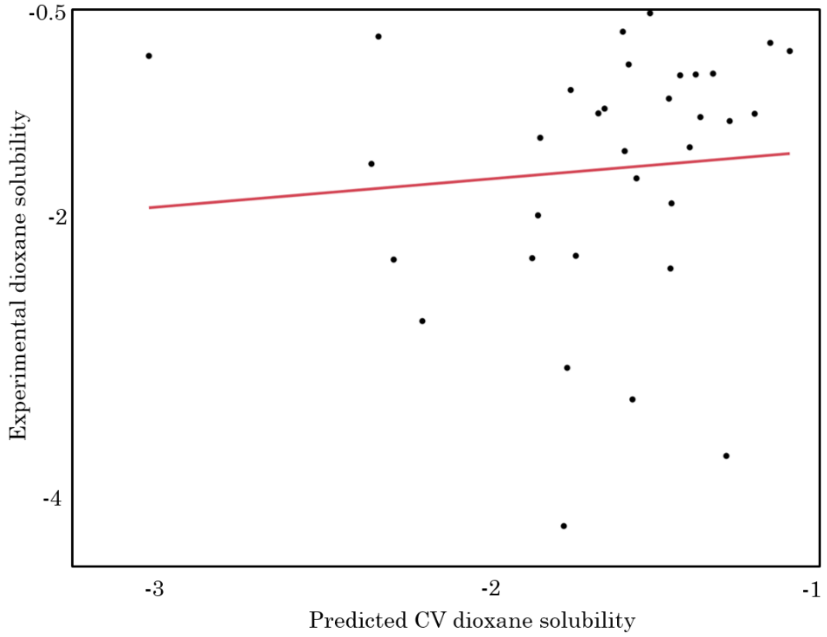


Figure 10.13 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the training set compounds in 1,4-dioxane.

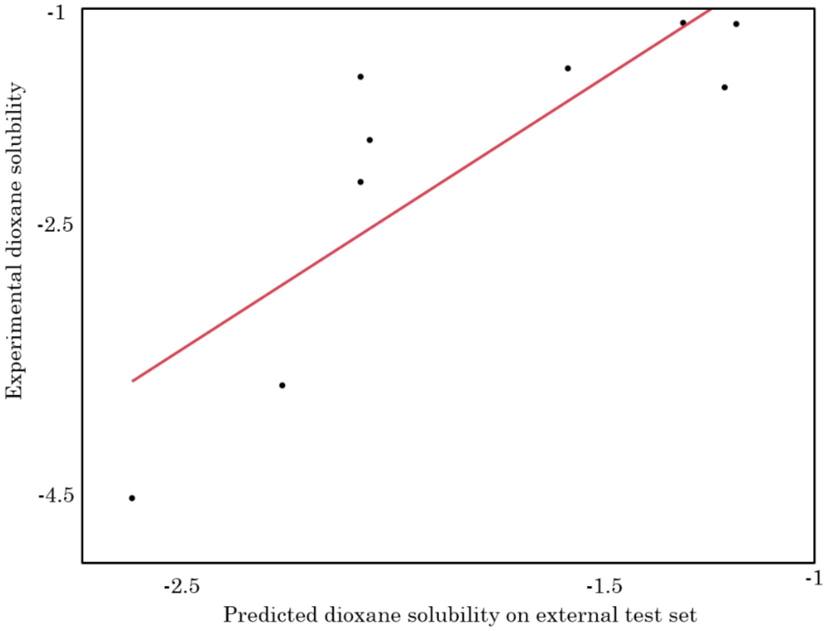


Figure 10.14 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in 1,4-dioxane.

10.3 APPENDIX 2 B

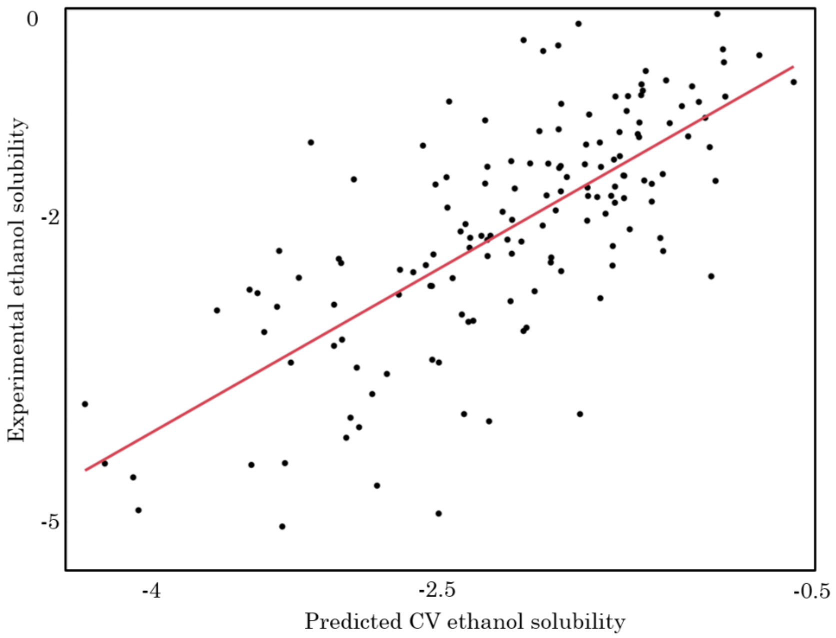


Figure 10.15 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in ethanol.

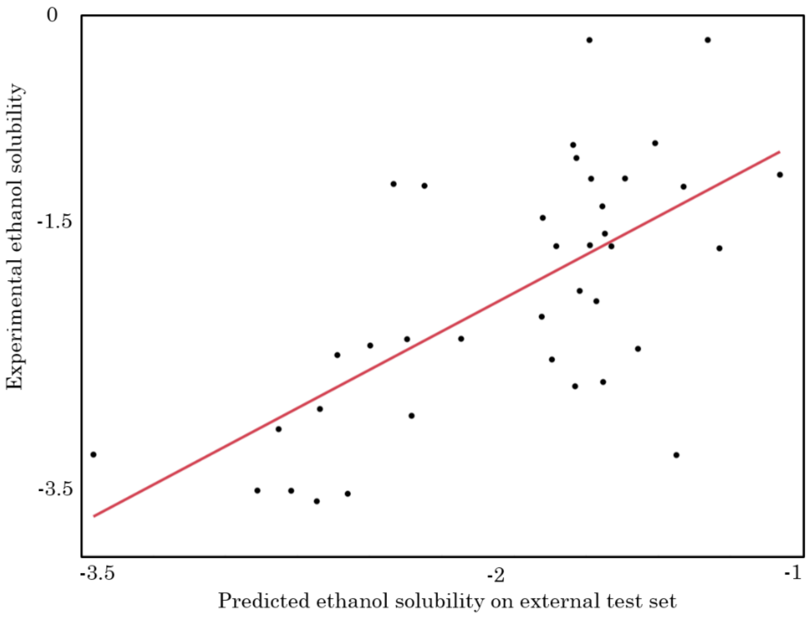


Figure 10.16 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in ethanol.

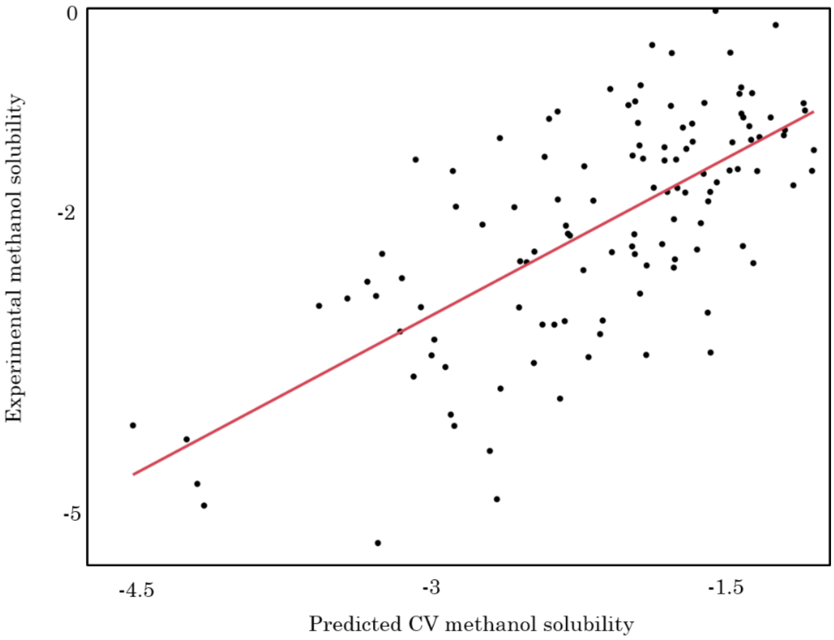


Figure 10.17 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in methanol.

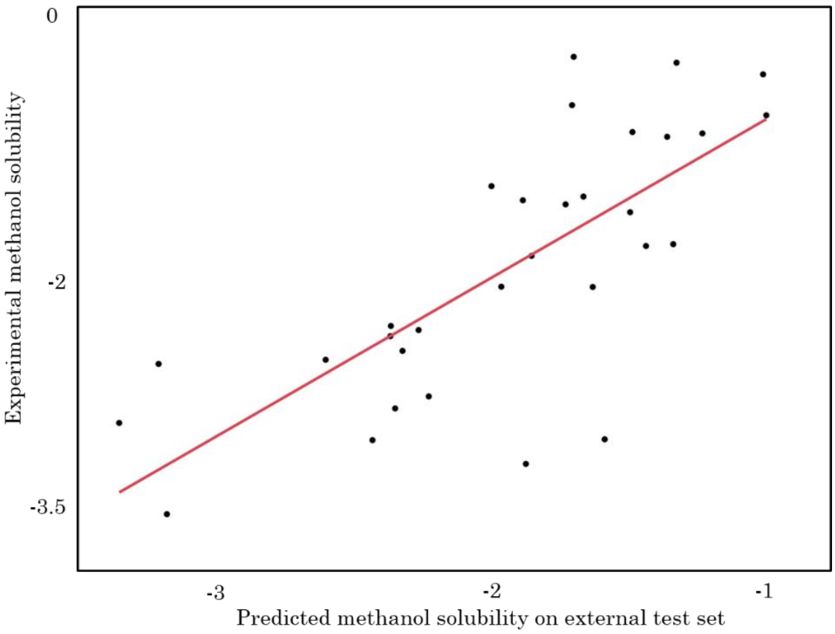


Figure 10.18 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in methanol.

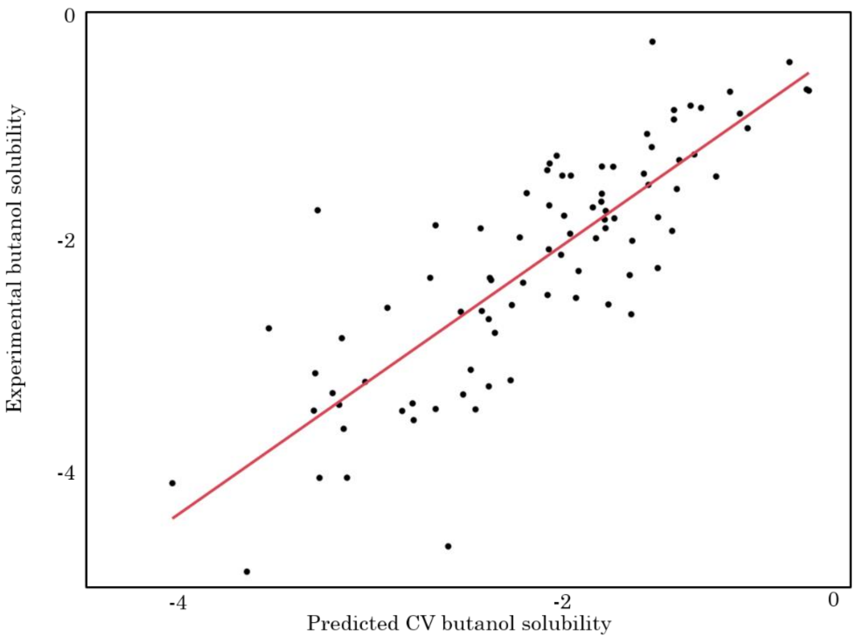


Figure 10.19 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in 1-butanol.

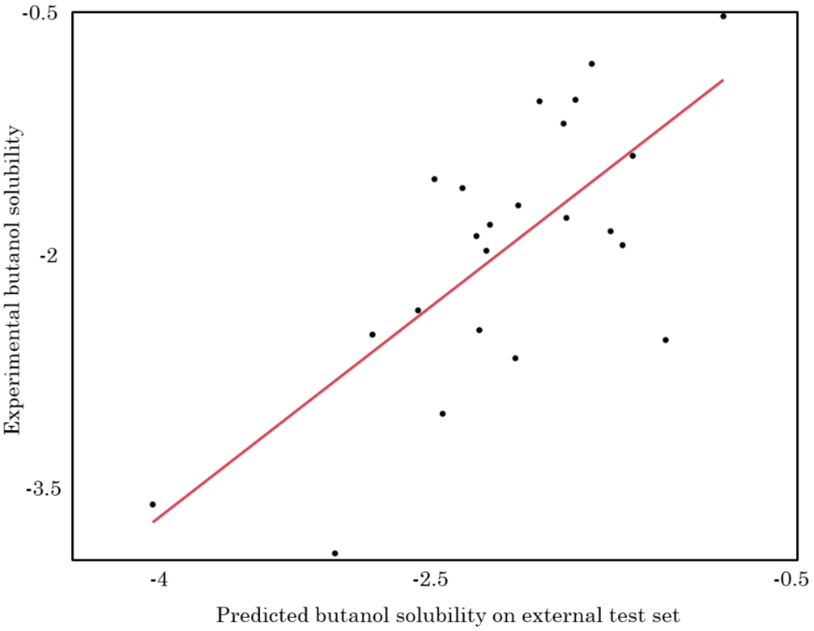


Figure 10.20 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using molecular descriptors on the external test compounds in 1-butanol.

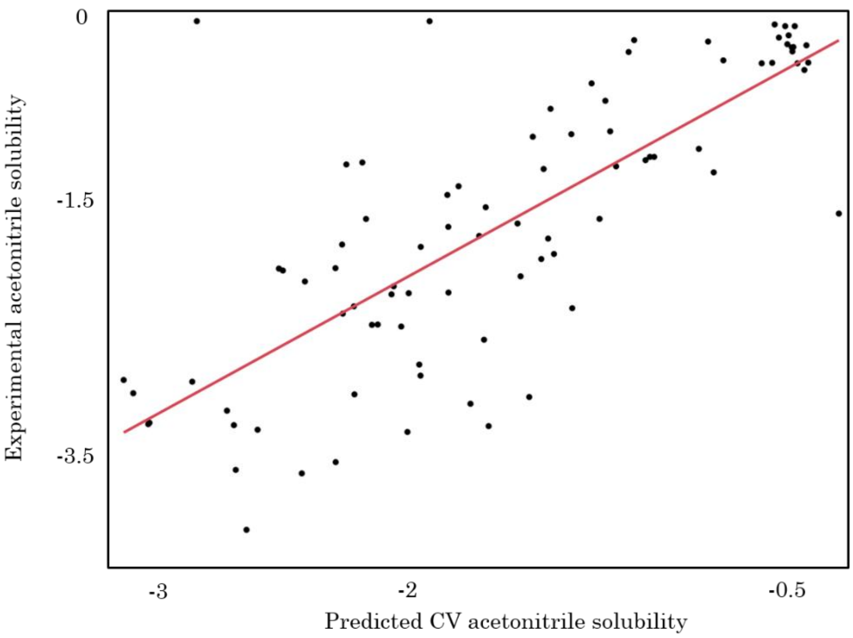


Figure 10.21 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in acetonitrile.

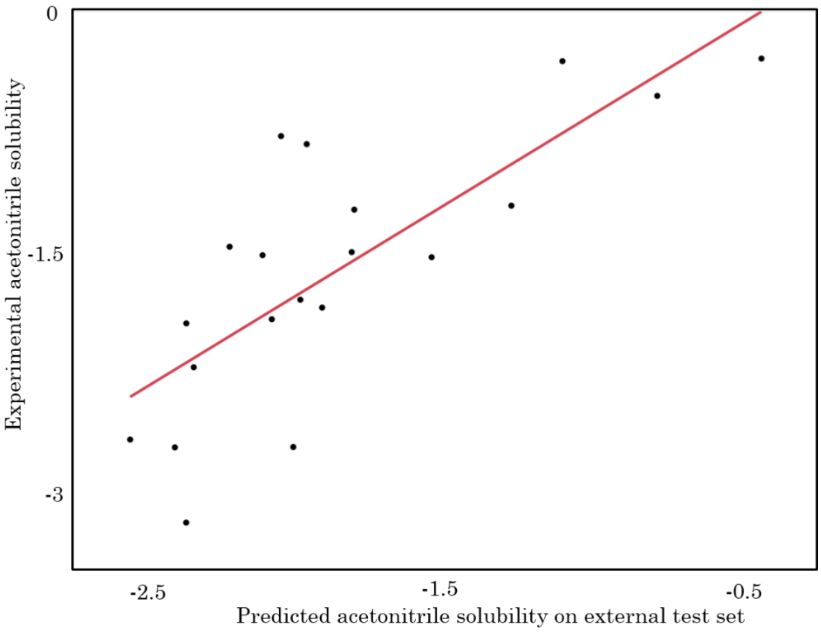


Figure 10.22 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in acetonitrile.

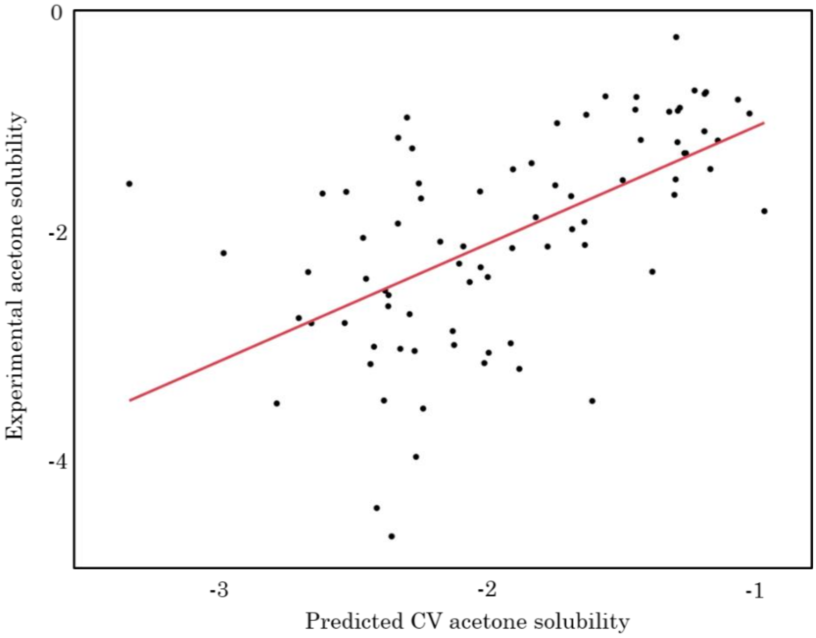


Figure 10.23 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in acetone.

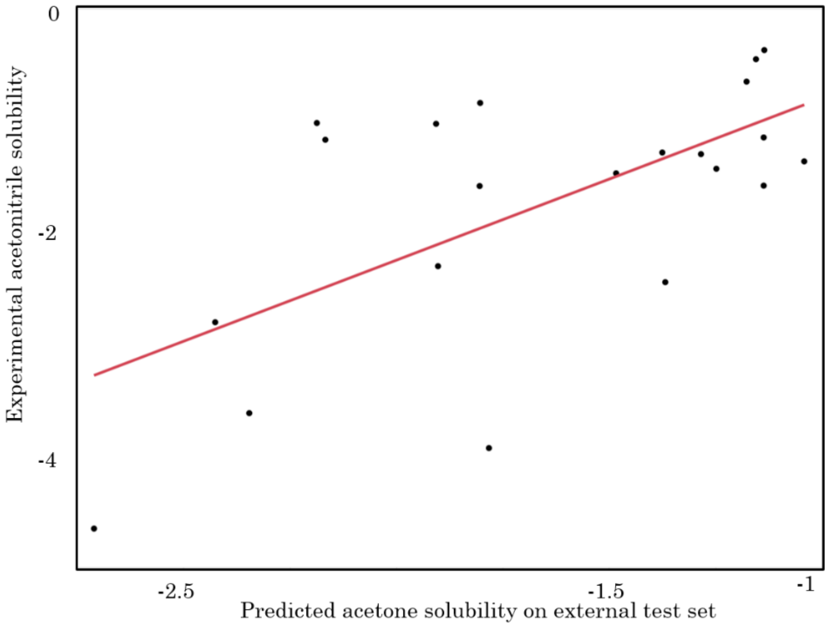


Figure 10.24 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in acetone.

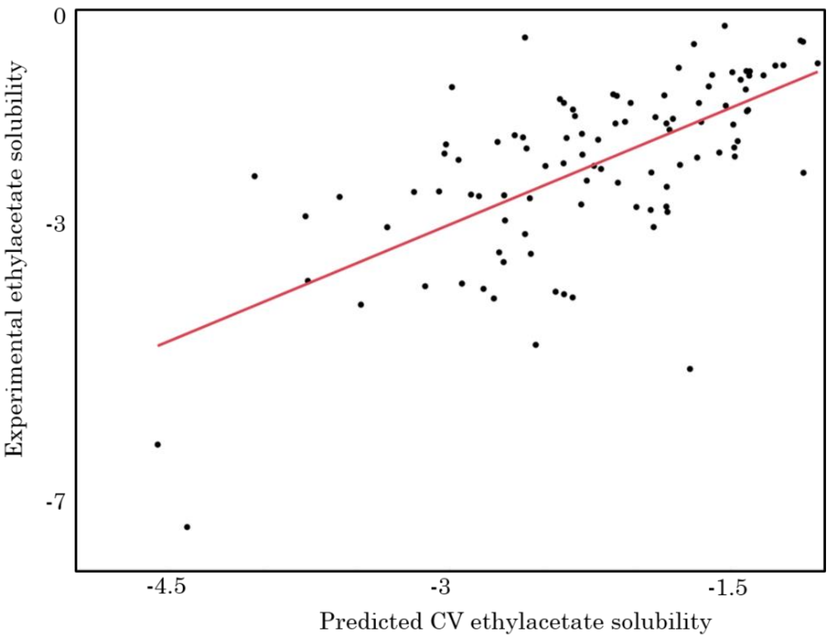


Figure 10.25 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in ethylacetate.

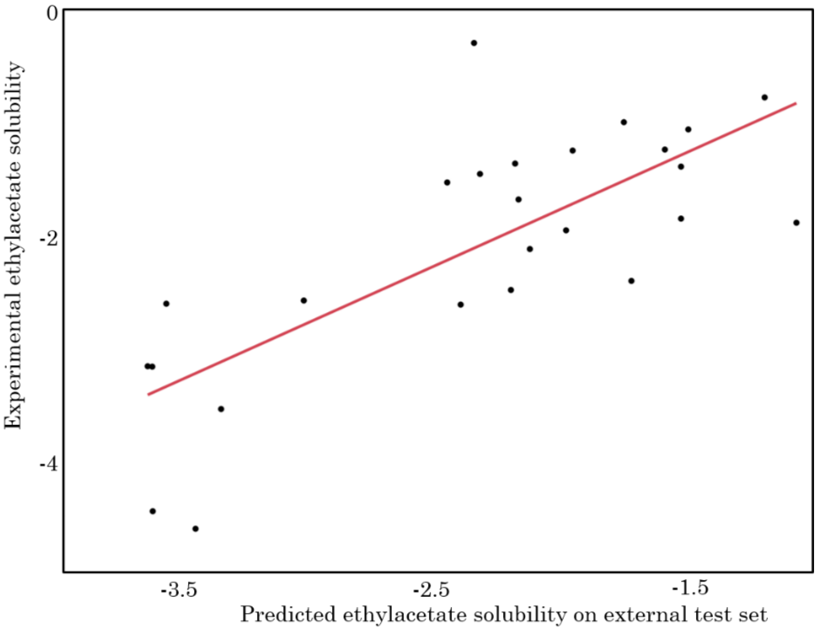


Figure 10.26 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in ethylacetate.

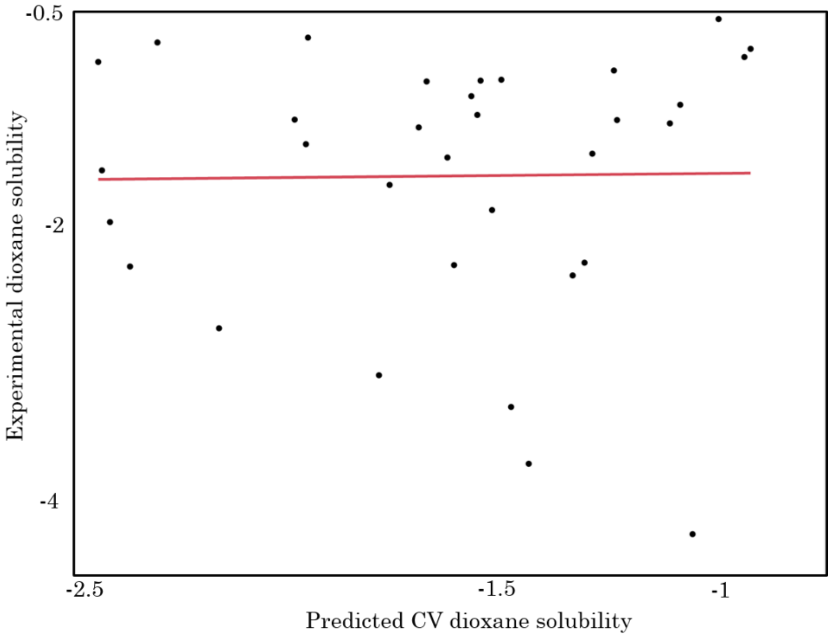


Figure 10.27 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the training set compounds in 1,4-dioxane.

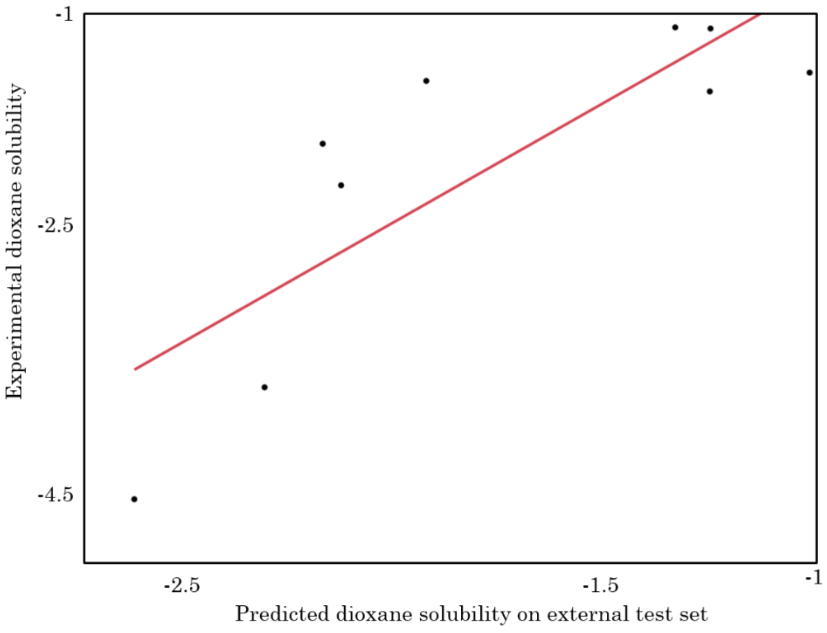


Figure 10.28 Regression plot of the experimental vs predicted solubility data (LogS) obtained from the RF model trained using MACCS fingerprint on the external test compounds in 1,4-dioxane.