Computational Studies of Integrin Inhibitors



Saleh S. Alarfaji

A thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

SEPTEMBER-2018

Abstract

Compounds containing amides play a key role in the pharmaceutical industry, with one area of interest, amongst many, being the development of new antifibrotic therapies. These compounds can bind to different integrins, which are essential activity regulators for transforming growth factor beta 1 (TGF-β_1), and biological activity can assessed using a cell adhesion assay. Two- and three-dimensional quantitative structure-activity relationships (2D/3D-QSARs) were used to model the correlation between the physicochemical properties of some amides and their biological activity to predict the activities of new molecules. An autocorrelation based method, topological maximum cross correlation (TMACC), was employed to build our 2D-QSAR models. We have generated models across four sets of data, ranging in size from 25 to 46 molecules on a number of integrins subtypes. The results were cross-validated using a leave-one-out (LOO) approach. Using partial-least-squares regression, a TMACC model with good predictive ability was generated based on training set of 25 compounds and showed satisfactory statistical results ($q^2 = 0.57$, $r^2 = 0.82$). Other TMACC models for 40 and 46 molecules were generated and showed similar but slightly higher predictivity: ($q^2 = 0.71$, $r^2 = 0.93$), and ($q^2 = 0.49$, $r^2 = 0.70$), respectively. 3D-QSAR models were also established for the same data sets using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). The CoMFA and CoMSIA models were not sensitive to changes in the orientation of the amide structure. The TMACC QSARs showed better

predictive ability than the 3D-QSARs. We also have employed a quantum chemical investigation into different datasets of benzamides in order to understand the relationship between the physicochemical properties of all datasets with their activity. Potential energy surface scans were calculated for each molecule using density functional theory (DFT). Several quantum descriptors were calculated and used to build our QSAR models. QSAR results were validated using the LOO and Y-randomisation methods. Statics were also employed to determine all possible correlation between the quantum descriptors and biological activity.

In a distinct study, the electronic transitions, in the aromatic side chain of the amino acid phenylalanine toluene was investigated. The toluene molecule was used as a model system for biological chromophores. The photophysics of toluene was studied to understand the photophysics of large biomolecules such as proteins better. The ground (S_0) and excited (S_1) states in the gas phase have been calculated using a multi-configurational CASSCF method. We also employ the state-averaged CASSCF approach to estimate the geometries and harmonic vibrational frequencies for the both S_0 and S_1 states too. The $\pi\pi^*$ transition is calculated to be at 4.69 eV, and the corresponding experimental energy is 4.65 eV. The experimental $\pi\pi^*$ transition dipole moment is 0.37 D. If four unoccupied a' orbitals are included in the active space, the calculated $\pi\pi^*$ transition dipole moment is 0.30 D. If the number of unoccupied a' orbitals is decreased to six orbitals, then the calculated $\pi\pi^*$ transition moment drops to 0.28 D.

In the future, further improvements of TMACC and 3D-QSAR are needed. For a given data set, the TMACC descriptor construction should be enlarged to include other essential and new atomic properties. Moreover, the output of TMACC and 3D-QSAR would be more beneficial if we could automate feature detection for any large data.

Acknowledgements

All praise is due to the Almighty God for the countless blessings He showered upon me throughout my journey in life, and for all the physical and mental strength. He bestowed on me during the activities of preparing this thesis and conducting the associated research.

First and foremost, my deepest gratitude and thanks go to my principal supervisor Prof Jonathan D. Hirst, who guided me through every single step along the way for the last four years, and helped me become the person who I am today. Without his great support, advice and above all the unparalleled patience of Prof Hirst, this thesis would not have seen the light.

I am grateful to GlaxoSmithKline for many useful discussions, particularly Simon Macdonald. Special thanks to Thomas Mcinally for useful meeting and discussions. Thanks to the members of the Hirst Group, especially Steven Oatley for helpful discussion. I am grateful to the School of Chemistry for the Computational Chemistry facilities and the University of Nottingham for providing access to the High Performance Computing facility. In addition, many thanks to King Khalid University in the Kingdom of Saudi Arabia for funding my project. I also would like to acknowledge the academic licenses made available from both OpenEye, which has enabled this work to proceed much further.

Finally, I am very grateful for the support from my family and friends to see this work through to the end. It is to them I dedicate this thesis.

Dedication

This work is dedicated to my parents, Saeed and Saleha, who have always loved me thoroughly and whose good examples have taught me to work hard for the things that I try to achieve.

This thesis is also dedicated to my family, Amal, Tameem, Rose, Abdulmalek, and each member of my lovely family for their endless love, support and encouragement who have always been a constant source of support and encouragement during the challenges of my whole PhD journey. Also to my brothers and sisters whom I am truly grateful for having in my life.

List of Figures

1.1 The steps that are essential for a drug to proceed through the research and development pipeline. 1

1.2 Molecular structures of pirfenidone (left) and Nintedanib (right) 3

1.3 An example shows the properties and their values for a drug-like molecule. 6

1.4 General schematic structure of $\alpha_v\beta$ integrin. The α- and β-subunits are indicated in blue and green, respectively. Cytoplasmic cells are circled in red. 13

1.5 Integrin families. This diagram outlines the 24 possible integrin dimer species, grouped by substrate type. 16

2.1 (a) Scatter diagram for molecular descriptor x and dependent variable y. (b) Straight-line relationship between molecular descriptor and dependent variable 23

2.2 A simple two-dimensional illustration of principal components analysis. . . 28

2.3 PLS model composes of two lines, one in the x-space (t_1, left panel) and one in the y-space (u_1, middle panel). Right panel shows the intercorrelation between t_1 and u_1 lines. 30

2.4 Example of a dendrogram or a hierarchical tree. This is based on the calculated ChelpG for 25 benzamide derivatives (The net charge on the ring). . 31

2.5 Schematic representation of leave-one-out cross-validation for 50 molecules; the first data point is removed and a model trained using the rest of the dataset. This model is used to predict the first data point. Then the second is removed and the process repeated until each data point was deleted once 33

2.6 Schematic representation of k-fold cross-validation; the dataset is split into 3 ubsets of equal size. Each of the subset is removed once, a model built with the reminder and the subset predicted with this model. 35
2.7 Neural network diagram 37
2.8 A two-dimensional plot for the support vector method. The separating hyperplane is shown as a line. The perpendicular length between the separating hyperplane and a hyperplane through the closest data points (the support vector) is so-called margin. 39
2.9 Examples of simple sigmoid functions 40
2.10 Example of a decision tree for a set of 25 molecules. In order to classify an unknown compound a path is followed through the tree according to the values of the relevant properties, until a terminal node is reached. 42
2.11 Hydrogen-suppressed graph of 2-methylhexane, with labelled vertices . . . 46
2.12 A template molecule in a 3D grid box. For presentation, only the grid lines at the surface of the box and only one molecule are shown instead of a superposition of all molecules. This grid includes $9 \times 9 \times 9 = 729$ grid points 48
2.13 CoMFA contour plots for 4-chloro-*N*-methyl-3-(trifluoromethyl)benzamide. For log$P$, yellow indicates regions where it is favourable to place a positive charge and white where it would be unfavourable. For log$S$, blue indicates regions where it would be favourable to place a positive charge group and red where it would be unfavourable. 49

3.1 Docked model of $\alpha_v\beta_1$ integrin, where α and β subunits are, respectively, in green and blue. 52
3.2 An example of (A) the whole structure (B) the corresponding fragment (*N*,2,6-trimethylbenzamide). 54
3.3 A Random Forest (RF) example shows steps of the ensemble method. Each empty box represents a leaf node that contains new molecules from the decision tree. P_n shows the real predicted value of each tree. 60

3.4 (A) Structural alignment of 27 benzamides according to OMEGA. (B) The template molecule (compound **2)** 61
3.5 The correlation between the experimental and predicted pIC_{50} for the Charge, log$S$, and MR model with 7 PLS components (G4 data set); a regression line has been drawn through the data (solid line). 64
3.6 Predictive performance (MAE) of SVM (RBFKernel), MLP, IBk, M5P, RF, and REPTree models assessed in LOO-CV on the training set of 90 models. 68
3.7 TMACC interpretation of log$S$ models with 6 PLS components for G1 and G2 groups. TMACC activity colour scheme: red for very negative activity; orange for negative activity; green for neutral activity; yellow for positive activity and blue for very positive activity. 69
3.8 TMACC interpretation of G4 group for two different models; Charge log$S$ MR with 7 components (a,c, e and g molecules), and log$S$ MR with 7 components (b, d, f and h molecules). 70
3.9 The variation of r^2 and q^2 with the number of components (latent variables) for the charge log$S$ MR model. 72
3.10 Visualisation of the TMACC/PLS charge log$S$ MR model of the G4 data set showing the effect of modifying a *para*-CF_3-substituted benzamides inhibitor molecule **15** (top, pIC_{50} = 7.5) to an *ortho*-CF_3-substituted molecule **14** (bottom, pIC_{50} = 8.1). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates that the CF_3-substitution becomes less potency-reducing in the *para*-position. 73
3.11 Visualisation of the TMACC/PLS model showing the effect of modifying an-ortho single Cl-substituted benzamide inhibitor molecule **10** (top, pIC_{50} = 8.2) to two *ortho*-Cl-substituted molecule molecule **2** (bottom, pIC_{50} = 8.8). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. 74

3.12 Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* single Cl-substituted benzamide inhibitor molecule **19** (top, pIC_{50} = 7.7) to two F-substituted molecule molecule **18** (bottom, $pIC_{50} = 7.2$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The yellow colour indicates that the F-substitution becomes less potency-increasing in the *meta*-position. 75

3.13 Visualisation of the TMACC/PLS model showing the nitrogen contribution to the activity for the highest and the less active inhibitors; The first inhibitor is molecule **28** (top, $pIC_{50} = 8.8$) and the other inhibitor is molecule **40** (bottom, $pIC_{50} = 6.1$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates that the N atoms becomes less potency-decreasing in the quinoline ring. 76

3.14 Visualisation of the TMACC/PLS model showing the sulfonamide contribution to the activity for molecule **26** (pIC_{50} = 7.1). Highlighted atoms indicated those that are most important for increasing (blue) the potency of inhibitors. 77

3.15 Experimental vs. predicted pIC_{50} for the charge, log$P$, and MR model . . 78

3.16 q^2 versus r^2 for the 96 y-randomisation models with q^2 values greater than zero (black dots). The real model is represented by the red triangle. 79

4.1 (a) Crystal structure of $\alpha_v\beta_6$ shows the binding site ligand (Ca^{2+} ion, in green). ASP 230.A, ASP 234.A, ASP 238.A, ASN 232.A, and ILE 236.A are shown to indicate the key interactions with Ca^{2+}. (b) Crystal structure of $\alpha_v\beta_6$ shows the pocket interaction with Mg^{2+} ion, in yellow. SER 125.B, GLU 223.B, and ASP 243.E are shown to indicate the key interactions with Mg^{2+}. The crystal structure is taken from the protein data bank (PDB code: 4UM9).[195] 83

4.2 The correlation between the experimental and predicted pIC_{50} for the log$S$ model with 4 PLS components (G5 data set); a regression line has been drawn through the data (solid line). 87
4.3 two subfigures MAE 89
4.4 TMACC interpretation of G5 group for two different models; log$S$ with 4 components (a, b, c, d, e, f, g, and h molecules). 91
4.5 Visualisation of the TMACC/PLS log$S$ model of the G5 data set showing the effect of modifying a *para*-CF_3-substituted molecule **15** (top, pIC_{50} = 5.3) to a *meta*-CF_3-substituted molecule **46** (middle, pIC_{50} = 5.9) and to an *ortho*-CF_3-substituted molecule **14** (bottom, pIC_{50} = 6.5). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates a slight negative contribution to the activity of the molecule. 93
4.6 Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* F-substituted molecule **19** (top, pIC_{50} = 5.8) to an-*ortho* F-substituted molecule **42** (bottom, pIC_{50} = 6.3). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. 94
4.7 Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* single Cl-substituted benzamide inhibitor molecule **10** (top, pIC_{50} = 6.6) to a di-*ortho*-Cl-substituted molecule **2** (bottom, pIC_{50} = 6.4). The orange colour provides a slight negative contribution to the activity of the molecule. 94
4.8 Visualisation of the TMACC/PLS model showing the nitrogen contribution to the activity for the highest and the less active inhibitors; The first inhibitor is molecule **28** (top, pIC_{50} = 7.4) and the other inhibitor is molecule **40** (bottom, pIC_{50} = 5.2). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. 95

4.9 (A) Structural alignment of 46 benzamides according to OMEGA. (B) The template molecule (compound **41)** 96
4.10 The correlation between the experimental and predicted pIC_{50} for the $\log S$ (G5 data set); a regression line has been drawn through the data (solid line). 97

5.1 The benzamide family, showing the scanned dihedral angle (ψ_1, C_1-C_2-C_3-N_4). 108
5.2 PES scans of molecules **20** and **22** 111
5.3 A histogram of calculated torsion angle distribution of molecule **23** with the calculated B3LYP/6-31G* PES. 113
5.4 A histogram of calculated torsion angle distribution of molecule **2** with the calculated B3LYP/6-31G* PES. 113
5.5 B3LYP/6-31G* PES for six molecules with different moieties where molecules: **2** in green solid-squares, **3** in brown circles, **4** in black squares, **21** in red triangle, **24** in blue pentagons, and **27** in cyan diamonds. 114
5.6 B3LYP/6-31G* PES for molecule (**9** in red) and (**13** in blue). The 0° orientation is the global minimum for **13** and a maximum for **9**. 115
5.7 Resonance structure, showing the orbital delocalisation between the methoxy, the phenyl ring and the amide group in 2-methoxy-*N*-methylbenzamide. . . 116
5.8 Two examples of PES profiles for 2-methoxy- and 2-cyano-*N*-methylbenzamides show the delocalisation effect on the stability of the global minima. They are presented in blue pentagons and red squares, respectively. 117
5.9 The B3LYP/6-31G* PES curves for molecules **5** and **27** show the atomic size effect of a substituent on the energy barrier. Molecule **5** is shown in blue pentagons and molecule **27** is in red squares. 118
5.10 Atomic partial charges calculated with the Mulliken and ChelpG methods for molecule **43** 119

6.1 A two-dimensional diagram of the staggered conformation of toluene defining the coordinate system 129

6.2 A molecular orbital diagram shows the six selected orbitals for the active space 130

A.1 The PES scan for benzamide molecules (**1 - 6**) 187

A.2 The PES scan for benzamide molecules (**7 - 12**) 188

A.3 The PES scan for benzamide molecules (**13 - 18**) 189

A.4 The PES scan for benzamide molecules (**19 - 24**) 190

A.5 The PES scan for benzamide molecules (**25 - 30**) 191

A.6 The PES scan for benzamide molecules (**31 - 36**) 192

A.7 The PES scan for benzamide molecules (**37 - 42**) 193

A.8 The PES scan for benzamide molecules (**43 - 46**) 194

List of Tables

3.1 Combinations of atomic properties. E = charge, P = log$P$, S = log$S$, and MR = Molar Refractivity. 55

3.2 Molecule number 10 in the G1 list. TMACC descriptors are based on topological distances. The maximum distance here is 9 between the methyl carbon and Cl. 56

3.3 Training and testing performance measures for Models Generated from Different Descriptors (G1 group) 63

3.4 The most predictive models for G1, G2, G3, and G4 data sets 63

3.5 The predicted activities for G4 group from the LOO-CV. 65

3.6 The highest predictive models using the SVM (RBFKernel), MLP, IBK, M5P, RF, and REPTree approaches on the G4 dataset. 67

3.7 q^2 Values for log$S$ only 72

3.8 Statistical data for 3D-QSAR analyses of benzamide derivatives (G1 data set). SEE is the standard error of estimate. 77

4.1 Training and testing performance measures for Models Generated from Different Descriptors (G5 dataset) 87

4.2 The predicted activities for G5 group from the LOO-CV. 88

4.3 The most predictive models using the SVM (RBFKernel), MLP, IBK, M5P, RF, and REPTree approaches. 90

5.1 Calculated and experimental geometric parameters at the optimised minimum for the *N*-methylbenzamide at different levels of theory. ψ_2 = C_2-C_3-N_4-C. Total energies at the minimum: $E(MP2) = -438.912476$ au: $E(B3LYP) = -440.2616384$ au, and $E(CCSD) = -438.963912$ au 110

5.2 Locations of local (LM) and global (GM) minima and relative energies for molecule **20** using the B3LYP level of theory and different basis sets. . . . 112

5.3 Locations of local (LM) and global (GM) minima and relative energies for molecule **22** using the B3LYP level of theory and different basis sets. . . . 112

5.4 The calculated descriptors and the predicted pIC_{50} for both the G4 and G5 datasets assessed in LOO-CV on the training set. 120

5.5 The most predictive models using the SVM (RBFKernel), MLP, IBk, M5P, RF and REPTree approaches for G4 in $\alpha_v\beta_1$ inhibitor. 122

6.1 Calculated bond distances in (Å), and angles in (°) for both the staggered and eclipsed toluene geometries of the S_0 state. 132

6.2 TD-DFT B3LYP/6-31G* calculated vertical transition energies and transition dipole moments for the eclipsed conformation; x, y, and z were defined in Figure 6.1. 133

6.3 TD-DFT B3LYP/6-31G* calculated vertical transition energies and transition dipole moments for the staggered conformation; x, y, and z were defined in Figure 6.1. 133

6.4 Permanent dipole moments of S_0 of the staggered conformation of toluene . 134

6.5 Calculated and experimental vibrational frequencies (cm^{-1}) of the S_0 state of the staggered toluene at the CASSCF(6, 6) level with the 6-31G*. [a] Scaling factor = 0.91 135

6.6 Calculated and experimental vibrational frequencies (cm^{-1}) of the S_0 and S_1 state of the staggered toluene at the CASSCF(6, 7) level with the 6-31G*. [a] Scaling factor = 0.91 137

6.7 Calculated and experimental vibrational frequencies (cm^{-1}) of S_1 state of the staggered toluene at the CASSCF(6, 7)/6-31G* and TD-B3LYP/aug-cc-pVTZ levels. 138

A.1 Datasets G1 and G2 comprised compounds 1 to 27 and datasets G3 and G4 comprised (1 - 40) compounds with their experimental pIC_{50} values. Dataset G5 comprised all compounds (1 - 46). 177

List of Abbreviations

ADME Absorption, Distribution, Metabolism, and Excretion

ATP Adenosine Triphosphate

BA Biological Activity

CASPT2 Multi-configurational Second-order Perturbation Theory

CASSCF Complete Active Space Self-consistent Field

CoMFA Comparative Molecular Field Analysis

CTGF Connective-tissue Growth Factor

DFT Density Functional Theory

ECM Extracellular Matrix

EGFR Epidermal Growth Factor Receptor

ERS Endoplasmic Reticulum Stress

ESS Explained Sum of Squares

FAK Transforming Growth Factor

FDA Food and Drug Administration

FGFR Fibroblast Growth Factor Receptor

FVC Forced Vital Capacity

HBA Hydrogen-bond Acceptors

HBD Hydrogen-bond Donors

HTS High Throughput Screening

IBk k-nearest Neighbours

IPF Idiopathic Pulmonary Fibrosis

LFER Linear Free-energy Relationship

LLE Ligand-lipophilicity Efficiency

LOO-CV Leave-One-Out Cross-Validation

MAE Mean Absolute Error

MLR Multi-linear Regression

MMFF Merck Molecular Force Field

MMP Matrix Metalloproteinases

MR Molar Refractivity

NCW Nottingham Cheminformatics Workbench

NN Neural Networks

PCA Principal Component Analysis

PDGF Platelet-derived Growth Factor

PES Potential Energy Surface

PK Pharmacokinetic

PLS Partial Least Squares

PRSS Predictive Residual Sum of Squares

QSAR Quantitative Structure-Activity Relationship

RBF Radial Basis Function

RF Random Forest

RGD Arginylglycylaspartic Acid

RMSE Root Mean Squared Error

RTK Receptor Tyrosine Kinases

SARs Structure-activity Relationships

SCOP Structure Classification of Proteins

SMO Sequential Minimal Optimisation

SSA Substructure Analysis

SSE Standard Error of Estimate

SVM Support Vector Machine

TGF Transforming Growth Factor

TMACC Topological Maximum Cross Correlation

TNF Tumour Necrosis Factor

TSS Total Sum of Squares

UPR Unfolded Protein Response

UV Ultraviolet

vEGFR2 Vascular Endothelial Growth Factor Receptor 2

Contents

Abstract i

Acknowledgements ii

Dedication iii

List of Figures iv

List of Tables vi

List of Abbreviations vii

1 Introduction 1

1.1 Drug discovery and industrial overview 1

1.2 Cheminformatics 4

1.2.1 Drug like and non drug like molecules 5

1.3 Machine learning approaches 7

1.4 QSAR models 8

1.4.1 History of QSAR 10

1.4.2 Electronic and steric effects 11

1.5 Biomedical relevance of Integrins 12

1.5.1 Structural features of integrins 15

1.6 Idiopathic pulmonary fibrosis 17

1.7 Goals of Thesis . . . 20

1.8 Thesis structure . . . 20

1.9 Conclusion . . . 21

2 Methods **22**

2.1 Regression-based methods . . . 22

2.1.1 Multiple Linear Regression . . . 24
2.1.2 Principal component analysis . . . 27
2.1.3 Partial Least Squares . . . 28
2.1.4 Cluster analysis . . . 30

2.2 Basic validation procedures . . . 32

2.2.1 Leave-One-Out cross-validation . . . 32
2.2.2 k-fold cross-validation . . . 34
2.2.3 Y-randomisation . . . 35
2.2.4 External validation . . . 36

2.3 Data mining methods . . . 37

2.3.1 Neural Networks . . . 37
2.3.2 Support vector machine . . . 38
2.3.2.1 Sequential Minimal Optimisation . . . 40
2.3.3 Decision trees . . . 41

2.4 Molecular descriptors . . . 43

2.4.1 Physicochemical properties . . . 44
2.4.2 Molar refractivity . . . 45
2.4.3 Topological indices . . . 45
2.4.4 Topological maximum cross correlation . . . 46

2.5 Comparative Molecular Field Analysis . . . 47

2.6 Conclusion . . . 49

3 QSAR for Inhibitors of $\alpha_v\beta_1$ integrin **51**

3.1 Introduction . . . 51

3.1.1 Data sets . . . 53
3.2 Methods . . . 54
3.2.1 2D-QSAR . . . 54
3.2.2 Machine learning methods . . . 56
3.2.2.1 Sequential Minimal Optimisation . . . 56
3.2.2.2 Support vector machine . . . 57
3.2.2.3 Multilayer Perceptron . . . 57
3.2.2.4 IBk . . . 58
3.2.2.5 Regression trees . . . 58
3.2.2.6 Random Forest . . . 59
3.2.3 3D-QSAR . . . 60
3.3 Results and discussion . . . 62
3.3.1 2D-QSAR with PLS analysis . . . 62
3.3.2 2D-QSAR with machine learning analysis . . . 66
3.3.3 2D-QSAR interpretation . . . 68
3.3.4 3D-QSAR Analysis . . . 77
3.3.4.1 Evaluating of 3D QSAR models . . . 77
3.4 Conclusions . . . 79

4 Inhibitors of $\alpha_v\beta_6$ integrin 81

4.1 Introduction . . . 81
4.1.1 From structure to function . . . 82
4.1.2 $\alpha_v\beta_6$ integrin–mediated TGF-β activation in IPF . . . 84
4.2 Data sets . . . 85
4.3 Methods . . . 85
4.3.1 2D-QSAR . . . 85
4.3.2 Machine learning methods . . . 86
4.4 Results and discussion . . . 86
4.4.1 2D-QSAR with PLS analysis . . . 86
4.4.2 2D-QSAR with machine learning analysis . . . 88
4.4.3 2D-QSAR interpretation . . . 89
4.4.4 3D-QSAR analysis . . . 96
4.4.4.1 Evaluating of 3D-QSAR models . . . 96

4.5 Conclusions . . . 97

5 DFT Study of Conformations of Benzamides **99**

5.1 Introduction . . . 99

5.1.1 A brief view of Density Functional Theory and Coupled Cluster methods . . . 100
5.1.2 Basis set . . . 103

5.2 Computational Details . . . 104

5.2.1 Datasets and molecular descriptors . . . 104
5.2.1.1 Boltzmann population of conformations . . . 105
5.2.1.2 Atomic charges . . . 106
5.2.1.3 Molecular orbital energies . . . 107
5.2.1.4 Dihedral angle . . . 107
5.2.2 Potential Energy Profiles of Methyl-Substituted Benzamide . . . 108
5.2.3 Machine Learning Methods . . . 109

5.3 Results and discussion . . . 110

5.3.1 Optimised parameters . . . 110
5.3.2 The PES profiles for the G4 and G5 datasets . . . 111
5.3.2.1 The PES curves for *ortho*-alkyl substituents . . . 111
5.3.2.2 The delocalisation effect on the minimum geometry . . . 116
5.3.2.3 Effect of halogens on internal rotation of aromatic amide 117
5.3.2.4 Mulliken vs. ChelpG . . . 118
5.3.3 QSAR with machine learning analysis . . . 119
5.3.3.1 MLR results . . . 119

5.4 Conclusions . . . 123

6 Excited States of Toluene **125**

6.1 Introduction . . . 125

6.2 Computational details . . . 129

6.3 Results and discussion . . . 131

6.3.1 Geometries . . . 131
6.3.1.1 TD-DFT . . . 133

6.3.2 Harmonic vibrational frequencies 134
6.4 Conclusions 138

7 Conclusions **140**

References **144**

Appendix A **177**

A.1 Data Set 177

Chapter 1

Introduction

1.1 Drug discovery and industrial overview

Drug discovery and bringing new drugs to the market through development processes, clinical trials, and being approved are challenging tasks with high risk and cost. In the Pharma environment, the story of a drug begins from associating a specific disease and how a group of molecules would affect biological targets in human cells, to several preclinical stages of tests before being approved from Food and Drug Administration (FDA). The estimated time to discover a new drug and bring it to the market takes between 12-15 years[1] (Figure 1.1).

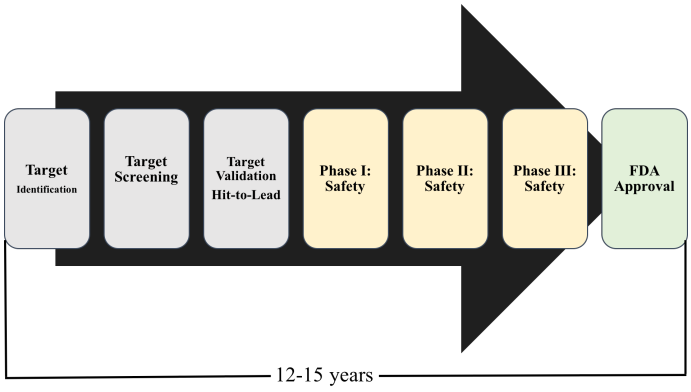


Figure 1.1: The steps that are essential for a drug to proceed through the research and development pipeline.

Figure 1.1 shows the standard steps for the completion of drug discovery. Four main levels such as target identification and validation, lead identification, lead optimisation and pre-clinical trials are used in the Drug discovery process. Target molecules can be genes, proteins or they could be possibly anything within a living organism for a specific disease.[2] In the target identification process, a ligand binds to the identified protein in its isolated form. Some methods involve few promising drugs which have the ability to inhibit the functionality of a disease. The target identification step usually corresponds to what is called the "Hit-to-lead" phase. High-throughput screening (HTS) is an experimental method that allows handling of chemicals or pharmacological tests. Within this method, one can quickly identify active and inactive molecules that control biomolecular pathways. During the HTS step, molecules which bind strongly to the targets undergo process to identify the lead molecules. Virtual screening, data mining, HTS, quantitative structure–activity relationship (QSAR), protein-ligand models, structure-based models are examples of approaches used for lead identification and optimisation. The preclinical stage involves Phase I - III which are important to verify whether the candidate molecule can be made into a drug which is not toxic and has minimum side effects.

Despite the enhanced spending on research and development (R&D), however, barely about 11% of drugs enrolled in clinical development approach the marketplace.[3] Some molecules failed to fulfil the drug properties which comprise absorption, distribution, metabolism, and excretion (ADME). These properties are important in processes that determine pharmacokinetic (PK) and toxicity, and are required for a drug to progress through the R&D cycle. course of treatment for a single patien The average cost of treatment for a single patient surpasses $100,000 annually in the United States.[4,5] In the US, the total expenditure from 2010 to 2020 is expected to increase by 26%.[6] This enormous cost is being identified as a global problem not only in the USA, but in the UK, and in every other country.

Idiopathic pulmonary fibrosis (IPF) is a fatal disease, which affects lungs resulting in the death of patients within 2-5 years from diagnosis.[7] Non-small cell lung cancer is similar to the 5-year survival rate of 20% - 40% associated with IPF. However, the latter is worse than that of many other cancers. In 2014, the FDA and US Food approved two drugs, pirfenidone and nintedanib, for the treatment of the IPF (Figure 1.2).[8] Pirfenidone and nintedanib were also approved by European regulatory agency for the treatment of the IPF patients. These two drugs mark a turning point for the IPF patients and physicians in the US and around the world.

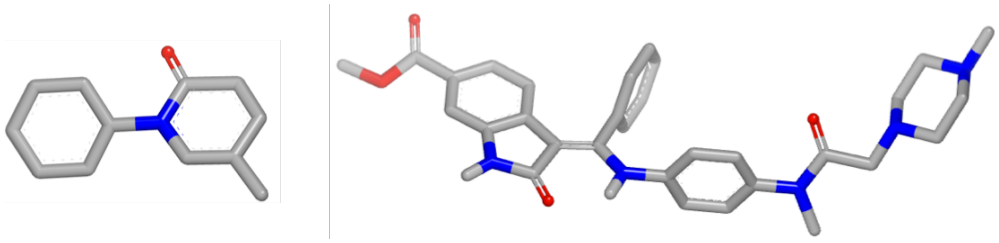


Figure 1.2: Molecular structures of pirfenidone (left) and nintedanib (right)

Both pirfenidone and nintedanib have been shown to decrease the rate of progression of IPF. Nintedanib is a dominant kinase inhibitor blocking the effects of growth factors implicated in the pathogenesis of IPF.[9] It has been clinically evaluated that the pirfenidone is safe and was used in various disorders such as IPF, chronic hepatitis C, multiple sclerosis and neurofibromatosis.[10] Pirfenidone is a pyridone derivative, which is characterised by anti-inflammatory, anti-fibrotic and anti-oxidant properties. This drug also is capable of slowing down the IPF disease progression.

IPF patients are monitored with so-called forced vital capacity (FVC), which measures the disease progression.[11] In a recent study,[12] pirfenidone treatment was applied in a total of 56 patients with IPF. Based on their rate of FVC decline in the 12-month pretreatment period, pirfenidone gives the ability for researchers to measure the patient population in rapid and slow progressors. This study confirmed that pirfenidone treatment reduces the rate of FVC decline in patients with IPF.

Anticancer and antifibrotic medicines are expensive. The average benefit obtained from these drugs is still limited. Out of 47 FDA anticancer therapeutics, hardly 9 matched the typical pattern of significant clinical benefit concerning overall survival set by American Society of Clinical Oncology (ASCO).[13] Their costs are justified by the high cost of R&D expenses.[14] Researchers at Tufts Medical Centre (TMC) settled that the total cost of developing a new medicine is roughly $2.6 billion.[14,15] Another research group reported that the cost of producing an approved drug reach $160 million, which is more than an order of size lower from the TMC group.[5] The TMC researchers suggested that the total cost is also accounted for the cost of high failure price.[5,14]

Many innovative platforms and methods such as Cheminformatics[16] need to be involved in the drug discovery to reduce the cost and to accelerate the pace of drug discovery. Distinguishing candidate failures ahead in the process is very helpful before jumping into costly later-phase steps. Significant contributions to chemoinformatics studies have been made by chemists, pharmacologists, bioinformaticists, and researchers in computer sciences to help the selection and optimisation of chemical structures for pre-synthesis steps. It also helps to reduce the chances of failure in the advanced levels of clinical trials.

1.2 Cheminformatics

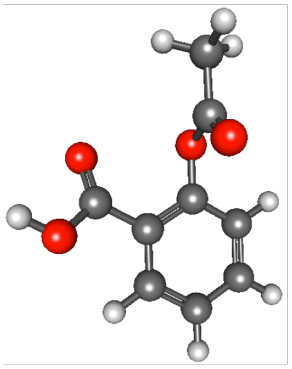
Cheminformatics is an *in silico* technique that uses computers and chemical information. The term chemoinformatics was first used in 1998.[17] In chemoinformatics, chemical structures can be generated in different forms: one-, two- or three-dimensional (1D, 2D or 3D), with the manipulations including a series of exploring, modelling, and mathematical methods.[18] Chemoinformatics covers two crucial scientific aspects. First, it can cover large databases of chemical compounds. For instance, it provides

appropriate technical methods to handle searching for some similar properties among these molecules. Second, chemoinformatics can deal with bioactive compounds that are of interest to industrial sectors such as pharmaceutical.

In 1957, Ray and Kirsch[19] published the first paper to define an algorithm for substructure searching, in which a user can specify a substructure to detect all molecules contained that query among a database. This work studied in graph theory applications and representation of the 2D and 3D-chemical structures. The 2D structure for a molecule in graph theory (chapter 2) is designated as nodes and edges that denote the atoms and bonds of a chemical compound, respectively. In 1961, the value of the computer for chemical information processing was acknowledged.[20]

1.2.1 Drug like and non drug like molecules

Various publications over the last two decades or so have shown the importance of different physicochemical properties, especially lipophilicity. In 1997, Lipinski et al.[21] published the novel "rule of five" principle in which they stated physicochemical property-based criteria for a drug to pass through a cell membrane (permeability) and solubility. Lipinski's group obtained a simple set of rules by profiling 2245 molecules from the Derwent Drug Index. They found that these compounds involve features related to: molecular size (such as molecular weight and molar refractivity), molecular complexity (such as the number of chiral centres, the number of rotatable bonds, and the number of aromatic rings), polarity (such as the number of hydrogen-bond donors (HBD), the number of hydrogen-bond acceptors (HBA), polar surface area, and molar refractivity), lipophilicity (i.e. log$P$ and log$D$ as will be detailed in chapter 2), as illustrated for the aspirin molecule in Figure 1.3. These physicochemical properties are easy to calculate.



Aspirin

Property	Value
Molecular weight	180 g/mol
Hydrogen-bond donor	1
Hydrogen-bond acceptors	4
log$P$	1.31
Molar refractivity	44.7 m^3/mol

Figure 1.3: An example shows the properties and their values for a drug-like molecule.

A drug-like molecule is usually characterised by: a molecular weight $\leq$ 500 Dalton, log$P \leq 5$, number of HBD ≤ 5, number of HBA ≤ 10, and a Molar refractivity of 40-130.[22] When a molecule fails to comply with two or more of the mentioned properties, it will fail to pass through the research and development pipeline. Search criteria in a large database were extended to involve "Lead-like" molecules. These leads are characterised by small molecular weights and have biological activities likely to be promising drugs, but require structure modification to adapt sufficiently to the biological targets. Including lead-like molecules in searching features of several databases led to an expansion in the use of in *silico* methods for drug design and discovery; for instance, researchers can extend their searching molecules to include those with calculated physio-chemical properties which might be lead- or drug-like.[23] The distinction between drug-like and lead-like has been reported in detail elsewhere.[24]

Computer power and its capability increase the tremendous growth of cheminformatics since computers are the core of any chemoinformatics approach. Chemoinformatics covers topics such as QSARs. Most techniques are available mainly from 2D all the way to 6D. In this thesis, we have considered 2D and 3D-QSAR methods, and they will be discussed in detail.

1.3 Machine learning approaches

Machine learning algorithms and/or data mining methods can speed up a few steps in this long journey. To discover useful patterns and correlations in large databases or datasets of molecules, we need to employ an automatic extraction method using high-level search techniques and algorithms. Such an approach is called data mining or machine learning. In data mining, different models that correlate a set of molecular properties to biological activity can be created. Then, the predicted models can be used to calculate the activity of new molecules. Data mining models can be either simple (simple equations obtained from linear methods) or complex (nonlinear models arising from nonlinear procedures).[25]

Substructural analysis (SSA) was the first application of machine learning in drug discovery.[26] Cramer et al. used the SSA method for the automated analysis of biological screening data.[26] In 1990s, a fast and simple scoring function to estimate free binding energy ΔG for a protein-ligand complex of known molecular structures was developed by Böhm.[27] This function was designed from several approaches (such as regression-based empirical scoring function, molecular mechanics, a potential of mean force, descriptors and protein-ligand atomic coordinates).

Machine-learning methods are composed of a few critical steps in general.[28] Choosing a proper set of properties or chemical descriptors, which are thought to carry the vital properties of the molecules are included. The descriptors selection is not an easy step. There are over 5000 molecular descriptors. Using more descriptors generates noise and might confuse the learning method. Some selected descriptors can work well for a specific data set. However, the exact descriptors might be irrelevant for a different data set. The properties that may distinguish the active from the inactive compounds need to be identified using the machine learning algorithm.[29] Typical algorithms for QSAR are linear regression, multi-linear regression (MLR),[30]

principal components analysis (PCA),[31–33] partial least squares (PLS),[34–40] decision trees, neural networks (NN),[41–43] random forest (RF),[44] and support vector machine (SVM).[45–49] Further details of these algorithms are detailed in chapter 2 and 3. The literature is full of examples of all these algorithms on many targets and several cases. For example, fifteen supervised machine learning algorithms were employed to classify a set of 11,360 pairs of protein domains.[50] It was found that the random forest is the most accurate classifier for their datasets. Another example, Jain and Hirst[41] used the random forest to predict the structure classification of proteins (SCOP).[51] The SCOP method is a common scheme for classification.

Drug discovery has been a common testbed for the application of machine learning for the past 15 years.[52,53] Reducing the amount of supervision needed to make useful predictions in drug discovery is a key challenge.[54] Oglic et. al. focus on discovering molecular structures from a design space defined with a proposal generator in which an integrin antagonist serves as a parent compound and all the designs are obtained by altering that compound. The designed-space size is of the order of 185,000 molecules, which is large enough to evaluate the efficiency of the used algorithm. This study discovered interesting compounds from the aspect of industrial medicinal chemists.[55]

In the next sections, we present a review on QSAR models. Also, we will show the biological activities of integrins with respect to their involvement in respiratory diseases such as pulmonary fibrosis. Our ultimate goal is to find novel and effective pharmaceuticals for the treatment of chronic lung disease.

1.4 QSAR models

From the fundamental chemical and physical theories, it is clear that there must be a relationship between the structure of a molecule and its biological activity. However, this connection is a complex, and it is doubtful that the biological activity can ever be

predicted entirely. For instance; when a drug is taken orally, it must first survive the digestive enzymes in the stomach, pass over into the bloodstream, diffuse and bind to the target enzyme at a distinct site with a large binding constant, and eventually have a proper half-life time in the organism before being degradable into non-toxic parts.

Correlating chemical structures, statistical and physical properties with biological activity by several methods is generally referred to as QSAR. One of the primary goals of QSAR studies is how to predict biological activity. However, the capacity of a proposed drug to act upon a specific target is pointless if the molecule cannot hit the target. A significant effort was made to explain the effects of molecules on living organisms. This means that understanding the interactions of compounds with cell membranes, organelles, etc. is of interest for many medicinal chemists. QSAR is an important paradigm for understanding such interactions. A general form of QSAR equation can be given as

$$\text{Activity} = f(P_1, P_2, \cdots P_n) \tag{1.1}$$

Activity is represented by (pIC_{50}). The pIC_{50} is a quantitative measure indicates how much of a particular inhibitor is needed to inhibit a given biological activity by half. Activity is also a function (f) of molecular properties ($P_1, P_2, \cdots P_n$), which comprises important chemical information about the molecule. In the first steps of developing a novel drug, the focus is normally on the binding to the target organism and having proper lipophilicity.[56]

QSAR models involve numerous steps of chemoinformatics and machine learning techniques. Current QSAR is challenging in the amount and diversity of descriptors considered. Normally, three main components are essential to a QSAR model: the dataset, the molecular descriptors and the chosen statistical procedure to find a re-

lationship between them. The datasets are molecules that need to be encoded in a format where the chemical structures or experimental data is included. As a second step, we have to choose the appropriate molecular descriptors to calculate. Final step, several statistical approaches are used to identify those descriptors that predict the correlation between the molecules and the biological activity.

1.4.1 History of QSAR

The thought of QSAR started from the idea of Crum-Brown and Fraser in 1868.[57] They proposed that biological response was a function of the chemical structure. One year later, the relationship between the toxicity of simple ethers and their water solubility was reported by Richardson.[58] Later on, Meyer[59] and Overton[60] were addressed a linear relationship between lipophilicity shown as oil-water partition coefficient or solubility and biological effects.

A novel paper in 1962, which attempted to correlate biological activity with molecular physicochemical properties using linear regression method was published by Hansch et al.[61] This paper was the starting point for QSAR. Subsequently, Hansch and Fujita[62] used lipophilicity as a physio-chemical property to predict the biological activity of phenoxyacetic acids. Lipophilicity, sometimes referred to as ligand-lipophilicity efficiency (LLE), is a physio-chemical parameter that links activity and lipophilicity in drug discovery to estimate drug. The log$P$ is a fundamental property for any promising drug. For a drug to be a cell-membrane permeable, the log$P$ should not be too low. Also when the log$P$ becomes very high, then it remains in situ, as will be described in 2.4.1. Other descriptors have also been used such as the molecular weights, polar surface area, number of hydrogen bonds, number of rotatable bonds, boiling point, melting points, chromatographic retention index, toxicity, IR frequencies, dipole moments, ... etc. Here we will discuss few obvious QSAR descriptors.

1.4.2 Electronic and steric effects

A diversity of molecular descriptors, in a classical QSAR, is involved besides using the log$P$. These include a measure of electronic and steric substituent parameters (Hammett and Taft constants),[63] and pKa/b values for acidic and basic moieties. Hammett linked electronic properties of benzoic acid derivatives, *meta*- and *para*-substituents, with their equilibrium constants and reactivity. This equation represents a linear free-energy relationship (LFER) with two parameters: a substituent constant (σ) and a reaction constant (ρ).

$$\log\frac{K}{K_0} = \sigma\rho \tag{1.2}$$

Here K, and K_0 for a given reaction are equilibrium and reference constants, respectively. Hammett found that the benzoic acid derivatives have σ equal to 0. Electron donating substituents gave negative values for σ while electron withdrawing substituents provided positive values due to the stabilising influence on the anion. Hammett also noticed that the same substituent in the *meta*- or *para*-positions have different σ values. This might due to the resonance influences at the *para*-position.

After the Hammett equation, the steric effects of substituents against activity (E_s) were proposed as a descriptor by Taft.[64] Taft examined the impacts of substituents in the hydrolysis of esters in an acidic solution as shown below

$$\mathrm{E}_s = \frac{K_R}{K_{Me}} \tag{1.3}$$

Where K_{Me} is the methyl derivative, and K_R represents the R-substituted molecule. The hydrolysis should be done under an acidic condition because the hydrolysis of esters depends only on the steric effect. To describe the properties upon the activity adequately, QSAR parameters are needed to investigate the correlation. In 1964,

Hansch[62] used more than one descriptor to predict the biological activity (BA).

$$\log(\mathrm{BA}) = c - a\pi^2 + b\pi + \rho\sigma + \mathrm{E}_s \tag{1.4}$$

Where a, b, and c are constants, π is a hydrophobic constant, which can be given as

$$\pi = \log P_x - \log P_H \tag{1.5}$$

Here P_x is the partition coefficient of the substituted compound and P_x represents the unsubstituted reference molecule.

The philosophy in QSAR studies is to combine few molecular descriptors that can be easily generated and then let the statistical approaches sort out which of these are relevant. It is clear that with many of the descriptors, MLR might not be suitable. Therefore, the PCA or PLS methods are better at detecting poor descriptors and dealing with the fact that the calculated activities usually have relatively large uncertainties. These statistical methods, as well as 2D/3D-QSAR approaches, will be discussed in detail in chapter 2.

1.5 Biomedical relevance of Integrins

Normal cells require signals usually delivered by ligands to control cell growth or apoptosis. However, abnormal signal transductions in cancer cells result in uncontrolled proliferation, loss of apoptosis, tissue invasion and metastasis (spread of cancer cells), and enhancing blood supply of tumour (angiogenesis).[65] Ligands or signals can be in the form of growth factors and inhibitors, extracellular matrix (ECM) or cell adhesion molecules. These signals are transmitted into the cell through proteins which are located on the surface of the cells called "integrins or receptors".

Integrins are cell-surface heterodimeric receptors which act as transmembrane

linkers between the ECM and the cytoskeleton. The integrin consists of three primary domains: an extracellular ligand-binding domain, transmembrane domain, and intracellular domain (Figure 1.4).

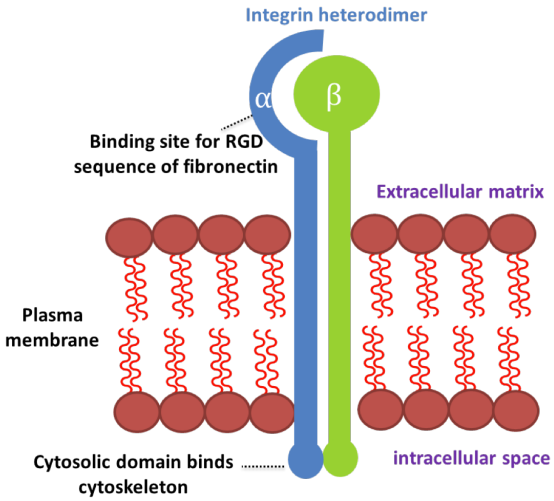


Figure 1.4: General schematic structure of $\alpha_v\beta$ integrin. The α- and β-subunits are indicated in blue and green, respectively. Cytoplasmic cells are circled in red.

Each ligand binds to an integrin. Integrins perform an essential role in cell signalling. They can modulate the cell signalling pathways of transmembrane protein kinases such as receptor tyrosine kinases (RTK). Binding a ligand to the ECM domain stimulates the receptor tyrosine kinase, which is capable of activating other proteins inside the cell. The tyrosine kinase is a protein that transfers a phosphate moiety from adenosine triphosphate (ATP) into the living cell.[66] A signal passes to the intracellular domain, when a ligand binds to the receptor, to activate the associated enzyme. Once this signal reaches the nucleus, the nucleus directs the cell to grow or to stop growing. Malignant cells have their own growth factors. This means the cancer cells can grow and divide with the reduced help from external growth stimulation. Some cells are able to produce their own growth factors and stimulate their own growth.[67] This is called "Autocrine Stimulation". The autocrine stimulation occurs in many tumour cells such as glioblastoma, lung cancer, and sarcomas. For

instance, glioblastoma multiform (GBM) expressed a platelet-derived growth factor (PDGF).[68] Lung cancer expressed epidermal growth factor receptor (EGFR) while sarcomas represented EGFR and transforming growth factor (TGF-β).[69] The production of cell-surface receptors in normal cells is limited by cellular restraints gene expression and protein translation. In cancer cells, however, mutations in the genes encoding for the receptors interrupts this finely attuned regulation, and many copies of genes are generated. This aspect is called gene amplification which leads to the excessive production of receptors (a higher than normal copies of receptors on the cell surface).[70]

When ligands overbind to the excess receptors, cancer cells enhanced the potential to become a tumour. The more receptors exposed, the more ligands can fit into the binding sites. Over-expression of growth factor receptors can occur in ligand-independent signalling where receptors are active in the absence of stimulating molecules. Structural changes to a receptor can also lead to ligand-independent activation. For example, truncated versions of the epidermal growth factor receptor where much of the intracellular domain is missing are consistently active. Epidermal Growth Factor (EGF) receptor also known as HER1 or c-ErbB-1 is a transmembrane receptor and a member of a subfamily of kind 1 receptor tyrosine kinases.[71] These receptors are found principally in the membranes of normal epithelial cells such as in skin, colon, lung, and other EGR receptors. The EGF receptors play a vital role in the control of cell reproduction, differentiation and remnant. EGFR is over-expressed in colon-, rectum-, lung, head, and neck-tumours. When a specific ligand binds to its receptor, this leads to changes in the receptor that forward a distinct signal into the cell. For example, the receptor tyrosine kinase is activated and begins a signalling pathway unique to that receptor. This appearance is called signal transduction. Activation of a signal transduction pathway forms a complicated series of events in the cytoplasm that ultimately drives into the cell nucleus where the regulating of cell

cycle progressions is stimulated resulting in cell proliferation.

A recent study[72] shows another therapeutic approach by targeting molecules downstream of integrins instead of integrin themselves. For example, Focal adhesion kinase (FAK) is enhanced in focal adhesions. FAK is a non-receptor tyrosine kinase associated in β-integrin signalling. Both FAK and Src kinase control many cellular functions such as cancer cell migration and cell invasion. Some clinical tests and experiments are continuing with FAK and Src inhibitors. These trials are under phase I-II safety and pharmacokinetic evaluation.

1.5.1 Structural features of integrins

Cancer is a complicated illness and grows within an aggressively developing ECM that regulates every phase of tumour cells. Various functions in cancer cells are regulated by integrin signalling such as apoptosis, proliferation, migration, and invasion.[65] Integrins are also detected in cancer host-cells such as fibroblasts, perivascular cells, and the vascular endothelium. Thus, signalling transactions in both the tumour cells and ECM can be blocked by integrin antagonists.[65] The conformational activation process in integrins is often required to motivate their adhesive function.[73] The matrix receptor ties the cell's cytoskeleton to the matrix. Integrins are the essential way that cells both bind to and respond to the ECM. Integrins are a species of 24 transmembrane heterodimers formed from a sequence of 18 α integrin and 8 β integrin subunits to date. Each heterodimer molecule consists of two non-covalently transmembrane glycoprotein subunits called α and β. The numerous sequences provide the variety of receptor species are shown in Figure 1.5.

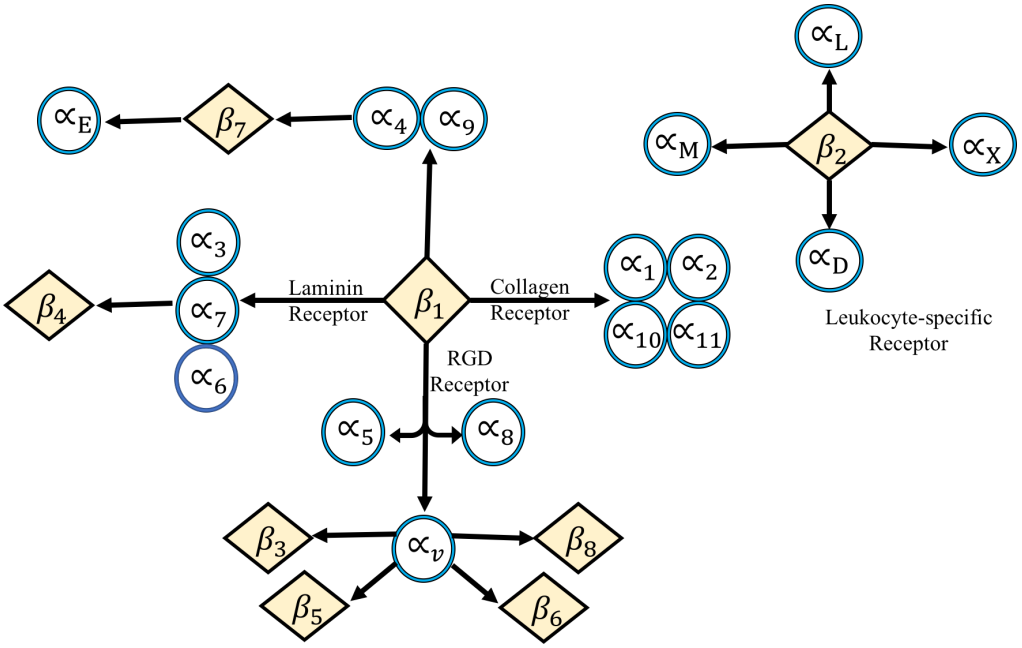


Figure 1.5: Integrin families. This diagram outlines the 24 possible integrin dimer species, grouped by substrate type.

Figure 1.5 shows the collagen-binding β_1 integrins and Arg-Gly-Asp acid RGD-binding β_1 and β_3 receptors. Binding was expressed as a two-step process when the α subunit connects to ligand first determined by the association kinetics and influencing ligand identification. The β moiety binds next, to determine dissociation kinetics and to form bond stabilisation, which is relevant to a divalent kinetics model with the α subunit denoting the high-affinity region and the β subunit as the low-affinity place.[74,75] A sequence from both subunits forms the ligand-binding site of integrin,[76] and the cytoskeleton is connected by its cytoplasmic domain.[77] The collection of integrins present on a given cell directs the extent to which that cell will adhere to and migrate on different forms.[65] Integrins of α_v have the RGD sequence on their corresponding ligands. The α_v inhibitors were first recognised by their capability to distinguish the RGD sequence.[78]

Every integrin-growth factor receptor pair is responsible for cell survival pathways. For instance, $\alpha_v\beta_3$ crosstalk with fibroblast growth factor receptor (FGFR) inhibits

apoptosis in endothelial cells.[65] Moreover, the cell death can be inhibited by a combination of $\alpha_v\beta_5$ and the vascular endothelial growth factor receptor 2 (vEGFR2).[65]

1.6 Idiopathic pulmonary fibrosis

Idiopathic pulmonary fibrosis (IPF) is a lung disease that causes progressive scarring of the lungs, and subsequently leads to death.[79] The IPF is a chronic condition, which is most commonly found in older adults (50 years old and above).[80] Although determined fibroblast activation and reproduction are thought to be targetable mechanisms, the pathogenesis of IPF is not entirely understood.[81] IPF for several years was thought an inflammatory disease in the lung cells. However, IPF was shown as an epithelial-driven condition in which an abnormally activated lung epithelium gives fibroblast migration, proliferation and differentiation into active myofibroblasts.[82] In the IPF, fibroblasts manifest uncontrolled proliferation, differentiate into myofibroblasts, and generate excessive ECM.[81] Other cellular sources such as bone marrow-derived fibrocytes and epithelial cells are involved in disease-associated fibroblasts.[83] Furthermore, fibrocytes can enter damaged tissue during chemokine-dependent mechanisms and develop into collagen-producing myofibroblasts.

Less women have been reported with IPF than men, and the majority of patients have a smoking history.[84,85] In the United Kingdom, an overall incidence rate of 4.6 per 100,000 person-years was reported and the annual incidence of IPF was estimated to increase by 11% between 1991 and 2003.[79,86] In the United States, a significant database of health care was used to determine the incidence of IPF.[87] They found that the IPF occurs between 6.8 and 16.3 per 100,000 person-years. Over the past two decades, death from pulmonary fibrosis has increased.[80] The IPF is an aggressive disease with a median survival time of 2-3 years from the time of diagnosis.[88] In the US, the fatality rate in 2003 was 54.5 per 1,000,000 in women and 61.2 deaths per

1,000,000 in men.[89] The death rate for IPF in Japan was estimated to be 24 per 1,000,000 in women and 33 per 1,000,000 in men.[90]

The common symptoms of the IPF are short breath, fatigue, unexplained weight loss for no known reason, a dry cough that does not get better, bibasal or bibasilar crackles, and a deformity of the finger (finger clubbing).[84–86] The IPF symptoms are also attributed to ageing, deconditioning or other comorbidities, such as cardiovascular disease and pulmonary emphysema.[82] The expected prevalence of IPF in the US is about 400 cases per 100,000 people for patients ageing more than 65 years.[91] As the number of IPF increases with age, a pathological specimen of IPF as a disease of age-related epithelial cell dysfunction has developed.[92] Age has been included into useful diagnostic algorithms clinically.[93] A cough in older individuals, besides age, should be seen as chance factors for IPF disease.

The lungs are endangered to many scars but exhibit an exceptional capability to recover through synchronised biological processes. However, repeated alveolar epithelial damage in a genetically ageing person could lead to pro-fibrotic epigenetic reprogramming, early and persistent epithelial cell catabiosis, extreme production of pro-fibrotic layers and continuous activation of mesenchymal cells, leading to a maladaptive response and consequent growth of IPF.[82]

Numerous risk agents have been associated with the pathogenesis of IPF, involving chronic viral diseases, environmental illnesses, smoking, and other disorders that an individual patient might have other than the primary IPF. Genetic variants[94] are of these factors, which lead to the IPF disease, and are estimated to explain up to one-third of the original individual risk of that disease.[95] Age is also another robust risk factor for IPF, which suggests that lung senescence is a deriving force for its progress.[93] The cellular signs of ageing occur early in IPF. Gas exchange takes place within alveoli in the lung. Air-filled sacs contained type 1 and type 2 epithelial cells (AEC1s and AEC2s). IPF leads to AEC2s apoptosis. In a healthy

lung, damage of AEC1s after an injury is replaced by proliferation and differentiation of AEC2s and stem cells. However, affected AEC2s generate damaged cells, and their capacity to renew normality is severely reduced: this is often happening in the IPF pathogenesis.[96] Certainly, IPF is a complicated illness rising from interactions between these risk factors and the relative contribution of each factor might differ among people.

Protein over-expression and endoplasmic reticulum stress (ERS) occur during a cellular activity. The ERS is a protective pathway, which arises when there is an inequality between the cellular need for protein synthesis and the endoplasmic reticulum. Another pathway is an unfolded protein response (UPR), which is a threat to all living cells and leads to cell apoptosis. The activation of UPR has various outcomes on cellular behaviour, which is not entirely understood. The UPR stimulates apoptosis by producing of profibrotic mediators such as TGF-β_1, connective-tissue growth factor (CTGF), many matrix metalloproteinases (involving MMP1, MMP7, MMP19, membrane-type MMP1 and membrane-type MMP2), tumour necrosis factor (TNF), angiotensinogen, PDGF, osteopontin, C-X-C motif chemokine 12 (CXCL12), and chemokine C-C motif ligand 2 (CCL2).[97]

The most crucial mediator in IPF pathogenesis is probably the TGF-β_1. Through $\alpha_v\beta_6$ activation, AEC2s may provide TGF-β_1 as an actin/myosin-mediated cytoskeletal contraction. One of the essential biological process is the $\alpha_v\beta_6$ integrin/TGF-β_1 pathway. TGF-β_1 can be activated also by lysophosphatidic acid (LPA), which can be regulated by autotaxin. TGF-β_1 mediator is able to increase the majority of cellular functions such as cell apoptosis, epithelial mesenchymal transition (EMT), cell migration, production of other profibrotic mediators into CTGF and other pathways. More details about integrins: $\alpha_v\beta_1$ and $\alpha_v\beta_6$, and integrin-mediated TGF-β activation in IPF will be discussed in chapter 3 and chapter 4.

1.7 Goals of Thesis

In our work, we have two primary goals. First of all, we aimed to build QSAR models against $\alpha_v\beta_1$ and $\alpha_v\beta_6$ inhibitors. We have used para- and non-parametric methods to correlate the molecular descriptors and the biological activity, and to build predictive models. The models can be used later to calculate the activity values of new compounds. The second goal of this thesis is studying a model system for biological chromophore such as toluene, which is the side chain of the amino acid phenylalanine. This investigation will provide insight into the spectroscopy of proteins in the near ultraviolet (UV) region.

1.8 Thesis structure

In this thesis, we have covered several topics of chemoinformatics such as QSAR, machine learning algorithms. In chapter 2, we present in detail all the methods which have been employed. We discuss several parametric and non-parametric methods. The models were validated using validation methods such as leave-one-out cross-validation (LOO-CV), Y-randomisation, and an external test set. The validation process is a very important step to make sure that there is no chance for under- or over-fitted data. We also present how to calculate the 2D- and/or 3D-molecular descriptors, and the software were employed in our calculations.

Chapter 3 covers the QSAR for inhibitors of $\alpha_v\beta_1$ integrin. In this chapter, we have shown how data sets were prepared. Different chemical structures (fragment or a whole structure) were considered in our calculations. A wide variety of 2D-QSAR descriptors were calculated using the NCW software, while Q-chem was used to calculate the quantum molecular descriptors. We also have employed different machine learning algorithms such as SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and

REPTree approaches. All models were validated using LOO-CV or Y-randomisation methods. The the 2D-QSAR models were visualised to show the effect of different substituents of the aromatic ring (Table A.1) against the activity. In chapter four, we have followed the same steps as chapter three, however, this time the data sets were different and all the QSAR models were developed against $\alpha_v\beta_6$ inhibitors.

In chapter 5, we discuss DFT and use it to compute some specific descriptors. We also have shown how more molecular descriptors: dihedral angles, electrostatic potential (ChelpG), molecular orbital energies (HOMO, LUMO), and Boltzmann populations were calculated. The potential energy surface scans (PES) for benzamide derivatives were investigated and local minimum for each molecule was determined. The alignment process is described.

In chapter 6, we have studied toluene as a model system for biological chromophores. Understanding the photophysics of such a system affords insight into the photophysics of biomolecules such as proteins. The multi-configuration ab initio method: CASSCF was employed to investigate the electronic $S_0 - S_1$ transition in toluene. We also have calculated the harmonic vibrational frequencies for both ground and excited states.

1.9 Conclusion

IPF is a fatal disease, but no conventional cure exists yet. Thus, QSAR with cheminformaticians are required to discover a future therapy for IPF. This can be accomplished by using suitable statistics that interpret the data and using the chemical structures into the QSAR models. Proper statistics are a subject of good practice to decide how beneficial a model is. Increasing confidence in models depends on robust statistics and reliable interpretation. Any modeller would seek the presence of readily available interpretation.

Chapter 2

Methods

Properties and activities of molecules depend on the presence of specific atoms and/or functional groups in their structures. To investigate the relationship between molecular properties and activities, we can use statistical techniques. In this chapter, we will discuss several statistical approaches. The main goal of these methods is to generate a model which explains an unknown relationship. Regression analysis is one such example of these important statistical techniques and has numerous applications in almost every field, including the chemical and physical sciences, engineering, economics, and biological sciences.

2.1 Regression-based methods

As a part of a problem in which regression investigation may be helpful, suppose that we need to analyse the relationship between the observed activity y and a property x for a data set of N molecules. We suspect that their observed activities are related somehow to that property. The N observations (for $N = 20$) are plotted in Figure 2.1. This graph is called a scatter diagram. This display clearly suggests a relationship between the experimental activities and the descriptor x; in reality, the impression is that the data points commonly, but not exactly, fall along a straight line. Figure 2.1

illustrates this straight-line relationship.

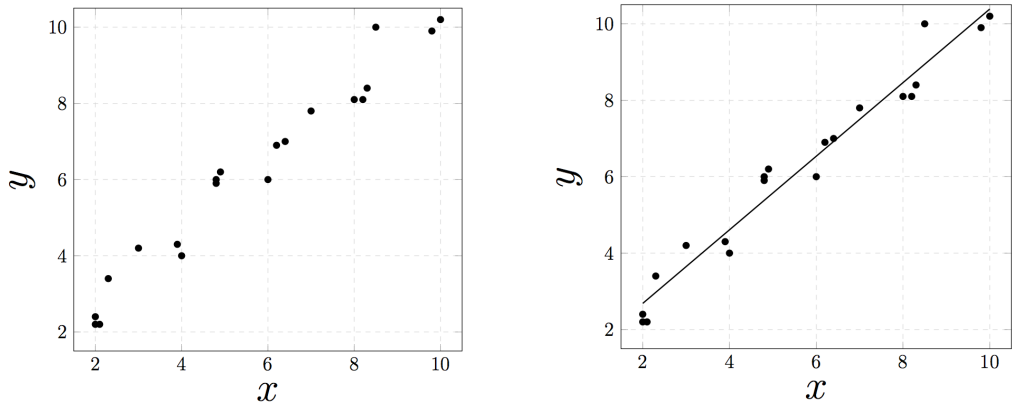


Figure 2.1: (a) Scatter diagram for molecular descriptor x and dependent variable y. (b) Straight-line relationship between molecular descriptor and dependent variable

If we let y represent the experimental activity and x represent the molecular property, then the equation of a straight line describing these two variables is

$$y = b + ax \tag{2.1}$$

Here b is the intercept and a is the slope or gradient of that line. The purpose of linear regression is to find values for the constant b and the coefficient a that minimise the sum of the differences between the values predicted by the equation and the experimental values. The values of a and b are given by the following equations (in which N is the number of data points):

$$a = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.2}$$

$$b = \bar{y} - a\bar{x} \tag{2.3}$$

where $\bar{x}$, and $\bar{y}$ represent the means of the independent and dependent variables, respectively.

Equation 2.1 is called a simple linear regression model. Customarily, x is called the independent variable and y is called the dependent variable.

Correlation of the activity data or property with molecular descriptors is a statistical concept. When the property data of molecules are quantitative, then a regression-based method can be employed to enable quantitative prediction of the property. This approach is most commonly used in MLR and PLS, which we discuss in more detail later in the chapter.

2.1.1 Multiple Linear Regression

MLR is an extension of linear regression to more independent variables (different molecular descriptors; e.g. descriptor 1, descriptor 2, and descriptor 3). A linear regression model may take the following form:

$$y = b + a_1x_1 + a_2x_2 + a_3x_3 \tag{2.4}$$

In the above expression, b is the y-intercept in the regression. The corresponding regression coefficients a_1, a_2 and a_3 are calculated through the use of estimators such as the least-squares method. The interpretation of contribution of each descriptor is straightforward depending on the corresponding coefficients and its algebraic sign. Equation 2.4 has a form which is adopted in Free-Wilson analysis,[98] which is a technique used in the early days of the QSAR. It generates an equation starting from a series of substituted compounds with their associated biological activities. The linear approach works well, when the phenomenon of interest is linear. For other systems, non-linear methods (e.g. artificial neural networks) need to be employed (see Data mining methods section). The quality of a regression equation can be assessed based upon a number of statistical parameters as described below.

Determination coefficient (r^2)

The canonical formulation of r^2, the coefficient of determination, is often used for regression models to assess the model predictivity. The determination coefficient (r^2) for the QSAR model can be calculated in the following manner:

$$r^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{2.5}$$

where i is an index for the N molecules in the dataset.

In the above expression, y_i stands for the experimental activity, while $\overline{y}$ is the corresponding average, and $\hat{y}_i$ are the predicted values. A high value of r^2 is not necessarily an indication of a predictive model. For example, increasing the number of descriptors in a model for a fixed number of observations will always increase r^2 values. However, the estimate of the reliability of the model can be corrected by using adjusted r^2 (r_a^2) in order to reflect the explained variance, as in the following equation:

$$r_a^2 = \frac{(N-1) \times r^2 - k}{N - 1 - k} \tag{2.6}$$

In the above equation, N designates the number of molecules and k is the number of descriptors.

Fisher statistic (F value)

The Fisher statistic F value is a variance-related parameter, which is used to measure the overall statistical significance of the regression coefficient.[99] The F value can be calculated as follows:

$$F = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2/k}{\sum(y_i - \hat{y}_i)^2/(N-k-1)} \tag{2.7}$$

A high value of F indicates that a significant correlation has been achieved.

Standard error of estimate (SEE)

The standard error of estimate (SSE) of any QSAR model should be low and can be calculated as follows:

$$SEE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - k - 1}} \tag{2.8}$$

r^2 is also can be measured using several quantities:[16] Total Sum of Squares (TSS), Explained Sum of Squares (ESS), and Predictive Residual Sum of Squares (PRSS). These three quantities are given in equations 2.9, 2.10, and 2.11 below:

$$TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{2.9}$$

$$ESS = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \tag{2.10}$$

$$PRSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.11}$$

where N is the number of compounds in the data, y_i is the experimental value of the activity, and $\hat{y}_i$ are the values predicted for each object when it is not in the training set. These quantities are also related to each other and can be written as:

$$TSS = ESS + RSS \tag{2.12}$$

Therefore, equation 2.5 can be modified to be:

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS} \tag{2.13}$$

2.1.2 Principal component analysis

Manipulation of the data may also involve a more complex technique such as principal components analysis (PCA). PCA decreases the dimensionality of a data set and results in a new set of descriptors with more desirable characteristics; it is often found that many of the correlations between some or all of the descriptors can be explained by a small number of these principal components.

Figure 2.2 illustrates the concept of principal components. As can be seen, there is a high correlation between the x_1 and x_2 values. This correlation can be explained by introducing a single variable that is a linear combination of these (i.e. $z = x_1 - x_2$). The new variable (z) is referred to as a principal component. In the case of many descriptors, each principal component is a multi-linear combination of these descriptors:

$$PC_i = c_{i,1}x_1 + c_{i,2}x_2 + \ldots c_{i,n}x_n = \sum_{j=1}^{n} c_{i,j}x_j \tag{2.14}$$

where PC_i is the ith principal component, $c_{i,j}$ is the coefficient of the descriptor x_j, and n is the number of descriptors. For example, PC_1 is the first principal component, which maximises the variance in the given data (Figure 2.2) and shows the greatest spread of these data. The maximum variance in the data that is not already explained by PC_1 will be accounted by the second principal component PC_2 and so on. It is important to point out that all of these components are orthogonal to each other, so PC_2 in 2.2 would define an axis at right angles to the first.

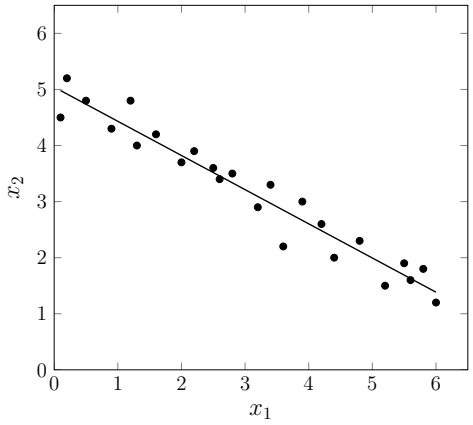


Figure 2.2: A simple two-dimensional illustration of principal components analysis.

2.1.3 Partial Least Squares

Partial Least Squares (PLS) is an expansion of MLR. It is useful when we need to predict a dependent variable from a set of independent variables (descriptors). The idea of PLS is to extract so-called latent variables, a set of orthogonal factors, accounting for as much of the variable variation as possible while modelling the responses. The main aim of PLS regression is to predict $\hat{y}_i$- from x-variables in order to describe their common structure. In PLS regression, the activity variables are often logarithmically transformed and x-variables should be scaled. These variables often have substantially different numerical ranges and it is important that they are scaled appropriately so that each variable has an equal chance of contributing to the overall analysis. Otherwise, descriptors with a large range of variation will bias the results. Auto-scaling is a common way to scale descriptors in which each descriptor value is divided by the standard deviation for that descriptor over all molecules.[16] This method is usually combined with mean centring in which the average value of a variable is subtracted from each individual value. Each scaled x-variable or descriptor then has variance of one:

$$x_i' = \frac{x_i - \bar{x}}{\sigma} \tag{2.15}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2}{N}} \tag{2.16}$$

where x_i' is the new transformed value of the original descriptor x_i, σ is the standard deviation, $\bar{x}$ is the corresponding mean, and N represents the number of data points.

The PLS technique is similar to PCA, with the vital variation that the computed quantities are taken to define not only the variation in the independent (x) variables but also the differences in the dependent ($\hat{y}_i$) variables as well. PLS regression becomes the same as the MLR when the number of latent variables is equal to the number of variables. PLS can be represented with equation 2.17, which explains the dependent variable in terms of latent variables. These latent variables are a combination of the independent variables with a weighting coefficient; the predicted dependent variable, y_i is then written as:

$$\hat{y}_i = c_1 t_1 + c_2 t_2 + ...c_n t_n \tag{2.17}$$

where c_n are coefficients of the latent variables, t_n. t_n is defined as

$$t_n = d_{i_1} x_1 + d_{i_2} x_2 + ...d_{i_n} x_n \tag{2.18}$$

where x_n are the independent variables. PLS has an important feature, which is its ability to deal with data sets with more than one descriptor. Here, two vectors can be constituted by the first component, one in the space of x variables (i.e. t_1 and the other in the y-space (usually written u_1)). Both t_1 and u_1 lines give good approximations to their respective data points but also have maximum correlation with each other (Figure 2.3).

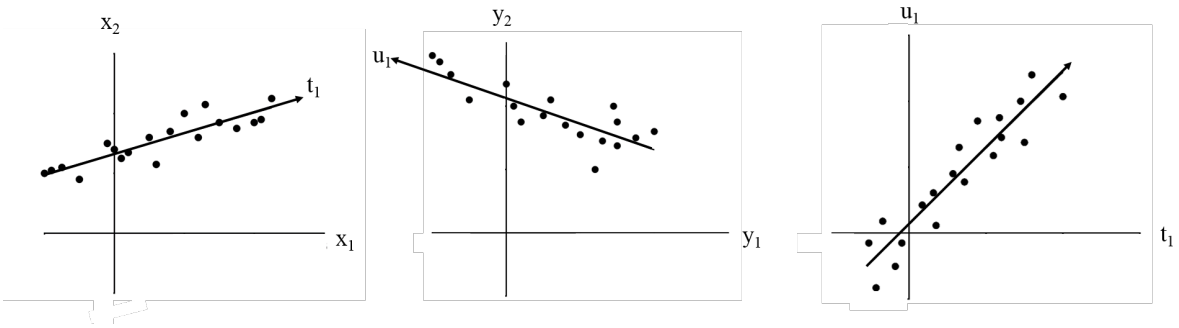


Figure 2.3: PLS model composes of two lines, one in the x-space (t_1, left panel) and one in the y-space (u_1, middle panel). Right panel shows the intercorrelation between t_1 and u_1 lines.

2.1.4 Cluster analysis

In cluster analysis,[16,100] there is no need to know which molecules belong to which clusters. These clusters are usually determined through an analysis of data. This analysis aims to distribute a collection of objects into clusters so that the objects within a cluster are alike but objects taken from different clusters are different. Cluster analysis maximises the similarity of cases within each cluster while maximising the variation between groups that are initially unknown. The clusters based on dissimilarities or distances among objects are determined by a hierarchical cluster analysis. Squared Euclidean or Euclidean distances are common methods to compute distances between objects in a multidimensional space. It begins with each case as a separate cluster and then links the clusters sequentially, decreasing the number of clusters at each level until only one cluster is left. There are several main steps in cluster-based compound selection. First of all, descriptors for each compound in the dataste need to be generated. Secondly, calculating the similarity or distance between between all of molecules. Then, grouping the molecules by using a clustering algorithm. Finally, from each cluster we can select a representative subset (one or more compounds). Therefore, a dendrogram (Figure 2.4) can be produced to show the linkage points: the clusters are connected at rising levels of dissimilarity. The hierarchical cluster-

ing clustered 25 individual molecules based on a complete linkage approach.[16] This method computes the distance between the clusters' most distant elements. The presented hierarchical start at the bottom of the dendrogram and proceed by merging the most similar clusters together in an iterative manner. In the first step, the two closest two molecules are merged into a single cluster. Then, as a next step, the closest two clusters are merged, and so on.



Figure 2.4: Example of a dendrogram or a hierarchical tree. This is based on the calculated ChelpG for 25 benzamide derivatives (The net charge on the ring).

2.2 Basic validation procedures

Cross-validation approaches provide a way to overcome some of the problems inherent in the use of a r^2 value alone. Cross-validation is a statistical method of dividing a sample of data into an initial single subset (training set), while the other subset (test set) retained for later on use to validate the models. Two types of cross-validation are often used in QSAR analysis: Leave-one-out, and k-fold cross-validation.

2.2.1 Leave-One-Out cross-validation

Leave-one-out cross-validation (LOO-CV) is preferred when the number of samples in the data set is limited, and it can be used for the internal validation of statistical models. The predictive ability of a model is estimated by using a reduced set of data. The training set is modified by eliminating one compound from the set each time, and a model is built from the reduced data. The activity of the deleted compound is computed based on the resulting QSAR equation. The process is repeated until all the molecules of the training set have been excluded and predicted once (Figure 2.5). Then, equation 2.19 can be used to calculated the predictivity of the model.

The r^2 belongs to the original model and it measures goodness-of-fit. Cross-validated data produce q^2 (the cross-validated r^2). The q^2 is normally lower than the simple r^2 but is a better guide to the predictive ability of the equation. The difference between r^2 and q^2 is that r^2 is calculated using RSS. However, q^2 is calculated using Predictive Residual Sum of Squares (PRSS), which is analogous to RSS, except the predicted, not the experimental, values are used (see equation 2.11).

$$q^2 = 1 - \frac{PRSS}{\sum(y_i - \bar{y})^2} \tag{2.19}$$

Note that $\bar{y}$ here is different than in equations 2.9 and 2.10. The mean $\bar{y}$ for q^2

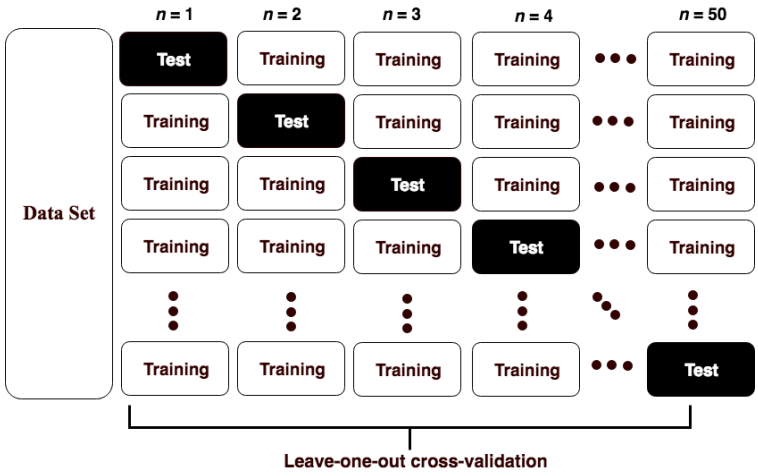


Figure 2.5: Schematic representation of leave-one-out cross-validation for 50 molecules; the first data point is removed and a model trained using the rest of the dataset. This model is used to predict the first data point. Then the second is removed and the process repeated until each data point was deleted once

is the mean for the cross-validated group and not for the whole data set (which is not an accurate representation when cross-validating).[16,101] However, the measures of model fit suggested can be misleading. For example, the r^2 value of 0.80 may sound good, but the real data could reflect several significant outliers, rather than a good correlation. Therefore, more than one single statistic should be used to judge a model. The two common error statistics: Root Mean Squared Error (RMSE, see equation 2.20), and Mean Absolute Error (MAE , equation 2.21) can be used as well to measure error from numeric predictions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{2.20}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{2.21}$$

where y_i is the experimental value of activity of interest, $\hat{y}_i$ is the calculated activity,

N is again the sample size in the model of interest. In the equations obtained during such methods, there should not be much variation for a well-behaved system and the r^2 value calculated should not be significantly larger than the mean q^2.[16] It is likely that the data has been over-fit and the predictivity of the equation will be suspect if there is a large discrepancy.

2.2.2 k-fold cross-validation

Another common approach to use in cross-validation is known as the k-fold cross-validation. The fold means a number of iterations. In this method, the original dataset is divided into k subsamples. A single subsample is only held as the validation data for testing the QSAR model and the remaining $n - 1$ subsamples are used as a training set. The cross-validation is repeated k times, with each of the k subsamples used exactly once as the validation data. For example, in a 3-fold cross-validation 1 fold is left out as the testing set while the remaining 2 folds are used as the training set for the QSAR model construction and then validated with the fold that was excluded (Figure 2.6).

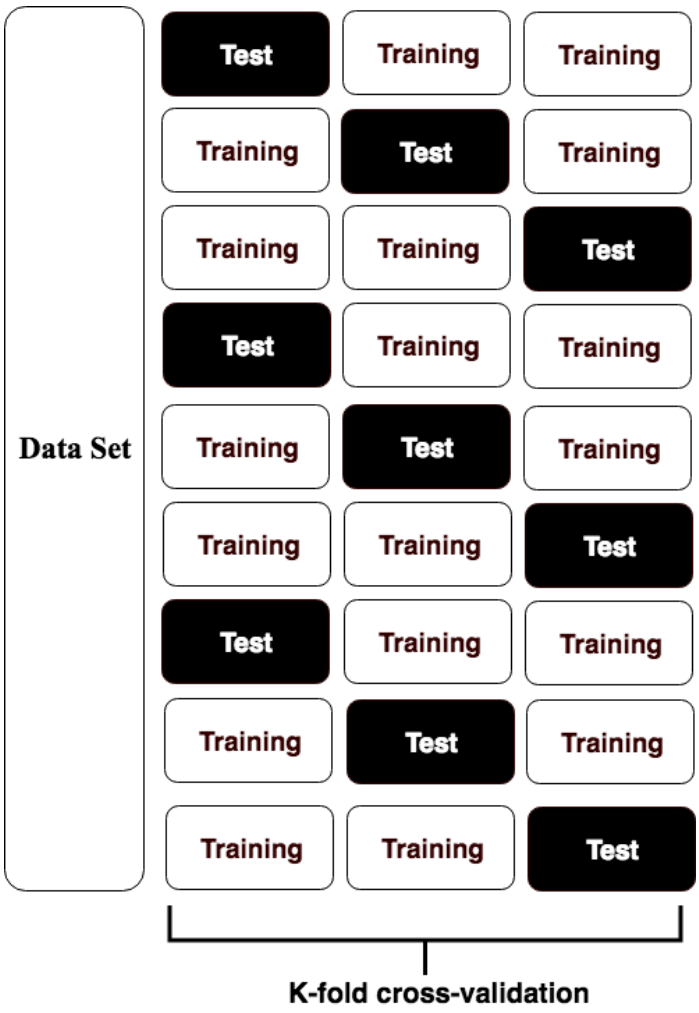


Figure 2.6: Schematic representation of k-fold cross-validation; the dataset is split into 3 ubsets of equal size. Each of the subset is removed once, a model built with the reminder and the subset predicted with this model.

2.2.3 Y-randomisation

Y-randomization[102] is a tool used in validation of QSAR models. The y-Randomisation is also called y-scrambling[100] and is a powerful validation procedure. For this, we built a number of independent models using the original descriptor pool and shuffled the activity values randomly. This established approach was used to evaluate the risk of chance correlation in the original model.[103] To get an acceptable QSAR model, the average correlation coefficient of randomised models should be less than the correla-

tion coefficient of the original non-randomised model.

Based on datasets taken from literature, Rücker and co-workers[103] have compared several alternatives of y-randomisation-including original response, permuted response and random pseudo-response together with original descriptors and replacing original descriptors with random pseudo-descriptors from MLR models. They found that a high r^2 value with random or permuted responses will arise when replacing the original descriptors with random (pseudo-descriptors). This was ascribed to the intercorrelation in the original descriptors that reduces the effective dimensionality of the descriptors.[104] In order to give confidence in the significance of the model, Rücker and co-workers suggest that a model r^2 should exceed both response and random pseudo-descriptor r^2. Other validation methods, such as external validation, can be used beside y-scrambling to give more confidence in the model predictivity. This validation approach will be presented in the next section.

2.2.4 External validation

The predictive power of the model can also be determined by using an external validation. This validation test requires one split of the complete data set into an external test set and a training set. The size of external set can be about 15-20% of the entire data set.[105] The outcome of a QSAR study, unfortunately, may depend on the way the dataset is split into the external test and the training set; this can cause problems usually for small datasets. Therefore, several methods have been put forward to create the external validation set, of which, random selection is one obvious method.

Many QSAR models cannot be considered reliable, due to the absence of any external validation.[106] Other reasons behind the failure of models are inappropriate use of models, procedures used to develop models, or false expectations.[107] Tropsha and co-authors[100,106,108] have demonstrated that several models with high q^2 values have poor

predictive power when applied for prediction of molecules in the external validation. They pointed out the importance of the external validation in model development and presented several statistic criteria of prediction that should be used for the validation of multiple QSAR models.

2.3 Data mining methods

2.3.1 Neural Networks

Neural Networks (NN) are widely used and applied to QSARs and other chemometric problems, which lead to a lot of successful applications. NN consist of a set of layered nodes inter-connected with each other analogous to the human brain. They often are composed of an input layer, one node per descriptor, one or more hidden layers, and an output layer (see Figure 2.7). The data is transformed to non-linear data using non-linear functions that operate between the nodes of the input layer and the hidden layers. The neural networks usually run iteratively using a distribution of optimised weighting parameters to link the input nodes to the hidden and output nodes. The correlation between molecular descriptors and activity can be then derived based on these optimised parameters.

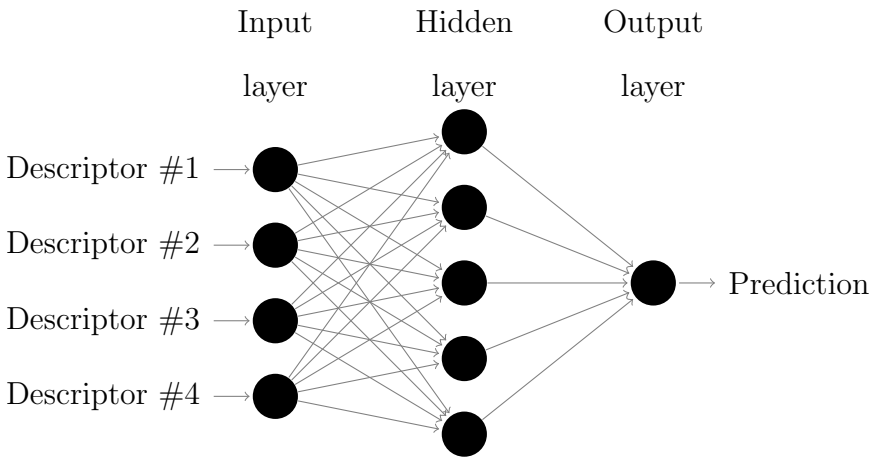


Figure 2.7: Neural network diagram

A supervised neural network is first trained; this is achieved by providing it with a set of inputs and the corresponding outputs from a training set.[16] For example, each node in the input layer may correspond to one of the descriptors that are used to characterise each compound. In the neural network initially, the weights and other parameters are set to random values. The weights and parameters, which govern the status of the nodes and the connections between them, are modified as a key step in the training phase. The minimisation of an error function that defined from the difference between the desired and obtained values is used to supervise the training of the network. After training of the neural network, it can be used to predict the values (activities) of new molecules.

A possible problem with neural networks is over-training. This problem leads to excellent results on training data but will have poor predictive capability on a previously unseen data set. This occurs because the network "memorises" the training data. To overcome such a problem, the training data needs to be split into two; one portion to be used for training as normal, and the other to assess the performance of the network during the training stage. Another challenge with the NN is that it is not feasible to ascertain why the network produces a particular result for a given input. This is due to the complex nature of the interconnections between the nodes; one cannot simply decide which properties are relevant.

2.3.2 Support vector machine

Support vector machine (SVM) is a promising classification and regression approach developed by Vapnik et al.[45] This powerful tool has gained much interest in bioinformatics[109–111] and chemistry.[112–116] In bioinformatics, SVMs have been used to solve classification and correlation challenges, such as breast cancer diagnosis,[117,118] lung cancer,[119] bioactivity of hepatitis C virus (HCV),[120] etc. SVMs have also been employed in chemistry, for example, the prediction of inhibitors based on activity cliffs

(ACs) and structure-activity relationships (SARs).[121] Simple implementation, good generalisation capability, and dimensional independence make SVMs superior when compared with traditional regression and NN.[46] SVMs are also characterised by flexibility in classification and their ability to approximate continuous function. Therefore, the SVM method is suitable and commonly used in many SAR and QSAR studies.

In the SVMs, a hyperplane is created to separate two regions of the training data (Figure 2.8). This hyperplane could be a line in a 2D-space. In three-dimensional space, it will be a 2D-plane in higher dimensions but will be hard to conceptualise.

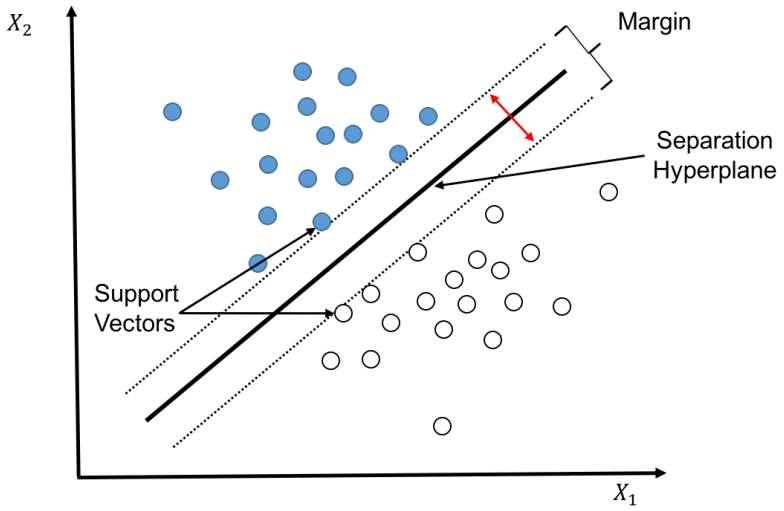


Figure 2.8: A two-dimensional plot for the support vector method. The separating hyperplane is shown as a line. The perpendicular length between the separating hyperplane and a hyperplane through the closest data points (the support vector) is so-called margin.

SVM method can be used for either classification or regression, and characterise its ability to model complex non-linear relationships by selecting a suitable kernel functions. The SVM uses a kernel function to map the data points into a higher dimensional feature space, and then an optimal hyperplane separating the two classes is constructed. The kernel functions used in the SVM can be linear, polynomial, radial basis function (RBF), and sigmoid function[122] (Figure 2.9).

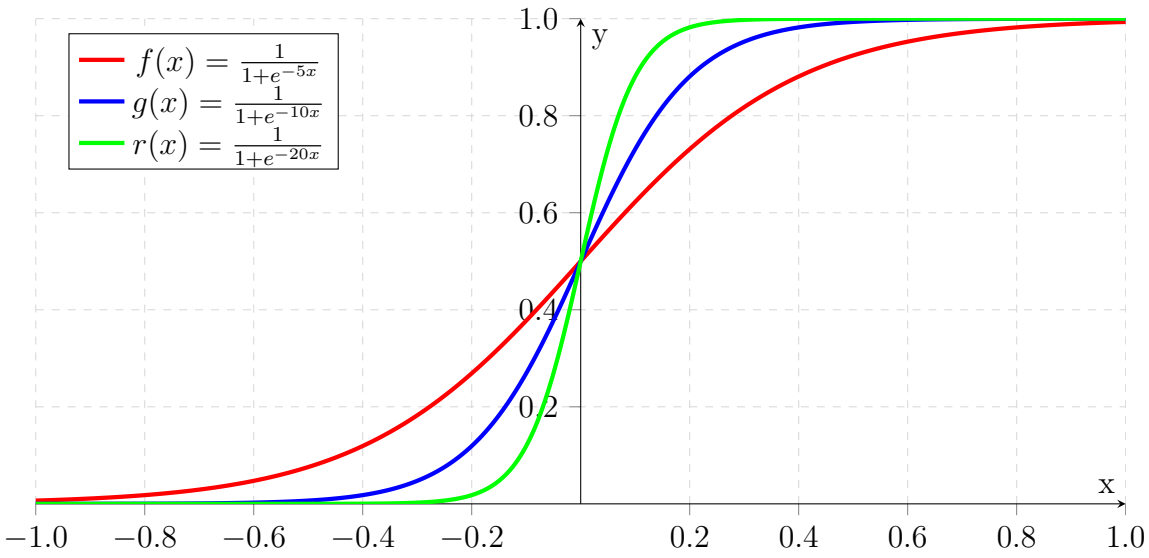


Figure 2.9: Examples of simple sigmoid functions

The linear model $f(x, \omega)$ in the feature space is given by

$$y = f(x, \omega) = \sum_{n=i}^{j} \omega_i g_i(x) + b \tag{2.22}$$

where $g_i(x)$, $i = 1, \ldots, j$ corresponds to a set of non-linear transformations, and b is the "bias" term. The coefficients ω and b are estimated by minimising

$$\mathrm{R}_{emp}(\omega) = \frac{1}{n} \sum_{i=1}^{n} L_\epsilon(y, f(x_i)) \tag{2.23}$$

where $\mathrm{R}_{emp}(\omega)$ is the empirical risk, $L_\epsilon(y, f(x_i))$ is ϵ-insensitive loss function that was proposed by Vapnik:[45]

$$L_\epsilon(y, f(x_i)) = \begin{cases} 0 & if \quad |y - f(x, \omega)| \leq \epsilon \\ |y - f(x, \omega)| - \epsilon & otherwise \end{cases} \tag{2.24}$$

2.3.2.1 Sequential Minimal Optimisation

SMO[123–125] implements a fast version of Platt's sequential minimal optimisation al-

gorithm for training a support vector regression model. Two kernel functions were employed, respectively:

1. SMO(1): puk (Pearson VII function-based universal kernel): The Pearson VII function was developed by Pearson[126] and is often used for curve fitting of X-ray diffraction scans.[127] This function is likewise the polynomial and RBF functions all belong to the class of valid kernel functions. A valid kernel function satisfies the so-called Mercer's theorem[128] in which a function must be symmetric, hence:

$$K(x,z) = \langle \phi(x).\phi(z) \rangle = \langle \phi(z).\phi(x) \rangle = K(z,x) \tag{2.25}$$

   where K is a kernel function of two arguments x and z (real numbers or vectors), and ϕ is a vector representing a non-linear mapping function.

2. SMO(2): NormalisedPolyKernel.

2.3.3 Decision trees

A decision tree is usually represented as a tree-like structure, with each node corresponding to a particular rule. Each rule may correspond to the lack or presence of a specific property (descriptor). An example of a decision tree is presented in Figure 2.10; this was one of several produced models in this study where the molecular descriptors were generated using density functional theory. The decision tree shown in Figure 2.10 is for some molecules, some of which possess inhibition properties relevant to fibrosis. This example is of particular interest because the active molecules are enclosed by less active ones in the descriptor space; such a condition can cause problems for regression-based approaches. The decision tree begins from the top root-node and follows the edge appropriate to the first rule. This proceeds until a terminal node is reached, at which point one can set the compound into the proper activity class.

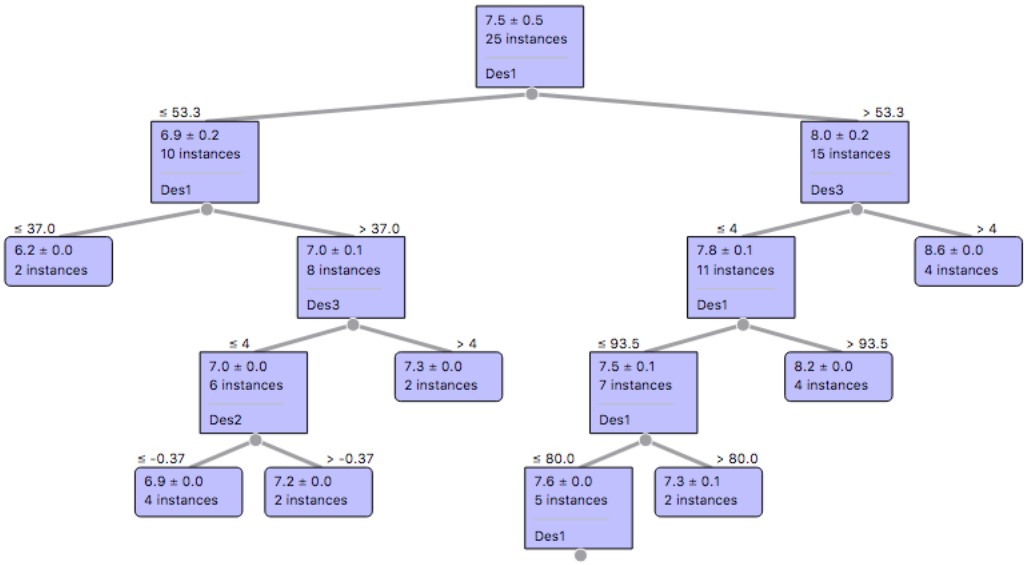


Figure 2.10: Example of a decision tree for a set of 25 molecules. In order to classify an unknown compound a path is followed through the tree according to the values of the relevant properties, until a terminal node is reached.

In the given example from Figure 2.10, Des1, Des2, and Des3 represent the first, second, and third descriptors, respectively. Des1 varies from 0 to 100. Des2 includes numbers between (-1 and +1). The last variable (Des3) will consider integers; 1 to 6. These descriptors (independent variables) were correlated to activity. Several approaches are available to build a decision tree. Most of them start with the entire data set and classify the descriptor that gives the best split. Thus, the data set will be split into two or more subsets. The same procedure is repeated to each of these subsets, and so on until all of the molecules have been adequately divided into their different classes. If there are $N = 25$ molecules divided into C_8 classes, each class C_i containing n_i compounds, then the information corresponding to this distribution can be written as:

$$I = \sum p_i ln p_i \tag{2.26}$$

where I corresponds to the entropy of the system and $p_i = \frac{n_i}{N}$. This result is employed to determine which molecular descriptor to use to construct the next rule. Thereby, reducing the number of molecules that needed to be screened is an important advantage for the decision tree when screening a vast data set from a combinational library. It was found that models constructed from the screening of a large data set can identify a significant part of the active compounds in the dataset.[16] Some decision tree approaches only consider multiple branches at each nodes; others can consider binary splits. Some methods require the descriptors either to be distinct categories (e. g. Des1 < 93.5; Des1 ≥ 93.5) or to be binary (e. g. indicating the presence or absence of a substructure). Other methods can use both. However, such methods may build very complicated trees. They also may produce an excellent fit for the training set but whose performance on unseen, test data might be less powerful.

2.4 Molecular descriptors

Physicochemical properties of a molecule such as geometrical, electronic, constitutional, hydrophobic, solubility, steric, and quantum chemical can be determined and used as so called molecular descriptors.[129–132] These molecular descriptors are numerical values which transform chemical information encoded within the molecular structures into a representation. Molecular descriptors vary in the complexity of the information they encode and in the time required to calculate them. For example, the molecular weight and number of H-bond acceptors do not convey much about molecule's properties but they are very rapid to compute. On the other hand, descriptors that are based on quantum mechanics or electronic structure may provide accurate representations of properties, but they are very time consuming to compute. There are commonly used molecular descriptors that can be applied to relatively large data sets. For example, Gasteiger partial charges (E)[133] to represent electro-

statics; Crippen-Wildman log$P$ parameters,[134] representing hydrophobicity; Crippen-Wildman molar refractivity (MR) parameters,[134] to represent steric properties and polarisability; and log$S$ parametres introduced by Xu and coworkers,[135] to represent solubility. There are also other descriptors that can be calculated from the 2D graph representation such as topological distances. In the next part, we will include some descriptors that are properties of the whole molecule rather than of substituents. We will briefly cover some molecular descriptors that were used in the development of QSAR models in Chapters 3 and 4.

2.4.1 Physicochemical properties

The activity of drugs can be influenced by hydrophobicity. This property shows how tightly a molecule binds to a protein and its ability to pass through a cell membrane. Hydrophobicity can be modelled using the logarithm of the partition coefficient P between n-octanol and water (log$P$). The experimental determination of lipophilicity can be difficult, especially for very lipophilic or polar molecules.[16] Therefore, there is a huge interest in the development of different methods for predicting log$P$ values. Many methods for estimating the log$P$ have been proposed. One method for estimating log$P$ was a fragment-based method.[136] This method is based on breaking molecules into fragments and then summing the substructure contributions to give the final log$P$. This method can be represented using the following equation:

$$\log P = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j \tag{2.27}$$

where a_i represents fragments of type i with f_i being the corresponding contribution and b_i occurrences of correction factor j with F being the correction factor. n and m correspond to the molecule fragments. There are many free software packages and open source code that can be used to calculate this type of descriptor such as

Chembench,[137] ChemAxon,[138] Datawarrior[139] and so on.

An alternative method to the fragment-based approach is an atom-based approach. In this method, log$P$ is calculated by summing contributions from individual atoms (equation 2.28).

$$\log P = \sum n_i a_i \tag{2.28}$$

where n_i is the number of atoms of atom type i and a_i is the atomic contribution. The drawback of this method is that a large number of atom types is needed to describe a variety of molecules. For instance, carbon, hydrogen, nitrogen, oxygen, sulphur and halogens in the Ghose and Crippen method,[140] are classified into 110 atom types based on their hybridisation state and chemical nature of the neighbouring atoms. Other physicochemical properties (e. g. solubility log$S$) have been developed using fragment- and/or atom- based methods as described above.

2.4.2 Molar refractivity

The steric bulk of a molecule is often measured by a molar refractivity (MR) and is given by:

$$\mathrm{MR} = \frac{n^2 - 1}{n^2 + 2}\frac{M_W}{d} \tag{2.29}$$

Here n is the refractive index, d is the density, and M_W is the molecular weight. The refractive index term represents polarisability of the molecule. In a similar way to the calculation of log$P$, MR can be calculated using the atomic contributions.

2.4.3 Topological indices

Topological indices (T_i) are 2D molecular descriptors that are calculated from the 2D graph representation of a molecule (usually excluding hydrogen atoms).[141–143] These

indices represent the overall shape of the molecule, its size, and degree of branching. As shown in Figure 2.11, the graphs are 2D. Atoms are termed 'vertices' and bonds are termed 'edges'.

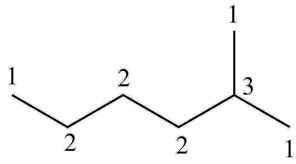


Figure 2.11: Hydrogen-suppressed graph of 2-methylhexane, with labelled vertices

To calculate a T_i, the values of adjacent vertices are multiplied, and then summed across all edges. So, for the 2-methylhexane molecule shown in Figure 2.11, a simple T_i would be $1\times2+2\times2+2\times2+2\times3+3\times1+3\times1$, which is 22. Some T_i take into account of double and triple bonds, weighted and multi-graph terms according to the relative importance of each to the topological indicies. Many T_i have proved useful in QSAR modeling.[144–147] These 2D descriptors are easily calculated and fast. They also have been shown to correlate well with numerous biological activities, suggesting that they are information-rich, and thus are qualified for use as descriptors in QSAR modelling.

2.4.4 Topological maximum cross correlation

Topological maximum cross correlation (TMACC) descriptors[148] are 2D descriptors, which can be used to generate an interpretable QSAR. The TMACC descriptors were inspired by the grid-independent descriptors (GRIND).[149] The TMACC descriptors do not require 3D coordinates; they are based only on properties derivable from atom types and 2D connection tables. They differ in two ways. Atomic physicochemical values replace force field interactions measured from a grid, and the 2D topological bond distances take the place of 3D distance between points. HQSAR[150] was used previously[148] to validate the TMACC descriptors. It was found the TMACC

method was comparable to the 3D QSAR techniques. The GRIND works similarly as the autocorrelation descriptor.[151] The TMACC only considers the maximum value calculated for any given distance across the molecule.

The atoms contribute to each molecular descriptors and are detailed in FGRAM file. For instance, the descriptor negative log$S$-negative log$S$ with a bond length of two has a value of 0.2. Coincidentally with the PLS coefficients, the FGRAM permits an activity contribution quantity to be designated to each atom. Beforehand, TMACC and FGRAM were generated using a javascript. To evaluate the interpretative capability of the approach, more time and effort need to be considered.

To overcome these difficulties, a GUI software called Nottingham Chemoinformatics Workbench (NCW) was implemented. NCW also has ChemAxon's Marvin[152] embedded, which gives the user the ability to draw and edit molecules. Models can be generated using the NCW through Weka. The descriptors can be interpreted as well using the PLS model. The final results can be visualised with a colour coding scheme in Marvin.

2.5 Comparative Molecular Field Analysis

One of the most significant developments in QSAR was the introduction of Comparative Molecular Field Analysis (usually abbreviated to CoMFA)[153] and Comparative Molecular Similarity Indices Analysis CoMSIA.[151] The aim of these methods are to derive a correlation between the activity of a series of molecules and their 3D shape, electrostatic and hydrogen bonds. A 3D molecular alignment is necessary step to start any CoMFA analysis and the predictive capability is directly dependent on the 3D molecular alignment of molecules with a reference compound. Therefore, one of the important steps in performing the 3D-QSAR analyses is how to select the template molecule. In general, the most active compound is chosen as a template molecule,

and remaining molecules are aligned to that template. The aligned molecule, which has high biological activity, is placed in a three-dimensional grid with the grid spacing and resolution of 2.0, and 1.0 Å, respectively. This grid is selected to generate points at the intersections of a 3D lattice. A few to several thousand points are generated based on the dimensions of the box and the chosen grid distance. Molecular properties were calculated on the regular grid with a grid resolution of 1.0 Å (Figure 2.12).

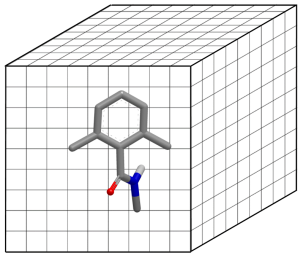


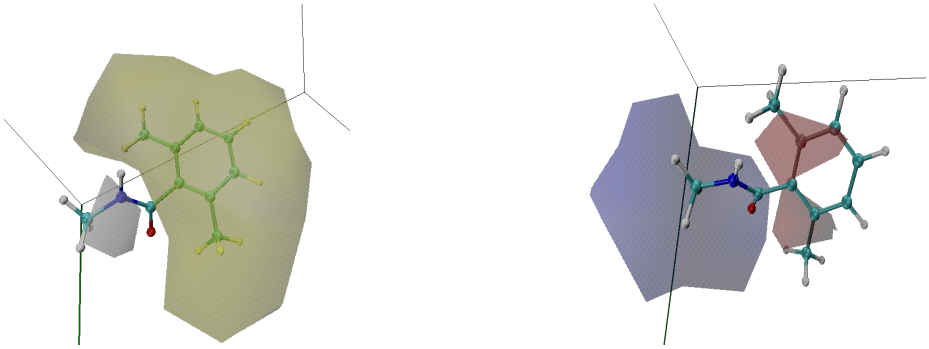
Figure 2.12: A template molecule in a 3D grid box. For presentation, only the grid lines at the surface of the box and only one molecule are shown instead of a superposition of all molecules. This grid includes $9 \times 9 \times 9 = 729$ grid points

The structure is placed within a regular lattice and the interaction energy is calculated between the molecule and a series of probe groups placed at each grid point. The electrostatic and steric energies are the standard fields that are used in CoMFA. The electrostatic potential is calculated using Coulomb's law and the steric energy using the Lennard-Jones potential. This process is similar to the GRIND method for use in structure-based design. The calculated energies from these fields are placed into a data matrix. Each column in the data matrix corresponds to the interaction energy of a specific probe at a particular grid point and each row in the matrix corresponds to a single molecule. An atomic probe, e.g., a carbon atom, is used to calculate field values in each grid point. If we have M grid points and N probes then we end up with $M \times N$ such columns. The activities can be added as a

final column. Then, PLS is used to analyse the matrix as seen in equation 2.30

$$\hat{y}_i = \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} E_{ij} + C \tag{2.30}$$

where $\hat{y}_i$ is the predicted activity, and each coefficient c_{ij} corresponds to placing probe j at grid point i with E_{ij} being the corresponding energy value. C is a constant. The significance of placing probe group at the grid points are determined by these coefficients. Later on, the data can be interpolated using a 3D contour plot that identifies regions of interest. For example, one may be able to identify regions where it would be favourable to place positively charged substituents or where increases in hydrophobicity or solubility would decrease the pIC_{50}. This is illustrated in Figure 2.13.



(a) log$P$ MAP (b) log$S$ MAP

Figure 2.13: CoMFA contour plots for 4-chloro-*N*-methyl-3-(trifluoromethyl)benzamide. For log$P$, yellow indicates regions where it is favourable to place a positive charge and white where it would be unfavourable. For log$S$, blue indicates regions where it would be favourable to place a positive charge group and red where it would be unfavourable.

2.6 Conclusion

In this chapter, we outlined the research methodology that has been used in this thesis. It reviewed the methods taken to conduct the research, starting with an overview of statistical techniques such as the MLR, and how to assess the quality

methods as a whole using a number of statistical parameters such as determination coefficient, F value, SEE, TSS, RMSE, and PRESS. We also described several validation procedures that can be used when building QSAR models using the 2D- and 3D-molecular descriptors. For instance, the validation was started from LOO-CV, k-fold cross-validation, external validation into the Y-randomisation methods. Then, we provided a brief description of a wide variety of machine learning approaches. We explained how such methods are helpful to extract non-linear correlations between the molecular descriptors and the biological activity.

Chapter 3

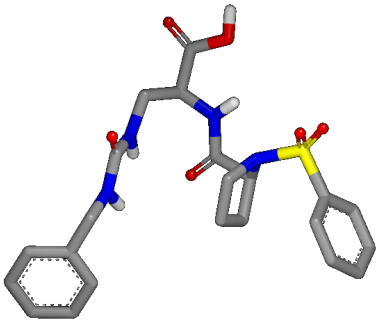
QSAR for Inhibitors of $\alpha_v\beta_1$ integrin

3.1 Introduction

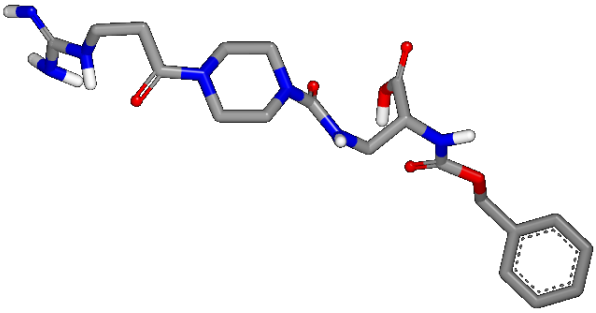
The $\alpha_v\beta_1$ integrin was identified more than two decades ago by combining the human α_v integrin with Chinese Hamster Ovary CHO cell β_1.[154] This integrin was reported to be a fibronectin receptor[76] and was suggested to mediate profileration, migration, adhesion, and invasion of ovarian cancer cells.[155] The $\alpha_v\beta_1$ integrin, Figure 3.1c, plays a critical role in tissue fibrosis and in mediating cell adhesion to fixed extracellular ligands and in the maintenance of tissue integrity.[156] The $\alpha_v\beta_1$ integrin binds to the latency-associated peptide of TGF-β_1 and mediates TGF-β_1 activation.[156]

Four α_v-containing integrins, $\alpha_v\beta_1$, $\alpha_v\beta_3$, $\alpha_v\beta_5$, and $\alpha_v\beta_8$ are expressed in tissue fibroblasts.[156] Previous studies showed that several α_v-containing integrins, including $\alpha_v\beta_1$, $\alpha_v\beta_3$, $\alpha_v\beta_5$, $\alpha_v\beta_6$, and $\alpha_v\beta_8$ can recognise the same arginine-glycine-aspartic acid (RGD) sequence, and can activate latent-TGF-β.[157–160] *In vitro*, deletion of the integrin α_v subunit was used to activate fibroblasts to demonstrate that loss of all α_v integrins from fibrosis in multiple mice organs was associated with reduced tissue

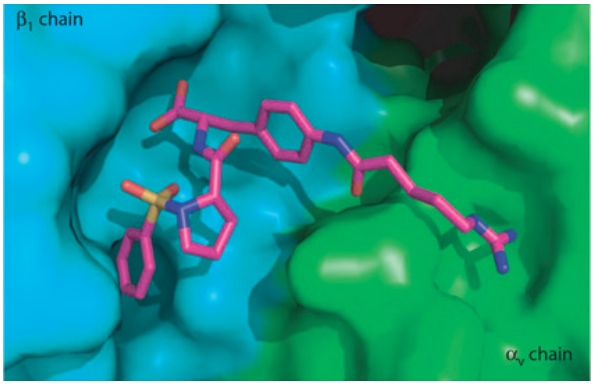
TGF-β signaling.[161] It was found that deletion of $\alpha_v\beta_3$, $\alpha_v\beta_5$, and $\alpha_v\beta_8$ integrins has no effect on organ fibrosis.[156] However, the contributions of $\alpha_v\beta_1$ to fibrosis are still not well understood.



(a) α_v-binding base compound



(b) β_1-binding sulfonamidoproline compound



(c) $\alpha_v\beta_1$ (Adopted from Ref.[156])

Figure 3.1: Docked model of $\alpha_v\beta_1$ integrin, where α and β subunits are, respectively, in green and blue.

The $\alpha_v\beta_1$ integrin is composed of α and β subunits that are both present in multiple heterodimers (5 in α_v and 12 in the case of β_1).[156] No crystal structure is available for the $\alpha_v\beta_1$ integrin. To model the structure of the complex, the crystal structures of other α_v and β integrins[162,163] were used. The model was used to develop a specific small-molecule inhibitor of $\alpha_v\beta_1$.[164] Previous studies[156,165] designed an $\alpha_v\beta_1$ inhibitor (Figure 3.1c) by combining a positively charged guanidine moiety from an $\alpha_v\beta_3$ integrin inhibitor (blue shading Figure 3.1b) and a sulfonamidoproline

moiety from an $\alpha_2\beta_1$ integrin inhibitor (green shading Figure 3.1a). These studies also suggest that the phenyl sulfonyl ring is situated to allow good shape complementarity. However, substitution on the phenyl ring might cause steric congestion in the pocket in the β_1 chain. Other studies[166–168] investigated how structural modification of the *N*-arylsulfonyl-L-prolyl group would influence the potency against $\alpha_v\beta_1$ and other β_1 integrins ($\alpha_2\beta_1$ and $\alpha_4\beta_1$).[166] It also was shown that a group of prolyl or aryl analogues in the same *N*-arylsulfonyl-L-prolyl structure were potent non-RGD β_1 integrin inhibitors.[166]

Bleomycin-induced plumonary fibrosis can be attenuated by the $\alpha_v\beta_1$ integrin.[156] GlaxoSmithKline (GSK) and the University of Nottingham have synthesised and assayed a series of novel compounds for activity against α_v integrins including $\alpha_v\beta_1$. This suggests that a structure-activity relationship for inhibitors of $\alpha_v\beta_1$ could be useful for developing new compounds towards a treatment of IPF. There have not been any QSAR studies of $\alpha_v\beta_1$ inhibitors. In this chapter, a 2D-QSAR has been derived for a series of benzamides as $\alpha_v\beta_1$ inhibitors. In addition, a 3D-QSAR has been established. Contour maps of properties were used to explain the impact of physicochemical and structural properties of compounds on $\alpha_v\beta_1$ inhibition.

3.1.1 Data sets

The aim of our QSAR studies is to generate a predictive model and to use it to screen (*in silico*) molecules that might be subsequently synthesised. The data set contained a range of 40 benzamide derivatives and experimental activities from GSK with potential experimental errors of ± 0.2 approximately. Models were built to relate a set of physico-chemical properties to the biological activity (pIC_{50}). We began by considering 27 of these molecules either as fragments (data set G1) or whole molecules (data set G2). The study was then extended to all 40 compounds, again as fragments (data set G3) or whole structures (data set G4). All data sets are shown in

Appendix Table A.1. An example of the molecules in G1 is 2-(5-(2-carboxy-2-(*N*,2,6-trimethylbenzamido)ethyl)amino)-5-oxopentyl)-5,6,7,8-tetrahydro-1,8-naphthyridin-1-ium (Figure 3.2a), and the corresponding fragment *N*,2,6-trimethylbenzamide (Figure 3.2b).

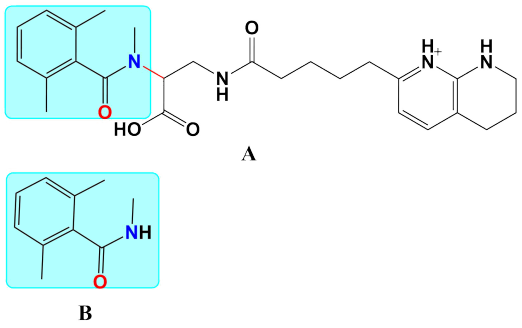


Figure 3.2: An example of (A) the whole structure (2-(5-(2-carboxy-2-(*N*,2,6-trimethylbenzamido)ethyl)amino)-5-oxopentyl)-5,6,7,8-tetrahydro-1,8-naphthyridin-1-ium) (B) the corresponding fragment (*N*,2,6-trimethylbenzamide).

3.2 Methods

3.2.1 2D-QSAR

To study the structure-activity relationships, we divided the benzamide structures into four regions: naphthyridine rings, the amide group, the carboxylic group, and the benzamide group. In this study, the substituents at the 2-, 3-, and 4-positions of the phenyl ring of the benzamide moiety were varied. The phenyl ring was then replaced by other rings. The other three regions: naphthyridine rings, the amide and the carboxylic groups were kept unchanged. Two-dimensional QSAR methods, described in chapter 2, were employed. TMACC descriptors were generated using the topological data of each molecule. All nonpolar hydrogen atoms were removed using Obabel 2.3.1[169] and polar hydrogen atoms were considered explicitly. Physicochemical properties were assigned to each atom. Four parameters were used (chapter 2):

Gasteiger partial charges E, log$S$, log$P$, and molar refractivity (MR). The number of models tested was the maximum available, a total of 15. The descriptors were calculated from 15 combinations of atomic properties (Table 3.1).

Combination Number	Combination type	Combination Number	Combination type
1	E	9	P and MR
2	P	10	S and MR
3	S	11	E, P, and S
4	MR	12	E, P, and MR
5	E and P	13	E, S, and MR
6	E and S	14	P, S, and MR
7	E and MR	15	E, P, S, and MR
8	P and S		

Table 3.1: Combinations of atomic properties. E = charge, P = log$P$, S = log$S$, and MR = Molar Refractivity.

Positive and negative values of properties were considered separately for each atomic property, but MR, as all atomic values for this property are only positive. The evaluated atomic refractivity should always be positive since the molar refractivity is related to the molar volume.[170] This gives seven atomic properties: positive charge, negative charge, positive log$S$, negative log$S$, positive log$P$, negative log$P$ and positive molar refractivity. Based on these seven properties, 28 possible combinations of pairs are possible. All 28 descriptors were calculated for each pair of atoms in the molecule. The different scales used by each set of atomic parameters were taken into account, and each contribution was scaled by the largest absolute value (positive and negative values took maximum values of +1 and -1, respectively). The descriptor was encoded with the topological bond distance. Therefore, the total number of descriptors is the maximum topological bond distance multiplied by 28. For a small molecule, e.g. molecule number 6 in the G1 set, with a maximum topological bond distance of nine, would give one block of descriptors for a given atomic property and given bond separation (Table 3.2). The descriptor positive charge-positive log$S$ is shown in Table 3.2, with nine potential bond distances. The total number of descriptors

for the shown molecule would be 252; that is 9 maximum topological bond distances multiplied by 28 TMACC descriptors.

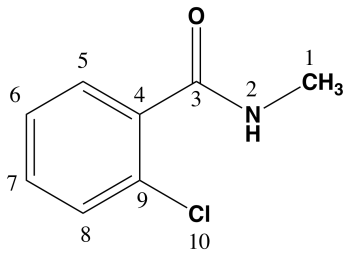
Structure	One block of atomic properties
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:1
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:2
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:3
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:4
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:5
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:6
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:7
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:8
	Maximum:ScaledAtomPartialPositiveCharge:ScaledAtomPartialPositiveLogS:9

Table 3.2: Molecule number 10 in the G1 list. TMACC descriptors are based on topological distances. The maximum distance here is 9 between the methyl carbon and Cl.

In-house Java-based software, the Nottingham Cheminformatics Workbench (NCW),[101] was used to perform all calculations. NCW is a package which generates the TMACC descriptors and interpretation. The PLS algorithm was used, using the SIMPLS algorithm of de Jong,[171] to find latent variables and construct the model. This algorithm is implemented in NCW and users can also access another machine-learning workbench Weka 3.9.1[172] from NCW.

3.2.2 Machine learning methods

Different machine learning algorithms implemented in Weka 3.9.1 were applied. A brief description for these methods was given in chapter 2, and more details on these algorithms are available elsewhere.[173] All these calculations are based on the 2D-QSAR descriptors using the G4 data set.

3.2.2.1 Sequential Minimal Optimisation

SMO implements a sequential minimal optimisation algorithm for training a support vector regression model. The Pearson VII function-based universal kernel (puk) and

the NormalisedPolyKernel functions were employed as SMO(1) and SMO(2), respectively. These methods were described in chapter 2.

3.2.2.2 Support vector machine

SVMs are support vector machines for regression. By transforming the 2D-QSAR data into a high dimensional feature space, nonlinear relationships can be shown in a linear form. This transformation is accomplished by using specific kernel functions. The values of parameters were set as follows: the SVM parameter C (C is a cost or complexity parameter = 50), kernel = RBFkernel $K(x_i, x_j) = \exp(-\gamma \left\|x_i - x_j\right\|^2)$. γ is a hyper parameter (also called the kernel bandwidth). In the SVM regression, SVM works by finding a line of best fit that minimises the error of a cost function. Molecules in the training dataset (support vectors) that are closest to the line of a best fit are only considered during an optimisation process. Often a margin is added around the line and its width is varied in order to improve the prediction accuracy.

3.2.2.3 Multilayer Perceptron

A feed-forward artificial neural network, a Multilayer Perceptron (MLP) regressor, was used to distinguish molecules. The MLP algorithm was inspired by a model of biological neural networks in the brain where neurones are organised into multiple layers. Well configured layers are able to approximate any function. In regression, our interest is to approximate a function that best fits the real output. A single output from multiple real-valued inputs is computed first by a perceptron. This forms a linear combination with the inputs and then predicts the output through a nonlinear decision surface. Sigmoid nodes were chosen in this network except for the output node. The back-propagation algorithm[174] was used to optimise the MLP. There are many MLP parameters that can only be tuned effectively through intuition. This makes the neural network a complex algorithm to predict a model. Therefore, the

Weka MLP options were set to the default as recommended.

3.2.2.4 IBk

IBk[175] is a k-nearest neighbours algorithm which supports both regression and classification. The IBk uses the brute force search algorithm for nearest neighbour search. Finding neighbours is performed by using a Euclidean distance function. An Euclidean distance matrix of all molecules in the training set is scanned for the k shortest distances to determine the nearest neighbours. The k parameter controls the size of the neighbourhood. For instance, if $k = 1$, then the single most similar training molecule is used to make the prediction. In Weka, the optimal k value can be calculated automatically by setting the crossValidate parameter to True. Thus, the best (k) value was five for the G4 data set. It is helpful in regression problems for k value to be an odd number as this decreases the chance of tied votes.[176] The other IBk parameters were set to the default.

3.2.2.5 Regression trees

M5P is used for generating M5 model trees. It was invented by Quinlan[177] and was improved by Wang and Witten.[178] The default M5P parameters were used and LOO-CV was carried out to validate the results. To generate a tree, molecules were partitioned from a parent node into two child nodes. A specific rule was followed to produce each node. This rule forces molecules below a certain descriptor's value to settle into one of the child nodes. The other molecules above fall into the other child node. M5 trees are analogous to piecewise linear functions, and have multivariate linear models at their leaves.[179] The Weka function (REPTree), a fast decision tree learner, is also used which generates a decision/regression tree using information increases and prunes it using reduced-error pruning.[180]

Pruning is one method of countering overfitting in decision tree, where subtrees

within the decision tree are replaced by leaf nodes. The size of the decision trees is reduced by removing parts that do not provide information relevant to the prediction of new molecules. REPTree is a good example of pruning. Reduced error pruning is performed by comparing the number of errors made by a non-leaf subtree to the error of the optimal leaf that could replace it. If the best leaf has fewer or an equal number of errors, the subtree is pruned and replaced. This is repeated for every non-leaf subtree in the tree until no more replacements can be done without increasing the error of the tree.[180] The default parameters were used: the seed used for randomising the data = 1, the minimum weight of the instances in a leaf (minNum) = 2, the amount of data used for pruning (numFolds) = 3 in which one fold is used for pruning and the rest for growing the rules.

3.2.2.6 Random Forest

Random Forest (RF) is an approach for classification and regression which was introduced by Breiman.[44] RF was used to construct a forest of random trees, and is an example of an ensemble method. The ensemble approach produces a weighted collection of decision trees. Often random vectors are generated; $p_1, p_2, \ldots p_n$ in order to grow these ensemble trees. By using a random sample of the data, random forests train each tree independently. This randomness makes the model less likely to overfit on the training data, and more robust than a single decision tree. In regression, the final ensemble prediction is calculated by combining results from the individual trees, $M_1, M_2, \ldots M_n$, and getting their average. Figure 3.3 below shows a simple example of an ensemble. The RF was trained upon all the calculated 2D descriptors.

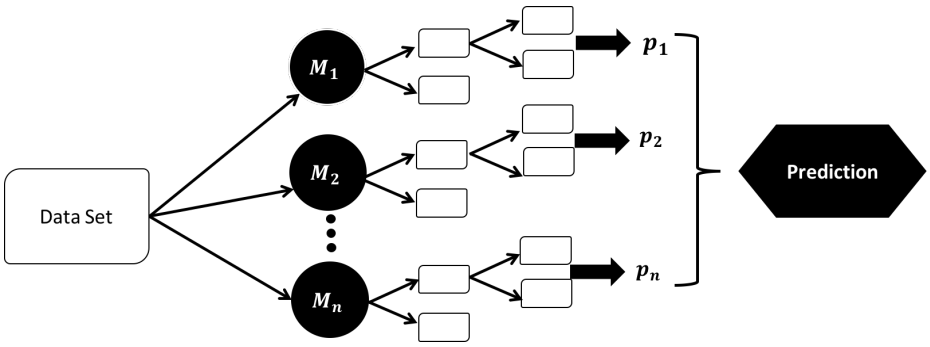


Figure 3.3: A Random Forest (RF) example shows steps of the ensemble method. Each empty box represents a leaf node that contains new molecules from the decision tree. P_n shows the real predicted value of each tree.

3.2.3 3D-QSAR

For the CoMFA models, the three-dimensional chemical structures of the G1 data set were generated using Obabel 2.3.1. This step was followed by generating 20 conformers for each molecule. The conformers were energy minimised using the Merck Molecular Force Field (MMFF).[181] The Gasteiger-Hückel method[133] was used to calculate the partial atomic charges. The 3D-QSAR calculations were performed using NCW and the Joelib2 package[182] accessed from NCW to calculate the molecular descriptors. The OMEGA algorithm[183] was used as a separate package to align the 3D molecular structures (see chapter 2). An active molecule should be chosen from the data set to be a template. This template is used to superpose all individual molecules and to afford a rational and consistent alignment. In our study, molecule **2** was used as a template as its activity was highest, and all of the remaining compounds were aligned to it to derive the subsequent 3D-QSAR analyses. The structure of the reference molecule with aligned molecules is shown in Figure 3.4.

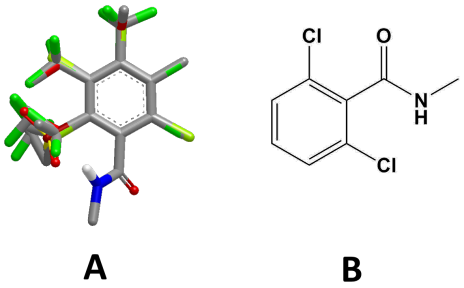


Figure 3.4: (A) Structural alignment of 27 benzamides according to OMEGA. (B) The template molecule (compound **2)**

The steric energy at each grid point is calculated by using the Lennard-Jones equation and the electrostatic energy is calculated by using the Coulomb potential. Energies exceeding a specified value (default 30 kcal/mol) are set to the cutoff value to minimise domination by large electrostatic and steric energies. To generate the steric and electrostatic fields, an sp^3 hybridised carbon atom with a charge of +1 is used as a probe atom. A distance-dependent Gaussian-type function (equation 3.1) was used to allocate each field value.

$$G_{pj}(c) = \sum_{i=1}^{n} P_i e^{-\alpha r_{ij}^2} \tag{3.1}$$

where G_{pj} is the field value of property p at grid point j for compound c; i is the atom index; p_i is the atomic property of atom i; α is the attenuation factor which is set at 0.3; r_{ij} is the distance between atom i and grid point j. The common distances between the probe atom and the atoms of the compounds in the aligned dataset should be taken into account to compute similarity at all grid points. Therefore, the Gaussian-type function is employed herein to describe this distance-dependence and to calculate the molecular property.

The resulting PLS regression equation is used to predict activity from values at the grid points. All models were validated using LOO-CV. To visualise the 3D-QSAR models, contour maps were generated by interpolating the product of the PLS

coefficient and the standard deviation of the 3D field descriptor. This interpolating step produces "stdev*coef" fields, which best explain the relationship between changes in the activity and variations in the property of a molecule. An equivalent percentage enclosing the lowest and highest field values was extracted from the fields, since we were interested only in descriptors with the most significant influence on the activity of a molecule.

3.3 Results and discussion

3.3.1 2D-QSAR with PLS analysis

The number of components used in the PLS model was varied from 1 to 10. This gives the total number of models generated as 150. Only models with q^2 of 0.50 or higher are shown in the tables. For the G1 data set the highest LOO-CV q^2 was 0.57 (Table 3.3). We also omit from the table models where r^2 is greater than 0.95 (as such models appeared to be overfit). We adopted the previous methodology for the G2, G3, and G4 data sets. A comparison of the most predictive models for all the data sets is presented in Table 3.4. For the G1 data set, the log$S$ model with 6 components represents the highest predictive model. For the G2 data set, the same model with the same number of PLS components shows the highest predictive model too with q^2 of 0.58 (Table 3.4). This means that the TMACC calculation on the G2 data set would not affect the accuracy of LOO-CV q^2 values when comparing with the fragmented structures in the G1 data set. However, the LOO-CV results were different for the G3 and G4 data sets. For the G4 data set, adding Gasteiger atomic partial charge and/or MR to the model increases accuracy. The highest scoring model, charge log$S$ MR, gives $q^2 = 0.72$ with 7 components.

Descriptors	Components	r^2	MAE_{train}	$RMSE_{train}$	q^2	MAE_{CV}	$RMSE_{CV}$
PSR	6	0.93	0.13	0.16	0.50	0.33	0.45
S	2	0.68	0.27	0.35	0.50	0.34	0.44
S	4	0.77	0.23	0.30	0.52	0.34	0.44
S	5	0.80	0.22	0.27	0.53	0.35	0.44
S	6	0.82	0.20	0.26	0.57	0.33	0.42
S	7	0.83	0.19	0.25	0.54	0.33	0.44
S	8	0.84	0.17	0.25	0.55	0.31	0.43

Table 3.3: Training and testing performance measures for Models Generated from Different Descriptors (G1 group): E = charge, P = log$P$, MAE_{train} = Mean absolute error for training set, $RMSE_{train}$ = Root-mean-square error for training set, MAE_{CV} = Cross-validated mean absolute error, and $RMSE_{CV}$ = Cross-validated root-mean-squared error.

Data set	Descriptors	Components	r^2	MAE_{train}	$RMSE_{train}$	q^2	MAE_{CV}	$RMSE_{CV}$
G1	S	6	0.82	0.20	0.26	0.57	0.33	0.42
G2	S	6	0.88	0.18	0.22	0.58	0.32	0.41
G3	S	10	0.80	0.21	0.33	0.57	0.36	0.50
G4	ESR	7	0.92	0.17	0.20	0.72	0.31	0.38

Table 3.4: The most predictive models for G1, G2, G3, and G4 data sets: E = charge, P = log$P$, S = log$S$, and R = MR, MAE_{train} = Mean absolute error for training set, $RMSE_{train}$ = Root-mean-square error for training set, MAE_{CV} = Cross-validated mean absolute error, and $RMSE_{CV}$ = Cross-validated root-mean-squared error.

It is clear that the model based on charge and solubility, and molar refractivity is the most predictive. The charge log$S$ MR model with seven components is characterised by $q^2 = 0.72$, $RMSE_{CV} = 0.38$, $r^2 = 0.92$ and $RMSE_{train} = 0.20$. It is noticeable that the r^2 for the ESR model is higher than for the other models in Table 3.4. The correlation between experimental and predicted pIC_{50} for the best predictive model is presented in Figure 3.5. The predicted versus observed activities for G4 data set as well as the residuals are listed in Table 3.5. For a QSAR model to be characterised as having high predictive power, a high value of LOO-CV q^2 is a necessary but not sufficient condition.[108] Golbraikh and Troposh[108] emphasised that the predictive ability of a QSAR model can be estimated using an external test set of molecules. These are molecules not used for building the model. However, due to the limited number of molecules in our data sets, we employed the Y-randomisation

method to validate our models. More details on this approach will be discussed later in this chapter.

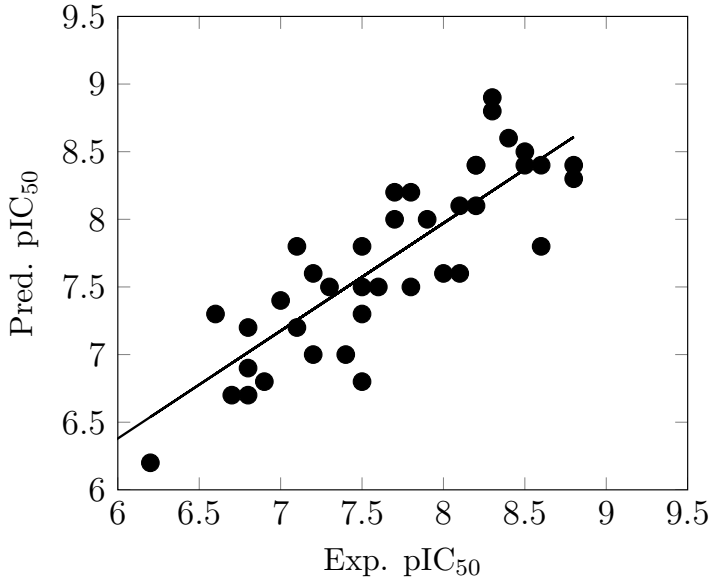


Figure 3.5: The correlation between the experimental and predicted pIC_{50} for the Charge, log$S$, and MR model with 7 PLS components (G4 data set); a regression line has been drawn through the data (solid line).

Of the 40 compounds in figure 3.5, the number of predictions with the residuals (difference between predicted and observed pIC_{50} values) larger than 0.6 are **3**, **20**, **22**, **25**, and **40**. The values of calculated performance metrics ($q^2 = 0.72$, $MAE_{CV} = 0.31$) indicate that the predictions of pIC_{50} values are made with satisfactory accuracy. Tropsha et al. have suggested that regression QSAR models are acceptable if they meet the following conditions: $r^2 > 0.6$, $q^2 > 0.5$, which are essential but not sufficient.[100] Models with a high difference (0.30) between its fitting and prediction are likely to be affected by overfitting.[184] For example, the best model from the G2 data set may be overfitted due to the high difference of 0.30 between its r^2 and q^2 values.

Molecule No.	Exp. pIC_{50}	Pred. pIC_{50}	Residual
1	8.3	8.8	0.5
2	8.8	8.3	-0.5
3	8.6	7.8	-0.8
4	8.5	8.5	0.0
5	8.5	8.4	-0.1
6	7.8	8.2	0.4
7	8.2	8.1	-0.1
8	8.6	8.4	-0.2
9	7.7	8.0	0.3
10	8.2	8.4	0.2
11	7.9	8.0	0.1
12	6.7	6.7	0.0
13	8.1	7.6	-0.5
14	8.1	8.1	0.0
15	7.5	7.5	0.0
16	7.3	7.5	0.2
17	7.6	7.5	-0.1
18	7.2	7.6	0.4
19	7.7	8.2	0.5
20	6.6	7.3	0.7
21	8.0	7.6	-0.4
22	7.5	6.8	-0.7
23	7.8	7.5	-0.3
24	8.2	8.4	0.2
25	7.1	7.8	0.7
26	7.1	7.2	0.1
27	8.3	8.9	0.6
28	8.8	8.4	-0.4
29	7.5	7.3	-0.2
30	8.4	8.6	0.2
31	7.0	7.4	0.4
32	7.4	7.0	-0.4
33	7.5	7.8	0.3
34	6.8	7.2	0.4
35	6.2	6.2	0.0
36	6.8	6.9	0.1
37	6.9	6.8	-0.1
38	7.2	7.0	-0.3
39	6.8	6.7	-0.1
40	6.1	6.6	0.7

Table 3.5: The predicted activities for G4 group from the LOO-CV.

3.3.2 2D-QSAR with machine learning analysis

As described in section 3.2.2, we used seven machine regression methods: SMO, SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and REPTree on the G4 data set. A total of 33 out of all 105 models afforded a q^2 higher than 0.50, and 62 models had MAE smaller than 0.5 for this LOO-validation test (Table 3.6). All SMO models gave q^2 values smaller than 0.5, and they performed worse than the other algorithms.

The performances in terms of MAE are assessed in LOO-CV for SVM (RBFKernel), MLP, IBk, M5P, RF, and REPTree and are given in Figure 3.6. The results indicate that most of the models were successful in training the experimental data for the 40-compound set.

RF models outperform REPTree models and are close to SVM models, although they are still less accurate than IBk models. MLP models are likely to be overfit. RF and SVM models are more stable and tolerant to noise in the input data.

As a supervised learning method, RF regression outperforms IBk, MLP, and M5P. We believe that the IBk performance could be improved by choosing a more appropriate set of IBk parameters. IBk algorithm uses distance-based comparisons that intrinsically assign equal weight to each attribute. It hence can be affected from low accuracy when provided irrelevant attributes. However, the method has been modified to incorporate attribute weighting and the pruning of noisy data tuples. The k value that yields the minimum error rate was selected ($k = 5$). This means the larger the number of training molecules, the larger the value of k will be.

Table 3.6: The highest predictive models using the SVM (RBFKernel), MLP, IBK, M5P, RF, and REPTree approaches on the G4 dataset.

	Training				Validation		
Methods	Model	r^2	MAE	RMSE	q^2	MAE	RMSE
SVM (RBFKernel)	log$S$	0.88	0.15	0.23	0.55	0.43	0.51
MLP	Charge log$P$ log$S$	0.92	0.25	0.33	0.58	0.42	0.58
	log$S$ MR	0.90	0.22	0.32	0.58	0.47	0.59
IBk	log$S$	0.77	0.31	0.39	0.61	0.39	0.51
	Charge log$P$ log$S$ MR	0.72	0.30	0.41	0.56	0.37	0.50
	log$S$ MR	0.76	0.31	0.42	0.55	0.40	0.54
	log$P$ log$S$ MR	0.74	0.32	0.45	0.55	0.40	0.55
M5P	log$S$	0.92	0.16	0.21	0.50	0.43	0.53
RF	log$S$	0.88	0.22	0.29	0.64	0.35	0.46
	log$S$ MR	0.88	0.23	0.28	0.62	0.37	0.47
	Charge log$P$ log$S$ MR	0.88	0.23	0.30	0.61	0.37	0.47
	Charge log$P$ log$S$	0.88	0.23	0.29	0.59	0.37	0.48
	Charge log$S$ MR	0.88	0.24	0.31	0.59	0.38	0.49
	log$P$ log$S$	0.88	0.22	0.28	0.56	0.37	0.49
	Charge log$S$	0.90	0.24	0.31	0.58	0.39	0.51
	log$P$ log$S$ MR	0.85	0.23	0.30	0.58	0.38	0.48
	Charge log$P$	0.88	0.23	0.30	0.53	0.39	0.51
	log$P$	0.85	0.23	0.31	0.53	0.38	0.50
	Charge log$P$ MR	0.86	0.23	0.30	0.52	0.39	0.51
REPTree	Charge log$S$	0.86	0.18	0.27	0.55	0.37	0.50
	Charge log$P$ log$S$	0.92	0.13	0.19	0.55	0.38	0.50
	Charge log$P$ log$S$ MR	0.92	0.13	0.19	0.53	0.39	0.51
	Charge log$P$	0.92	0.13	0.19	0.52	0.38	0.52
	Charge log$P$ MR	0.92	0.13	0.19	0.50	0.39	0.53
	log$S$	0.86	0.20	0.27	0.50	0.43	0.52

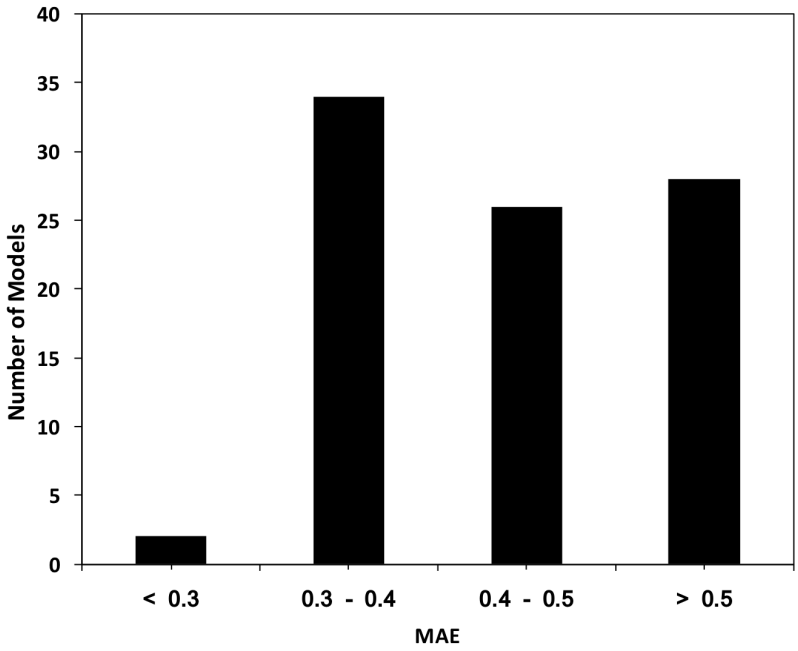
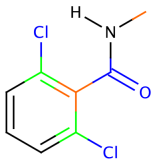


Figure 3.6: Predictive performance (MAE) of SVM (RBFKernel), MLP, IBk, M5P, RF, and REPTree models assessed in LOO-CV on the training set of 90 models (all models).

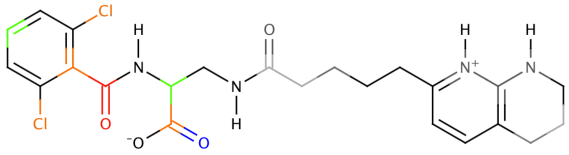
3.3.3 2D-QSAR interpretation

A TMACC interpretation was generated for each molecule, as described in chapter 2, leading to the colouring of each atom according to its contribution to activity. Activity increased upon moving from a *para*-substituted phenyl ring to an *ortho* substitution, for both chloro- and fluoro-substituted compounds. Some examples are shown in Figures 3.7 and 3.8.

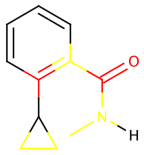
Below can be found a short comparison between two models, log$S$ with 6 components and a q^2 of 0.57 (G1 data set) and log$S$ with 6 components and a q^2 of 0.58 (G2 data set). For the 2,6-dichloro substituents the models do not agree with each other. In the G1 model, both Cl moieties contribute positively to the activity, but they negatively contribute to activity in the case of G2 data set. For the 2-cyclo-propyl substituents, both are suggesting a slight positive contribution to activity but with a small variation in the magnitude (Figures 3.7c and 3.7d).



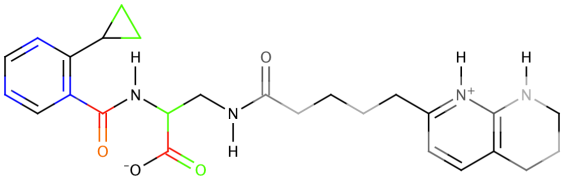
(a) G1; Molecule 2



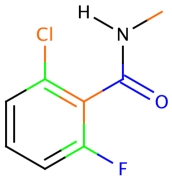
(b) G2; Molecule 2



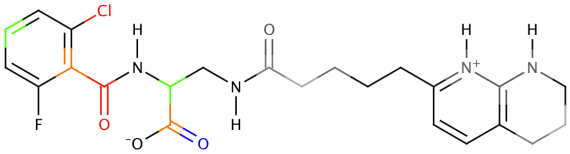
(c) G1; Molecule 22



(d) G2; Molecule 22

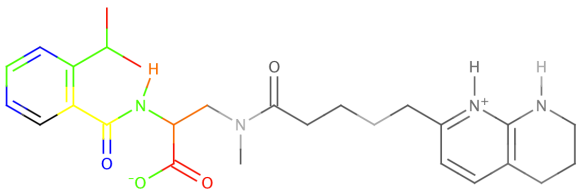


(e) G1; Molecule 5

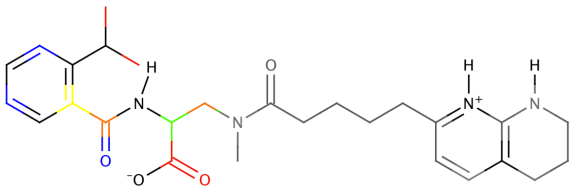


(f) G2; Molecule 5

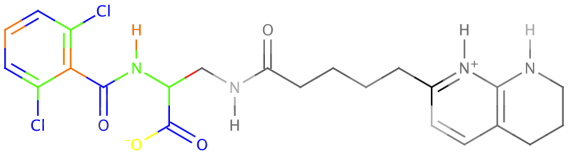
Figure 3.7: TMACC interpretation of log$S$ models with 6 PLS components for G1 and G2 groups. TMACC activity colour scheme: red for very negative activity; orange for negative activity; green for neutral activity; yellow for positive activity and blue for very positive activity. **a)** The Cl substituents on the *ortho*-position both provide a positive contribution to the activity. **b)** A negative contribution for both Cl groups to the activity. **c)** The cyclo-propyl moiety provides a positive contribution to the activity. **d)** The cyclo-propyl group is shown in green, indicating that it provides no contribution to the activity of the molecule. **e)** F and Cl provide a negative and positive contribution to the activity, repectively. **f)** The Cl provides a negative contribution to the activity.



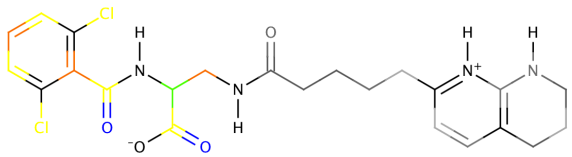
(a) Charge log$S$ MR; Molecule 20



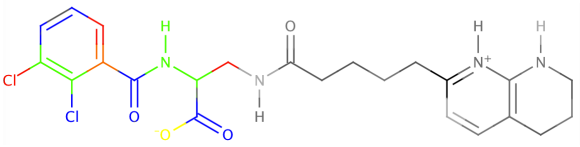
(b) log$S$ MR; Molecule 20



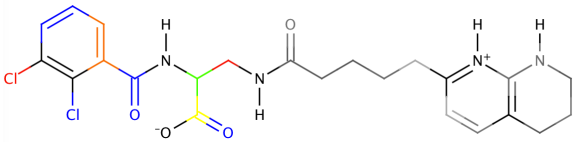
(c) Charge log$S$ MR; Molecule 2



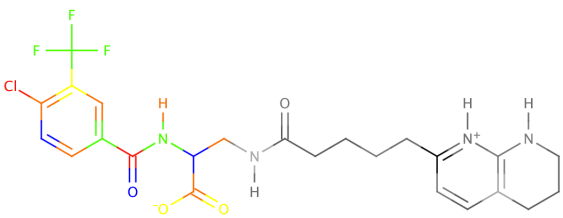
(d) log$S$ MR; Molecule 2



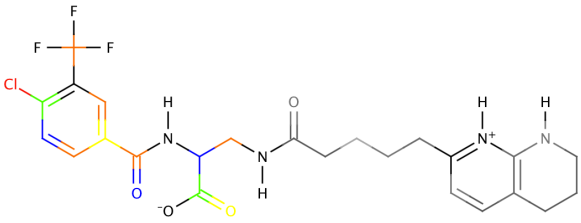
(e) Charge log$S$ MR; Molecule 9



(f) log$S$ MR; Molecule 9



(g) Charge log$S$ MR; Molecule 12



(h) log$S$ MR; Molecule 12

Figure 3.8: TMACC interpretation of G4 group for two different models; Charge log$S$ MR with 7 components (a,c, e and g molecules), and log$S$ MR with 7 components (b, d, f and h molecules). (a) The iso-propyl moiety provides a slight negative contribution to the activity of the molecule. (b) More negative contribution for the iso-propyl group to the activity. (c) The Cl groups provide a slight positive contribution to the activity. (d) The Cl substituents are shown in blue, indicating that they provide a very positive contribution to the activity. (e) The Cl group on the *ortho*-position and *meta*-position provides a positive, and negative contribution, respectively, to the activity. (f) The Cl provides the same contribution effect to the activity of the molecule as same as in (e). (g) The *para*-chloro substituent provides a negative contribution to the activity while no effect is noted for CF_3 group in the *meta*-position. (h) A positive activity-contribution for the CF_3 group in the *meta*-position

Both Cl substituents on the *ortho*-positions in Figures 3.7e and 3.7f provide a negative contribution to the activity. According to Figure 3.7, the comparison between the TMACC results for the same property (log$S$) with the same number of PLS components, but in different data sets shows how substituents would contribute in different ways to activity. This also suggests that the NCW is sensitive to the small variation in similar, but not exact data sets.

For the 2,6-dichloro substituent, in Figure 3.8, both models agree. However, they suggest a different level of positive contribution to the activity, (Figures 3.8c and 3.8d). For the 2,3-dichloro substituent both models also agree, with both suggesting a positive contribution (Figures 3.8e and 3.8f). The models also do agree on this positive contribution, (Figures 3.8g and 3.8h). The trifluoromethyl group on the *meta* position in the log$S$ MR and Charge log$S$ MR models does not contribute to the activity while existing of the Cl moiety on the *meta*-position does.

To study the behaviour of r^2 and q^2 for a single model when adding more PLS components, we plotted (Figure 3.9) the number of PLS components against the r^2 and q^2 values for the most predictive model of G4. The statistical measures associated with 1 to 15 components are shown in Table 3.7. The model quality was expected to increase with an increasing number of PLS components up to a specific number. Then, q^2 starts to decrease, although r^2 continues to rise. However, the statistical measures of the predictive performance of the charge log$S$ MR model seem to plateau.

PLS Components	r^2	MAE_{train}	$RMSE_{train}$	q^2	MAE_{CV}	$RMSE_{CV}$
1	0.59	0.37	0.46	0.38	0.44	0.56
2	0.76	0.28	0.35	0.51	0.39	0.51
3	0.81	0.26	0.31	0.58	0.38	0.47
4	0.88	0.21	0.25	0.65	0.36	0.44
5	0.90	0.18	0.23	0.65	0.37	0.43
6	0.91	0.17	0.21	0.69	0.34	0.41
7	0.92	0.17	0.20	0.72	0.31	0.38
8	0.95	0.13	0.17	0.69	0.34	0.42
9	0.95	0.12	0.16	0.69	0.35	0.42
10	0.96	0.11	0.14	0.69	0.34	0.42
11	0.96	0.11	0.13	0.69	0.34	0.42
12	0.97	0.09	0.12	0.68	0.34	0.43
13	0.98	0.08	0.10	0.67	0.36	0.44
14	0.98	0.07	0.10	0.66	0.36	0.45
15	0.98	0.06	0.09	0.65	0.37	0.45

Table 3.7: Training and testing performance measures for charge log$S$ MR as a combined descriptor (G4 group): MAE_{train} = Mean absolute error for training set, $RMSE_{train}$ = Root-mean-square error for training set, MAE_{CV} = Cross-validated mean absolute error, and $RMSE_{CV}$ = Cross-validated root-mean-squared error.

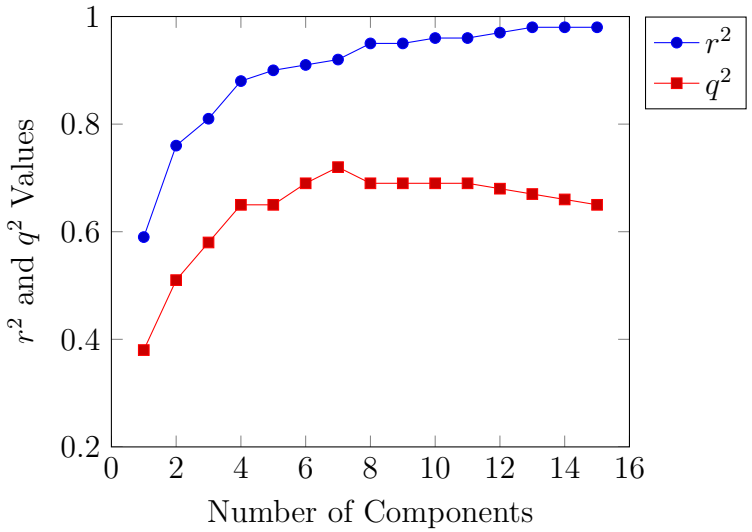


Figure 3.9: The variation of r^2 and q^2 with the number of components (latent variables) for the charge log$S$ MR model.

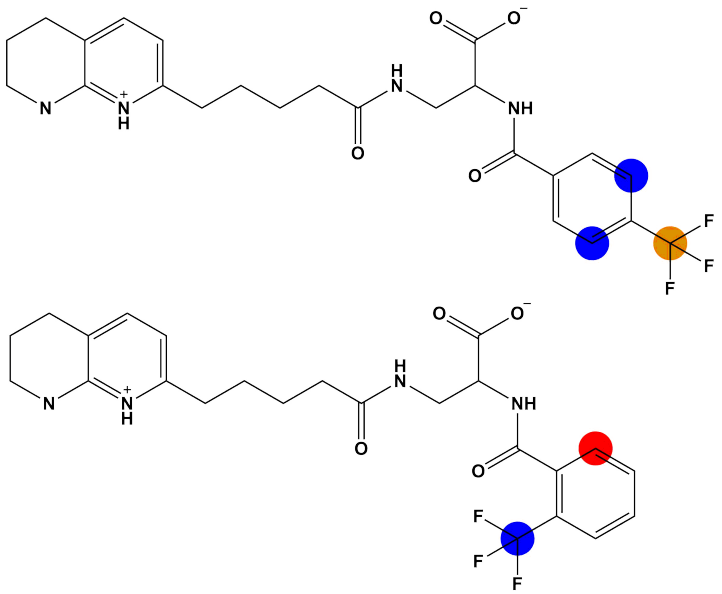


Figure 3.10: Visualisation of the TMACC/PLS charge log$S$ MR model of the G4 data set showing the effect of modifying a *para*-CF_3-substituted benzamides inhibitor molecule **15** (top, $pIC_{50} = 7.5$) to an *ortho*-CF_3-substituted molecule **14** (bottom, $pIC_{50} = 8.1$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates that the CF_3-substitution becomes less potency-reducing in the *para*-position.

For the G4 data set, the activity increased upon moving from a *para*-substituted phenyl ring to an *ortho* substitution, for CF_3- , Cl-, and F-substituted molecules. Figure 3.10 displays molecule **14**, and **15**, the *para*- and *ortho*-substituted compounds, respectively. The atoms that most contribute to activity are coloured in blue, and those that most reduce activity are given in red. We also display the colouring of the CF_3 group. In the interest of simplicity, we do not show the contributions of the other atoms as their impact on the activity is not especially influential. When the *ortho*-substituted molecules are examined, the contribution of the *ortho*- Cl, F, and CF_3 groups are coloured in blue; hence, this group has been identified as an activity-increasing group. However, the *meta*- and/or *para*-substituted groups are coloured orange; this also indicates an activity-decreasing group, but their effects are more moderate. Moreover, when a molecule has two activity-increasing groups such as Cl substituents on the *ortho*-positions of a phenyl ring, their impact on activity will

increase (Figure 3.11). This figure shows a similar effect from *ortho* substitution, but this time with two chloro groups, illustrated by compound **2** and **10**. By examining the pIC_{50} values of both molecules, adding an extra Cl substitution increases the pIC_{50} by 0.6 log units. Both Cl groups have to be at the *ortho*-position to increase activity. However, when only one Cl group is located at the *meta*-position and the other substituent is kept in the *ortho*-position (molecule **9**), the pIC_{50} values were reduced by 1.0 log units when comparing with molecule **2**.

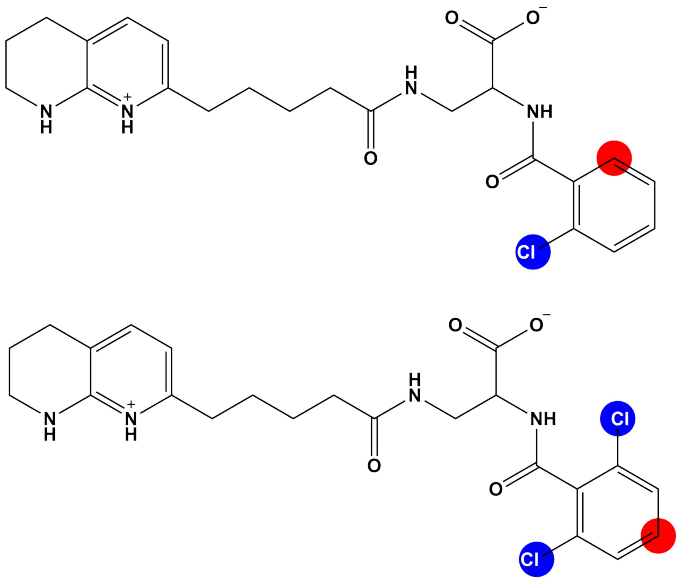


Figure 3.11: Visualisation of the TMACC/PLS model showing the effect of modifying an-ortho single Cl-substituted benzamide inhibitor molecule **10** (top, $pIC_{50} = 8.2$) to two *ortho*-Cl-substituted molecule molecule **2** (bottom, $pIC_{50} = 8.8$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors.

The impact of fluoro groups on activity is similar to the Cl substituents when they are at the *ortho*-position, but slightly lower (Figure 3.12). Their effect on activity decreases at the *meta*- and/or *para*-positions. The pIC_{50} values were reduced by 0.5 log units when one F group is in the *meta*-position and the other substituent to the *para*-position (molecule **18**). The F substitution in the *meta* position changed from blue (as in *ortho*-case) to yellow, indicating a change from activity-increasing

to moderately activity-increasing. By examining the pIC_{50} values of the molecules, changing from an *ortho*-chloro to a *para*-chloro substitution decreases the pIC_{50} by 1.6 log units approximately, while changing from *ortho*-fluoro to *para*-fluroro substitution decreases the activity by 0.5 log units. Further mono-*ortho* and di-*ortho* substituents, (for instance, Cl, F, CH_3, and C_2H_5 groups), may be important for increasing the $\alpha_v\beta_1$ activity. However, these data do not indicate mono-substitution in the *ortho* position is useful for increasing activity. These observations can be rationalised by noting that the steric effects of the substituents may critical for increasing activity. Additionally, the *meta*- and *para*-substituted molecule clearly indicates that a lack of substitution at the 2 position of the aryl ring decreases activity.

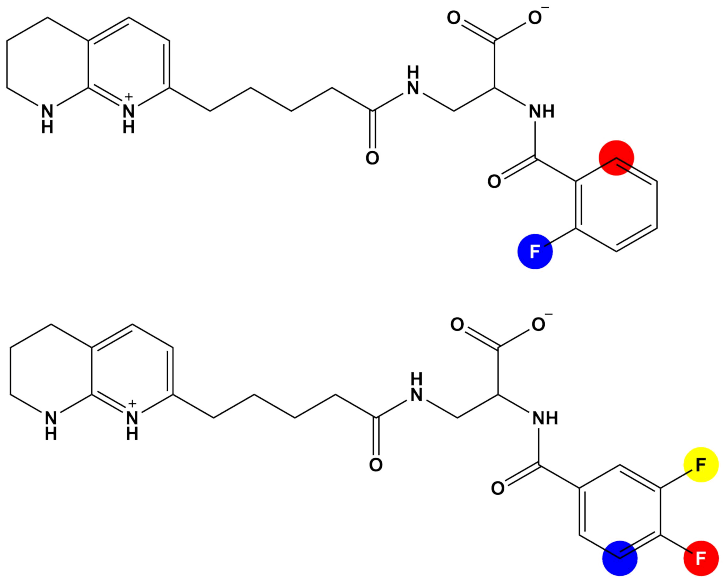


Figure 3.12: Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* single Cl-substituted benzamide inhibitor molecule **19** (top, $pIC_{50} = 7.7$) to two F-substituted molecule molecule **18** (bottom, $pIC_{50} = 7.2$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The yellow colour indicates that the F-substitution becomes less potency-increasing in the *meta*-position.

Among the studied molecules we also have noticed that presence of a nitrogen in a heterocyclic benzamide will reduce activity, especially when the nitrogen atom is on the *para*-position of the phenyl moiety. To clarify this point, we were interested

to see our TMACC interpretation for two different inhibitors (Figure 3.13). The first is an inhibitor with high affinity (molecule **28**, top) with a measured pIC_{50} of 8.8, and the other is a less active inhibitor (molecule **40**, bottom), which had a measured pIC_{50} of 6.1. We also highlight those atoms that our TMACC/PLS analysis

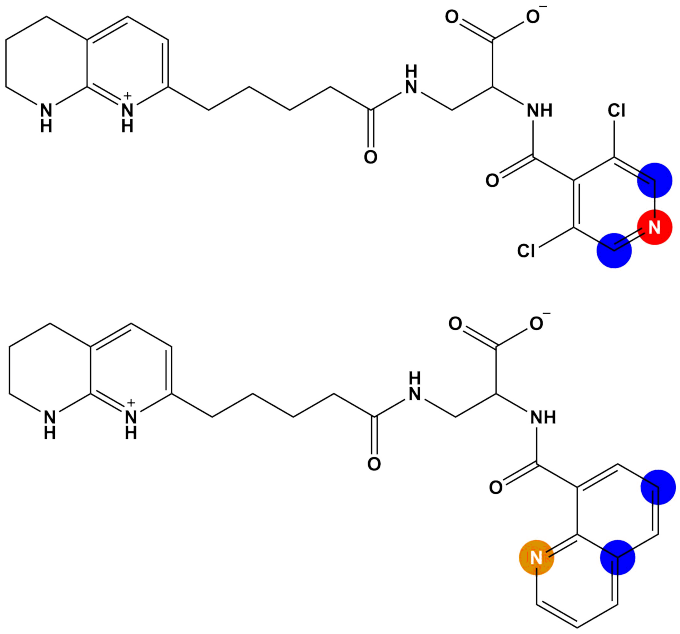


Figure 3.13: Visualisation of the TMACC/PLS model showing the nitrogen contribution to the activity for the highest and the less active inhibitors; The first inhibitor is molecule **28** (top, $pIC_{50} = 8.8$) and the other inhibitor is molecule **40** (bottom, $pIC_{50} = 6.1$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates that the N atoms becomes less potency-decreasing in the quinoline ring.

indicated contributed the most to increasing activity (blue) and decreasing activity (red). Figure 3.13 also shows that the TMACC highlights the other carbonyl carbons in molecule **28** and different aliphatic carbons of molecule **40** as activity-increasing. There is more than a 10-fold range in activity against $\alpha_v\beta_1$ (pIC_{50} 7.0 - 8.1) with an *ortho*-CF_3 **14** showing the most potent activity and an *ortho*-CF_3 in a heterocyclic ring with an *ortho*-nitrogen atom **31** the least, suggesting that the electronic property of the nitrogen atom (the high electronegativity of the nitrogen) may decrease the activity. Despite varying size of substituents, a similar pattern is seen with **29** and

10. The TMACC also highlights the sulfonamide moiety in molecule **26** as activity enhancer (Figure 3.14).

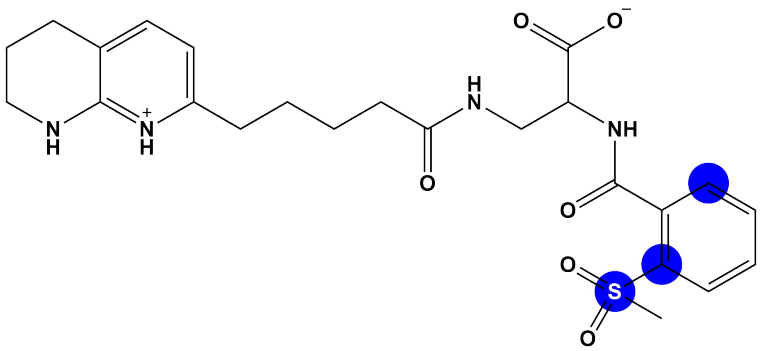


Figure 3.14: Visualisation of the TMACC/PLS model showing the sulfonamide contribution to the activity for molecule **26** ($pIC_{50} = 7.1$). Highlighted atoms indicated those that are most important for increasing (blue) the potency of inhibitors.

3.3.4 3D-QSAR Analysis

3.3.4.1 Evaluating of 3D QSAR models

Models were generated to consider all four property types individually: charge, log$P$, log$S$, and MR, and for all possible combinations to yield a total of 15 models. Statistics for all models are provided in an Appendix; they are shown in Table 3.8 for the top five models. These models were selected based on their predictive ability (q^2). All five gave comparable q^2 values, differing by 0.10 approximately.

	E + log$P$ + MR	E + log$P$ + log$S$ + MR	E + log$P$	E + log$P$ + log$S$	log$P$ + MR
q^2	0.48	0.46	0.41	0.38	0.36
r^2	0.54	0.55	0.46	0.46	0.42
SEE	0.49	0.50	0.52	0.53	0.54

Table 3.8: Statistical data for 3D-QSAR analyses of benzamide derivatives (G1 data set). SEE is the standard error of estimate.

All top five models were selected with PLS less than five components since the

r^2 values, in general, increase with the number of components used to construct the model.[185] The cross-validated standard error and the standard error (SEE) of the model residuals can be used as measures of goodness of predictive and fit ability.[103,185] This suggested that the model with three properties was the best model, ($q^2 = 0.48$; SEE 0.49). The most predictive 3D-QSAR model for the G2 dataset is the charge, log$P$, and MR model, and is presented in Figure 3.15. This model provides insight into the spatial features that are important for activity. Therefore, we focus on this model hereafter. To justify the robustness of the predictive model, the models built using randomised activities should have much lower q^2 and r^2 values than the model with the current activities.[103] Over the 350 models, the average q^2 and r^2 values were 0.18 and 0.42, respectively. Out of the 350 permutated activity models 96 gave q^2 values greater than zero. Figure 3.16 shows the relationship between the r^2 and q^2 for the scrambled activity models that had non-zero q^2 values. Only four of these models had q^2 greater than 0.40, the highest of which was 0.46. Therefore, we can be confident our real model is predictive and not due to chance of correlation.

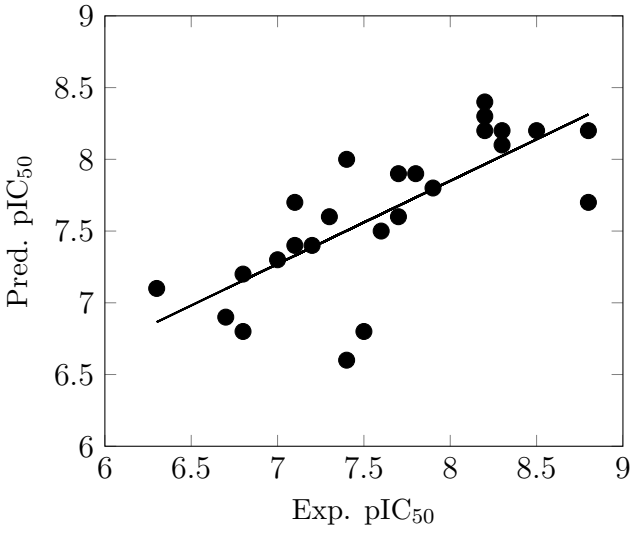


Figure 3.15: Experimental vs. predicted pIC$_{50}$ for the charge, log$P$, and MR model

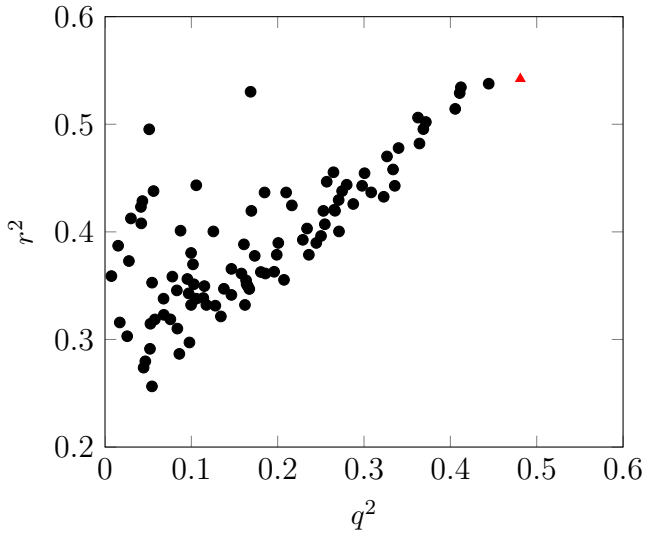


Figure 3.16: q^2 versus r^2 for the 96 y-randomisation models with q^2 values greater than zero (black dots). The real model is represented by the red triangle.

3.4 Conclusions

Described here are 2D/3D-QSARs against the $\alpha_v\beta_1$ integrin with the aim of ultimately identifying an orally bioavailable selective and potent $\alpha_v\beta_1$ inhibitor for treating IPF. Different models using four data sets: G1, G2, G3 and G4 were validated and the best model indicates the importance of the combined three properties: charge, log$S$, and molar refractivity for a model. Having only the TMACC values for each atom pair provides better results compared with the 3D-QSAR method. Comparison to the 3D-QSAR demonstrates that the TMACC performs better and is superior. The best 2D-QSAR model established a good predictive accuracy (q^2 of 0.72) for the G4 data set, using 2D descriptors generated from the four property types. The 2D-QSARs suggests the mono-substitution in the *ortho*-position is more influential on activity against $\alpha_v\beta_1$. A greater range in activity was seen with disubstituted analogues compared with mono-substituted analogues, such as 2,6-dichloro in molecules **2** and **28**, and 2,3-dichloro **9** and **7**. These results are in agreement the available experimen-

tal work.[187] Visualising the model with the NCW using a colour scheme for each property identified key features which influence the activity of the molecules. This 2D interpretation also suggested several positions on the benzamides where modifications may enhance activity. Hence, the study has provided us with predictive and diagnostic value for the design of a future set of benzamides analogues. This study provides evidence for the potential of the TMACC approach for the development of QSAR models. Compounds which are partially selective for $\alpha_v\beta_6$ have also been studied and further studies will be reported in the next chapter. In this study, we also have illustrated the utility of parametric and non-parametric regressors in the analysis of structure-activity data. The SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and REPTree methods appear not be the most profitable strategy for the data set under consideration. The greater accuracy of PLS over the non-parametric methods is clear. The data do not afford any very obvious or simple relationships. The best non-parametric model (RF model) achieves cross-validated q^2 values of 0.64, corresponding to predictive utility.

Chapter 4

Inhibitors of $\alpha_v\beta_6$ integrin

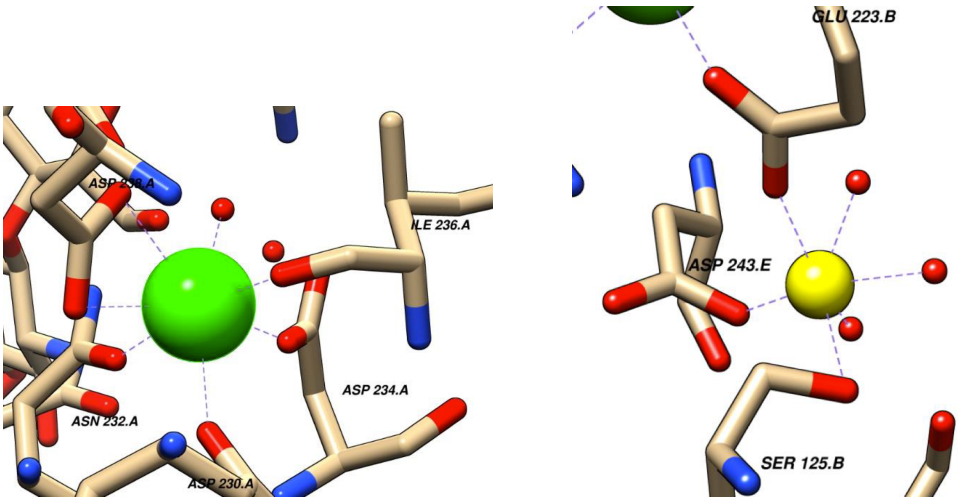
4.1 Introduction

The $\alpha_v\beta_6$ integrin is a cell surface heterodimeric receptor, which links the α_v and β_6 proteins by a non-covalently bound matrix.[186,187] This receptor is mainly expressed in fibrosis and selective $\alpha_v\beta_6$ antibodies are in development to control this disorder.[188] The $\alpha_v\beta_6$ integrin activates TGF-β in models of fibrosis in the lung, kidney and biliary tract.[161] Cross-screening data against $\alpha_v\beta_1$ and/or $\alpha_v\beta_6$ from the selectivity profiles of other integrin antagonists in the literature are limited.[188] An antifibrotic agent against the epithelium-specific $\alpha_v\beta_6$ (STX-100) is in a phase II study in IPF. The subunit β_6 presents in a number of epithelial cell lines.[189] The integrin heterodimer, $\alpha_v\beta_6$, on the surface of two human carcinoma cell lines was identified using a polyclonal antibody raised against the carboxyl-terminal peptide of β_6.[189] It also was demonstrated that the $\alpha_v\beta_6$ binds to fibronectin using affinity chromatography of lysates from the pancreatic carcinoma cells.[189] In response to bleomycin and due to deficiency of collagen subunit Col4A3, the $\alpha_v\beta_6$ integrin was identified as promoting lung and kidney fibrosis, respectively.[190,191]

4.1.1 From structure to function

The $\alpha_v\beta_6$ inhibitor was found to be a fibronectin-binding protein in human pancreatic carcinoma cell by using affinity chromatography.[189] The α_v integrin does not interact with immobilised collagen I, which is the central structural protein in the extracellular-matrix. However, the immobilised collagen I binds β_1 containing integrins. The $\alpha_v\beta_6$ integrin is eluted from its respective immobilised ligand by a hexa-peptide containing the sequence Arg-Gly-Asp (RGD).[189] The RGD in presence of Ca^{2+} is effective at ligand binding more than in the presence of Mg^{2+}. This makes $\alpha_v\beta_6$ function as an RGD-dependent fibronectin receptor in carcinoma cells.[189] The integrin appears to be the primary mediator of cell-extracellular matrix adhesion, and can be considered as one of the many families of molecules active in cell-cell adhesion. The matrix receptor is a cell surface heterodimeric receptor because it always has one alpha and one beta chain. These two subunits contribute to the binding of a ligand, which depends on extracellular divalent cations Ca^{2+} or Mg^{2+} (Figure 4.1).[192]

The divalent cation-binding cations Ca^{2+} or Mg^{2+} are attached to the extracellular domains of integrins. While high-affinity conformations of the integrin are promoted by Mg^{2+} under physiological conditions, the low-affinity ones are promoted by the Ca^{2+} cation.[193] The ligand-binding sites of integrins, which are expressed on blood cells: leukocytes or platelets, are important targets for late-stage clinical drugs. Conformational changes or "activation processes" occur in leukocyte and platelet integrins. The leukocyte and platelet integrins play an important role in the inside-out signalling.[73] For example, the affinity of leukocyte integrins can be changed by about 10,000-fold.[194] I-domain-containing integrin is a domain that has homology to the von Willebrand factor A and can be located in the extracellular portion of the α-subunit.[186] There are three known TGF-β isomers, TGF-β_1, TGF-β_2 and TGF-β_3. The TGF-β_1 is a homodimeric pro-TGF-β, which consists of TGF-β linked with a



(a) (b)

Figure 4.1: (a) Crystal structure of $\alpha_v\beta_6$ shows the binding site ligand (Ca^{2+} ion, in green). ASP 230.A, ASP 234.A, ASP 238.A, ASN 232.A, and ILE 236.A are shown to indicate the key interactions with Ca^{2+}. (b) Crystal structure of $\alpha_v\beta_6$ shows the pocket interaction with Mg^{2+} ion, in yellow. SER 125.B, GLU 223.B, and ASP 243.E are shown to indicate the key interactions with Mg^{2+}. The crystal structure is taken from the protein data bank (PDB code: 4UM9).[195]

covalent bond to the latency associated protein (LAP).[196] The cytokine TGF-β_1 plays a central role in lung fibrosis. It is inactive within its normal state and cannot bind to its receptor. However, the TGF-β_1 can be activated when pathways involve integrins such as the $\alpha_v\beta_6$. The $\alpha_v\beta_6$ inhibitors have been particularly described in the pathogenesis of lung fibrosis.[197] Each TGF-β protein is processed intracellularly and assembled as a complex of two disulfide-linked homodimers. These dimers are formed from the shorter C-terminal fragment (the active cytokine) and the longer N-terminal fragment (LAP). The two homodimers are non-covalently linked to form a small latent complex. The mature C-terminal fragment is prevented to bind its receptors by the associated LAP homodimer. The associated LAP is also linked to other different proteins called the latent TGFβ binding proteins (LTBPs). The large latent complex is formed when one of these three proteins, LTBP 1, 3 and 4 (all contain cysteine residues) forms a disulfide bond with a conserved cysteine in TGF-β_1 , 2 and 3.[198,199] Thus, the extracellular matrix such as fibronectin attaches that large latent complex

and is cross-linked to TGFβ families through the action of extracellular tissue. An understanding of basic mechanisms of cell adhesion is needed to achieve success in the drug delivery world. The adaptor molecules such as Talin 1[200] and the kindlins (kindlin 1, kindlin 2 and kindlin 3)[201] are the best known regulators of $\alpha_v\beta_6$ integrin activation. In addition to their significant importance in adhesion, $\alpha_v\beta_6$ integrin is also associated in signal transduction. It was found that $\alpha_v\beta_6$ integrins are able to activate signal transduction pathways inside cells.

4.1.2 $\alpha_v\beta_6$ integrin–mediated TGF-β activation in IPF

In the pathogenesis of tissue fibrosis, the $\alpha_v\beta_6$ inhibitor activates the TGF-β. Development of fibrosis can be prevented by interference with $\alpha_v\beta_6$ integrin-mediated TGF-β activation.[202] However, over-expression of the $\alpha_v\beta_6$ integrin in the presence of injury can promote the development of tissue fibrosis.[203]

Thrombin and Lysophosphatidic acids are injurious signals that induce cytoskeleton cells. This results in $\alpha_v\beta_6$-mediated TGF-β activation. There is no release of free TGF-β into the extracellular matrix. As a result, the activated TGF-β interacts only with TGF-β receptors on adjacent epithelial cells under normal circumstances. The activated TGF-β also starts transcription of the integrin β subunit gene, which increases the expression of $\alpha_v\beta_6$ on the epithelial cell surfaces. Permitting activated TGF-β to interact with β receptors disrupts the basement membrane and initiates the IPF on mesenchymal cells. This drives cells to release more profibrotic mediators, transdifferentiate into myofibroblasts, and place new extracellular matrix that further expands the profibrotic medium.[202]

A series of novel compounds for activity against α_v integrins including $\alpha_v\beta_6$ has been synthesised and assayed by GSK and the University of Nottingham. The $\alpha_v\beta_6$ inhibitors are able to activate TGF-β *in vitro* and *vivo* by a pathway that involves cytoskeletal changes. In this chapter, a structure-activity relationship for inhibitors

of $\alpha_v\beta_6$ could be useful for developing new compounds towards a treatment of IPF. 2D- and 3D-QSARs have been derived for a series of benzamides as $\alpha_v\beta_6$ inhibitors. In addition, the impact of physicochemical and structural properties of compounds on $\alpha_v\beta_6$ inhibition was explained by contour maps.

4.2 Data sets

The aim of our QSAR study in this chapter is to build predictive models and to use them to screen *in silico* compounds that might be eventually synthesised. A range of 46 benzamide derivatives and their experimental activities were provided by GSK. They have been used to build models in order to relate a set of physicochemical properties to the biological activity (pIC_{50}). For simplicity and since there was no large variation in accuracy between the G1, G2, G3, and G4 datasets in the previous chapter, in this chapter, we only considered a single dataset comprising all 46 structures. This G5 dataset is shown in Appendix **A**; Table A.1. For molecule **25**, the experimental pIC_{50} value was given as < 5. However, the NCW software cannot deal with such imprecision. So, we have considered its activity to be 5.

4.3 Methods

4.3.1 2D-QSAR

The methods described in chapter 2 were employed. The molecules were studied by dividing them into four regions: naphthyridine rings, the amide group, the carboxylic group, and the benzamide group in order to study the structure-activity relationship. The substituents varied at the 2-, 3-, and 4-positions of the phenyl ring of the benzamide moiety. The other three regions were unchanged. The phenyl ring was also substituted by other rings. In chapters 2 and 3, two-dimensional QSAR methods were

described in detail and were employed herein. The topological data of each molecule was used to generate the TMACC descriptors. All nonpolar hydrogen atoms were excluded from the sdf file of the G5 data set while polar hydrogen atoms were considered explicitly by using Obabel 2.3.1. Gasteiger partial charges E, log$S$, log$P$, and molar refractivity (MR) were also used to assign the physicochemical properties to each atom, and 15 combinations of atomic properties were employed to calculate the TMACC descriptors (Table 3.1).

4.3.2 Machine learning methods

In this chapter, we have applied the six different machine learning algorithms using Weka 3.9.1: SMO, SVM, MLP, IBk, RT and RF. These algorithms have been described elsewhere (chapter 2 and chapter 3). The parameters in these methods were kept the same for consistency.

4.4 Results and discussion

4.4.1 2D-QSAR with PLS analysis

In each PLS model, the number of components used was varied from 1 to 10, giving 150 models number in total. We only considered those models with q^2 of 0.50 or higher (Table 4.1). The log$S$ model is most accurate and gives q^2 of 0.50 with 4 components. The model is characterised by $q^2 = 0.50$, $\text{RMSE}_{CV} = 0.54$, $r^2 = 0.66$, $\text{RMSE}_{train} = 0.45$. The correlation between the experimental and predicted pIC_{50} is shown in Figure 4.2.

Descriptors	Components	r^2	MAE_{train}	$RMSE_{train}$	q^2	MAE_{CV}	$RMSE_{CV}$
ES	4	0.68	0.35	0.44	0.47	0.47	0.56
S	3	0.64	0.38	0.46	0.46	0.48	0.57
S	4	0.66	0.37	0.45	0.50	0.46	0.54
S	5	0.67	0.36	0.44	0.49	0.47	0.56
S	6	0.68	0.34	0.43	0.50	0.47	0.56
S	7	0.69	0.33	0.42	0.48	0.48	0.57
S	8	0.70	0.33	0.42	0.49	0.47	0.56
S	9	0.71	0.33	0.41	0.48	0.48	0.58

Table 4.1: Training and testing performance measures for Models Generated from Different Descriptors (G5 dataset): E = charge, S = log$S$, MAE_{train} = Mean absolute error for training set, $RMSE_{train}$ = Root-mean-square error for training set, MAE_{CV} = Cross-validated mean absolute error, and $RMSE_{CV}$ = Cross-validated root-mean-squared error.

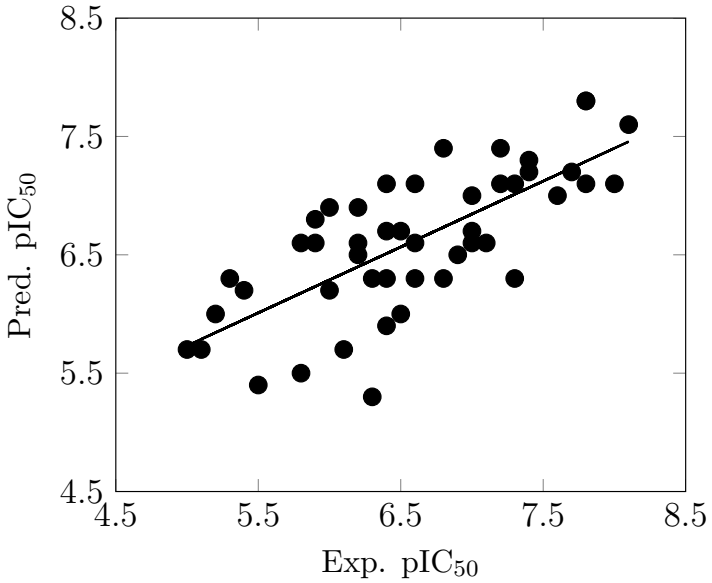


Figure 4.2: The correlation between the experimental and predicted pIC_{50} for the log$S$ model with 4 PLS components (G5 data set); a regression line has been drawn through the data (solid line).

The predicted versus observed activities for G5 data set as well as the residuals are listed in Table 4.2. The predictions of the 46 molecules in Figure 4.2 with the residuals larger than 0.8 are **3**, **8**, **15**, **42**, **45**, and **46**. The predictions of pIC_{50} values are made with only moderate accuracy (q^2 = 0.50, MAE_{CV} = 0.46). Thus, noticeable differences between predicted and observed pIC_{50} values were expected. Large differences between r^2 and q^2 can indicate overfitting.[204] However, the difference

Molecule No.	Exp. pIC_{50}	Pred. pIC_{50}	Residual	Molecule No.	Exp. pIC_{50}	Pred. pIC_{50}	Residual
1	7.0	6.6	-0.4	24	6.2	6.9	0.7
2	6.4	7.1	0.7	25	< 5.0	5.7	0.7
3	7.3	6.3	-1.0	26	6.2	6.5	0.3
4	7.8	7.1	-0.7	27	6.8	6.3	-0.5
5	7.6	7.0	-0.6	28	7.4	7.3	-0.1
6	7.2	7.1	-0.1	29	6.4	6.7	0.3
7	7.8	7.8	0.0	30	7.2	7.4	0.2
8	8.0	7.1	-0.9	31	6.5	6.0	-0.5
9	7.8	7.8	0.0	32	6.3	6.3	0.0
10	6.6	6.6	0.0	33	5.8	5.5	-0.3
11	7.7	7.2	-0.5	34	5.1	5.7	0.6
12	6.4	6.3	-0.1	35	6.0	6.2	0.2
13	7.4	7.2	-0.2	36	6.4	5.9	-0.5
14	6.5	6.7	0.2	37	6.6	7.1	0.5
15	5.3	6.3	1.0	38	8.1	7.6	0.5
16	7.0	7.0	0.0	39	6.8	7.4	0.6
17	6.9	6.5	-0.4	40	5.2	6.0	0.8
18	6.2	6.6	0.4	41	5.9	6.6	0.7
19	5.8	6.6	0.8	42	6.3	5.3	-1.0
20	5.4	6.2	0.8	43	6.1	5.7	-0.4
21	7.3	7.1	-0.2	44	5.5	5.4	-0.1
22	6.6	6.3	-0.3	45	6.0	6.9	0.9
23	7.1	6.6	-0.5	46	5.9	6.8	0.9

Table 4.2: The predicted activities for G5 group from the LOO-CV.

between r^2 and q^2 in our model is quite low (0.16).

4.4.2 2D-QSAR with machine learning analysis

The SMO, SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and REPTree machine regression methods were used as described in the previous chapter. Only 30 out of 105 models afforded a q^2 higher than 0.50, and 65 models had MAE smaller than 0.50 for the LOO-CV test (Table 5.5). The SMO algorithms show models with q^2 values smaller than 0.50, and they performed worse than the other algorithms. These SMO results are quite similar to our results in the previous chapter with the $\alpha_v\beta_1$

inhibitors. For the 46-molecule set, the results indicate that most of the models were successfully predicted. RF models are close to SVM and outperform REPTree models. Our cross-validated 2D-QSAR results showed the log$S$ models as predictive models. The predictive model is presented among the top models; (q^2 of 0.63 in IBk, q^2 of 0.61 in RF, q^2 of 0.55 in both SVM and REPTree, and q^2 value of 0.53 in M5P algorithm). Both RF and SVM methods are characterised with more stability to noise in the input data. The k value in IBk algorithm was calculated automatically by Weka to give the minimum error rate. The MAEs from LOO-CV for the machine learning algorithms are shown in Figure 4.3.

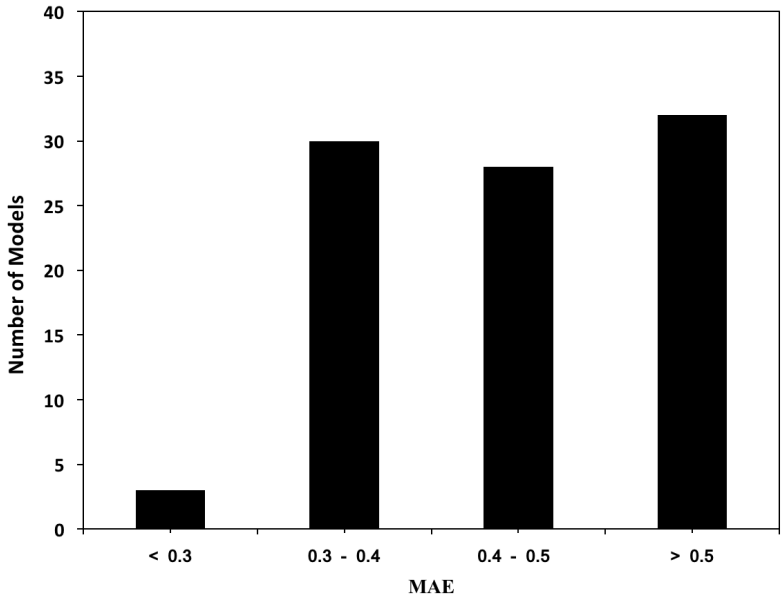


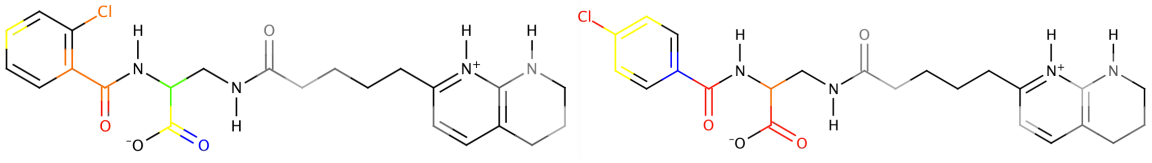
Figure 4.3: Predictive performance (MAE) of SVM (RBFKernel), MLP, IBk, M5P, RF, and REPTree models assessed in LOO-CV on the training set of 90 models: (15 cross-validated models for each machine learning algorithms).

4.4.3 2D-QSAR interpretation

Figure 4.4 shows different molecules from the best predicted model (log$S$ with 4 components and a q^2 of 0.50). For the *ortho*- and *para*-chloro substituents, both molecules contribute negatively to the activity with different magnitude. In the *ortho*-

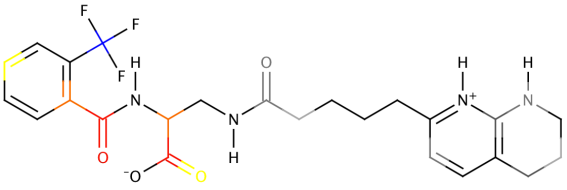
Table 4.3: The most predictive models using the SVM (RBFKernel), MLP, IBK, M5P, RF, and REPTree approaches.

	Training				Validation		
Methods	Model	r^2	MAE	RMSE	q^2	MAE	RMSE
SVM (RBFKernel)	Charge log$S$	0.85	0.14	0.31	0.61	0.39	0.51
	log$S$	0.79	0.24	0.36	0.55	0.43	0.52
MLP	Charge log$S$	0.78	0.40	0.49	0.58	0.39	0.51
	Charge MR	0.69	0.36	0.47	0.57	0.37	0.49
	Charge log$P$	0.88	0.33	0.42	0.56	0.37	0.50
	Charge	0.70	0.39	0.47	0.52	0.39	0.51
	Charge log$P$ log$S$	0.94	0.16	0.19	0.51	0.42	0.53
IBk	log$S$	0.81	0.39	0.49	0.63	0.37	0.47
	Charge log$P$ log$S$ MR	0.76	0.42	0.53	0.60	0.38	0.47
	Charge log$P$ log$S$	0.79	0.39	0.50	0.58	0.39	0.51
	Charge log$S$	0.79	0.42	0.52	0.55	0.43	0.52
	Charge log$P$	0.72	0.43	0.54	0.54	0.39	0.51
	Charge log$P$ MR	0.73	0.43	0.54	0.51	0.42	0.53
	log$P$ log$S$ MR	0.72	0.43	0.55	0.50	0.43	0.53
M5P	log$S$ MR	0.85	0.19	0.30	0.58	0.38	0.52
	log$S$	0.84	0.22	0.31	0.53	0.38	0.50
RF	log$S$	0.84	0.29	0.36	0.61	0.39	0.47
	log$P$ log$S$	0.83	0.30	0.38	0.61	0.37	0.49
	log$P$ log$S$ MR	0.81	0.31	0.39	0.61	0.37	0.47
	Charge log$P$ log$S$ MR	0.86	0.30	0.38	0.59	0.37	0.48
	Charge log$P$ log$S$	0.85	0.31	0.38	0.59	0.39	0.49
	Charge log$S$	0.87	0.31	0.39	0.58	0.40	0.51
	Charge log$S$ MR	0.86	0.31	0.39	0.58	0.39	0.51
	log$S$ MR	0.83	0.30	0.37	0.57	0.37	0.49
	Charge log$P$	0.83	0.32	0.40	0.56	0.37	0.51
	log$P$ MR	0.80	0.31	0.39	0.53	0.39	0.51
	log$P$	0.79	0.32	0.40	0.53	0.39	0.50
	Charge log$P$ MR	0.83	0.31	0.40	0.51	0.42	0.52
REPTree	log$S$	0.71	0.31	0.42	0.55	0.43	0.52
	log$S$ MR	0.83	0.23	0.32	0.54	0.39	0.51

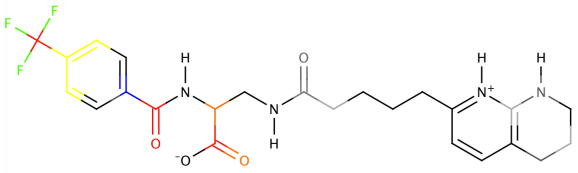


(a) Molecule **10**

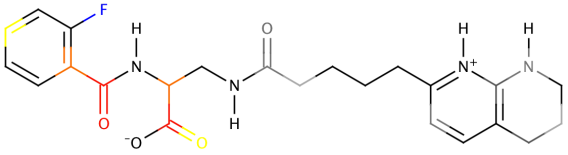
(b) Molecule **43**



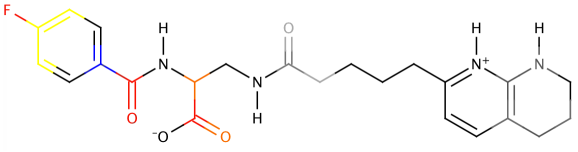
(c) Molecule **14**



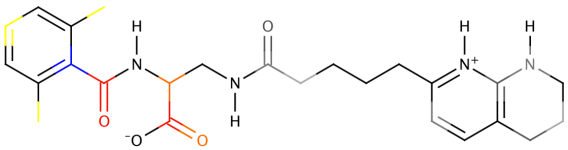
(d) Molecule **15**



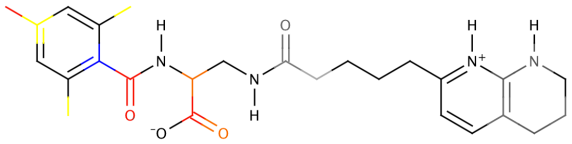
(e) Molecule **19**



(f) Molecule **42**



(g) Molecule **24**



(h) Molecule **21**

Figure 4.4: TMACC interpretation of G5 group for two different models; log$S$ with 4 components (a, b, c, d, e, f, g, and h molecules). (a) The Cl group on the *ortho*-position provides a slight negative contribution to the activity of the molecule. (b) More negative contribution for the Cl in the *para*-position group to the activity. (c) The CF_3 group on the *ortho*-position provides a positive contribution to the activity. (d) The CF_3 group on the *para*-position is shown in red, indicating that it provides a very negative contribution to the activity. (e and f) The F group on the *ortho*-position and *para*-position provides a very positive, and very negative contribution, respectively, to the activity. (g) The *di*-methyl substituents provide a slight positive contribution to the activity. (h) A positive activity-contribution for the *tri*-CH_3 groups to the activity

chloro compound, the Cl moiety contributes negatively to the activity while it shows more negative contribution to the activity but in the case of the *para*-position. It also was found that the *para*-fluoro group provides negative contribution similar to that of *para*-chloro (Figure 4.4f). However, the F group on the *ortho*-position is indicated in blue that means it provides a very positive contribution to the activity. This positive contribution in the case of *ortho*-F is unlike that of *ortho*-Cl. The *ortho*-Cl contributes negatively to the activity as shown in Figure 4.4e and Figure 4.4a. The *ortho*-CF_3 substituent behaves similarity with *ortho*-F substituent with both indicating a very positive contribution to the activity (Figures 4.4c and 4.4e). The trifluoromethyl group on the *para* position in the log$S$ model contributes negatively to the activity.

Figure 4.5 displays molecules **15**, **46** and **14**, the *para*-, *meta*- and *ortho*-substituted compounds, respectively. The atoms that contribute less to activity are coloured in red, and those that most increase activity are given in blue. The contribution of those atoms, which their impact on the activity are not influential, were not shown for simplicity. We also display the influence of the CF_3 group. The activity is increased upon moving from a *para*- and *meta*-substituted phenyl ring to an *ortho* substitution, for CF_3-substituted molecules (Figure 4.5). Hence, CF_3 groups have been identified as activity-increasing substituents.

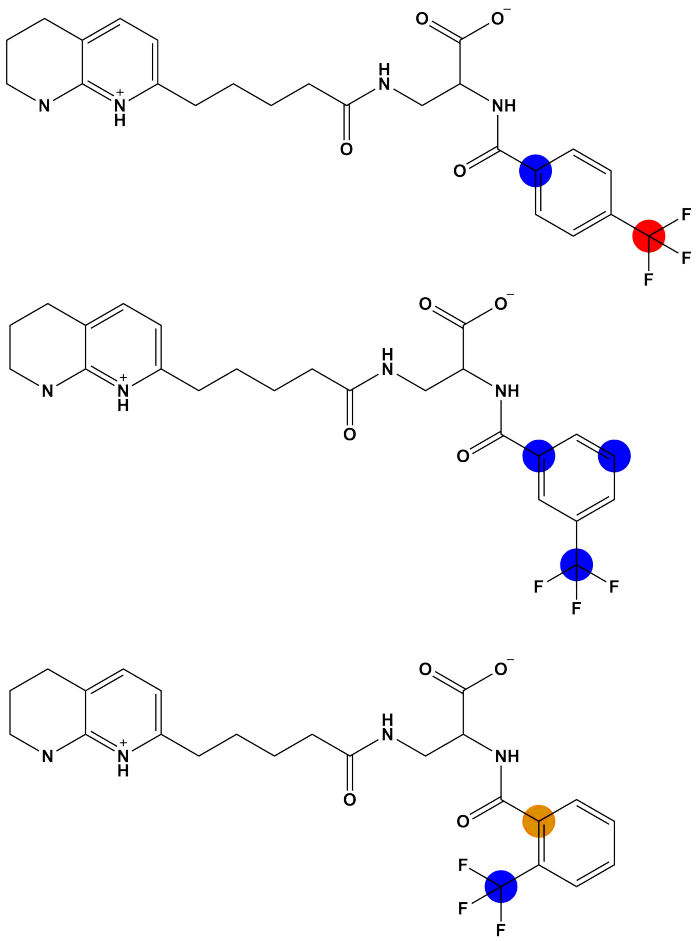


Figure 4.5: Visualisation of the TMACC/PLS log$S$ model of the G5 data set showing the effect of modifying a *para*-CF_3-substituted molecule **15** (top, $pIC_{50} = 5.3$) to a *meta*-CF_3-substituted molecule **46** (middle, $pIC_{50} = 5.9$) and to an *ortho*-CF_3-substituted molecule **14** (bottom, $pIC_{50} = 6.5$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors. The orange colour indicates a slight negative contribution to the activity of the molecule.

Another similar example is the F-substituted molecules (Figure 4.6). When the *ortho*-substituted molecule is studied, the contribution of the *ortho*-F group is coloured in blue; hence, this substituent has been recognised as an activity-increasing group. In the *para*-position, this group is coloured in red which means it was identified as activity-decreasing substituent.

Figure 4.7 shows the different effects of the *ortho*-substituted Cl groups on activity. All Cl groups are coloured in orange. This indicates an activity-decreasing group,

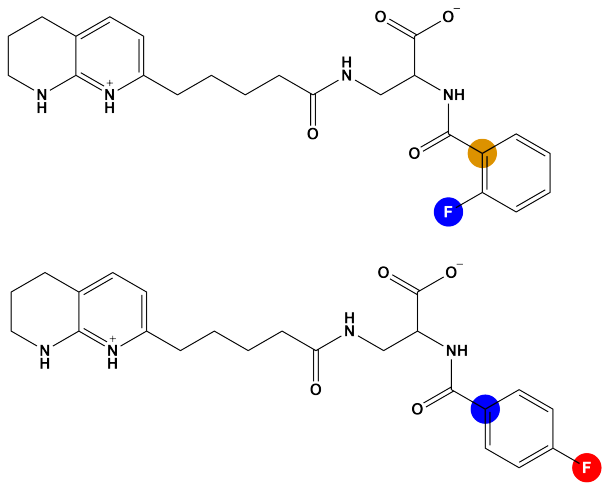


Figure 4.6: Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* F-substituted molecule **19** (top, pIC_{50} = 5.8) to an-*ortho* F-substituted molecule **42** (bottom, pIC_{50} = 6.3). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors.

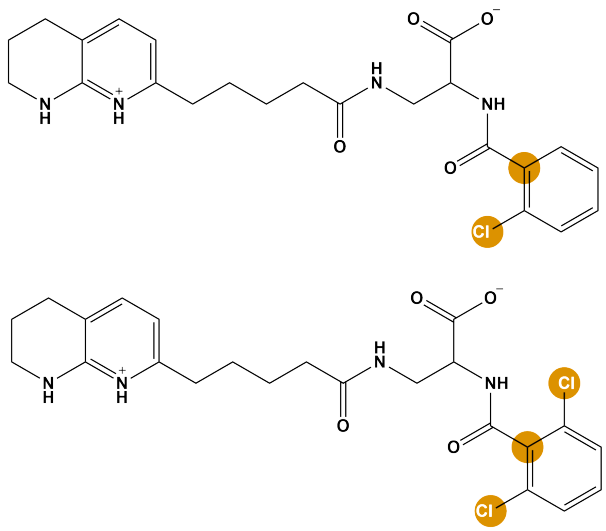


Figure 4.7: Visualisation of the TMACC/PLS model showing the effect of modifying an-*ortho* single Cl-substituted benzamide inhibitor molecule **10** (top, pIC_{50} = 6.6) to a di-*ortho*-Cl-substituted molecule **2** (bottom, pIC_{50} = 6.4). The orange colour provides a slight negative contribution to the activity of the molecule.

but their effects are more moderate. The *ortho* single Cl-substituted molecule **10** is characterised by a pIC_{50} of 6.6 while the di- *ortho*-Cl-substituted molecule **2** has an activity of 6.4. That means the pIC_{50} is reduced by 0.2 log units, which is not significant within the experimental error of the assay. However, this reduced activity

might be due to the extra Cl moiety on the *ortho*-position. This result is in contrast with the previous results in chapter 3, which suggested that when a molecule has two activity-increasing groups such as Cl substituents on the *ortho*-positions of a phenyl ring, their impact on activity will increase (Figure 3.11). This could reflect that there may be other influenced molecular properties beside the four studied in this 2D-QSAR study. Or the lower accuracy of our best model might be lead to this conflict. Or it could be that the β_1 and β_6 subunits have different binding requirements.

The presence of a nitrogen in a heterocyclic benzamide will influence $\alpha_v\beta_6$ activity, especially when the nitrogen atom is at the *para*-position of the phenyl ring. TMACC interpretation of two different inhibitors (Figure 4.8) is shown to explain this point. The first is an inhibitor with high affinity (molecule **28**, top) with a measured pIC_{50} of 7.4, and the other is a less active inhibitor (molecule **40**, bottom), which had a measured pIC_{50} of 5.2.

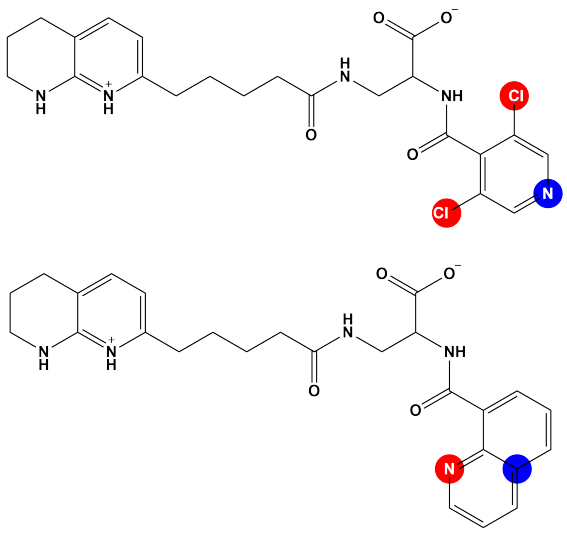


Figure 4.8: Visualisation of the TMACC/PLS model showing the nitrogen contribution to the activity for the highest and the less active inhibitors; The first inhibitor is molecule **28** (top, $pIC_{50} = 7.4$) and the other inhibitor is molecule **40** (bottom, $pIC_{50} = 5.2$). Highlighted atoms indicated those that are most important for decreasing (red) and increasing (blue) the potency of inhibitors.

Figure 4.8 also shows that the TMACC highlights the quinoline 10 carbon atom in

molecule **40**, in blue, as activity-increasing. The electronic properties of nitrogen atom may increase the activity. For example, a 1.0 log unit range in activity against $\alpha_v\beta_6$ (pIC_{50} 6.4 - 7.4) with molecule **2** showing the less potent activity and molecule **28** with a heterocyclic ring the most, suggesting that the electronic properties of the nitrogen atom may increase the activity.

4.4.4 3D-QSAR analysis

4.4.4.1 Evaluating of 3D-QSAR models

We have followed the alignment procedure previously (see chapter 2). The aligned molecules of the G5 data set are presented in Figure 4.9. Molecule **41** was used a reference for alignment of the 3D molecular structures.

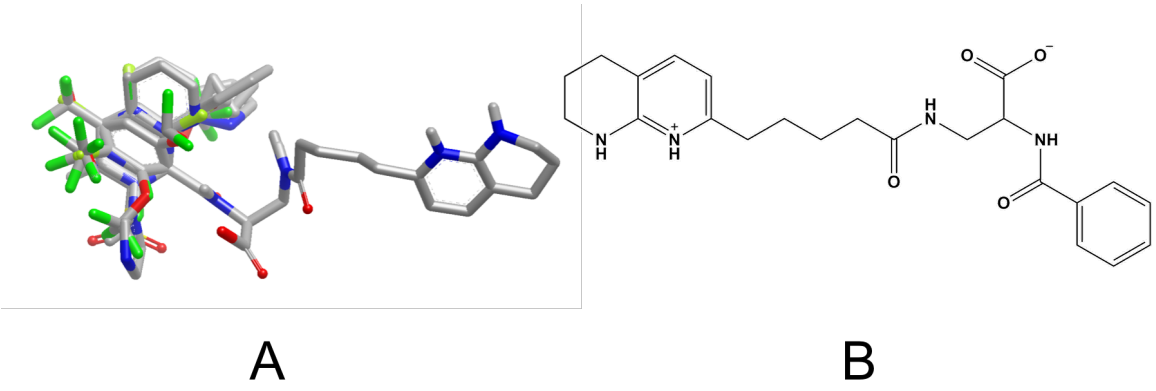


Figure 4.9: (A) Structural alignment of 46 benzamides according to OMEGA. (B) The template molecule (compound **41)**

To generate our models, we have considered all four property types individually: charge, log$P$, log$S$, and MR, and for all possible combinations to yield a total of 15 models. The strongest relationship was log$S$ model with a r^2 and q^2 of 0.32, and 0.12, respectively (Figure 4.10).

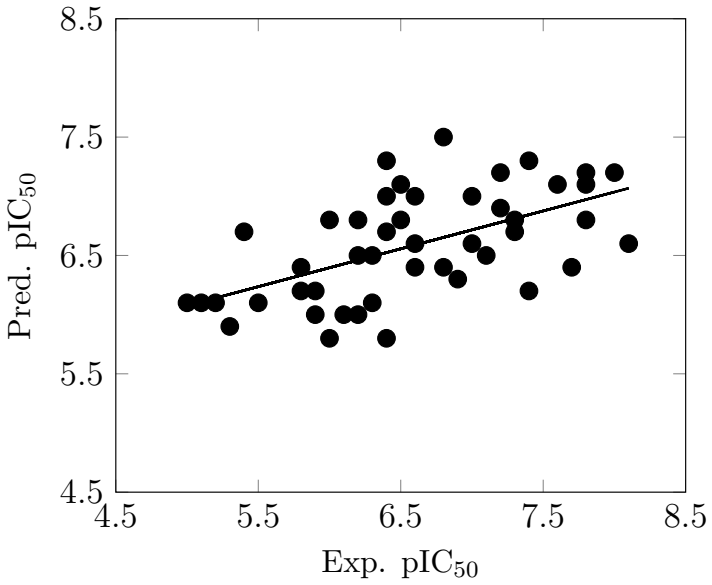


Figure 4.10: The correlation between the experimental and predicted pIC_{50} for the log$S$ (G5 data set); a regression line has been drawn through the data (solid line).

4.5 Conclusions

The present work derived 2D/3D-QSAR models to investigate the effect of each substituent of 46 compounds (G5 data set) on inhibition of the $\alpha_v\beta_6$ integrin. All 2D- and 3D-QSAR models were validated using LOO-CV approach. The NCW software was employed for both 2D- and 3D-QSAR methods.

The 2D-QSAR showed promising predictivity ($r^2 = 0.66$, $q^2 = 0.50$, $MAE_{CV} = 0.46$) and was noticeably better than the 3D-QSAR model. The best model indicates the importance of the log$S$ property. The NCW-2D interpretation using a colour scheme for each property showed key features that influence the activity of the compound. The visualised 2D-QSAR models indicate several positions on the phenyl ring where modifications may increase activity. For example, the activity is increased upon moving CF_3 substituents from a *para*- and *meta*-position to an *ortho*-position in molecules **15**, **46** and **14**. The 2D-QSARs suggest the mono-substitution in the *ortho*-position, (molecules (**14**, **19**) is influential for activity against $\alpha_v\beta_6$.

However, this is not always true. For example, molecules **10** and **26** show a negative contribution to the activity. The mono-substituents in the *para*-position such as molecules **15** and **43** are always gave negative contributions to the activity. This study has confirmed the potential of the TMACC approach for the development of QSAR models.

In contrast, the 3D-QSAR showed weak predictive results ($q^2 = 0.12$). The calculated statistical parameters (low r^2 and q^2 values) illustrated the low predictive ability of the constructed 3D-QSAR models so that no further external validation such as Y-randomisation was performed.

The utility of parametric and non-parametric regressors in the QSARs was also studied to generate predictive models for potent $\alpha_v\beta_6$ inhibitor. The SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and REPTree algorithms were used and validated using the LOO-CV approach. The best non-parametric model (IBk model) achieves a cross-validated q^2 value of 0.63, corresponding to predictive utility. This model suggests the importance of log$S$ to the activity.

Chapter 5

DFT Study of Conformations of Benzamides

5.1 Introduction

Benzamide derivatives have been investigated in the pharmaceutical industry and in diagnosing malfunctions in dopaminergic neurotransmission.[205] They have also been used widely in the treatment of diabetes, limiting tumour growth, and in the treatment of IPF.[206] The amide moiety is one of the most important functional groups in medicinal chemistry. Peptides and proteins are composed of subsets of these amides. In benzamides, the interaction between the amide and the phenyl groups is of fundamental interest. In QSAR studies, the biological responses elicited by drugs are often correlated with a combination of physicochemical properties. From chapter 3 and chapter 4, we have noticed how drug–receptor interactions have dependencies on factors such as log$S$, log$P$ and MR. QSAR methods can use properties calculated by quantum chemical methods (chapter 2). Quantum chemical properties can help find reliable relationships between descriptors and the biological activity for a series of compounds so that these rules can be applied to predicting new chemical

entities. A large number of molecular quantities and well-defined physical information can be defined using quantum chemical approaches such as geometry, binding properties, charges, energies of the highest occupied molecular orbital (HOMO), energies of the lowest unoccupied molecular orbital (LUMO), HOMO-LUMO energy E_{Gap}, etc. Many quantum descriptors are introduced and discussed in detail elsewhere.[207–209] Although there is some natural overlap between experimental measurements and quantum chemical descriptors, the latter do not have statistical errors. Instead, the quantum descriptors have constitutional errors. These inherent errors are linked to the assumptions required to serve the calculations, and are considered approximately constant throughout the group of molecules.[207]

Calculating quantum chemical descriptors is time-consuming. Yet, the rising computational power of hardware and reducing computational cost motivates chemists to calculate various quantum descriptors for a wide variety of datasets. It is worth mentioning that only a small subset of calculated descriptors provides valuable information with which to build a successful model. However, descriptor selection is a challenging task.

5.1.1 A brief view of Density Functional Theory and Coupled Cluster methods

Density functional theory (DFT) was developed by Hohenberg and Kohn in 1964[210] and has been used for calculating a vast number of quantum and mechanical descriptors. DFT is a kind of electronic structure calculation used to obtain the energetics, structure and properties of atoms and/or molecules at lower cost than traditional *ab initio* wave function techniques. There are many electronic structure theories besides DFT, e.g. the Hartree–Fock (HF) method,[211,212] Møller–Plesset perturbation theory (MP),[213] configuration interaction (CI),[214] etc. The critical difference between these theories and DFT is how each theory treats electron–electron interactions. HF, MP

and CI methods rely on computation of the atomic wave function, while DFT does not require a wave function. Instead, DFT requires computation of the total electron density.

Coupled cluster with single and double excitations (CCSD)[215] method is useful for the study of small systems; however, for large molecules, the use of high-level CCSD approach is costly. CCSD level geometry optimisation was performed on molecule **41** to compare the accuracy of other methods. In terms of the least accurate and low cost methods, HF treats electron–electron interactions in a very approximate way. DFT provides higher accuracy than HF calculations, but at a higher computational cost than HF and a lower cost than CCSD. DFT become a powerful method for studying large biologically relevant structures.

In this brief review we summarise the enormous potential of DFT and how DFT can calculate quantum mechanical descriptors, which are very useful in QSAR studies.

The Schrödinger wave equation is the basic starting point for electronic structure theory.

$$\hat{H}\Psi = E\Psi \tag{5.1}$$

where $\hat{H}$ is the Hamiltonian operator. Ψ is the wave function, which describes the system of particles. E the total energy of the system. The Hamiltonian consists of two main parts: the kinetic and potential energy operators ($\hat{H} = \hat{T} + \hat{V}$). For a Coulombic system when invoking Born-Oppenheimer approximation, $\hat{H}$ is given by

$$\hat{H} = -\sum_{i}^{N} \frac{1}{2}\nabla_i^2 - \sum_{iA}^{N} \frac{Z_A}{|r_i - r_A|} + \sum_{i>j}^{N} \frac{1}{r_{ij}} \tag{5.2}$$

Z_A represents a number of nuclei with charges Z_A, and positions r_A. The number of electrons is N. r_i represents the position of electron i and r_{ij} is the distance between electrons i and j. $\hat{T}$ in classical mechanics can be expressed in terms of the momentum: $\hat{T} = \frac{1}{2}mv_x^2$; in quantum mechanics, the momentum is replaced by the momentum operator: $mv_x = -i\hbar\frac{d}{dx}$. Then, $\hat{T} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2}$

The beginning of DFT dates back to the Hohenberg–Kohn theorem, which states that potential energy is a functional of the ground-state density. Later on, Kohn and Sham (KS)[216] simplified the ground-state energy of DFT

$$E[\rho] = T_s[\rho] + V_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{5.3}$$

where each functional was expressed explicitly in equations: 5.4-5.8. $T_s[\rho]$ is the kinetic energy of the electrons, $\rho(r)$ is the electron density, $V_{ne}[\rho]$ represents the potential energy of the nucleus and electron due to their Coulombic attractions, $J[\rho]$ is the classical electron–electron repulsion energy, and, finally, the $E_{xc}[\rho]$ term stands for the exchange–correlation energy.[217,218]

$$T_s[\rho] = \sum_i \left\langle \phi_i \left| -\frac{1}{2}\nabla^2 \right| \phi_i \right\rangle \tag{5.4}$$

$$\rho(r) = \sum_i |\phi_i(r)|^2 \tag{5.5}$$

$$V_{ne}[\rho] = \int \rho(r) v(r) dr \tag{5.6}$$

$$J[\rho] = \frac{1}{2} \int \int \frac{\rho(r)\rho(r')}{|r - r'|} dr dr' \tag{5.7}$$

$$E_{xc}[\rho] = (T[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{5.8}$$

The $E_{xc}[\rho]$ term corrects the overly simplistic $J[\rho]$ term, which is based on the average repulsion between electrons. The beauty of DFT is its simplicity. However, its simplicity should not become a thoroughly empirical method, nor should it be as complicated as the full configuration interaction (FCI) approach. One of the sim-

plest functionals was local density approximation (LDA), which performed well with geometries for small molecules. DFT was used convincingly when the seminal work of Becke, Perdew, Langreth, and Parr introduced the first derivative of the density in the form of the generalised gradient approximation (GGA).[219,220] The following significant improvement came with the addition of a fraction of Hartree-Fock exact exchange in the functional, as described by Becke in the 1990s.[220,221] Later on, the B3LYP[222,223] was introduced, the most extensively employed of all the functionals.[220]

5.1.2 Basis set

A basis set, in general, is a set of functions called basis functions that are used to represent the electronic wave function. Molecular orbitals are made as a linear combination of atomic orbitals (LCAO):

$$\psi_i = \sum_{n=1}^{n} c_{in}\phi_n. \tag{5.9}$$

where ψ_i is the molecular orbital, c_{in} are the coefficients of linear combination, ϕ_n is the atomic orbital, and n is the number of atomic orbitals. There are two types of basis functions commonly used in electronic structure calculations: Slater Type Orbitals (STO) and Gaussian Type Orbitals (GTO). STO are mainly used for atomic and diatomic systems where high accuracy is needed, and they have the functional form shown in equation 5.10

$$\psi^{STO}_{a,b,c}(x,y,z) = Nx^a y^b z^c e^{-\zeta r} \tag{5.10}$$

Here N is a normalisation constant, a, b, and c control angular momentum ($L = a + b + c$), and x, y, z, and r measure a nucleus-electron distance. ζ represents width of orbital: large ζ value indicates a tight function while small value gives diffuse function. The GTO basis functions for molecules are most often chosen to be cartesian

Gaussian functions and can be shown in equation 5.11

$$\psi_{a,b,c}^{GTO}(x,y,z) = Nx^a y^b z^c e^{-\zeta r^2} \tag{5.11}$$

STO basis functions are more accurate than GTOs, but they are time-consuming. This difficulty can be resolved by using a linear combination of n numbers of GTO's for achieving a certain accuracy compared with STO. This linear combination is often called STO-nG. The basis set employed has usually been the 6-31G* family by Pople et al.[224] This basis set is a Gaussian-type orbital which is the most often used method. The following basis sets were also used:[225–227] 6-311G*, 6-311G** and 6-311+G*.

In this chapter, we focus on the geometry of each benzamide derivative when binds to the $\alpha_v\beta$ inhibitor. Our hypothesis states that the biological activity of each molecule would vary based on the rotation of the phenyl ring which has different substituents. To achieve that, the DFT was employed to calculate different molecular descriptors. We also investigate the potential energy surfaces (PES) for each molecule of our datasets. The correlations between the molecular descriptors and activities were studies using QSAR models with several machine learning methods.

5.2 Computational Details

5.2.1 Datasets and molecular descriptors

The datasets comprise the G4 and G5 (chapter 3) sets of molecules. In chapter 2, we have discussed in detail molecular descriptors and their importance in building useful QSAR models. In this section, we will focus on six quantum molecular descriptors that have been used to build our QSAR models.

5.2.1.1 Boltzmann population of conformations

Benzamides have many conformational degrees of freedom. They also have dynamic and thermodynamic properties, which can be determined by the nature of their PES. Conformational analysis is used to find stable structures of conformers. PES can be characterised by their minima, which correspond to stable configurations, and by the transition state regions or the saddle points which determine the minima. The Boltzmann equation can be written in term of the fractional population f and energy E as follows:

$$f = \exp(-\beta E) \tag{5.12}$$

where β is the Boltzmann constant and has its usual form:

$$\beta = \frac{1}{kT} \tag{5.13}$$

The Boltzmann fractional population (f) is an important statistical thermodynamic concept in which the numbers of molecules in each state of a system can be calculated at any given temperature. The Boltzmann relationship can be written as

$$N_i = \frac{Ne^{-\beta E_i}}{q} \tag{5.14}$$

N_i is the number of molecules in a specific state, i with energy E_i, N is the total number of molecules, T is the absolute temperature in Kelvin, q is the molecular partition function, and k is a constant, 1.381×10^{-23} J/K.

The molecular partition function q can be defined in 5.14 as follow:

$$q = \sum_{i=1}^{N_{states}} e^{-\beta E_i} \tag{5.15}$$

The first exponential relation of the form of 5.12 for side-chain torsion angles of

residues with aromatic side chains was proposed by Pohl.[228] Such relations have subsequently been considered for residues in secondary structures.[229] The distribution of benzamide conformations, in this study, is derived from energies calculated with DFT.

5.2.1.2 Atomic charges

Due to the importance of electron densities and charges in many physicochemical properties, charge based descriptors have been used as indices of intermolecular interactions. In this study, two separate charge analysis methods were used: the Mulliken charges calculated with the default setting in Q-Chem; and the electrostatic-potential-based approach (ChelpG), wherein grid points with a smoothing function are used to generate the molecular electrostatic potential. Charges derived from the electrostatic potential using a grid-based approach (ChelpG)[230,231] is a widely used approach for estimating atomic charges from the electrostatic potential. A molecule is placed in a centre of a cube, then atomic charges are fitted to generate the molecular electrostatic potential (MEP) at a number of points around the molecule (fitting approach). It uses multi-spaced points around the investigated molecule at a distance of 2.8 Å away from the nuclei. To avoid large distortions due to the nuclei, the points that are located inside the molecular van der Waals radius are excluded from the fitting approach.[232]

Martin and Zipse[233] have used different quantum mechanical methods and basis sets to investigate the charge distribution in the water molecule. The atomic charges for the oxygen atom in a water molecule and the correlated dipole moment were calculated by HF, DFT and MP2 methods. The results were compared to the experimentally measured water dipole moment. They found that ChelpG charge distribution can be predicted using minimal Pople basis sets such as the 6-31G* basis set. However, choosing larger correlated-consistent basis sets with the B3LYP functional

would be more reliable. The B3LYP calculations predict consistently smaller oxygen charges compared to the HF method.[233] The Mulliken charges cover a very broad range of values from -0.35 to -0.87 at the B3LYP.[233] In this study, we calculated the ChelpG charges for each molecule. We have focused on calculating the net atomic charges of the aromatic ring, which is the total net charge on each phenyl moiety and use that as a molecular descriptor.

5.2.1.3 Molecular orbital energies

The highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) are frontier molecular orbitals (FMO)[234] the energies of which direct the chemical reactivity of chemical compounds. They have been used as a simple indicator of kinetic stability[207,208,235] in which a small HOMO-LUMO gap (E_{Gap}) designates low kinetic stability and high chemical reactivity. The HOMO energy (E_{HOMO}) is linked to the ionisation potential and describes the sensitivity of the molecule toward attack by electrophiles.[207] The LUMO energy (E_{LUMO}) is associated with the electron affinity and defines the sensitivity of the molecule toward attack by nucleophiles.[207]

5.2.1.4 Dihedral angle

Our attention was focused on the torsional, (ψ_1, C_1-C_2-C_3-N_4), preferences for each structure in G4 and G5 data sets. For all of the systems considered in this work, there is one symmetrical minimum and two asymmetrical minima. For some molecules such as **1**, **2**, **7**, **21**, **24**, and **27**, the dihedral angle is 180° at the symmetrical minimum and around +90° and -90° at the two asymmetrical minima.

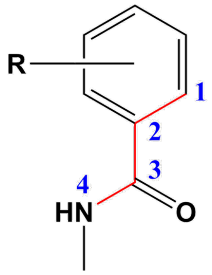


Figure 5.1: The benzamide family, showing the scanned dihedral angle (ψ_1, C_1-C_2-C_3-N_4).

We started our calculations on *N*-methylbenzamide (Figure 5.1) by performing full geometry optimisation using different levels of theory. The structures were optimised using MP2, DFT, with the B3LYP functional, and CCSD approaches. The 6-31G* basis set with polarisation functions, which is common and often considered the best compromise between accuracy and speed, was adopted throughout this work. All calculations were performed with Q-Chem[236] and the results were compared with the experimental X-ray crystallographic data.[237,238] Once the molecular descriptors had been calculated, they served as independent variables in the construction of the QSAR model.

5.2.2 Potential Energy Profiles of Methyl-Substituted Benzamide

Relaxed conformational energy scans at the B3LYP/6-31G* level of theory were carried out on the fragmented molecules G4 and G5 datasets. This PES involves the rotation around the dihedral angle ψ_1, Figure 5.1, in order to locate the minima on the surface. The torsion angle was scanned from -180° to 165° at 15° intervals. Starting from SMILES[239] strings and SDF files, the PESs for *N*-methylbenzamide derivatives were obtained by constraining ψ_1 to a series of values and fully optimising the remaining geometrical degrees of freedom. The influence of substituents on conformational stability was investigated and we determined the most stable geometrical structure

for each compound. DFT was used to calculate the energies of the highest occupied molecular orbital (E_{HOMO}), the lowest unoccupied molecular orbital (E_{LUMO}) and the gap energy (E_{Gap}) between them. The harmonic vibrational frequencies were calculated at the same theory level so as to confirm that minima had been detected (no imaginary frequencies were seen and all frequencies were real). The (E_{HOMO}), the (E_{LUMO}) and the (E_{Gap}) energies were used as molecular descriptors in addition to the net ring charge from ChelpG (δ), the torsional angle of the global minimum (ψ) and the Boltzmann population of the global minimum (f), to build our regression models.

5.2.3 Machine Learning Methods

MLR was used to investigate the possible correlation of input parameters and the biological activity. We have built 126 different MLR models for both the G4 and G5 datasets using 63 descriptor combinations. All possible combinations were generated using a simple Python3.6 script, and the MLR models were built using Weka 3.9.1.

The previous six algorithms, chapter 2, were employed on both the G4 and G5 datasets. For each algorithm, 63 models were built using all combinations of the six descriptors. Six different algorithms were used to build 441 models in total. The predictive performances of the models have been assessed using the leave-one-out (LOO) procedure.

The default setting of Weka were employed for all algorithms except the MLP neural networks. Three main steps, as mentioned in Chapter 2, are required to construct the MLP: choosing the radial basis function centroids, selecting the suitable radius value, and determining the number of hidden layers. The neural network created for this work was composed of an input layer, with the number of hidden layer nodes being equal to that of the training set (e.g. 5 nodes in G4 and G5 datasets, respectively), and an output layer with only one normalised output node.

The quantum chemical descriptors were used as inputs. The input layer has six input nodes (six quantum parameters).

5.3 Results and discussion

5.3.1 Optimised parameters

Several levels of theory were employed on molecule **41** (Table A.1), as a primary example for our dataset, so as to compare the speed and accuracy of the methods. The optimised geometries are presented in Table 5.1. The optimised geometries at the B3LYP/6-31G* level showed good agreement with the crystal structure conformations.[237,238] As a result, the B3LYP/6-31G* was used in this work to calculate the PESs for the remainder of our molecules.

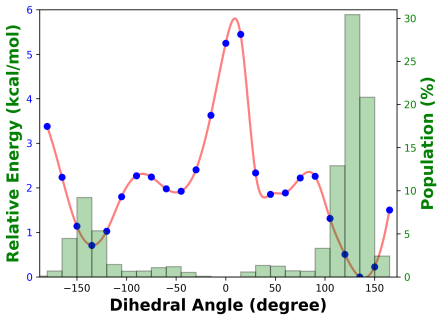
Parameter	Gas-phase			Crystal
	MP2/6-31G*	B3LYP/6-31G*	CCSD/6-31G*	X-ray (Ref.[237,238])
$r_{C_1-C_2}$ / Å	1.39	1.40	1.39	1.38
$r_{C_2-C_3}$ / Å	1.49	1.50	1.50	1.50
$r_{C=O}$ / Å	1.23	1.23	1.22	1.23
$r_{C_3-N_4}$ / Å	1.36	1.37	1.36	1.34
$r_{N_4-C_3}$ / Å	1.45	1.45	1.45	1.46
$< C_1-C_2-C_3$ / °	117.5	117.4	117.6	118.0
$< C_2-C_3-O$ / °	122.2	121.8	122.0	121.0
$< C_2-C_3-N_4$ / °	115.4	116.2	115.8	117.0
$< N_4-C_3-O$ / °	122.2	121.9	122.2	122.0
ψ_1 / °	28.5	22.1	27.5	22
ψ_2 / °	175.1	175.8	174.8	175

Table 5.1: Calculated and experimental geometric parameters at the optimised minimum for the *N*-methylbenzamide at different levels of theory. ψ_2 = C_2-C_3-N_4-C. Total energies at the minimum: $E(MP2) = -438.912476$ au: $E(B3LYP) = -440.2616384$ au, and $E(CCSD) = -438.963912$ au

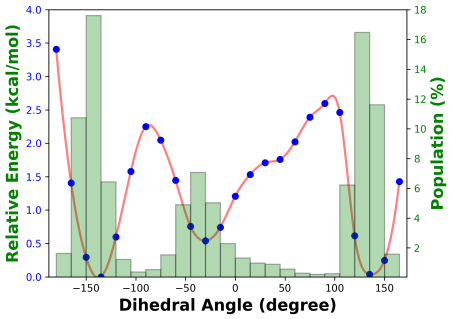
5.3.2 The PES profiles for the G4 and G5 datasets

5.3.2.1 The PES curves for *ortho*-alkyl substituents

Additional relaxed PES scans, computed at the B3LYP level of theory and several split-valence basis sets with polarisation functions were explored:[240,241] 6-31G*, 6-311G*, 6-311+G* and 6-311G**, to investigate the variation in the PES between 2-isopropyl-*N*-methylbenzamide **20** (Figure 5.2a) and 2-cyclopropyl-*N*-methylbenzamide **22** (Figure 5.2b), respectively. Molecule **22** has a pIC_{50} of 7.5 against $\alpha_v\beta_1$, which is 0.9 larger than for molecule **20**. These molecules are sterically similar but electronically different. Thus, we need to explain the effect of such substituents on the PES and activity. The results of calculations were fairly sensitive to basis sets (as discussed later) in the case of **20** and insensitive in the case of **22**. The steric interactions between the N–H bond and the phenyl hydrogens in **20** and **22** led to asymmetrical PES curves.



(a) Molecule **20**



(b) Molecule **22**

Figure 5.2: PES scans of molecules **20** and **22**

We expected to see similar PESs for **20** and **22**, but they were different. Therefore, the PESs were calculated using a variety of different basis sets: 6-31G*, 6-311G*, 6-311+G* and 6-311G**. The energy differences (ΔE) are the calculated energy differences between the energy at the global minimum (E_{GM}) and the local minimum energy (E_{LM}), and are shown in Tables 5.2 and 5.3, respectively.

Property	B3LYP			
	6-31G*	6-311G*	6-311+G*	6-311G**
$\psi_{GM}/^{\circ}$	+150	+135	+135	+135
$\psi_{LM}/^{\circ}$	-135	-135	-120	-135
ΔE / kcal/mol	0.76	0.56	0.36	0.60

Table 5.2: Locations of local (LM) and global (GM) minima and relative energies for molecule **20** using the B3LYP level of theory and different basis sets.

Property	B3LYP			
	6-31G*	6-311G*	6-311+G*	6-311G**
$\psi_{GM}/^{\circ}$	-135	-135	+135	+135
$\psi_{LM}/^{\circ}$	+135	+135	-135	-135
ΔE / kcal/mol	0.04	0.03	0.09	0.02

Table 5.3: Locations of local (LM) and global (GM) minima and relative energies for molecule **22** using the B3LYP level of theory and different basis sets.

A global minimum at +150° was determined for molecule **20** and a maximum at +15°. The barrier height of **20**, relatively to **22**, was shifted upwards to 5.8 kcal/mol at 15°. The global minimum of **20** was characterised by a non-planar geometry wherein the N–H bond was in an opposite direction to the isopropyl moiety. Both CH_3 groups at the minimum are twisted. The geometry of **20** reflects the large steric effect when the C=O/phenyl torsion angle is close to planarity. Molecule **22** shows that the N–H bond of the amide group is aligned in the same plane of the benzene ring and the carbonyl bond is trans to the cyclopropyl ring. At the minimum for compound **22**, the steric strain is minimised by rotating the C=O/phenyl torsion angle to -135°. By comparing the calculated energy differences, ΔE between **20** and **22**, the effect of H atoms might play an important role by shifting the energy differences downwards to 0.36 kcal/mol between the two minima in the case of **20** (Table 5.2). On the other hand, as in the case of **22**, such an effect is small and can be neglected ($\Delta E < 0.1$ kcal/mol, Table 5.3). The global minimum for **23**, Figure 5.3, was located at -135° with a twisted geometry, in which the C=O group was trans to the benzene ring in

order to minimise the steric effect. The ethyl group at the minimum tends to be in-plane and oriented away from the carbonyl bond.

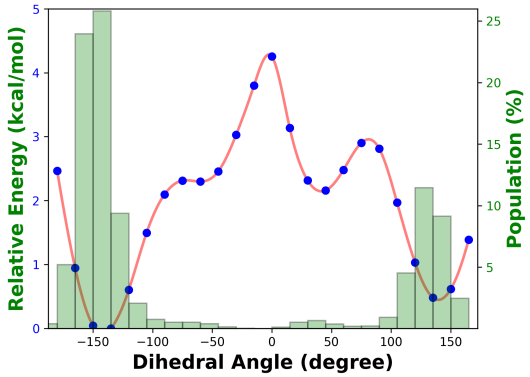


Figure 5.3: A histogram of calculated torsion angle distribution of molecule **23** with the calculated B3LYP/6-31G* PES.

All PESs for both datasets are presented in Figure A.1. The histograms and relative PES energies were generated for all molecules. The energy at each i on the PES is subtracted from the energy at the global minimum. All histograms are superimposed on the calculated B3LYP/6-31G* torsional profiles of the corresponding model compounds (Figure A.1).

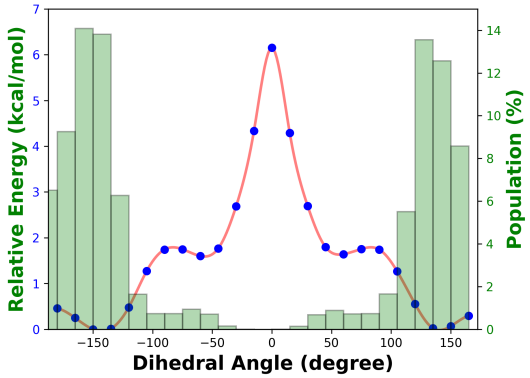


Figure 5.4: A histogram of calculated torsion angle distribution of molecule **2** with the calculated B3LYP/6-31G* PES.

Figure 5.4 presents the calculated energy curve and the histogram for molecule **2**.

The energy curve has four minima, corresponding to the minimum at -165° and +165°, and the other minima at approximately +65° and -65°. In general, the calculated Boltzmann populations are bimodal through the PES curve. However, the population spreads across the surface when the energy barrier between two minima becomes low (Figure 5.5)

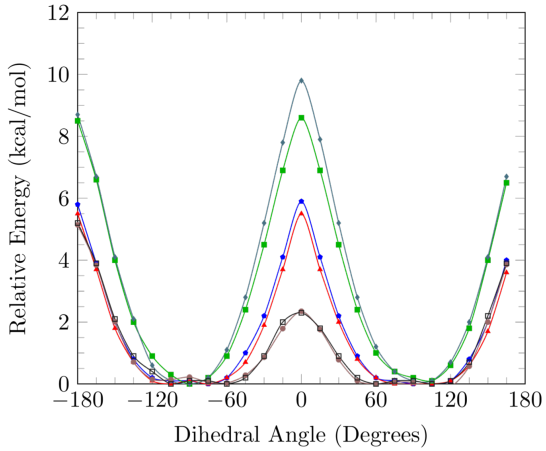


Figure 5.5: B3LYP/6-31G* PES for six molecules with different moieties where molecules: **2** in green solid-squares, **3** in brown circles, **4** in black squares, **21** in red triangle, **24** in blue pentagons, and **27** in cyan diamonds.

Figure 5.5 shows that molecules **2**, **3**, **4**, **21**, **24** and **27** can adopt low energy conformations near the twisted geometries at -90° and +90° approximately. They were predicted to possess unfavourable orientations at 0°.

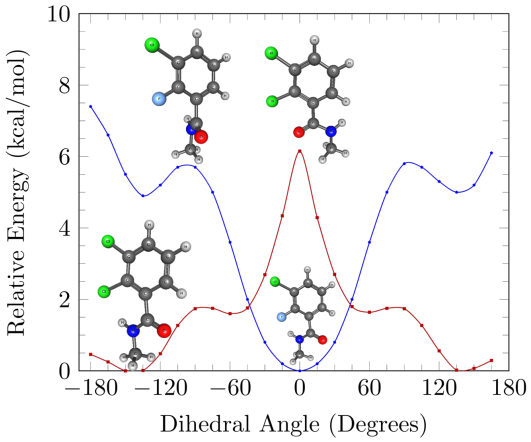


Figure 5.6: B3LYP/6-31G* PES for molecule (**9** in red) and (**13** in blue). The 0° orientation is the global minimum for **13** and a maximum for **9**.

The O-Cl syn conformation ($\psi = 0°$) for molecule **9** is predicted to lie 6.3 kcal/mol higher in energy above its -150° minimum (Figure 5.6). Conversely, molecule **13** would increase the population of the (0°) conformer via replacing the *ortho*-chloro with *ortho*-bromo ($\psi = 0°$; Figure 5.6). Molecules **1**, **21** and **24** have symmetrical PESs with two twisted geometries at the global minima, in which 98.5–99.8% of the population is located at (±105°) for **21** and **24**, and (±95°) for **1**. The twisted geometries for **21**, **24** and **1** can be rationalised by the large steric effect of the methyl and chlorine substituents on the *ortho*-positions. Compounds **21** and **24** have very similar PESs, which means that the extra methyl group on the *para*-position for **21** has an insignificant effect on the energy profile. The atomic ring net charges varied from -0.43 to -0.36 a.u. for **21** and **24**, respectively. Adding a methyl substituent at the *ortho*-position decreases the number of minima to only two distinctive minima at ±105° and the maximum is located at 0°. The steric effect which resulted from the addition of methyl groups on *ortho*-positions was not affected by more methyl groups to the *meta*- and/or *para*-substituents on the PES. The steric effect can be minimised in **24** and **21** by aligning the amide group perpendicular to the aromatic ring. The global minima turn out to be planar geometries for **13** and **19** and are located at ψ_1

= 0° and ψ_1 = -180° for **13** and **19**, respectively, with almost 100% of the population. This planarity might occur due to the possibility of a hydrogen bond between the NH and the fluorine. The strength of the hydrogen bond donation and the polarity of the fluorine increase the chance of making the hydrogen bond for both molecules.

5.3.2.2 The delocalisation effect on the minimum geometry

The global minima for 2-methoxy-*N*-methylbenzamide and molecule **44** are located at $\psi_1 = 0°$ and ±15°, respectively, with a very high percentage of the population located there. The distribution at these global minima is due to the orbital delocalisation between the methoxy group (-OCH_3) in the *ortho*- and *para*-positions of the amide group with the amide moiety itself, and the phenyl ring (Figure 5.7). The calculated and experimental distances between the methoxy protons and N–H bond are 3.10 Å (this work) and 2.97-3.06 Å based on the NMR study of Liu et al.[242] Our calculations agree with this report and confirm the presence of a persistent intramolecular H bond between the methoxy and the N–H moieties.

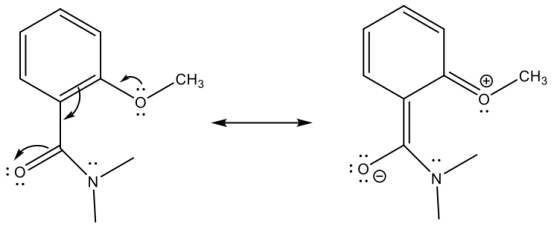


Figure 5.7: Resonance structure, showing the orbital delocalisation between the methoxy, the phenyl ring and the amide group in 2-methoxy-*N*-methylbenzamide.

We have used an external molecule, which is not in G4 or G5 datasets, as an example (2-cyano-*N*-methylbenzamide) to confirm whether delocalisation is the reason behind the planar geometry. Both PESs for 2-methoxy- and 2-cyano-*N*-methylbenzamides show planar geometry is the global minimum energy (Figure 5.8).

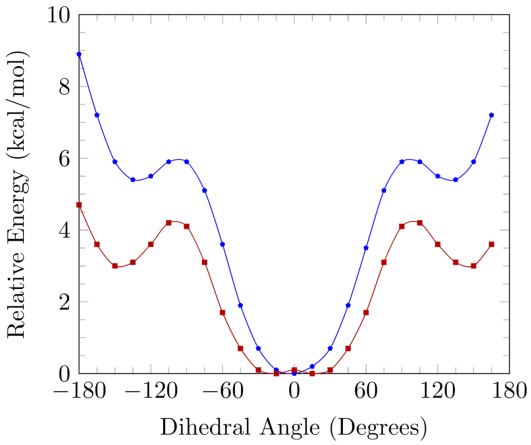


Figure 5.8: Two examples of PES profiles for 2-methoxy- and 2-cyano-*N*-methylbenzamides show the delocalisation effect on the stability of the global minima. They are presented in blue pentagons and red squares, respectively.

5.3.2.3 Effect of halogens on internal rotation of aromatic amide

We considered halogens in three different positions — *ortho*-, *meta*- and *para*- — on the benzene rings. In the *ortho*-positions, the energy barrier varied from 5.8 kcal/mol in **10** to 7.4 kcal/mol in **19**. Two global minima were determined for each molecule: at -180° for **19**, ±150° for **10**, and ±105° for **1**. The global minima are planar for **10** and **19** where the amide hydrogen interacts with the halogens, and two distinctive twisted geometries were also located at each local minimum. The global maxima were planar for all of the mentioned derivatives. At each maximum, the repulsion forces are larger due to the steric interactions between the oxygen and the halogen. For molecule **19**, the global minimum was 5.6 kcal/mol lower than the local one. The local minimum for **10**, on the other hand, was only 2.2 kcal/mol higher than the global minimum. By comparing **10** and **19**, it was found that the large size of the chlorine atom prevents **10** from being planar at the global minimum and shifts the torsional angle to -150°.

We have also noticed that the atomic size of the substituent on the *ortho*-position plays a critical role in the energy barrier. For instance, a general PES comparison

between molecules **5** and **27** was performed. The large atomic size for bromine leads to approximately double the energy barrier ($\Delta E_{\mathbf{5}} = 5.3$ kcal/mol for **5** and $\Delta E_{\mathbf{27}} = 9.8$ kcal/mol for **27**) when compared with chlorine (Figure 5.9).

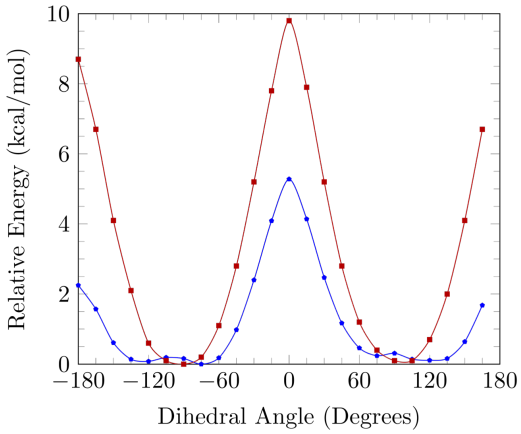


Figure 5.9: The B3LYP/6-31G* PES curves for molecules **5** and **27** show the atomic size effect of a substituent on the energy barrier. Molecule **5** is shown in blue pentagons and molecule **27** is in red squares.

When substituting different halogens — Cl^-, Br^- and F^- — in the *meta*- and *para*-positions, we found that the substituents in these positions did not affect the PES much, as expected. Less than a 0.02 kcal/mol change in the barrier height was calculated among them. The PES curves for molecules **12**, **15**, **18** and **25** confirm the insignificant influence of *meta*- and/or *para*-substituents. Regardless of the size of the substituent in *meta*- or *para*-positions to the amide bond: *p*-Cl and $m-CF_3$ in **12**, $p-CF_3$ in **15**, $m-F$ and $p-F$ in **18**, and ($m-OCH_3$ and $p-OCH_3$) in **25**, the effect on the PES is insignificant (see the PES profiles for molecules **42** and **43** Figure A.7)

5.3.2.4 Mulliken vs. ChelpG

As expected, the Mulliken charges showed a broader range of values when compared with the ChelpG charges. For example, it was found that the carbonyl oxygen charge

(using Mulliken) varied from -0.47 to -0.57 a.u. In ChelpG, the relative charge varied from -0.45 to -0.54 a.u. among the G4 and G5 molecules. Figure 5.10 shows a comparison between the two different population methods. Molecule **43** was chosen as an example to show the charge distributions for 19 atoms in the molecule. We can see that ChelpG shows a narrower distribution of atomic charges against the Mulliken method.

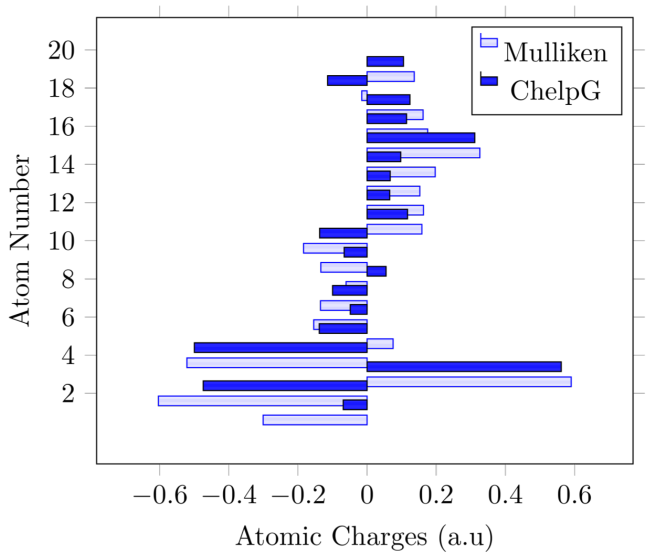


Figure 5.10: Atomic partial charges calculated with the Mulliken and ChelpG methods for molecule **43**

5.3.3 QSAR with machine learning analysis

5.3.3.1 MLR results

MLR was used to investigate the possible correlation of input parameters and the biological activity. The MLR results are shown in Table 5.4. From Table 5.4 it can be seen that there is no simple linear correlation between the predicted and experimental activities. The most predictive models were (ψ, f, and E_{Gap}) and (ψ, f, and E_{HOMO}) models with q^2 values of 0.20 and 0.13 for activity against the $\alpha_v\beta_1$ and $\alpha_v\beta_6$ integrins, respectively.

Compound	δ	ψ	f	E_{HOMO}	E_{LUMO}	E_{Gap}	$\alpha_v\beta_1$		$\alpha_v\beta_6$	
							Exp. pIC_{50}	Pred. pIC_{50}	Exp. pIC_{50}	Pred. pIC_{50}
1	-0.39	-60	18	-0.244	-0.030	-0.214	8.3	7.8	7.0	6.6
2	-0.53	-90	18	-0.251	-0.032	-0.219	8.8	7.9	6.4	6.8
3	-0.04	105	10	-0.246	-0.025	-0.221	8.6	7.8	7.3	6.8
4	0.15	-60	9	-0.251	-0.042	-0.209	8.5	7.9	7.8	6.9
5	-0.23	-75	10	-0.249	-0.032	-0.217	8.5	8.0	7.6	6.8
6	-0.03	-90	7	-0.245	-0.026	-0.219	7.8	8.1	7.2	6.9
7	-0.17	-105	10	-0.253	-0.041	-0.212	8.2	7.9	7.8	6.9
8	0.16	135	10	-0.251	-0.049	-0.202	8.6	8.0	8.0	6.8
9	-0.55	-150	14	-0.253	-0.049	-0.204	7.7	7.4	7.8	6.8
10	-0.59	-150	16	-0.250	-0.040	-0.210	8.2	7.8	6.6	6.5
11	-0.40	45	19	-0.247	-0.033	-0.214	7.9	7.8	7.7	6.7
12	-0.55	15	15	-0.259	-0.059	-0.200	6.7	7.8	6.4	6.9
13	-0.24	0	33	-0.253	-0.050	-0.203	8.1	7.5	7.4	6.9
14	-0.58	-45	21	-0.248	-0.046	-0.202	8.1	7.7	6.5	6.7
15	-0.60	165	38	-0.255	-0.055	-0.200	7.5	7.7	5.3	6.6
16	-0.28	165	30	-0.250	-0.055	-0.195	7.3	7.4	7.0	6.7
17	-0.29	-180	37	-0.247	-0.042	-0.205	7.6	7.4	6.9	6.4
18	0.09	-165	11	-0.252	-0.042	-0.210	7.2	8.0	6.2	6.7
19	-0.30	-180	38	-0.248	-0.039	-0.209	7.7	7.8	5.8	6.5
20	-0.47	135	31	-0.238	-0.021	-0.217	6.6	7.7	5.4	6.5
21	-0.03	105	10	-0.227	-0.006	-0.221	8.0	8.1	7.3	6.6
22	-0.47	-135	18	-0.228	-0.020	-0.208	7.5	5.9	6.6	6.6
23	-0.47	-135	27	-0.238	-0.022	-0.216	7.8	7.7	7.1	6.5
24	-0.25	105	11	-0.233	-0.009	-0.224	8.2	8.1	6.2	5.2
25	0.10	-165	20	-0.206	-0.021	-0.185	7.1	7.5	<5.0	6.3
26	-0.74	-180	86	-0.255	-0.064	-0.191	7.0	5.7	6.2	6.9
27	-0.47	-90	18	-0.251	-0.032	-0.219	8.3	7.9	6.8	6.8
28	-0.44	-90	15	-0.264	-0.055	-0.209	8.8	7.8	7.4	6.5
29	-0.54	60	17	-0.243	-0.045	-0.198	7.5	7.4	6.4	6.6
30	-0.50	75	14	-0.256	-0.048	-0.208	8.4	7.8	7.2	6.9
31	-0.54	60	20	-0.247	-0.048	-0.199	7.0	7.7	6.5	6.7
32	-0.39	-45	18	-0.243	-0.038	-0.205	7.4	7.8	6.3	6.8
33	-0.42	-135	14	-0.107	-0.011	-0.096	7.5	7.8	5.8	5.0
34	-0.42	150	34	-0.110	-0.010	-0.100	6.8	7.4	5.1	5.1
35	-0.28	-180	36	-0.228	-0.067	-0.161	6.2	7.0	6.0	6.9
36	-0.50	-135	17	-0.231	-0.035	-0.196	6.8	6.0	6.4	6.7
37	-0.31	30	17	-0.236	-0.045	-0.191	6.9	7.6	6.6	6.8
38	-0.31	-15	24	-0.218	-0.026	-0.192	7.2	7.5	8.1	6.3
39	-0.29	150	38	-0.238	-0.037	-0.201	6.8	7.7	6.8	6.3
40	-0.07	-120	21	-0.226	-0.060	-0.166	6.1	7.4	5.2	6.5
41	-0.67	165	11	-0.243	-0.033	-0.210			5.9	6.9
42	-0.29	165	11	-0.247	-0.034	-0.213			6.3	6.7
43	-0.59	-30	33	-0.248	-0.041	-0.207			6.1	6.5
44	-0.29	-15	14	-0.224	-0.022	-0.202			5.5	6.6
45	-0.54	-15	11	-0.255	-0.054	-0.201			6.0	6.3
46	-0.60	15	13	-0.253	-0.05	-0.203			5.9	6.8

Table 5.4: The calculated descriptors and the predicted pIC_{50} for both the G4 and G5 datasets assessed in LOO-CV on the training set.

A nonlinear method neural network was applied to predict the pIC_{50} values. The structure of the neural network is 6-5-1 in the case of the G4 dataset and 6-5-1 in the G5 dataset. The most predictive MLP model was characterised by a q^2 of 0.44. According to the employed machine learning algorithms in the G4 and G5 datasets, only 28 out of 441 models provided a q^2 higher than 0.40 (Table 5.5).

The SMO, MLP, IBk, M5P and REPTree algorithms gave models with q^2 values smaller than 0.50, and they performed worse than the RF and SVM algorithms. Unfortunately, all built models for the G5 dataset show only a weak correlation with the biological activities with cross-validated q^2 below 0.30.

Table 5.5: The most predictive models using the SVM (RBFKernel), MLP, IBk, M5P, RF and REPTree approaches for G4 in $\alpha_v\beta_1$ inhibitor.

	Training				Validation		
Methods	Model	r^2	MAE	RMSE	q^2	MAE	RMSE
SVM (RBFRegressor)	ψ, f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.68	0.33	0.41	0.50	0.43	0.52
	E_{HOMO}, E_{LUMO}, E_{Gap}	0.65	0.33	0.43	0.50	0.38	0.51
	E_{HOMO}, E_{Gap}	0.61	0.34	0.45	0.49	0.41	0.52
	ψ, δ, E_{HOMO}, E_{Gap}	0.68	0.31	0.40	0.47	0.40	0.53
	ψ, δ, f, E_{LUMO}, E_{Gap}	0.65	0.34	0.42	0.45	0.43	0.54
MLP	δ, f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.67	0.37	0.44	0.46	0.44	0.55
	f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.64	0.35	0.43	0.45	0.46	0.63
	ψ, f, E_{LUMO}, E_{Gap}	0.69	0.33	0.40	0.42	0.51	0.71
	δ, f, E_{HOMO}, E_{Gap}	0.62	0.35	0.44	0.42	0.51	0.71
	δ, E_{Gap}	0.54	0.39	0.49	0.41	0.44	0.56
IBk	ψ, f, E_{HOMO}, E_{Gap}	0.66	0.35	0.44	0.46	0.42	0.54
	ψ, f, E_{Gap}	0.68	0.35	0.42	0.45	0.43	0.54
	δ, E_{HOMO}, E_{Gap}	0.65	0.38	0.45	0.42	0.44	0.55
	E_{HOMO}, E_{LUMO}, E_{Gap}	0.60	0.39	0.47	0.40	0.47	0.57
	E_{HOMO}, E_{Gap}	0.61	0.36	0.45	0.40	0.44	0.56
M5P	δ, f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.67	0.35	0.42	0.40	0.43	0.56
	f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.63	0.36	0.45	0.40	0.44	0.56
	f, E_{HOMO}, E_{Gap}	0.63	0.36	0.45	0.40	0.45	0.56
RF	ψ, f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.95	0.15	0.19	0.51	0.41	0.50
	ψ, E_{HOMO}, E_{LUMO}, E_{Gap}	0.93	0.17	0.20	0.48	0.42	0.52
	ψ, δ, f, E_{HOMO}, E_{LUMO}, E_{Gap}	0.95	0.15	0.19	0.47	0.44	0.52
	ψ,f, E_{HOMO}, E_{Gap}	0.94	0.16	0.20	0.47	0.43	0.53
	f, E_{HOMO}, E_{Gap}	0.93	0.17	0.20	0.47	0.44	0.53
REPTree	E_{Gap}	0.62	0.35	0.44	0.44	0.44	0.54
	ψ, δ, E_{Gap}	0.54	0.38	0.49	0.43	0.43	0.55
	ψ, E_{Gap}	0.54	0.38	0.49	0.43	0.43	0.54
	E_{LUMO}, E_{Gap}	0.60	0.37	0.45	0.41	0.46	0.55
	δ, E_{Gap}	0.54	0.38	0.49	0.41	0.45	0.55

Table 5.5 shows only the best models of the $\alpha_v\beta_1$ integrin data generated by each of the methods examined in this study. Up to five models are shown per approach, with the best defined with respect to each of the three predictive measures in turn, q^2, RMSE and the MAE. As described previously, all possible sets of the six descriptors were analysed, i.e. six one-descriptor models, 15 two-descriptor models, 20 three-descriptor models, 15 four-descriptor models, six five-descriptor models, and a model using all six descriptors, giving a total of 63 models. A similar number of combinations and models were generated for the $\alpha_v\beta_6$ inhibitor, but, no predictive model was found. In Table 5.5, we have only shown the predictive models among the $\alpha_v\beta_1$. The M5P performs worse than the MLP, IBk models or the REPTrees. The SVM and the other nonparametric approaches outperform the MLR with respect to cross-validated parameters q^2 and MAE. A model may be considered significant if $q^2 > 0.5$. In this case, RF (one out of 63) is the most accurate and M5P (three out of 63) is the least robust. We have used two different distance functions, i.e. Euclidean or Manhattan. The Euclidean function calculates the square root of the sum of the squares of the differences of the coordinates. For example, if x=(a,b) and y=(c,d), the Euclidean distance between x and y is $\sqrt{(a-c)^2+(b-d)^2}$. In Manhattan, we take the sum of the absolute values of the differences of the coordinates. For instance, if x=(a,b) and y=(c,d), the Manhattan distance between x and y is $|a-c|+|b-d|$. However, the IBk method did not change the overall quantitative picture when using these distance functions. The IBk remains poorer than the SVM and RF methods.

5.4 Conclusions

A quantum chemical investigation into 40 benzamide derivatives (G4) and 46 other compounds (G5) was performed in order to understand the relationships between the physicochemical properties of all the datasets with their pIC_{50} values. PES scans

show twisted geometries with a higher population at each global minimum for compounds **1**, **5**, **14**, **21**, **24**, **28**, **29** and **30**, (Figure A.1). The data presented here on compounds **10**, **18** and **19** show the effect of *ortho*-substituents on the PES. Upon adding weaker H-bonding acceptors such as (Cl^- and F^-)[243–245] to the *ortho*-position, the barrier increases to 0.5 and 5.8 kcal/mol for **18** and **10**, respectively. It also increases from 0.5 kcal/mol in compound **43** to 7.2 kcal/mol in compound **9**.

In the gas phase, our results indicate that F^- and OCH_3 *ortho*-substituents are mainly in the H-bonded structures with planar geometries. However, higher flexibility was exhibited in the Cl^- *ortho*-substituent, as only 22% of the conformation is intramolecularly H-bonded. The presented PESs for **10** and **19** are in good agreement with the NMR and DFT results which were reported by Galan et al.[246] and Abraham et al.[238]

With this study, we have demonstrated the efficacy of quantum chemistry in the generation of descriptors for QSAR studies with benzamide derivatives. We have further illustrated how to perform QSAR models through SOM, SVM, IBk, MLP, RF and REPTrees. Structure information encoded in the six molecular descriptors — ψ, δ, f, E_{HOMO}, E_{LUMO}, and E_{Gap} — indicates a satisfactory correlation between our descriptors and the observed pIC_{50}. From the presented analysis, we can conclude that the molecular descriptors have some usefulness in QSAR studies.

Chapter 6

Excited States of Toluene

I am grateful to Dr David Robinson whom I have had the pleasure to work with on this chapter. Many thanks to him for his contribution and collaboration.

6.1 Introduction

Small benzene derivatives have been used widely as biological chromophores due to their unique biochemistry and photophysical properties.[247] Toluene is an example of these chromophores and has been the subject of many spectroscopic investigations.[248–251] Most proteins have many chromophores such as phenylalanine, tyrosine, and tryptophan, which contribute to their near-UV spectroscopy. To calculate the near-UV CD spectra of proteins, knowledge of the electronic transitions of model chromophores like phenylalanine or 2-amino-3-phenylpropanoic acid is required. The wave functions of the ground and excited states can only be calculated using *ab initio* methods. Electronic interactions between ground and excited states in the toluene

would provide insight into the electronic structure of phenylalanine. This can be investigated by using multi-configurational methods.

The fluorescence excitation spectra of the ground state (S_0) and excited (S_1) states of toluene have been measured by a supersonic free jet-cooled spectroscopy[249,250,252–256] and with resonance-enhanced multiphoton ionisation (REMPI), zero-kinetic-energy (ZEKE), and time-resolved slow-electron velocity map imaging (tr-SEVI) spectroscopy.[256–259] These experimental studies provided the data to which one may compare our calculations.

The ground state geometry of toluene has been determined by microwave spectroscopy,[249] gas-phase electron diffraction,[260] liquid-phase Raman backscattering spectra,[261] and by IR-UV double resonance spectroscopy.[249,254] In the S_0 state, eight bands have been identified in the UV-IR spectrum.[254] These bands are classified into two groups. The first group appears above 3000 cm^{-1}, which is assigned as phenyl a_1 CH stretches:[253,254] intense bands at 3087, 3063, and 3055 cm^{-1}.[253] The other group is the vibrations which appear below 3000 cm^{-1}. These lower energy vibrations are the CH_3 stretches of the methyl group.

A fluorescence excitation spectrum for the S_1-S_0 transition of toluene in a jet-cooled molecular beam was measured by Hickman et al.[253] Hickman's group suggested three main band spectra in the S_0 state excluding the phenyl a_1 CH stretches (region 1). Region 2 has the b_2 modes between 1250 and 1500 cm^{-1}. In this region, four fundamental vibrations were detected: one associated with the CH_3 group and three with the phenyl ring. In region 3, various studies[262,263] reported the b_2 methyl asymmetric bend between 967 and 1060 cm^{-1}. In an early study, Hopkins et al.[252] determined the excitation of the first singlet-singlet band in jet-cooled of toluene, the $1^1A_1 - 1^1B_2$ transition, at around 4.68 eV.[252] Another study[264] has determined the S_0 band to be at 4.65 eV[249,264] and 5.7 eV for the first excited state S_1 by time-resolved photoelectron imaging spectroscopy. In another study, Borst and Pratt investigated the

structure and dynamical behaviour of toluene in its ground and excited states.[249] In the methyl torsional mode, a small barrier height in the ground state of toluene has been determined to be 4.87 cm^{-1} for the S_0 state and 26.4 cm^{-1} for the S_1 state.[249] The S_0 state was determined at 4.65 eV and the internal rotor constant F is: 5.29 cm^{-1}.[249] Only six bands were identified by Minejima et al.[254] Vibrations that appear above 3000 cm^{-1} were assigned to the aromatic CH stretches. The observed frequencies for CH stretching vibrations in the S_1 of toluene are higher than the corresponding modes in the S_0 state.[254]

We now briefly review relevant theoretical studies. Many studies have considered different labelling systems for vibrational states of toluene.[265] Heickman et al.[253] employed the Mulliken[266] labelling scheme, which is the method we consider in this work. This scheme was originally used by Herzberg.[267] In this method, the vibrations were listed in descending order within each irreducible representation based on the suggested symmetry. Toluene was assumed to have C_2v symmetry and its vibrations would be arranged in terms of a_1, a_2, b_1, and b_2 modes. The methyl group is treated as a point mass. This means that one vibrational mode of the 39 normal vibrational modes of toluene is probably better considered as a torsion mode.

Several theoretical studies of the excited states of toluene have been conducted. For example, Hameka and Jensen[263] performed multi-configurational self-consistent-field (MCSCF) calculations, using a 6-31G* basis set, which placed the S_0 transition at 4.58 eV and the S_1 transition at 5.20 eV. In a complete active space self-consistent field (CASSCF)[268,269] approach, three sets of orbitals are defined. Closed orbitals, also known inactive orbitals, are doubly occupied and can be considered as the lowest orbitals. On the other hand, the highest virtual orbitals represent unoccupied orbitals. The active orbitals lie between the closed and the virtual unoccupied orbitals. In the work of Hameka and Jensen,[263] the active space was characterised by four electrons in four orbitals. They also estimated the zero-point energies to be 365.2

and 340.4 kJ/mol for the S_0 and S_1 states, respectively. Another study[250] reported a CASSCF(6, 6) calculation with a 6-31G* basis to investigate the S_0 and S_1 states of toluene. Six electrons in six orbitals were chosen as an active space and the $S_1 - S_0$ transition was predicted to be at 4.65 eV. The barrier heights of the methyl torsional frequency, V_6, are calculated to be 1.8 for the S_0 state and 22.6 cm^{-1} for the S_1 state. Cogan et al.[270] have also reported CASSCF calculations with the correlated-consistent polarised double zeta cc-pVDZ basis set. The 0-0 transition was calculated to be 3.59 eV. The active space was characterised by seven orbitals with eight electrons; the six π orbitals (three occupied and three unoccupied) and the occupied π pseudo orbital of the CH_3 group. Another theoretical study[270] reported an intersystem crossing at singlet conical intersection (ISC) in toluene. The ISC, which is an intermediate between the ground state and the first excited singlet, was located at 4.67 eV.[270] This conical intersection (ConInt) is believed to occur due to the electronic transition from the higher singlet state S_1 to the ground S_0 state.[270]

A recent study[256] reported DFT and time-dependent density functional theory (TD-DFT) calculations for the S_0 and S_1 states of toluene, respectively. In order to obtain all real vibrational frequencies, the methyl group for the S_0 and S_1 states had to be oriented in a staggered conformation; one of the C-H bonds is perpendicular to the benzene ring. In our study, we employed the DFT, TD-DFT, CASSCF, and multi-configurational second-order perturbation (CASPT2)[272,273] approaches to better understand the $\pi\pi^*$ transitions in toluene. State averaged calculations were performed over the excited states of interest when using the CASSCF method. We used CASPT2 to account for the dynamical electron correlation. Our goal was to perform calculations on toluene within a consistent framework, in which no special treatment is required for a particular state, to provide a clear basis for investigation of phenylalanine.

6.2 Computational details

Our calculations in this work were performed with the Q-Chem[236] and MOLPRO[271] suite of programs with different levels of theory and basis sets. Two different toluene conformations (eclipsed and staggered) were investigated (Figure 6.1).

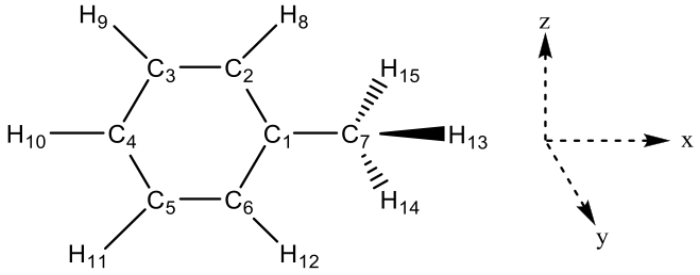


Figure 6.1: A two-dimensional diagram of the staggered conformation of toluene defining the coordinate system

The eclipsed form represents the conformation where one of the CH bonds of methyl group is parallel to the benzene ring. In the staggered conformation, one of the CH bonds of the methyl group is perpendicular to the benzene ring. Both geometries were optimised using a variety of *ab initio* calculations. Figure 6.1 shows the two-dimensional structure of the staggered toluene conformation with the numbering system used in our calculations. The ground electronic state for both conformations were optimised using MP2, B3LYP, CASSCF, and CASPT2 approaches. A 6-31G* basis set has been employed in all optimisations. Geometrical parameters for the ground states were firstly optimised using MP2 and B3LYP. These have been used in the subsequent CASSCF, and CASPT2 calculations as initial coordinates when calculating the excited state geometries and harmonic vibrational frequencies.

For the calculations of the excited state, we have used TD-DFT[274] to optimise structures for the first five excited states (S_1-S_5), of both geometries at the ΩB97X-D[275]/6-31G* level. These calculations included five roots and excluded the ground state. Symmetry was imposed in the TD-DFT calculations: the staggered form has C_s

symmetry. Energies and the transition dipole moments of both conformations were calculated.

Because we are concerned with the $\pi\pi^*$ transitions in toluene, the obvious choice of active space is to include all the π orbitals and electrons involved in the valence transitions (6, 6) as a first choice. The configurations of symmetry and spin in a CASSCF calculation, are generated from all possible arrangements of the active electrons among the active orbitals. A state-averaged procedure can provide molecular orbitals for several electronic states in which all of these states are simultaneously treated. The ground state of toluene has the electronic configuration $\tilde{X}^1A_1$. Different active spaces were used in our calculations. In the first CASSCF and CASPT2 calculations, the active space was six electrons in six orbitals (three occupied and three virtual). Symmetry was imposed in each calculation and the active space was (28; 22). $28a'$ orbitals were treated as closed and $22a'$ orbitals were occupied. i.e., three unoccupied orbitals ($26a'$ - $28a'$) and three occupied orbitals ($23a'$ - $25a'$) See Figure 6.2.

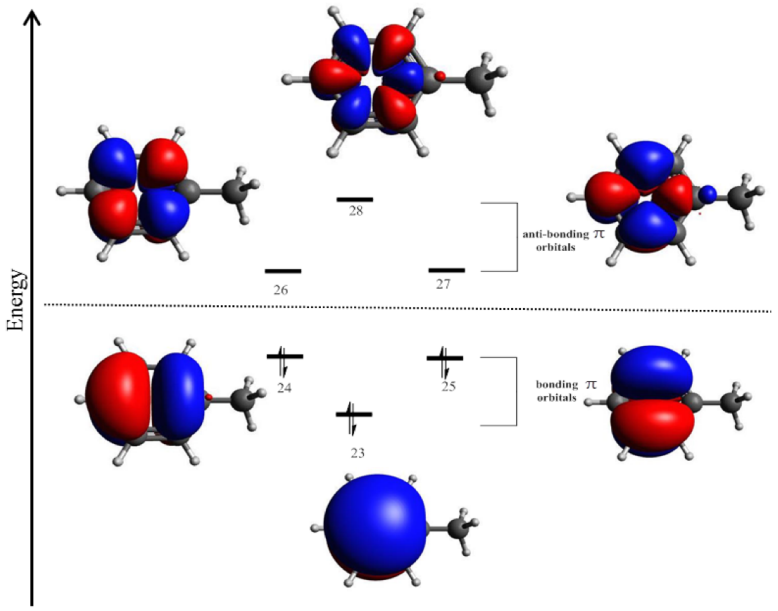


Figure 6.2: A molecular orbital diagram shows the six selected orbitals for the active space

Energy and permanent dipole moments for the ground state of the staggered toluene were calculated. The active space may be defined as six electrons in seven a' orbitals (three occupied and four virtual). The additional virtual orbital in the active space is included to identify Rydberg orbitals and to minimise any valence mixing.

6.3 Results and discussion

6.3.1 Geometries

The $\tilde{X}^1A_1$ equilibrium geometries for each conformation were optimised at the MP2, B3LYP, CASSCF, and CASPT2 levels, and the optimised parameters are shown in Table 6.1. At the CASSCF level the optimised bond lengths for the eclipsed toluene are $R_{C_1-C_7}$= 1.512 Å, $R_{C_1-C_2}$= 1.397 Å, $R_{C_2-H_8}$= 1.076 Å, and $R_{C_7-H_{14}}$= 1.085 Å. These values are in good agreement with the experiment, which yielded $R_{C_1-C_7}$= 1.523 Å, $R_{C_1-C_2}$= 1.398 Å, $R_{C_2-H_8}$= 1.106 Å, and $R_{C_7-H_{14}}$= 1.097 Å. At the CASPT2 level, in which the dynamical correlation is considered, the results are $R_{C_1-C_7}$= 1.508 Å, $R_{C_1-C_2}$= 1.402 Å, $R_{C_2-H_8}$= 1.088 Å, and $R_{C_7-H_{14}}$= 1.095 Å. These values are in excellent agreement with the experimental values.[260] At the CASPT2 level, the calculated bond length becomes closer to the experimental values. Our calculated MP2 optimised geometry is also in agreement with the CASSCF(6, 6) geometries.

Property	MP2/6-31G*		B3LYP/6-31G*		CASSCF/6-31G*		CASPT2/6-31G*		[a]Expt.
	Staggered	Eclipsed	Staggered	Eclipsed	Staggered	Eclipsed	Staggered	Eclipsed	Eclipsed
C_1-C_7	1.508	1.508	1.512	1.512	1.512	1.512	1.485	1.508	1.523
C_1-C_2	1.402	1.401	1.402	1.400	1.400	1.397	1.400	1.402	1.398
C_2-C_3	1.397	1.398	1.396	1.398	1.395	1.398	1.39	1.4	1.398
C_3-C_4	1.397	1.396	1.397	1.395	1.395	1.394	1.391	1.398	1.398
C_2-H_8	1.089	1.089	1.088	1.088	1.076	1.076	1.096	1.088	1.106
C_3-H_9	1.088	1.088	1.087	1.087	1.076	1.075	1.095	1.087	1.106
C_4-H_{10}	1.087	1.087	1.087	1.087	1.075	1.075	1.095	1.087	1.106
C_7-H_{13}	1.096	1.094	1.099	1.095	1.085	1.084	1.098	1.093	1.097
C_7-H_{14}	1.094	1.095	1.096	1.098	1.085	1.085	1.098	1.095	1.097
C_7-H_{15}	1.094	1.095	1.096	1.098	1.087	1.086	1.099	1.095	1.097
C_6-C_1-C_2	118.4	118.4	118.2	118.2	118.3	118.3	119.5	118.4	-
C_1-C_2-C_3	121	121	121	121	121	121	120.1	120.9	-
C_2-C_3-C_4	120.1	120.1	120.2	120.1	120.1	120.1	120.1	120.1	-
C_3-C_4-C_5	119.5	119.5	119.4	119.4	119.5	119.5	119.9	119.5	-

Table 6.1: Calculated bond distances in (Å), and angles in (°) for both the staggered and eclipsed toluene geometries of the S_0 state. [a]Experimental parameters were taken from Ref.,[260] in which the structural parameters have been determined by gas-phase electron diffractions and this study suggested the eclipsed conformation.

6.3.1.1 TD-DFT

For both conformations, the energies and transition dipole moments were calculated for five excited states and are presented in Table 6.2 and Table 6.3

Excited State	Transition Energy (eV)	Transition dipole moment (D)			Oscillator Strength
		μ_x	μ_y	μ_z	
S_1	5.43	0.07	0.12	0.00	0.003
S_2	6.38	0.35	0.07	0.00	0.021
S_3	7.75	2.10	0.94	0.00	1.005
S_4	7.81	-1.19	1.79	0.00	0.888
S_5	8.12	0.00	0.00	-0.05	0.001

Table 6.2: TD-DFT B3LYP/6-31G* calculated vertical transition energies and transition dipole moments for the eclipsed conformation; x, y, and z were defined in Figure 6.1.

Excited State	Transition Energy (eV)	Transition dipole moment (D)			Oscillator Strength
		μ_x	μ_y	μ_z	
S_1	5.35	0.00	-0.11	0.00	0.002
S_2	6.33	-0.37	0.00	0.04	0.022
S_3	7.66	-2.39	0.00	0.00	1.076
S_4	7.68	0.00	1.96	0.00	0.723
S_5	7.98	0.00	-0.14	0.00	0.004

Table 6.3: TD-DFT B3LYP/6-31G* calculated vertical transition energies and transition dipole moments for the staggered conformation; x, y, and z were defined in Figure 6.1.

6.3.2 Harmonic vibrational frequencies

The permanent dipole moments (D) at the electronic ground state were computed for the most stable conformer of toluene and the results are presented in Table 6.4. The vibrational frequencies of the staggered conformation were calculated at the CASSCF(6, 6) and CASPT2(6, 6) levels of theory. We have used the Mulliken nomenclature. 39 real vibrational modes are shown in Table 6.5 for the S_0 state. In this labelling scheme, the vibrations were listed depending on the suggested symmetry. The vibrational modes are listed in decreasing order in four irreducible representations; a_1, a_2, b_1, and b_2.

Level of theory	Energy (Hartree)	Permanent dipole moment (D)		
		μ_x	μ_y	μ_z
CASSCF(6, 6)	-269.81230	0.03	0.20	-0.02
CASPT2(6, 6)	-270.61952	0.05	0.27	-0.03
CASPT2(6, 7)	-270.62034	0.05	0.29	-0.03

Table 6.4: Permanent dipole moments of S_0 of the staggered conformation of toluene

Even though there is a little quantitative effect on the calculated energies when using different active spaces, it is a quite important to take into account using such method to overcome some convergence problems in our calculations. Calculated vibrational frequencies (this work) and experimental values[253] are shown in Table 6.5.

Symmetry	Mulliken mode	Assignment	CASSCF(6, 6)/6-31G* [a]scaled	Exp.[253]
a_1	1	C-H Stretch	3075	3087
	2	C-H Stretch	3063	3063
	3	C-H Stretch	3054	3055
	4	C-H3 symmetric Stretch	2911	2921
	5	C-C Stretch	1601	1605
	6	C-C Stretch	1502	1494
	7	C-H Bend	1428	1379
	8	C-CH_3 Stretch	1195	1210
	9	C-H Bend	1168	1175
	10	C-H Bend	1009	1030
	11	C-C Bend	977	1003
	12	C-C Bend	763	785
	13	C-C Bend	504	521
a_2	14	C-C Bend	901	964
	15	C-H Bend	794	843
	16	C-C Bend	394	407
b_1	17	C-H_3 anti-symmetric Stretch	2978	2979
	18	C-C Stretch	1490	1450
	19	C-H Bend	1045	1040
	20	C-C Bend	975	978
	21	C-H Bend	845	895
	22	C-H Bend	694	728
	23	C-C Bend	656	695
	24	C-C Bend	443	464
	25	C-CH3 Wag	198	216
b_2	26	C-H Stretch	3042	3039
	27	C-H Stretch	3040	3029
	28	C-H_3 anti-symmetric Stretch	2961	2952
	29	C-C Stretch	1574	1586
	30	C-H_3 Bend	1494	1463
	31	C-C Stretch	1447	1445
	32	C-H Bend	1339	1312
	33	C-H Bend	1211	1280
	34	C-H Bend	1110	1155
	35	C-H Bend	1060	1080
	36	C-C Bend	929	967
	37	C-CH_3 Bend	610	623
	38	C-CH_3 Wag	330	342
	39	C-CH_3 Rotation	14	-

Table 6.5: Calculated and experimental vibrational frequencies (cm^{-1}) of the S_0 state of the staggered toluene at the CASSCF(6, 6) level with the 6-31G*. [a] Scaling factor = 0.91

The calculated vibrational frequencies are in good agreement with experiment.[253,254] So, we have also employed an active space of six electrons and seven orbitals i. e. (29; 22). Also, symmetry was imposed in each calculation. The 29a' orbitals were treated as closed and 22a' orbitals were occupied. The additional active orbital was important in computing the vibrational frequencies for the first excited state of toluene, which could not be achieved using the (6, 6) active space. Calculated and experimental[253] frequencies of the S_0 and S_1 states at the CASSCF(6, 7)/6-31G* level are shown in Table 6.6.

Symmetry	Mulliken mode	Assignment	CASSCF(6, 7)/6-31G* S_0	Exp.[253]	CASSCF(6, 7)/6-31G* S_1	Exp.[253]
a_1	1	C-H Stretch	3075	3087	3118	-
	2	C-H Stretch	3063	3063	3106	-
	3	C-H Stretch	3054	3055	3097	-
	4	C-H_3 symmetric Stretch	2913	2921	2917	-
	5	C-C Stretch	1598	1605	1736	-
	6	C-C Stretch	1500	1494	1529	-
	7	C-H Bend	1428	1379	1432	-
	8	C-CH_3 Stretch	1195	1210	1179	1193
	9	C-H Bend	1167	1175	1018	1021
	10	C-H Bend	1008	1030	905	935
	11	C-C Bend	978	1003	949	966
	12	C-C Bend	763	785	722	736
	13	C-C Bend	505	521	462	457
a_2	14	C-C Bend	909	964	652	687
	15	C-H Bend	802	843	897	-
	16	C-C Bend	397	407	260	228
b_1	17	C-H_3 anti-symmetric Stretch	2980	2979	2969	-
	18	C-C Stretch	1490	1450	1497	-
	19	C-H Bend	1046	1040	1045	-
	20	C-C Bend	973	978	1018	-
	21	C-H Bend	847	895	897	-
	22	C-H Bend	700	728	652	-
	23	C-C Bend	663	695	455	423
	24	C-C Bend	448	464	321	320
	25	C-CH_3 Wagg	201	216	145	157
b_2	26	C-H Stretch	3042	3039	3087	-
	27	C-H Stretch	3041	3029	2994	-
	28	C-H_3 anti-symmetric Stretch	2961	2952	2917	-
	29	C-C Stretch	1569	1586	1555	-
	30	C-H_3 Bend	1495	1463	1491	-
	31	C-C Stretch	1444	1445	1437	-
	32	C-H Bend	1338	1312	1394	-
	33	C-H Bend	1209	1280	1157	-
	34	C-H Bend	1116	1155	1147	-
	35	C-H Bend	1058	1080	1045	-
	36	C-C Bend	937	967	905	-
	37	C-CH_3 Bend	608	623	539	532
	38	C-CH_3 Wag	330	342	324	332
	39	C-CH_3 Rotation	16	-	36	-

Table 6.6: Calculated and experimental vibrational frequencies (cm^{-1}) of the S_0 and S_1 state of the staggered toluene at the CASSCF(6, 7) level with the 6-31G*. [a] Scaling factor = 0.91

Our harmonic frequencies for the S_1 states are in better agreement with experimental data when compared with the calculated harmonic frequencies at the TD-B3LYP/aug-cc-pVTZ level by Wright's group[256] and shown in Table 6.7. Our calculated harmonic frequencies for the S_1 state at the CASSCF(6, 7)/6-31G* level for a_1 mode are about 14 cm^{-1} smaller than the experimental data.

Symmetry	Mulliken mode	CASSCF(6, 7)/6-31G* (This work) S_1	TD-B3LYP/aug-cc-pVTZ Ref.[256] S_1	Exp.[253]
a_1	8	1179	1221	1193
	9	1018	1177	1021
	10	905	974	935
	11	949	990	966
	12	722	768	736
	13	462	466	457
a_2	14	652	719	687
	16	260	217	228
b_1	23	600	450	423
	24	411	322	320
	25	192	150	157
b_2	37	499	538	532
	38	326	339	332

Table 6.7: Calculated and experimental vibrational frequencies (cm^{-1}) of S_1 state of the staggered toluene at the CASSCF(6, 7)/6-31G* and TD-B3LYP/aug-cc-pVTZ levels.

6.4 Conclusions

The ground and excited states of toluene were investigated using the MP2, B3LYP, TD-DFT, CASSCF and CASPT2 levels of theory. Toluene is assumed to have C_{2v} symmetry and the methyl moiety is treated as a point mass for the labelling. The ground state geometry is in good agreement with the experimental data; all MP2, B3LYP, and CASSCF bond lengths are the same ($\Delta < 0.01$Å), and bond angles are the same ($\Delta < 0.5°$).

At the CASSCF level, different active spaces were used in our calculations even though many of the variations in active space had very little quantitative effect on the cal-

culated result. The transition is calculated to be at 4.69 eV, which is in good agreement with the corresponding experimental energy, 4.65 eV. The calculated transition dipole moment was insignificant when changing the number of unoccupied orbitals. For instance, the transition dipole moment equals 0.30 D when using four unoccupied orbitals in the active space. The transition dipole moment becomes 0.28 D if the number of unoccupied orbitals is increased to six orbitals. Dynamical electron correlation was taken into account by employing the CASPT2 calculations. With the 6-31G* basis set especially, our CASPT2 calculations show similar results to CASSCF when computing the valence excited states of toluene. Additional virtual orbitals included in the active space are important to minimise any artifactual Rydberg-valence mixing.

Including augmented and diffuse functions in our future calculations might allow the basis set enough flexibility to describe Rydberg states and to better describe the excited states. For the S_0 and S_1 states, calculating anharmonic vibrational frequencies at the CASSCF level even with a small basis set may give frequencies closer to experiment. Moreover, locating the toluene conical intersection is of particular interest to obtain more details about the excited states of toluene.

Chapter 7

Conclusions

Drug design is the creative method of discovering new compounds based on the knowledge of biological targets such as $\alpha_v\beta$ integrins for treating lung fibrosis. IPF is a deadly lung disease, with a fatality rate that surpasses that of many malignancies. Recently, many studies have been shown new lead compounds for IPF treatment.[276,277] However, the results have mostly been discouraging.[276]

In silico procedures can assist in early development of lead compunds. Reducing development time and the cost is a critical priority in drug design and discovery. QSAR is one of the numerous techniques available from chemoinformatics. This thesis has reported several aspects of QSAR from model selection, to molecular descriptors calculations and suitable statistics. Every decision made before building a QSAR model is significant as it influences the possible consequence and hence usefulness of it. However, there are thousands to choose from and picking the proper descriptor is a real challenge. Some molecular descriptors are easy to be calculated, but others are time-consuming.

The data sets: G1, G2, G3 and G4 were used to build our 2D/3D-QSARs against the $\alpha_v\beta_1$ integrin. Comparison to 3D-QSAR shows that the TMACC performs better. The 2D-QSAR results revealed the importance of the three properties: charge, log$S$,

and molar refractivity. The best 2D-QSAR model presents a good predictive accuracy (q^2 of 0.72) for the G4 data set. According to the 2D-QSARs, mono-substitution in the *ortho*-position on the phenyl ring has more influence on activity against $\alpha_v\beta_1$. The NCW software allows visualising the models using a colour scheme for each property distinct features which affect the activity of the molecules. The 2D interpretation also proposed many positions on the benzamides where alterations may improve activity. Therefore, the study has granted us with predictive and diagnostic value for the design of a promised set of benzamides analogues. This work affords indication for the potential of the TMACC method for the development of QSAR models.

For the $\alpha_v\beta_6$ inhibitor, the existing work also derived 2D/3D-QSAR models to investigate the effect of each substituent of 46 compounds on activity. The NCW software was employed for both 2D- and 3D-QSAR methods, and all 2D- and 3D-QSAR models were validated using LOO-CV. The 2D-QSAR exhibited encouraging predictivity ($r^2 = 0.66$, $q^2 = 0.50$, $\text{MAE}_{CV} = 0.46$) and was noticeably better than the 3D-QSAR model. The best model reflects the importance of the log$S$ property. The visualised 2D-QSAR models show certain positions on the phenyl ring where changes may enhance activity. The activity is also improved upon moving CF_3 substituents from a *para*- and *meta*-position to an *ortho*-position in molecules **15**, **46** and **14**. The 2D-QSARs suggest the mono-substitution in the *ortho*-position, (molecules **14**, **19**) is influential for activity. However, this is not universally true. For example, molecules **10** and **26** display a negative contribution to the activity. The mono-substituents in the *para*-position such as molecules **15** and **43** regularly produced negative contribution to the activity. In contrast, the 3D-QSAR showed weak predictive results ($q^2 = 0.12$). The calculated statistical parameters (low r^2 and q^2 values) represented the low predictive ability of the constructed 3D-QSAR models.

In this study, we also have used the parametric and non-parametric regressors in the analysis of structure-activity data for both $\alpha_v\beta_1$ and $\alpha_v\beta_6$ inhibitors. For instance,

the SVM (RBFKernel, PUK), MLP, IBk, M5P, RF, and REPTree methods against the $\alpha_v\beta_1$ integrin seem not be the most successful strategy for the data set under consideration. The higher accuracy of PLS over the non-parametric techniques is clear. The data do not afford any very obvious or straightforward relationships. The best non-parametric model (RF model) produces cross-validated q^2 values of 0.64. For $\alpha_v\beta_6$ inhibiton, the best non-parametric model (IBk model) achieves a cross-validated q^2 value of 0.63, corresponding to predictive utility. This model suggests the importance of log$S$ to the activity.

A further quantum chemical investigation was performed for the (G4) and (G5) data sets. These minima are characterised by twisted geometries. The effect of *ortho*-substituents on the PES is clear and is present in molecules such as **10**, **18** and **19**. The barrier increases to 0.5 and 5.8 kcal/mol for **18** and **10**, respectively due to adding weaker H-bonding acceptors such as (Cl^- and F^-)[243–245] to the *ortho*-position. Planar geometries at the minima of both F^- and OCH_3 *ortho*-substituents indicate the possibility of H-bonded structures. Nevertheless, higher flexibility was shown in the Cl^- *ortho*-analogue, as only 22% of the conformation is intramolecularly H-bonded. The PESs for **10** and **19** are in good agreement with the NMR and DFT results which were stated by Galan et al.[246] and Abraham et al.[238] We have demonstrated the ability of quantum chemistry in the generation of descriptors for QSAR studies with benzamide derivatives. The six molecular descriptors — ψ, δ, f, E_{HOMO}, E_{LUMO}, and E_{Gap} — showed some correlation with the observed pIC_{50}. We also have generated QSAR models through SOM, SVM, IBk, MLP, RF and REPTrees. We can conclude that the molecular descriptors have some usefulness in QSAR studies in the presented study.

The mono-substituted benzene, toluene, has been studied and served as a model system for biological chromophores. The ground and excited states of toluene were studied using the MP2, B3LYP, TD-DFT, CASSCF and CASPT2 levels of theory.

The ground state geometry is in good agreement with the experimental data; all MP2, B3LYP, and CASSCF bond lengths are the same ($\Delta < 0.01$Å), and bond angles are the same ($\Delta < 0.5^{\circ}$). Various active spaces at the CASSCF level of theory were employed in our calculations but many of the variations in active space had little quantitative effect. The transition is determined to be at 4.69 eV, which is in excellent agreement with the corresponding experimental energy; 4.65 eV. Including dynamic electron correlation to our calculations, using the CASPT2/6-31G*, does not improve the calculated frequencies of toluene. Determining of universal scale factors for harmonic vibrational frequencies for the CASPT2 method might improve the results.

Since original thought of QSAR by Hansch, the studies in this field have grown extensively. Numerous successful QSAR applications were provided. This work aims to proceed with this development further. This study also showed the advantages of using the NCW and TMACC platforms without the need to rewrite the base code. We also demonstrated the interpretation feature of the NCW. This feature provides users with a better understanding of the more influential physical properties for their built QSAR models.

Further developments of TMACC and NCW are needed. The TMACC descriptor construction should be extended to involve other new atomic properties, which are considered as important for a given data set. The output of TMACC would be more beneficial if we could automate feature detection for a large data sets.

References

[1] A. D. Roses, Pharmacogenetics in drug discovery and development: a translational perspective, *Nat. Rev. Drug Discov.*, 2008, **7**, 807.

[2] B. F. Begam and J. S. Kumar, A study on cheminformatics and its applications on modern drug discovery, *Procedia Eng.*, 2012, **38**, 1264–1275.

[3] A. K. Ghose, T. Herbertz, J. M. Salvino and J. P. Mallamo, Knowledge-based chemoinformatic approaches to drug discovery, *Drug Discov. Today*, 2006, **11**, 1107–1114.

[4] S. Mailankody and V. Prasad, Five years of cancer drug approvals: innovation, efficacy, and costs, *JAMA Oncol.*, 2015, **1**, 539–540.

[5] V. Prasad, K. De Jesús and S. Mailankody, The high price of anticancer drugs: origins, implications, barriers, solutions, *Nat. Rev. Clin. Oncol.*, 2017, **14**, 381.

[6] A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer and M. L. Brown, Projections of the cost of cancer care in the United States: 2010–2020, *J Natl. Cancer Inst.*, 2011, **103**, 117–128.

[7] M. J. Strand, D. Sprunger, G. P. Cosgrove, E. R. Fernandez-Perez, S. K. Frankel, T. J. Huie, A. L. Olson, J. Solomon, K. K. Brown and J. J. Swigris, Pulmonary function and survival in idiopathic vs secondary usual interstitial pneumonia, *Chest*, 2014, **146**, 775–785.

[8] L. Richeldi, R. M. Du Bois, G. Raghu, A. Azuma, K. K. Brown, U. Costabel, V. Cottin, K. R. Flaherty, D. M. Hansell and Inoue, Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis, *N. Engl. J. Med.*, 2014, **370**, 2071–2082.

[9] P. Bonniaud, P. J. Margetts, M. Kolb, J. A. Schroeder, A. M. Kapoun, D. Damm, A. Murphy, S. Chakravarty, S. Dugar, L. Higgins *et al.*, Progressive transforming growth factor β1–induced lung fibrosis is blocked by an orally active ALK5 kinase inhibitor, *Am. J. Respir. Crit. Care Med.*, 2005, **171**, 889–898.

[10] P. Cho, An Anti-Fibrotic and Cytoprotective Agent as Therapy for Progressive Kidney Disease, *Expert Opin. Investig. Drugs.*, 2011, **19**, 1–13.

[11] R. M. Du Bois, D. Weycker, C. Albera, W. Z. Bradford, U. Costabel, A. Kartashov, T. E. King Jr, L. Lancaster, P. W. Noble and Sahn, Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference, *Am. J. Respir. Crit. Care Med.*, 2011, **184**, 1382–1389.

[12] D. Biondini, E. Balestro, D. Lacedonia, S. Cerri, R. Milaneschi, F. Luppi, E. Cocconcelli, E. Bazzan, E. Clini, M. P. F. Barbaro *et al.*, Pretreatment rate of decay in forced vital capacity predicts long-term response to pirfenidone in patients with idiopathic pulmonary fibrosis, *Sci. Rep.*, 2018, **8**, 5961.

[13] H. Kumar, T. Fojo and S. Mailankody, An appraisal of clinically meaningful outcomes guidelines for oncology clinical trials, *JAMA Oncol.*, 2016, **2**, 1238–1240.

[14] J. Avorn, The $2.6 billion pill—methodologic and policy considerations, *N. Engl. J. Med.*, 2015, **372**, 1877–1879.

[15] J. A. DiMasi, H. G. Grabowski and R. W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, *J. Health Econ.*, 2016, **47**, 20–33.

[16] A. R. Leach and V. J. Gillet, *An introduction to chemoinformatics*, Springer Science & Business Media, 2007.

[17] F. K. Brown *et al.*, Chemoinformatics: what is it and how does it impact drug discovery, *Annu. Rep. Med. Chem.*, 1998, **33**, 375–384.

[18] P. Willett, Chemoinformatics: a history, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2011, **1**, 46–56.

[19] L. C. Ray and R. A. Kirsch, Finding chemical records by digital computers, *Science*, 1957, **126**, 814–819.

[20] P. Willett, Aslib proceedings, 2008, pp. 4–17.

[21] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.*, 2012, **64**, 4–17.

[22] C. A. Lipinski, Lead-and drug-like compounds: the rule-of-five revolution, *Drug Discov. Today*, 2004, **1**, 337–341.

[23] J. G. Cumming, A. M. Davis, S. Muresan, M. Haeberlein and H. Chen, Chemical predictive modelling to improve compound quality, *Nat. Rev. Drug Discov.*, 2013, **12**, 948.

[24] T. I. Oprea, A. M. Davis, S. J. Teague and P. D. Leeson, Is there a difference between leads and drugs? A historical perspective, *J Chem. Inf. Comput. Sci.*, 2001, **41**, 1308–1315.

[25] Y. Yang, S. J. Adelstein and A. I. Kassis, Target discovery from data mining approaches, *Drug Discov. Today*, 2012, **17**, S16–S23.

[26] R. D. Cramer III, G. Redl and C. E. Berkoff, Substructural analysis. Novel approach to the problem of drug design, *J. Med. Chem.*, 1974, **17**, 533–535.

[27] K. L. Damm-Ganamet, R. D. Smith, J. B. Dunbar Jr, J. A. Stuckey and H. A. Carlson, CSAR benchmark exercise 2011–2012: evaluation of results from docking and relative ranking of blinded congeneric series, *J. Chem. Inf. Model.*, 2013, **53**, 1853–1870.

[28] E. Gawehn, J. A. Hiss and G. Schneider, Deep learning in drug discovery, *Mol. Inf.*, 2016, **35**, 3–14.

[29] P. Willett, The calculation of molecular structural similarity: principles and practice, *Mol. Inf.*, 2014, **33**, 403–413.

[30] M. Kutner, *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005.

[31] S. Wold, K. Esbensen and P. Geladi, Principal component analysis, *Chemometr. Intell. Lab.*, 1987, **2**, 37–52.

[32] H. Abdi and L. J. Williams, Principal component analysis, *Wiley interdisciplinary reviews: computational statistics*, 2010, **2**, 433–459.

[33] I. T. Jolliffe, in *Principal component analysis*, Springer, 1986, pp. 115–128.

[34] S. Wold, M. Sjöström and L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109 – 130.

[35] P. Geladi and B. R. Kowalski, Partial least-squares regression: a tutorial, *Anal. Chim. Acta.*, 1986, **185**, 1–17.

[36] J. F. Hair Jr, G. T. M. Hult, C. Ringle and M. Sarstedt, *A primer on partial least squares structural equation modeling (PLS-SEM)*, Sage Publications, 2016.

[37] B.-H. Mevik, V. H. Segtnan and T. Næs, Ensemble methods and partial least squares regression, *J. Chemometrics*, 2004, **18**, 498–507.

[38] P. P. Roy and K. Roy, On some aspects of variable selection for partial least squares regression models, *Mol. Inform.*, 2008, **27**, 302–313.

[39] T. Mehmood, K. H. Liland, L. Snipen and S. Sæbø, A review of variable selection methods in partial least squares regression, *Chemom. Intell. Lab. Syst.*, 2012, **118**, 62–69.

[40] D. T Stanton, QSAR and QSPR model interpretation using partial least squares (PLS) analysis, *Curr. Comput. Aided. Drug. Des.*, 2012, **8**, 107–127.

[41] P. Jain and J. D. Hirst, Automatic structure classification of small proteins using random forest, *BMC Bioinform.*, 2010, **11**, 364.

[42] R. E. Smith, M. K. Jiang, J. Bacardit, M. Stout, N. Krasnogor and J. D. Hirst, A learning classifier system with mutual-information-based fitness, *Evol. Intell.*, 2010, **3**, 31–50.

[43] M. Eklund, U. Norinder, S. Boyer and L. Carlsson, Choosing feature selection and learning algorithms in QSAR, *J. Chem. Inf. Model.*, 2014, **54**, 837–843.

[44] L. Breiman, Random forests, *Mach. Learn.*, 2001, **45**, 5–32.

[45] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, Wiley New York, 1998, vol. 1.

[46] A. J. Smola and B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.*, 2004, **14**, 199–222.

[47] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, 1995, **20**, 273–297.

[48] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya and K. R. K. Murthy, Improvements to the SMO algorithm for SVM regression, *IEEE trans. Neural. Netw.*, 2000, **11**, 1188–1193.

[49] W. S. Noble, What is a support vector machine?, *Nat. Biotech.*, 2006, **24**, 1565–1567.

[50] P. Jain, J. M. Garibaldi and J. D. Hirst, Supervised machine learning algorithms for protein structure classification, *Comput. Biol. Chem.*, 2009, **33**, 216–223.

[51] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 1995, **247**, 536–540.

[52] G. Schneider and U. Fechner, Computer-based de novo design of drug-like molecules, *Nat. Rev. Drug Discov.*, 2005, **4**, 649.

[53] K. Baumann and G. Schneider, Big Data and Deep Learning: A New Age of Molecular Informatics?, *Mol. Inf.*, 2017, **36**, 1780132.

[54] H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, Low data drug discovery with one-shot learning, *ACS Cent. Sci.*, 2017, **3**, 283–293.

[55] D. Oglic, S. A. Oatley, S. J. Macdonald, T. Mcinally, R. Garnett, J. D. Hirst and T. Gärtner, Active Search for Computer-aided Drug Design, *Mol. Inf.*, 2018, **37**, 1700130.

[56] F. Jensen, *Introduction to computational chemistry*, John wiley & sons, 2017.

[57] A. C. Brown and T. R. Fraser, On the connection between chemical constitution and physiological action; with special reference to the physiological action of the

salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia, *MOJ Anat. Physiol.*, 1868, **2**, 224.

[58] B. Richardson, Physiological research on alcohols, *Med. Times and Gazzette*, 1869, 703–706.

[59] H. Meyer, On the theory of alcohol narcosis, *Arch. Exp. Pathol. Pharmakol.*, 1901, **46**, 338–346.

[60] C. E. Overton, *Studies on anesthesia at the same time a contribution to general pharmacology*, Fischer, 1901.

[61] C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients, *Nature*, 1962, **194**, 178.

[62] C. Hansch and T. Fujita, p-σ-π Analysis. A method for the correlation of biological activity and chemical structure, *J. Am. Chem. Soc.*, 1964, **86**, 1616–1626.

[63] C. Hansch, A. Leo and R. Taft, A survey of Hammett substituent constants and resonance and field parameters, *Chem. Rev.*, 1991, **91**, 165–195.

[64] E. S. Lewis, Steric Effects in Organic Chemistry., *J. Am. Chem. Soc.*, 1957, **79**, 1014–1014.

[65] J. S. Desgrosellier and D. A. Cheresh, Integrins in cancer: biological implications and therapeutic opportunities, *Nat. Rev. Cancer*, 2010, **10**, 9.

[66] A. Ullrich and J. Schlessinger, Signal transduction by receptors with tyrosine kinase activity, *Cell*, 1990, **61**, 203–212.

[67] J. H. Walsh, W. Karnes, F. Cuttitta and A. Walker, Autocrine growth factors and solid tumor malignancy., *West J. Med.*, 1991, **155**, 152.

[68] C. E. Hart, J. W. Forstrom, J. D. Kelly, R. A. Seifert, R. A. Smith, R. Ross, M. J. Murray and D. F. Bowen-Pope, Two classes of PDGF receptor recognize different isoforms of PDGF, *Science*, 1988, **240**, 1529–1531.

[69] M. A. Smith, R. Hall, K. Fisher, S. M. Haake, F. Khalil, M. B. Schabath, V. Vuaroqueaux, H.-H. Fiebig, S. Altiok, Y. A. Chen *et al.*, Annotation of human cancers with EGFR signaling–associated protein complexes using proximity ligation assays, *Sci. Signal.*, 2015, **8**, 1–11.

[70] D. G. Albertson, Gene amplification in cancer, *Trends Genet.*, 2006, **22**, 447–455.

[71] R. S. Herbst, Review of epidermal growth factor receptor biology, *Int. J. Radiat. Oncol. Biol. Phys.*, 2004, **59**, S21–S26.

[72] H. Hamidi, M. Pietilä and J. Ivaska, The complexity of integrins in cancer and new scopes for therapeutic targeting, *Br. J. Cancer*, 2016, **115**, 1017.

[73] K. Ley, J. Rivera-Nieves, W. J. Sandborn and S. Shattil, Integrin-based therapeutics: biological basis, clinical use and new drugs, *Nat. Rev. Drug Discov.*, 2016, **15**, 173–183.

[74] S. V. Hudson, C. E. Dolin, L. G. Poole, V. L. Massey, D. Wilkey, J. I. Beier, M. L. Merchant, H. B. Frieboes and G. E. Arteel, Modeling the kinetics of integrin receptor binding to hepatic extracellular matrix proteins, *Sci. Rep.*, 2017, **7**, 12444.

[75] E. F. Plow, T. A. Haas, L. Zhang, J. Loftus and J. W. Smith, Ligand binding to integrins, *J. Biol. Chem.*, 2000, **275**, 21785–21788.

[76] B. E. Vogel, G. Tarone, F. Giancotti, J. Gailit and E. Ruoslahti, A novel fi-

bronectin receptor with an unexpected subunit composition ($\alpha_v\beta_1$), *J. Biol. Chem.*, 1990, **265**, 5934–5937.

[77] C. A. Otey, F. M. Pavalko and K. Burridge, An interaction between alpha-actinin and the beta 1 integrin subunit in vitro., *J. Cell Biol.*, 1990, **111**, 721–729.

[78] R. Pytela, M. D. Pierschbacher and E. Ruoslahti, Identification and isolation of a 140 kd cell surface glycoprotein with properties expected of a fibronectin receptor, *Cell*, 1985, **40**, 191–198.

[79] G. Raghu, H. R. Collard, J. J. Egan, F. J. Martinez, J. Behr, K. K. Brown, T. V. Colby, J.-F. Cordier, K. R. Flaherty, J. A. Lasky *et al.*, An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management, *Am. J. Respir. Crit. Care Med.*, 2011, **183**, 788–824.

[80] D. M. Mannino, R. A. Etzel and R. G. Parrish, Pulmonary fibrosis deaths in the United States, 1979-1991. An analysis of multiple-cause mortality data., *Am. J. Respir. Crit. Care Med.*, 1996, **153**, 1548–1552.

[81] G. Trujillo, A. Meneghin, K. R. Flaherty, L. M. Sholl, J. L. Myers, E. A. Kazerooni, B. H. Gross, S. R. Oak, A. L. Coelho, H. Evanoff *et al.*, TLR9 differentiates rapidly from slowly progressing forms of idiopathic pulmonary fibrosis, *Sci. Transl. Med.*, 2010, **2**, 57ra82–57ra82.

[82] F. J. Martinez, H. R. Collard, A. Pardo, G. Raghu, L. Richeldi, M. Selman, J. J. Swigris, H. Taniguchi and A. U. Wells, Idiopathic pulmonary fibrosis, *Nat. Rev. Dis. Primers*, 2017, **3**, 1–19.

[83] G. J. Laurent, R. J. McAnulty, M. Hill and R. Chambers, Escape from the

matrix: multiple mechanisms for fibroblast activation in pulmonary fibrosis, *Proc. Am. Thorac. Soc.*, 2008, **5**, 311–315.

[84] W. W. Douglas, J. H. Ryu and D. R. Schroeder, Idiopathic pulmonary fibrosis: impact of oxygen and colchicine, prednisone, or no therapy on survival, *Am. J. Respir. Crit. Care Med.*, 2000, **161**, 1172–1178.

[85] T. E. King Jr, J. A. Tooze, M. I. Schwarz, K. R. Brown and R. M. Cherniack, Predicting survival in idiopathic pulmonary fibrosis: scoring system and survival model, *Am. J. Respir. Crit. Care Med.*, 2001, **164**, 1171–1181.

[86] J. Gribbin, R. B. Hubbard, I. Le Jeune, C. J. Smith, J. West and L. J. Tata, Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK, *Thorax*, 2006, **61**, 980–985.

[87] G. Raghu, D. Weycker, J. Edelsberg, W. Z. Bradford and G. Oster, Incidence and prevalence of idiopathic pulmonary fibrosis, *Am. J. Respir. Crit. Care Med.*, 2006, **174**, 810–816.

[88] R. M. Rudd, R. J. Prescott, J. Chalmers and I. D. Johnston, British Thoracic Society Study on cryptogenic fibrosing alveolitis: response to treatment and survival, *Thorax*, 2007, **62**, 62–66.

[89] A. L. Olson, J. J. Swigris, D. C. Lezotte, J. M. Norris, C. G. Wilson and K. K. Brown, Mortality from pulmonary fibrosis increased in the United States from 1992 to 2003, *Am. J. Respir. Crit. Care Med.*, 2007, **176**, 277–284.

[90] K. Iwai, T. Mori, N. Yamada, M. Yamaguchi and Y. Hosoda, Idiopathic pulmonary fibrosis. Epidemiologic approaches to occupational exposure., *Am. J. Respir. Crit. Care Med.*, 1994, **150**, 670–675.

[91] G. Raghu, S.-Y. Chen, W.-S. Yeh, B. Maroni, Q. Li, Y.-C. Lee and H. R. Collard, Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001–11, *Lancet. Respir. Med.*, 2014, **2**, 566–572.

[92] P. J. Wolters, H. R. Collard and K. D. Jones, Pathogenesis of idiopathic pulmonary fibrosis, *Annu. Rev. Pathol.*, 2014, **9**, 157–179.

[93] M. L. Salisbury, M. Xia, S. Murray, B. J. Bartholmai, E. A. Kazerooni, C. A. Meldrum, F. J. Martinez and K. R. Flaherty, Predictors of idiopathic pulmonary fibrosis in absence of radiologic honeycombing: a cross sectional analysis in ILD patients undergoing lung tissue sampling, *Respir. Med.*, 2016, **118**, 88–95.

[94] B. Ley and H. R. Collard, Epidemiology of idiopathic pulmonary fibrosis, *Clin. Epidemiol.*, 2013, **5**, 483.

[95] S. K. Mathai, C. A. Newton, D. A. Schwartz and C. K. Garcia, Pulmonary fibrosis in the era of stratified medicine, *Thorax*, 2016, thoraxjnl–2016.

[96] W. R. Coward, G. Saini and G. Jenkins, The pathogenesis of idiopathic pulmonary fibrosis, *Ther. Adv. Respir. Dis.*, 2010, **4**, 367–388.

[97] G. Sgalla, B. Iovene, M. Calvello, M. Ori, F. Varone and L. Richeldi, Idiopathic pulmonary fibrosis: pathogenesis and management, *Respir. Res.*, 2018, **19**, 32.

[98] H. Kubinyi, Free Wilson analysis. Theory, applications and its relationship to Hansch analysis, *Mol. Inform.*, 1988, **7**, 121–133.

[99] T. J. Archdeacon, *Correlation and regression analysis: a historian's guide*, Univ of Wisconsin Press, 1994.

[100] A. Tropsha, P. Gramatica and V. K. Gombar, The importance of being earnest:

validation is the absolute essential for successful application and interpretation of QSPR models, *Mol. Inform.*, 2003, **22**, 69–77.

[101] C. L. Bruce, *PhD Thesis*, The University of Nottingham, 2010.

[102] G. Klopman and A. N. Kalos, Causality in structure—activity studies, *J. Comput. Chem.*, 1985, **6**, 492–506.

[103] C. Rücker, G. Rücker and M. Meringer, y-Randomization and its variants in QSPR/QSAR, *J. Chem. Inf. Model.*, 2007, **47**, 2345–2357.

[104] A. C. Rencher and F. C. Pun, Inflation of R^2 in Best Subset Regression, *Technometrics*, 1980, **22**, 49–53.

[105] J.-L. Faulon and A. Bender, *Handbook of chemoinformatics algorithms*, CRC press, 2010.

[106] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, *Mol. Inform.*, 2010, **29**, 476–488.

[107] T. R. Stouch, J. R. Kenyon, S. R. Johnson, X.-Q. Chen, A. Doweyko and Y. Li, In silico ADME/Tox: why models fail, *J. Comput. Aided. Mol. Des.*, 2003, **17**, 83–92.

[108] A. Golbraikh and A. Tropsha, Beware of q^2, *J. Mol. Graph. Model.*, 2002, **20**, 269–276.

[109] R. Cao, Z. Wang, Y. Wang and J. Cheng, SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines, *BMC Bioinform.*, 2014, **15**, 120.

[110] R. Kumar, A. Srivastava, B. Kumari and M. Kumar, Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine, *J. Theor. Biol.*, 2015, **365**, 96–103.

[111] G. A. Khoury, P. Tamamis, N. Pinnaduwage, J. Smadbeck, C. A. Kieslich and C. A. Floudas, Princeton TIGRESS: Protein geometry refinement using simulations and support vector machines, *Proteins: Struct., Funct., Bioinf.*, 2014, **82**, 794–814.

[112] B. L. Ingle, B. C. Veber, J. W. Nichols and R. Tornero-Velez, Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability, *J. Chem. Inf. Model.*, 2016, **56**, 2243–2252.

[113] F. R. Burden and D. A. Winkler, Relevance vector machines: sparse classification methods for QSAR, *J. Chem. Inf. Model.*, 2015, **55**, 1529–1534.

[114] K. Heikamp and J. Bajorath, Support vector machines for drug discovery, *Exp. Op. Drug Disc.*, 2014, **9**, 93–104.

[115] R. L. M. Robinson, A. Palczewska, J. Palczewski and N. Kidley, Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets, *J. Chem. Inf. Model*, 2017, **57**, 1773–1792.

[116] R. P. Sheridan, Debunking the Idea that Ligand Efficiency Indices Are Superior to pIC_{50} as QSAR Activities, *J. Chem. Inf. Model.*, 2016, **56**, 2253–2262.

[117] B. Zheng, S. W. Yoon and S. S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert. Syst. Appl.*, 2014, **41**, 1476–1482.

[118] A. Ahmadi and P. Afshar, Intelligent breast cancer recognition using particle swarm optimization and support vector machines, *J. Exp. Theor. Artif. Intell.*, 2016, **28**, 1021–1034.

[119] T. Sun, J. Wang, X. Li, P. Lv, F. Liu, Y. Luo, Q. Gao, H. Zhu and X. Guo, Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set, *Comput. Methods. Programs. Biomed.*, 2013, **111**, 519–524.

[120] Z. Qin, M. Wang and A. Yan, QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM), *Med. Chem. Lett.*, 2017, **27**, 2931–2938.

[121] D. Horvath, G. Marcou, A. Varnek, S. Kayastha, A. de la Vega de Leon and J. Bajorath, Prediction of activity cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression, *J. Chem. Inf. Model.*, 2016, **56**, 1631–1640.

[122] X. Yao, A. Panaye, J.-P. Doucet, R. Zhang, H. Chen, M. Liu, Z. Hu and B. T. Fan, Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression, *J. Chem. Inf. Model.*, 2004, **44**, 1257–1266.

[123] J. Platt, in *Advances in Kernel Methods - Support Vector Learning*, ed. B. Schoelkopf, C. Burges and A. Smola, MIT Press, 1998.

[124] S. Keerthi, S. Shevade, C. Bhattacharyya and K. Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Comput.*, 2001, **13**, 637–649.

[125] T. Hastie and R. Tibshirani, Advances in Neural Information Processing Systems, 1998.

[126] K. Pearson, X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material, *Phil. Trans. R. Soc. Lond. A*, 1895, **186**, 343–414.

[127] A. De Weijer, C. B. Lucasius, L. Buydens, G. Kateman, H. Heuvel and H. Mannee, Curve fitting using natural computation, *Anal. Chem.*, 1994, **66**, 23–31.

[128] J. Mercer, Functions of positive and negative type, and their connection with the theory of integral equations, *Philos. Trans. Royal Soc. A*, 1909, **209**, 415–446.

[129] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, John Wiley & Sons, 2008, vol. 11.

[130] M. Karelson, *Molecular descriptors in QSAR/QSPR*, Wiley-Interscience, 2000.

[131] J. Devillers and A. T. Balaban, *Topological indices and related descriptors in QSAR and QSPR*, Amsterdam, 1999.

[132] Danishuddin and A. U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, *Drug Discov. Today*, 2016, **21**, 1291–1302.

[133] J. Gasteiger and M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, *Tetrahedron*, 1980, **36**, 3219–3228.

[134] S. A. Wildman and G. M. Crippen, Prediction of physicochemical parameters by atomic contributions, *J. Chem. Inf. Model.*, 1999, **39**, 868–873.

[135] T. Hou, K. Xia, W. Zhang and X. Xu, ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach, *J. Chem. Inf. Model.*, 2004, **44**, 266–275.

[136] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome and Y. Matsushita, Simple method of calculating octanol/water partition coefficient, *Chem. Pharm. Bull.*, 1992, **40**, 127–130.

[137] T. Walker, C. M. Grulke, D. Pozefsky and A. Tropsha, Chembench: a cheminformatics workbench, *Bioinform.*, 2010, **26**, 3000–3001.

[138] L. Weber, *JChem Base-ChemAxon*, 2008.

[139] T. Sander, J. Freyss, M. von Korff and C. Rufener, DataWarrior: an open-source program for chemistry aware data visualization and analysis, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.

[140] A. K. Ghose and G. M. Crippen, Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity, *J. Comput. Chem.*, 1986, **7**, 565–577.

[141] A. T. Balaban, Topological indices based on topological distances in molecular graphs, *Pure and Applied Chemistry*, 1983, **55**, 199–206.

[142] J. Devillers and A. T. Balaban, *Topological indices and related descriptors in QSAR and QSPAR*, CRC Press, 2000.

[143] M. Randić and J. Zupan, On interpretation of well-known topological indices, *J. Chem. Inf. Model.*, 2001, **41**, 550–560.

[144] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou and C. T. Supuran, QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors using topological information indices, *Bioorg. Med. Chem.*, 2006, **14**, 1108–1114.

[145] O. Ivanciuc and A. Balaban, The graph description of chemical structures, *Topological indices and related descriptors in QSAR and QSPR*, 1999, 59–167.

[146] R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, John Wiley & Sons, 2009.

[147] K. Roy, S. Kar and R. N. Das, *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*, Academic Press, 2015.

[148] J. L. Melville and J. D. Hirst, TMACC: Interpretable Correlation Descriptors for Quantitative Structure- Activity Relationships, *J. Chem. Inf. Model.*, 2007, **47**, 626–634.

[149] M. Pastor, G. Cruciani, I. McLay, S. Pickett and S. Clementi, GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors, *J. Med. Chem.*, 2000, **43**, 3233–3243.

[150] T. W. Heritage and D. R. Lowis, *Molecular hologram QSAR*, ACS Publications, 1999.

[151] G. Klebe, U. Abraham and T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, *J. Med. Chem.*, 1994, **37**, 4130–4146.

[152] C. Marvin, *5.5. 1.0 program, Budapest, Hungary, 2011*, https://www.chemaxon.com/.

[153] R. D. Cramer, D. E. Patterson and J. D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, 1988, **110**, 5959–5967.

[154] Z. Zhang, A. O. Morla, K. Vuori, J. S. Bauer, R. Juliano and E. Ruoslahti, The $\alpha_v\beta_1$ integrin functions as a fibronectin receptor but does not support fibronectin matrix assembly and cell migration on fibronectin, *J. Cell Biol.*, 1993, **122**, 235–242.

[155] N. Ahmed, C. Riley, G. Rice and M. Quinn, Role of integrin receptors for fibronectin, collagen and laminin in the regulation of ovarian carcinoma functions in response to a matrix microenvironment, *Clin. Exp. Metastasis*, 2005, **22**, 391–402.

[156] N. I. Reed, H. Jo, C. Chen, K. Tsujino, T. D. Arnold, W. F. DeGrado and D. Sheppard, The $\alpha_v\beta_1$ integrin plays a critical in vivo role in tissue fibrosis, *Sci. Transl. Med.*, 2015, **7**, 1–17.

[157] D. Mu, S. Cambier, L. Fjellbirkeland, J. L. Baron, J. S. Munger, H. Kawakatsu, D. Sheppard, V. C. Broaddus and S. L. Nishimura, The integrin $\alpha_v\beta_8$ mediates epithelial homeostasis through MT1-MMP–dependent activation of TGF-β_1, *J. Cell Biol.*, 2002, **157**, 493–507.

[158] J. S. Munger, X. Huang, H. Kawakatsu, M. J. Griffiths, S. L. Dalton, J. Wu, J.-F. Pittet, N. Kaminski, C. Garat, M. A. Matthay and D. S. D. B. Rifkin, A mechanism for regulating pulmonary inflammation and fibrosis: the integrin $\alpha_v\beta_6$ binds and activates latent TGF β_1, *Cell*, 1999, **96**, 319–328.

[159] Y. Asano, H. Ihn, K. Yamane, M. Jinnin, Y. Mimura and K. Tamaki, Increased expression of integrin $\alpha_v\beta_3$ contributes to the establishment of autocrine TGF-β signaling in scleroderma fibroblasts, *J. Immunol.*, 2005, **175**, 7708–7718.

[160] Y. Asano, H. Ihn, K. Yamane, M. Jinnin and K. Tamaki, Increased expression of integrin $\alpha_v\beta_5$ induces the myofibroblastic differentiation of dermal fibroblasts, *Am. J. Pathol.*, 2006, **168**, 499–510.

[161] N. C. Henderson, T. D. Arnold, Y. Katamura, M. M. Giacomini, J. D. Rodriguez, J. H. McCarty, A. Pellicoro, E. Raschperger, C. Betsholtz and P. G. Ruminski, Targeting of α_v integrin identifies a core molecular pathway that regulates fibrosis in several organs, *Nat. Med.*, 2013, **19**, 1617–1624.

[162] M. Nagae, S. Re, E. Mihara, T. Nogi, Y. Sugita and J. Takagi, Crystal structure of $\alpha_5\beta_1$ integrin ectodomain: Atomic details of the fibronectin receptor, *J. Cell Biol.*, 2012, **197**, 131–140.

[163] J.-P. Xiong, T. Stehle, R. Zhang, A. Joachimiak, M. Frech, S. L. Goodman and M. A. Arnaout, Crystal structure of the extracellular segment of integrin $\alpha_v\beta_3$ in complex with an Arg-Gly-Asp ligand, *Science*, 2002, **296**, 151–155.

[164] M. W. Miller, S. Basra, D. W. Kulp, P. C. Billings, S. Choi, M. P. Beavers, O. J. McCarty, Z. Zou, M. L. Kahn and J. S. Bennett, Small-molecule inhibitors of integrin $\alpha2\beta1$ that prevent pathological thrombus formation via an allosteric mechanism, *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 719–724.

[165] N. I. Reed, Y.-Z. Tang, J. McIntosh, Y. Wu, K. S. Molnar, A. Civitavecchia, D. Sheppard, W. F. DeGrado and H. Jo, Exploring N-arylsulfonyl-L-proline scaffold as a platform for potent and selective $\alpha_v\beta_1$ integrin inhibitors, *ACS Med. Chem. Lett.*, 2016, **7**, 902–907.

[166] S. Choi, G. Vilaire, C. Marcinkiewicz, J. D. Winkler, J. S. Bennett and W. F. DeGrado, Small molecule inhibitors of integrin $\alpha_2\beta_1$, *J. Med. Chem.*, 2007, **50**, 5457–5462.

[167] W. K. Hagmann, P. L. Durette, T. Lanza, N. J. Kevin, S. E. de Laszlo, I. E. Kopka, D. Young, P. A. Magriotis, B. Li, L. S. Lin, G. Yang, T. Kamenecka, L. L. Chang, J. Wilson, M. MacCoss, S. G. Mills, G. V. Riper, E. McCauley, L. A. Egger, U. Kidambi, K. Lyons, S. Vincent, R. Stearns, A. Colletti, J. Teffera, S. Tong, J. Fenyk-Melody, K. Owens, D. Levorse, P. Kim, J. A. Schmidt and R. A. Mumford, The discovery of sulfonylated dipeptides as Potent VLA-4 antagonists, *Bioorg. Med. Chem. Lett.*, 2001, **11**, 2709 – 2713.

[168] L. S. Lin, T. Lanza, E. McCauley, G. V. Riper, U. Kidambi, J. Cao, L. A. Egger, R. A. Mumford, J. A. Schmidt, M. MacCoss and W. K. Hagmann, Specific and dual antagonists of $\alpha_4\beta_1$ and $\alpha_4\beta_7$ integrins, *Bioorg. Med. Chem. Lett.*, 2002, **12**, 133 – 136.

[169] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminform.*, 2011, **3**, 33.

[170] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar and R. K. Robins, Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 163–172.

[171] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, *Chemom. Intell. Lab. Syst.*, 1993, **18**, 251–263.

[172] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, The WEKA data mining software: an update, *SIGKDD Explor.*, 2009, **11**, 10–18.

[173] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.

[174] D. Rumelhart, David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, *Nature*, 1986, **323**, 533–536.

[175] D. W. Aha, D. Kibler and M. K. Albert, Instance-based learning algorithms, *Mach. Learn.*, 1991, **6**, 37–66.

[176] J. B. Mitchell, Machine learning methods in chemoinformatics, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 468–481.

[177] J. R. Quinlan, 5th Australian joint conference on artificial intelligence, 1992, pp. 343–348.

[178] Y. Wang and I. H. Witten, Proceedings of the Ninth European Conference on Machine Learning, 1997, pp. 128–137.

[179] F. Pereira, K. Xiao, D. A. Latino, C. Wu, Q. Zhang and J. Aires-de Sousa, Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals, *J. Chem. Inf. Model.*, 2016, **57**, 11–21.

[180] J. R. Quinlan, Simplifying decision trees, *Int. J. Man. Mach. Stud.*, 1987, **27**, 221–234.

[181] T. A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, *J. Comp. Chem.*, 1996, **17**, 490–519.

[182] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner and E. L. Willighagen, The Blue Obelisk—interoperability in chemical informatics, *J. Chem. Inf. Model.*, 2006, **46**, 991–998.

[183] P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.

[184] R. Todeschini, V. Consonni, A. Mauri and M. Pavan, Detecting "bad" regression models: multicriteria fitness functions in regression analysis, *Anal. Chim. Acta*, 2004, **515**, 199–208.

[185] R. Perkins, H. Fang, W. Tong and W. J. Welsh, Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology, *Environ. Toxicol. Chem.*, 2003, **22**, 1666–1679.

[186] R. O. Hynes, Integrins: bidirectional, allosteric signaling machines, *Cell*, 2002, **110**, 673–687.

[187] J. Adams, E. C. Anderson, E. E. Blackham, Y. W. R. Chiu, T. Clarke, N. Eccles, L. A. Gill, J. J. Haye, H. T. Haywood, C. R. Hoenig, M. Kausas, J. Le, H. L. Russell, C. Smedley, W. J. Tipping, T. Tongue, C. C. Wood, J. Yeung, J. E. Rowedder, M. J. Fray, T. McInally and S. J. F. Macdonald, Structure Activity Relationships of α_v Integrin Antagonists for Pulmonary Fibrosis by Variation in Aryl Substituents, *ACS Med. Chem. Lett.*, 2014, **5**, 1207–1212.

[188] M. Allison, Stromedix acquisition signals growing interest in fibrosis, *Nat. Biotech.*, 2012, **30**, 375–376.

[189] M. Busk, R. Pytela and D. Sheppard, Characterization of the integrin $\alpha_v\beta_6$ as a fibronectin-binding protein., *J. Biol. Chem.*, 1992, **267**, 5790–5796.

[190] J. S. Munger, J. G. Harpel, F. G. Giancotti and D. B. Rifkin, Interactions between growth factors and integrins: latent forms of transforming growth factor-β are ligands for the integrin $\alpha_v\beta_1$, *Mol. Biol. Cell*, 1998, **9**, 2627–2638.

[191] K. Hahm, M. E. Lukashev, Y. Luo, W. J. Yang, B. M. Dolinski, P. H. Weinreb, K. J. Simon, L. C. Wang, D. R. Leone, R. R. Lobb, A. N. E. McCrann, Donald J., F. A. Horan, Gerald S., S. I. C. F. Kalluri, Raghu, G. H. A. Sheppare, Dean and S. M. Violette, $\alpha_v\beta_6$ integrin regulates renal fibrosis and inflammation in Alport mouse, *Am. J. Pathol.*, 2007, **170**, 110–125.

[192] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular biology of the cell (Garland Science, New York, NY)*, 2002.

[193] T. Xiao, J. Takagi, B. S. Coller, J.-H. Wang and T. A. Springer, Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics, *Nature*, 2004, **432**, 59–67.

[194] M. Shimaoka, T. Xiao, J.-H. Liu, Y. Yang, Y. Dong, C.-D. Jun, A. McCormack, R. Zhang, A. Joachimiak, J. Takagi, J.-H. Wang and T. A. Springer, Structures

of the αL I domain and its complex with ICAM-1 reveal a shape-shifting pathway for integrin regulation, *Cell*, 2003, **112**, 99–111.

[195] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, Springer, 2006, pp. 675–684.

[196] A. Bandyopadhyay and S. Raghavan, Defining the role of integrin $\alpha_v\beta_6$ in cancer, *Curr. Drug Targets*, 2009, **10**, 645–652.

[197] A. Goodwin and G. Jenkins, Role of integrin-mediated TGFβ activation in the pathogenesis of pulmonary fibrosis, *Biochem. Soc. Trans.*, **4**, 849–854.

[198] D. Sheppard, Roles of αv integrins in vascular biology and pulmonary pathology, *Curr. Opin. Cell Biol.*, 2004, **16**, 552–557.

[199] J. S. Munger, J. G. Harpel, P.-E. Gleizes, R. Mazzieri, I. Nunes and D. B. Rifkin, Latent transforming growth factor-β: structural features and mechanisms of activation, *Kidney Int*, 1997, **51**, 1376–1382.

[200] S. Tadokoro, S. J. Shattil, K. Eto, V. Tai, R. C. Liddington, J. M. de Pereda, M. H. Ginsberg and D. A. Calderwood, Talin binding to integrin ß tails: a final common step in integrin activation, *Science*, 2003, **302**, 103–106.

[201] M. Moser, K. R. Legate, R. Zent and R. Fässler, The tail of integrins, talin, and kindlins, *Science*, 2009, **324**, 895–899.

[202] A. L. Tatler and G. Jenkins, TGF-β activation and lung fibrosis, *Proceedings of the American Thoracic Society*, 2012, **9**, 130–136.

[203] L. Häkkinen, L. Koivisto, H. Gardner, U. Saarialho-Kere, J. M. Carroll, M. Lakso, H. Rauvala, M. Laato, J. Heino and H. Larjava, Increased expression

of β6-integrin in skin leads to spontaneous development of chronic wounds, *Am. J. Pathol.*, 2004, **164**, 229–242.

[204] G. Schüürmann, R.-U. Ebert, J. Chen, B. Wang and R. Kühne, External validation and prediction employing the predictive squared correlation coefficient Test set activity mean vs training set activity mean, *J. Chem. Inform. Model.*, 2008, **48**, 2140–2145.

[205] D. Stark, M. Piel, H. Hübner, P. Gmeiner, G. Gründer and F. Rösch, In vitro affinities of various halogenated benzamide derivatives as potential radioligands for non-invasive quantification of D 2-like dopamine receptors, *Bioorganic Med. Chem.*, 2007, **15**, 6819–6829.

[206] C. Vancheri, M. Failla, N. Crimi and G. Raghu, Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology, *Eur. Respir. J.*, 2010, **35**, 496–504.

[207] M. Karelson, V. S. Lobanov and A. R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, *Chem. Rev.*, 1996, **96**, 1027–1044.

[208] Z. Zhou and R. G. Parr, Activation hardness: new index for describing the orientation of electrophilic aromatic substitution, *J. Am. Chem. Soc.*, 1990, **112**, 5720–5724.

[209] J. Gasteiger and T. Engel, *Chemoinformatics: a textbook*, John Wiley & Sons, 2006.

[210] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.*, 1964, **136**, B864.

[211] D. R. Hartree, Mathematical Proceedings of the Cambridge Philosophical Society, 1928, pp. 89–110.

[212] J. C. Slater, A simplification of the Hartree-Fock method, *Phys. Rev.*, 1951, **81**, 385.

[213] C. Møller and M. S. Plesset, Note on an approximation treatment for many-electron systems, *Phys. Rev.*, 1934, **46**, 618.

[214] I. Shavitt, in *Methods of electronic structure theory*, Springer, 1977, pp. 189–275.

[215] G. D. Purvis III and R. J. Bartlett, A full coupled-cluster singles and doubles model: The inclusion of disconnected triples, *J. Chem. Phys.*, 1982, **76**, 1910–1918.

[216] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *J. Phys. Rev.*, 1965, **140**, A1133.

[217] L. Zhang, G.-F. Hao, Y. Tan, Z. Xi, M.-Z. Huang and G.-F. Yang, Bioactive conformation analysis of cyclic imides as protoporphyrinogen oxidase inhibitor by combining DFT calculations, QSAR and molecular dynamic simulations, *Bioorg.Med. Chem.*, 2009, **17**, 4935–4942.

[218] B. F. Rasulev, N. D. Abdullaev, V. N. Syrov and J. Leszczynski, A Quantitative Structure-Activity Relationship (QSAR) Study of the Antioxidant Activity of Flavonoids, *Mol. Inform.*, 2005, **24**, 1056–1065.

[219] J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.*, 1996, **77**, 3865.

[220] A. J. Cohen, P. Mori-Sánchez and W. Yang, Challenges for density functional theory, *Chem. Rev.*, 2011, **112**, 289–320.

[221] A. D. Becke, A new mixing of Hartree–Fock and local density-functional theories, *J. Chem. Phys.*, 1993, **98**, 1372–1377.

[222] A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

[223] B. Miehlich, A. Savin, H. Stoll and H. Preuss, Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr, *Chem. Phys. Lett.*, 1989, **157**, 200–206.

[224] P. C. Hariharan and J. A. Pople, The influence of polarization functions on molecular orbital hydrogenation energies, *Theoret. Chim. Acta*, 1973, **28**, 213–222.

[225] M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements, *J. Chem. Phys.*, 1982, **77**, 3654–3665.

[226] R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions, *J. Chem. Phys.*, 1980, **72**, 650–654.

[227] T. Clark, J. Chandrasekhar, G. W. Spitznagel and P. V. R. Schleyer, Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+ G basis set for first-row elements, Li–F, *J. Comp. Chem.*, 1983, **4**, 294–301.

[228] F. Pohl, Empirical protein energy maps, *Nat. New Biol.*, 1971, **234**, 277.

[229] O. Gascuel and J. Golmard, A simple method for predicting the secondary structure of globular proteins: implications and accuracy, *Bioinform.*, 1988, **4**, 357–365.

[230] L. E. Chirlian and M. M. Francl, Atomic charges derived from electrostatic potentials: A detailed study, *J. Comput. Chem.*, 1987, **8**, 894–905.

[231] C. M. Breneman and K. B. Wiberg, Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis, *J. Comput. Chem.*, 1990, **11**, 361–373.

[232] J. Rigby and E. I. Izgorodina, Assessment of atomic partial charge schemes for polarisation and charge transfer effects in ionic liquids, *Phys. Chem. Chem. Phys.*, 2013, **15**, 1632–1646.

[233] F. Martin and H. Zipse, Charge distribution in the water molecule—a comparison of methods, *J. Comput. Chem.*, 2005, **26**, 97–105.

[234] I. Fleming, *Frontier orbitals and organic chemical reactions*, Wiley, 1976.

[235] J. I. Aihara, Reduced HOMO- LUMO gap as an index of kinetic stability for polycyclic aromatic hydrocarbons, *J. Phys. Chem. A*, 1999, **103**, 7487–7495.

[236] Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kús, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio Jr., H. Dop, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Lu-

enser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, P. A. Pieniazek, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, N. Sergueev, S. M. Sharada, S. Sharmaa, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, V. Vanovschi, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhou, B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xua, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill and M. Head-Gordon, Advances in molecular quantum chemistry contained in the Q-Chem 4 program package, *Mol. Phys.*, 2015, **113**, 184–215.

[237] R. Vargas, J. Garza, D. Dixon and B. P. Hay, C (sp2)- C (Aryl) Bond Rotation Barrier in N-Methylbenzamide, *J. Phys. Chem. A*, 2001, **105**, 774–778.

[238] R. J. Abraham, N. Aboitiz, M. Filippi, E. Genesio, P. Piaggio and F. Sancassan, Conformational analysis, part 43. A theoretical and LIS/NMR investigation of the conformations of substituted benzamides, *Magn. Reson. Chem.*, 2015, **53**, 498–508.

[239] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Model*, 1988, **28**, 31–36.

[240] R. Falcioni, L. Cimino, M. P. Gentileschi, I. D'Agnano, G. Zupi, S. J. Kennel

and A. Sacchi, Expression of $\beta 1$, $\beta 3$, $\beta 4$, and $\beta 5$ integrins by human lung carcinoma cells of different histotypes, *Exp. Cell. Res.*, 1994, **210**, 113–122.

[241] S. M. Wilting, R. A. van Boerdonk, F. E. Henken, C. J. Meijer, B. Diosdado, G. A. Meijer, C. le Sage, R. Agami, P. J. Snijders and R. D. Steenbergen, Methylation-mediated silencing and tumour suppressive function of hsa-miR-124 in cervical cancer, *Mol. Cancer*, 2010, **9**, 167.

[242] Z. Liu, R. C. Remsing, D. Liu, G. Moyna and V. Pophristic, Hydrogen bonding in ortho-substituted arylamides: The influence of protic solvents, *J. Phys. Chem. B*, 2009, **113**, 7041–7044.

[243] R. Banerjee, G. R. Desiraju, R. Mondal and J. A. Howard, Organic Chlorine as a Hydrogen-Bridge Acceptor: Evidence for the Existence of Intramolecular OH ClC Interactions in Some gem-Alkynols, *Chem. Eur. J.*, 2004, **10**, 3373–3383.

[244] J. A. Howard, V. J. Hoy, D. O'Hagan and G. T. Smith, How good is fluorine as a hydrogen bond acceptor?, *Tetrahedron*, 1996, **52**, 12613–12622.

[245] J. D. Dunitz, Organic fluorine: odd man out, *ChemBioChem*, 2004, **5**, 614–621.

[246] J. F. Galan, J. Brown, J. L. Wildin, Z. Liu, D. Liu, G. Moyna and V. Pophristic, Intramolecular hydrogen bonding in ortho-substituted arylamide oligomers: a computational and experimental study of ortho-fluoro-and ortho-chloro-N-methylbenzamides, *J.Phys. Chem. B*, 2009, **113**, 12809–12815.

[247] D. W. Pratt, High resolution spectroscopy in the gas phase: even large molecules have well-defined shapes, *Ann. Rev. Phys. Chem.*, 1998, **49**, 481–530.

[248] P. Breen, J. Warren, E. Bernstein and J. I. Seeman, A study of nonrigid aromatic molecules by supersonic molecular jet spectroscopy. I. Toluene and the xylenes, *J. Chem. Phys.*, 1987, **87**, 1917–1926.

[249] D. R. Borst and D. W. Pratt, Toluene: Structure, dynamics, and barrier to methyl group rotation in its electronically excited state. A route to IVR, *J Chem. Phys.*, 2000, **113**, 3658–3669.

[250] A. L. East, H. Liu, E. C. Lim, P. Jensen, I. Déchene, M. Z. Zgierski, W. Siebrand and P. Bunker, Toluene internal-rotation: Measurement and simulation of the high-resolution S 1–S 0 fluorescence excitation spectrum at 0.5 K, *J. Chem. Phys.*, 2000, **112**, 167–175.

[251] C. Serralheiro, D. Duflot, F. F. da Silva, S. V. Hoffmann, N. C. Jones, N. J. Mason, B. Mendes and P. Limao-Vieira, Toluene valence and Rydberg excitations as studied by ab initio calculations and vacuum ultraviolet (VUV) synchrotron radiation, *J. Phys. Chem. A*, 2015, **119**, 9059–9069.

[252] J. Hopkins, D. Powers and R. Smalley, Vibrational relaxation in jet-cooled alkyl benzenes. I. Absorption spectra, *J. Chem. Phys.*, 1980, **72**, 5039–5048.

[253] C. G. Hickman, J. R. Gascooke and W. D. Lawrance, The S_1-S_0 ($1B_2$-$1A_1$) transition of jet-cooled toluene: Excitation and dispersed fluorescence spectra, fluorescence lifetimes, and intramolecular vibrational energy redistribution, *J. Chem. Phys.*, 1996, **104**, 4887–4901.

[254] C. Minejima, T. Ebata and N. Mikami, C–H stretching vibrations of benzene and toluene in their S 1 states observed by double resonance vibrational spectroscopy in supersonic jets, *Phys. Chem. Chem. Phys.*, 2002, **4**, 1537–1541.

[255] J. A. Davies, A. M. Green and K. L. Reid, Deducing anharmonic coupling matrix elements from picosecond time-resolved photoelectron spectra: application to S 1 toluene at low vibrational energy, *Phys. Chem. Chem. Phys.*, 2010, **12**, 9872–9883.

[256] A. M. Gardner, A. M. Green, V. M. Tamé-Reyes, V. H. Wilton and T. G. Wright, Vibrations of the low energy states of toluene (X~ 1 A 1 and A~ 1 B 2) and the toluene cation (X~ 2 B 1), *J. Chem. Phys.*, 2013, **138**, 134303.

[257] P. T. Whiteside, A. K. King, J. A. Davies, K. L. Reid, M. Towrie and P. Matousek, Photoelectron spectroscopy of S 1 toluene: II. Intramolecular dynamics of selected vibrational levels in S 1 toluene studied by nanosecond and picosecond time-resolved photoelectron spectroscopies, *J. Chem. Phys.*, 2005, **123**, 204317.

[258] T. Suzuki, Ultrafast Internal Conversion of Aromatic Molecules Studied by Photoelectron Spectroscopy using Sub-20 fs Laser Pulses, *Molecules*, 2014, **19**, 2410–2433.

[259] E. A. Virgo, J. R. Gascooke and W. D. Lawrance, The intensity of forbidden torsional transitions in electronic spectra of molecules with a 6-fold barrier: Application to toluenes, *J. Chem. Phys.*, 2014, **140**, 154310.

[260] T. Iijima, An Electron-Diffraction Investigation of the Molecular Structure of Toluene, *Z. Naturforsch A*, 1977, **32**, 1063–1064.

[261] J. Kapitan, L. Hecht and P. Bouř, Raman spectral evidence of methyl rotation in liquid toluene, *Phys. Chem. Chem. Phys.*, 2008, **10**, 1003–1008.

[262] P. Chen and C. Wu, Ab Initio Molecular Orbital Study of Toleune and Its Derivatives, *J. Phys. Chem.*, 1995, **99**, 15023–15027.

[263] H. F. Hameka and J. O. Jensen, Calculations of the energies, geometries and vibrational frequencies of the ground and first excited singlet states of toluene and p-cresol, *J. Mol. Struct.*, 1995, **331**, 203–214.

[264] Y.-I. Suzuki, T. Horio, T. Fuji and T. Suzuki, Time-resolved photoelectron imaging of S2→ S1 internal conversion in benzene and toluene, *J. Chem. Phys.*, 2011, **134**, 184313.

[265] A. M. Gardner and T. G. Wright, Consistent assignment of the vibrations of monosubstituted benzenes, *J. Chem. Phys.*, 2011, **135**, 114305.

[266] R. S. Mulliken, Electronic population analysis on LCAO–MO molecular wave functions. I, *J. Chem. Phys.*, 1955, **23**, 1833–1840.

[267] G. Herzberg, *Infrared and Raman spectra of polyatomic molecules*, D. Van Nostrand Company.; New York, 1945.

[268] B. O. Roos, P. R. Taylor, P. E. Si *et al.*, A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach, *J. Chem. Phys.*, 1980, **48**, 157–173.

[269] H.-J. Werner and P. J. Knowles, A second order multiconfiguration SCF procedure with optimum convergence, *J. Chem. Phys.*, 1985, **82**, 5053–5063.

[270] S. Cogan, Y. Haas and S. Zilberg, Intersystem crossing at singlet conical intersections, *J. Photochem. Photobiol. A*, 2007, **190**, 200–206.

[271] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz, Molpro: a general-purpose quantum chemistry program package, *Interdiscip. Rev. Comput. Mol. Sci.*, 2012, **2**, 242–253.

[272] H.-J. Werner, Third-order multireference perturbation theory The CASPT3 method, *Molecular Physics*, 1996, **89**, 645–661.

[273] P. Celani and H.-J. Werner, Multireference perturbation theory for large restricted and selected active space reference wave functions, *J. Chem. Phys.*, 2000, **112**, 5546–5557.

[274] E. Runge and E. K. Gross, Density-functional theory for time-dependent systems, *Phys. Rev. Lett.*, 1984, **52**, 997.

[275] J.-D. Chai and M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.

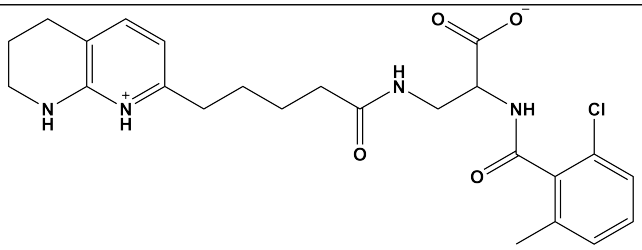
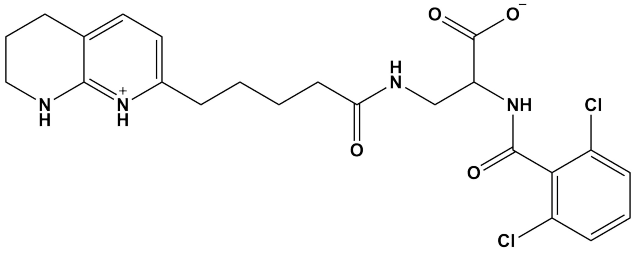
[276] R. Du Bois, Strategies for treating idiopathic pulmonary fibrosis, *Nat. Rev. Drug Discov.*, 2010, **9**, 129.

[277] Z. Huang, H. Li, Q. Zhang, F. Lu, M. Hong, Z. Zhang, X. Guo, Y. Zhu, S. Li and H. Liu, Discovery of Indolinone-Based Multikinase Inhibitors as Potential Therapeutics for Idiopathic Pulmonary Fibrosis, *ACS Med. Chem. Lett.*, 2017, **8**, 1142–1147.

Appendix A

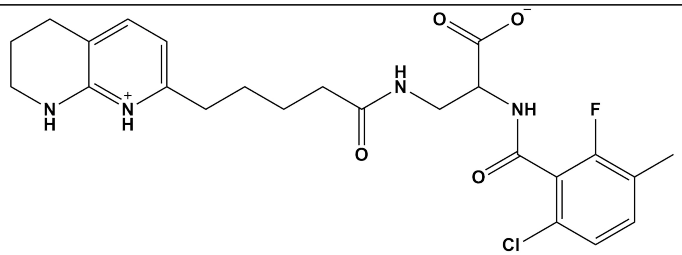
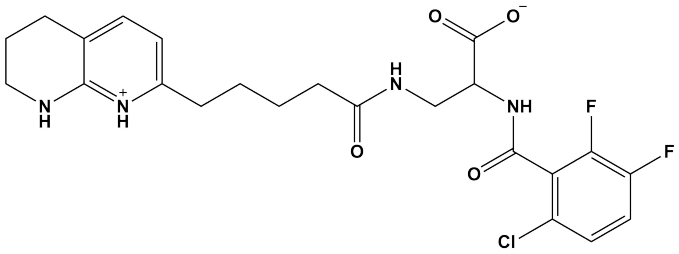
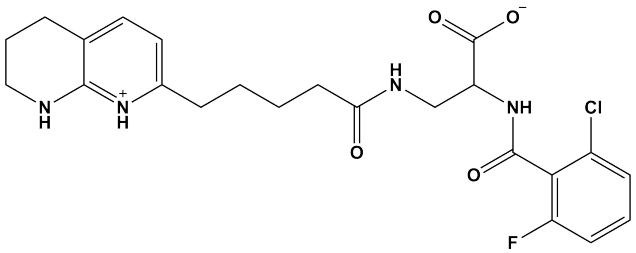
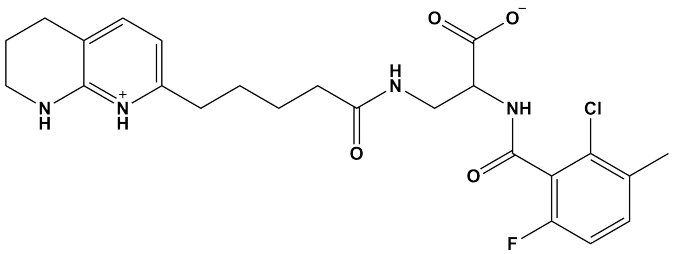
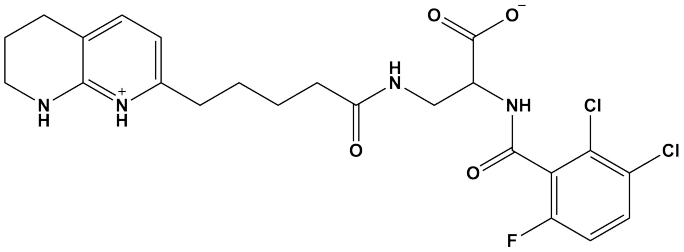
A.1 Data Set

Table A.1: Datasets G1 and G2 comprised compounds 1 to 27 and datasets G3 and G4 comprised (1 - 40) compounds with their experimental pIC_{50} values. Dataset G5 comprised all compounds (1 - 46).

Number	**Structure**	$\alpha_v\beta_1$	$\alpha_v\beta_6$
1		8.3	7.0
2		8.8	6.4

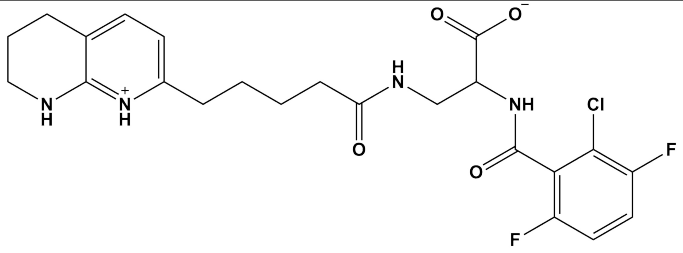
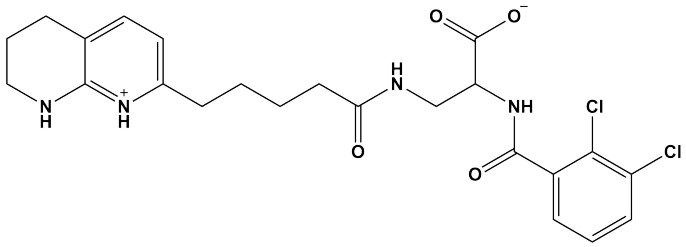
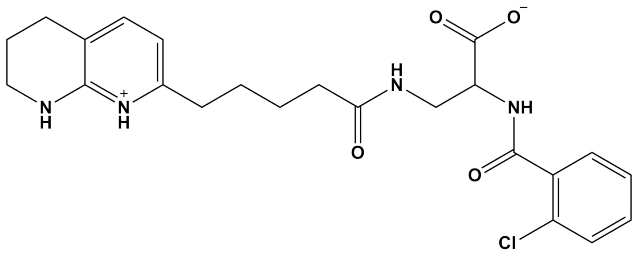
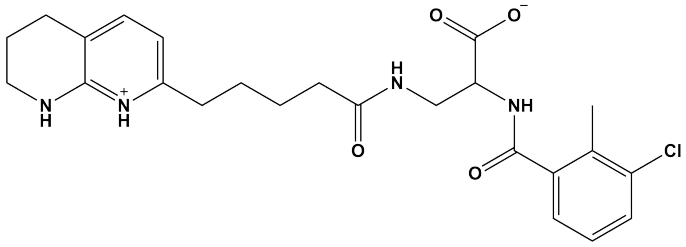
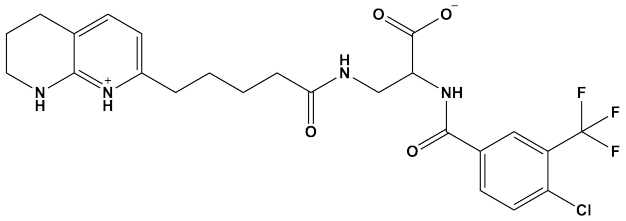
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
3		8.6	7.3
4		8.5	7.8
5		8.5	7.6
6		7.8	7.2
7		8.2	7.8

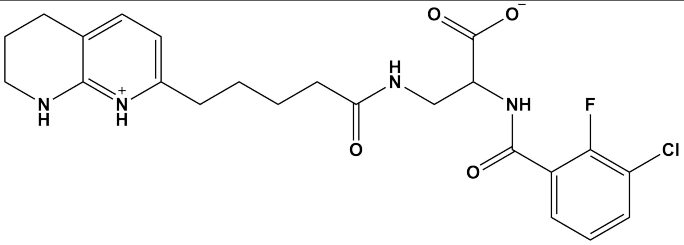
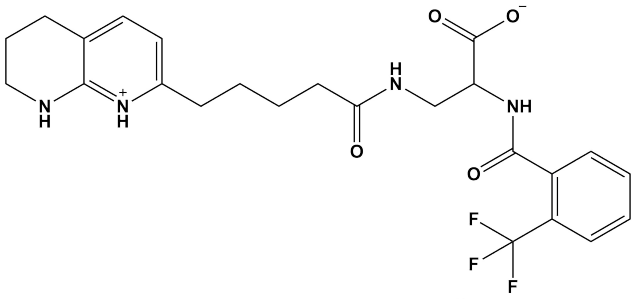
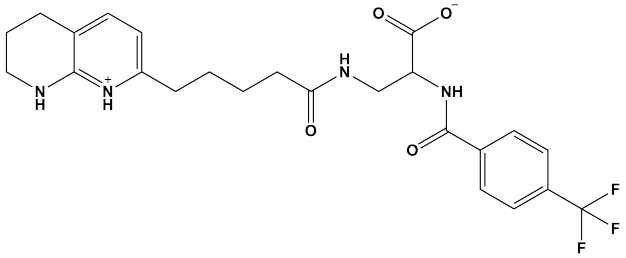
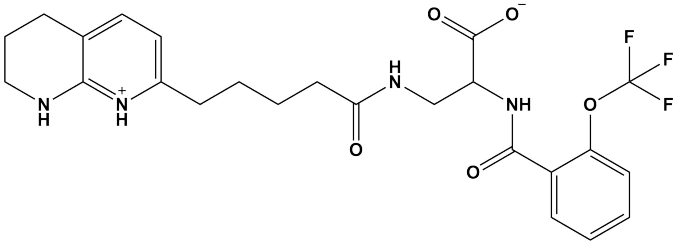
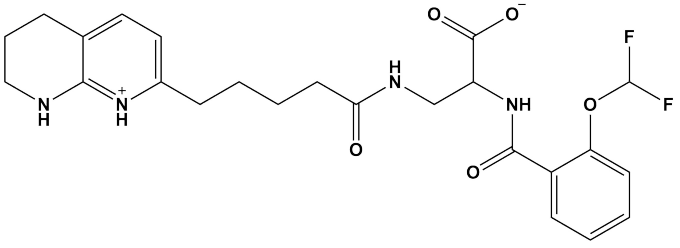
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
8		8.6	8.0
9		7.7	7.8
10		8.2	6.6
11		7.9	7.7
12		6.7	6.4

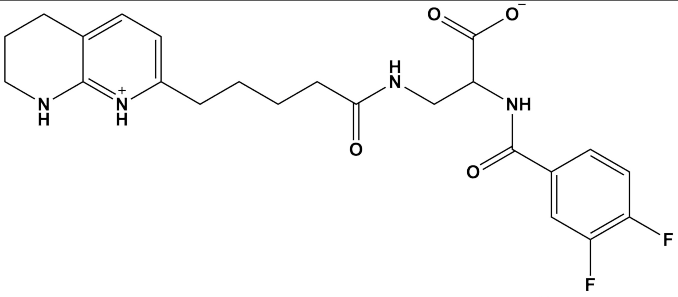
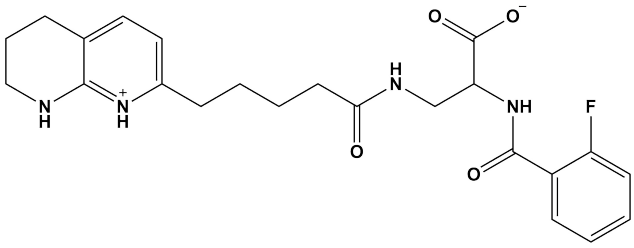
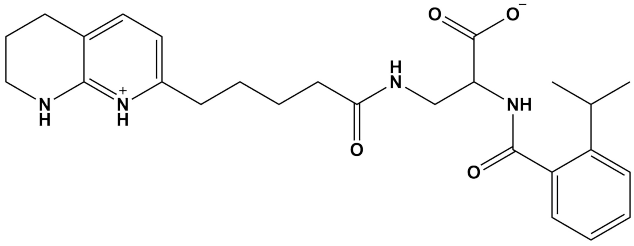
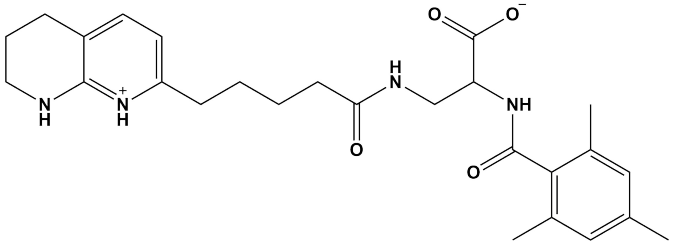
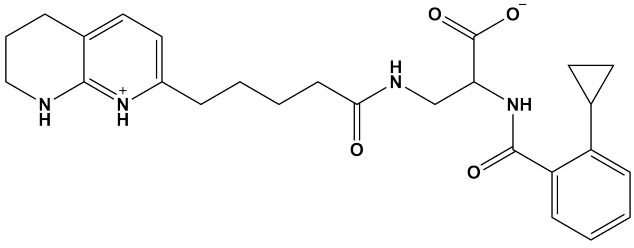
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
13		8.1	7.4
14		8.1	6.5
15		7.5	5.3
16		7.3	7.0
17		7.6	6.9

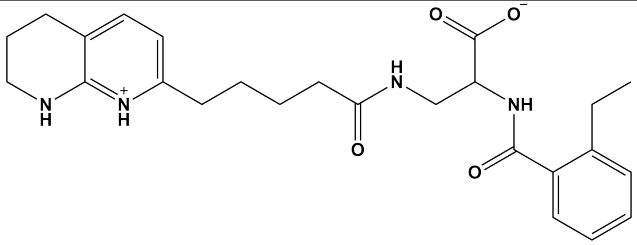
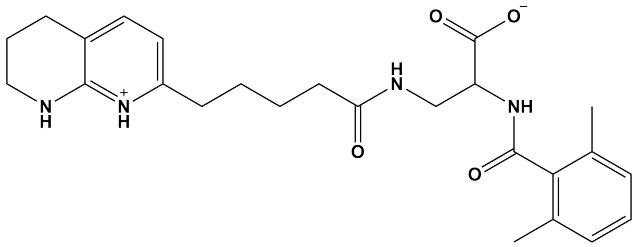
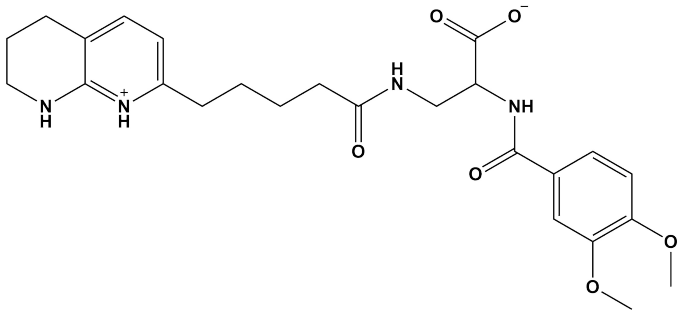
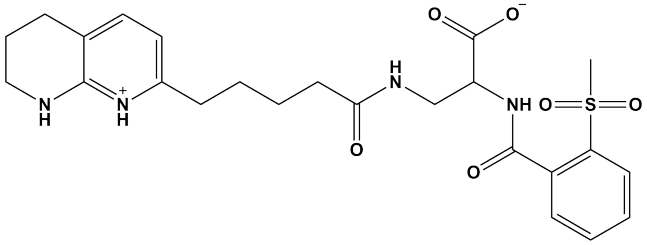
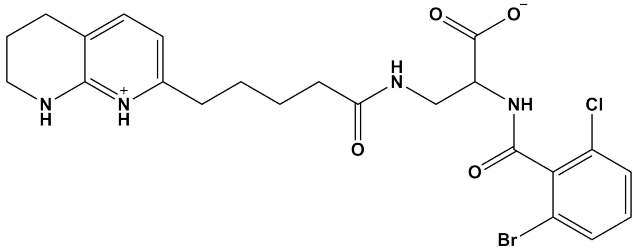
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
18		7.2	6.2
19		7.7	5.8
20		6.6	5.4
21		8.0	7.3
22		7.5	6.6

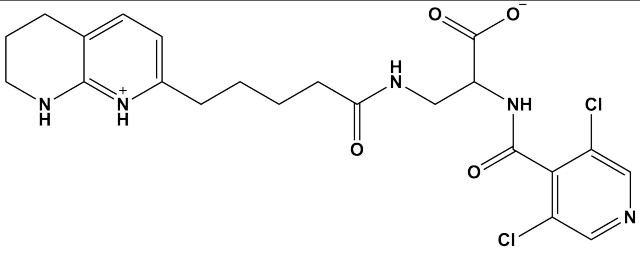
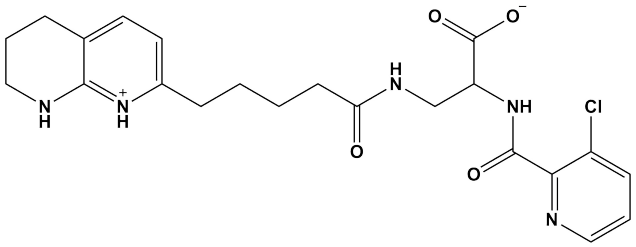
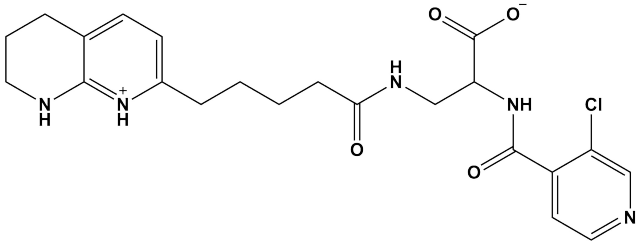
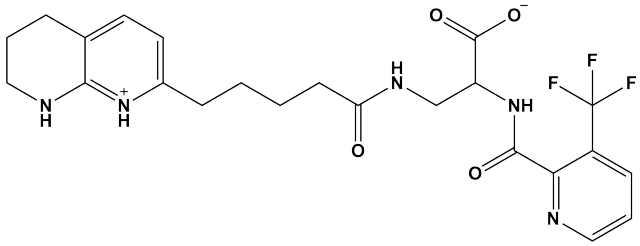
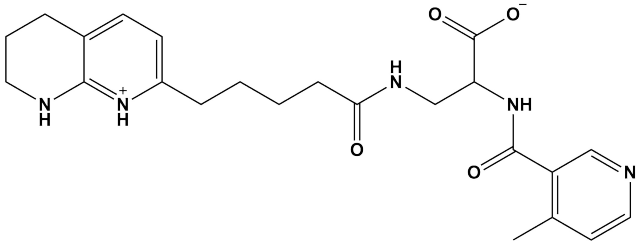
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
23		7.8	7.1
24		8.2	6.2
25		7.1	< 5.0
26		7.0	6.2
27		8.3	6.8

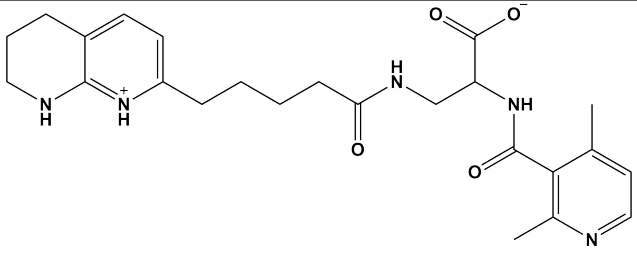
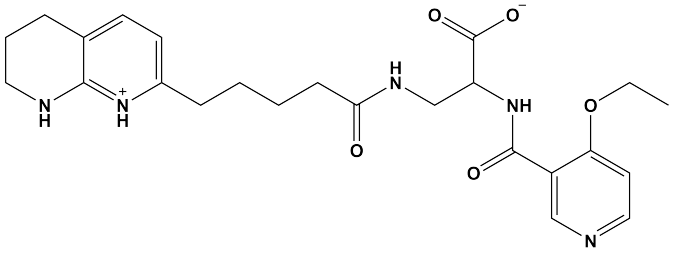
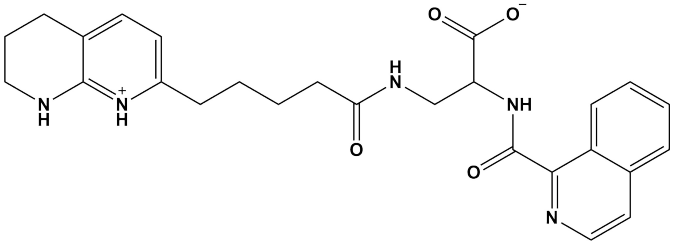
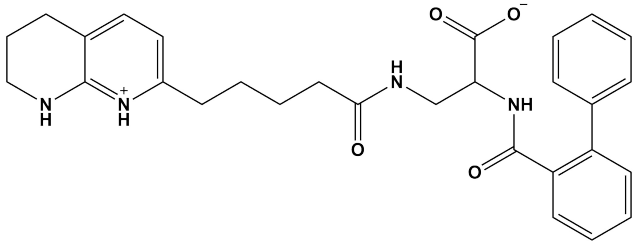
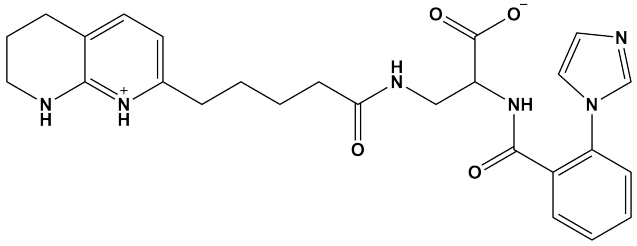
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
28		8.8	7.4
29		7.5	6.4
30		8.4	7.2
31		7.0	6.5
32		7.4	6.3

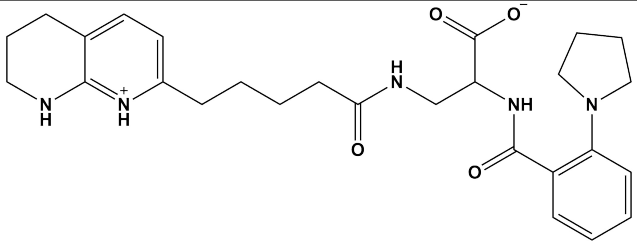
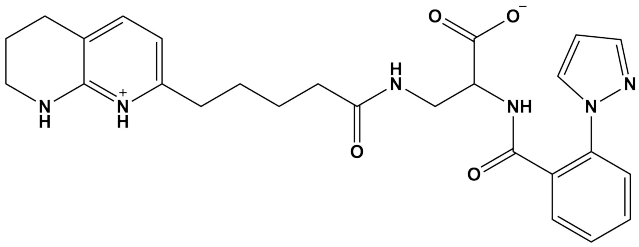
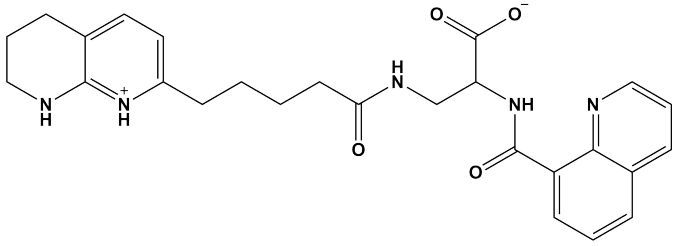
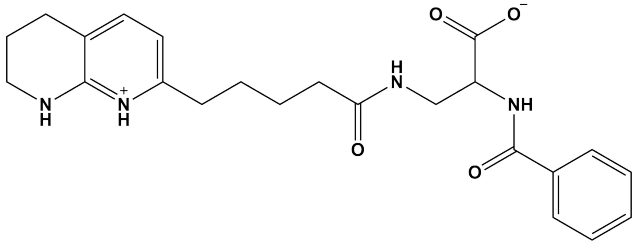
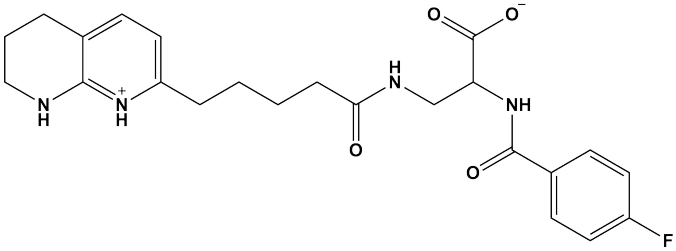
Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
33		7.5	5.8
34		6.8	5.1
35		6.2	6.0
36		6.8	6.4
37		6.9	6.6

Continued on next page

Table A.1 – continued from previous page

Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
38		7.2	8.1
39		6.8	6.8
40		6.1	5.2
41		-	5.9
42		-	6.3

Continued on next page

Table A.1 – continued from previous page

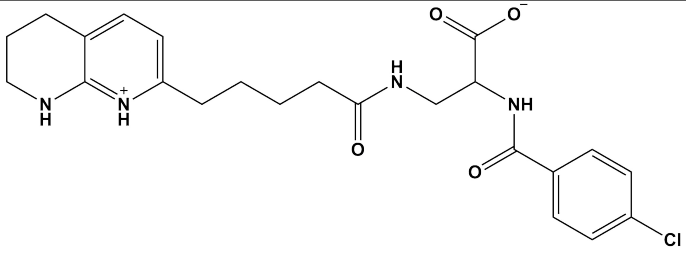
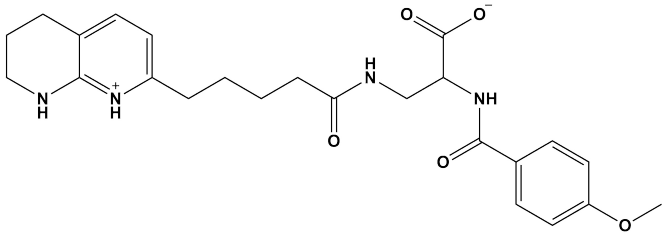
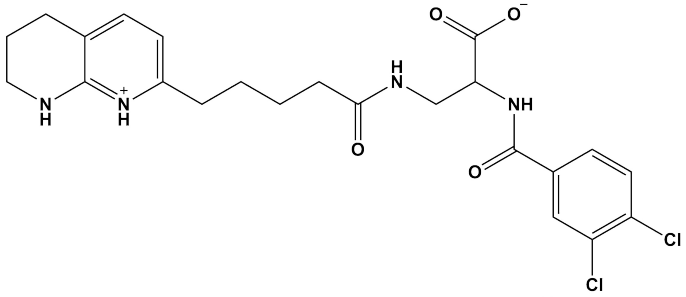
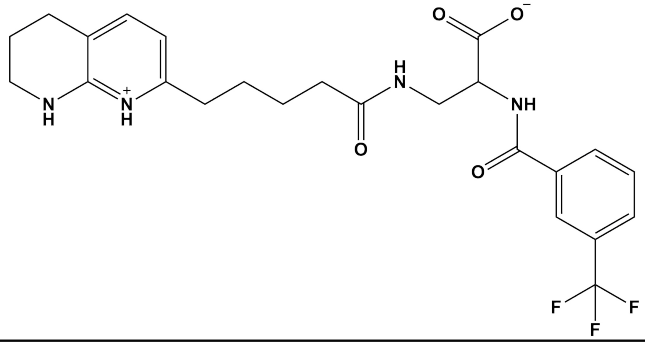
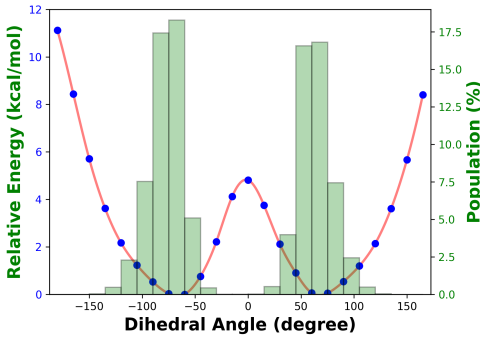
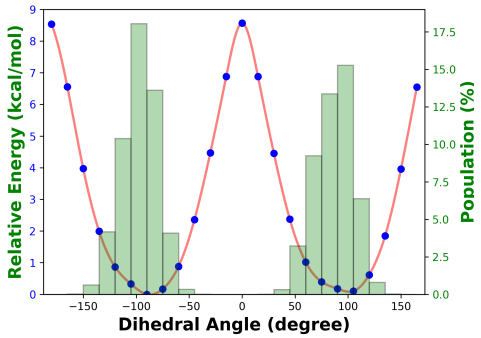
Number	Structure	$\alpha_v\beta_1$	$\alpha_v\beta_6$
43		-	6.1
44		-	5.5
45		-	6.0
46		-	5.9

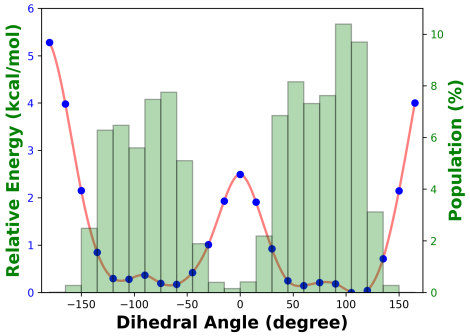
Figure A.1: The PES scan for benzamide molecules (**1 - 6**)



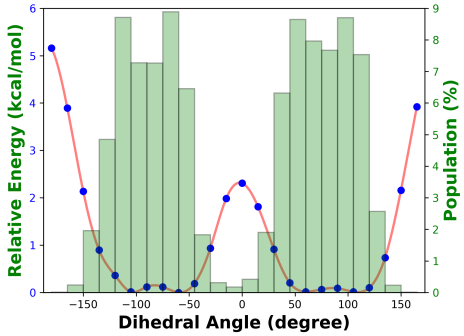
(a) Molecule **1**



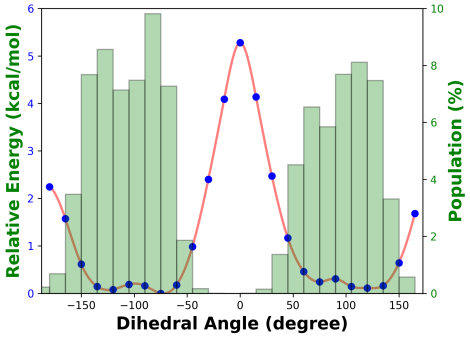
(b) Molecule **2**



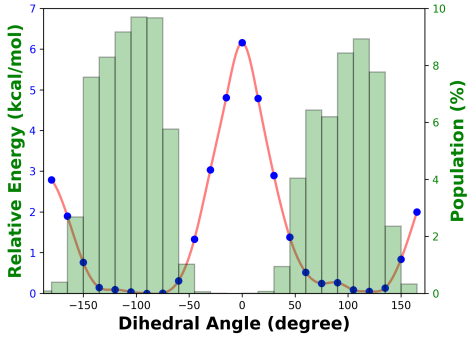
(c) Molecule **3**



(d) Molecule **4**

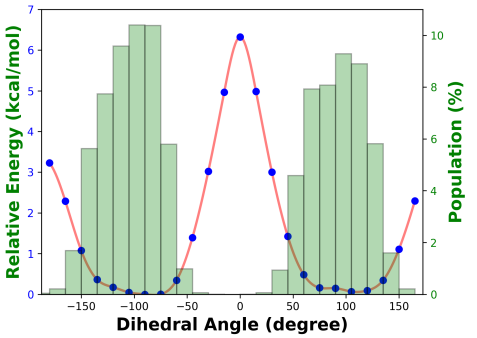


(e) Molecule **5**

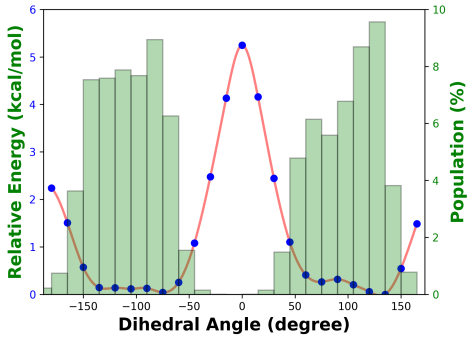


(f) Molecule **6**

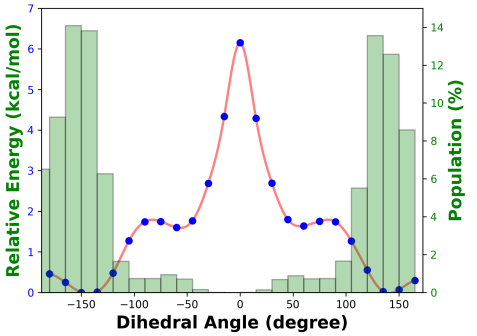
Figure A.2: The PES scan for benzamide molecules (**7 - 12)**



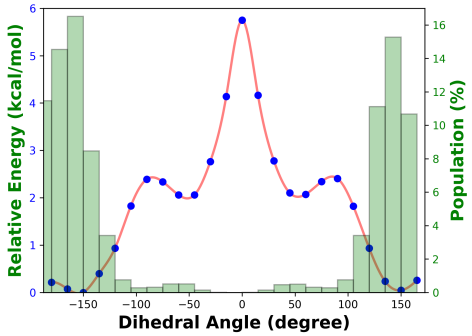
(a) Molecule **7**



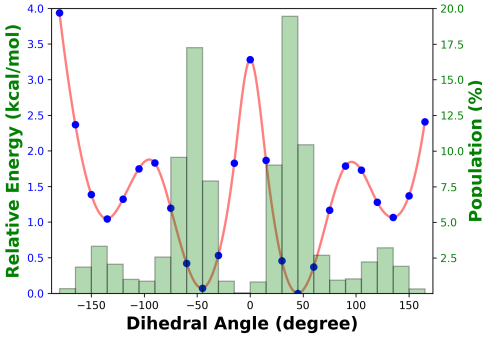
(b) Molecule **8**



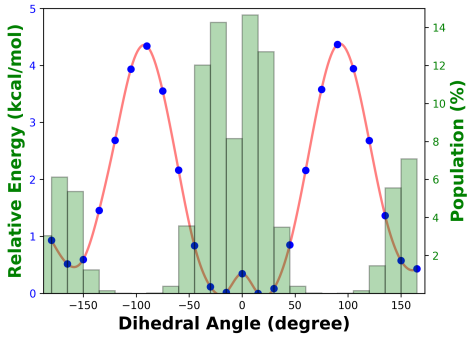
(c) Molecule **9**



(d) Molecule **10**

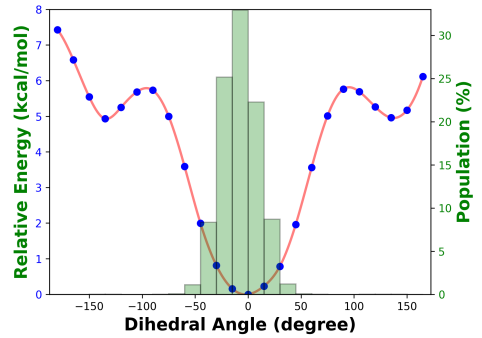


(e) Molecule **11**

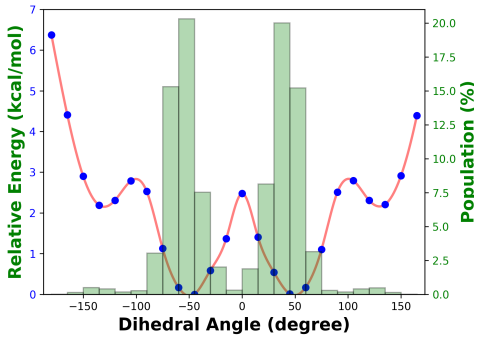


(f) Molecule **12**

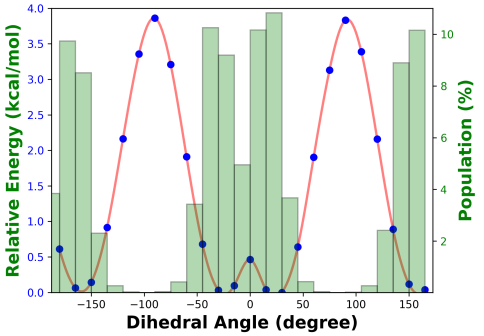
Figure A.3: The PES scan for benzamide molecules (**13 - 18**)



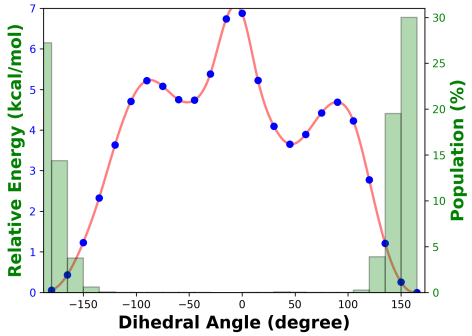
(a) Molecule **13**



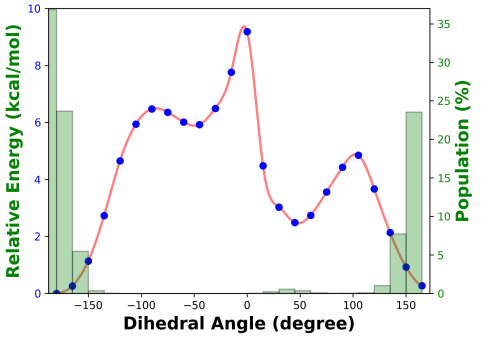
(b) Molecule **14**



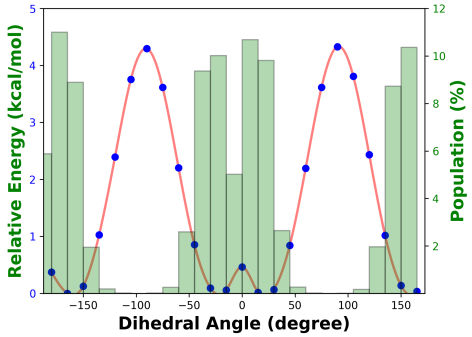
(c) Molecule **15**



(d) Molecule **16**

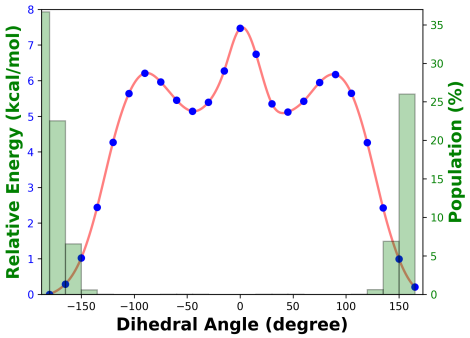


(e) Molecule **17**

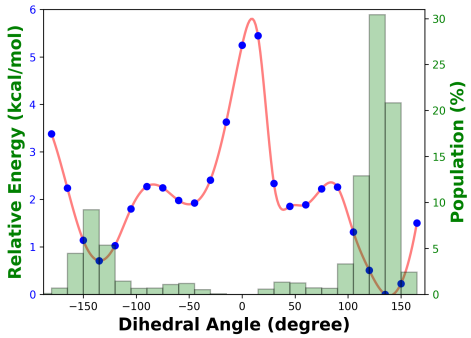


(f) Molecule **18**

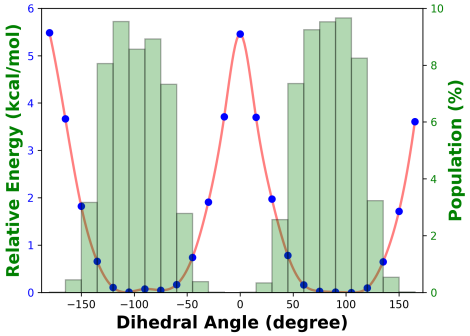
Figure A.4: The PES scan for benzamide molecules (**19 - 24**)



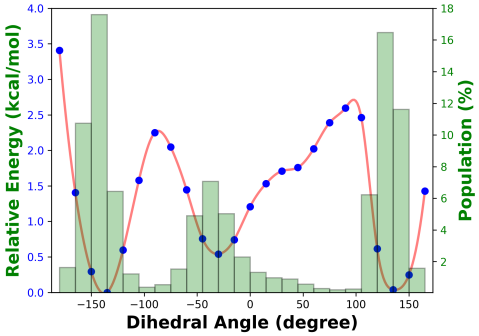
(a) Molecule **19**



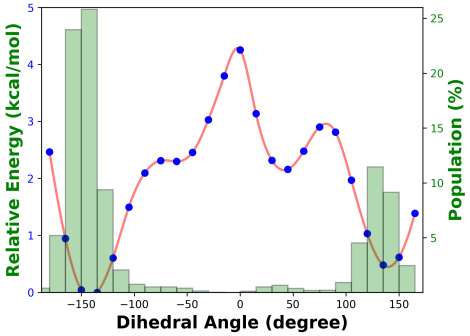
(b) Molecule **20**



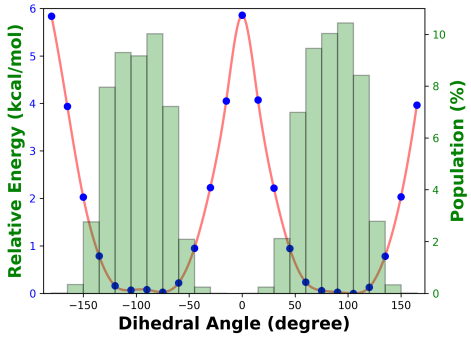
(c) Molecule **21**



(d) Molecule **22**

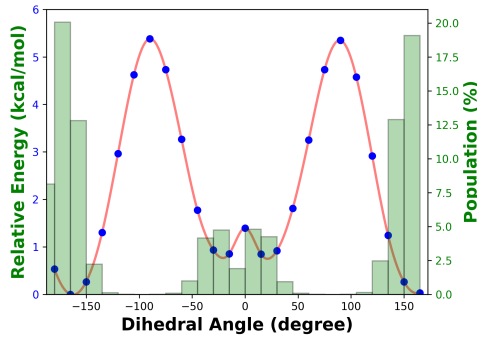


(e) Molecule **23**

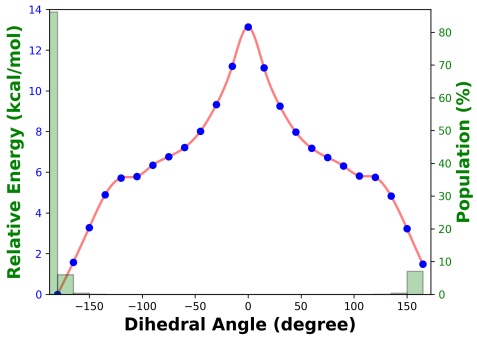


(f) Molecule **24**

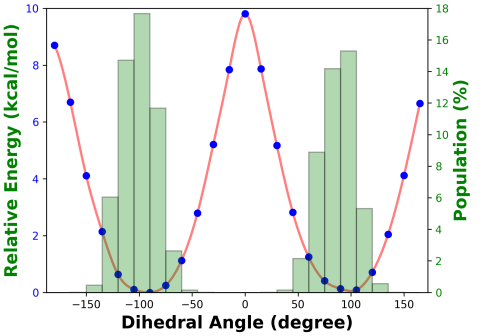
Figure A.5: The PES scan for benzamide molecules (**25 - 30**)



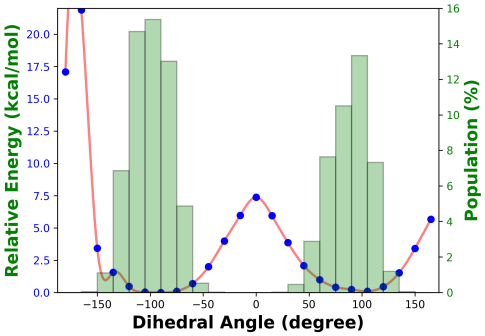
(a) Molecule **25**



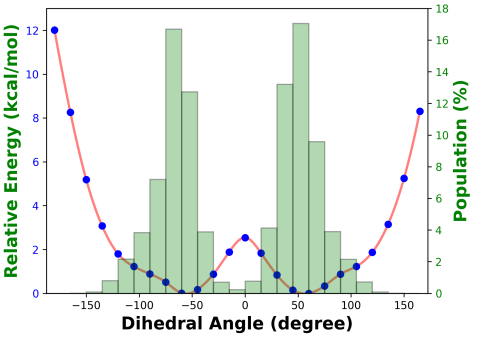
(b) Molecule **26**



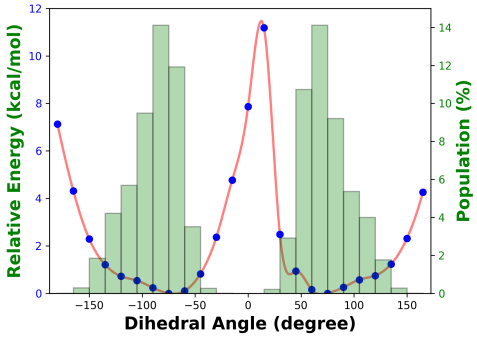
(c) Molecule **27**



(d) Molecule **28**

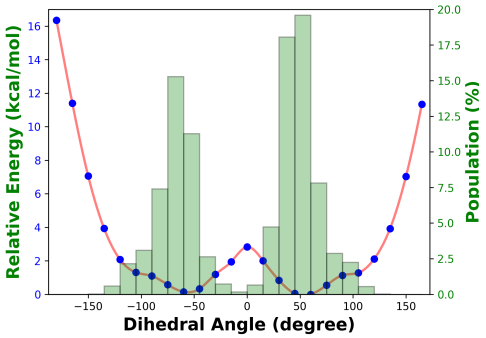


(e) Molecule **29**

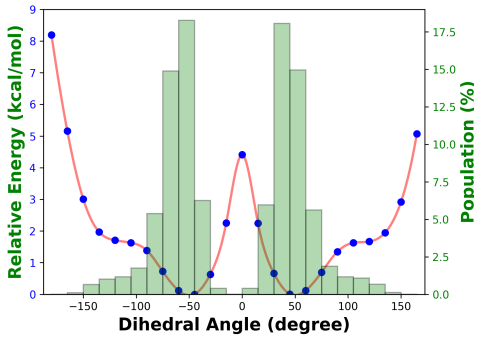


(f) Molecule **30**

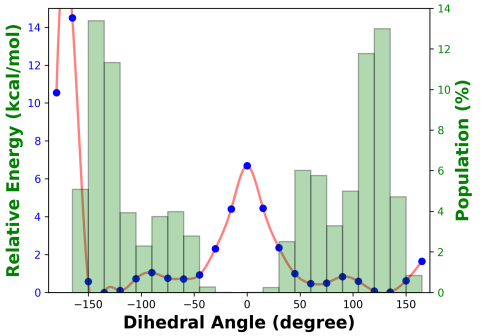
Figure A.6: The PES scan for benzamide molecules (**31 - 36**)



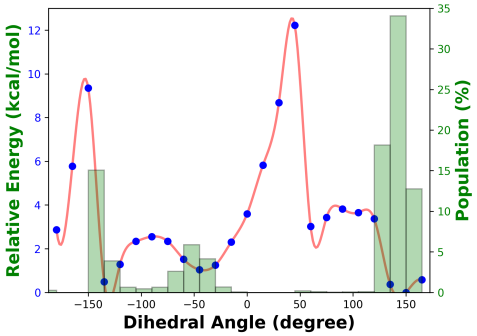
(a) Molecule **31**



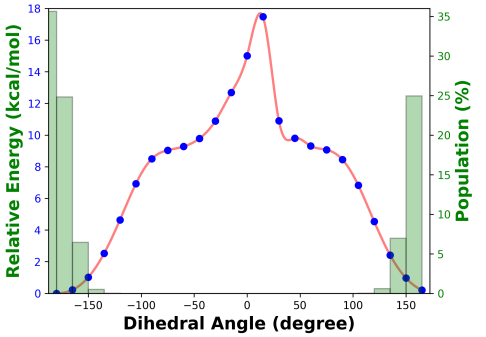
(b) Molecule **32**



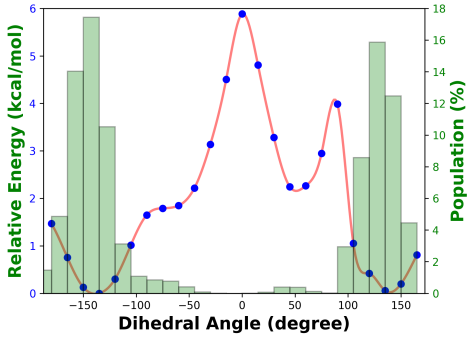
(c) Molecule **33**



(d) Molecule **34**

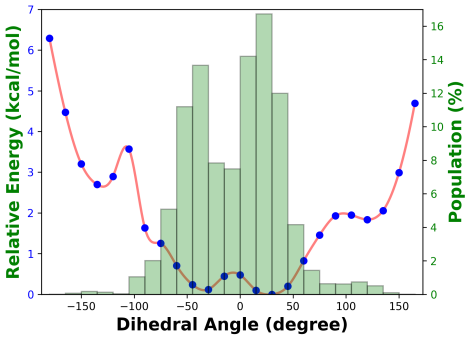


(e) Molecule **35**

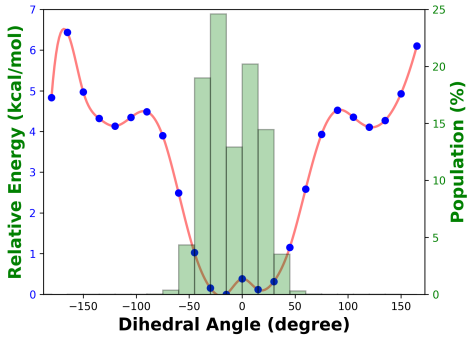


(f) Molecule **36**

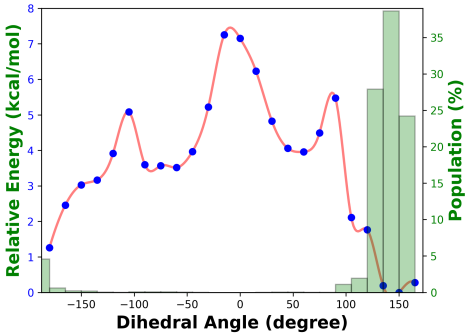
Figure A.7: The PES scan for benzamide molecules (**37 - 42**)



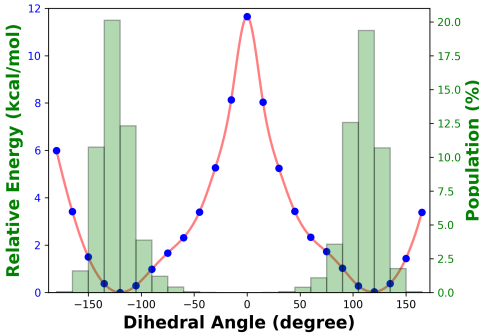
(a) Molecule **37**



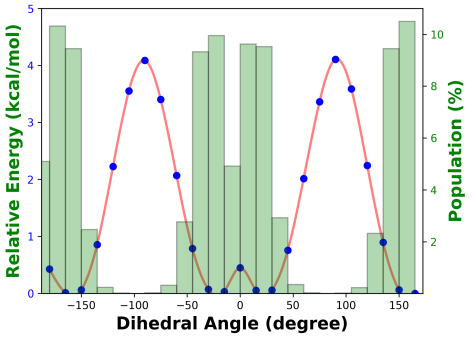
(b) Molecule **38**



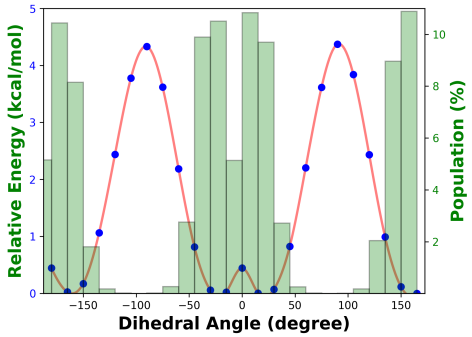
(c) Molecule **39**



(d) Molecule **40**

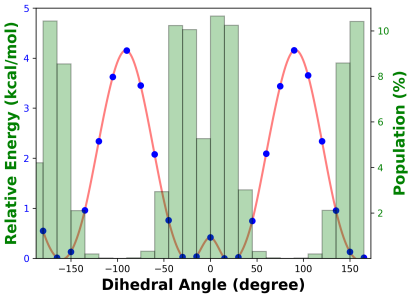


(e) Molecule **41**

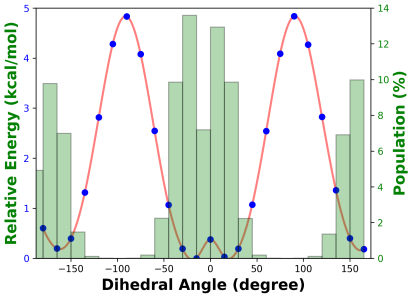


(f) Molecule **42**

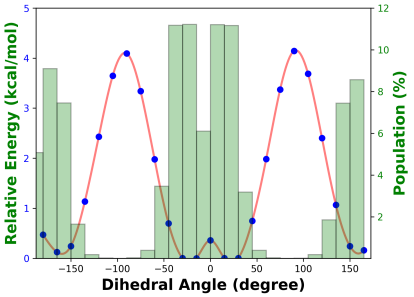
Figure A.8: The PES scan for benzamide molecules (**43 - 46**)



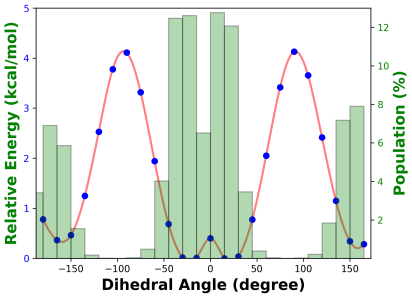
(a) Molecule **43**



(b) Molecule **44**



(c) Molecule **45**



(d) Molecule **46**