The Prediction of Mutagenicity and pK_a for Pharmaceutically Relevant Compounds Using "Quantum Chemical Topology" Descriptors

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

2010

Alexander Paul Harding

School of Chemistry

Contents

Preliminary Pages

Word count = 54,000

List of Tables

List of Figures

Abstract

Quantum Chemical Topology (QCT) descriptors, calculated from *ab initio* wave functions, have been utilised to model pK_a and mutagenicity for data sets of pharmaceutically relevant compounds. The pK_a of a compound is a pivotal property in both life science and chemistry since the propensity of a compound to donate or accept a proton is fundamental to understanding chemical and biological processes. The prediction of mutagenicity, specifically as determined by the Ames test, is important to aid medicinal chemists select compounds avoiding this potential pitfall in drug design. Carbocyclic and heterocyclic aromatic amines were chosen because this compounds class is synthetically very useful but also prone to positive outcomes in the battery of genotoxicity assays.

The importance of pK_a and genotoxic characteristics cannot be overestimated in drug design, where the multivariate optimisations of properties that influence the Absorption-Distribution-Metabolism-Excretion-Toxicity (ADMET) profiles now features very early on in the drug discovery process.

Models were constructed using carboxylic acids in conjunction with the Quantum Topological Molecular Similarity (QTMS) method. The models produced Root Mean Square Error of Prediction (RMSEP) values of less than 0.5 pK_a units and compared favourably to other pK_a prediction methods. The ortho-substituted benzoic acids had the largest RMSEP which was significantly improved by splitting the compounds into high-correlation subsets. For these subsets, single-term equations containing one *ab initio* bond length were able to accurately predict pK_a. The pK_a prediction equations were extended to phenols and anilines.

Quantitative Structure Activity Relationship (QSAR) models of acceptable quality were built based on literature data to predict the mutagenic potency (LogMP) of carbo- and heterocyclic aromatic amines using QTMS. However, these models failed to predict Ames test values for compounds screened at GSK. Contradictory internal and external data for several compounds motivated us to determine the fidelity of the Ames test for this compound class. The systematic investigation involved recrystallisation to purify compounds, analytical methods to measure the purity and finally comparative Ames testing. Unexpectedly, the Ames test results were very reproducible when 14 representative repurified molecules were tested as the freebase and the hydrochloride salt in two different solvents (water and DMSO). This work formed the basis for the analysis of Ames data at GSK and a systematic Ames testing programme for aromatic amines. So far, an unprecedentedly large list of 400 compounds has been made available to guide medicinal chemists. We constructed a model for the subset of 100 meta-/para-substituted anilines that could predict 70% of the Ames classifications. The experimental values of several of the model outliers appeared questionable after closer inspection and three of these have been retested so far. The retests lead to the reclassification of two of them and thereby to improved model accuracy of 78%. This demonstrates the power of the iterative process of model building, critical analysis of experimental data, retesting outliers and rebuilding the model.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.

Alexander Paul Harding

Copyright Statement

List of Abbreviations

AA	Carbocyclic and Heterocyclic Primary Aromatic Amine
ADMET	Absorption-Distribution-Metabolism-Excretion-Toxicity
AIM	Atoms in Molecules
BCP	Bond Critical Point
CV	Cross-Validation
CYP	Cytochrome
GSK	GlaxoSmithKline
IF	Induction Factor
LV	Latent Variable
MAE	Mean Absolute Error
MIA	1-Methyl-Imidazole-2-Amine
MLR	Multilinear Regression
MP	Mutagenic Potency
PCA	Principle Component Analysis
PCR	Principle Component Regression
PLS	Partial Least Square
QCT	Quantum Chemical Topology
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship
QTMS	Quantum Topological Molecular Similarity
RBFNN	Radial Basis Function Neural Networks
RMSE	Root Mean Squared Error
RMSEE	Root Mean Squared Error of Estimation
RMSEP	Root Mean Square Error of Prediction
S9	Rat Liver Homogenate
SAR	Structure-Activity Relationships
SVM	Support Vector Machines
VIP	Variable Importance in the Projection

Acknowledgements

I would like to extend my sincere thanks to all those who have helped me throughout the course of this doctorate. In particular, I wish to thank my supervisors Paul Popelier and Michael Kranz for their continuous support and helpful advice. I would also like to thank GlaxoSmithKline and the EPSRC for the funding. I thank all colleagues, both past and present, at University and GSK for all their help. Our fruitful discussions, both scientific and otherwise, have made the past four year rewarding and enjoyable. Finally, I would like to acknowledge both family and friends. Without their patience and support this thesis would not have been possible.

Preface

After completing my Masters in Chemistry at The University of Manchester in 2006, I subsequently enrolled, at the same institute, on the Engineering Doctorate (EngD) Programme. The EngD differs from a PhD in a number of ways. The EngD programme is four years in duration and involves company sponsorship; in this case by GlaxoSmithKline. A stipulation of the EngD is that the students spend an extended period of time working in the sponsoring company. Over the last four years, I have spent the majority of time working in the Computational Chemistry department at GSK, Stevenage.

To successfully graduate, EngD students are required to complete a postgraduate diploma (PGDip) in Management Science. The diploma involves the completion of eight modules which are assessed by a combination of examinations, group presentations and coursework. The eight modules comprise Production Systems, Managerial Economics, Total Quality Management, Individuals, Groups & Organisations, Industrial Relations, Marketing, Management Accounting and Logistics & Supply. I completed the modules in the first two years of my EngD and obtained a grade 1 classification from Manchester Business School in 2008.

Professional development is another key area of the EngD programme. The structured programme of professional development includes attendance at numerous courses, for example, Effective Communication, Management of Projects, Negotiation Skills and Industrial Law. I obtained a level 3 qualification from the Institute of Leadership and Management by completing a Leadership and Team Development one week residential course. I am a member of the Royal Society of Chemistry and a graduate member of the American Chemical Society. I have volunteered for schemes including the Institute of Physic's 'Lab in a Lorry' scheme and the RSC's e-mentoring programme which aim to interest young students in science.

Publication List

Harding, A. P.; Wedge, D. C.; Popelier, P. L. A., pK_{a} Prediction from “Quantum Chemical Topology” Descriptors. *J.Chem.Inf.Mod.* **2009,** 49, 1914–1924. (Appendix A)

Kar, S.; Harding, A. P.; Roy, K.; Popelier, P. L. A., QSAR with Quantum Topological Molecular Similarity Indices: Toxicity of Aromatic Aldehydes to *Tetrahymena Pyriformis*. *SAR and QSAR in Environmental Research* **2010,** 21, 149-168. (Appendix A)

Harding, A. P.; Popelier, P. L. A., pK_{a} Prediction from an ab initio Bond Length Part 1: Application to Phenols . *J.Chem.Inf.Mod.* (to be submitted)

Harding, A. P.; Popelier, P. L. A., pK_{a} Prediction from an ab initio Bond Length Part 2: Application to Benzoic Acids and Anilines. *J.Chem.Inf.Mod.* (to be submitted)

Harding, A. P.; Popelier, P. L. A.; Harvey, J.; Giddings, A.; Foster G.; Kranz, M., How Clean is Your Aniline? – Fidelity of the Ames Test. *Environmental and Molecular Mutagenesis.* (to be submitted)

Chapter 1

Introduction and Industrial Context

1.1 Introduction

An Engineering Doctorate (EngD) was created in collaboration with GlaxoSmithKline (GSK) and the School of Chemistry at The University of Manchester. The project focus was to use computational methods to impact on the progress of drug discovery. A specific aim of the project was to extend the use of 'Quantum Chemical Topology'[1, 2] (QCT) descriptors[3] for property predictions. Subsequently pK_a and toxicity were identified as two important molecular properties to the industrial sponsor and the wider scientific community. These selected properties enabled the extension and robust testing of QCT descriptors and also provided interesting results from which several conclusions have been drawn.

1.2 Industrial Context

1.2.1 GlaxoSmithKline (GSK)

GlaxoSmithKline (GSK) was formed on the 27th December 2000 as a result of the merger of SmithKline plc and Glaxo Wellcome plc. It is now one of the largest pharmaceutical companies in the world[4] with 100,000 employees in 114 different countries and a market share in 150 countries. Sales in 2009 amounted to £28.4 billion with profits before taxation of £8 billion. It is world leader in research-based pharmaceuticals engaged in the creation, discovery, development, manufacture and marketing of pharmaceuticals and consumer health-related products to create value for stakeholders.

The complexity of the organisation means there are numerous stakeholders including employees, customers, suppliers, regulators, charities, aid workers, environmental and animal rights campaigners, health and safety authorities, governments, local areas and site neighbours, unions, academia and competitors All have unique needs and interests. GSK's mission statement (below) places patients as central stakeholders.

> *"We have a challenging and inspiring mission to improve the quality of human life by enabling people to do more, feel better and live longer."*

At times, this may cause conflicts with other stakeholders, particularly shareholders, who have a financial interest in the company. GSK has to be profitable but has set itself the scientific goal to improve human health and therefore is actively involved in researching treatments for rare and unprofitable diseases. GSK is also involved in patent pools, where they forgo their patent rights, to provide essential medicines to developing countries. Local companies make the medicines generically at a mutually-agreed licence fee.

1.2.2 Key Products

GSK's product portfolio is divided into Pharmaceuticals and Consumer Healthcare. The principal pharmaceutical products are grouped into nine main therapeutic areas (Table 1.1). Respiratory, anti-virals and vaccines represent the three therapeutic areas which provide the largest turnover for GSK. In 2009[5] GSK had two of the world's top 60 pharmaceutical products (*Seretide/Advair* (respiratory) and *Valtrex* (anti-viral)) compared to eight in 2005[6]. A broader range of products are currently maintained to provide growth and stability as opposed to a reliance on blockbuster drugs. Sales were boosted by the H1N1 global influenza pandemic in 2009. Sales of GSK's pandemic vaccine, amounting to £883 million, contributed to the total vaccine turnover. Similarly, a significant increase in sales of *Relenza* contributed to the anti-virals turnover. The consumer healthcare portfolio comprises of over-the-counter medicines such as cold remedies and nicotine replacement therapy, oral healthcare such as *Aquafresh* and *Sensodyne* and nutrition healthcare including *Lucozade*, *Horlicks* and *Ribena*. The percentage turnover that the consumer healthcare portfolio contributes to the total is small (Table 1.2) compared to the sales generated from pharmaceutical product, but the diverse product range represents important revenue streams to GSK; which continue to grow.

Table 1.1. GSK pharmaceutical turnover by therapeutic area for 2009 and 2008.

Therapy area	**% of total 2009**	**2009**	**2008**
		£m	**£m**
Respiratory	29	6,997	5,817
Anti-virals	18	4,150	3,206
Central nervous system	8	1,870	2,897
Cardiovascular and urogenital	10	2,298	1,847
Metabolic	5	1,181	1,191
Anti-bacterial	7	1,592	1,429
Oncology and emesis	3	692	496
Vaccines	16	3,706	2,539
Other	4	1,063	959
		23,466	**20,381**

Table 1.2. GSK consumer healthcare turnover for 2009 and 2008.

	% of total 2009	2009 £m	2008 £m
Over-the-counter medicines	50	2,319	1,935
Oral healthcare	32	1,484	1,240
Nutritional healthcare	18	851	796
		4,654	**3,971**

1.2.3 Key Markets

Although GSK has a market share in 150 countries, the biggest single market is the United States. The US market grew by 3.6% in 2009 however, sales of GSK's pharmaceuticals declined by 13%. To combat the reduction in sales, GSK has placed emphasis on improving sales in the Emerging Markets (i.e. Brazil, Russia, India and China). This resulted in a 20% increase in sales in 2009 to £3 billion, representing 10% of the total sales. Operating in different markets means GSK has to be adaptable to local rules and cultures. It means the company has to liaise with different regulatory bodies such as the Food and Drug Administration (FDA) in the US and the European Medical Agency. Furthermore, direct advertising of prescription drugs is forbidden in Europe, meaning marketing is directed largely at doctors and health authorities who decide which drugs to prescribe. In the US, pharmaceutical companies directly target patients with marketing campaigns. It is expected that the recent US health care reform will have huge implication for the pharmaceutical industry in the near future.

Table 1.3. Breakdown of the value of the pharmaceutical market by geographical region.

World market by geographical region	Value £bn	% of total	Growth %
USA	**187**	**40**	**3.6**
Europe	131	**28**	4
France	25	5	-
Germany	24	5	-
Italy	16	3	-
UK	12	3	-
Rest of World	150	**32**	**9.9**
Emerging markets	55	14	-
Asia Pacific	20	4	-
Japan	50	11	-
Canada	11	2	-
Total	**468**	**100**	

1.2.4 Research and Development at GSK

The average cost of developing a new drug is estimated to be $868 million, but can vary between $500 and $2000 million depending on the therapy[7]. It also takes around twelve years for a drug to reach market and therefore the structure of GSK is vital in supporting the organisation's purpose. At a high level the business areas responsible for delivering new drugs are shown in Figure 1.1.

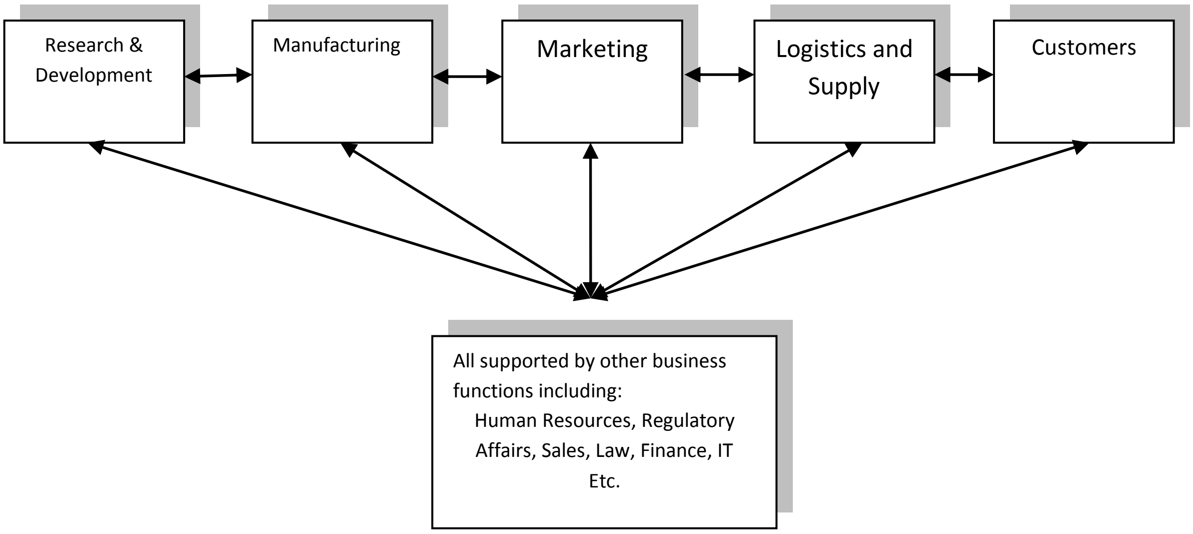

Research & Development
Manufacturing
Marketing
Logistics and Supply
Customers
All supported by other business functions including:
Human Resources, Regulatory Affairs, Sales, Law, Finance, IT
Etc.

Figure 1.1. The business functions and inter-relationships responsible for delivering new drugs to customers.

The function of Research and Development (R&D) is to deliver new drugs to the pipeline. The pressure due to this has intensified over the last few years as the increase in the discovery of novel drugs has not risen as predicted, regulatory approval has decreased and patent expiry has meant increased erosion of market share by generic drug manufacturers. To maintain a competitive pipeline, research is organised into six Centres of Excellence for Drug Discovery (CEDDS) each focusing on defined therapy areas. Separate R&D functions, for example biological pharmaceutical research, vaccine research and R&D China each have their own defined roles.

In 2008, the new Chief Executive Officer of GSK, Andrew Witty, took the CEDD model one step further and initiated the creation of a number of smaller Discovery Performance Units (DPUs) within each CEDD, to focus on a particular disease area. The DPU model was based on the success of small biotechnology companies over the last decade. There are now thirty-six DPUs, each being a compact, fully-empowered, focused and integrated team, which is responsible for a small part of the drug pipeline. Some standalone DPUs not linked to CEDDs have also been created to explore new therapy areas or new ways of working, such as the academic DPU which forms drug discovery collaborations with academia. Each DPU has a three-year business plan defining the overall budget and clear objectives. The DPUs have to present to the Drug Discovery Investment Board, which is made up of internal and external experts, who review the success and decide on future investment. This creates internal competition and an entrepreneurial culture with the potential to enhance both the scientific basis and commercial value of the business. One outcome of this initiative is that GSK and the healthcare providers are now in direct dialogue. Developing drugs that will reimburse the cost of getting the drug to market is fundamental.

Molecular Discovery Research (MDR) is an R&D function that supports the CEDDs in the entire drug discovery process. Within MDR is Computational and Structural Chemistry (CSC). The function of CSC is to provide computational chemistry support to the CEEDs and DPUs throughout the drug discovery process. The aim is to improve the quality of drug candidates, reduce attrition and as a consequence reducing costs and adding value to the business. In the United Kingdom CSC is divided into two groups, Lead Generation (LG) and Lead Optimisation (LO), each with its own manager responsible for around fifteen employees. LG and LO are sub divided into specific teams of around six employees aligned to the therapy area of each of the CEDDs with line managers responsible for each group.

Patent expiries, regulatory issues and price pressures from healthcare providers have created an environment where the sector is associated with lower growth and higher risk. Shortly after becoming CEO, Andrew Witty outlined three strategic priorities to transform GSK into a company that delivers more growth, less risk, and improved financial performance to overcome the unprecedented challenges in the pharmaceutical industry. The focus was to grow a diversified global business, deliver more products of value and simplify the operating model. The strategic priorities have caused changes across GSK and in particular R&D. The business has created a more balanced portfolio over the last two years and moved away from the emphasis to discover the next ‘blockbuster’ drug. GSK has and continues to improve its pipeline by acquiring, collaborating, and in-licensing promising compounds from other organisations. This has meant the externalisation of approximately 30% of discovery research with 47 external partners where the risks associated with drug discovery are shared[5]. The operating model is also evolving to reduce complexity and improve efficiency to ultimately reduce cost. The evolution has seen the closure of a number of R&D sites across Europe but an expansion into China in line with the strategic priorities.

1.2.5 Drug Design Process

The process of drug discovery is time consuming, expensive and highly risky[8]. It begins with the identification of an unmet medical need or a judgement on the adequacy of existing therapies. At this stage, a decision will be made whether to proceed with a discovery project based on a multitude of information including, the potential for reimbursement from a new drug, how many patients would use the drug, the existing expertise within the organisation, how likely a tractable target can be found, the cost of failure and ultimately the chance of success. If the decision is made to pursue the therapeutic indication, then a research team is formed and the objectives of the project are set.

The next stage is to identify a suitable target implicated in the mechanism of a particular illness. These include ion channels, kinases, nuclear receptors and other enzymes and proteins that are crucial to the survival of the cell or able to restore the functional capabilities of malfunctioning cells. If possible a structure of the target is generated using X-ray crystallography and/or NMR. Assay development is also undertaken to develop tests that will be able to detect biological activity *in vitro*.

High-throughput screening (HTS) is applied to a compound library to identify biologically active compounds (i.e. a 'hit'). When a number of hits have been identified, structure-activity relationships (SAR) are investigated and biological (e.g. P450 inhibition) and pharmacokinetic properties (e.g. solubility) are measured to identify promising chemotypes for the start of chemistry. Only hits that have attractive properties are declared lead compounds. Lead optimisation is then focused on generating analogues of the lead compounds to optimise the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties while simultaneously improving the potency of the compounds. This is a multidimensional optimisation procedure where properties such as solubility, ionisation and lipophilicity can impact on the ADMET profile. Improving one aspect of the ADMET profile can have a detrimental affect on another. Generally, one series of compounds will be optimised and, where possible, exposed to different assays to enhance the understanding of the ADMET. Representative compounds from the series will be screened in low throughput assays (e.g. the Ames test) as early as possible to provide reassurance that issues relating to toxicity will not curtail the drug discovery programme in the future. If undesirable properties in the chemical series are identified and cannot be rectified, lead optimisation starts on alternative compounds. Some lead optimisation may have already been undertaken on these compounds however, switching to a new lead series is costly, both in terms of time and expenditure, and therefore undesirable. If lead optimisation identifies a compound that has desirable target potency and ADMET profile, further safety assessment and *in vivo* studies are conducted. If the *in vitro* assays have successfully predicted the efficacy and safety *in vivo,* then the candidate will progress to clinical trials in humans. If the candidate makes it through all three phases of clinical trials, regulatory approval for market access is sought.

Despite the careful considerations taken before project initiation, less than 1 in 50 projects get a drug to market[9]. In the discovery phase only 1 in 5 projects gets as far as selecting a compound for clinical trials. These failures are attributed to biological problems which include poorly validated targets and chemistry problems such as chemical instability or toxicity. The advent of the target-based approach to drug discovery was expected to drive the discovery of new medicines. HTS has considerably expanded the number of compounds that can be evaluated for their biological activity[10] however, in the last two decades the number of new drugs approved has not risen[9].

Recently, there has been increased emphasis on fragment-based drug design. It is based on screening a smaller number of molecules, typically several thousand fragments with a molecular weight between 100 and 250, in the hope of finding low-affinity ligands with activity in the high micromolar to millimolar range. The fragments probe key binding interactions in the protein, but are small enough to minimise the chances of unfavourable interactions that can prevent larger molecules from binding efficiently. In comparison, conventional screening campaigns evaluate a million or more compounds in the hope of finding relatively potent drug leads. The active fragments are then grown into a lead compounds with a binding affinity that is the sum of the individual parts[11]. GSK have entered into a collaboration with Astex Therapeutics to apply fragment-based methods to multiple targets identified by GSK[12] with the aim of increasing discovery productivity.

Computers play a pivotal role throughout the drug discovery process[13]. Beyond structural drawing, structure conversion, molecular visualisation, data handling and other techniques, several computation techniques have improved the efficiency of the drug design process. At the lead generation stage, HTS requires a compound library to be screened against for activity. One approach is to screen all the compounds in the organisation's corporate library. This was performed with two million GSK compounds to identify compounds that could be developed to inhibit the malaria parasite Plasmodium. The screen took five scientists around a year to complete but identified 13,500 structures for potential further investigation[14].

Molecular modellers use methods to reduce the size of screening libraries making them more focused. If the structure of the biological target is known or can be predicted by homology modelling, then virtual screening using docking can be applied to select the most appropriate compounds for screening. *De novo* design may also be used to build inhibitors from scratch given the target binding site. If the structure is not known but a number of active compounds have been identified, then pharmacophore models can be constructed for the positioning of key features like hydrogen-bonding and hydrophobic groups. Such models can be used as a template to select the most promising compounds. Similarity searching against active compounds is also useful when the target structure is unknown.

Rules of thumb can be used to create libraries based on drug-likeness. The rule-of-five[15] highlights that most orally administered drugs have a molecular weight of 500 or less, a LogP no higher than 5, five or fewer hydrogen-bond donor sites, and 10 or fewer hydrogen-bond acceptor sites. Extending the rule-of-five, Veber and co-workers[16] suggest that a maximum of seven rotatable bonds is optimal for bioavailability and Clark and Picket[17] indicate polar surface area as another key property. For fragment libraries design, a molecular weight less than 300, 3 or less hydrogen-bond donors and acceptors, and LogP less than three is advised for efficient lead discovery[18]. Gleeson[19] also highlighted that almost without exception absorption, distribution, metabolism, excretion and toxicity (ADMET) liabilities increase with increasing molecular weight and/or LogP. Ionization states also play either a beneficial or detrimental role depending on the property in question. Filtering a library based on rules can significantly reduce the number of compounds to screen and increase the chances that an identified hit will progress to market.

In the lead optimisation phases, predictions need be made to investigate how changing molecular or structural features will affect physical properties. They are usually predicted from Quantitative Structure Activity Relationships (QSARs) that have been trained on experimental data. Bioisosteric replacements[20, 21] may be suitable to apply to lead compounds. Bioisosteres are substituents or groups with similar physical or chemical properties which produce broadly similar biological properties to a chemical compound. The idea is to preserve the desired activity without making significant changes in chemical structure. Bioisosteric replacements may be used to improve properties such as solubility and hydrophobicity. They may also be used to reduce undesirable features such as compound toxicity.

The number of possible compounds in the small-molecule universe is estimated at 10^{40} - 10^{100}. Typically a drug discovery programme tests 10^5 - 10^7 compounds[22]. Despite this small coverage of chemical space, the fact that lead compounds are identified and optimised to drug molecules is impressive. The process has undoubtedly improved with the use of computers at all stages.

1.3 Project Brief

This project will extend the use of QCT descriptors for property predictions using a method known as Quantum Topological Molecular Similarity[23] (QTMS). Two distinct properties important in the ADMET profiles of drugs will be used to create models to predict the properties of new compounds. Although pK_a and mutagenicity are both important in ADMET and can be linked here we treat them separately as they were used to satisfy different objectives. QCT descriptors will be used to predict pK_a for large data sets of drug-like compounds. The models will be extensively validated and the results compared to software frequently being used today. QCT descriptors will also be applied to the prediction of toxicity, specifically mutagenicity, in an attempt to create models that enable chemists at GSK to predict the results for new compounds, giving an early indication of the likely outcome in an assay. This will enable them to make informed choices about progressing with lead compounds and also rank molecules for priority testing in low throughput assays.

1.3.1 Project Objectives

The specific objectives to be completed are as follows:

In relation to pK_a

1. Apply QTMS to a large data set of drug-like molecules.
2. Investigate the use of different machine learning methods.
3. Improve the validation of the results from QTMS beyond the uses of the cross-validation statistic q^2.
4. Compare the results to publically and commercially available pK_a prediction tools.
5. Investigate whether only *ab initio* bond lengths can be used to predict pK_a.
6. Test the models by predicting the pK_a of drug molecules.

In relation to mutagenicity

7. Apply QTMS to predict mutagenic potency of carbocyclic and heterocyclic primary aromatic amines.
8. Establish the experimental requirements for a reliable data set.
9. Verify the reliability of the computational classification scheme by getting some outliers retested experimentally.

1.3.2 Project Flow

A number of sub-projects have emerged that attempt to satisfy the objectives. These projects represent stages in the research when findings highlighted areas for further research.

1) pK_a Prediction from "Quantum Chemical Topology" Descriptors
2) pK_a Prediction from a Single *Ab Initio* Bond Length
3) Prediction of the Mutagenic Potency of Primary Aromatic Amines Using "Quantum Chemical Topology" Descriptors.
4) Experimental and Computational Investigations into the Mutagenicity of Carbocyclic and Heterocyclic Primary Aromatic Amines.
5) Predicting Mutagenic and Non-Mutagenic Carbocyclic and Heterocyclic Primary Aromatic Amines.

Chapter 2

Background Theory

2.1 Quantum Mechanics

Electronic structure methods use the laws of quantum mechanics rather than classical physics as the basis for their computations. Quantum mechanics states that the electronic structure and properties of any given molecule, in any of its available states, may be determined in principle by solutions of Schrӧdinger's equation. Equation 2.1 describes the wave function of a particle:

$$\left\{\frac{-h^2}{8\pi^2}\nabla^2+\mathbf{V}\right\}\Psi(\vec{r},t)=\frac{ih}{2\pi}\frac{\partial\Psi(\vec{r},t)}{\partial t} \qquad \text{Equation 2.1}$$

where Ψ is the wave function, m is the mass of the particle, h is Plank's constant, and $\boldsymbol{V}$ is the potential field in which the particle is moving. ∇^2, is a differential operator, where ∇ is equivalent to partial differentiation with respect to the particle's coordinates. The Schrödinger equation for a collection of particles like a molecule is very similar. In this case, Ψ would be a function of the coordinates of all particles in the system. As stated, the energy and many other properties of the particle can be obtained by solving Ψ. Many different wave functions are solutions to the equation, corresponding to different states of the system.

For most molecular *ab initio* calculations the time-independent Schrödinger's equation, which takes the simplified form:

$$H(\vec{r},\vec{R})\Psi(\vec{r},\vec{R})=E\Psi(\vec{r},\vec{R}) \qquad \text{Equation 2.2}$$

Here Ψ represents the wave function of the position of the electrons and nuclei within a molecule, which are denoted as $\vec{r}$ and $\vec{R}$ respectively. E is the allowed energies of the system and H is the Hamiltonian operator, which is made up of kinetic and potential energy terms:

$$H=T+V \qquad \text{Equation 2.3}$$

The kinetic energy in three dimensions is a summation of the Laplacian, denoted by ∇^2, over all the particles in the molecule:

$$T = -\frac{h^2}{8\pi^2}\sum_k \frac{1}{m_k}\left(\frac{\partial^2}{\partial x_k^2}+\frac{\partial^2}{\partial y_k^2}+\frac{\partial^2}{\partial z_k^2}\right) \qquad \text{Equation 2.4}$$

The potential energy component is the Coulomb interaction between each pair of charged entities where each atomic nucleus is treated as a single charged mass:

$$V = \frac{1}{4\pi\varepsilon_o}\sum_j \sum_{k<j} \frac{e_j e_k}{\Delta r_{jk}} \qquad \text{Equation 2.5}$$

where Δr_{jk} is the distance between the two particles, and e_j and e_j are the charges on the particles *j* and *k*. For an electron, the charge is -*e*, while for a nucleus, the charge is Z*e*; where Z is the atomic number for that atom. Thus, the equation that represents the electron-nuclear attraction, the electron-electron repulsion, and the nuclear-nuclear repulsion is given by:

$$V = \frac{1}{4\pi\varepsilon_o}\left(-\sum_i^{elecrons}\sum_I^{nuclei}\left(\frac{Z_I e^2}{\Delta r_{iI}}\right)+\sum_i^{electrons}\sum_{j<i}\left(\frac{e^2}{\Delta r_{ij}}\right)+\sum_I^{nuclei}\sum_{J<I}\left(\frac{Z_I Z_J e^2}{\Delta R_{ij}}\right)\right) \qquad \text{Equation 2.6}$$

The Schrödinger equation is an eigenvalue equation with the solutions being a spectrum of eigenvalues (E) and corresponding eigenfunctions (Ψ). It is solved to find the wave function "Ψ" from which chemical properties for the system can be determined. Currently there is no clear chemical interpretation of Ψ, however, $|\Psi|^2$ gives the probability of finding an electron at a given point. From Ψ, we can therefore generate a probability density *P*(**r**) by integrating Ψ over all spatial coordinates except the set of coordinates describing one electron, and by summing over all spin coordinates. This must be carried out as each electron is described by four coordinates (three spatial coordinates and a spin coordinate) and so the integration renders the wave function into three dimensions from originally residing in a high dimensional space. The *P*(**r**) is multiplied by the number of electrons *N* to give the electron density ρ(**r**), or just 'ρ'.

The equation solved is time-independent, therefore the solutions are for a frozen structure at zero Kelvin. The system is usually isolated, so best represents a molecule in the gas phase. The gas-phase, zero Kelvin approximation is used to reduce the computational cost of the *ab initio* calculation. It is important to consider the impact of such approximations when applying *ab intio* data to real-world systems, which are dynamic and interact with a wealth of other species in the local environment. Often solvent interactions are considered, either implicitly or explicitly, in an attempt to account for a proportion of these interactions.

2.1.1 Level of Theory

It is possible to model reactions, molecular structures and dynamic processes *in silico* using quantum chemistry. Due to computational limitations a number of approximations must be made before reaching practical molecular orbital methods that can provide approximate solutions to the Schrödinger equation. There are two main classes of electronic structure methods:

- Semiempirical methods, such as AM1, incorporate experimental parameters in an attempt to predict molecular properties. Such methods are computationally very fast, however, their success is determined by having appropriate experimental input for the system under investigation.
- *Ab initio* methods rely on no experimental values; but are based solely on fundamental constants. The ability of these methods to predict molecular properties solved exclusively by equations means that they are computationally very demanding.

The 'Born-Oppenheimer approximation' (BOA) simplifies the general molecular problem experienced by *ab initio* calculations by separating the nuclear and electronic motion. The motivation behind this is based on the fact that there is a large difference between the masses of electrons and nuclei, allowing electrons to respond almost instantaneously to the motion of the nuclei. This means there is a large difference in timescale of electronic and nuclear motion, therefore allowing the electronic motion to be treated as occurring in a field of fixed nuclei, and so the Schrödinger equation is solved as a parametric function of the nuclear coordinate.

The most basic *ab initio* method in common use is the Hartree-Fock 'Self Consistent Field' (SCF) method. As well as utilising the BOA a further approximation is used. The Schrödinger equation is initially solved for a single electron in a system, which experiences all remaining electrons via an average field of negative charge. An initial guess is used to create the first field, and when the wave function for the first electron has been solved it is used to calculate the field for the next electron. When the wave functions for all electrons have been calculated the process starts again and iterates in this manner until there is a negligible change in the wave function and the field has become consistent. This method of using an electron cloud to represent all other electrons is known as the Hartree-Fock Approximation.

The Linear Combination of Atomic Orbitals (LCAO) is also introduced to make calculations of molecules practical. The approximation is based on the idea that orbitals are not just centred on one nucleus, but on every nucleus in the molecule. Basis sets are introduced to mathematically represent the atomic orbitals within a molecule, which in turn combine to approximate the total electronic wave function.

The above approximations generate a scheme that provides solutions to the Schrödinger equation. These methods, however, neglect the Coulomb correlation energy experienced by electrons of opposite spin. Coulomb correlation arises from charge repulsion and its neglect causes electrons to move too close together, therefore the energy calculated will always remain above the exact energy. The difference between these two energies is known as the correlation energy. Fermi correlation between electrons of the same spin arises from the Pauli Exclusion Principle, and is included in the 'exchange' term in Hartree-Fock calculations.

'Density Functional Theory' (DFT) and Moller-Plesset Perturbation Theory also use the Born-Oppenheimer and LCAO approximations. Unlike HF calculations, they partially account for the Coulomb correlation. Moller-Plesset Perturbation Theory takes the estimates offered by HF calculations and adds a corrective contribution from the Coulomb correlation. DFT is based on the theorems of Hohenberg-Kohn[24] which state that the ground state energy of a non degenerate system, as well as its electronic properties, are solely defined by its electron density. As such, DFT does not use the wave function but an electron probability density function which refers to the probability of finding an electron in a volume centred on a point with coordinates x, y and z. DFT methods take a different approach to HF and Moller-Plesset methods by incorporating Coulomb correlation with an exchange-correlation energy term based directly on the electron density. These methods partition the energy into several terms:

$$E^{DFT} = E^{nuclear} + E^{core} + E^{Coulomb} + E^{exchange\text{-}correlation}$$

whereas HF theory contains:

$$E^{HF} = E^{nuclear} + E^{core} + E^{Coulomb} + E^{exchange}$$

The tem $E^{exchange\text{-}correlation}$ is the only one that is not determined directly because of its unknown mathematical formulation. Usually $E^{exchange\text{-}correlation}$ is described as a sum of the exchange term and the electronic correlation. The exchange term can be calculated by using approximations that applies a homogenous electron density, such as the Local Density Approximation[25] and the Local Spin Density Approximation[26], by using gradient corrected functional such as the Generalised Gradient Approximation (GGA) methods (e.g. Becke95[27] (B95) and Lee-Yang-Parr[28] (LYP)). $E^{\textit{exchange-correlation}}$ can also be calculated using hybrid density functionals, which combines a conventional GGA method with a percentage of Hartree-Fock exchange. Examples of hybrid density functional include B3LYP[29] and B3PW91[30].

2.1.2 Basis Sets

Basis sets are functions used to represent atomic orbitals, which in turn combine to form molecular orbitals. Using a linear combination of atomic orbitals a single molecular orbital (ϕ_i) can be constructed:

$$\phi_i = \sum_{\mu=1}^{N} c_{\mu i}\,\chi_\mu \qquad \text{Equation 2.7}$$

where N is the set of functions used, χ_μ refers to an arbitrary basis function, and each has associated with it some coefficient $c_{\mu i}$. The total molecular wave function is then calculated as the antisymmetric product of the single molecular orbitals.

Most *ab initio* programs use Gaussian-type atomic functions as basis functions. Gaussian functions have the general form:

$$g(\alpha, \vec{r}) = cx^n y^m z^l e^{-ar^2} \qquad \text{Equation 2.8}$$

where $\vec{r}$ is composed of x, y, and z, α is a constant determining the shallowness of the function, and *n, l,* and m are integers that determine the directional dependence. If $n + l + m = 0$ then there is no directional dependence and the Gaussian function represents an ‘s-type’ orbital. When $n + l + m = 1$ then the Gaussian function lies along one of the axes and represents a ‘p-type’ orbital. A single Gaussian function, however, does not give an accurate representation of an orbital, so a combination of several is required. The more atomic orbitals used to construct the basis set, the closer the energy approaches the exact value for a given molecule. It would seem sensible to use very large basis sets in calculation, however, due to the large computational time and cost, the size of the basis sets that can be used are limited.

Minimal basis sets use one (single) basis function for each type of atomic orbital occupied in the separate atoms. Larger basis sets more accurately approximate the atomic orbitals by imposing fewer restrictions on the locations of the electrons in space, hence it is often necessary to use larger basis sets than the minimal. The double and triple zeta basis sets use two and three basis functions for each type of atomic orbitals in atoms, respectively. In a molecule, the electron density around the atom will be different to that around the separate atom, using two or three functions per orbital allows for variation due to bonding and other interactions. Split valence basis sets are used to reduce the computational time and cost and are based on the assumption that the core electrons of an atom are less affected by the chemical environment than the valence electrons. In other words the core electrons retain their atomic characteristics and so the flexibility needed to account for bonding is not so crucial. This leads to split valence basis sets consisting of a minimal representation of the core electrons combined with double or triple zeta representations of the valence electrons.

Adding polarization and diffuse functions is a further modification that improves accuracy. Polarization functions add orbitals of higher angular momentum to atoms in molecules that are not normally occupied in the separate atoms. Using these functions improves the flexibility of the basis sets and better represents the electron density in bonding regions between atoms by allowing the electron density to be polarized. Diffuse basis functions are extra functions that are added to basis sets to represent very broad electron distributions. They allow a better representation of the electron density when it is spread over a large region in mid to large sized molecules.

2.2 Quantum Chemical Topology

The theory of Quantum Chemical Topology (QCT), sometimes referred to as Atoms in Molecules (AIM), is a partitioning method pioneered by Richard Bader and co-workers in the early 1970s. Essentially the theory forms a bridge between quantum mechanics and working chemical concepts such as the atom and the bond. The electron density, obtained from the wave function, is partitioned into atomic portions, with each having its own unique properties. This method has a solid basis as the electron density can also be observed experimentally, for example using X-ray crystallography as well as being derived computationally from ab intio calculation previously outlined. The applications of QCT are continually expanding and being applied to new areas of chemistry.

2.2.1 Partitioning the Electron Density

Figure 2.1 displays a relief map of the electron density (ρ) in the symmetry plane of the furan molecule. Figure 2.2 shows the same furan molecule but the electron density is represented by a 2-dimensional contour map. Inside each contour line lies a set of non-intersecting contour lines of higher electron density; these are nested contours. In Figure 2.1 and Figure 2.2 it can be observed that the electron density is high around the atom centres and quickly drops when the distance from the atom centres increases. Bonding interactions can be observed along the ridges of the high electron density that run between atom centres.

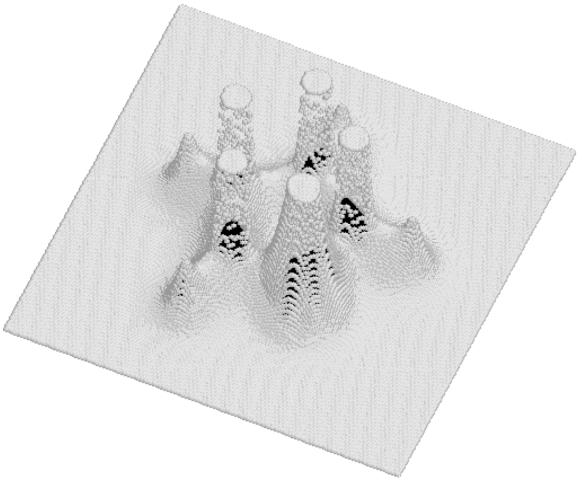


Figure 2.1. A relief map of the electron density of the furan ring in the symmetry plane. Ridges of electron density can be observed running between atoms.

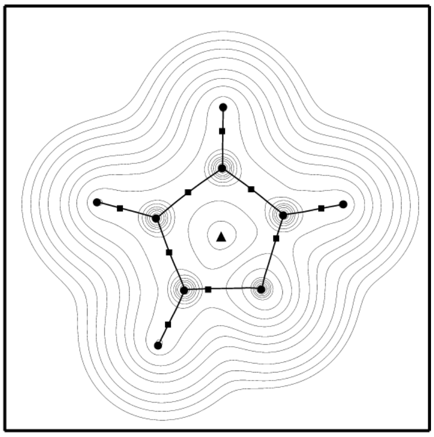


Figure 2.2. A contour plot of the electron density in the symmetry plane of furan.

The gradient vector of the electron density ($\nabla\rho$) points in the direction of the greatest increase in ρ. A succession of infinitesimally small gradient vectors forms the gradient path. A gradient path has three important properties; its always orthogonal to the contour surface, gradient paths never cross except when $\nabla\rho = 0$, and they have a beginning and an end. An infinite collection of gradient paths forms the gradient vector field (Figure 2.3). In reality, gradient vector fields of ρ are shown using a finite number of gradient vector paths. As gradient vector paths start at infinity, they are traced backwards, from a small circle (the nuclei) with a certain number of equally spaced points, following the gradient vectors of decreasing ρ.

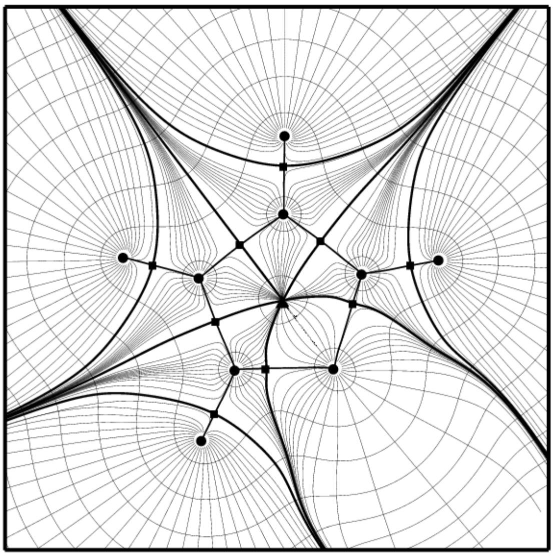


Figure 2.3. The gradient vector field superimposed on a contour map of ρ in the symmetry plane of furan.

In Figure 2.3 there are gradient vectors paths that do not terminate at a nucleus; the places where they do terminate are known as critical points. The special points between two bonded nuclei where gradient vector paths terminate are known as 'bond critical points' (BCPs). A collection of vector paths in 3-dimensional space that terminate at a BCP defines the Interatomic Surface (IAS). The IAS, sometimes called the *zero-flux surface*, distinguishes itself from other arbitrary surfaces in that at every vector $\nabla\rho$ has no component through the IAS. This is equivalent to saying that the gradient vector field must be parallel to the IAS at every point on its surface. The IAS defines the boundary between two atoms (Figure 2.4). Not all atoms in a molecule will be completely enclosed with IAS formed between neighbouring atoms and as gradient paths extend to infinity it is useful to cap atoms using an outer shell to give them a finite volume. The 'atomic basin' defines the region of an atom bound by IAS and an outer envelope (Figure 2.5).

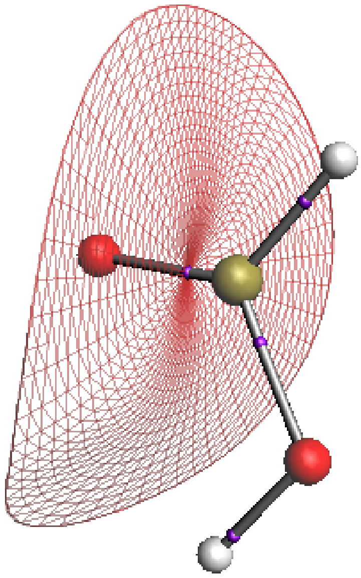


Figure 2.4. The IAS is formed from the gradient vector field lines terminating at the Bond Critical Point (purple sphere). The example shows the interatomic surface defining the boundary between oxygen and carbon atoms of formic acid.

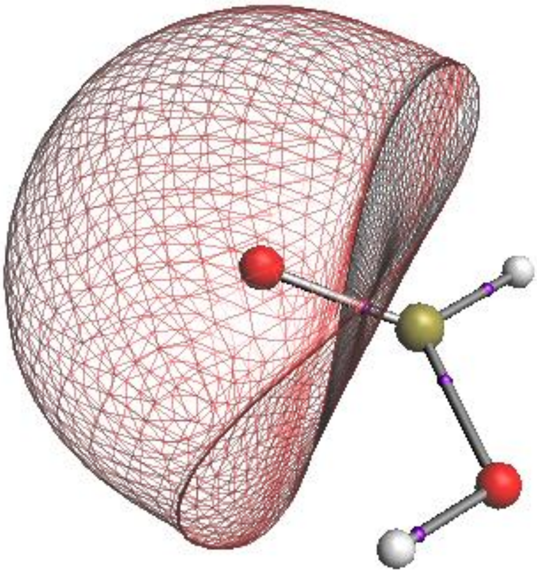


Figure 2.5. The 3-dimentional basin of the oxygen atom of formic acid. The basin is surrounded by the red interatomic surface and capped. Bond Critical Points are shown as purple spheres lying along the grey bond paths.

2.2.2 Critical Point

The idea of bond critical points has already been introduced. A BCP forms when the gradient paths terminate at a point in space between two bonded nuclei, this means the BCP is the centre of the IAS. The BCP is also the point where ρ reaches a maximum within the IAS, any displacement away from the BCP within the IAS reduces the value of ρ compared with the electron density at the BCP, denoted as ρ_b. It is important to note that ρ_b is not a maximum in all directions; if this were true then the BCP would be identical to a nucleus. In fact, ρ_b is a minimum in the orthogonal direction to the IAS meaning it is a saddle point. The orthogonal direction to the IAS at the BCP is referred to as the bond path (BP), and is a special type of gradient path connecting two nuclei. Figure 2.4 shows the BCPs between bonded nuclei. The gradient paths forming the IAS terminate at the BCP, not shown are the gradient paths that originate at the BCP and terminate at nuclei. These gradient paths form the atomic interaction line (AIL). The AIL is found between every pair of nuclei whose atomic basins share a common IAS. However, the presence of an AIL between two nuclei does not necessarily mean they are bonded, for example, AILs can be found between two noble gas atoms at any separation. An extra condition is required, namely that the molecule is in an energy minimum on its energy surface. It is also important to note that, except for symmetrical bonds, BCPs can occur anywhere along the bond. In addition to BCPs there are other types of CPs. These are the ring critical point, cage critical point and nuclear or non-nuclear attractor. Each CP can be categorised by three eigenvalues, λ_1, λ_2 and λ_3, calculated from the Hessian matrix at the CP. At CP, the three eigenvalues are always non-zero and the signs determine the types of CP being considered. A nuclear CP is characterised by three negative eigenvalues, meaning a maximum of electron density in all directions. A ring critical point is a saddle point, having two positive curvatures (i.e. two negative eigenvalues) and one negative curvature (i.e. positive eigenvalues). At the ring critical point the electron density is a minimum in the plane of the ring and a maximum perpendicular to it. A cage critical point has three positive eigenvalues meaning it is a minimum of the electron density in all directions. Cage critical points appear within structures bonded by two or more rings.

2.2.2.1 Bond Critical Point Properties

QCT provides a means of evaluating several properties of BCPs using the electron density alone. Using all the BCPs in a molecule, these properties can be utilized to provide a compact representation of a given molecule. The properties that have been evaluated and used in this work are listed below.

1. The electron density (ρ_b) derived from quantum mechanics, is the first. It has been used to derive bond orders[31] and also displays strong correlations with bond energy[32].

2. At the BCP, the Hessian of the electron density has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive one ($\lambda_3 > 0$). The eigenvector associated with λ_3 is tangential to the bond, and so λ_3 describes curvature along the bond. The eigenvectors corresponding to λ_1 and λ_2 are orthogonal to the bond, and so λ_1 and λ_2 describe curvature perpendicular to the bond.

3. The sum of the three eigenvalues is the Laplacian of the electron density, denoted by $\nabla^2\rho$, which gives a measure of the local charge concentration or depletion at the BCP (Equation 2.9). If the negative eigenvalues λ_1 and λ_2 dominate, then an accumulation of charge takes place in the plane perpendicular to the bond. This is common for shared interactions, such as covalent bonds. This results in a negative value for the Laplacian. If the positive eigenvalues $\lambda_3$dominates, then the electron density accumulates along the bond towards the nuclei. This is common for closed-shell interactions such as ionic, hydrogen and van der Waals bonds. This results in a positive value for the Laplacian.

$$\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3 \qquad \text{Equation 2.9}$$

4. The ellipticity of the electron density, denoted by ε and defined in Equation 2.10, provides a further useful property associated with BCPs. If ρ protrudes more in one of two directions perpendicular to a bond, then an oval pattern appears, such as pure double bonds. This ovality is measured by the ellipticity. Single bonds are characterized by λ_1 and λ_2 being nearly identical, and hence ε is near zero.

$$\varepsilon = \left(\frac{\lambda_1}{\lambda_2}\right) - 1 \qquad \text{Equation 2.10}$$

5. Two types of kinetic energy density, denoted by K(**r**) and G(**r**)[33] can also be obtained as further BCP properties. Interpreting K(**r**) in chemical terms is not straightforward, however a useful formula describing its link to the Laplacian and G(**r**) is given by Equation 2.11.

$$\mathrm{K(r)} = \mathrm{G(r)} - {}^{1}\!/_{4}\, \nabla^2 \rho \qquad \text{Equation 2.11}$$

6. The equilibrium bond length, denoted by R_e, is not strictly a BCP property. However, it can be considered as the sum of the distances between the BCP and one nucleus and the distance between the same BCP and the other nucleus, neglecting any deviation from a straight line the bond path may exhibit and hence, can artificially be turned into a BCP property.

2.2.2.2 Atomic Properties

In order to give the full overview of QCT, it is important to mention atomic properties although full details will not be provided. Unlike bond properties which are evaluated at a single point, volume integration over an atomic basin yields the atomic properties associated with that particular basin. Obtaining the atomic properties is much more computationally demanding than the calculation of BCP properties. The most commonly used atomic properties include the population, volume, different types of energy, and electrostatics. It is important to note that as QCT strictly partitions the atoms in a molecule, summations of a group of atoms can give fragment properties, while summations over all atoms give the molecular properties. This concept has led to much work being undertaken in the transferability of atoms and functional groups (fragments) with similar QCT properties. Although not strictly related to this work, substituent effects demonstrate that changing the chemical environment in one part of a molecule can induce significant changes in surrounding moieties. This principle forms the basis for this work, since changes can be predicted using QCT and will be encapsulated in the wave function and thus in the properties calculated at the BCP used to predict the property of interest.

2.3 Quantitative Structure-Activity Relationships

It has been known for millennia that different chemicals have different biological effects. However, it was not until the science of chemistry had become sufficiently developed to assign structures to compounds that it became possible to speculate about the causes of such biological properties. These developments allowed researchers to link the structure of molecules to certain activities and propose structure-activity relationships (SARs). There are however, problems associated with SARs, the main one being that the relationships are empirical and are semi-quantitative, in that the changes in structure are represented as 'all or nothing' effects and the relationship only applies to the set of compounds from which it was derived[34].

Modern quantitative structure-activity relationship (QSAR) methods owe their origins to the work of Hansch and Fujita in the 1960s[35]. They successfully used an approach based on applying linear free-energy relationships (LFERs)[36] to correlate suitable physiochemical parameters to biological activities. Their work was a bold extension to the work of Hammett[37] almost three decades before. Hammett discovered that pK_a of benzoic acids and phenylacetic acid in aqueous solution was solely dependent on the substituent and a proportionality constant fixed by the solvent and temperature. These relationships turned out to be universal and can be used to predict pK_a for different ring substituents with known σ constants[38]. Since this advance, QSAR have been successfully applied to optimisation problems in drug and agrochemical design. In 1995 Hansch *et al*. estimated that somewhere between 15 000 and 20 000 chemical QSARs had been published, while about 6000 biological equations have been published[39]. These numbers must have significantly grown over the last 15 years.

QSAR is the process by which the chemical structure of a series of molecules is quantitatively correlated with a well-defined process, such as biological activity (e.g. toxicity and biodegradability) or physiochemical properties (e.g. solubility and pK_a). The ultimate aim of all QSAR studies is to determine an equation of the following form:

$$Activity = f\,(x_1, x_2, ..., x_n) \qquad \text{Equation 2.12}$$

In Equation 2.12, *f* is a mathematical function, which is usually determined using an appropriate statistical technique and *x* represents the molecular descriptors that provide information about aspects of the molecular structure. This is where the fundamental difference between SARs and QSARs lies. The Q in QSAR refers to the way the molecular structure is quantitatively represented by descriptors. Therefore, the main challenges in QSAR are to find appropriate descriptors to represent the molecular structure and a function that relates the activity to those descriptors. The choice of descriptors is not a simple process, since there are thousands available. These are loosely characterized as theoretical, empirical or derived from readily available experimental characteristics of the structure[40]. A good QSAR will produce an excellent correlation between the activity with the molecules used in the training set but should also be able to predict the activity of molecules outside that data set. QSARs can however end up as a futile fitting exercise without true predictive power. Some QSAR models use hundreds of descriptors to predict the required activity. The high correlation between the number of brooding storks and newborn babies in West Germany between 1965 and 1980 highlights the potential pitfalls of using the wrong descriptors[41]. There is a higher probability of accidental correlation the larger the number of independent variables in the model. Furthermore, using a large number of descriptors may produce good predictive models but chemical understanding is difficult, and so it is argued that the use of fewer descriptors with known physical meanings provides easier interpretation of the models and mechanistic insight into the activity that is being predicted.

2.4 Quantum Topological Molecular Similarity (QTMS)

2.4.1 Background

Popelier[3] defined a novel similarity measure by utilising information available from QCT. BCP properties for the drug Haloperidol were mapped into abstract space. The distance d_{ij} between two BCPs i and j is defined as simply their Euclidean distance (Equation 2.13). Note that here only the values of ρ, $\nabla^2\rho$, and ε are used meaning the abstract space has three dimensions.

$$d_{ij} = \left[(\rho_i - \rho_j)^2 + (\nabla^2\rho_i - \nabla^2\rho_j)^2 + (\varepsilon_i - \varepsilon_j)^2\right]^{1/2} \qquad \text{Equation 2.13}$$

The distance between two BCPs gives an idea about how similar the different bonds are. Taking this idea further, it is suggested that BCP space can compactly and reliably describe the electronic structure of a molecule meaning molecules can be compared for similarity. The distance $d(A,B)$ between two molecules A and B is then defined as the sum of the BCP distance as calculated from Equation 2.14. The lower the value of $d(A,B)$, given by Equation 2.14, the more similar the two molecules are.

$$d(A,B) = \sum_{i\in A}\sum_{j\in B} d_{ij} \qquad \text{Equation 2.14}$$

In principle every BCP in molecule A could be compared with every BCP in molecule B under this definition. In the original work, a set of congeneric molecules was used so only corresponding BCPs in A and B were compared. This is a priori matching procedure, but is a natural mode of operation for molecules typical for QSAR. The practical use of this technique was demonstrated by reproducing the acidities of two sets of benzoic acids (as expressed via the Hammett equation) and at the same time recovering the portion of the molecule responsible for the acidity. Following these promising results the method was developed further by O'Brien and Popelier[42] and evolved into what is currently known as QTMS.

2.4.2 QTMS Analysis

Figure 2.6 summarises the main computational modules involved in a QTMS analysis together with the corresponding names of the computer programs that are most commonly used. The details of data generation, machine learning techniques and chemometric analysis applied to different data sets are specifically provided in the relevant Chapters. As with all QSAR studies, the first steps are to select the property of interest and a suitable data set. Young *et al*.[43] have highlighted the importance of chemical data curation in the context of QSAR modelling and others have demonstrated that the type of chemical descriptors has a much greater influence on the prediction performance of QSAR models than the nature of the optimisation techniques[44, 45]. Clearly, small structural errors or wrongly assigned experimental values within the data set can lead to significant loss of predictive ability of QSAR models. Recently, Fourches *et al*. provided a procedure to prepare a data set to be as accurate and consistent as possible[46]. The steps include removal of inorganics and mixtures, structural conversion and cleaning, normalization of specific chemotypes, removal of duplicates and finally manual checking. Many of the data sets used in this work were taken from the literature therefore many of the steps in the outlined procedure had already been performed, however manual checking was carried out and errors were detected. The errors and subsequent corrections are discussed in relevant chapters.

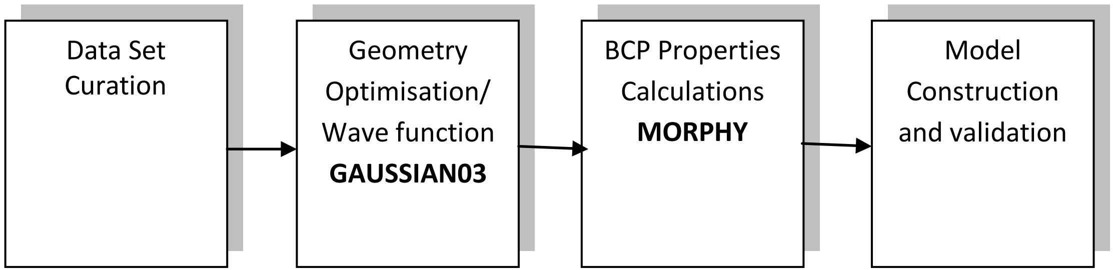

Data Set Curation
Geometry Optimisation/ Wave function GAUSSIAN03
BCP Properties Calculations MORPHY
Model Construction and validation

Figure 2.6. Chart representing the main modules involved in a QTMS analysis. The bond text represents the names of the programmes used in this work.

The second step in QTMS is the generation of molecular geometries and wave functions. An approximation of the geometry of each molecule is provided by MOLDEN[47]. Using the programme GAUSSIAN03,[48] geometries are optimised at a certain level of theory. The optimization steps are by far the most computationally expensive stage in the QTMS process. This expense is governed by the level of theory selected which can depend on time and computational resources available. The electronic wave function calculated by GAUSSIAN is then passed on to a local version of the programme MORPHY98[49], which locates the BCPs using a robust algorithm[50]. This yields a property vector for each BCP, providing a discrete "quantum fingerprint" for each molecule, when all BCPs appearing in a molecule are combined. A common skeleton associated with each molecule is defined to allow the location of descriptors important to the statistical analysis to be indentified in each molecule. In addition, the common skeleton allows the BCPs in one molecule to be mapped onto the BCPs in other molecules. This is not a fundamental requirement of the method, and the constraint of a common skeleton can be relaxed[51]. A two-dimensional data matrix is then constructed to display and allow easy manipulation of the data. Usually the first column contains unique identifications (e.g. chemical names) for each molecule in the data set. The second column contains the corresponding activities. The remaining columns contain the BCP properties associated with particular bonds selected from the common skeleton. For example, if all eight BCP properties are used to describe the molecules and there are four bonds in the common skeleton then each molecule is described by 32 descriptors (8 descriptors x4 bonds). BCP properties are inexpensive to obtain computationally and can be calculated for a data set of one hundred compounds on a standard PC in minutes. The next step is to construct a model using a suitable machine learning technique.

2.4.2.1 Partial Least Squares

Projection to latent structures by means of partial least square (PLS)[52] is a multivariate linear regression technique, which attempts to correlate independent variables with the dependent variable of interest. PLS utilises latent variables (LVs) to reduce the dimensionality of the data set, whilst retaining the majority of the variance in the independent variables that describe the dependent variable. LVs are constructed from linear combinations of the original independent variables into a set of new variables, subject to a condition that all newly constructed LVs are orthogonal to one another.

$$\begin{aligned} LV_1 &= a_{1,1}x_1 + a_{1,2}x_2 + a_{1,n}x_n \\ LV_2 &= a_{2,1}x_1 + a_{2,2}x_2 + a_{2,n}x_n \\ LV_q &= a_{q,1}x_1 + a_{q,2}x_2 + a_{q,n}x_n \end{aligned}$$ Equation 2.15

Each coefficient, $a_{i,j}$, represents the contribution of each variable to a particular LV. The sign of the coefficient indicates whether a particular variable makes a positive or negative contribution to the LV, and the magnitude of the coefficient shows how much the variable contributes to the LV. LVs are constructed in such a manner that the first extracted variable explains the maximum variance in the data set. The second variable then explains the maximum part of the remaining variance in the data set and so on. The maximum number of LVs can never exceed the smaller of the number of descriptors or compounds used to construct the model. The LVs are then combined in a linear fashion to correlate with the variable of interest creating a model.

$$y = a_1LV_1 + a_2LV_2 + \dots a_nLV_n$$ Equation 2.16

PLS may seem very similar to principal component regression (PCR), which involves principal component analysis (PCA) followed by multilinear regression (MLR). They differ in that the PLS algorithm is actually an iterative procedure[53]. Unlike, PCR, where PCA takes place followed by regression analysis, LVs are constructed so as to maximize their correlation with the dependent variables and LVs will only enter the PLS equation in the order, one, two, three etc. The main advantage of PLS over PCR is its ability to handle cases where the number of independent descriptors greatly exceeds the number of dependent variables. It is also capable of handling large numbers of noisy, incomplete, and collinear descriptors in a data set.

2.4.2.2 Parameters to Assess the Quality of the Model

In this work, PLS models have been constructed using the programme SIMCA-P[54] to fit the descriptors to the experimental values of interest. The predefined criterion for determining the significant number of LVs was used. If the value of q^2 (Equation 2.21) of the newly constructed LV is less than 0.097, then that LV is not considered significant, and no more LVs are computed; the PLS regression is then deemed complete. The construction of models using SIMCA-P provides three statistics (i.e. r^2, q^2 and RMSEE) to give an indication of goodness-of-fit and goodness-of prediction. We provide a general overview of these statistics here. The uses of statistical measure are explicitly stated in each Chapter, where applicable. The first statistic is the squared correlation coefficient (r^2),

$$r^2 = \frac{\sum_{i=1}^{n}\left(y_{calc,i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_{obs,i} - \bar{y}\right)^2} \qquad \text{Equation 2.17}$$

where n is the number of observations in the entire data set, $y_{calc,i}$ is the calculated value for molecule i from the regression equation, $y_{obs,i}$ is the corresponding experimental value, and $\bar{y}$ is the mean experimental value of the entire data set. r^2 can take the value of 0, where the PLS model is explaining none of the variance up to 1 where the regression explains all of the variance. The Root Mean Squared Error of Estimation (RMSEE) is calculated as,

$$\mathrm{RMSEE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{obs,i} - y_{calc,i}\right)^2}{n - 1 - a}} \qquad \text{Equation 2.18}$$

where n is the number of observations in the entire data set and a is the number of LVs used to construct the PLS model.

Another common error measure is the Root Mean Squared Error (RMSE), which is defined as,

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{obs,i}} - \mathrm{y}_{\mathrm{calc,i}}\right)^2}{\mathrm{n} - 1}} \qquad \text{Equation 2.19}$$

In an alternative expression sometimes encountered in the literature the denominator in the RMSE equation is sometimes set to n. Because the denominator of the RMSEE is always smaller than that of RMSE, RMSEE is always larger then RMSE. Hence, RMSEE penalises the deviation between observed and calculated data more than RMSE, and is therefore a more severe error measure. In turn, the RMSE is more severe than the Mean Absolute Error (MAE), which is given by,

$$\mathrm{MAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\mathrm{y}_{\mathrm{obs,i}} - \mathrm{y}_{\mathrm{calc,i}}\right|}{\mathrm{n}} \qquad \text{Equation 2.20}$$

In summary, the RMSEE is always larger than the RMSE, which is in turn always larger than the MAE, or RMSEE > RMSE > MAE. This is important to keep in mind when comparing results to the literature, where RMSE and MAE frequently appear. The MAE is also less sensitive to larger outliers than both the RMSE and RMSEE since all the individual differences are weighted equally in the average. For the RMSE and RMSEE, the errors are squared before they are averaged. This means that these two measures give a relatively higher weight to large errors compared to the MAE.

It should be noted, however, that high values of r^2 and low error values do not always indicate a predictive model, and can be quite misleading. The statistic r^2 is employed in conjunction with q^2 which provides an indication of the predictive ability of the model. Using *K*-fold Cross-Validation (CV), this cross-validated r^2 is calculated as

$$q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{obs,i} - \hat{y}_{pred,i}\right)^2}{\sum_{i=1}^{n}\left(y_{obs,i} - \bar{y}\right)^2} \qquad \text{Equation 2.21}$$

In *K*-fold CV the original set is partitioned in *K CV subsets.* The default setting for *K* in SIMCA-P is exactly seven, provided there are more than 7 data points in the total data set. Hence, if the total number of compounds n is divisible by 7 then there are n/7 compounds in each of the 7 CV subsets. If n is not divisible by 7 then the remaining compounds will be evenly distributed over the 7 CV subsets (note that the number of subsets does not vary). The compounds in the first CV subset are predicted from a model constructed from the remaining six CV subsets, all combined in one training set. The compounds in the second CV subset are then predicted from a different model, now constructed from the new remaining six CV subsets, excluding the second CV subset. Again these six subsets are all combined in one (new) training set. This process is repeated for the third and higher CV subsets, until each compound has been excluded exactly once. Each compound will then have been predicted by its corresponding training set. The predicted pK_a value for compound i, denoted by $\hat{y}_{pred,i}$, is obtained from the regression equation constructed from each training set. The automatically generated q^2 is based on 'leave-one-seventh' of the data out rather than 'leave-one-out', which is not recommended because of its known pitfalls[34, 55]. However, if the model is constructed from seven or less compounds then the CV is 'leave-one-out'. The default 'leave-one-seventh' CV can be altered at the discretion of the user in the newer version of SIMCA-P. One may question if the use of CV is justified against an assessment based on splitting data sets into a training and test set. Hawkins *et al*. recommend that, when the data set is small, then CV may be better than splitting the data set[56]. The majority of the data sets used to create models in this work are small in the sense of Hawkins *et al*. who advocates CV when the data set is less than 100 compounds.

A useful statistics which are not automatically provided when models are constructed in SIMCA-P is given below. The Root Mean Square Error of Prediction (RMSEP) is provided by

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{obs,i} - \hat{y}_{\ pred,i}\right)^2}{n}} \qquad \text{Equation 2.22}$$

and differs from the RMSEE, looking at the numerator, because $\hat{y}_{pred,i}$ is obtained from the models constructed from the training set during CV.

2.4.2.3 Further Considerations

As with any data analysis, PLS models work best when the data is relatively symmetrical in its distribution. If the observations lie in two distinct clusters in the y-variable, then an artificially good QSAR may be obtained that predicts well when the observations lie within one cluster but poorly if it lies in between. An observed versus predicted plot gives insight into the distribution of the data so the problem can be avoided.

PLS analysis provides a number of ways to identify data points with large residual errors. The easiest way to identify outliers is to examine the observed versus predicted plot. Observations that do not seem to fit the trend should be inspected. This can lead to the identification of errors in the calculations of descriptors, which can easily be rectified. If no errors are obvious then major structural or mechanistic differences relative to the rest of the data set may be responsible and the observation may be excluded. If there are a number of observations that have to be excluded for the same reason then the use of more local models may be more appropriate.

As with all regression techniques, there is a desire to reduce the number of variables used to construct models with the aim of improving the models predictions, reducing the number of calculations, and simplifying the interpretability and understanding of the models. SIMCA-P provides VIP (variable importance in the projection) values, which offer a concise summary of the importance of each descriptor. VIP plots can be examined and descriptor variables with a VIP value less than unity are considered unimportant to the model and can be removed[52]. Other variable reduction techniques can be employed and these are discussed in subsequent chapters.

2.4.3 Applications of QTMS

The applications of QTMS have been numerous. The exact implementations of QTMS have varied over the years but, almost without exception, they follow the procedure in Figure 2.6. After the incorporation of a firm statistical framework, the results of the study on para-benzoic acids presented in Section 2.4.1 were confirmed by more rigorous statistical treatment, which included CV and randomisation of the y-variable[23]. Popelier *et al.*[51] showed that the one-to-one mapping between molecules in less congeneric series of molecules can be achieved. Para- and meta-benzoic acids were combined to model Hammet σ constants. Furthermore, with a small modification para-substituted phenylacetic acids were added to the set.

In further extensions to the QTMS methodology, VIP plots were introduced to highlight the bonds that appeared to contribute the most to the activity. Consistently, the bonds with the highest VIP values were associated with the mode of activity for the data set under investigation. For example, QTMS highlighted the bonds of the carboxylic group in carboxylic acids, the O-H bond in phenols, and the C-N bond in anilines as the most important to the prediction of pK_a[57]. These results are consistent with the mechanism of dissociation for the acids. The 'active centre', highlighted by VIP plots, has also been shown to correctly move around in a data set of para-substituted phenols depending on the activity or property being predicted. Without any prior knowledge of the systems or mechanism of action, QTMS is able to reproduce the property of interest. QTMS was used to predict the mutagenic potency for a set of triazenes with two proposed mechanisms for their metabolism to genotoxic metabolites[58]. The study suggested a preferred mechanism; something which has yet to be confirmed experimentally. However, if this information is known, then the common skeleton can, in the first instance, be reduced to the bonds that are expected important reducing the number of descriptors. Ideally, the active centre is well localised, but there have been cases where it is turns out to be rather diffuse or "contaminated" with bonds one would not associate with the activity[59]. No systematic investigation into this contamination has been performed however, this issue is revisited in Chapter 4.

The QTMS method is closely related to another technique known as StruQT[60]. This method has been used to predict wavelengths of the lowest UV transitions for a set of anthocyanidins and to distinguish between reaction pathways for the electrophilic addition of hydrochloric acid to propene[61]. StruQT has been combined with inductive logic programming to include background knowledge and remove the need for molecular alignment[62]. This method was tested on a large set of mutagenic compounds but only produced slightly better results than the original StruQT analysis.

Over the years, QTMS methods have produced excellent results of relevance to biology[58, 63], medicinal[51, 59, 64, 65], environmental[66], industrial and physical organic chemistry[60, 61, 67, 68]. In all cases the models had excellent validation statistics and also provided information about the active centre or region of the compounds thought to be important to the activity. From the aforementioned QTMS publications it has become clear that QTMS descriptors are effective at capturing electronic effects. Therefore we deduce that when QTMS fails, electronic effects are not as important to the predicted property or activity as, for example, solubility or steric effects. This was the case for the predictive QSAR models of phenols[65] where LopP was introduced to describe the importance of hydrophobicity for hepatocyte toxicity prediction in conjunction with QTMS descriptors capturing the important electronic effects. A lipophilicity descriptor, $\log K_{o/w}$, also had to be introduced to predict the toxicity of aromatic aldehydes[69]. The model suggests that lipophilicity dominated but electronic factors are also important (this publication is in Appendix A). Another example is the QTMS study[70] on a remarkable and unusual set of ortho alkyl substituted phenols, known for their cytotoxicity and previously investigated by the Hansch group[71]. The QTMS results do not support their proposal that a steric factor is important in the determination of the cytotoxicity. In fact, QTMS results suggest no steric contribution whatsoever.

Chapter 3
pK_a Prediction from "Quantum Chemical Topology" Descriptors for Carboxylic Acids

3.1 Introduction

The publication relating to the work in this Chapter is provided in Appendix A.

In chemistry and biochemistry, the acid dissociation constant, the acidity constant, or the acid-ionization constant (K_a) is a specific type of equilibrium constant that indicates the extent of dissociation of hydronium ions from an acid, which is represented by

$$HA_{(aq)} + H_2O_{(l)} \overset{K_a}{\rightleftharpoons} H_3O^+{}_{(aq)} + A^-{}_{(aq),}$$

where

$$K_a = \frac{[H_3O^+][A^-]}{[HA]}.$$

As this constant differs for each acid or base, and varies over many degrees of magnitude, it is represented by the symbol pK_a, where

$$pK_a = -log_{10}K_a.$$

In general, a larger value of K_a (or smaller value of pK_a) indicates a stronger acid, since the extent of dissociation is large at the same concentration.

The pK_a of a compound is an important property in both life sciences and chemistry since the propensity of a compound to donate or accept a proton is fundamental to understanding chemical and biological processes. As the pK_a value of a molecule also determines the amount of protonated and deprotonated species at a specific pH, for example at physiological pH, knowing the pK_a of a molecule gives insight into pharmacokinetic properties. These properties can be strongly effected by pK_a, which therefore influences a compound's ADMET profile. A drug generally has to pass through at least one biomembrane via passive diffusion, or by carrier-mediated uptake, before it can produce any biological effect. Neutral molecules are easily absorbed by phospholipid membranes while these lipid bilayers in the cell walls have very low permeability for ions and highly polar molecules. The solubility of charged molecules is 3 to 4 log units higher than that of neutral counterparts, whereas the reverse is true for lipophilicity, which is 3-4 log units lower if the compound is charged[72]. Ionisable groups also affect the ability of molecules to interact with biological targets as they can influence binding orientation in protein active sites. Many biological systems also use proton-transfer reactions to communicate between the intra- and extracellular media and the rate of the proton-transfer reaction depends, in-part, on the pK_a values of the species involved[73]. It is estimated that ninety-five percent of medicinal compounds are ionisable, to some extent at physiological pH[74], while approximately sixty percent of drug molecules listed in the World Drug Index can be ionised between pH 2 and 12[75]. Beyond ADMET profiles, pK_a can be important in drug formulation and chemical synthesis. The benefit of *in silico* p$\mathrm{K_a}$ prediction is that physical samples are not needed. Therefore, predictions can influence decision-making in a drug development process before expensive and time-consuming synthetic work is undertaken.

There are a number of well established experimental techniques[76], such as spectroscopy, potentiometry, conductometry, competitive reactions and titrometry that can accurately determine pK_a values for a molecule. However, experimental determinations of the acidity of a specific part of a large biological molecule, such as a protein, is not a straightforward task[77] and is often associated with large uncertainties in the results. For small molecules, the accuracy of pK_a measurements can be affected by choice of experimental method, pH meter calibration, temperature control, solvent composition and chemical stability[78]. The benefits of a technique that accurately predicts the dissociation constant without the need for “wet” experiments are clear. The chemical industry, in particular the pharmaceutical and agrochemical sectors, screen thousands of compounds during the discovery process for many properties simultaneously, including the dissociation constant. More efficient techniques are required because of the logistics of measuring the pK_a values of these compounds. There are also problems associated with certain techniques. For example, high-throughput UV absorption measurements can often miss groups not in close proximity to a UV-chromophore[79].

pK_a estimation continues to receive much attention. A recent perspective by Lee and Crippen[80] highlighted the importance of the equilibrium constant and the multitude of methods available for predicting pK_a values for both proteins and small molecules. In the context of small molecules the methods generally fall into two main categories: (i) predictive models, using a range of descriptors and learning methods[79, 81-86], and (ii) *ab initio* quantum chemical methods based on different thermodynamic cycles[73, 87-92]. The first category was reviewed by Lee and Crippen and includes linear free energy relationships and quantitative structure activity/property relationships (QSAR/QSPR). This category of models relies on choosing the right descriptors to model the pK_a of a particular dataset. Structural, physiochemical, topological, geometrical, constitutional, electrostatic, quantum-chemical and thermodynamic descriptors have all been used to predict pK_a with varying success. Gruber and Buss[93] performed semiempirical calculations on some 190 phenols and carboxylic acids. They used multi-linear regression, with descriptors such as heats of formation, molecular orbital energies and charge densities, to produce a three-term equation for the benzoic acid derivatives (r^2=0.67) and a four-term equation for the aliphatic acids ($r^2$0.80). Citra[94] criticized this linear free energy relationship-based approach for lacking scientific bases and vast use of correction factors favouring quantum mechanical methods. A three-term equation for 57 benzoic acids ($r^2 = 0.80$), was reported using a method similar to that of Gruber and Buss. Gross and Seybold[95] rejected the use of semiempirical methods, instead using density functional theory, after a set of survey calculations demonstrated it performed significantly better for the descriptors they employed. After studying phenols they found that atomic charge (r^2=0.89) and, the difference between the HOMO and LUMO energies (r^2=0.95), correlated with pK_a. After rejecting the use of *ab initio* methods as too computationally demanding, Tehan[96, 97] and co-workers produced QSARs for numerous classes of acids and bases using semiempirical descriptors. Xing and Glen[83] fashioned a novel structure tree representation of atoms to align molecules. Twenty-four atom types and nine group types that were of biological interest were used in conjunction with PLS to produce a QSAR for a large set of acids and bases (r^2=0.93 and q^2=0.85). Following this procedure, Xing[84] and co-workers reduced the number of atom types and increased the group types used, also noting that splitting the dataset improved their results. The approach introduced by Xing has been taken up by numerous authors[82, 86, 98]. For example, Jelfs *et al.*[79] utilized the tree fingerprint method to develop a prediction method using semiempirical chemical properties, such as partial charge and electrophilic superdelocalizability of atoms undergoing protonation or deprotonation, to produce an online pK_a prediction web-tool at Novartis.

In the second category of methods, *ab initio* quantum chemical methods based on different thermodynamic cycles[89, 90] have started to receive more attention[91, 99]. The method involves calculating the standard change in Gibbs energy related to the dissociation of a proton from the compound under study in water. The method utilizes gas phase and aqueous phase *ab initio* calculations but depending on the thermodynamic cycle used, involves at least four separate geometry optimizations for each prediction. The choice of thermodynamic cycle, level of theory, and solvation model can all affect pK_a calculation[100]. Ho and Coote[100] suggest that a realistic error margin should be in the vicinity of 2 pK_a units, including a partial cancellation of errors. The calculation of pK_a based entirely on first principles can be criticised for being too computationally demanding as it requires thermodynamic analysis and high levels of theory[101]. Pulay *et al.*[102] produced a number of one-term equations to predict pK_a. These equations rely on entropic effects cancelling each other out and use only the enthalpy energy difference between the protonated and deprotonated forms, in conjunction with the COSMO continuum solvation model[103], to describe the solvation. After an initial investigation using 34 molecules to compare methods (B3LYP, OLYP, HF and PW91) and basis sets (3-21G(d) to 6-311++G(3df,3pd)) for geometry optimisation and single point energy calculations, they concluded that OLYP/3-21G(d) for geometry optimisation, and OLYP/6-311+G** for the energy calculation, were the best compromise between computational expense and accuracy[101]. Extending the work to a dataset of 370 different organic acids, including carboxylic acids, phosphonic acids, alcohols, thiols, and oximes, they produced linear regression equations for individual classes of compounds with mean absolute deviations of 0.4 pK_a units[102]. Out of all the commercial packages available for pK_a prediction, Schrödinger's Jaguar[104] application is the only tool that employs this method. The Jaguar package uses empirical correction terms, where calculated values are fitted to experimental values stored in a database, to repair deficiencies in both, the *ab initio* calculations and solvation models. Namazian[77, 105-108] and co-workers used an equation[73, 87] that relates the standard change of Gibbs energy to the pK_a of 66 carboxylic acids. They achieved an r^2 of 0.81 and a MAE of 0.48.

Here we evaluate QTMS descriptors in an extension to previous QTMS studies, which have shown good predictive ability for pK_a [57, 109]. Adam[110] obtained impressive results by incorporating QCT into his study. Using transferability between similar molecules, an idea at the origin of QCT[111] where any molecular property is the sum of the values of the property for the individual partitioned atoms , he obtained an r^2 for aliphatic and benzoic acids greater than 0.84, in most cases, using the energy of the dissociating proton in solution as the only descriptor. On the other hand, QTMS descriptors have been successfully employed for carboxylic acids and anilines[57], and phenols in aqueous[112] and polar solvents[109].

Over recent years there have been a number of publications comparing pK_a prediction methods. Dearden *et al.*[113] compared ten prediction software packages (ADME Boxes[114], VCCLAB[115], ADMET Predictor[116], Pipeline Pilot[117], SPARC[118], Marvin[119], QikProp[120], ACD/Labs[121], Pallas[122], ChemSilico pK_a [123]) using an undivulged test set of 653 molecules and found a package called ADME Boxes to be the most accurate judged by r^2 and the *mean absolute error* (MAE). As Lee and Crippen highlighted[80], the VCCLAB predictions were actually made by ADME Boxes, since VCCLAB links to ADME Boxes to make the predictions. The differing results for these two packages were attributed to the difference in SMILES handling. Pharma-algorithms, the company responsible for ADME Boxes, has merged with ACD/Labs keeping the ACD/Labs company name. Therefore, VCCLAB now uses ACD/Labs pK_a predictions. The r^2 and MAE range for the ten packages were 0.96 to 0.57, and 0.32 to 1.48, respectively. This comparison was based on a test set provided by ChemSilico. ChemSilico had verified that none of the compounds were part of their training set, which was not the case for the other packages. This may be one of the reasons for ChemSilico performing the worst. Meloun *et al.*[124] used the REGDIA regression diagnostics algorithm, in the package S-Plus, to compare the pK_a predictions of 64 drug molecules from four packages: ACD/Labs, Marvin, Pallas and SPARC. They found that ACD/Labs achieved the best predictive power and the most accurate results. Balogn *et al.*[125] used 248 drugs, agrochemicals and intermediates to compare ACD/Labs, Epik[126], Marvin, Pallas and VCCLAB. It is clear from their paper that at the time the predictions were made, VCCLAB was still using ADME Boxes predictions. VCCLAB was found to be the most predictive. However, it was suggested that ACD/Labs and Marvin are the most suitable methods for medicinal chemistry as VCCLAB only calculates pK_a for the most acidic and basic groups. The r^2 and MAE ranged from 0.95 to 0.49, and 0.30 to 1.79, respectively. Liao and Nicklaus[127] have compared nine programs to predict pK_a, both commercially available and free. They used 197 pharmaceutical substances with 261 pK_a values and found ADME Boxes, ACD/Labs and SPARC to rank the highest based on r^2 and MAE. The r^2 and MAE for all nine programs ranged from 0.94 to 0.58, and from 0.39 to 1.28, respectively. It is interesting to note that when pK_a was predicted for sites for which the experimental pK_a was determined to be between medicinally more relevant interval of 5.4 to 9.4 log units, the r^2 ranged from 0.68 to 0.35, and the MAE from 0.45 to 1.04. The relatively poor performance of Jaguar[128] confirms the discussion above on the second category of methods. The Jaguar method uses quantum mechanics to calculate the free energy change in going from the protonated to the deprotonated state. Empirical correction terms are then used to repair deficiencies in both the *ab initio* calculations and the solvent models, which brings the MAE below 2 pK_a units. This error is suggested as satisfactory for this type of methods.

In the past, we tended to utilize the interpretative ability of PLS at the chemometric stage to identify the active centre. While this is a compelling feature of QTMS, we concentrate here on highly predictive models for pK_a estimation. For this reason, we have used other statistical methods, such as support vector machines (SVM) and radial basis function neural networks (RBFNN), which yield models that are not so interpretable but possibly more accurate. We have also extensively cross-validated our models and moved away from relying on SIMCA-P for validation. The 228 carboxylic acid compounds included in this study is the largest set of compounds investigated with QTMS. This large set of diverse carboxylic acids facilitates the aim of extending the domain of applicability and producing more predictive models.

3.2 Methods and Computational Details

3.2.1 Data Set

We seek to predict the pK_a of molecules or fragments with pharmaceutical relevance. We therefore used the dataset of Tehan *et al.*[97], who had previously applied a variety of filters in order to remove non-drug like molecules from their dataset. We selected carboxylic acids as compounds of interest because we want to apply the QTMS methodology to a large set of diverse compounds and extending previous applications of QTMS. After iodine-containing molecules had been removed, since the basis sets were not readily available, our dataset contained 228 carboxylic acids with a pK_a range of 0.51-6.20. This included 44 meta- and para-substituted benzoic acids, 50 ortho-substituted benzoic acids and 134 aliphatic carboxylic acids. The observed pK_a values for all 228 carboxylic acids are listed in Appendix B.

3.2.2 Data Generation

The data generation process for QTMS can be found in Chapter 2 and a previous publication[23]. In short, an approximation of the geometry of each molecule is provided by MOLDEN[47]. Using the programme GAUSSIAN03[48] geometries were optimized successively at five different levels of theory: AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31+G(d,p) and B3LYP/6-311+G(2d,p). The levels are denoted by letters *A, B, C, D* and *E,* respectively, for consistency with previous publications. The wave function calculated by GAUSSIAN is then passed on to a local version of the programme MORPHY98[49], which locates the BCPs. In this study the electron density (ρ), the three eigenvalues of the Hessian of the electron density (λ_1, λ_2 and λ_3), the two types of kinetic energy (K(**r**) and G(**r**)) and the equilibrium bond lengths (R_e) have been used to describe each BCP. Since $\nabla^2\rho$ and ε are calculated using λ_1, λ_2 and λ_3, we have chosen to exclude the former. The numbering scheme given to atoms in the common skeleton of each molecule is shown in Figure 3.1. This allows the location of descriptors important to the statistical analysis to be identified in each molecule. Secondly, the scheme allows BCPs in one molecule to be mapped onto corresponding BCPs in other molecules. We end up with five data matrices (one for each level of theory) consisting of 228 observations (i.e. measured pK_a values) and 21 descriptors for each observation, that is, 7 descriptors obtained for each of the three bonds in the common skeleton.

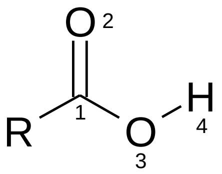


Figure 3.1. Numbering scheme of the common skeleton of the carboxylic acids.

3.2.3 Machine Learning and Chemometric Analysis

3.2.3.1 Partial Least Squares

Partial least squares (PLS)[52] analysis was carried out to fit the BCP descriptor variables to the experimental pK_a values. The programme SIMCA-P[54] was used along with its predefined criterion for determining the significant number of Latent Variables (LVs) to appear in the PLS equation.

First, the initial model is constructed, involving all descriptors at each level of theory. Then the VIP plots are examined because they offer a concise summary of the importance of each of the descriptors. Descriptor variables with a VIP value less than unity are considered unimportant to the model and hence discarded[52, 129]. The models are then reconstructed with the reduced set of variables. We also built models using just the R_e of the three bonds in the common skeleton to demonstrate that using BCP properties as descriptors provides more information than R_e alone. As well as producing global models for the carboxylic acids, we repeated the PLS analysis after splitting the dataset into aliphatic and benzoic acids, which were further split into meta/para-substituted and ortho-substituted sets. Altogether there are three subsets.

3.2.3.2 Support Vector Machines

Support vector machines (SVM), originally proposed by Vapnik[130] to solve pattern recognition problems[131], were extended in 1996 for linear and nonlinear support vector regression (SVR)[132]. SVM have found numerous applications in chemistry including drug design, QSAR/QSPR, chemometrics, sensors, chemical engineering and text mining[133].

As with other multivariate statistical methods, the performance of SVM for regression depends on the combination of several parameters. We employed a Gaussian Radial Basis Function kernel for SVR because of its effectiveness and speed in the training process[134]. This function contains an extra parameter γ (a constant) that controls the amplitude of the Gaussian function, thereby controlling the generalization ability of the SVM to some extent. The user-prescribed parameters (i.e. γ, ε of the ε-insensitive loss function and capacity parameter *C*) were chosen based on the lowest root mean squared error (RMSE) of the training data. The SVR programmes were written by former group member C.X. Xue in an R-file, based on a script written in the R language for SVM, which utilized the e1071 package[135]. The scripts were compiled using the R 2.5.1 compiler[136] and run on a Pentium D PC with 1GB RAM.

3.2.3.3 Radial Basis Function Neural Networks (RBFNN)

The subject of neural networks is covered in depth in the work by Haykin[137] and Gurney[138]. The theory of RBFNNs as applied to QSARs has been extensively described in the paper of Yao *et al.*[139]. The training procedure involved the forward subset selection routine, which selected the centres for the RBF one at a time and adjusted the weights between the hidden layer and the output layer after the addition of each centre, using a least-squares[140] solution. One third of the training set was randomly selected and 'held back' as test data, and training was terminated when the error on the test data showed no further improvement. RBFNN training was carried out using a range of RBF widths between 0.2 and 5.0 and the width yielding the lowest error on the test set was selected.

3.2.3.4 Comparison of the Methods

To compare these three machine learning methods we used *k*-fold cross validation (CV), where the datasets were divided in 4, 7 (as implemented in SIMCA-P) and 10 CV groups. The division of the datasets was carried out using systematic sampling where the compounds were ordered according to their pK_a values and assigned to a group accordingly. For example, for 4-fold CV the first compound was grouped with the fifth, ninth, thirteenth, etc. In random sampling method, the compounds were ordered by random numbers and divided into groups of different sizes depending of the *k*-fold CV being used (e.g. for 10-fold CV for the 50 ortho-substituted benzoic acids the first five compounds were group one, the next five group 2 etc.). Each CV group was excluded in turn so that each compound was excluded from the training data exactly once, and the RMSEP of prediction and q^2 were calculated for the CV group according to Equation 2.22 and Equation 2.21, respectively. The method that produced the lowest RMSEP in conjunction with the highest q^2 was considered to be the most accurate.

3.3 Results and Discussion

3.3.1 Choice of Level of Theory

Table 3.1 shows a summary of the initial PLS analysis at the five different levels of theory for the dataset and subsets. At each level, three different models were generated: a bond length only model, a model with bond length involving all BCP descriptors (amounting to 21, which derives from 3 bonds and 7 descriptors per bond), and a model including only those descriptors from the bond length and BCP descriptor model having a VIP score greater than one. At level *A*, only a bond length model can be generated. This is because semiempirical (AM1) wave functions do not contain core densities, which corrupts the topology by affecting the position or even appearance of BCPs[3, 100]. The level *A* results are rather disappointing compared to the previous QTMS study of carboxylic acids[57], where an r^2 and q^2 of 0.920 and 0.891 were obtained, respectively, and compared to the results in the literature discussed in the introduction earlier.

Table 3.1. Summary of the initial PLS analysis to determine the level of theory to use for the comparison of learning methods[b].

		All acids			ortho subset			para/meta subset			aliphatic subset		
Level	Descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.554	0.537	1	0.733	0.717	1	0.768	0.756	2	0.795	0.767
B	bond lengths	2	0.506	0.484	1	0.741	0.717	1	0.664	0.573	2	0.728	0.693
	BCP properties	1	0.600	0.595	1	0.761	0.716	1	0.683	0.601	2	0.771	0.731
	BCP properties (VIP>1)	3	0.638	0.616	1	0.757	0.718	1	0.708	0.651	2	0.726	0.713
C	bond lengths	2	0.593	0.581	1	0.770	0.754	1	**0.783**	**0.758**	2	0.729	0.696
	BCP properties	2	0.660	0.637	1	0.770	0.747	1	0.779	0.750	3	0.805	0.744
	BCP properties (VIP>1)	2	0.661	0.648	1	0.783	0.763	1	0.779	0.752	4	**0.823**	**0.791**
D	bond lengths	2	0.448	0.431	1	0.745	0.728	1	0.720	0.672	2	0.704	0.674
	BCP properties	6	**0.783**	**0.742**	1	0.769	0.752	1	0.759	0.724	5	0.815	0.768
	BCP properties (VIP>1)	7	0.696	0.646	1	0.794	0.775	1	0.772	0.742	2	0.788	0.758
E	bond lengths	2	0.462	0.446	1	0.767	0.757	1	0.737	0.700	2	0.700	0.672
	BCP properties	6	0.766	0.731	1	0.767	0.754	1	0.749	0.710	5	0.813	0.763
	BCP properties (VIP>1)	7	0.728	0.693	1	**0.792**	**0.782**	1	0.750	0.712	4	0.808	0.778

[a] Number of latent variables. [b] The bold text highlights the best models for each set (i.e. all, ortho, para/meta and aliphatic) based on the highest q^2.

Outlier detection was undertaken using the subset models as they were more easily distinguishable from correct predictions than in the models containing all the carboxylic acids. In all the para- and meta-substituted models, compound 37 (3,4-diamino-benzoic acid) was always an outlier. This compound was one of three zwitterions in the subset including compound 15 (3-aminobenzoic acid) and 21 (4-aminobenzoic acid). These compounds were predicted reasonably well (observed pK_a values of 4.74 and 4.85 and predicted pK_a values of 4.14 and 4.82, respectively) according to the best para- and meta-substituted model marked in bold in Table 3.1. However, compound 37 was predicted consistently poorly (observed pK_a of 3.49 and a predicted pK_a of 4.62 according to the same model). The zwitterions were all modelled in their neutral form, which sufficed for the monoamino-benzoic acids but was not appropriate for diamino-benzoic acid, which was predicted to have a larger pK_a value due to the fact that the increased stability in its zwitterionic form is not encapsulated in the BCP descriptors. As in this work, Tehan[97] and co-workers struggled to model this effect and had few problems with the monoamino-benzoic acids. In line with Tehan's approach we omitted 37 as an outlier. In the ortho-substituted models, two compounds were identified as outliers and removed from the models, namely 72 (2,6-

dihydroxybenzoic acid) and 79 (2-hydroxy-3,5-dinitro-benzoic acid) whose pK_a values were predicted to be 2.60 and 1.86 compared to observed values of 1.05 and 0.70, respectively. The hydroxyl group at the ortho position(s) in these compounds could be held responsible for their over prediction because internal hydrogen bonding in the anionic form could increase the stability of the ion therefore decreasing their pK_a values. This effect is not encapsulated in the BCP descriptors and therefore absent from the models. The issue with this reasoning is that there are a further 11 compounds in the subset that are hydroxyl-substituted at the 2 position and they are predicted well.

Four further compounds were identified as outliers from the aliphatic carboxylic acid models and removed. They were 124 (4-[(4-chloro-2-methylphenyl)oxy]butanoic acid), 150 (cyanoacetic acid), 155 (9-hydroxy-9H-fluorene-9-carboxylic acid/flurenol) and 228 (4-(cyclopropylcarbonyl)-3,5-dioxocyclohexanecarboxylic acid)). Tehan[97] and co-workers brought into question the reliability of the observed pK_a value of compound 124, with which we concur. Compound 155 (Figure 3.2a) was excluded from their model on the basis that the proximity of the carboxyl group to the two aromatic rings and the presence of an α-hydroxy group make its pK_a difficult to predict. Alternatively, the observed pK_a is incorrect. We can confirm that it is the *observed* pK_a value that is the most likely cause for discrepancy. Our best model predicts the pK_a to be 2.87 while the experimental value given by Tehan (our source data set) is 1.09. A different source[141] gives the observed pK_a value of 2.96, which is close to our predicted value and the value predicted by ACD/Laboratories[142] as 3.04. Furthermore, there is a similar structure (104, hydroxy(diphenyl)acetic acid) in the dataset that has an observed pK_a of 3.05 (Figure 3.2b). A literature value for the observed pK_a values of these compounds (155 and 104) with their respective hydroxyl group removed was found to be 3.61[143] for 155 (Figure 3.2c) and 3.9[144] for 104 (Figure 3.2d), a difference of only 0.3 log units. The difference of 1.96 log units (=3.05-1.09) between compounds 155 and 104 generated by the addition of a hydroxyl group at the same position in each is unlikely when considering the difference is 0.3 log units between the analogous compounds, thus further supporting a wrong observed pK_a. No reason for compounds 150 and 228 being outliers can be offered but they were both excluded.

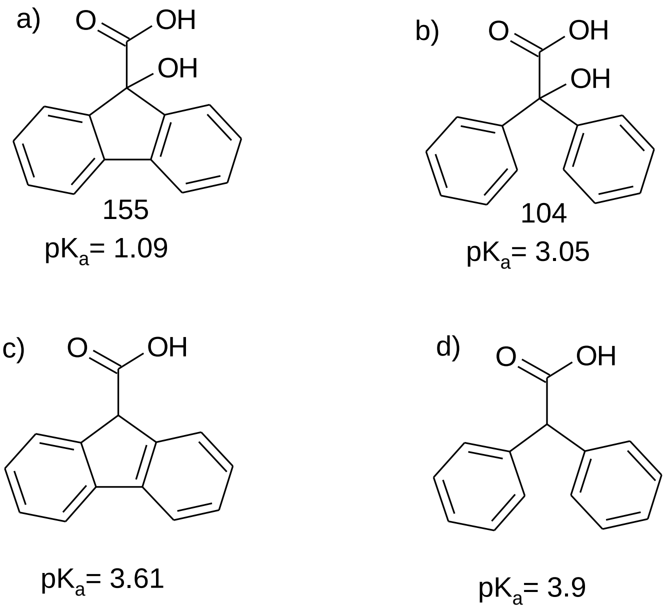


Figure 3.2. Structures and experimental p*K*$_a$ values for compounds 155 and 104 and their analogues fluorine-9-carboxylic acid and diphenyacetic acid.

Table 3.2 shows the results of the PLS analysis after the outliers were removed. Here, one of the BCP property models always outperforms the bond length model at each level of theory in terms of both r^2 and q^2. Generally, the models improved when the VIP<1 "cut-off" was used. After considering the r^2 and q^2 for all of the models, we found that level *C* gave the optimum results for the subsets. However, level *E* gave the best results for a model involving *all* carboxylic acids, although this model is constructed from 6 LVs.

Table 3.2. Summary of the initial PLS analysis after the outliers had been removed.

		All acids			ortho subset			para/meta subset			aliphatic subset		
Level	Descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.630	0.612	2	0.797	0.787	1	0.900	0.896	2	0.795	0.748
B	bond lengths	2	0.601	0.581	2	0.838	0.801	1	0.905	0.902	2	0.837	0.828
	BCP properties	4	0.769	0.730	1	0.837	0.810	1	0.912	0.911	2	0.863	0.852
	BCP properties (VIP>1)	4	0.764	0.749	1	0.835	0.815	1	0.909	0.908	6	0.895	0.861
C	bond lengths	2	0.696	0.688	1	0.841	0.832	1	0.931	0.929	1	0.875	0.875
	BCP properties	3	0.802	0.779	1	0.837	0.819	1	0.931	0.928	3	0.897	0.873
	BCP properties (VIP>1)	2	0.787	0.782	1	**0.848**	**0.836**	1	**0.933**	**0.932**	4	**0.911**	**0.901**
D	bond lengths	2	0.486	0.455	1	0.777	0.762	1	0.912	0.906	3	0.807	0.784
	BCP properties	6	0.860	0.839	2	0.851	0.813	1	0.923	0.919	3	0.882	0.869
	BCP properties (VIP>1)	4	0.738	0.689	1	0.842	0.833	1	0.927	0.925	2	0.890	0.881
E	bond lengths	2	0.535	0.498	1	0.803	0.794	2	0.917	0.907	2	0.810	0.788
	BCP properties	6	**0.879**	**0.844**	2	0.866	0.818	1	0.917	0.913	2	0.875	0.858
	BCP properties (VIP>1)	7	0.845	0.818	1	0.844	0.835	1	0.918	0.913	3	0.897	0.886

[a] Number of latent variables. [b] The bold text highlights the best models for each group based on having the highest q^2.

Figure 3.3 shows a comparison of the CPU time needed for optimization for nine of the compounds (three from each subset) that converged at each level of theory without any restarts versus the highest q^2 from each level of theory. This demonstrates that apart from the models involving all the carboxylic acids, there is no improvement in q^2 above level *C* despite a significant increase in computational expense. Level *A* and *B* are inferior in all cases. At this stage, we are in a position to explore different statistical learning methods to predict pK_a values based of BCP properties. To test the suitability of Level *C* in pK_a prediction and to examine whether the initial results still hold, we carried out the analysis with level *C* and the more computationally expensive level *E*.

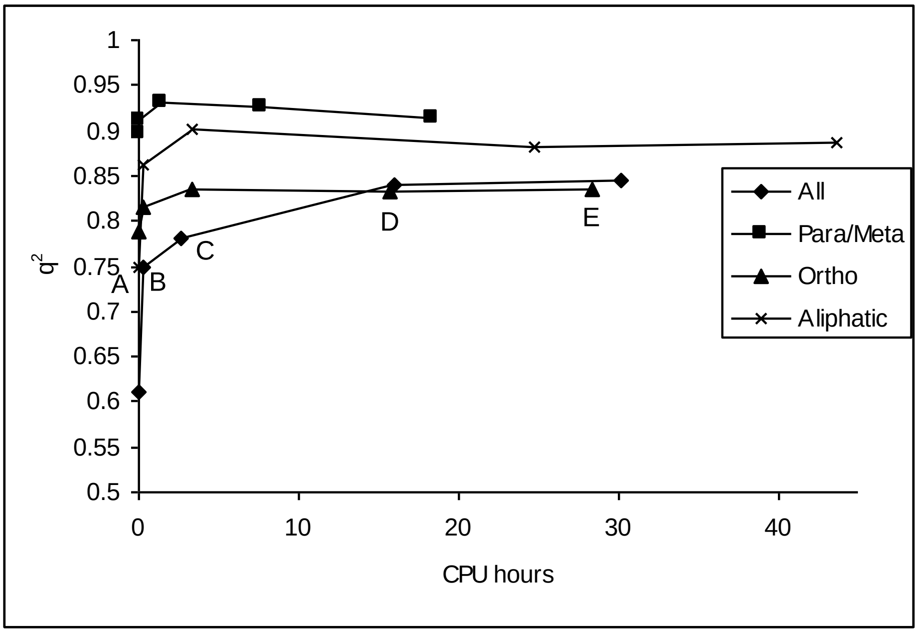


Figure 3.3. Comparison of the CPU times needed to optimize the compounds versus the highest q^2 value at each level of theory (A, B, C, D and E).

3.3.2 Comparison of the Statistical Learning Methods

The data for all comparisons in Section 3.3.2 to Section 3.3.5 can be found in Table 3.3. For each learning method, there are four groups of compounds to test: all carboxylic acids, aliphatic acids, meta/para- and ortho-substituted acids. There are three different sizes of CV sets for both the systematic and random sampling method. This gave 24 values (= 4 groups of compounds × 3 CV set sizes × 2 sampling methods) of q^2 and RMSEP to compare for each learning method employed. At level *C*, SVM gave the lowest RMSEP value 16 times out of 24 comparisons, PLS 7 out of 24 comparisons, and RBFNN 2 times out of 24 comparisons (the RMSEP for the random sampling para/meta 10-fold CV models were identical for PLS and SVM). Again out of 24 comparisons but at level *E* this time, SVM gave the lowest RMSEP 21 times, PLS 3 times and RBFNN 2 times (the RMSEP for the systematic sampling para/meta 4-fold CV models and the random sampling 7-fold CV models were identical for PLS and SVM). This clearly demonstrates that SVM is superior to both PLS and RBFNN, at both levels of theory.

Table 3.3. Summary of the results obtained from the three machine learning methods at level *C* and *E*. The values in bold are the lowest RMSEP and the values in italics are the next lowest RMSEP.

			PLS		SVM		RBFNN	
	k-fold validation	set	q^2	RMSEP	q^2	RMSEP	q^2	RMSEP
Level *C*								
systematic sampling	4-fold	all	0.782	0.422	**0.897**	**0.288**	*0.878*	*0.316*
		ortho	*0.812*	*0.408*	**0.797**	**0.398**	0.746	0.472
		para/Meta	*0.917*	*0.121*	**0.915**	**0.117**	0.847	0.147
		aliphatic	*0.893*	*0.288*	**0.899**	**0.28**	0.874	0.317
	7-fold	all	0.778	0.424	*0.895*	*0.291*	**0.927**	**0.256**
		ortho	**0.844**	**0.361**	*0.819*	*0.391*	0.774	0.437
		para/Meta	**0.934**	**0.104**	*0.915*	*0.126*	0.900	0.129
		aliphatic	**0.920**	**0.265**	0.904	0.27	*0.910*	*0.267*
	10-fold	all	0.779	0.420	**0.888**	**0.295**	*0.884*	*0.308*
		ortho	*0.845*	*0.355*	**0.848**	**0.352**	0.832	0.372
		para/Meta	*0.939*	*0.101*	**0.942**	**0.096**	0.942	0.107
		aliphatic	0.905	0.335	**0.907**	**0.262**	*0.889*	*0.298*
random sampling	4-fold	all	0.775	0.440	**0.891**	**0.295**	*0.860*	*0.338*
		ortho	*0.836*	*0.379*	**0.836**	**0.378**	0.807	0.407
		para/Meta	**0.927**	**0.120**	*0.910*	*0.129*	0.235	0.325
		aliphatic	**0.907**	**0.267**	*0.900*	*0.282*	0.883	0.305
	7-fold	all	0.778	0.426	*0.860*	*0.290*	**0.912**	**0.282**
		ortho	**0.801**	**0.382**	*0.807*	*0.395*	0.753	0.460
		para/Meta	*0.928*	*0.112*	**0.925**	**0.105**	0.789	0.177
		aliphatic	*0.903*	*0.273*	**0.915**	**0.251**	0.881	0.308
	10-fold	all	0.778	0.423	**0.885**	**0.300**	*0.855*	*0.345*
		ortho	*0.827*	*0.369*	**0.843**	**0.352**	0.761	0.412
		para/Meta	**0.929**	**0.097**	**0.930**	**0.097**	0.878	0.135
		aliphatic	*0.903*	*0.269*	**0.911**	**0.261**	0.891	0.295
Level *E*								
systematic sampling	4-fold	all	0.814	0.391	**0.873**	**0.269**	*0.849*	*0.352*
		ortho	*0.806*	*0.417*	**0.842**	**0.372**	0.690	0.520
		para/Meta	**0.916**	**0.131**	**0.916**	**0.131**	0.920	0.126
		aliphatic	0.875	0.313	**0.900**	**0.275**	0.902	0.280
	7-fold	all	0.821	0.375	*0.885*	*0.301*	**0.907**	**0.289**
		ortho	0.836	0.370	*0.851*	*0.354*	**0.852**	**0.350**
		para/Meta	*0.906*	*0.133*	**0.913**	**0.125**	0.869	0.159
		aliphatic	*0.894*	*0.285*	**0.909**	**0.263**	0.893	0.293
	10-fold	all	0.816	0.377	**0.881**	**0.303**	*0.859*	*0.339*
		ortho	*0.834*	*0.371*	**0.861**	**0.342**	0.749	0.423
		para/Meta	*0.917*	*0.122*	**0.919**	**0.119**	0.892	0.146
		aliphatic	*0.883*	*0.299*	**0.906**	**0.262**	0.762	0.435
random sampling	4-fold	all	0.787	0.407	**0.888**	**0.298**	*0.864*	*0.334*
		ortho	*0.825*	*0.390*	**0.846**	**0.366**	0.712	0.487
		para/Meta	*0.870*	*0.160*	**0.908**	**0.137**	0.760	0.220
		aliphatic	*0.877*	*0.307*	**0.901**	**0.280**	0.876	0.314
	7-fold	all	0.814	0.387	**0.886**	**0.289**	*0.897*	*0.304*
		ortho	**0.879**	**0.398**	*0.788*	*0.405*	0.641	0.554
		para/Meta	**0.911**	**0.129**	**0.913**	**0.129**	0.848	0.167
		aliphatic	0.884	0.299	**0.905**	**0.269**	*0.903*	*0.279*
	10-fold	all	0.814	0.383	**0.886**	**0.299**	*0.882*	*0.311*
		ortho	*0.879*	*0.380*	**0.832**	**0.368**	0.704	0.483
		para/Meta	*0.908*	*0.119*	**0.914**	**0.116**	0.853	0.139
		aliphatic	*0.883*	*0.295*	**0.904**	**0.269**	0.878	0.312

3.3.3 Comparison of the Level of Theory

Comparing the results of each q^2 and RMSEP values obtained at both levels of theory, level *C* produces the highest q^2 46 times out of 72 (=3×24) comparisons and the lowest RMSEP 45 times out of the 72 comparisons, with one value being the same for level *C* and *E*. This suggests that level *C* is superior to level *E* because it has the highest number of higher q^2 and RMSEP values.

3.3.4 Comparison of the Validation Set Selection Method

At each level of theory, there are 36 RMSEP and q^2 values to compare because there are 3 machine learning methods, 3 values for *r* in *k*-fold CV and 4 compound groups (36=3×3×4). At level *C*, there are 24 higher q^2 values for systematic sampling compared to 11 for random sampling and 1 q^2 value that is the same (36=24+11+1). There are 22 RMSEP values lower for systematic sampling compared to 13 lower RMSEP values for random sampling and 1 RMSEP value that is the same (36=22+13+1). At level *E*, there are 19 higher q^2 values for systematic sampling compared to 16 for random sampling and 1 q^2 value that is the same (36=19+16+1). There are 21 RMSEP values that are lower for systematic sampling compared to 15 values for random sampling (36=21+15). The better results gained from systematic sampling is not surprising because this method ensures the maximum value for the denominator in the q^2 equation. Using systematic sampling means that the most dissimilar compounds in relation to their pK_a values are excluded and so it is more likely that the training set will contain similar compounds to the CV set therefore leading to better predictions[145]. When random sampling is used one cannot ascertain that the CV sets contain all similar compounds in terms of their pK_a values. Hence, the models are possibly trained using compounds dissimilar to the CV set, which can lead to poor prediction statistics. If the CV sets contain similar compounds in terms of their pK_a values, then this can lead to a small denominator in the q^2, thus increasing the q^2 value.

3.3.5 Comparison of Validation Set Size

With regards to the *k*-fold validation, 10-fold validation generally provided the highest q^2 and the lowest RMSEP values, although this was not always the case. When the difference is calculated between the highest and lowest q^2 values for the *k*-fold CV set size for each subset and each method, the mean difference is 0.057 (standard deviation (SD) of 0.127) and 0.039 (SD=0.042) for level *C* and *E*, respectively. When the same is calculated for the RMSEP values, then the mean difference is 0.041 (SD=0.040) and 0.042 (SD=0.042) for level *C* and *E*, respectively. When the poor result for 4-fold-para/meta-random sampling using RBFNN at level *C* is omitted, then the mean difference for q^2 is 0.032 (SD=0.025) and the mean value for RMSEP is 0.035 (SD=0.024). The *k*-fold CV does give different results but the difference is small. It is not surprising that 10-fold CV generally gives the best validation statistics because the smaller the CV groups, the more compounds there are to train the models. What these results do suggest is that the models still provide good predictions even when 25% of the compounds are omitted in training when using 4-fold CV.

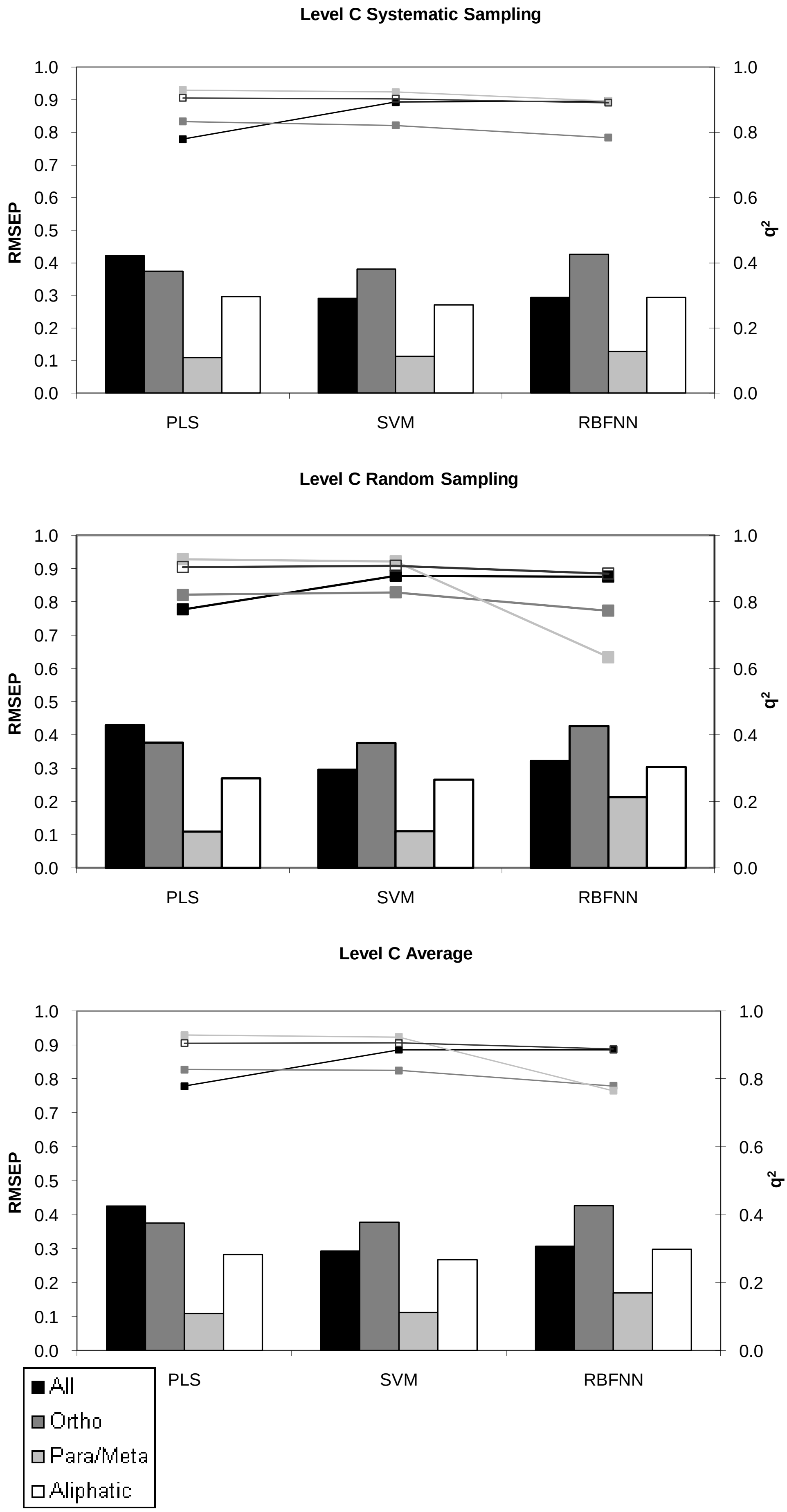
3.3.6 Confirmation of Finding Based on Averages

Table 3.4, Figure 3.4 and Figure 3.5 show the results of all the subsets and learning methods when the *k*-fold CV results have been averaged and the results of the subsets and learning methods when the sampling methods have been averaged from Table 3.3. These results confirm what had previously been suggested. There is little difference between the CV statistics for the random and systematic sampling methods. Excluding the RBFNN para/meta results at level *C*, the largest difference in q^2 and RMSEP values is 0.021 and 0.029, respectively, for the RBFNN all acid models. At level *E*, the largest difference in q^2 and RMSEP is 0.078 and 0.077, respectively, for the RBFNN ortho models. Based on the final average results of the *k*-fold and sampling methods, it is clear that SVM generally gives the best CV statistics. However, at level *C* the PLS statistics for the para/meta and ortho models provide the highest q^2 value and lowest RMSEP. However, these are close to the q^2 value and RMSEP value provided by SVM. Comparing the “all acid” models and the aliphatic models, the q^2 value and RMSEP for SVM are much better than the q^2 and RMSEP for PLS. At level *E*, SVM provides the highest q^2 and lowest RMSEP in all cases. Although PLS was better than SVM in two cases at level *C,* we chose SVM as the best learning method. This decision is based on the fact that, when PLS was superior, the difference between the statistics was small and, when SVM was better, then the difference between the statistics was large. Comparing the averages of systematic and random sampling for SVM level *E* only provides better CV statistics for the ortho dataset. The difference between the statistics is small but this may suggest that the more expensive level *E* accounts more for the steric effects than level *C* in some way.

Improved models were created when the dataset was split into aliphatic and aromatic subsets, which were further split into meta/para and ortho-substituted carboxylic acids. The subset models provided significant improvements on the “all acid” model. The para/meta model was the most accurate in prediction, followed by the aliphatic and then the ortho model. The excellent CV results of the para/meta models are not surprising because the pK_a difference caused by substituent changes can be accurately predicted by the Hammett equation and BCP properties display a strong correlation with Hammett’s sigma parameter[3]. The lower q^2 value and higher RMSEP for the ortho model can be explained in terms of steric effects[146]. Whereas the pK_a of meta- and para-substituted carboxylic acids is affected mainly by inductive and resonance contributions, the pK_a of ortho-substituted carboxylic acids is highly sensitive to steric contributions. Primary steric hindrance to deprotonation is important where there are bulky groups around the acidic centre. Secondary steric effects may either be acid-weakening (if there is steric hindrance to solvation), or acid-strengthening (if there is steric inhibition of resonance in the neutral molecule). We have already stated that when QTMS fails, one can be certain that steric effects are very important. Since the BCP properties do not account for steric effects, we can be confident that this is the reason for the poorer results. Since the results of the aliphatic subset are an improvement on the ortho subset, the steric effects must be less important for the former subset (Table 3.4).

Table 3.4. Average values of the results from Table 3.3.

		PLS		**SVM**		**RBFNN**	
	set	q^2	RMSEP	q^2	RMSEP	q^2	RMSEP
Level *C*							
systematic sampling	all	0.780	0.422	0.893	0.291	0.897	0.293
	ortho	0.834	0.375	0.821	0.380	0.784	0.427
	para/Meta	0.930	0.109	0.924	0.113	0.896	0.128
	aliphatic	0.906	0.296	0.903	0.271	0.891	0.294
random sampling	all	0.777	0.430	0.879	0.295	0.875	0.322
	ortho	0.821	0.377	0.829	0.375	0.774	0.426
	para/Meta	0.928	0.110	0.922	0.110	0.634	0.212
	aliphatic	0.904	0.270	0.909	0.265	0.885	0.303
average of systematic	all	0.778	0.426	0.886	0.293	0.886	0.307
and random sampling	ortho	0.828	0.376	0.825	0.378	0.779	0.427
	para/Meta	0.929	0.109	0.923	0.112	0.765	0.170
	aliphatic	0.905	0.283	0.906	0.268	0.888	0.298
Level *E*							
systematic sampling	all	0.817	0.381	0.880	0.291	0.872	0.327
	ortho	0.825	0.386	0.851	0.356	0.763	0.431
	para/Meta	0.913	0.129	0.916	0.125	0.893	0.144
	aliphatic	0.884	0.299	0.905	0.267	0.852	0.336
random sampling	all	0.805	0.392	0.887	0.295	0.881	0.316
	ortho	0.861	0.389	0.822	0.380	0.686	0.508
	para/Meta	0.896	0.136	0.912	0.127	0.820	0.176
	aliphatic	0.881	0.300	0.903	0.273	0.886	0.301
average of systematic	all	0.811	0.387	0.883	0.293	0.876	0.322
and random sampling	ortho	0.843	0.388	0.837	0.368	0.724	0.469
	para/Meta	0.905	0.132	0.914	0.126	0.857	0.160
	aliphatic	0.883	0.300	0.904	0.270	0.869	0.319



Level C Systematic Sampling

Level C Random Sampling

Level C Average

Figure 3.4. A graphical representation of Table 3.4 for level *C*. The bar charts represent the RMSEP, and the lines represent the q^2 values obtained.

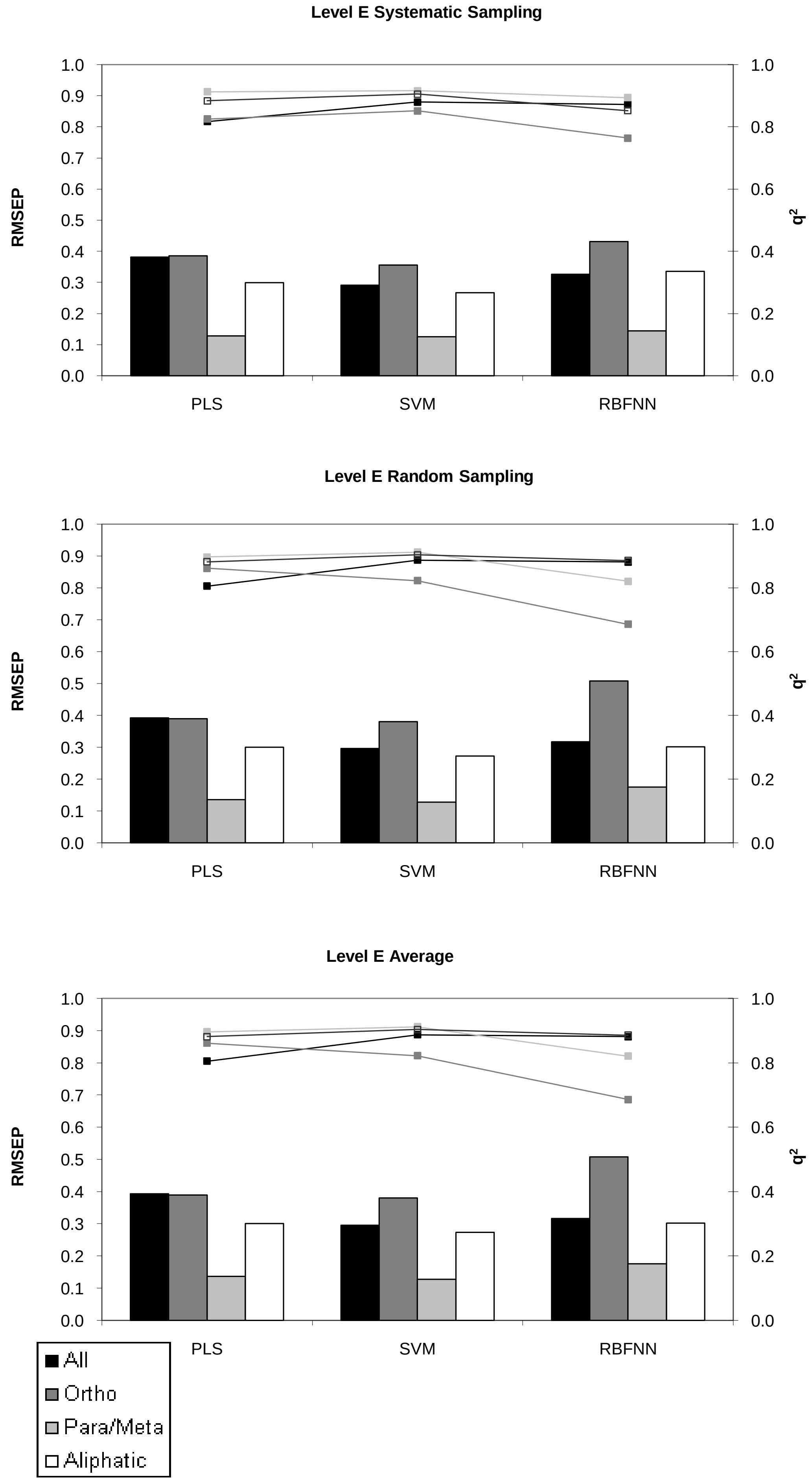


Figure 3.5. A graphical representation of Table 3.4 for level *E*. The bar charts represent the RMSEP, and the lines represent the q^2 values obtained.

3.3.7 Comparison to pK_a Prediction Software

To compare pK_a prediction by the QTMS method with available (commercial) software, we have used the methods provided by a number of organizations, namely ACD/Labs' pK_a DB[142], the SPARC[147] online calculator (SPARC Performs Automated Reasoning in Chemistry), VCCLAB's web-based ALOGPS 2.1 program[148], and ChemAxon's pK_a Plugin[149] for the Marvin software package. Recently, it was stated that the web-version of the SPARC performs 50,000-100,000 calculations per month[124]. Each software package enables the user to input the structures in SMILES format[150].

We removed each of the compounds in turn from the global and subset carboxylic acid models built using PLS, SVM and RBFNN, for both level *C* and *E* and rebuilt the models using the new model to predict the pK_a of the compound omitted[85]. Using this method, we acknowledge that the compounds are not an external test set (e.g. they have been used for initial variable and parameter selection in some cases) nor can we be sure that they have not been used to train the packages we investigated. Table 3.5 gives the RMSEP for the methods based on leave-one-out. The RMSEP obtained from testing the alternative computer programs are also given in Table 3.5. These results confirm that SVM provides the best models to predict pK_a. The SVM models have the lowest RMSEP in all the LOO cases apart from the level *C* ortho and para/meta carboxylic acid models, where the PLS models have the lowest RMSEP of 0.388 for the ortho model and 0.121 for the para/meta model, compared to 0.407 and 0.123 for the SVM models, respectively. Comparing levels of theory confirms that level *C* is the best as it generally provides the lowest RMSEP for the models. There are some exceptions to this. For example, the ortho model at level *E*, using SVM has an RMSEP of 0.367 compared to 0.407 for level *C*. Where level *E* provides a lower RMSEP the largest difference observed in RMSEP between level *C* and *E* is 0.04 for the SVM ortho models and the PLS all carboxylic acid models. In fact, the difference between the models at the different levels judged by RMSEP (based on LOO) is negligible when considering the large increase in CPU time needed to optimize the compounds at level *E* (See Figure 3.3).

Table 3.5. The RMSEP for the three learning methods based on leave-one-out. The RMSEP for the commercial computer programmes are given at the bottom of the table.

	RMSEP		
method	PLS	SVM	RBFNN
Level *C*			
QTMS (all)	0.427	0.293	0.363
QTMS (subset avg.)	0.285	0.276	0.321
QTMS ortho subset	0.388	0.407	0.477
QTMS para/meta subset	0.121	0.123	0.138
QTMS aliphatic subset	0.278	0.252	0.291
Level *E*			
QTMS (all)	0.396	0.301	0.323
QTMS (subset avg.)	0.311	0.278	0.323
QTMS ortho subset	0.407	0.367	0.486
QTMS para/meta subset	0.136	0.131	0.168
QTMS aliphatic subset	0.313	0.276	0.285
Compared software/tools	RMSE		
ACD/Laboratories	0.263		
VCCLAB	0.279		
SPARC	0.356		
ChemAxon	0.398		

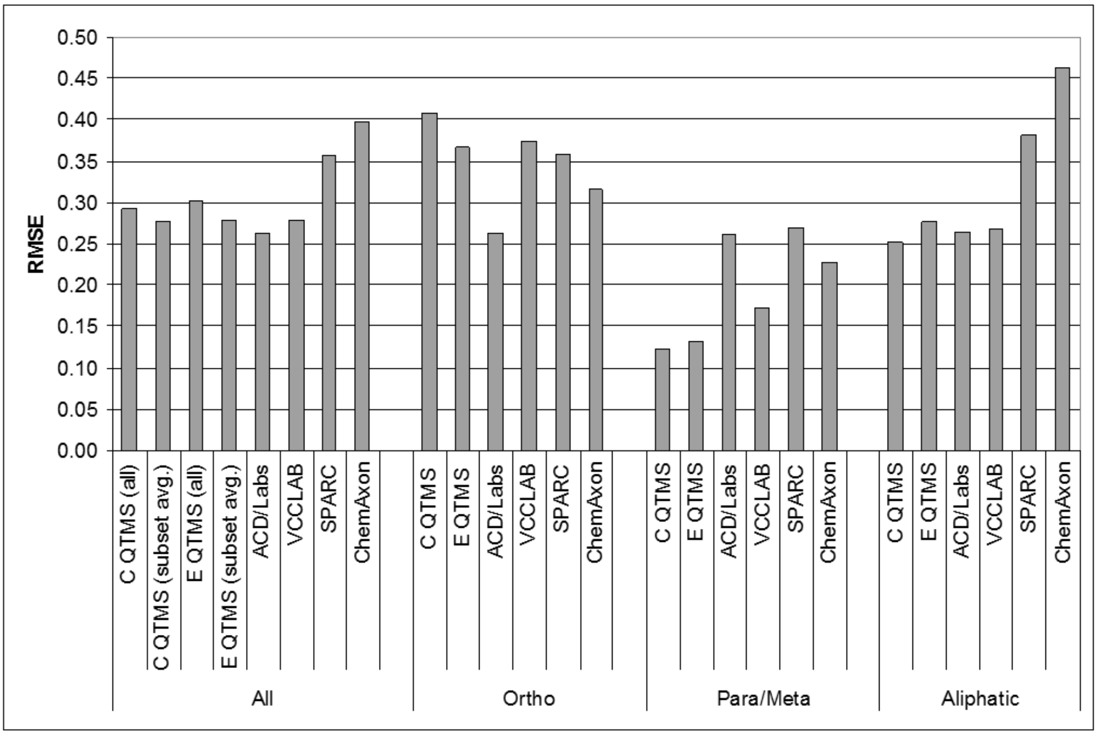


Figure 3.6. Comparison between QTMS and other pK_a prediction software, based on the RMSEP.

Figure 3.6 graphically compares our QTMS SVM models to the results obtained from the commercial predictions. Recently, Meloun and Bordovská[124] have rigorously compared the same packages using 64 drugs and other organic molecules with complex and diverse structural patterns. Although we only base our ranking on the RMSEP, we too found ACD/pK_a to be the most accurate method. This conclusion contradicts the findings of Dearden *et al.*[113] who compared ten prediction tools using a test set of 653 compounds. They found that ACD/pK_a was the *least* accurate out of the four programs we investigated, with and without inclusion of tautomeric compounds. The other three programs were ordered in the same way as in our study: VCCLAB being the most accurate followed by SPARC and then ChemAxon. Apart from ACD/Labs, which is consistent across all the subsets, the other methods vary in their prediction ability. Out of all the methods, QTMS has the lowest RMSEP for the para/meta substituted benzoic acids and aliphatic carboxylic acids, but has the highest RMSEP for the ortho-substituted benzoic acids. As has been previously pointed out[59, 70], QTMS fails when steric effects are important, which is the case for the ortho substituted benzoic acids.

3.4 Summary

The results presented in this systematic study indicate that BCP descriptors are effective in predicting the pK_a of small- to large-sized carboxylic acids of pharmaceutical relevance. Furthermore, extensive cross validation shows that there is no need to use the computationally more expensive level *E* when level *C* provides similar, if not superior, CV statistics. More predictive models were gained from splitting the dataset. Generally, SVM provides the best learning method although the lack of interpretability may mean it is not necessarily the most suitable method when mechanistic understanding is important. Finally, we have also demonstrated that predictions from our QTMS method compete with frequently used pK_a prediction tools.

Chapter 4
pK_a Prediction from an *ab initio* Bond Length for Phenols, Benzoic Acids and Anilines

4.1 Introduction

QTMS[51, 59, 63] is a new approach to solving QSAR/QSPR problems using properties defined by QCT[1, 2, 151]. QCT defines so-called critical points inside a given molecule, where quantum mechanical functions such as the electron density are evaluated. These and other values are QTMS descriptors. In Chapter 3 and our publication[152] we modelled the pK_a of 228 carboxylic acids using the QTMS methodology, in which equilibrium bond lengths are usually added to the descriptor pool[152]. Indeed, as early as 2002, *ab initio* equilibrium bond lengths featured in the rationalisation of antitumor activity of (*E*)-1-phenylbut-1-en-3-ones[63]. Better models were achieved using the descriptors defined by QCT than with bond lengths alone. This has generally been the case in previous QTMS studies that predicted pK_a and other properties[57, 69, 153, 154]. Superior models were achieved when the benzoic acids were split into ortho- and meta-/para-substituted groups. However, we believe that if the focus is placed on accuracy rather than globality, which means splitting chemical classes beyond the common aliphatic, ortho-, para- and meta-substituted groups, then strong correlations between a single *ab initio* bond length and pK_a are achievable, without the need for the computation of QCT descriptors. This is the approach and strategy in this Chapter. Furthermore, simple linear equations using just one bond length will be constructed and shown to be equal to if not better than using several bond lengths in more sophisticated multi-term equations. Quantum mechanical methods are becoming standard in computational drug design[155] and the equations presented here offer a simple and practical way to predict pK_a using information generated from first principles.

Using the accuracy of first-principle methods, Han and co-workers studied the complete series of chlorophenols[156]. Using B3LYP/6-311++G(d,p) for geometry optimisation, in conjunction with a molecular probe to simulate the acid-base interaction, they found that several molecular parameters correlated well with the acidity of the phenols. They found ammonia to be a better molecular probe than water because it is a stronger base and induces larger measurable changes in the molecular properties. The C-O bond length (*r*(C-O)), O-H bond length (*r*(O-H)) and O-H…N hydrogen bond length (*r*(O-H...N)) all correlated well with the experimental pK_a, with correlation coefficients (r^2) ranging from 0.89 to 0.97 for the phenol-ammonia complexes. The complete series of bromophenols, fluorophenols and hydroxybenzoic acids was also investigated using the same methods and similar correlations were noted[157, 158]. The authors of these papers demonstrated that weaker correlations were observed with the molecular properties of monomeric phenols without using the molecular probe[159, 160]. Strong correlations (r^2 > 0.92) were also found for aliphatic and carboxylic acids[161] with the use of a molecular probe. These correlations demonstrated that specific bond lengths could be used to predict pK_a beyond the complete series of halogen phenols, although ortho-substituted benzoic acids had to be modelled separately and all the benzoic acids were mono-substituted, apart from one compound. It was suggested that bond lengths were more practical to use because the calculation of vibration frequencies was computationally more demanding.

Yu *et al.*[162] have recently compared the semiempirical approach to predict pK_a, originally purposed by Tehan and co-workers[97], to ACD/Labs[121] and SPARC[118]. The semiempirical method performed significantly better when the data set was split into compound class-specific subsets. However, the overall performance was inferior to both that of ACD/Labs and SPARC. The authors suggest that improvements may be possible using higher-level quantum chemical methods to calculate the descriptors and the exploration of other quantum chemical parameters.

With the end-user in mind, we will demonstrate that the required accuracy in pK_a prediction can be achieved with a relatively low level of *ab initio* theory. This offers the opportunity for pK_a predictions of large data sets within an acceptable time. A comparison with previous work will show that the use of the probe molecule is unnecessary. The correlations are generated for phenols, carboxylic acids, and anilines, and subsequently used to predict the pK_a values of drug molecules. The advantage of single-term linear regression equations over multi-term equations will also be discussed. One advantage is the easier detection of outliers, allowing that the experimental data can be challenged. A second advantage is a reduced potential of over-fitting. Finally, we will describe a procedure that can be followed to predict pK_a. While this work is limited to three classes of compounds, the procedure is expected to be generic and hence applicable to a diverse range of compounds.

4.2 Methods and Computational Detail

4.2.1 Data Sets

Table 4.1 provides the constitution of the data sets for the phenols, benzoic acids and anilines. The experimental pK_a values for the phenols and benzoic acids were taken from a paper by Tehan *et al.*[97] while the anilines' experimental pK_a values were taken another paper[96] by Tehan *et al.*, unless otherwise stated. These authors had previously applied a variety of filters in order to remove non-druglike molecules. Where we have used other sources for experimental pK_a values to correct experimental values from our original data source, expanded the data set or tested our models, we explicitly highlight these occurrences in the text. The experimental pK_a values are listed in Appendix C with the corresponding chemical name and the identification numbers used in this Chapter.

Table 4.1. A summary of the data sets investigated.

Compound Class		# of compounds
Phenols		**171**
	Meta/Para[a]	55
	Ortho[b]	90
	Ortho, capable of forming Internal Hydrogen Bonds (ortho-phenols-IHB)[c]	26
Benzoic Acids		**94**
	Meta/Para[d]	44
	Ortho[e]	50
Anilines		**52**
	Meta/Para[f]	24
	Ortho[g]	28

[a] Two iodine containing compounds removed (4-iodophenol (compound 35) and 3-iodophenol (compound 50). 4-hydroxyacetophenone (compound 58) was also removed since the CAS number and name provided did not match, therefore causing ambiguity. The name given was hydroxyacetophenone whilst the CAS number relates to 4-hydroxyphenylacetaldehyde. The experimental pK_a quoted is 8.05, which is the same as that of 4-hydroxyacetophenone (compound 12).

[b] One iodine containing compound removed 2-iodophenol (compound 134). 2-methyl-4-chlorophenol (compound 90) corrected since the name and CAS number provided did not match. The name provided for compound 90 was already in the dataset (compound 174) so the CAS number was trusted and the structure corrected to 2-chloro-4-methylphenol.

[c] The name provided for compound 83 was incorrect and was corrected to 3,5-4′-trichloro-2′-nitro salicylanilide.

[d] Two iodine containing compounds removed (3-iodobenzoic acid (compound 207) and 4-iodobenzoic acid (compound 212)).

[e] Three iodine containing compounds removed (2-iodobenzoic acid (compound 233), 2-hydroxy-5-iodo-benzoic acid (compound 247) and 3,5-diiodosalicylic acid (compound 249)) .

[f] Two iodine containing compounds (3-iodoaniline (compound 297) and 4-iodoaniline (compound 298)) were removed. The macro pK_a value (3.07) for 3-aminobenzoic acid (compound 275) was provided in the original data set. The micro pK_a value (4.53) was found in the literature[72] and subsequently adopted.

[g] One iodine-containing compound removed (2-iodoaniline (compound 325)).

4.2.2 Data Generation and Analysis

The phenols, benzoic acids and anilines were treated separately. The discussion below provides a general overview of the data generation and analysis. More details about the exact analysis of each data set are given in the results (Section 4.3). An initial guess of the geometry of each compound was provided by MOLDEN[47]. Using the programme GAUSSIAN03[48], geometries were optimised at HF/6-31G(d) level. The bond lengths of interest were then extracted and a PLS[52] analysis was carried out to fit the bond lengths to the experimental pK_a values. SIMCA-P[54] was used for the majority of the data analysis. Models using all the bond lengths of interest were initially created using the predefined criterion for determining the significant number of Latent Variables (LVs) to appear in the PLS equation. If the value of q^2 of the newly constructed LV is less than 0.097, then no more LVs are computed; the PLS regression is then deemed complete. Separate models were also created for the ortho-, para- and meta-substituted compounds. Variable Important in the Projection (VIP) plots for the models were subsequently examined. VIP plots provide a condensed summary of the relative importance of each variable to the model, in this case the contribution of specific bond lengths. The bond lengths that contributed the most to the models were then used to construct one-term bond length models for the compound classes and the results analysed. Attempts were then made to separate these models into chemically meaningful groups of compounds where one common bond length showed high correlation with the experimental pK_a values. We refer to these groups of compounds as *high-correlation subsets.* Through the analysis of the single-bond-length equations, the influence of conformation was investigated. Outliers and errors were detected and where possible corrected. Higher levels of *ab initio* theory were examined and comparisons of the results with and without an ammonia probe were made for a selection of the high-correlation subsets. The predictions made from the high-correlation subsets were compared to the predictions made from models constructed using all the bond lengths and more diverse training sets. Models were validated using leave-many-out and compared using a variety of statistics discussed in Section 4.2.3 below. Finally, we tested the power of these models to predict the pK_a of drug molecules[127].

4.2.3 Statistics

In this work we report r^2, RMSEE and q^2 values for the constructed models. We assess our models by means of RMSEE, which is the strictest criterion of quality, both in terms of outlier assessment and the series of inequalities mentioned in Chapter 2. We base our comparisons on RMSEE and use this as a guide to indicate which models should be cross validated. We only performed full *k*-fold CV on the most promising models. In the current work *k* is set to exactly seven throughout, provided there are more than 7 data points in the total data set.

One may question whether the use of CV is justified against an assessment based on splitting the data set into a training and test set. Hawkins *et al*.[56] recommend that, when the data set is small, then CV may be better than splitting the data set into training and test sets. The high-correlation subsets we use to create models are small in the sense of Hawkins *et al*. who advocate CV when the data set contains less than 100 compounds. CV should involve using a suitable variable selection technique to select the variables important to the training set, each time a CV subset is excluded. This procedure renders a 'true q^2' rather than a 'naïve q^2', where variable selection is not performed each time a CV subset is removed. Below we argue that our assessment procedure is generating a true q^2. Essentially, the main argument is that variable selection does not apply to our way of setting up a model. This is because we only consider either an all-bond-length model or a single-bond-length model. The latter type of model was based on the most important variable in the VIP plot of the all-bond-length model. During the CV of the all-bond-length model, the VIP plots for the seven models created in CV were monitored and in the vast majority of cases, the most important bond length remained the same for all the models. Furthermore, SIMCA-P automatically selects the number of LVs to construct the all-bond-length models. Because a new model is constructed for each of the seven training sets, variable selection is performed by default. For these reasons we consider the q^2 value quoted to be the 'true q^2'. By default, SIMCA-P automatically produces a q^2 value when models are constructed.

CV also provides a means to calculate the RMSEP (Equation 2.22) for the cross-validated models. The squared correlation coefficient also obtained through CV and denoted by r^2_{CV}, which is not be confused with q^2, is calculated as,

$$\mathrm{r}^2_{\mathrm{CV}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\hat{\mathrm{y}}_{\mathrm{pred,i}} - \bar{\mathrm{y}}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{obs,i}} - \bar{\mathrm{y}}\right)^2} \qquad \text{Equation 4.1}$$

where the variables have already been explained.

We also use a further metric denoted as r^2_m, which is calculated as,

$$\mathrm{r}^2_{\mathrm{m}} = \mathrm{r}^2_{\mathrm{CV}} \times (1 - \sqrt{\mathrm{r}^2_{\mathrm{CV}} - \mathrm{r}^2_{\mathrm{CV,0}}}) \qquad \text{Equation 4.2}$$

Here, r^2_{CV} and $r^2_{CV,0}$ are the squared correlation coefficient values between the observed (X-variable) and predicted (Y-variable) pK_a values, obtained through CV, with intercept *not* set to zero and set to zero, respectively.

A high r^2_{CV} value does not necessarily indicate that the predicted values are very close to the experimental values. There may be considerable numerical differences between the observed and predicted values in spite of the presence of a good overall correlation. When this is the case there will be substantial differences between r^2_{CV} and $r^2_{CV,0}$, which the r^2_m statistic penalises heavily. Mitra *et al.*[154] have shown that in the case of small data sets, r^2_m calculated from a CV when variable selection is performed at each CV step, reflects the external validation characteristics of the developed model. Based on the reasoning above about 'true q^2' we believe that our quoted r^2_m values are 'true $r^2_{m\,(leave-one-seventh-out)}$'. Ultimately, we judge the performance of the models based on RMSEP. However, we stress that RMSEE is used to decide which models to perform full CV on.

4.3 Results

4.3.1 Phenols

Figure 4.1 shows the common skeleton and bonds screened to predict the pK_a of the phenol compounds (Table 4.1). In previous QTMS studies, conformation has not been taken into account with the knowledge that the substitution effects have a greater influence on the models. We show that conformation is important in some cases (discussed later). We initially investigated the use of the 8 phenol bond lengths of the common skeleton in order to model all the phenol compounds together and the common groupings of ortho, meta/para and ortho-phenols that we deemed capable of forming internal hydrogen bonds[97]. During the course of this work, suspected errors in the initial dataset were corrected and the importance of conformation was examined.

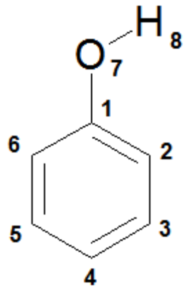


Figure 4.1. The eight bond lengths used to predict the pK_a of the phenol compounds. The main text refers to bond lengths 1-7 and 7-8 as r(C-O) and r(O-H), respectively. Reference to other bonds makes use of this numbering scheme to distinguish between the C-C bonds, e.g. r(C_1-C_2).

Inspection of the VIP plot modelling all the phenol compounds showed that r(C-O) and r(O-H) contributed most to the model. Therefore, these bonds were monitored to see if they could model the 171 phenol compounds individually (Table 4.2), in line with our motivation discussed in the Introduction. The r^2 decreased and the RMSEE increased when these two bond lengths were used individually. The reduction in the quality of the model was less for r(C-O) on its own than for r(O-H) on its own. We determined what influence splitting the data set into meta-/para- and ortho-substituted phenols (common for compounds of this type) had on the quality of the models. The RMSEE for the model constructed using all the bond lengths for meta-/para-substituted phenols decreased by approximately 50% but the RMSEE for the ortho-substituted phenols increased. The reduction in the quality of the models of the meta-/para-substituted phenols when using r(C-O) on its own or r(O-H) on its own is also small compared to the all-bond-length model. This suggests that the meta-/para-substituted phenols can be modelled using just one bond length. The RMSEE for the ortho-substituted model increased when using r(C-O) on its own or r(O-H) on its own. This is not surprising since different ortho-substituents can affect the pK_a of compounds because of their close proximity to the acidic hydrogen. These effects include steric hindrance to protonation or deprotonation and internal hydrogen bonding.

To investigate the large deterioration of the ortho-substituted phenol model, when using all the bond lengths compared to using just r(C-O) or r(O-H) on their own, we inspected the predicted versus observed pK_a plots. Figure 4.2a shows such a plot for all 171 phenols using a regression model using only r(C-O). Inspection of Figure 4.2a suggests subsets of phenols that have a higher r^2 value than the full set of 171 phenols. It was rewarding to find that such *high-correlation subsets*, identified by eye, later turned out to be meaningful chemical subsets. For example, in Figure 4.2b it is clear that o-halogen phenols (shown in dark blue) and o-nitro (shown in light blue) phenols are separate high-correlation subsets. This was seen for other o-phenols depending on the o-substituent. It appeared that meta-/para-phenols were a high-correlation subset irrespective of the different substituents. A number of compounds appeared to be outliers from the high-correlation subset to which they would have been expected to belong to. An example of this is shown in Figure 2b for 4,6-dinitro-o-cresol (compound 135). 4,6-dinitro-o-cresol appeared to belong to the o-halogen high-correlation subset (dark blue in Figure 4.2b), which was inconsistent with the compound's structure. Inspection of the optimised structures showed that this was caused by the anti conformation of the acidic proton being used instead of the syn conformation that was found for the other o-nitrophenols. When this compound was optimised as the syn conformer, it correctly moved into the o-nitrophenol high-correlation subset (light blue in Figure 4.2b). This example illustrates the situation for a number of other compounds that appeared to belong to high-correlation subsets different to the chemically meaningful subsets we had identified. All the ortho-phenols were subsequently optimised in the syn and anti form and the energies were used as a guide to decide which high-correlation subset they belonged to. Because of symmetry conformation plays no role in di-ortho-substituted phenols with identical substituents. However, for the asymmetrical di-ortho-substituted and the mono-ortho-substituted compounds, the orientation of the acidic hydrogen can have a large influence on bond lengths. The results of the detailed modelling of the o-phenols and identification of high-correlation subsets are reported in the following section.

Table 4.2. The results of the phenol compounds modelled with the bond lengths calculated at the HF/6-31G(d) level of theory.

Subsets	# LV	# Bonds	# Compounds	r^2	q^2	RMSEE
All	4	All	171	0.92	0.88	0.67
All	1	r(C-O)	171	0.86	0.85	0.88
All	1	r(O-H)	171	0.52	0.51	1.62
Meta/Para	2	All	55	0.91	0.87	0.34
Meta/Para	1	r(C-O)	55	0.87	0.85	0.41
Meta/Para	1	r(O-H)	55	0.84	0.83	0.45
Ortho	4	All	116	0.92	0.86	0.72
Ortho	1	r(C-O)	116	0.85	0.85	0.99
Ortho	1	r(O-H)	116	0.47	0.46	1.84
Ortho without Ortho-IHB	5	All	90	0.94	0.88	0.65
Ortho without Ortho-IHB	1	r(C-O)	90	0.88	0.87	0.94
Ortho without Ortho-IHB	1	r(O-H)	90	0.59	0.58	1.72

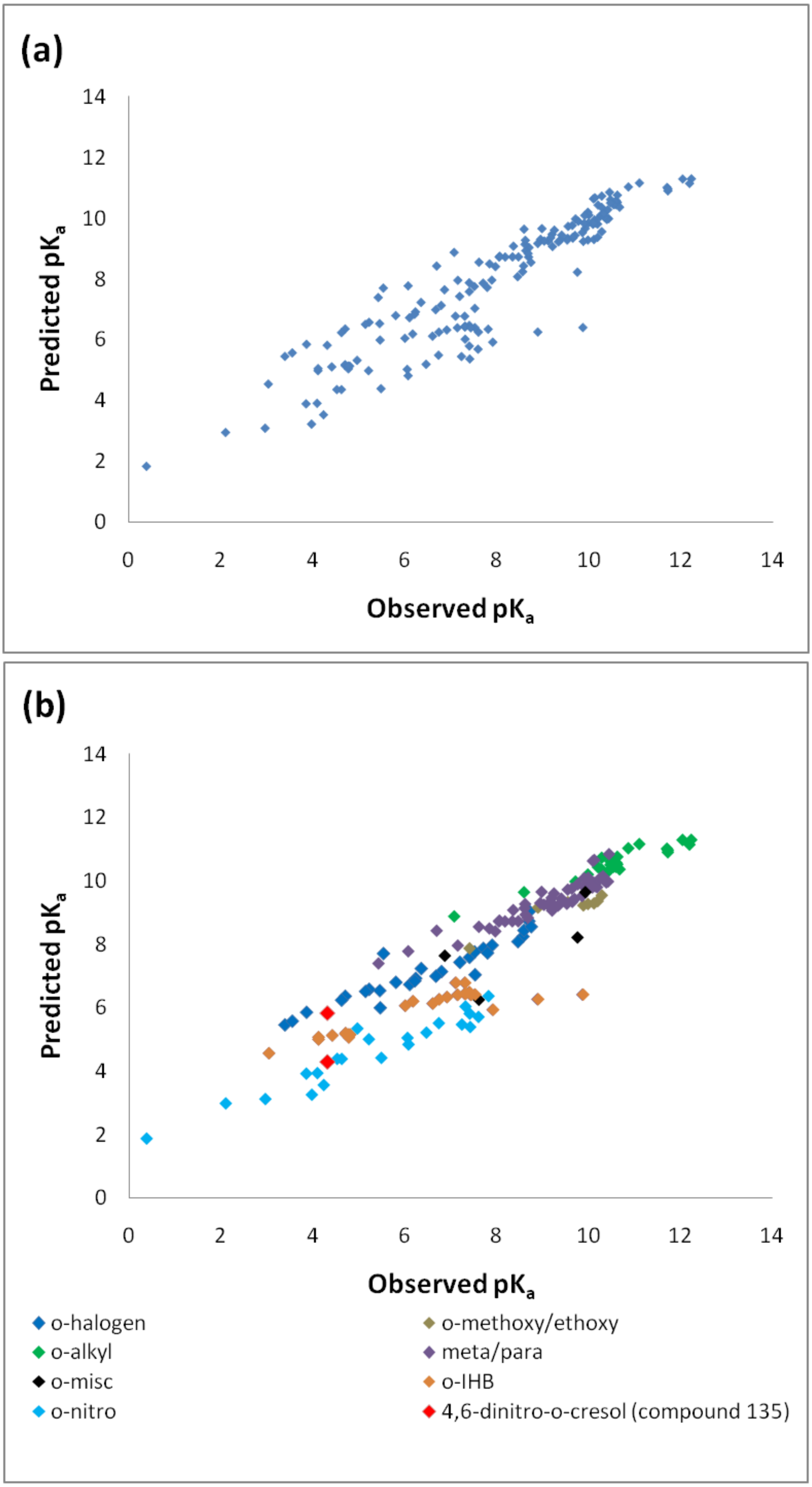


Figure 4.2. (a) Plot of predicted vs. observed pK_a for the phenols using r(C-O). (b) Plot of the predicted vs. observed pK_a for the phenols using r(C-O) separated by colour into chemically meaningful high-correlation subsets. The different pK_a values of 4,6-dinitro-o-cresol calculated from r(C-O) for the syn and anti conformer is highlighted in red as an example.

4.3.2 Ortho-Phenols

By Inspection of the structures of the compounds belonging to high-correlation subsets and their energies, rules were determined based on the o-phenols in the data set to assign the compounds to specific o-phenol high-correlation subsets. These rules were confirmed by the detailed investigation of the high-correlation subsets. In Sections 4.3.2.1 to 4.3.2.6 we discuss the results that allowed us to state the rules here. In the case of the o-phenols it was fortuitous that the energies could be used as a guide, without exception. These rules are encapsulated in the flow chart below (Figure 4.3) showing which high-correlation subset a phenol of interest should be predicted from.

It should be noted that certain phenols can belong to different high-correlation subsets. For example, 2-nitro-6-chlorophenol (compound 141) can be predicted by the o-nitro and the o-halogen models depending on the direction of the acidic hydrogen (i.e. syn and anti). We will show that the better prediction is made by the o-nitro model because nitro substituents decrease the pK_a more than chlorine substituents as the former are more electron-withdrawing. We will also show that for meta-/para-substituted phenols the influence of conformation on the quality of the models is minimal. We screened the compounds in search of high-correlation subsets from different classes of compounds in the ortho subset, previously (see RMSEE values larger than 0.5 in Table 4.2) shown to produce poor correlations when modelled together. We compare all-bond-length models to single-bond-length models using CV discussed in Section 4.2.3.

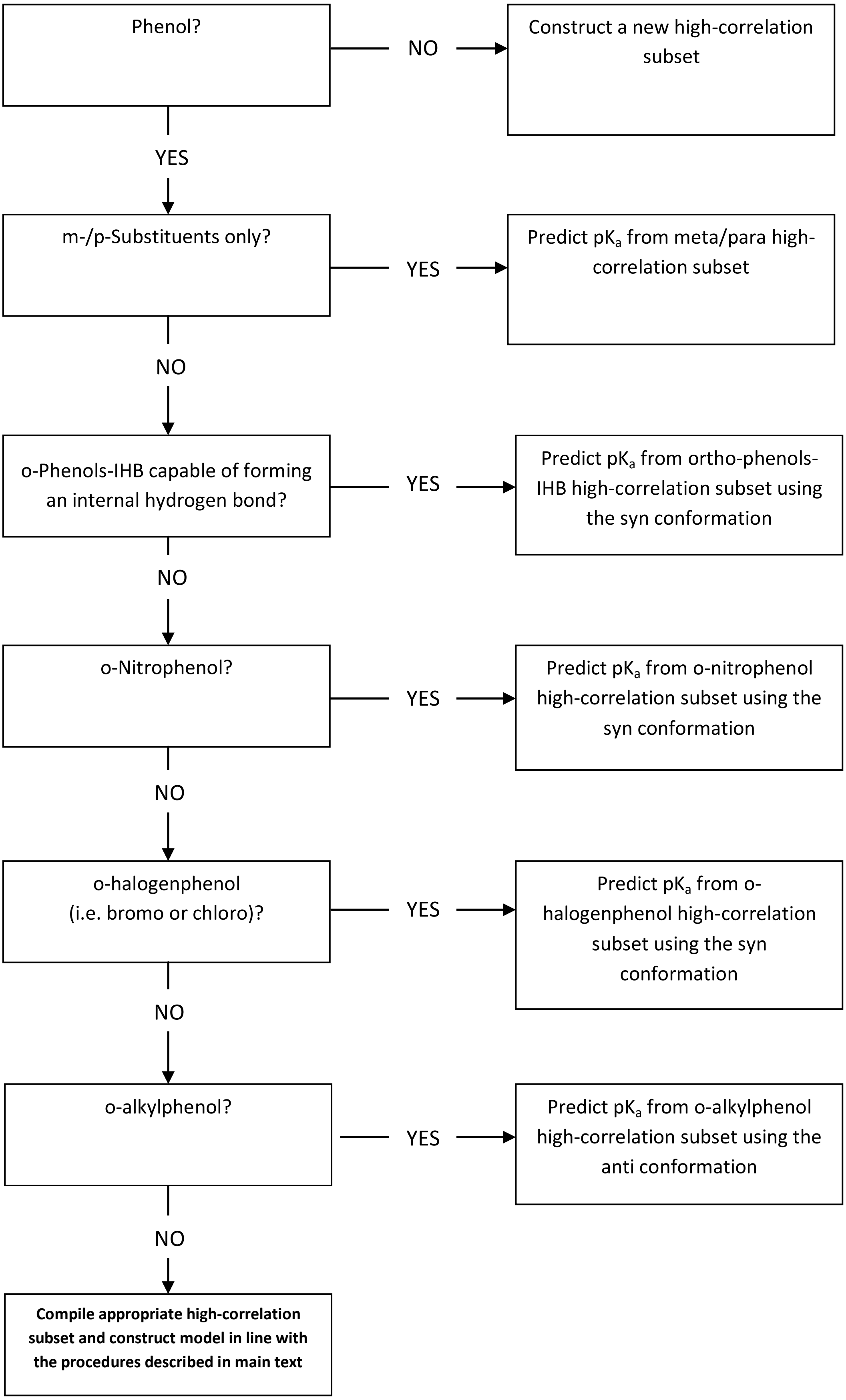

Phenol?
NO
Construct a new high-correlation subset
YES
m-/p-Substituents only?
YES
Predict pK$_a$ from meta/para high-correlation subset
NO
o-Phenols-IHB capable of forming an internal hydrogen bond?
YES
Predict pK$_a$ from ortho-phenols-IHB high-correlation subset using the syn conformation
NO
o-Nitrophenol?
YES
Predict pK$_a$ from o-nitrophenol high-correlation subset using the syn conformation
NO
o-halogenphenol (i.e. bromo or chloro)?
YES
Predict pK$_a$ from o-halogenphenol high-correlation subset using the syn conformation
NO
o-alkylphenol?
YES
Predict pK$_a$ from o-alkylphenol high-correlation subset using the anti conformation
NO
Compile appropriate high-correlation subset and construct model in line with the procedures described in main text

Figure 4.3. Flow chart describing which high-correlation subset a new compound should be predicted from.

4.3.2.1 o-Nitrophenols

Table 4.3. The statistical details of the models created for the o-nitrophenols.

					Model Statistics				**CV Statistics**		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Syn	All	4	23	0.97	0.37	0.91	0.65	0.88	0.87	0.79
HF	Syn	r(C-O)	1	23	0.91	0.58	0.91	0.58	0.91	0.90	0.82
HF	Syn	r(O-H)	1	23	0.85	0.77	0.84	0.78	0.83	0.80	0.70
HF	Syn	All	2	22	0.97	0.38	0.93	0.51	0.94	0.94	0.93
HF	**Syn**	**r(C-O)**	**1**	**22**	**0.94**	**0.48**	**0.94**	**0.50**	**0.93**	**0.93**	**0.87**
HF	Syn	r(O-H)	1	22	0.88	0.71	0.87	0.73	0.85	0.84	0.75
HF	Anti	All	2	23	0.98	0.33	0.88	0.81	0.82	0.80	0.70
HF	Anti	r(C-O)	1	23	0.79	0.90	0.78	0.91	0.77	0.71	0.58
B3LYP	Syn	All	2	23	0.94	0.50	0.90	0.61	0.89	0.89	0.82
B3LYP	Syn	r(C-O)	1	23	0.91	0.58	0.91	0.59	0.90	0.89	0.81

The results from CV are reported in Table 4.3 to allow comparisons between models. Inspection of the VIP plot for the all-bond model using all the o-nitrophenols revealed that r(C-O) was the most important descriptor followed by r(O-H). For this reason we created separate models for each of these two bond lengths. Looking at r^2 it is surprising that this value remains high for either of the single-bond-length models compared to the all-bond-length model. Inspection of the plot showing observed versus predicted pK_a values for the single-bond-length models caused suspicion about the experimental pK_a of 2,3-dinitrophenol (compound 87, experimental pK_a given as 4.96). Another source[109] quoted the experimental pK_a of this compound to be 5.24. This increase in pK_a moves it towards a value of approximately 6 log units predicted by our different models. Removal of compound 87 from the fitting procedure improved the model statistics. When compound 87 was removed during CV of the all-bond-length model, the resulting model used only two LVs compared to the three LVs making up the models with compound 87 included. This suggests that the program SIMCA-P had added a LV to fit compound 87. This was not the case for the single-bond-length models as the fitting was minimal here. During CV, the VIP plots of the models were inspected when each CV group was removed in turn. r(C-O) followed by r(O-H) were the most important bonds to all the models in CV. The r(C-O) model, when compound 87 was removed, produced the lowest RMSEP (0.50) and a high r^2_m (0.87). This was pleasing considering only one bond length is used.

Table 4.3 also provides the statistics relating to the models built using anti conformations. The all-bond-length model has the highest r^2 value in conjunction with the lowest RMSEE. However, the model is shown to be weaker when CV is performed compared to that constructed using the lower energy syn conformations. The r(C-O) model using the anti conformations is also poorer than when the syn conformers are used. The VIP plot for the all-bond-length anti model showed r(C-O) to be the most important, however, r(C_5-C_6) was the next most important and r(O-H) was ranked sixth.

The use of a higher level of theory, B3LYP/6-311+G(2d,p), did not improve the results (Table 4.3) and produced very similar statistics suggesting the more economical HF/6-31G(d) is sufficient. This conclusion is supported by a study of 2,4-dinitrophenol where the HF/6-31G(d) level of theory performed very well for predicting the geometrical parameters[163]. In that work, the HF method failed to reproduce the vibration frequency of the O-H bond stretch. However, this is of no importance as we only use the bond lengths.

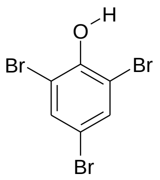
4.3.2.2 *o-Halogen Phenols*

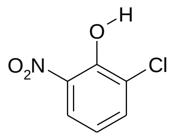
Table 4.4. The statistical details of the models created using o-halogen phenols.

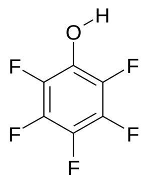
					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Syn	All	1	32	0.75	0.89	0.71	1.05	0.64	0.55	0.45
HF	Syn	r(C-O)	1	32	0.88	0.61	0.87	0.62	0.87	0.86	0.79
HF	Syn	r(O-H)	1	32	0.63	1.08	0.63	-[a]	-	-	-
HF	Syn	All	2	26	0.95	0.40	0.91	0.44	0.93	0.91	0.81
HF	**Syn**	**r(C-O)**	**1**	**26**	**0.97**	**0.27**	**0.97**	**0.29**	**0.97**	**0.97**	**0.94**
HF	Syn	r(O-H)	1	26	0.60	1.06	0.58	-	-	-	-
HF	Anti	All	2	32	0.83	0.74	0.70	-	-	-	-
HF	Anti	rC-O)	1	32	0.83	0.72	0.81	-	-	-	-
HF	Anti	r(O-H)	1	32	0.61	1.12	0.53	-	-	-	-
HF	Anti	All	2	26	0.94	0.40	0.91	0.46	0.92	0.91	0.81
HF	Anti	r(C-O)	1	26	0.96	0.35	0.96	0.35	0.95	0.95	0.91
HF	Anti	r(O-H)	1	26	0.78	0.79	0.76	-	-	-	-
B3LYP	Syn	All	1	32	0.71	0.97	0.65	-	-	-	-
B3LYP	Syn	r(C-O)	1	32	0.87	0.64	0.86	-	-	-	-
B3LYP	Syn	r(O-H)	1	32	0.52	1.24	0.44	-	-	-	-
B3LYP	Syn	All	2	26	0.93	0.46	0.86	0.67	0.83	0.81	0.70
B3LYP	Syn	r(C-O)	1	26	0.96	0.32	0.96	0.33	0.96	0.96	0.92
B3LYP	Syn	r(O-H)	1	26	0.72	0.89	0.70	-	-	-	-

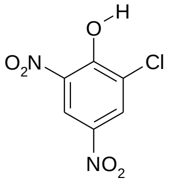
[a] A dash in this Table 4.5 to Table 4.33 indicates that various cross-validation statistics were not collected as justified in the main text.

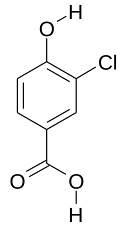
The statistics relating to the o-halogen phenols are given in Table 4.4. The o-halogen phenols were initially modelled as syn conformations. The models built from all the compounds were inspected. The VIP plot for the all-bond-length model showed r(C-O) was the most important followed by r(O-H). The r(C-O) model gave better statistics than the all-bond-length model. Inspection of the observed versus predicted plot for this model revealed six suspicious data points. The structures of the compounds that represent these points are shown in Figure 4.4. We will discuss each outlier in turn starting with 2,4,6-tribromophenol. A different experimental pK_a of 6.1 for 2,4,6-tribromophenol (compound 120) was found[109] instead of the value of 6.8 given in the source we used for the experimental pK_a values. The value of 6.1 was much closer to that predicted from our correlation and was hence adopted.


120


141


156


160


171

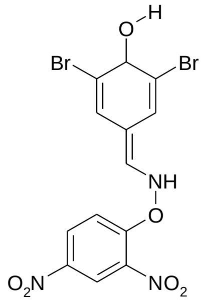

175

Figure 4.4. Structures (and their compound numbers) that belong to the data points that seemed to be outliers in the o-halogenphenol models.

Next we explain why 6-chloro-2-nitrophenol (compounds 141) and 6-chloro-2,4-dinitrophenol (compound 160) should only be predicted from the o-nitro high-correlation subset model. We note that these compounds have both o-nitro and o-halogen substituents and so the pK_a could be predicted by either the o-nitro or o-halogen high-correlation subsets. To obtain a reasonable prediction from the latter for compounds 141 and compound 160, we used the conformation in which the acidic hydrogen points towards the halogen, which we note is *not* the lowest energy, but is consistent with the conformations used for the other o-halogen phenols. 6-chloro-2,4-dinitrophenol (compound 160) was predicted reasonably well by the r(C-O) model (experimental pK_a of 1.6 compared to a predicted pK_a of 2.1). However, 6-chloro-2-nitrophenol (compound 141) had an error of 1 pK_a. These two compounds were predicted more accurately by the o-nitro high-correlation subset, where the acidic hydrogen points towards the nitro groups, which were the lowest energy conformations. From these two compounds, we conclude that phenols with a nitro and a halogen substituent in either ortho positions should be predicted from the o-nitro high-correlation subset and not the o-halogen high-correlation subset.

Pentafluorophenol (compound 156) showed the largest discrepancy between observed and predicted pK_a. 2-fluorophenol (compound 129) was the only other compound in our data set that had an o-fluoro substituent and appeared to belong to the o-halogen high-correlation subset. It was clear that o-chlorophenols and o-bromophenols formed a single high-correlation subset. However, because we only had two o-fluorophenols in the dataset it was impossible to establish if this class of compounds needed to be modelled separately or if the experimental pK_a of pentafluorophenol (compound 156) should be challenged. The experimental value we adopted from the work of Tehan *et al*. was verified against an alternative literature source[157], where the same pK_a value of 5.53 was used to produce excellent correlations. This check confirmed that the experimental value used is accurate. We therefore calculated the bond lengths of a further three o-fluorophenols, for which we had experimental pK_a values, to verify that they produce a separate high-correlation subset. The r(C-O) of 2,4-difluorophenol (compound 330), 2,6-difluorophenol (compound 331) and 2,3,5,6-tetrafluorophenol (compound 332) were calculated and the correlation between r(C-O) of the five o-fluorophenols and pK_a was checked. An r^2 of 0.91 and RMSEP of 0.40 suggested that o-fluorophenols indeed produce their own high-correlation subset and cannot be included with the other o-halogen compounds. To confirm that this was not a fortuitous result based on the HF/6-31G(d) level of theory, we compared our result to that obtained by Han and Toa[157] using B3LYP/6-311++G(d,p) and an ammonia probe. Using their r(C-O) equation to predict the pK_a for the same five o-fluorophenols, we obtained an r^2 and RMSEP of 0.90 and 0.41, respectively. After this confirmation we removed 2-fluorophenol (compound 129) and pentafluorophenol (compound 156) from subsequent analysis of the o-halogen high-correlation subset because it was clear they produced a separate o-fluorophenol high-correlation subset. Reasons for 3-chloro-4-hydroxybenzoic acid (compounds 171) and bromofenoxim (compound 175) having large residuals were unclear but they were also excluded.

Now we focus on the influence of the omission of outliers. Table 4.4 shows how the models improved when the six compounds were removed, resulting in good CV results. Table 4.4 shows that the single-bond-length models benefit approximately equally from this omission compared to the all-bond-length models. For example, upon omission of six outliers the RMSEE for the r(C-O) model roughly halves, from 0.61 to 0.27. Equally, the RMSEE for all-bond-length model also halves from 0.89 to 0.40. A similar trend is observed for RMSEP. The most dramatic improvement due to the omission of outliers is seen in the r_m^2 statistic. For the all-bond-length models with outliers included, a r_m^2 value of 0.45 suggests that poor predictions are made in CV, while reasonable predictions are made for the single-bond-length model, suggested by an r_m^2 of 0.79. After outlier omission, the r_m^2 value for the all-bond-length model improves to 0.81, suggesting a large improvement in prediction. However, the r(C-O) model without outliers is superior based on an r_m^2 of 0.94. It is interesting to note that the r(C-O) models are always superior to the all-bond-length models in the original model fit and in CV. The r(O-H) models were not cross-validated as they were inferior to the other models based on the original fitting statistics. This is why the corresponding CV statistics are not listed in Table 4.4.

High-correlation subsets using the anti conformations were also investigated with and without the identified outliers. The outliers were still suspicious data points in the inspected correlations. These were removed and models with all bond lengths and r(C-O) were cross-validated to compare to the models constructed using the syn conformations. Using just the r(C-O) provided a better model than using all bond lengths, as with the syn conformer models, but not as good as the models where the syn conformers were used. Using r(O-H) once again provided a poor correlation. Inspection of the observed pK_a versus predictive pK_a plot from r(O-H) revealed high-correlation subsets different to those seen when all the phenols were modelled with r(C-O). The structures of the phenols producing these separate high-correlation subsets were inspected and showed that di-o-bromophenols, di-o-chlorophenols and mono-orthophenols (i.e. those substituted with a chlorine or bromine at the ortho position) belong to their own subsets. This is not surprising, as r(O-H) is affected by the substituent that it points towards, resulting in separate models for the di-o-phenols and a single high-correlation subset for the mono-o-phenols as the acid hydrogen points towards a hydrogen in each case. This observation is confirmed by 2-chloro-6-methylphenol (compound 90) not belonging to any high-correlation subset as the methyl group has a different influence to that of a hydrogen. These results confirm the success of r(C-O) and the syn conformation models. The statistics of the models constructed with and without the six outliers using B3LYP/6-311+G(2d,p) geometries and the syn conformation offer no improvement to those created using HF/6-31G(d).

4.3.2.3 o-Alkylphenols

Table 4.5. The statistical details of the models created using o-alkylphenols.

					Model Statistics				CV Statistics		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	anti	All	3	29	0.96	0.44	0.94	-	-	-	-
HF	anti	r(C-O)	1	29	0.93	0.58	0.3	-	-	-	-
HF	anti	r(O-H)	1	29	0.79	1.00	0.78	-	-	-	-
HF	anti	All	2	25	0.94	0.28	0.83	0.42	0.79	0.76	0.64
HF	**anti**	**r(C-O)**	**1**	**25**	**0.91**	**0.34**	**0.9**	**0.37**	**0.89**	**0.87**	**0.78**
HF	anti	r(O-H)	1	25	0.36	0.90	0.36	0.91	0.30	-1.06	-0.05
B3LYP	anti	All	3	29	0.95	0.52	0.91	-	-	-	-
B3LYP	anti	r(C-O)	1	29	0.92	0.62	0.92	-	-	-	-
B3LYP	anti	r(O-H)	1	29	0.83	0.91	0.83	-	-	-	-
B3LYP	anti	All	2	25	0.92	0.33	0.76	-	-	-	-
B3LYP	anti	r(C-O)	1	25	0.92	0.32	0.91	-	-	-	-
B3LYP	anti	r(O-H)	1	25	0.37	0.90	0.35	-	-	-	-

The anti conformation is the lowest energy for the o-alkylphenols, which is opposite to the syn conformation favoured by the o-nitro and o-halogenphenols. For symmetrical 2,6-substituted phenols the conformation is irrelevant. However, the acidic hydrogen pointing towards an alkyl group is more stable than it pointing out of plane between the two ortho substituents. For asymmetrical 2,6-substituted phenols, e.g. 2-(1,1-dimethylethyl)-4,6-dimethylphenol (compound 164), the conformation with the hydrogen pointing towards the methyl group has the lowest energy. Three compounds that also had o-nitro and o-halogen substituents were initially included in the modelling (Table 4.6) as the conformation where the acidic hydrogen pointed towards the alkyl substituent, which we note is not the lowest energy. These compounds fitted into the alkyl high-correlation subset relatively well. However, they are modelled better in the lower energy conformation by the o-nitro and o-halogen high-correlation subsets. Therefore, the compounds were excluded from CV in agreement with two rules listed in Section 4.3.2. The models using all the bond lengths are comparable to the model using just r(C-O) (Table 4.5). The correlation obtained using r(O-H) was inferior to that obtained using r(C-O). There was no notable improvement when using B3LYP/6-311+G(2d,p) generated bond lengths compared to those calculated with HF/6-31G(d).

4.3.2.4 *o-Phenols Capable of Forming Internal Hydrogen Bonds.*

Table 4.6. The statistical details of the models created using o-phenols capable of forming internal hydrogen bonds.

					Model Statistics				**CV Statistics**		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	O-H-O	All	2	26	0.83	0.71	0.8	-	-	-	-
HF	O-H-O	r(C-O)	1	26	0.77	0.81	0.76	-	-	-	-
HF	O-H-O	r(O-H)	1	26	0.30	1.41	0.27	-	-	-	-
HF	O-H-O	r(C_2-C_3)	1	26	0.72	0.90	0.69	-	-	-	-
HF	O-H-O	All	1	23	0.88	0.50	0.83	0.79	0.74	0.74	0.71
HF	**O-H-O**	**r(C-O)**	**1**	**23**	**0.95**	**0.32**	**0.95**	**0.33**	**0.94**	**0.94**	**0.92**
HF	O-H-O	r(O-H)	1	23	0.35	1.16	0.23	1.31	0.15	-1.50	-0.04
HF	O-H-O	r(C_2-C_3)	1	23	0.85	0.56	0.82	0.62	0.80	0.78	0.69

Twenty-three out of the 26 compounds in this high-correlation subset had the same common skeleton with different substituents around both aromatic rings (Figure 4.5). Two different internal hydrogen bonds can be formed i.e. O…H-N and O-H…O. Because the latter structure corresponds to the lowest energy this was the only conformation considered. The three compounds that did not have the same common skeleton are shown in (Figure 4.6).

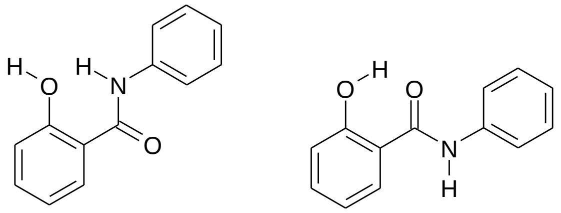
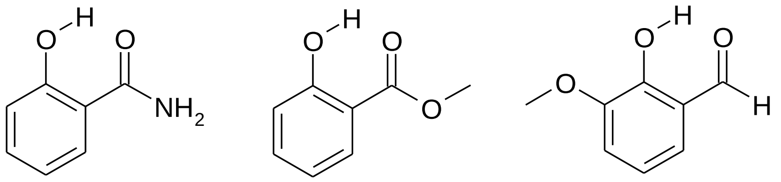


Figure 4.5. The common skeletons and different hydrogen bonds than can be formed by the o-phenols capable of forming internal hydrogen bonds.



59 61 62

Figure 4.6. The three compounds in the class of phenols capable of forming internal hydrogen bonds that did not have the same common skeleton as the rest of the compounds.

The observed versus predicted plots for the models containing all 26 compounds were inspected. The three compounds, 2-hydroxybenzamide (compounds 59), methyl salicylate (compound 61) and 2-vanillin (compound 62), which did not have the same common skeleton as the majority of the compounds in this high-correlation subset, appeared to be outliers. Methyl salicylate was identified as an outlier by Tehan *et al.*[97]. One would not be surprised if this compound did not fit the models because of its similarity to compounds 59 and 62. The reason for these outliers could be the lack of structural similarity between these three compounds and the rest of the o-phenols capable of forming an internal hydrogen bond. We suggest that these compounds belong to their own high-correlation subset needed to predict pK_a using just r(C-O). This was confirmed by an r^2 value of 0.95 for the correlation of pK_a and r(C-O), although more data points for these types of compounds are needed to confirm this. The pK_a of the remaining 23 compounds were modelled using all the bond lengths (Table 4.6). Once again the r(C-O) was most important in the VIP plot, however, it was followed by r(C_2-C_3) and not r(O-H). The r(C-O) model gave the best statistics for the original model and CV statistics compared to the all-bond-length model and the other single-bond-length models.

4.3.2.5 o-Methoxy/ethoxyphenols

This high-correlation subset consisted of only eight compounds. The pK_a range was small (7.4 for vanillin (compound 124) to 10.28 for 4-methyl-2-methoxyphenol (compound 104)). Removing vanillin, which had a much lower pK_a value than the rest, resulted in a range of only 0.74 log units. The syn conformation is the lowest energy in all cases. According to the statistics, the models deteriorate when vanillin is included (Table 4.7). To increase the size of the dataset we sourced 21 compounds from Ragnar *et al.*[164]. Five of these compounds were already present in our dataset. A comparison of the given pK_a values in that publication and in our dataset showed they were in good agreement, the largest difference being 0.05 pK_a units. We used the 16 remaining compounds as a test set for the syn models. It was pleasing to note that including vanillin gave lower values for RMSEP in all cases and that the r(C-O) bond length model gave the lowest RMSEP (Table 4.8). The models created without vanillin had rather poor CV statistics (i.e. q^2) because of the small pK_a range. However, these models actually produced reasonable predictions for the test set, which included extrapolation outside the range of pK_a values used to create the models. We added the 16 compounds to the Tehan compounds and created new models containing more compounds to increase the domain of applicability of the model (Table 4.7).

Table 4.7. The statistical details of the model created using o-methoxy/ethoxyphenols.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	syn	All	1	8	0.76	0.52	0.25	0.75	0.42	0.25	0.24
HF	syn	r(C-O)	1	8	0.87	0.37	0.85	0.63	0.83	0.78	0.64
HF	syn	r(O-H)	1	8	0.82	0.44	0.82	0.44	0.77	0.67	0.52
HF	syn	All	1	7	0.76	0.26	0.14	-	-	-	-
HF	syn	r(C-O)	1	7	0.47	0.39	0.17	-	-	-	-
HF	syn	r(O-H)	1	7	0.15	0.49	-0.10	-	-	-	-
HF	anti	All	1	8	0.82	0.44	0.29	-	-	-	-
HF	anti	r(C-O)	1	8	0.88	0.36	0.86	-	-	-	-
HF	anti	r(O-H)	1	8	0.91	0.31	0.90	-	-	-	-
HF	anti	All	1	7	0.82	0.23	0.44	-	-	-	-
HF	anti	r(C-O)	1	7	0.51	0.37	0.09	-	-	-	-
HF	anti	r(O-H)	1	7	0.64	0.32	0.28	-	-	-	-
HF	syn	All	1	24	0.84	0.39	0.79	0.82	0.39	0.19	0.21
HF	**syn**	**r(C-O)**	**1**	**24**	**0.91**	**0.29**	**0.89**	**0.53**	**0.69**	**0.57**	**0.45**
HF	syn	r(O-H)	1	24	0.85	0.37	0.83	0.58	0.61	0.43	0.35

Table 4.8. The results of testing 16 methoxyphenols in the methoxy/ethoxyphenol models.

					Model Statistics			16 Compound Test Set
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP
HF	syn	All	1	8	0.76	0.52	0.25	0.43
HF	syn	r(C-O)	1	8	0.87	0.37	0.85	0.30
HF	syn	r(O-H)	1	8	0.82	0.44	0.82	0.41
HF	syn	All	1	7	0.76	0.26	0.14	0.53
HF	syn	r(C-O)	1	7	0.47	0.39	0.17	0.77
HF	syn	r(O-H)	1	7	0.15	0.49	-0.10	0.42

4.3.2.6 Miscellaneous o-Phenols

2-cyanophenol, 2-hydroxybiphenyl, 2-amino-4-nitrophenol and 2-aminophenol are the only representatives of these classes of o-phenol compounds. It is expected that these would produce separate high-correlation subsets but as there are few examples, this was not investigated.

4.3.3 Meta- and Para-Phenols

The meta/para phenol models were already of high quality using just r(C-O) with an r^2 value of 0.87 and an RMSEE of 0.41, without taking into account conformation (Table 4.2). The r(C-O) and r(O-H) were the most important to the all-bond-length model according to the VIP plot. We investigated conformations to see if it was important as seen in the case of the o-phenols. Different conformations are only possible for the asymmetrical meta- and meta-/para-phenols. Different conformations based on the direction of the acidic hydrogen were optimised and an r(C-O) model was created using all the conformations and all the compounds. The differences between the predicted pK_a values for the same compounds in the different conformations were calculated. The average difference was found to be less than 0.1 log unit. For this reason we decided that conformational differences would not be considered in the subsequent investigations for the meta- and para-phenols. After modelling the para-phenols and meta-phenols separately and finding little improvement to the models, we investigated high-correlation subsets between similar compound classes. The dataset contained 6 nitrophenols, including 3-

trifluoromethyl-4-nitrophenol and 3-nitro-4-cresol, 14 halogen phenols, including 3-trifluoromethylphenol, 4-trifluoromethylphenol, 4-chloro-3,5-dimethylphenol, 3-methyl-4-chlorophenol, 15 alkylphenols, 5 methoxy/ethoxyphenols, 2 hydroxybenzaldehydes, 2 hydroxyacetophenones, and 11 compounds we classed as miscellaneous, which included compounds such as m/p-cyanophenol, m/p-phenylphenol and m/p-aminophenol. We investigated the nitro, halogen and alkylphenols to see if treating these classes of compounds separately produced high-correlation subsets.

4.3.3.1 *m-/p-Nitrophenols*

Table 4.9. The statistical details of the model created using m-/p-nitrophenols.

					Model Statistics				**CV Statistics**		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	6	0.98	0.25	0.91	0.52	0.83	0.83	0.83
HF	N/A	r(C-O)	1	6	0.90	0.45	0.89	0.47	0.83	0.81	0.69
HF	N/A	r(O-H)	1	6	0.98	0.21	0.97	0.27	0.95	0.95	0.94

The r(O-H) model produced the highest correlation and the lowest RMSEE (Table 4.9). As there were only 6 compounds, we tested the model using 5 compounds for which pK_a values could be found in the literature. These were 3-methyl-4-nitrophenol (compound 351), 3,5-dimethyl-4-nitrophenol (compound 352) and 3-chloro-4-nitrophenol (compound 353), 3-fluoro-4-nitrophenol (compound 349) and 3,5-difluoro-4-nitrophenol (compound 350)[165-167]. The RMSEP for these compounds is shown in (Table 4.10). The results are above the 0.5 pK_a unit threshold that we aim for but it must be considered that there are no halogen-substituted compounds in the training set and that the predictions for 3-fluoro-4-nitrophenol and 3,5-difluoro-4-nitrophenol are extrapolations as there are no stronger acids in this high-correlation subset.

Table 4.10. The statistics relating to the 5 compound test set.

					Model Statistics			**5 Compound Test Set**
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**
HF	N/A	All	2	6	0.98	0.25	0.91	0.62
HF	N/A	r(C-O)	1	6	0.90	0.45	0.89	0.64
HF	N/A	r(O-H)	1	6	0.98	0.21	0.97	0.62

4.3.3.2 *m-/p-Halogen Phenols*

The all-bond-length model produced the best statistics, however, the r(C-O) model was very similar in terms of RMSEE (Table 4.11). Three compounds were tested by the models with the RMSEP shown in Table 4.12. The test compounds were 3-chloro-4-nitrophenol (compound 353), 3,5-difluoro-4-nitrophenol (compound 350) and 3-fluoro-4-nitrophenol (compound 349). Their predictions are extrapolations because no compounds in the training set are stronger acids. The predictions are poor, suggesting that the nitro group has the greatest effect and they should be predicted by the nitro model.

Table 4.11. The statistical details of the models created using m-/p-halogen phenols.

					Model Statistics		CV Statistics				
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	14	0.93	0.17	0.78	0.32	0.74	0.73	0.66
HF	N/A	r(C-O)	1	14	0.86	0.23	0.81	0.30	0.76	0.75	0.69
HF	N/A	r(O-H)	1	14	0.79	0.29	0.70	0.38	0.65	0.63	0.55

Table 4.12. Results of the 3 compound test set.

					Model Statistics			3 Compound Test Set
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP
HF	N/A	All	2	14	0.93	0.17	0.78	1.05
HF	N/A	r(C-O)	1	14	0.86	0.23	0.81	1.06
HF	N/A	r(O-H)	1	14	0.79	0.29	0.70	1.46

4.3.3.3 m-/p-Alkylphenols

Modelling of the alkyl phenols was attempted but as Table 4.13 shows, proved unsuccessful because of the small pK_a range (0.53 pK_a units) of this class. For these compounds the best prediction would come from using the mean pK_a value of this high-correlation subset (10.2) knowing that the error is approximately 0.25 pK_a units.

Table 4.13. The statistical details of the models created using m-/p-alkylphenols

					Model Statistics			CV Statistics			
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	15	0.38	0.13	0.13	0.32	0.18	0.02	-1.94
HF	N/A	C-O	1	15	0.19	0.15	0.15	0.30	0.15	0.09	-3.04
HF	N/A	O-H	1	15	0.02	0.16	-0.04	0.38	0.17	0.04	-19.52

4.3.3.4 Comparison of the models created for the high-correlation subsets of phenols to those constructed using different subsets of all the phenols.

Table 4.14 provides the statistics for different subsets of o-phenols to compare to the predictions from the high-correlation subset models constructed separately for o-nitro, o-halogen, o-alkyl, o-methoxy/ethoxy and the o-phenols capable of forming internal hydrogen bonds. We performed this analysis to prove that the predictions from the single-bond-length high-correlation subset models were better than those made by models constructed using all the o-phenols and all bond lengths. The eight outliers that were identified from the high-correlation subsets have been removed to give a fair comparison. The lowest energy conformation was used for all the compounds. The models created for all the o-phenols with the eight outliers removed (116 compounds – 8 outliers = 108 compounds) have lower RMSEEs than the models created with the outliers included (Table 4.2). Removal of the miscellaneous compounds has only a small effect on the statistics, however, the models improve slightly when the o-phenols capable of forming internal hydrogen bonds are removed. In all cases the internal statistics and CV statistics are the best for the models created using all the bond lengths compared to those created using r(C-O) and r(O-H). The CV statistics confirm that the models created using r(C-O) are better than those that created using r(O-H).

Table 4.14. The statistics relating to the models constructed for subsets of o-phenols.

					Model Statistics				**CV Statistics**		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r_{CV}^2	$r_{CV,0}^2$	r_m^2
HF	Lowest energy	All	4	108 outliers identified in previous sections removed)	0.93	0.67	0.90	0.81	0.90	0.89	0.83
HF	Lowest energy	r(C-O)	1	108	0.88	0.89	0.88	0.89	0.88	0.86	0.77
HF	Lowest energy	r(O-H)	1	108	0.53	1.76	0.52	1.76	0.52	0.14	0.20
HF	Lowest energy	All	4	104 (ibid but without miscellaneous compounds)	0.95	0.60	0.92	0.72	0.92	0.92	0.87
HF	Lowest energy	r(C-O)	1	104	0.89	0.89	0.89	0.88	0.88	0.87	0.78
HF	Lowest energy	r(O-H)	1	104	0.54	1.76	0.54	1.76	0.53	0.18	0.22
HF	Lowest energy	All	5	81 (ibid without o-phenols IHB)	0.96	0.57	0.93	0.67	0.94	0.93	0.88
HF	Lowest energy	r(C-O)	1	81	0.89	0.92	0.89	0.91	0.89	0.87	0.79
HF	Lowest energy	r(O-H)	1	81	0.62	1.68	0.62	1.68	0.61	0.41	0.34

Table 4.15. The statistics relating to the models constructed for subsets of m-/p-phenols.

					Model Statistics				**CV Statistics**		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r_{CV}^2	$r_{CV,0}^2$	r_m^2
HF	N/A	All	2	55	0.91	0.34	0.87	0.37	0.89	0.88	0.81
HF	**N/A**	**r(C-O)**	**1**	**55**	**0.87**	**0.41**	**0.85**	**0.43**	**0.85**	**0.83**	**0.72**
HF	N/A	r(O-H)	1	55	0.84	0.45	0.83	0.48	0.82	0.78	0.66
HF	N/A	All	2	35 (only nitro, halogen, alkyl compound)	0.96	0.26	0.94	0.30	0.94	0.94	0.88
HF	N/A	r(C-O)	1	35	0.95	0.30	0.94	0.31	0.94	0.94	0.89
HF	N/A	r(O-H)	1	35	0.95	0.29	0.95	0.30	0.95	0.94	0.90

Table 4.15 provides the statistics for different subsets of m/p-phenols to compare to the predictions from the high-correlation subsets. The internal model statistics are the same as those given in Table 4.2 since no outliers were indentified. Here we also provide the CV statistics for these models. The all-bond-length model has the best statistics followed by the r(C-O) and the r(O-H) models, respectively. The CV statistics confirm that models of high quality have been generated and the RMSEP is below 0.5 pK_a units for all the models. An improvement in the models is noticeable when only the m-/p-nitro, halogen and alkyl phenols are investigated as

high-correlation subsets are used to construct models. The all-bond-length, r(C-O) and r(O-H) models have virtually the same statistics.

Table 4.16. The statistics relating to the models constructed using all the phenols

					Model Statistics				**CV Statistics**		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r_{CV}^2	$r_{CV,0}^2$	r_m^2
HF		ALL 8	4	163	0.93	0.63	0.91	0.68	0.92	0.91	0.84
HF		C-O	1	163	0.89	0.80	0.88	0.80	0.88	0.87	0.78
HF		O-H	1	163	0.57	1.55	0.56	1.55	0.56	0.24	0.25

Table 4.16 provides the statistics for the models created using all the phenols without the eight o-phenol outliers identified from the high-correlation subsets. Small improvements in the internal statistics are observed compared to the models with the eight outliers included in Table 4.2. The all-bond-length models has the best CV statistics followed by r(C-O) and r(O-H), respectively.

We now compare the predictions made from the high-correlation subsets to those made by the models constructed from combinations of compounds from the high-correlation subsets and the models constructed with all the phenol compounds (171-8(outliers)=163). To compare the predictions from the high-correlation-subsets to those obtained from the different subset models of the o-phenols, m-/p-phenols and all the phenols shown in Table 4.14, Table 4.15 and Table 4.16, respectively, we use the RMSEP. From the CV of the different models we calculated the RMSEP for only the compounds that belonged to high-correlation subsets. For the o-phenols this involved five high-correlation subsets (i.e. o-nitro, o-halogen, o-alkyl, o-phenols capable of forming IHB and o-methoxy/ethoxy) and three different models (i.e. all-bond-length, r(C-O) and r(O-H)). We then calculated the mean RMSEP from all-bond-length, r(C-O) and r(O-H) models constructed using all the compounds that belonged to high-correlation subsets. This provided three average RMSEP values. This was repeated for only the compounds that formed high-correlation subsets from the models constructed from the o-phenols without the o-phenols capable of forming internal hydrogen bonds and miscellaneous compounds, all the o-phenols without the miscellaneous compounds, all the o-phenols (Table 4.14) and all the phenols (Table 4.16). The average RMSEP values obtained for the compounds belonging to high-correlation subsets are given in Table 4.17. Note that the models constructed without the o-phenols capable of forming internal hydrogen-bonds were not used to calculate an RMSEP for this high-correlation subset. Therefore the value is based on the remaining four high-correlation subset compounds. The RMSEP from the high-correlation subsets is lower in all cases and the r(C-O) models provide the lowest RMSEP compared to the all-bond-length models and the r(O-H) model. This indicated that better predictions are made by the high-correlation sets and proves the ability of a single *ab initio* bond length to predict pK_a of o-phenols.

Table 4.17. The average RMSEP for the o-phenols predicted from the relevant models.

	Compounds used to build models				
# Bonds	All phenols	All o-phenols	All o-phenols without miscellaneous o-phenols	All Ortho without miscellaneous o-phenols and o-phenols capable of forming IHB	High-correlation subsets
All	0.69	0.77	0.72	0.73	0.58
r(C-O)	0.83	0.81	0.81	0.84	0.42
r(O-H)	1.67	1.68	1.70	1.66	0.85

The same was performed for the m-/p-phenols for the compounds identified as forming high-correlation subsets. The results are shown in Table 4.18. The improvement in the RMSEP from using the predictions made by the high-correlation subsets is much less than that observed for the o-phenols. However, it is interesting to note that both the r(C-O) and r(O-H) models have lower RMSEP than the models constructed from all the bond lengths. These results suggest that either r(C-O) or r(O-H) can be used to predict the pK_a for m-/p-phenols.

Table 4.18. The average RMSEP for the m/p-phenols predicted from the relevant models.

	Compounds used to build models			
# Bonds	All phenols	All m-/p-phenols	Nitro, halogen and alkyl phenols	High-correlation subsets
All	0.41	0.36	0.32	0.34
r(C-O)	0.67	0.39	0.34	0.31
r(O-H)	1.21	0.43	0.30	0.27

4.3.4 Benzoic Acids

The pK_a of these compounds have been previously modelled using QTMS[152] where inevitably, the importance of the COOH group was identified. Reasonable bond length models were created using r(C=O), r(C-O) and r(O-H), but these models were inferior to those using QTMS descriptors. We added the bond linking the carboxylic group to the benzene ring, r(C-C), as this type of bond produced strong correlations for the phenols. The common skeleton of the benzoic acids is shown in Figure 4.7 and the constitution of the data set is shown in Table 4.1. All-bond-length models and single-bond-length models were constructed for all the compounds and the subsets of meta-/para- and ortho substituted benzoic acids to investigate which bonds correlated the strongest with pK_a (Table 4.19). For all the benzoic acids, the all-bond-length model produced the best correlation followed by r(O-H). For the o-benzoic acids, r(O-H) produced the best correlations followed by the all-bond-length-model. For the m-/p-benzoic acids the all-bond-length-model and r(O-H) models were very similar and so we focus here on the high-correlation subsets for the o-benzoic acids since little improvement was found using high-correlation subsets for the m-/p-phenols. A brief discussion of the modelling of the m-/p-benzoic acids is provided in

Section 4.3.5.4. Outliers in the benzoic acid data set have previously been detected[152], so in line with our motivation, we investigated high-correlation subsets and identified these compounds.

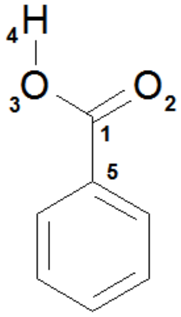


Figure 4.7. The 4 bond lengths used to predict the p*K*$_a$ of the benzoic acids. The main text refers to bond lengths 1-2, 1-3, 3-4, and 1-5 as r(C=O), r(C-O), r(O-H) and r(C-C), respectively.

Table 4.19. The results for the benzoic acids.

Subsets	# LV	# Bonds	# Compounds	r^2	q^2	RMSEE
All	2	All	94	0.83	0.72	0.40
All	1	r(O-H)	94	0.77	0.76	0.47
All	1	r(C-O)	94	0.57	0.56	0.64
All	1	r(C-C)	94	0.39	0.34	0.76
All	1	r(C=O)	94	0.15	0.07	0.90
Ortho	1	All	50	0.59	0.53	0.66
Ortho	1	r(O-H)	50	0.69	0.68	0.57
Ortho	1	r(C-O)	50	0.36	0.33	0.82
Ortho	1	r(C-C)	50	0.28	0.13	0.87
Ortho	1	r(C=O)	50	0.27	0.17	0.88
Meta/Para	2	All	44	0.79	0.74	0.22
Meta/Para	1	r(O-H)	44	0.79	0.78	0.21
Meta/Para	1	r(C-O)	44	0.72	0.67	0.25
Meta/Para	1	r(C=O)	44	0.70	0.68	0.26
Meta/Para	1	r(C-C)	44	0.49	0.45	0.34

4.3.5 Ortho-Benzoic Acids

Inspection of the observed versus predicted plot for the all-bond-length model highlighted 2,6-dihydroxybenzoic acid (compounds 252) and 2-hydroxy-3,5-dinitro-benzoic acid (compound 259) as outliers. When these compounds were removed all the models improved (Table 4.20). An explanation of these outliers is included in the o-hydroxybenzoic acid section. According to the VIP plot, r(C-C) now contributed most to the all-bond-length model and also produced the lowest RMSEE when the single-bond-length models were constructed. High-correlation subsets seen in the observed versus predicted plots were not as pronounced as for the phenols, but inspection of the structures in the data set revealed that o-halogen benzoic acids and o-hydroxybenzoic acids formed high-correlation subsets. These observations were partly due to the number of examples of these classes of compounds in the data set. There were 15 o-halogen benzoic acids and 14 o-hydroxybenzoic acids. There were only 3 examples of o-nitrobenzoic acids and a further 18 compounds that were classed as miscellaneous, including o-amine, o-alkyl and o-methoxy substituted benzoic acids. The sufficiently large number of o-halogen benzoic acids and o-hydroxybenzoic allowed us to investigate the performance of simple linear regression against PLS using all the bond lengths.

Table 4.20. The statistical details of the models created for the o-benzoic acids.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	48	0.80	0.43	0.79	-	-	-	-
HF	N/A	r(C-C)	1	48	0.69	0.53	0.69	-	-	-	-
HF	N/A	r(O-H)	1	48	0.65	0.56	0.65	-	-	-	-
HF	N/A	r(C-O)	1	48	0.60	0.60	0.60	-	-	-	-
HF	N/A	r(C=O)	1	48	0.55	0.64	0.53	-	-	-	-

4.3.5.1 o-Halogen Benzoic Acids

Considering the carboxyl group, the planar syn conformation was used due to the stability resulting from intramolecular hydrogen bonding. Depending on the nature and number of the ortho substitutions the carboxyl group can adopt different conformations. Full geometry optimisation was performed using the HF/6-31G(d) level of theory on the different possible conformations and the energies inspected. For 2-monohalogen-substituted benzoic acids, two minimum energy conformations were found. In one conformation the carboxylic O-H was closest to the halogen, while in the other the carboxylic C=O was closest. For symmetrical benzoic acids, the carboxylic groups were almost perpendicular to the aromatic ring. For asymmetric o-halogen benzoic acids, the lowest energy conformation could be either that with the carboxylic O-H being closest to the halogen or vice versa.

For seven of the 2-monohalogen-substituted benzoic acids, the carboxylic O-H closest to the halogen was the most stable. For o-bromobenzoic acid (compound 234) and 2-chloro-5-nitrobenzoic acid (compound 269) the carboxylic C=O closest to the halogen was the most stable. The anomaly with o-bromobenzoic acid is interesting as the opposite conformation is favoured by 2-chlorobenzoic acid (compound 245) and 2-fluorobenzoic acid (compound 254) and is presumably due to steric hindrance caused by the bulky bromine substitution. It should also be noted that the difference in energies between the two conformations of o-bromobenzoic acid was less than 0.15 kJmol^{-1} calculated at the HF/6-31G(d) level of theory. The presence of the 5-nitro substitution in 2-chloro-5-nitrophenol (compound 269) could be the cause of this compound favouring a different conformation. For asymmetric o-halogen benzoic acids, the carboxylic OH closest to the halogen was found to be the lower energy conformation. The influence the conformation can have on the modelling can be seen from the observed versus predicted plot in Figure 4.8. The different conformations of the same compound cause the predicted pK_a to be quite different.

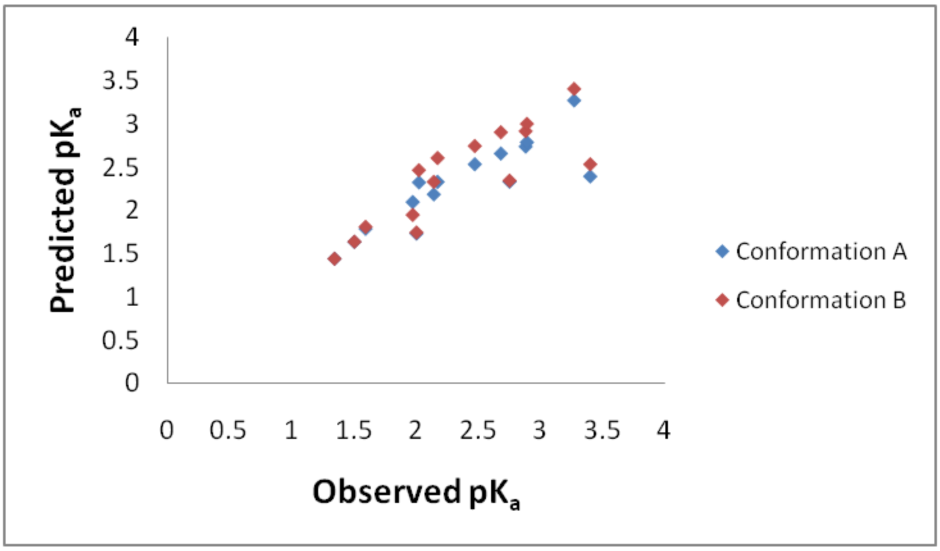


Figure 4.8. Observed versus predicted plot of the o-halogen benzoic acids from the PLS model constructed using all the bond lengths and different conformations for the same compound where applicable.

Models were constructed using the all-bond-length and single-bond-length models employing the lowest energy conformations (Table 4.21). The observed versus predicted plot was inspected and 3-amino-2,5-dichlorobenzoic acid (compound 250) was identified as an outlier, which may also exist as a zwitterion. Both forms were modelled but it remained an outlier and so was removed from all subsequent analysis. r(C-C) and r(O-H) were identified as the most important bonds from the VIP plot for the all-bond-length model and so were the only bonds considered for further analysis (Table 4.22). In the r(C-C) model, 2-chloro-5-nitrobenzoic acid (compound 269) and 2-chloro-6-methyl-benzoic acid (compound 273) had the largest errors from their experimental values. When these compounds were removed the models improved. 2-chloro-6-methyl-benzoic acid had a predicted pK_a of 2.20, which is lower than the experimental value of 2.75. It is known that any ortho substitutions increase the acidity of benzoic acids, regardless of the electronic effect of the ortho substitution on the benzoic acids. Toa and co-workers demonstrated that this effect was not captured by the molecular properties, including bond lengths, in the set of benzoic acids they investigated and led to a higher predicted pK_a for 2-methylbenzoic acids compared to experiment. However, this does not explain the lower pK_a predicted by r(C-C), which must be due to steric interference from the o-methyl substitution. 2-chloro-5-nitrobenzoic was one of the two compounds that had a lower energy with the carboxylic C=O bond being closer to the halogen than the OH. Using the bond length from this conformation appears to corrupt the correlation. Using r(C-C) from the higher energy conformation moved this compound back into the correlation. 2-Bromobenzoic acid (compound 234), which also had a different low energy conformation compared to the majority of the halogen benzoic acids, fitted the correlation when either conformation was used. For the r(O-H) model, 2-chloro-5-nitrobenzoic acid and 2-chloro-6-methyl-benzoic acid did not fit the correlation, for what appeared to be the same reasons given for the r(C-C) model. However, these compounds were not the worst predicted because 2,3,5,6-tetrafluoro-4-methyl-benzoic acid (compound 262) also had a predicted value of 1.56 compared to an experimental pK_a of 2.00. When 2-chloro-5-nitrobenzoic acid and 2-chloro-6-methyl-benzoic acid were removed, the models improved. The reason for 2,3,5,6-tetrafluoro-4-methyl-benzoic acid not fitting the correlation is unclear, but for the o-halogen phenol high-correlation subset, o-fluorophenols could not be included, which was probably the case here. When 2,3,5,6-tetrafluoro-4-methyl-benzoic acid was removed, the correlation improved further. When 2-chloro-5-nitrobenzoic acid was modelled using the higher energy conformation and the one that was consistent with the majority of the other compounds, it fitted the correlation. The same was true for 2-bromobenzoic acid with an improvement in the model observed. These results suggest that r(O-H) is more sensitive to conformation than r(C-C) and also demonstrates that inconsistent conformations can corrupt the correlations.

To attempt to solve the conformation issues with 2-bromobenzoic acid and 2-chloro-5-nitrobenzoic acid we reoptimised the halogen benzoic acids at the HF/6-31G(d,p) level of theory, that is, now adding *p* functions on the hydrogen atoms. 2-bromobenzoic acid, with the carboxylic OH closest to the bromine, became the lowest energy conformer; however, the lowest energy conformation for compound 2-chloro-5-nitrobenzoic acid did not change. The models created using the geometries calculated using additional polarisation functions were very similar to those created using HF/6-31G(d) bond lengths (data not shown).

The best models obtained from r(C-C) and r(O-H) were exposed to CV. The r(C-C) model produced the best CV statistics, which included the lowest RMSEP compared to the r(O-H) model. It is also noteworthy that this model included 2,3,5,6-tetrafluoro-4-methyl-benzoic acid (compound 262), which was not included in the r(O-H) model.

Table 4.21. The statistical details of the models created for the o-halogen benzoic acids.

					Model Statistics			CV Statistics			
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	2	15	0.72	0.35	0.68	-	-	-	-
HF	Lowest Energy	r(C-C)	1	15	0.63	0.40	0.59	-	-	-	-
HF	Lowest Energy	r(O-H)	1	15	0.61	0.41	0.57	-	-	-	-
HF	Lowest Energy	r(C-O)	1	15	0.61	0.41	0.57	-	-	-	-
HF	Lowest Energy	r(C=O)	1	15	0.05	0.64	-0.04	-	-	-	-
HF	Lowest Energy	All	2	14 (-250)	0.86	0.22	0.84	-	-	-	-
HF	Lowest Energy	r(C-C)	1	14 (-250)	0.83	0.25	0.81	-	-	-	-
HF	Lowest Energy	r(O-H)	1	14 (-250)	0.70	0.33	0.66	-	-	-	-
HF	Lowest Energy	r(C-O)	1	14 (-250)	0.68	0.34	0.62	-	-	-	-
HF	Lowest Energy	r(C=O)	1	14 (-250)	0.09	0.57	-0.01	-	-	-	-

Table 4.22. The statistical details of the models created using r(C-C) and r(O-H) for the o-halogen benzoic acids.

					Model Statistics			CV Statistics			
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	r(C-C)	1	14 (-250)	0.84	0.25	0.83	-	-	-	-
HF	Lowest Energy	r(C-C)	1	13 (-250, 273)	0.90	0.20	0.88	-	-	-	-
HF	Lowest Energy	r(C-C)	1	12 (-250, 273 and 269)	0.94	0.16	0.92	-	-	-	-
HF	Mixed	r(C-C)	1	13 (-250, 273)(269 conf a)	0.93	0.16	0.91	-	-	-	-
HF	**Mixed**	**r(C-C)**	**1**	**13 (-250, 273)(269 conf a, 234 conf a)**	**0.93**	**0.16**	**0.92**	**0.19**	**0.90**	**0.89**	**0.82**
HF	Lowest Energy	r(O-H)	1	14 (-250)	0.70	0.33	0.66	-	-	-	-
HF	Lowest Energy	r(O-H)	1	13 (-250, 273)	0.76	0.30	0.71	-	-	-	-
HF	Lowest Energy	r(O-H)	1	12 (-250, 273 and 269)	0.81	0.28	0.77	-	-	-	-
HF	Lowest Energy	r(O-H)	1	12 (-250, 273 , 269 and 262)	0.89	0.22	0.85	-	-	-	-
HF	Mixed	r(O-H)	1	12 (-250, 273 , and 262) (269 conf a)	0.88	0.22	0.85	-	-	-	-
HF	Mixed	r(O-H)	1	12 (-250, 273 , and 262) (269 and 234 conf a)	0.92	0.18	0.90	0.26	0.84	0.83	0.79

4.3.5.2 o-Hydroxybenzoic Acids

As with the o-halogen benzoic acids, the syn conformation was used for the carboxylic group. Unconstrained optimisation of the carboxylic group resulted in it being coplanar with the aromatic ring. Taking the simplest o-hydroxybenzoic acid in the data set, salicylic acid (compound 23), four stable conformations were found (Figure 4.9). Full geometry optimisation was performed on the 14 o-hydroxybenzoic acids in each of the four conformations. The most stable conformation was always the one where the hydroxyl hydrogen forms an intramolecular hydrogen bond with the C=O oxygen (Figure 4.9d). The bond lengths from these conformations were used in modelling (Table 4.23). 2,6-Dihydroxybenzoic acid (compound 252) and 2-hydroxy-3,5-dinitro-benzoic acid (compound 259) were omitted as outliers in Chapter 3 and our publication[152] on the basis of the presence of the hydroxyl group at the ortho position(s). We were therefore cautious with including these compounds, but interested to investigate whether the previous omissions were justified.

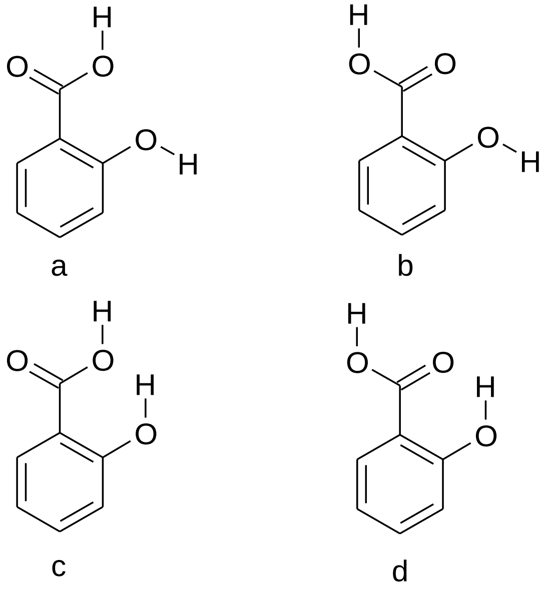


Figure 4.9. The stable conformations of salicylic acid.

Table 4.23. The statistical details of the models created for the o-hydroxybenzoic acids.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	3	14	0.97	0.16	0.91	0.74	0.73	0.67	0.55
HF	**Lowest Energy**	**r(O-H)**	**1**	**14**	**0.98**	**0.13**	**0.97**	**0.15**	**0.96**	**0.96**	**0.92**
HF	Lowest Energy	r(C=O)	1	14	0.54	0.59	0.45	-	-	-	-
HF	Lowest Energy	r(C-C)	1	14	0.15	0.80	-0.07	-	-	-	-
HF	Lowest Energy	r(C-O)	1	14	0.00	0.87	-0.10	-	-	-	-
HF	Lowest Energy	All	2	12 (-252 and 259)	0.92	0.12	0.82	-	-	-	-
HF	Lowest Energy	r(O-H)	1	12 (-252 and 259)	0.92	0.12	0.90	-	-	-	-
HF	Lowest Energy	r(C-O)	1	12 (-252 and 259)	0.87	0.15	0.79	-	-	-	-
HF	Lowest Energy	r(C=O)	1	12 (-252 and 259)	0.72	0.21	0.62	-	-	-	-
HF	Lowest Energy	r(C-C)	1	12 (-252 and 259)	0.55	0.27	0.37	-	-	-	-
HF	Lowest Energy	All	3	13 (-252)	0.97	0.13	0.87	-	-	-	-
HF	Lowest Energy	r(O-H)	1	13 (-252)	0.97	0.14	0.94	-	-	-	-
HF	Lowest Energy	r(C-O)	1	13 (-252)	0.89	0.25	0.78	-	-	-	-
HF	Lowest Energy	r(C=O)	1	13 (-252)	0.63	0.46	0.45	-	-	-	-
HF	Lowest Energy	r(C-C)	1	13 (-252)	0.49	0.54	0.30	-	-	-	-
HF	Lowest Energy	All	3	13 (-259)	0.98	0.12	0.69	-	-	-	-
HF	Lowest Energy	r(O-H)	1	13 (-259)	0.97	0.11	0.97	-	-	-	-
HF	Lowest Energy	r(C=O)	1	13 (-259)	0.42	0.51	0.38	-	-	-	-
HF	Lowest Energy	r(C-O)	1	13 (-259)	0.19	0.60	-0.1	-	-	-	-
HF	Lowest Energy	r(C-C)	1	13 (-259)	0.03	0.66	-0.1	-	-	-	-

In all the single-bond-length models involving all compounds, r(O-H) produced the best correlations compared to any other single-bond-length model or the all-bond-length model. For the all-bond-length model and the r(O-H) model, 2,6-dihydroxybenzoic acid and 2-hydroxy-3,5-dinitro-benzoic acid were found lying on the regression line. However, these two compounds did not lie on the regression line corresponding to single-bond-length models other than r(O-H). Also, when these compounds were excluded, these models did improve. Sequentially removing these compounds demonstrated that 2,6-dihydroxybenzoic acid had the greatest influence in reducing the correlation statistics. This is probably due to the fact that 2,6-dihydroxybenzoic acid is the only 2,6-substituted benzoic acid. Indeed, the r(C=O), r(C-O) and r(C-C) models suffer from the presence of the second substitution, which is not the case for the r(O-H) model. r(O-H) and r(C-O) produce good models when 2-hydroxy-3,5-dinitro-benzoic acid is included. The two highly electron-withdrawing nitro groups in the meta positions deteriorated the correlation for the r(C=O) and r(C-C) models. It is interesting to note that r(C-C) produces the worst correlations for o-hydroxybenzoic acids but the best for the o-halogen benzoic acids. We have demonstrated that r(O-H) gives good correlations with pK_a when 2,6-dihydroxybenzoic acid and 2-hydroxy-3,5-dinitro-benzoic acid are included but models obtained from bond lengths other than r(O-H) suffer from their inclusion. The current work and its conclusion can be used to explain why we had to exclude these two compounds in Chapter 3 and our publication[152] where we did not restrict ourselves to a single-bond-length model, nor concentrated on particular high-correlation subsets. In this previous publication, properties from bonds other than r(O-H) were used to model the pK_a. We have shown above that these bonds cause deterioration in the models for 2,6-dihydroxybenzoic acid and 2-hydroxy-3,5-dinitro-benzoic acid. This is confirmed by the CV statistics for the all-bond-length model.

We also constructed similar models with the different conformations. It was interesting to find that strong correlations were found but different bond lengths became important according to the VIP plot. Good models can also be constructed using different conformations and all the bond lengths in a PLS model. However, when only one bond length is used, the models drastically deteriorate.

4.3.5.3 Comparison of the models created for the high-correlation subsets of o-benzoic acids to those constructed using different subsets of all the benzoic acids.

A comparison between the RMSEP values, as performed from the o-phenols, is not worthwhile here because different bonds gave better results.

4.3.5.4 Meta-/Para-Benzoic Acids

The models and statistics for the m-/p-benzoic acids are given in Table 4.24. Inspection of the observed versus predicted plot for the all-bond-length model and all the single-bond-length models highlighted 3,4-diamino-benzoic acid as a large outlier. This compound was excluded by Tehan *et al*. and also in Chapter 3 and our publication[152] for the reason that it may be partially in a zwitterionic form in solution. After excluding this compound, the models significantly improved. The all-bond-length model and four single-bond-length models were cross-validated, constructed from the remaining 43 m-/p-benzoic acids. The results in Table 4.24 show that only r(O-H) is required to predict the pK_a for these compounds. The corresponding statistics are almost exactly the same as those of the all-bond-length model. Furthermore, the r(C-O) model demonstrated good predictive ability, which decreases for the r(C=O) model and further for the r(C-C) model.

Table 4.24. The statistical details of the models created for the m-/p-benzoic acids.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	N/A	All	2	44	0.79	0.22	0.74	-	-	-	-
HF	N/A	r(O-H)	1	44	0.79	0.21	0.78	-	-	-	-
HF	N/A	r(C-O)	1	44	0.72	0.25	0.67	-	-	-	-
HF	N/A	r(C=O)	1	44	0.70	0.26	0.68	-	-	-	-
HF	N/A	r(C-C)	1	44	0.49	0.34	0.45	-	-	-	-
HF	N/A	All	2	43	0.93	0.12	0.92	0.13	0.92	0.92	0.85
HF	**N/A**	**r(O-H)**	**1**	**43**	**0.92**	**0.13**	**0.91**	**0.13**	**0.91**	**0.91**	**0.84**
HF	N/A	r(C-O)	1	43	0.90	0.15	0.90	0.15	0.90	0.89	0.80
HF	N/A	r(C=O)	1	43	0.83	0.19	0.83	0.20	0.82	0.78	0.65
HF	N/A	r(C-C)	1	43	0.61	0.29	0.59	0.30	0.57	0.29	0.27

4.3.6 Anilines

Figure 4.10 shows the common skeleton and bonds screened to predict pK_a of the aniline compounds. Table 4.1 provides the constitution of the data set for the aniline compounds. Here we use the pK_a values associated with the dissociation of the hydrogen from the conjugated acid. Of course, the protonation of the substituted anilines are associated with pK_b values but because pK_a + pK_b = 14, all the pK_a values are related to the pK_b values. Therefore, using either pK_a or pK_b does not change the models' statistics. In contrast to phenol and benzoic acid, aniline is symmetrical and therefore most compounds have only one stable conformation. We used the conformation with the lowest energy. All-bond-length models were created for all the anilines and for o-aniline and m-/p-aniline subsets (Table 4.25). It was pleasing to note that the VIP plot for the all-bond-length model with all the anilines highlighted r(C-N) as the most important bond followed by r(N_7-H_8) and r(N_7-H_9). Models were subsequently constructed using a three-bond-length model r(CNH_2) and single-bond-length models using r(C-N), r(N_7 -H_8) and r(N_7 -H_9). The RMSEE for these models was higher than the RMSEE obtained using all the bond lengths. However, r(C-N) gave a lower RMSEE than r(CNH_2). The r(N_7 -H_8) and r(N_7 -H_9) gave similar statistics. Models were constructed for the m-/p- and o-aniline subsets using the same bonds to investigate whether the models improved. For the m-/p-anilines all the RMSEE values decreased compared to the same models built from all the compounds. The opposite occurred for the o-anilines where the RMSEE increased compared to the models built using all the anilines suggesting that the o-anilines caused the high RMSEE in these models. In both cases the r(N-H) models were very similar but were inferior to those constructed using r(C-N). We also noted that the RMSEE for the majority of the models was greater than 0.5 and the statistics for the m-/p-aniline models were especially poor compared to those obtained for the phenols and benzoic acids. Next we attempted to identify high-correlation subsets in line with our motivation.

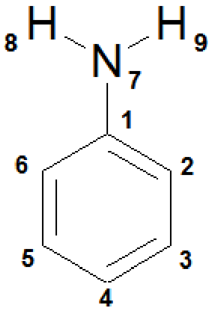


Figure 4.10. The nine bond lengths used to predict the pK_a of the aniline compounds. The main text refers to bond lengths 1-7 as r(C-N), and bond lengths 7-8 and 7-9 as r(N_7-H_8) and r(N_7-H_9), respectively. References to other bonds make use of this numbering scheme to distinguish between the C-C bonds, e.g. r(C_1-C_2). Where just the bonds associated to the NH_2 group are used, i.e. r(C-N), r(N_7-H_8) and r(N_7-H_9), then r(CNH_2) is used.

Table 4.25. The results of the aniline compounds modelled with the bond lengths calculated at the HF/6-31G(d) level of theory.

Subsets	# LV	# Bonds	# Compounds	r^2	q^2	RMSEE
All	3	All	52	0.89	0.84	0.70
All	1	r(CNH_2)	52	0.74	0.73	1.06
All	1	r(C-N)	52	0.78	0.78	0.96
All	1	r(N_7-H_8)	52	0.67	0.65	1.18
All	1	r(N_7-H_9)	52	0.66	0.63	1.21
Meta/Para	4	All	24	0.90	0.76	0.40
Meta/Para	2	r(CNH_2)	24	0.77	0.71	0.58
Meta/Para	1	r(C-N)	24	0.66	0.63	0.69
Meta/Para	1	r(N_7-H_8)	24	0.59	0.55	0.76
Meta/Para	1	r(N_7-H_9)	24	0.58	0.55	0.77
Ortho	3	All	28	0.89	0.81	0.79
Ortho	1	r(CNH_2)	28	0.69	0.64	1.29
Ortho	1	r(C-N)	28	0.76	0.73	1.14
Ortho	1	r(N_7-H_8)	28	0.60	0.54	1.46
Ortho	1	r(N_7-H_9)	28	0.59	0.49	1.49

4.3.7 Ortho-Anilines

The observed versus predicted plots and the VIP plots for all the o-aniline models in Figure 4.11 were inspected. In contrast to the all-bond-length model constructed from all the anilines, the VIP plot for the all-bond-length model constructed from only the o-anilines gave r(C_2-C_3) as the second most important bond. Pentafluoroaniline (compound 313) has an experimental pK_a of -0.28, however, the model predicted it to be 1.21. When this compound was removed the VIP plot gave r(C_2-C_3) as the fourth most important bond behind the three bonds associated with the C-NH_2 group. It is not surprising that this compound caused problems since o-fluorophenols appear to form their own high-correlation subset and 2,3,5,6-tetrafluoro-4-methyl-benzoic acid had been excluded from the benzoic acid modelling. As pentafluoroaniline was the only o-fluoroaniline in the data set, we excluded it and rebuilt the models (Table 4.26). The models for the o-anilines improved compared to those in Table 4.25. This time the correlation between r(C-N) and pK_a

(Figure 4.11) was inspected in conjunction with the structures of the compounds in an attempt to identify high-correlation subsets. Ten o-halogen anilines, five o-nitro anilines and four o-alkyl halogen allowed us to investigated high-correlation subsets for these compounds.

Table 4.26. The statistical details of the models created for the o-anilines.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	3	27	0.91	0.72	0.83	1.16	0.76	0.74	0.66
HF	Lowest Energy	r(CNH_2)	1	27	0.74	1.19	0.64	1.40	0.61	0.56	0.48
HF	Lowest Energy	r(C-N)	1	27	0.80	1.06	0.74	1.20	0.72	0.68	0.57
HF	Lowest Energy	r(N_7-H_8)	1	27	0.66	1.37	0.54	1.58	0.51	0.43	0.36
HF	Lowest Energy	r(N_7-H_9)	1	27	0.65	1.38	0.53	1.61	0.50	0.43	0.36

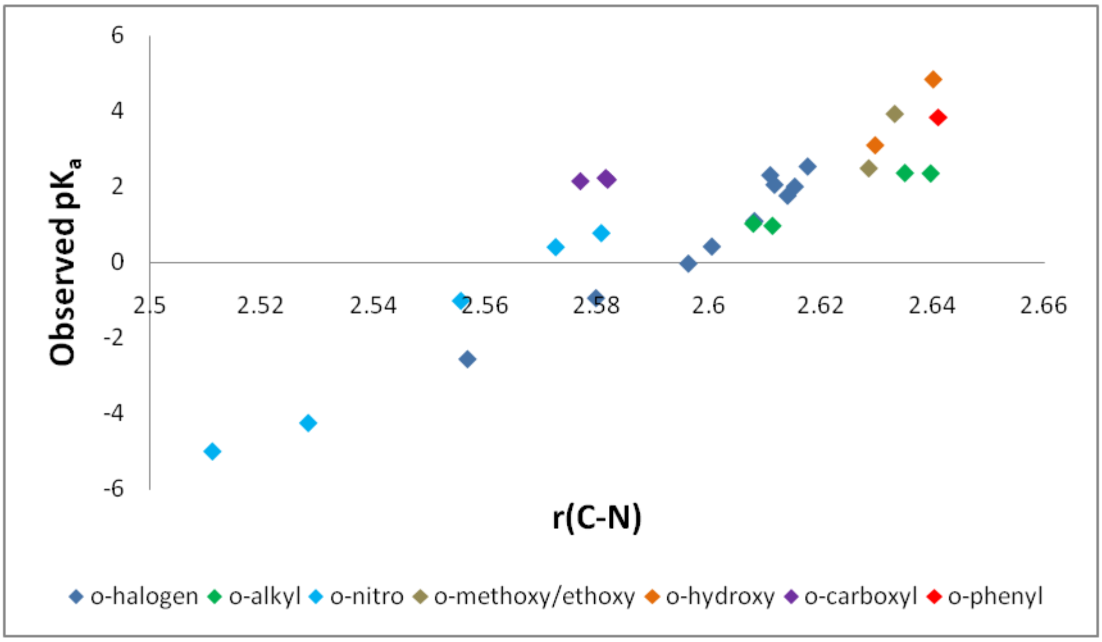


Figure 4.11. Plot of r(C-N) versus p*K*$_a$ for the o-anilines.

4.3.7.1 o-Halogen Anilines

Table 4.27 gives the statistics for the models created for the o-halogen anilines. The RMSEE and RMSEP for all the models are below 0.5 pK_a units and are much lower than those obtained from the models constructed with all the o-anilines. The r(CNH_2) model has the lowest RMSEP followed by the r(C-N) model. The statistics for all the models are very similar and their predictive ability is confirmed by high r^2_m values.

Table 4.27. The statistical details of the models created for the o-halogen anilines.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	1	10	0.95	0.38	0.94	0.43	0.92	0.92	0.87
HF	Lowest Energy	r(CNH_2)	1	10	0.97	0.32	0.96	0.42	0.93	0.93	0.90
HF	**Lowest Energy**	**r(C-N)**	**1**	**10**	**0.95**	**0.39**	**0.94**	**0.44**	**0.93**	**0.92**	**0.90**
HF	Lowest Energy	r(N_7-H_8)	1	10	0.94	0.44	0.93	0.48	0.91	0.90	0.85
HF	Lowest Energy	r(N_7-H_9)	1	10	0.94	0.42	0.94	0.44	0.92	0.92	0.87

4.3.7.2 o-Nitro Anilines

These models were constructed with only 5 o-nitro anilines, which is probably the reason for the large variations in the quality of the models (Table 4.28). r(CNH_2) gave the lowest RMSEE and RMSEP followed by r(C-N). It is interesting to note that for o-nitro anilines both the r(N-H) models were poor in contrast to the o-halogen anilines where they produced good models.

Table 4.28. The statistical details of the models created for the o-nitro anilines.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	2	5	0.99	0.37	0.92	1.45	0.66	0.65	0.61
HF	Lowest Energy	r(CNH_2)	2	5	1.00	0.14	1.00	0.34	0.98	0.98	0.98
HF	**Lowest Energy**	**r(C-N)**	**1**	**5**	**0.98**	**0.40**	**0.97**	**0.54**	**0.95**	**0.95**	**0.95**
HF	Lowest Energy	r(N_7-H_8)	1	5	0.80	1.37	0.77	1.58	0.66	0.64	0.56
HF	Lowest Energy	r(N_7-H_9)	1	5	0.01	3.06	-0.1	4.42	0.76	-0.12	0.05

4.3.7.3 o-Alkyl Anilines

The small number of examples of o-alkyl anilines was not ideal. However, good models were constructed for these compounds and using only r(C-N) provided a lower RMSEP compared to any of the other models Table 4.29. We note that the r(N_7-H_8) and r(N_7-H_9) models are very different. This can also be observed for the o-nitro anilines (Table 4.28). For the o-alkyl anilines this difference is caused by two compounds having two ortho-methyl substituents in the 2 and 6 positions, whereas the other two compounds are only mono-ortho substituted. This reason also applies to the o-nitro anilines where only one compound has nitro substituents in the 2 and 6 position. These results highlight the drawback of using the r(N-H) bond length, a problem that does not apply to the r(C-N) model. This difference is not observed for the r(N-H) o-halogen anilines model (Table 4.29), which was constructed using 10 compounds, including three di-ortho halogen substituted compounds.

Table 4.29. The statistical details of the models created for the o-alkyl anilines.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	1	4	0.99	0.08	0.77	0.25	0.93	0.78	0.57
HF	Lowest Energy	r(CNH_2)	2	4	0.96	0.26	0.86	0.95	0.54	0.52	0.47
HF	**Lowest Energy**	**r(C-N)**	**1**	**4**	**0.97**	**0.16**	**0.95**	**0.23**	**0.90**	**0.90**	**0.88**
HF	Lowest Energy	r(N_7-H_8)	1	4	0.17	0.87	-0.1	1.22	0.98	-2.75	-0.90
HF	Lowest Energy	r(N_7-H_9)	1	4	0.52	0.66	0.36	1.64	0.28	0.27	0.26

4.3.8 Meta-/Para- Anilines

The RMSEEs for the m-/p-anilines (Table 4.25) were much higher than those obtained for the phenols and benzoic acids. Similarly to the o-anilines, the VIP plot for the all-bond-length model gave a bond length, r(C_1-C_2), not part of the CNH_2 group, as the second most important bond. The VIP plots and observed versus predicted plots were inspected for the r(CNH_2) and single-bond-length models. These revealed 3,5-dinitroaniline (compound 280) as a suspicious data point. When this compound was removed, the VIP plot for the all-bond-length model returned r(C-N), r(N_7-H_8) and r(N_7-H_9) as the most important bonds in that order.

Models were constructed without 3,5-dinitroaniline and the statistics improved (Table 4.30). The RMSEE for the r(C-N) model was below 0.5 pK_a units. The observed versus predicted plot revealed that the meta-substituted anilines were predicted far worse than the para-substituted anilines. We separated the m-/p-anilines and constructed models of the separate classes (Table 4.31 and Table 4.32). The r(C-N) model for the p-anilines gave an RMSEE of 0.27. However, the r(C-N) model for the m-anilines had an RMSEE of 0.59. This was unexpected as the single-bond-length models for the p-/m-phenols and p-/m-carboxylic acids provide good correlations. Inspection of the observed versus predicted plot did not reveal any suspicious data points nor any chemically meaningful high-correlation subsets. *Ab initio* bond lengths generated at the HF/6-31G(d) level of theory have previously been used in our group to model the pK_a of a different data set of 36 m-/p-substituted anilines[57]. An r^2 value of 0.92 and a q^2 value of 0.88 were obtained using PLS. However, all the bond lengths in the common skeleton were used and the bond lengths linking the substituents to the aromatic ring were also included. Twenty m-substituted anilines were added to our data set in an attempt to improve the r(C-N) model (Table 4.33). The RMSEE for the r(C-N) model increased to 0.74 compared to 0.59 obtained previously. Once again, no obvious chemically meaningful high-correlation subsets could be found.

Table 4.30. The statistical details of the models created for the m-/p-anilines with 3,5-dinitroaniline removed.

					Model Statistics				CV Statistics		
Method	Conformation	# Bonds	# LV	# Compounds	r^2	RMSEE	q^2	RMSEP	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	3	23	0.92	0.32	0.82	0.55	0.72	0.68	0.57
HF	Lowest Energy	r(CNH_2)	2	23	0.85	0.42	0.79	0.58	0.69	0.64	0.53
HF	**Lowest Energy**	**r(C-N)**	**1**	**23**	**0.80**	**0.48**	**0.76**	**0.54**	**0.73**	**0.68**	**0.57**
HF	Lowest Energy	r(N_7-H_8)	1	23	0.74	0.54	0.70	0.60	0.66	0.59	0.48
HF	Lowest Energy	r(N_7-H_9)	1	23	0.73	0.56	0.69	0.60	0.66	0.56	0.45

Table 4.31. The statistical details of the models created for the p-anilines.

					Model Statistics				CV Statistics		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	1	11	0.90	0.38	0.87	0.50	0.80	0.79	0.70
HF	Lowest Energy	r(CNH_2)	2	11	0.97	0.24	0.93	0.37	0.90	0.89	0.81
HF	**Lowest Energy**	**r(C-N)**	**1**	**11**	**0.95**	**0.27**	**0.93**	**0.33**	**0.92**	**0.92**	**0.92**
HF	Lowest Energy	r(N_7-H_8)	1	11	0.93	0.33	0.88	0.44	0.87	0.87	0.86
HF	Lowest Energy	r(N_7-H_9)	1	11	0.93	0.33	0.90	0.40	0.89	0.89	0.89

Table 4.32. The statistical details of the models created for the m-anilines.

					Model Statistics				CV Statistics		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	4	12	0.97	0.21	0.91	0.53	0.67	0.46	0.36
HF	Lowest Energy	r(CNH_2)	2	12	0.80	0.47	0.75	0.65	0.54	0.41	0.34
HF	**Lowest Energy**	**r(C-N)**	**1**	**12**	**0.66**	**0.59**	**0.61**	**0.63**	**0.56**	**0.41**	**0.34**
HF	Lowest Energy	r(N_7-H_8)	1	12	0.58	0.65	0.52	0.70	0.45	0.21	0.23
HF	Lowest Energy	r(N_7-H_9)	1	12	0.54	0.69	0.50	0.72	0.42	0.03	0.16

Table 4.33. The statistical details of the models created for the m-anilines with the twenty additional m-anilines.

					Model Statistics				CV Statistics		
Method	**Conformation**	**# Bonds**	**# LV**	**# Compounds**	r^2	**RMSEE**	q^2	**RMSEP**	r^2_{CV}	$r^2_{CV,0}$	r^2_m
HF	Lowest Energy	All	4	32	0.81	0.49	0.57	0.86	0.35	0.04	0.16
HF	Lowest Energy	r(CNH_2)	2	32	0.72	0.56	0.67	0.61	0.63	0.52	0.42
HF	Lowest Energy	r(C-N)	1	32	0.49	0.74	0.44	0.78	0.41	-0.03	0.14
HF	Lowest Energy	r(N_7-H_8)	1	32	0.41	0.81	0.34	0.83	0.32	-0.38	0.05
HF	Lowest Energy	r(N_7-H_9)	1	32	0.37	0.83	0.34	0.85	0.30	-0.64	0.01

4.3.9 Comparison of the Correlation Obtained With and Without an Ammonia Probe

As mentioned in the Introduction, the complete series of chlorophenols, bromophenols and fluorophenols has previously been investigated separately for correlations of molecular properties with pK_a [156, 157, 159, 160]. Having demonstrated that strong correlations between pK_a and one bond length can be achieved for halogen phenols, we investigated whether better results could be achieved using an ammonia probe and a higher level of theory. For all the monomeric o-halogen phenols in the dataset, the syn conformation was the lowest in energy apart from the two 2-halogen-6-nitrophenols. It is suggested that when a probe molecule is introduced, the anti conformer (where the hydroxyl hydrogen points away from the closest halogen) is more stable[156, 157]. This can only be the case for mono-ortho-substituted halogen phenols. The ammonia, as a probe molecule, is positioned with its lone pair at the hydroxyl hydrogen of the halogen phenols, conserving the C_s symmetry if other meta-/para-substituents are ignored (Figure 4.12).

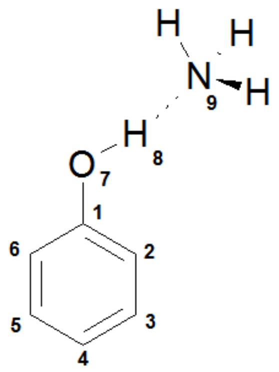


Figure 4.12. The general structure and number scheme for the phenol-ammonia complex.

Full geometry optimisations were performed at the HF/6-31G(d) and B3LYP/6-311++G(d,p) level of theory on the data set of o-halogen phenols (30 compounds), without the two 2-halogen-6-nitrophenols, in the presence of the ammonia probe. For the asymmetric halogen phenols, geometry optimisations were preformed on both the syn and anti conformers. Contrary to the calculated energies of the monomeric halogen phenols, where the syn conformation was consistently lower in energy, the same was *not* found for the halogen phenol-ammonia complexes. For the asymmetric halogen phenols, at both levels of theory, we generally found the syn conformation to be most stable. However, in some cases the lowest energy conformer was not consistent at both levels of theory for the same compound. These findings were contrasting with the work of Han *et al.*[156]. For example, we find the syn conformer of 2-chlorophenol to be most stable using both levels of theory and including the basis-set superposition error (BSSE) in the HF calculation. For this reason we constructed models based on the three possible combinations: each compound being in its lowest energy conformer, each in its syn conformer and each in its anti conformer (Table 4.34) (symmetrically substituted compounds such as 2,6-dichlorophenol can of course not be assigned anti or syn but this fact did not exclude them from the dataset). The four compounds previously identified as outliers were still outliers in the models even after the introduction of the probing ammonia. This can be seen by an improvement in all the models statistics when the outliers are removed.

Table 4.34 The statistics relating to the models constructed for the phenols and an ammonia probe.

			HF				B3LYP			
Conformation	Bonds	# Compounds	# LV	r^2	q^2	RMSEE	# LV	r^2	q^2	RMSEE
Lowest Energy										
	All	30	4	0.97	0.92	0.31	2	0.94	0.88	0.42
	C-O	30	1	0.90	0.90	0.51	1	0.91	0.90	0.50
	O-H	30	1	0.81	0.78	0.71	1	0.73	0.70	0.84
	O-H...N	30	1	0.54	0.44	1.11	1	0.56	0.49	1.09
	All	26	3	0.99	0.98	0.18	2	0.98	0.96	0.26
	C-O	26	1	0.97	0.97	0.27	1	0.95	0.95	0.36
	O-H	26	1	0.94	0.93	0.42	1	0.88	0.87	0.59
	O-H...N	26	1	0.72	0.70	0.89	1	0.77	0.76	0.81
Syn										
	All	30	4	0.97	0.92	0.30	2	0.95	0.89	0.37
	C-O	30	1	0.90	0.90	0.51	1	0.91	0.91	0.49
	O-H	30	1	0.82	0.79	0.69	1	0.75	0.72	0.82
	O-H...N	30	1	0.59	0.48	1.04	1	0.57	0.51	1.08
	All	26	2	0.98	0.96	0.25	2	0.99	0.97	0.20
	C-O	26	1	0.98	0.98	0.25	1	0.96	0.96	0.34
	O-H	26	1	0.93	0.92	0.44	1	0.90	0.89	0.54
	O-H...N	26	1	0.87	0.86	0.61	1	0.78	0.78	0.78
Anti										
	All	30	1	0.87	0.84	0.58	2	0.93	0.87	0.43
	C-O	30	1	0.91	0.91	0.49	1	0.91	0.91	0.49
	O-H	30	1	0.85	0.82	0.62	1	0.85	0.82	0.65
	O-H...N	30	1	0.46	0.38	1.21	1	0.61	0.55	1.02
	All	26	3	0.98	0.95	0.24	2	0.97	0.95	0.28
	C-O	26	1	0.97	0.97	0.29	1	0.95	0.94	0.39
	O-H	26	1	0.90	0.84	0.54	1	0.93	0.90	0.46
	O-H...N	26	1	0.50	0.30	1.18	1	0.70	0.60	0.91

These results indicate that there is no need to use the more expensive B3LYP/6-311++G(d,p) level of theory as the models generated using HF/6-31G(d) are of equal and sometimes of superior quality. The lowest RMSEEs are produced by the all-bond-length models followed by the r(C-O), r(O-H) and r(O-H…N) models, respectively. This is confirmed by the VIP plot for the all-bond-length models ranking the importance of these bond lengths to the models in the same order. In most cases the difference between the statistics for the all-bond-length models and the r(C-O) models is small, suggesting that the single r(C-O) models are suitable for predicting pK_a of halogen phenols. Considering the single-bond-length models, the syn conformation generally produces the lowest RMSEEs. This is because the influence of the o-halogen substitution is constant for each complex considered. For the lowest-energy and the anti-conformation models, the influence is not constant and therefore corrupts the correlations. For example, the anti conformation models created using r(O-H…N) have the highest RMSEE. This is caused by 2-chloro-6-methylphenol (compound 174). The presence of the methyl group causes the O-H…N bond length to be much larger than it is for the other compounds, presumably because of steric hindrance and repulsions between the hydrogens, and therefore has a predicted pK_a much higher than the other compounds when the experimental pK_a is not the highest of all halogen phenols. Han and co-workers[156, 157] found that separate correlations were required for pK_a with r(O-H…N) for di-ortho halogen phenols because of steric interference. By using the syn conformation, every compound is exposed to steric interference from the o-halogen substitution, although it appears that separate correlations may still be needed for r(O-H…N) and possibly r(O-H). Pentabromophenol (compound 142) corrupted the correlations for all the r(O-H…N) models. Han and Tao[157] excluded this compound from their equations on the basis that the full geometry optimisation of its complex with ammonia had not converged (note that our geometry optimisation of this complex did converge though). Removing this compound from the r(O-H…N) model with the syn conformations, which is the best r(O-H…N) correlation, did not improve it enough to be better than the r(C-O) model.

This investigation into using an ammonia probe demonstrates that single bond-lengths can be used to predict the pK_a of o-halogen phenols. The results obtained from HF/6-31G(d) and B3LYP/6-311++G(d,p) are comparable. The use of r(C-O) with the syn conformation produced the best statistics for the single-bond-length models and has the advantage of avoiding erratic predictions caused by non-halogen ortho-substitutions for di-orthophenols and the need for separate correlations for di-ortho-halogenated phenols. However, comparing the r(C-O) model to that obtained using the monomeric phenols where an r^2, q^2 and RMSEE of 0.97, 0.97, and 0.27, were obtained, respectively, the use of an ammonia probe is unnecessary considering the increase in time taken to perform the geometry optimisation. Large improvements are seen in the models using the r(O-H) bond length obtained from the o-halogen phenol-ammonia complex compared to those models obtained from the monomeric halogen phenols. However, the improvements are not strong enough to make the use of a probe the preferred option over the monomeric r(C-O) model.

Zhang and co-workers performed density functional calculations (B3LYP/6-311++G(d,p)) on the complete series of hydroxybenzoic acids[158]. They concluded that, for the twelve compounds that had experimental pK_a values, the use of an ammonia probe produced stronger correlation than the monomer. We have repeated the calculations with and without the probing ammonia and compared the correlations to those obtained using HF/6-31G(d) (Table 4.35). The conformation with the carboxylic OH and ammonia probe on the opposite side from the ortho substitution was used to minimise steric hindrance[161]. The ammonia probe was placed collinearly with the acidic OH bond to form a hydrogen bonded complex with the hydroxybenzoic acids (Figure 4.13).

Using a higher level of theory and an ammonia probe produce slightly better correlations than those obtained with HF/6-31G(d). The results obtained for the monomer using HF/6-31G(d) do not seem to suffer a deterioration when the probe is removed, like those obtained using B3LYP, which supports the use of the monomer and the lower level of theory.

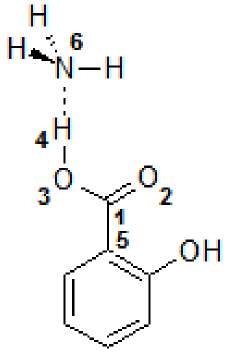


Figure 4.13. The general structure and number scheme for the hydroxybenzoic acid-ammonia complex.

Table 4.35. The statistical details of the correlations for the hydroxybenzoic acids modelled by Zhang *et al.*[158]

		B3LYP		HF	
Compounds	**Bonds**	r^2	**RMSEE**	r^2	**RMSEE**
Hydroxybenzoic acids with probe					
	r(O-H)	0.98	0.15	0.97	0.20
	r(O-H…N)	0.99	0.12	0.98	0.17
Hydroxybenzoic acids monomer					
	r(O-H)	0.85	0.44	0.96	0.23

Nine of the fourteen o-hydroxybenzoic acids from the benzoic acid data set (modelled above) were not considered by Zhang *et al.* as they had different substitutions in the meta and para positions (i.e. methyl, bromine, benzene, nitro, chlorine, amine). Our current work expands on theirs by showing that the correlations obtained including these nine compounds with hydroxybenzoic acids retains the very good statistics of the single-bond length models, as shown in Table 4.36. The results are consistent with those for the twelve acids and suggest that only small improvements are achieved by using a probing molecule compared to using the monomer in conjunction with the HF/6-31G(d) level of theory.

Table 4.36. The statistical details of the correlations for the hydroxybenzoic acids modelled by Zhang *et al.*[158] including nine o-hydroxybenzoic acids with non-hydroxyl substitutions in the meta and para positions.

		B3LYP		HF	
Compounds	**Bonds**	r^2	**RMSEE**	r^2	**RMSEE**
Hydroxybenzoic acids with probe					
	r(O-H)	0.97	0.19	0.96	0.21
	r(O-H…N)	0.98	0.16	0.97	0.19
Hydroxybenzoic acids monomer					
	r(O-H)	0.78	0.48	0.94	0.25

4.4 Discussion: Summary and Application

Data sets of phenols, benzoic acids and anilines have been deconstructed and high-correlation subsets have been identified where one *ab initio* bond length, calculated at the HF/6-31G(d) level of theory, can be used to predict pK_a. For the phenol dataset, lower RMSEEs were found when they were modelled as o-phenols and m-/p-phenols separately. The o-phenols had an RMSEE greater than 0.5 pK_a units when they were modelled together using all the bond lengths. When r(C-O) and r(O-H) were used alone to predict pK_a, the models drastically reduced in quality. Subsequent analysis of the observed versus predicted plots for these two bond lengths revealed the possibility of improving the predictions by further deconstructing the data set in to high-correlation subsets. Notably, that the high-correlation subsets, identified in the predicted versus observed correlation plots, were chemically meaningful. These high-correlation subsets, which included o-nitrophenols, o-halogenphenols, o-alkylphenols, o-phenols capable of forming internal hydrogen bonds and o-methoxy/ethoxyphenols, were fully analysed by comparing all-bond-length models to single-bond-length models. All-bond-length models differ from single-bond-length models in their capacity to highlight outliers. Outliers are readily exposed in single-bond-length models, where they cannot benefit from the fitting flexibility offered by all-bond-length models. In other words, the simplicity of the single-bond-length models calls for the obligatory investigation of a number of suspicious compounds. The majority of outliers could be explained by wrong conformations, erroneous experimental pK_a values and structural differences with the rest of the compounds in the high-correlation subset. In most cases, r(C-O) models were the best, compared to all-bond-length and r(O-H) models.

The m-/p-phenol models for r(C-O) and r(O-H), constructed using all the compounds, were comparable to the all-bond-length model, which was not the case for the o-phenols. However, because improved models were found by separating the o-phenols into high-correlation subsets, the same separation was carried out for the m-/p-phenols. Small improvements were noted but these were not comparable to the improvements seen for the high-correlation subsets of o-phenols. For all the phenols, r(C-O) consistently provided the best models.

Through analysis of the phenol data set, we proposed rules to decide in which conformation the phenols need to be optimised in order to make the best possible prediction. Secondly, these rules also determined which high-correlation subset scores the best prediction. In the cases of the phenols, these rules were decided based on the energy of each compound. We note that no compound violated these rules.

The benzoic acids were subjected to a similar analysis as the phenols. From previous work on this data set in Chapter 3, it was known that the bonds making up the carboxylic acid group produced the best correlation with pK_a and therefore we limited the analysis to just these four bonds. The m-/p-benzoic acids showed excellent statistics for single-bond-length models compared to the all-bond-length model. Hence, no high-correlation subsets were determined for these compounds. For the m-/p-benzoic acids, r(C-O) provided the highest r^2 value and the lowest RMSEE. High-correlation subsets were identified from the o-benzoic acids, for o-halogenbenzoic acids and o-hydroxybenzoic acids.

The energy of the different o-halogenbenzoic acids could not be used as a definitive guide to determine from which conformation the bond lengths should be taken. Two single-bond-length models, r(C-C) and r(O-H), showed comparable statistics to the all-bond-length model. The r(C-C) produced the lowest RMSEP when outliers had been removed. Furthermore, the low RMSEP obtained for the r(C-C) model required two compounds to be optimised in the syn conformation, which we know was not the lowest energy. However, the syn conformation was required to comply with the rest of the high-correlation subset.

The lowest energy conformers of the o-hydroxybenzoic acids were of the same conformation across the series and were used to build the models for this high-correlation subset. The four bonds investigated for single-bond-length models produced varying statistics. The r(O-H) model produced the lowest RMSEE out of all the single-bond-length models and it was also lower than the all bond-length model. By investigating this high-correlation subset with different bond lengths we were able to explain why two compounds (2,6-dihydroxybenzoic acid (compound 252) and 2-hydroxy-3,5-dinitro-benzoic acid (compound 259)) were determined as outliers in Chapter 3 and our publication. By using r(O-H) alone the pK_a of these compounds can be predicted from the remaining compounds in this high-correlation subset.

The analysis of all the bond lengths in the common skeleton of the anilines highlighted that the bonds associated with the CNH_2 group contributed the most to the all-bond-length model. As was seen for the o-phenols and o-benzoic acids, high-correlation subsets could be extracted from the o-anilines, which were chemically meaningful and produced good models using just one bond length. r(C-N) consistently produced the best models for the o-halogen, o-nitro and o-alkylaniline. The models created with r(N_7-H_8) and r(N_7-H_9) were always inferior to those created with r(C-N).

In contrast to the single-bond-length models for m-/p-phenols and benzoic acids, the RMSEE for the m-/p-aniline model using the r(C-N) was much higher than that of the all-bond-length model and nearly greater than 0.5 pK_a units. Constructing separate models for the m-anilines and p-anilines highlighted that it was the m-anilines that caused the high RMSEE. This unusual finding remained when a further twenty m-anilines where added to the data set. We were unable to identify high-correlation subsets from the models constructed with the thirty-two m-anilines and presently are unable to explain this interesting discrepancy.

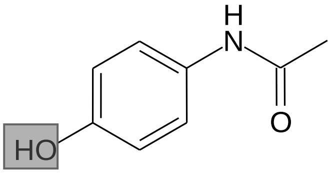
In this study our emphasis has been on accuracy rather than globality. The results demonstrate that one bond length from the group of compounds of interest can be used to predict pK_a. In order for this to work however, high-correlation subsets need to be identified and treated separately. Generally, properties of ortho-substituted compounds are notoriously more difficult to predict than m-/p-substituted compounds. This has been demonstrated by the vast improvements in the statistics of the ortho models when the high-correlation subsets were identified. While high-correlation subsets were identified for the m-/p-phenols, the improvements in modelling these separately were minor.

Some may highlight certain drawbacks in attempting to model pK_a with a single bond-length. These include the need for many highly localised models that require experimental pK_a values to construct models and may not be applicable to a compound for which a pK_a prediction is needed. Furthermore, the need to consider different conformations may be regarded an unnecessary hindrance. However, we argue that these are advantages of using a single bond-length. We have shown that the predictions from high-correlation subsets are more accurate, modelled with a single bond-length, compared to using multiple bond lengths and combining more diverse compounds. The use of high-correlation subsets and a single bond-length revealed compounds with wrongly assigned experimental pK_a values that were not clearly obvious from all-bond-length and all compound models. In the majority of cases the lowest energy conformer for each compound was the same as that of the other compounds and so was used as a guide to determine which conformer should be used. Where this was not the case, e.g. for two of the halogenbenzoic acids, the lowest energy conformation of the other compounds in the high-correlation subset was used to correct the conformation from which the bond lengths were calculated and restore consistency for all compounds. The energy can also be used to determine which high-correlation subset a compound belongs to or should be predicted from. We have even demonstrated the ability of the single bond- length models to extrapolate outside the pK_a range of the compounds used to construct it.

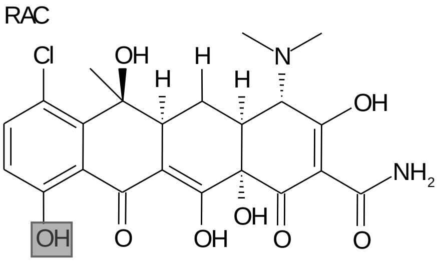
The division of the data sets in order to find high-correlation subsets was not a trivial task. However, the resulting procedure offers a practical and simple approach to predict pK_a and can be applied to more complex structures. Drug discovery programmes generate numerous, similar structures when a SAR is being explored. If pK_a is important to the activity and the compounds form a high-correlation subset then this procedure can be used. The most time consuming step would be the determination of the high-correlation subsets and identification of the bond length that correlates the strongest with pK_a. Producing a pK_a prediction tool that contains hundreds of equations and programming it to know which equation is the right one to use is straightforward and computationally undemanding.

The use of the HF/6-31G(d) level of theory stems from recent work[69, 152] where using higher levels of theory did not improve the models created. We confirmed this by comparing the correlations obtained using B3LYP/6-311+G(2d,p) for some of the high-correlation subsets and the statistics did not improve. The use of an ammonia probe also did not improve the results either. The most time consuming step in predicting the pK_a of a new compound is the geometry optimisation. We compared the time for an example compound (benzoic acid), at HF/6-31G(d) and the higher level B3LYP/6-311++G(d,p) using the ammonia probe. For the monomer the optimisation took 7 mins using HF/6-31G(d) and 1 hour 23 mins using B3LYP/6-311++G(d,p). The calculation time increased to 21 mins and 1 hour 29 mins , when the ammonia probe was included, for HF/6-31G(d) and B3LYP/6-311++G(d,p), respectively. We have demonstrated that the prediction of pK_a using a single bond-length can be applied beyond, for example chlorophenols, bromophenols and fluorophenols. Furthermore, a single bond-length can be used to predict poly-substituted compounds with different substituents provided a suitable high-correlation subset model has been constructed.

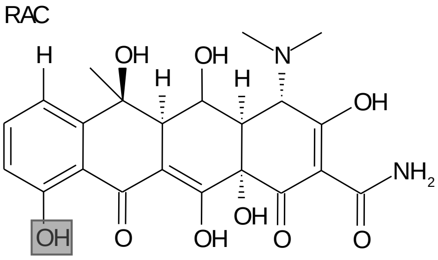
To test the use of single bond lengths to predict p$K_{a,}$ we chose 24 compounds from a data set of drug molecules[127], that were part of a 197 compound set used to compare nine pK_a prediction packages (Figure 4.14). It contains 20 phenols and 4 anilines for which micro pK_a values had been measured. There were also 7 carboxylic acids in the original data set but we had already considered four of them and we had no suitable subset models from which we could predict the remaining three. The structures of the 24 molecules were taken from the supplementary information provided by Liao and Nicklaus and optimised at the HF/6-31G* level of theory. In line with our results, we optimised the ortho-substituted molecules in the conformation determined by our models and arbitrary conformations for meta-/para-substituted molecules. We only considered the neutral forms of the drugs. The relevant bond lengths were extracted and subjected to relevant bond length models to predict the pK_a.



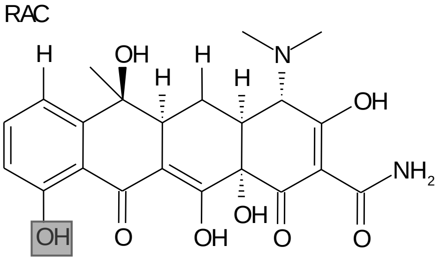
385 Actetaminophen



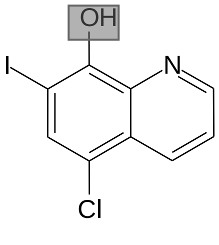
386 Chlorotetracycline



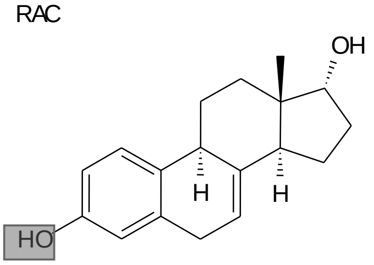
387 Oxytetracycline



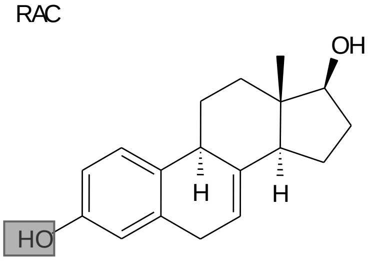
388 Tetracycline



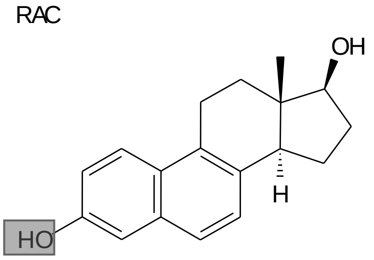
389 Clioquinol



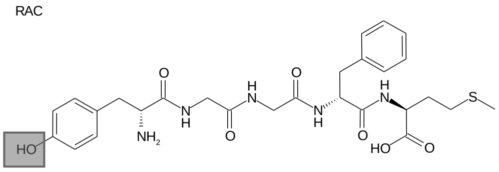
390 17α-Dihydroequilin



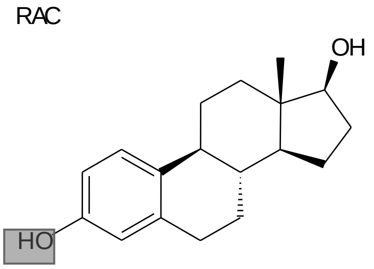
391 17β-Dihydroequilin



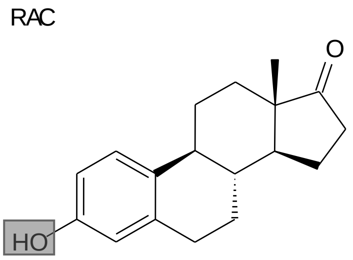
392 17β-Dihydroequilenin



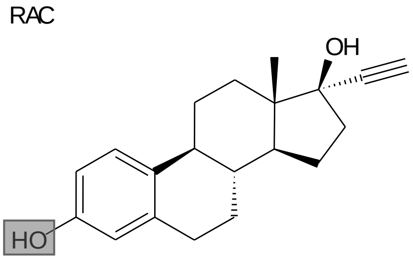
393 Enkephalin



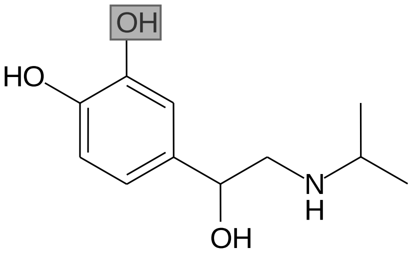
394 17β-Estradiol



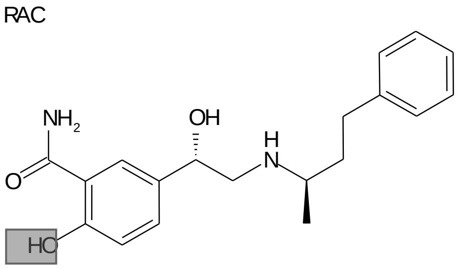
395 Estrone



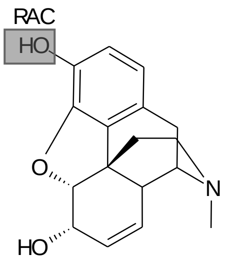
396 Ethinyl estradiol



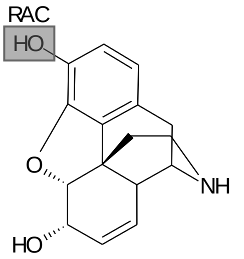
397 Isoproterenol



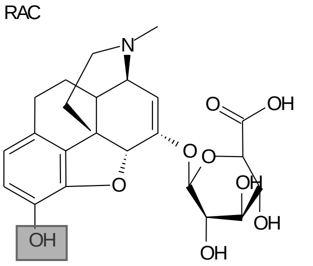
398 Labetalol



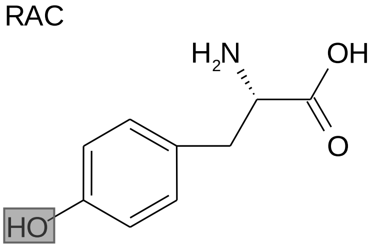
399 Morphine



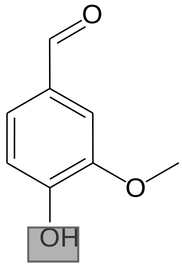
400 Normorphine



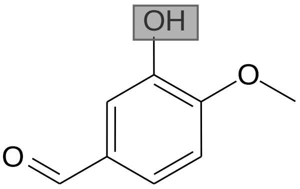
401 Morphine-6-glucuronide



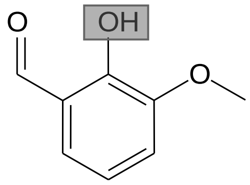
402 Tyrosine



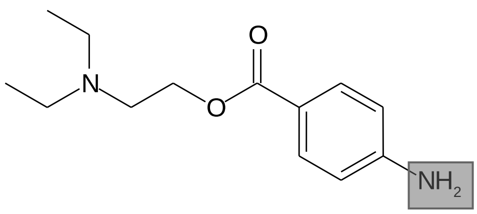
403 Vanillin



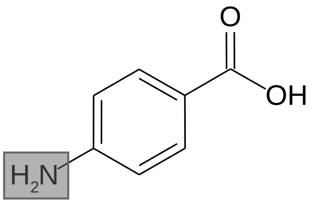
404 Iso-Vanillin



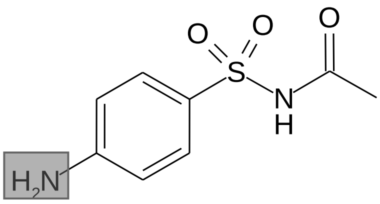
405 ortho-Vanillin



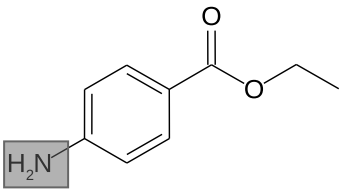
408 Procaine



406 4-Aminobenzoic acid



409 Sulfacetamide



407 Benzocaine

Figure 4.14. The structures ID and name, taken from Liao and Nicklaus[127], of the drug molecules we predicted the p*K*$_a$ for. The micro p*K*$_a$ of interest is highlighted in grey.

Table 4.37 provides the experimental p*K*$_a$ of the 24 molecules and Table 4.38 gives the predicted p*K*$_a$ values from the nine p*K*$_a$ prediction programmes compared by Liao and Nicklaus[127]. Some of the molecules were already present in the data set that we used to create single bond-length models. These are highlighted in Table 4.37 with the experimental p*K*$_a$ used by Tehan *et al.*[96, 97] included. We also state which compound class the molecules belong too (i.e. either phenol or aniline), as well as from which high correlation-subset the prediction was made and which bond length was used. Where the molecules were already in our models, we removed them in turn and reconstructed the single bond-length model to make the prediction. We also used the experimental p*K*$_a$ provided by Liao and Nicklaus[127].

We will first discuss the predictions for the phenols. Acetaminophen (compound 385) was predicted to be 9.01 compared to the experimental p*K*$_a$ of 9.63. An experimental value of 9.38 was provided by Tehan *et al.*, which is closer to the prediction made by the model constructed from the meta-/para-phenol high-correlation subset. The predictions for the tetracyclic compounds 386, 387 and 388, which all have similar structures, were very poor. We used the ortho-phenol model constructed from the ortho-phenols capable of forming internal hydrogen bonds. The errors indicate that this was not a suitable model and a new model would be needed to make predictions for these compounds. Compounds 390 and 391 only differed by the stereochemistry of a distant hydroxyl group separated by many bonds from the phenolic group of interest for the pK_a predictions. One would expect such a distant change to have little effect on the pK_a, which is why it is not surprising that all the pK_a programmes (apart from Jaguar, see Table 4.38) predicted the pK_a of these two drugs to be identical. However, the experimental pK_a for these compounds curiously turned out to be 10.29 and 9.77, respectively. Such an unexpectedly large pK_a difference of 0.52 pK_a units is difficult to accept and therefore needed resolving. Our original data source of Liao and Nicklaus's work used the experimental pK_a values published in a book[78] in 2007. A check of this source confirmed that Liao and Nicklaus correctly adopted the pK_a values and structures from this book. We therefore checked the primary reference[168] where the experiments to determine the pK_a values of these two compounds were performed. The pK_a of compound 390 was the same (i.e. 10.29) as that quoted in Liao and Nicklaus's work. The pK_a for compound 391 did not relate to that compound and actually was that of 17β-dihydroequilenin[168]. The structure of 17β-dihydroequilenin is shown in Figure 4.14 and labelled 392. We did predict a small difference for compounds 390 and 391, but when the correct structure was used the prediction was much better. We used the correct structure and predicted the pK_a from three (ACD, Marvin and SPARC) of the nine programmes used in the pK_a comparisons. All three programmes made an accurate prediction when the correct structure was used. The absolute error (AE) (calculated as $\left|y_{\mathrm{obs,i}} - y_{\mathrm{calc,i}}\right|$) for compound 392 was 0.35 compared to AEs for the nine programmes that ranged from 0.05 to 1.36. The MAD of our predictions for steroids 393, 394 and 395 was 0.37. We also correctly predicted the order of the pK_a values, which only Epik, Jaguar and SPARC achieved as well. It is interesting to note that Marvin, Pallas and Pipeline Pilot predicted the same pK_a values for all three compounds suggesting they do not detect the different structures.

Compound 397 had an AE of 0.74. We predicted it from the model constructed from the ortho-phenols capable of forming internal hydrogen bonds. The structure was very different from those used to construct this model, which probably explains the poor predictions. The opiates 398, 399, 400 were predicted from the model constructed from ortho-methoxy substituted phenols, although they do not strictly have o-methoxy substitutions. Considering this, the predictions for compounds 398 (AE = 0.58) and 399 (AE = 0.62) were reasonable, but the prediction for compound 400 (AE = 1.96) was poor. The predictions from seven (no prediction was provided by Jaguar or SPARC) of the pK_a programmes were mixed, with AEs between the range of 0.07-1.83. Compound 401 had a very accurate prediction with an AE of 0.07. We did not make a prediction for compound 404 as we had previously excluded it as an outlier. The prediction for compounds

402 (AE = 0.06) was very good and only reasonable for compound 403 (0.61). The AE of these three compounds for the predictions from the nine programmes ranged from 0 to 1.51.

We now focus on the four aniline compounds. Two of the four substituted anilines were already in our original data set. They were all predicted from the model constructed from the para-anilines. Table 4.39 gives the MAE for our predictions and for the nine programmes. For all four anilines, the MAE ranged from 0.11 for SPARC to 1.21 for Pipeline Pilot. The prediction from the r(C-N) model was 0.29, which ranks it the fifth lowest MAE prediction in Table 4.39. An MAE of 0.26 was obtained for the two anilines not present in our original data set. This MAE was ranked third lowest. The MAE ranged from 0.1 for ADME Boxes and SPARC to 1.79 for Pipeline Pilot. These results are encouraging considering the small number of para-substituted anilines and the use of only r(C-N).

Table 4.37. The experimental pK_a values for the 24 drug molecules. The model used to make the prediction from is also highlighted.

Compound Class	ID	Exp. pK_a Liao	Exp pK_a Tehan	High-Correlation subset	Bond length used	Predicted pK_a	Absolute Error
Phenols							
	385	9.63	9.38	Meta/Para	r(C-O)	9.01	0.62
	386	9.30	-	Ortho-IHB	r(C-O)	5.06	4.24
	387	9.11	-	Ortho-IHB	r(C-O)	5.76	3.35
	388	9.69	-	Ortho-IHB	r(C-O)	6.06	3.63
	389	8.16	-	Iodine containing molecules were not considered in this work			
	390	10.29	-	Meta/Para	r(C-O)	10.15	0.14
	391	9.77	-	Meta/Para	r(C-O)	10.11	0.34
	392	9.77	-	Meta/Para	r(C-O)	10.00	0.23
	393	9.89	-	Meta/Para	r(C-O)	9.54	0.35
	394	10.71	-	Meta/Para	r(C-O)	9.95	0.76
	395	10.34	-	Meta/Para	r(C-O)	10.06	0.28
	396	10.40	-	Meta/Para	r(C-O)	10.33	0.07
	397	10.07	-	No suitable high-correlation subset -see structure			
	398	7.41	-	Ortho-IHB	r(C-O)	6.67	0.74
	399	9.40	-	Methoxy	r(C-O)	9.98	0.58
	400	9.80	-	Methoxy	r(C-O)	10.42	0.62
	401	9.36	-	Methoxy	r(C-O)	11.32	1.96
	402	10.27	-	Meta/Para	r(C-O)	10.14	0.13
	403	7.40	7.40	Methoxy	r(C-O)	7.46	0.06
	404	8.90	8.89	Methoxy	r(C-O)	9.50	0.61
	405	7.91	7.91	Identified as outlier previously			
Anilines							
	406	2.50	2.38	Para	r(C-N)	2.02	0.48
	407	2.52	2.51	Para	r(C-N)	2.33	0.19
	408	2.29	-	Para	r(C-N)	2.48	0.19
	409	1.76	-	Para	r(C-N)	1.46	0.30

Table 4.38. The predicted pK_a for the 24 drug molecules made by nine programmes. The values were taken from the supplementary material provided by Lioa and Nicklaus[127].

ID	Liao p*Ka*	ACD[121]		ADME Boxes[114]		ADMET Predictor[116]		Epik[126]		Jaguar[128]		Marvin[119]		Pallas[122]		Pipeline Pilot[117]		SPARC[118]	
		Pred	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**	**Pred**	**AE**
385	9.63	9.86	0.23	9.8	0.2	10.02	0.39	10.12	0.49	10.0	0.4	9.46	0.17	9.48	0.15	9.50	0.13	9.49	0.14
386	9.3	8.03	1.27	11.4	2.1	8.53	0.77	7.55	1.75	a		10.97	1.67	7.56	1.74	7.33	1.97	b	
387	9.11	8.48	0.63	12.5	3.4	8.87	0.24	7.43	1.68	a		11.43	2.32	8.04	1.07	7.79	1.32	b	
388	9.69	8.50	1.19	12.0	2.3	8.90	0.79	8.49	1.20	a		11.44	1.75	8.06	1.63	7.81	1.88	b	
389	8.16	7.23	0.93	7.9	0.3	7.91	0.25	6.15	2.01	8.3	0.1	7.34	0.82	3.83	4.33	7.90	0.26	7.80	0.36
390	10.29	10.15	0.14	10.5	0.2	10.24	0.05	11.01	0.72	10.8	0.5	9.41	0.88	10.28	0.01	11.12	0.83	10.40	0.11
391	9.77	10.15	0.38	10.5	0.7	10.24	0.47	11.01	1.24	10.7	0.9	9.41	0.36	10.28	0.51	11.12	1.35	10.40	0.63
392	9.77	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
390*		10.08										9.41						10.41	
391*		10.08										9.41						10.41	
392*		9.78										9.79						9.91	
393	9.89	9.97	0.08	9.8	0.1	9.94	0.05	10.15	0.26	a		9.51	0.38	9.71	0.18	8.53	1.36	10.00	0.11
394	10.71	10.27	0.44	10.5	0.2	10.41	0.30	11.42	0.71	10.9	0.2	10.33	0.38	10.37	0.34	13.18	2.47	10.54	0.17
395	10.34	10.25	0.09	10.4	0.1	10.30	0.04	11.19	0.85	10.7	0.4	10.33	0.01	10.37	0.03	13.18	2.84	10.48	0.14
396	10.4	10.24	0.16	10.4	0.0	10.41	0.01	11.41	1.01	10.8	0.4	10.33	0.07	10.37	0.03	13.18	2.78	10.51	0.11
397	10.07	9.60	0.47	9.5	0.6	9.81	0.26	9.75	0.32	a		9.81	0.26	9.38	0.69	10.01	0.06	10.49	0.42
398	7.41	8.21	0.80	7.7	0.3	9.28	1.87	8.20	0.79	a		8.05	0.64	8.04	0.63	8.54	1.13	9.98	2.57
399	9.4	9.48	0.08	9.4	0.0	9.71	0.31	11.22	1.82	a		10.26	0.86	10.38	0.98	8.16	1.24	10.34	0.94
400	9.8	9.54	0.26	9.7	0.1	8.54	1.26	9.89	0.09	a		9.77	0.03	9.22	0.58	9.67	0.13	10.96	1.16
401	9.36	9.43	0.07	9.6	0.2	9.70	0.34	11.19	1.83	a		10.25	0.89	10.43	1.07	8.16	1.20	b	
402	10.27	10.01	0.26	10.2	0.1	10.02	0.25	10.38	0.12	a		9.79	0.48	9.43	0.84	10.47	0.21	10.47	0.21
403	7.396	7.78	0.38	7.6	0.2	8.91	1.51	7.62	0.22	7.6	0.2	7.81	0.41	7.59	0.19	7.25	0.15	7.79	0.39
404	8.889	9.25	0.36	8.7	0.2	9.53	0.64	9.12	0.23	9.2	0.3	9.39	0.50	8.98	0.09	8.89	0.00	9.18	0.29
405	7.912	8.18	0.27	7.9	0.0	7.98	0.07	8.25	0.34	7.7	0.2	8.74	0.83	8.17	0.26	7.91	0.00	7.58	0.33
406	2.5	2.51	0.01	2.1	0.4	2.53	0.03	2.84	0.34	a		2.69	0.19	2.04	0.46	2.19	0.31	2.36	0.14
407	2.515	2.51	0.01	2.6	0.1	2.67	0.16	2.03	0.49	2.7	0.2	2.78	0.27	2.51	0.01	3.46	0.95	2.39	0.13
408	2.29	2.12	0.17	2.4	0.1	2.63	0.34	1.94	0.35	1.1	1.2	2.70	0.41	3.23	0.94	3.46	1.17	2.22	0.07
409	1.76	0.93	0.83	1.9	0.1	1.81	0.05	1.51	0.25	1.8	0.0	2.14	0.38	1.87	0.11	4.16	2.40	1.64	0.12

a Site was not calculated because of complexity.

b Programme failed to predict value.

Table 4.39. MAE for all four aniline molecules and for the two anilines not in our original dataset.

Method	**MAE (4 anilines)**	**MAE (2 anilines)**
r(C-N)	0.29	0.26
ACD	0.25	0.50
ADME Boxes	0.2	0.1
ADMET Predictor	0.14	0.20
Epik	0.36	0.30
Jaguar	0.5	0.6
Marvin	0.31	0.40
Pallas	0.38	0.53
PP	1.21	1.79
SPARC	0.11	0.10

In Table 4.40 we present the single bond-length equations that can be applied to the prediction of pK_a for suitable compounds. We encourage an extension of this list of equations to cover more chemical space using the information and protocols we have devised.

Table 4.40. The high-correlation subsets we investigated with their associated equation to predict pK_{a}. The number of compounds used to construct the equation and relevant statistics are also provided.

Compound Class	High-Correlation-Subset	Equation	# compounds	r^2	RMSEE	q^2	RMSEP	r_{CV}^2	$r^2_{CV,0}$	r^2_m
Phenols	o-nitro phenols	pK_{a}= 137.575r(C-O) – 337.575	22	0.94	0.48	0.94	0.50	0.93	0.93	0.87
	o-halogen phenols	pK_{a} = 147.411r(C-O) – 365.540	26	0.97	0.27	0.97	0.29	0.97	0.97	0.94
	o-alkyl phenols	pK_{a} = 162.106r(C-O) – 405.207	25	0.91	0.34	0.9	0.37	0.89	0.87	0.78
	o-phenols-IHB	pK_{a} = 160.912r(C-O) – 397.578	23	0.95	0.32	0.95	0.33	0.94	0.94	0.92
	o-methoxy/ethoxyphenols	pK_{a} = 128.767r(C-O) – 318.646	24	0.91	0.29	0.89	0.53	0.69	0.57	0.45
	m-/p-phenols	pK_{a} = 122.985r(C-O) – 304.553	55	0.87	0.41	0.85	0.43	0.85	0.83	0.72
Benzoic Acids	o-halogen benzoic acids	pK_{a} = -53.316r(C-C) + 153.475	13	0.93	0.16	0.92	0.19	0.90	0.89	0.82
	o-hydroxy benzoic acids	pK_{a} = -975.258r(O-H) + 1758.265	14	0.98	0.13	0.97	0.15	0.96	0.96	0.92
	m-/p-benzoic acids	pK_{a} = -770.717r(O-H) +1390.747	43	0.92	0.13	0.91	0.13	0.91	0.91	0.84
Anilines	o-halogen anilines	pK_{a} = 83.401r(C-N) – 216.089	10	0.95	0.39	0.94	0.44	0.93	0.92	0.90
	o-nitro anilines	pK_{a} = 89.774r(C-N) – 230.712	5	0.98	0.40	0.97	0.54	0.95	0.95	0.95
	o-alkyl anilines	pK_{a} = 47.563r(C-N) – 123.095	4	0.97	0.16	0.95	0.23	0.90	0.90	0.88
	m-/p-anilines	pK_{a} = 67.932r(C-N) – 175.213	23	0.80	0.48	0.76	0.54	0.73	0.68	0.57
	p-anilines	pK_{a} = 80.644r(C-N) – 208.635	11	0.95	0.27	0.93	0.33	0.92	0.92	0.92
	m-aniline	pK_{a} = 56.090r(C-N) – 144.080	12	0.66	0.59	0.61	0.63	0.56	0.41	0.34

4.5 Summary

We have investigated the prediction of pK_a for phenols, carboxylic acids and anilines. We aimed to construct models that were able to predict pK_a within 0.5 pK_a units using a single bond length from a monomer geometry optimised by an affordable and sufficiently reliable *ab initio* method, which was determined to be HF/6-31G(d). We achieved this by grouping molecules into high-correlation subsets, which were visually identified from observed-versus-predicted plots. It was pleasing to note that the structures in each subset contained a common substitution pattern, e.g. an OH group adjacent to a NO_2 group. We have shown that improvements in model statistics are small for high-correlation subsets of meta-/para-substituted compounds compared to one model containing all these compounds. However, for ortho-substituted compounds, the statistics of high-correlation subsets improve much compared to a single model for all ortho compounds.

In the majority of cases, the models constructed from a single bond length were superior or, at the very least, similar to the models constructed using all the bond lengths. In each all-bond-length model, the most important bonds (i.e. those with the highest VIP value) were associated with the functional group where the deprotonation occurs. However, the most important bond differed between high-correlation subsets. For example, for the o-halogen benzoic acids, r(C-C) produced the best statistics, but for the o-hydroxybenzoic acids, r(O-H) was the best. The use of an ammonia probe or a higher level of theory for the o-halogen phenols and the hydroxybenzoic acids provided no advantage over the use of single bond lengths generated for the monomer at HF/6-31G(d). The constructed models were used to predict the pK_a for a set of 24 drug molecules. The predictions were accurate (MAE 0.06 to 0.62) for all the molecules, apart from those where a reason for the poor prediction was identified. For example, one reason could be that a suitable high-correlation subset model was not available or that the experimental pK_a turned out to be wrong. We have listed fifteen single-bond-length equations from which the pK_a of relevant compounds can be predicted. We encourage an extension of this list of equations, which can be constructed in a relatively small amount of CPU time on a standard PC.

Chapter 5
Prediction of Ames Mutagenicity of Carbocyclic and Heterocyclic Primary Aromatic Amines Using *ab initio* Charge Densities

5.1 Introduction

5.1.1 Drug Toxicity and Attrition Rates

Toxicology is the study of adverse effects of chemicals on living organisms. Historically, toxicology formed the basis of therapeutic and experimental medicine. Today, it continues to develop and expand by assimilating knowledge and techniques from branches of biology, chemistry, mathematics and physics.

Safety issues relating to drug toxicity occur throughout the drug discovery process. Since the number of drugs discovered in pharmaceutical research has fallen in the last decade, it has become crucial to look at the reasons for failures[169].

There are a number of reasons why toxicity has become an important issue. Two decades ago, a major problem for the pharmaceutical industry was unpredictable metabolism and pharmacokinetics in humans[170]. Figure 5.1 shows that increased knowledge about the basic and practical aspects of human metabolism and distribution have helped a great deal. Far fewer drugs fail in development due to the iterative process of chemical synthesis, target screening, and *in vitro* metabolism studies.

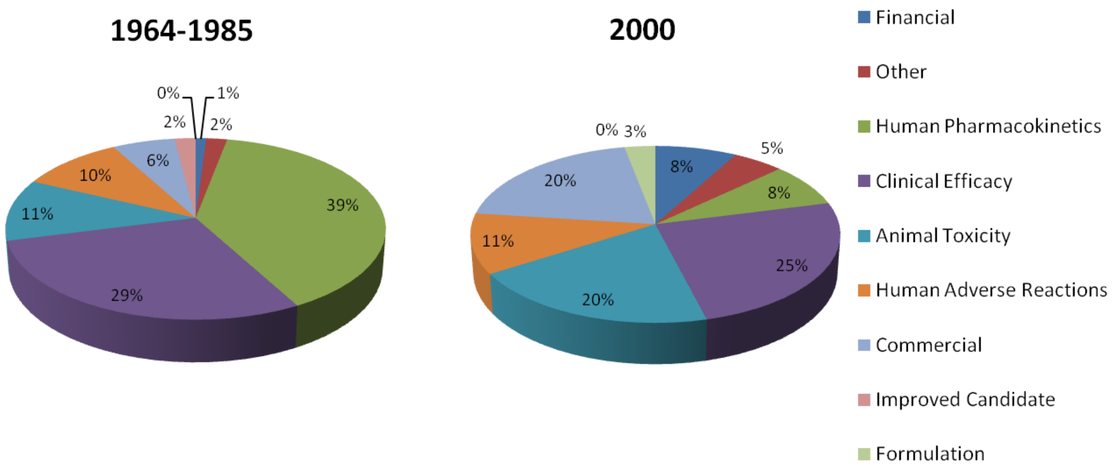


Figure 5.1. Reasons for the termination of drug candidates in development between 1964-1985[170] and in 2000[171].

Less progress has been made in accurately predicting human toxicity problems with drugs and the challenge remains considerable. Human adverse drug reactions (ADRs) are one of the most common causes of pharmaceutical product recall. An estimated 100,000 deaths per year are attributed to ADRs, making it the sixth leading cause of death in the United States[172].

5.1.2 Genetic Toxicology and Primary Aromatic Amines

Genetic toxicology is the study of the ability of chemicals to cause heritable or somatic genetic defects in humans. Genotoxicity encompasses DNA damage (chemical modification), mutagenicity (point mutation), clastogenicity (chromosome breakage) and aneugenicity (chromosome loss) caused by a chemical or its metabolites.

Carbocyclic and heterocyclic primary aromatic amines (AAs) are one of the most important classes of industrial compounds as they are widely used in the cosmetic, dye, pesticide, petrochemical and pharmaceutical industries. They are also known to be present in tobacco smoke and cooked meats. The chemical reactivity of AAs is an asset to drug synthesis but a bane for biological systems. While this compound class is ubiquitous in drug molecules (e.g. Ropivicaine, Lidocaine, Piroxicam, Lornoxicam, Tenooxicam, Atorvastatin, Leflunomide, Sorafenib and Acomplia), their genotoxic tendencies are perceived as a serious risk. Cleavage of the appropriate amide bond in these compounds means these drugs have the potential to release aromatic amines *in vivo.*

Cytochromes (CYPs) P450s are a family of haem-containing proteins that catalyse the metabolism of a broad range of molecules[173, 174], including the majority of drugs[175]. Fifty-seven P450s have been identified and they account for approximately three-quarter of the enzymes involved in drug metabolism. Of the fifty-seven human P450s, five are involved in ~95% of these reactions (Figure 5.2). They are implicated in toxicological events because they have the ability to metabolise molecules that present no risk to human health into compounds that are toxic. CYP P450 1A2 (CYP1A2) accounts for approximately 13% of hepatic P450s[176] and is implicated in the metabolism of drug molecules such as paracetamol and caffeine[177], as well as AAs.

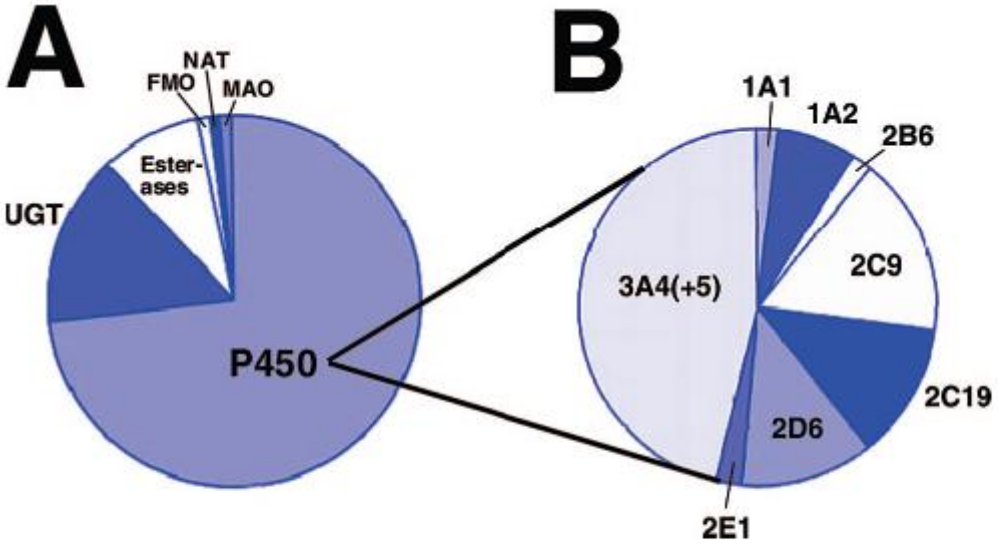


Figure 5.2. Contributions of enzymes to the metabolism of marketed drugs. The results are from a study of Pfizer drugs, and similar percentages have been reported by other pharmaceutical companies. (A) Fraction of reactions on drugs catalyzed by various human enzymes. (B) Fractions of P450 oxidations on drugs catalysed by individual P450 enzymes[173].

Figure 5.3 shows the widely accepted mechanism for the conversion of AAs to DNA-reactive metabolites. The initial step in the activation of AAs is enzymatic N-oxidation by CYP1A2 to yield an N-hydroxylamine product. The N-hydroxyl species can be further converted to highly reactive N-acetoxy or N-sulphate esters that permit a more facile heterolysis of the N-O bond or undergo direct N-O bond cleavage. This process generates highly reactive electrophilic nitrenium ions, which can bind to DNA to form adducts, resulting in genetic damage, mutations and, ultimately carcinogenesis.

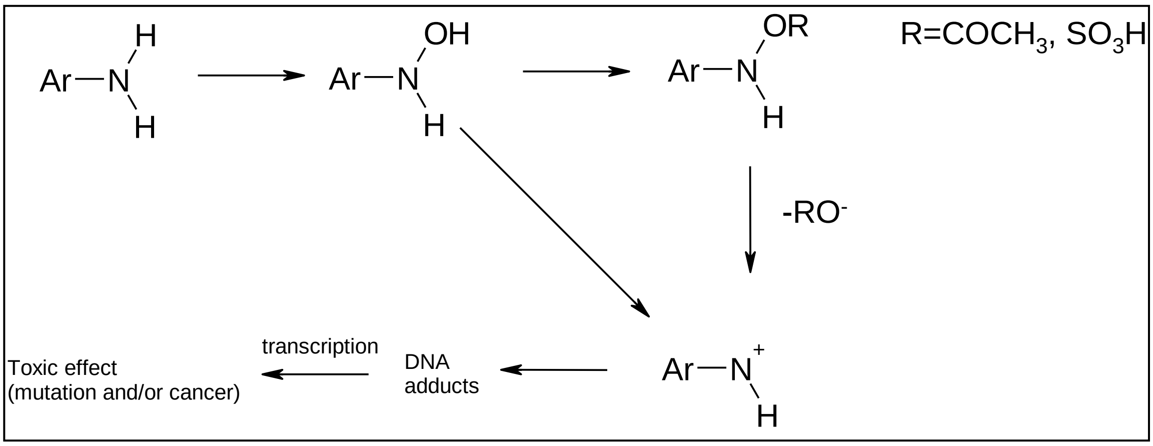


Figure 5.3. A scheme representing the metabolic activation of AAs and subsequent DNA binding[178].

5.1.3 Using Carbocyclic and Heterocyclic Primary Aromatic Amines in Drug Candidates

The regulatory agencies for pharmaceutical preparations require that all drug candidates are tested for the potential of causing genetic mutations. The impracticality and cost of long-term animal tests means that an early judgement on compound toxicology needs to be based on a regulatory battery of *in vitro* tests[179]. The three obligatory assays determine 1) genetic mutations in bacteria (i.e. the Ames test[180]); 2) chromosomal damage to mammalian cells (e.g. the chromosome aberration assay or the mouse lymphoma assay); and 3) chromosomal damage using rodent haematopoietic cells. Not all positive results in genotoxicity assays preclude further development of drug candidates, however, there is a concomitant high probability of rodent carcinogenicity and therefore, by implication, human carcinogenicity[181, 182]. The highest concern is attached to positive outcomes in the Ames test because of the assay's high specificity in relation to rodent carcinogenicity which shows a positive correlation of between 63-90%[183]. Occasionally, genotoxic drugs can progress to market if their therapeutic benefit outweighs the risk associated with them, e.g. some oncology drugs.

The risks associated with AA drug fragments have not halted their use since 89 drugs with this moiety are approved and an additional 131 molecules have entered clinical trails[183]. International regulatory guidelines require a computational SAR assessment of a drug molecule[184, 185] as part of the development process, or at very least an investigation of chemical structure to ensure the standard battery of International Committees on Harmonization (ICH) genotoxicity tests will be suitable for the drug in question[186]. Guidance on SAR assessment of putative drug metabolites is less clear, however metabolites that trigger SAR alerts for genotoxicity or that are known mutagens would obviously represent a cause for concern in any drug development programme[187]. Most attention is focussed on predicting the likely outcome of AAs in the Ames test because of its ability to detect rodent carcinogens. An introduction to the principles of the Ames test is given in Section 5.1.4. Experimental details are provided in the next Chapter. Section 5.1.5 contains a review and discussion on literature and commercial methods for predicting the Ames test result for AAs.

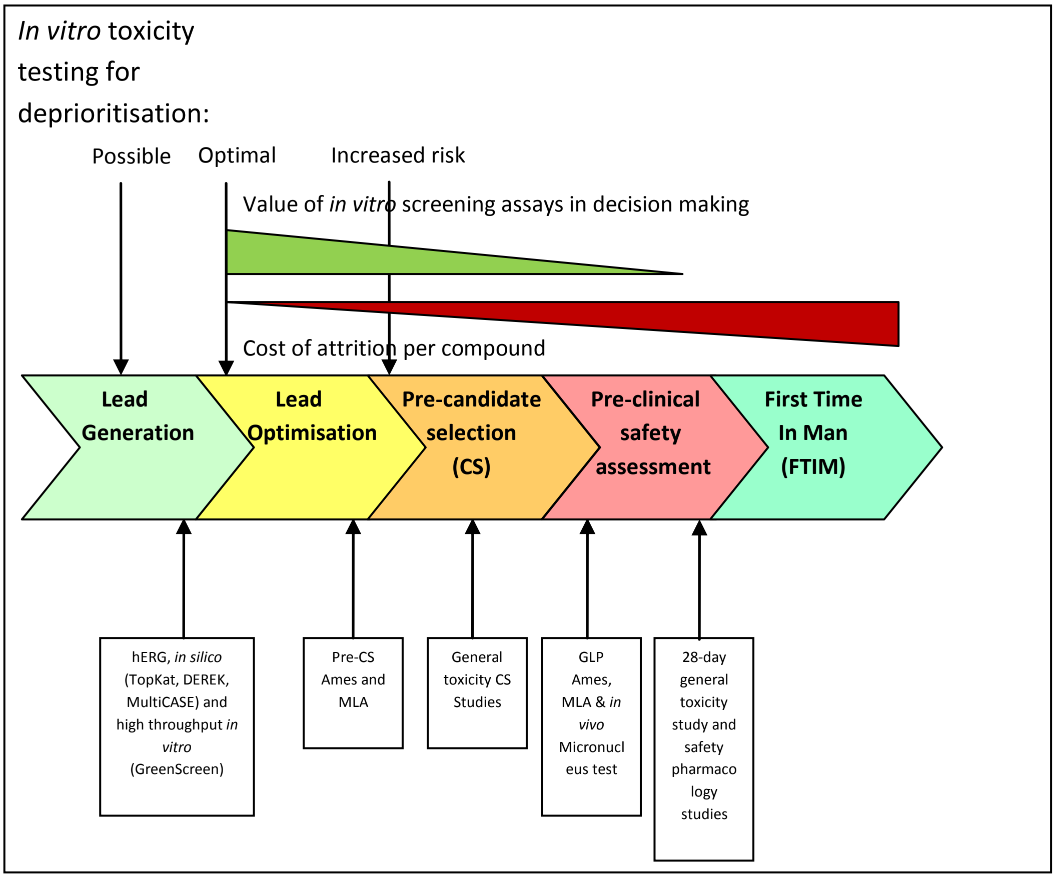
5.1.4 The Ames Test

The Ames test uses amino acid-requiring strains of *Salmonella typhimurium* and *Escherichia coli* to detect point and frameshift mutations, which involves substitution, addition or deletion of one or a few DNA base pairs[180]. The principle of this test is that it detects compounds that mutate mutant bacteria back to the wild type and thereby restore the functional capability of the bacteria (revertants) to synthesise an essential amino acid. The revertant bacteria are detected by their ability to grow in the absence of the amino acid required by the mutant strain. Often it is the metabolites of compounds that are responsible for their mutagenic activity, therefore the Ames test is performed in the presence (+S9) and absence (-S9) of rat liver homogenate, known as S9, to mimic mammalian metabolism.

This bacterial reverse mutation test is long-established, relatively easy to conduct by trained persons and inexpensive[188, 189]. However, in comparison to other commonly used toxicity assays[190], a complete assay is low-throughput, results take at least three days to generate and gram quantities of test compound can be required, which are not necessarily available in the early stages of drug development . Further considerations associated with the Ames test include:

- The test utilises prokaryotic cells, which differ from mammalian cells in such factors as uptake, metabolism, chromosome structure and DNA repair process.
- The test, conducted *in vitro*, requires the use of exogenous sources of metabolic activation, which cannot mimic entirely the mammalian *in vivo* conditions. Thus the test does not provide direct information on the mutagenic and carcinogenic potency of a substance in mammals.
- Although there is a correlation between Ames positive results and mutagenic activity in other tests, there are examples where mutagenic compounds are not detected in the Ames test and vice versa. Reasons for this include the nature of the endpoint detected, differences in metabolic activation, or differences in bioavailability.
- The Ames test is not appropriate for the evaluation of certain types of compounds, for example highly bactericidal compounds.
- Although many compounds that are positive in the test are mammalian carcinogens, the correlation is not absolute.

The factors discussed above prevent the extensive and early use of the Ames test in the lead optimisation stage, along with the other *in vitro* battery of tests for genotoxicity (Figure 5.4). There is the possibility to employ cut-down versions of the regulatory tests. These typically reduce the scale of the test, for example by reducing the Ames test to only two strains of bacteria instead of the five required for a full assessment to provide early indications, however, the number of compounds that can be screened is still limited. In line with modern drug discovery, there is a desire to screen for genotoxicity early[191]. Identifying compounds with possible genotoxic liabilities early means high investment optimising the compound, that will most likely fail later on, will not be made. Higher throughput assays have been developed[192, 193] to give early indications on the likely outcome for a greater number of compounds in the battery of genotoxicity tests, but their predictions are not absolute and they are not considered a replacement.



In vitro toxicity testing for deprioritisation:
Possible
Optimal
Increased risk
Value of *in vitro* screening assays in decision making
Cost of attrition per compound
Lead Generation
Lead Optimisation
Pre-candidate selection (CS)
Pre-clinical safety assessment
First Time In Man (FTIM)
hERG, *in silico* (TopKat, DEREK, MultiCASE) and high throughput *in vitro* (GreenScreen)
Pre-CS Ames and MLA
General toxicity CS Studies
GLP Ames, MLA & *in vivo* Micronucleus test
28-day general toxicity study and safety pharmacology studies

Figure 5.4. The generic path of drug discovery and development[194], and where safety assessment takes place.

Before a compound is selected as a candidate, all chemicals involved in the complete synthetic route (including intermediates) and potential metabolites of the target molecule will be exposed to *in silico* toxicity prediction tools. If any of them are flagged as possible Ames-positive, then they will be tested unless reliable literature data can be sourced. For Ames-positive AAs, it is important to consider two issues regarding their use in drug design. Firstly, if the AA is an intermediate in a synthetic route, its levels as an impurity in the final product must be controlled. European guidelines state that genotoxic impurities can only be present up to a limit of 1.5 µg/day[195]. Remaining below this threshold is relatively easy for low dose drugs but becomes increasingly difficult when high doses are required. Secondly, a drug candidate that has an embedded Ames-positive AA likely to be released through metabolism acting on the parent compound carries high risk and identifying this early is paramount.

5.1.5 Methods for predicting the mutagenicity of AAs

A recent review discussed the vast number of non-commercial QSAR models that have been published to predict mutagenicity and carcinogenicity[196]. The models can be divided into two categories: a) "local" models, i.e. QSAR models for congeneric classes; and b) "general," or global or non-local models, i.e. QSARs for non-congeneric sets of chemical. Predictions of mutagenicity for AAs are usually made from local models, although there are examples of general models being applied to AAs[197-201]. Benigni and Bossa[202] highlighted that non-local QSARs for non-congeneric chemicals are more prone to erratic predictions because modelling large sets of chemicals acting by different mechanisms makes it unavoidable for the use of large numbers of descriptors. Local models for AAs can be divided further into models for the graduation of mutagenic potency and the discrimination between positive and negative compounds. Mutagenic potency (LogMP) is usually the dependent variable (y-variable) used for continuous QSAR models. It is calculated as the logarithm of the number of revertants per nanomole of chemical from the linear portion of the dose-response curve. For discriminant models, compounds are assigned positive and negative labels based on an induction factor (IF) calculated from the Ames test results. This factor is the quotient of the number of revertants per plate divided by the number of revertants per plate of the negative control. Substances with an IF above two or three, depending on the bacterial strain, are considered positive.

Numerous papers have been published attempting to predict the mutagenic potency of AAs using a variety of descriptors and statistical learning methods. Debnath *et al.*[203] gathered from the literature a set of 95 AAs acting on *Salmonella typhimurium* TA98 and TA100 +S9. The authors presented correlations for a subset of 88 of these amines based on hydrophobicity (LogP), energies of the highest occupied (HOMO) and lowest unoccupied molecular orbitals (LUMO), and an indicator variable that designates the presence of three or more fused aromatic rings. Their best result for TA98 has an r=0.898 (r^2=0.806[204]), although the quality of the model deteriorated when all 95 compounds were used. A subset of 67 AAs were modelled for TA100 and produced a model with an r =0.87. The importance of LogP was found to be nearly identical for both models and the electronic descriptors played a minor role. Since the compilation of this data set, it has been revisited by many authors using different classes of descriptors. Basak *et al.*[205] used topological and geometric descriptors to model all 95 compounds with similar results i.e. r=0.893 (r^2=0.797[204]). Maran *et al.*[206] calculated an extensive set of constitutional, geometrical, topological, electrostatic, and quantum mechanical descriptors for this data set. These workers produced a six-parameter correlation with r =0.913 (r^2=0.834[204]). The use of quantum mechanical calculations was criticised by Cash[207] as too demanding on computational resource. Therefore, electrotopological state indices were used to produce an r=0.876 (r^2=0.767[204]). Basak and co-workers[205] revisited the data set and found that the majority of the variance could be explained with topological parameters and the inclusion of LogP. Geometric and quantum chemical parameters did not result in significantly improved predictive models. All these studies used linear techniques e.g. MLR and PLS. Vracko *et al*.[208] attempted to detect the possibility of non-linear relationships between structure and mutagenicity of these compounds using artificial neural networks in conjunction with topological descriptors. They achieved an r^2= 0.751; lower compared to the linear methods. Volkova *et al*.[209] studied how to select the minimal training set, which covers the information space efficiently. They produced an impressive model with neural networks trained on only 30 of the compounds with an r^2=0.986. The test set twice larger than the training set has a correlation between experimental and predicted activity of r^2=0.816. Cash *et al*.[210] revisited the QSAR equation constructed from electrotopological state (E-state) indices to perform external validation using a data set of 29 aromatic amines. The r^2 of the regression between predicted and experimental Log MP was 0.27, indicating that their original model had poor predictive accuracy for compounds in the test set. Examination of the training and test sets revealed that only three of the eight descriptors in their original model were represented in the test set. Both data sets were combined and then randomly split into a new training and test set. The new six-term equation had an r^2 of 0.77 and q^2 of 0.70. However, an r^2 for the test set of 0.44 indicated that the new model provided little improvement. It was concluded that the descriptors used could describe the training set relatively well but the results from external validation indicated possible over-fitting. It was also noted that compounds only differing in an alkyl substituent far away from the amine function greatly affected the LogMP. This is something that the E-state indices cannot account for. Bhat *et al*.[211] increased the size of the Debnath data set to 181 AAs using data from a variety of literature sources. Using AM1 optimised geometries, they calculated hundreds of molecular descriptors. Using multiple linear regression techniques they accounted for 66% of the observed variation in the mutagenic potency. Using neural networks they were able to account for 90% of the variation.

Hatch *et al*.[212] constructed a data set of 80 AAs of diverse structure and a range of 10 orders of magnitude in mutagenic potency. They investigated numerous types of descriptors including, structural and quantum chemical ones. The results were interpreted to show that a main determinant of mutagenic potency was the extent of the aromatic π-system. Small contributions were made by dipole moments and the calculated stability of the nitrenium ion.

Many of the methods discussed above highlight lipophilicity, and the HOMO and LUMO energies as important in determining the mutagenic potency of AAs. Lipophilicity, usually included as LogP, is commonly considered a measure of the propensity of chemicals to be absorbed and transported, as well as of being able to interact with the receptors responsible for metabolic activation. The inclusion of the HOMO energies of the parent AAs agrees with the fact that the majority of AAs are mutagenic only in the presence of the S9 microsomal preparation, by which they are oxidised to the mutagenic metabolite. The importance of the LUMO energies is less straight forward. Given the first step in the metabolic activation of AAs is oxidation, it is the HOMO energy of the parent (or the LUMO energy of the oxidized nitrenium form), rather than the LUMO energy of the parent amine, that might be expected to correlate with the mutagenic potency. To our knowledge the reason for the importance of the LUMO energy is unresolved. It may point to the importance of the stability of the nitrenium ion (discussed below). However, the LUMO energy of the nitrenium ion is not necessarily coincident with that of the parent amine[213]. Felton *et al.*[214] evaluated a set of 23 amino-carbolines related to cooked food mutagens. Their results showed a reasonably strong correlation (r^2=0.80) between LUMO energy and the observed mutagenic potential of several heterocyclic amine mutagens. They reasoned, a lower LUMO energy means a higher electron affinity. Using the reasoning that electron withdrawing groups should lead to lower LUMO energies, two novel, highly mutagenic heterocyclic amine analogues were proposed, but not tested. The HOMO-LUMO energy gap is also considered to be an indication of stability[213]. A large HOMO-LUMO energy gap suggests low reactivity.

LogP, HOMO and LUMO energies are clearly important in predicting the mutagenic potency of AAs. However, it should be noted that not all QSAR analysis of the Debnath dataset found these to be the best predictors of LogMP. Indeed, Maran *et al.*[206] did not include any of these descriptors, instead finding the most important descriptor to be the number or aromatic rings. Other workers[201, 212] have also questioned the importance of LogP to predict mutagenic potency because they did not find good correlation with LogMP, which suggests it may be data set dependent. QSAR equations, constructed from topological parameters, can be difficult to interpret and they tend do not to lend themselves to easy comparison and generalisation. However, some of these have also highlighted the number of aromatic rings as important[213].

Since the nitrenium ions have been implicated as the active electrophile in the reaction with DNA bases, it is reasonable to expect that LogMP may correlate with properties of these ions. Ford and co-workers calculated heats of formation (Δ*H)* for Equation 5.1 by the semiempirical AM1 method for a series of AAs, and observed a negative correlation of LogMP with Δ*H*[215, 216] .

$$ArNH_2 + PhNH^+ \rightarrow ArNH^+ + PhNH_2 \qquad \text{Equation 5.1}$$

An explanation of the correlation provided by Δ*H* of Equation 5.1 is that these enthalpy values indirectly relate to the lifetime and, therefore the selectivity of the nitrenium ions. According to this explanation, the more stable and, therefore, longer lived the ion, the greater the mutagenic potency. This hypothesis was tentatively confirmed by Novak and Rajagopal[217] who indirectly measured the lifetimes of nitrenium ions in solution for 18 AAs and observed a strong correlation with Δ*H* (r^2=0.74). However, the importance of nitrenium ion stability in determining the potency of AAs has been questioned on larger data sets[212].

Borosky studied a set of 17 AAs using higher level DFT calculation, compared to semiempirical AM1 and Hartree-Fock *ab intio* methods previously used. It was found that the formation of the nitrenium ion was more plausible through the N-O dissociation reaction calculated from acetic and sulphuric ester of the parent amine than from the related hydroxylamine. Correlations with the calculated stability of the nitrenium ion were only found when the compounds classified as aromatic, imidazo-carbocyclic and imidazo-heterocylic, were separated. The low correlations with *ab initio* nitrenium ion stabilities in previous work could be caused by the diversity of the data set considered[212]. In this study, water was included implicitly as a solvent and was found to significantly stabilize the nitrenium ions but the same trend was observed for the gas-phase calculations. In the follow-up study Borosky[218] used only gas-phase calculations on a set of 43 AAs for this reason. This time, the data set had to be split into six classes; aromatic; heteroaromatic; imidazocarbocyclic; imidazoheterocylic; dipyridoimidazole; and quinoxalines; to detect correlations. Mutagenic potential was found to increase with nitrenium ion stability and an increased negative charge on the exocyclic nitrogen on the ion. Correlations were found to be non-linear for certain classes of compounds, which in some cases contained as little as three data points. The role of hydrophobicity was also investigated, although positive correlations were observed for each series of compounds, they were not as strong as the other descriptors.

During lead optimisation, medicinal/computational chemists are usually interested in QSAR models that describe how the potency of a series of compounds is modulated by small changes in structure. However, when considering mutagenic potential, it is also important to consider what distinguishes the Ames-positive compounds from those that are Ames-negative. Discriminant models attempt to categorise these compounds using features of the molecular structure. In relation to AAs, it has been demonstrated that the factors that modulate the mutagenic potency are usually different from those that make the difference between positive and negative outcomes in the Ames test[219, 220].

Leach *et al.*[221] studied a number of reactions implicated in causing the mutagenicity of AAs. The reaction energies, computed with (U)B3LYP/6-31G* for 312 compounds, involving the formation of the nitrenium ion in Equation 5.2 and Equation 5.3 were the only energies that were able to significantly discriminate between active and inactive compounds. However, there was an overlap between the energies for both, Ames-positive and Ames-negative compounds. This method was able to discriminate the AAs and because only one parameter was used, it was possible to assign a probability to a compound of being Ames positive. For example, a computed high reactivity should be avoided when designing new compounds. Leach *et al*. also compared the discrimination achieved to that achieved by a large number of in house descriptors for the same compounds. Only two descriptors were able to provide a similar discrimination but their interpretation was less straight forward. It was interesting to note that LogP descriptors were worse than a larger number of other descriptors.

$$ArNHOH + H_3O^+ \rightarrow ArNH^+ + 2H_2O \qquad \text{Equation 5.2}$$

$$ArNHOAc \rightarrow ArNH^+ + AcO^- \qquad \text{Equation 5.3}$$

The use of nitrenium ion stabilities, introduced by Ford *et al*.[215], has recently been extended to discriminate between 257 Ames-positive and negative AAs[183]. From Equation 5.1, one can calculate relative energies ΔΔ*E* according to Equation 5.4.

$$\Delta\Delta E = \Delta E_{ArNH^+} + \Delta E_{PhNH_2} - \Delta E_{ArNH_2} + \Delta E_{PhNH^+} \qquad \text{Equation 5.4}$$

A negative value for ΔΔ*E* indicates that the nitrenium ion for the AA of interest is more stable than that for the reference aniline, and a positive value for ΔΔ*E* indicates a less stable nitrenium aniline. By implication, a negative value of ΔΔ*E* should correlate with Ames-positive result, whereas positive values of ΔΔ*E* should correlate with Ames-negative compounds. The authors compared four different levels of theory and found AM1 provided a good balance between speed and accuracy. They were able to correctly classify 85% of the data set. Similar to the Leach *et al*. approach, their method is based on a continuous spectrum of ΔΔ*E* values, which allowed them to observe and rationalize SAR for some series of molecules. For example, ΔΔ*E* captured the trends going from Ames-positive to Ames-negative for some para-substituted anilines.

Another approach is to use structure alerts originally purposed by Ashby and subsequently revised by Ashby and Tennant[222] to highlight risks of mutagenicity. Kazius *et al*.[223] applied 29 toxicophores to a data set of 4337 diverse chemicals. They defined toxicophores as substructures that indicate an increased potential for mutagenicity, whether this is caused by DNA reactivity or not. They were able to correctly classify 82% of the training set and 85% for an external test set of 535. The statistics for 441 compounds with the specific aromatic amine toxicophore shows an accuracy of 86%. The use of toxicophores for discriminating Ames results for AAs was taken further by Casalegno *et al*.[224]. They extracted diatomic fragments from the Debnath data set using MLR and neural networks to produce models with r^2 values ranging from 0.77 to 0.91. However, the leave-one-out q^2 values were inferior to those obtained by Maran *et al.*[206] with a QSAR equation for the same data set.

There are numerous commercial software packages available for the prediction of Ames mutagenicity, the most well known being DEREK[225], TOPKAT[226] and MultiCASE[227]. The underlying algorithms range from rule-based expert systems to QSAR based methods. DEREK is a knowledge- and rule-based expert system that makes semi-quantitative estimations as to whether or not an Ames positive moiety is present in the input chemical structure. The output consists of a prediction about the presence and probability of a specific sub structure to result in a positive Ames result ranked with adjectives certain, probable, plausible etc. to impossible. On the other hand, TOPKAT provides a probability of mutagenicity based QSAR models constructed from electrotopological descriptors. It also provides a measure of similarity between the molecule of interest and the chemical space covered by the models. MultiCASE dissociates each input molecule into 2-10 atom fragments and statistically evaluates the strength of association between the fragments (toxicophores) and similar ones from the database, assigning mutagenicity scores. The quantitative prediction of mutagenicity is further refined by taking into account physicochemical properties. While predictions from these programmes are widely used and often requested by regulators, their accuracy is limited[228]. Furthermore, AAs have been highlighted as a class of compounds for which the Ames predictions from DEREK and TOPKAT were particularly poor[229].

It was the importance of AAs in a medicinal chemistry context and the discrepancies over the importance of descriptors that motivated us to investigate this class of compound. QTMS has previously been shown to produce good models for a set of 23 halogenated hydroxyfuranone derivatives that do not require metabolic activation to cause mutations in the Ames test. More importantly, QTMS was successfully applied to a set of 23 triazenes that are only active in the Ames test with the addition of S9. A model, including LogP as a descriptors, produced an r^2 and q^2 (leave-one-seventh out) of 0.86 and 0.74, respectively. The remainder of this Chapter discusses the modelling of literature mutagenic potency for AAs using QTMS descriptors.

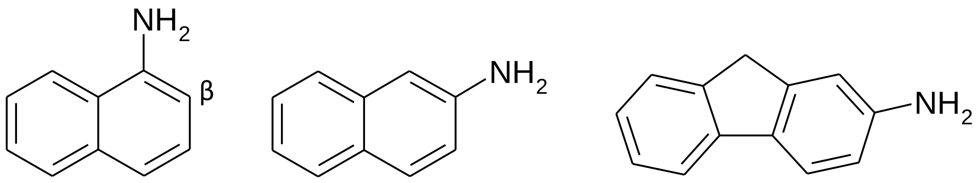
5.2 Methodology

The QCT descriptors were used to investigate correlations with LogMP for a number of literature data sets. Precise details about the investigations are provided in each section. In short, compounds of interest were optimised with Gaussian03 using DFT (B3LYP/6-311+G(2d,p)) to produce the wave function from which the BCP properties were extracted. All the BCP properties were considered, which included the electron density ρ, the eigenvalues of the Hessian of ρ, λ_1, λ_2, and λ_3, the Laplacian $\nabla^2\rho$, the ellipticity ε, the kinetic energy density $K(r)$, a more classical kinetic energy $G(r)$ and the equilibrium bond length R_e.

5.3 Results and Discussion

5.3.1 Sasaki Data Set

As a first step, we selected three compounds that displayed a wide range in mutagenic potency[230] (Figure 5.5).



1-NA 2-NA 2-AF

Figure 5.5. 1-naphthylamine (1-NA) with the β-carbon labelled, 2-naphthylamine (2-NA), and 2-aminofluorene (2-AF).

Experimental results[231] indicate that 1-NA is almost exclusively ring oxidised at the β-carbon, 2-AF is almost exclusively N-oxidised and 2-NA is both ring and N-oxidised, with ring oxidation rates generally much higher than N-oxidation. As ring oxidation is a detoxification mechanism, it means that the mutagenic potency increase according to 1-NA < 2-NA < 2-AF.

Sasaki *et al.*[230] suggested that the nitrenium ion stability is an important factor in explaining the mutagenic potency of these compounds. For this reason, we calculated the BCP properties of the nitrenium ions as well as the parent compounds. In line with experimental results, it was expected that the BCP properties of the common bonds in the three compounds would follow the mutagenic potency order once mapped onto each other. However, this was not the case for either the parent compounds or their nitrenium ions, as 2-AF was the median value in many cases. Different LogMP values were found in the literature which ordered the mutagenic potency of these compounds as 2-NA < 1-NA < 2-AF in contrast to Guengerich's 1-NA < 2-NA < 2-AF. However the difference between LogMP for 1-NA and 2-NA was only 0.07 log units. It is important to note that the experimental results in the first instance were collected in the same laboratory, while the latter results were collected from various sources (1-NA[232], 2-NA[203], 2-AF[233]). As the BCP properties for 2-AF were the median values of the three compounds, in most cases they still failed to predict the mutagenic order, even when the order of 1-NA and 2-NA were swapped.

5.3.2 Hatch Data Set

Hatch *et al.*[212] published a data set of 80 AAs (Appendix D). We constructed a variety of PLS models with different combinations of BCP descriptors from the bonds of the common skeleton shown in Figure 5.6.

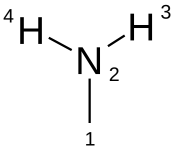


Figure 5.6. The common skeleton and numbering scheme used for the 80 aromatic and heteroaromatic amines.

All the models constructed had an r^2 below 0.25 and a q^2 below 0.21. An example observed versus predicted plot is shown in Figure 5.7.

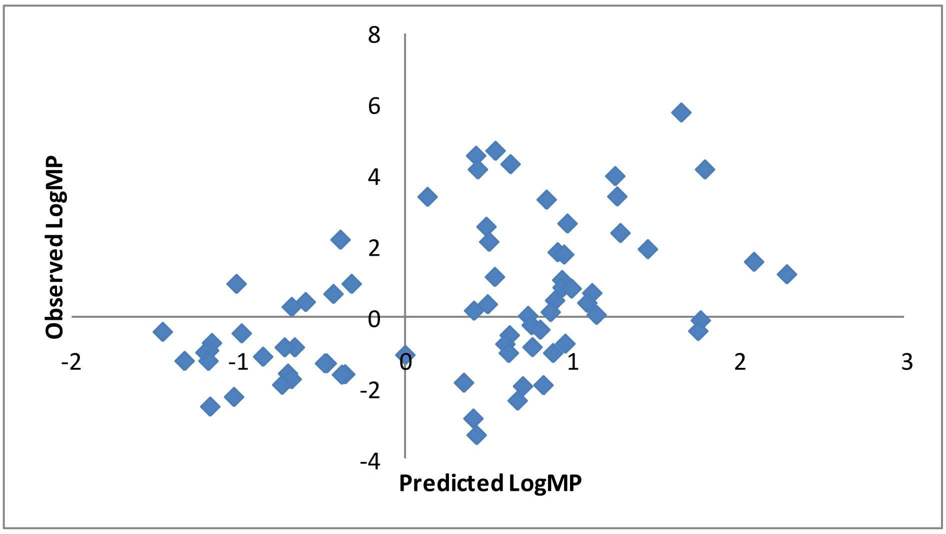


Figure 5.7. An observed versus predicted plot from one of the models constructed for the 80 aromatic and heteroaromatic amines using the C-NH$_2$ bonds.

These poor correlations led us identify subsets of compounds that would provide better correlations. The data set contained 25 aromatic amines and 25 1-methyl-imidazole-2-amine (MIA) derivatives. We attempted to model these separately. The results are reported in the following sections.

5.3.2.1 *Aromatic Amines*

The increased size of the common skeleton for the aromatic amines allowed us to include more bonds and therefore more BCP descriptors. The common skeleton included 6 C-C bonds, 1 C-N bond, 2 N-H bonds and 5 C-R bonds, where R is hydrogen if no substituent is present.

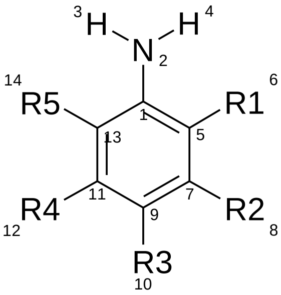


Figure 5.8. The common skeleton and numbering scheme used for the aromatic amines.

Different PLS models were constructed with different combinations of compounds and bonds in an attempt to explain the differences in LogMP for these compounds. Different observed versus predicted plots were inspected.

Table 5.1. PLS models generated for the aromatic amines. The compounds related to the numbering can be found in Appendix D.

	Model no.	No. of Compounds	No. LVs	r^2	q^2	Comment
All aromatic amines	M1	19	3	0.85	0.61	C-C and C-N bonds
	M2	25	1	0.41	0.07	All bonds
	M3	25	1	0.36	0.09	C-C and C-N and N-H bonds
	M4	25	1	0.37	0.08	C-C and C-N
	M5	19	1	0.58	0.22	All bonds – Outliers 80, 76, 73, 43, 41, 21 removed
	M6	19	1	0.45	0.11	As M5 but C-C and C-N bonds
	M7	18	3	0.96	0.32	As M5 – 20 also removed
Diamine Removed	M8	15	2	0.75	0.30	C-C and C-N bonds – diamines 38, 42, 43, 52, 54, 58, 59, 65, 66, 68 are removed.
	M9	15	1	0.61	0.21	As M8 – but all bonds
	M10	15	1	0.59	0.24	As M8 - but C-C, C-N, and C-H bonds
Diamines	M11	10	2	0.77	0.61	C-C and C-N bonds – only diamines 38, 42, 43, 52, 54, 58, 59, 65, 66, 68 included
	M12	10	2	0.91	0.56	As M11 – but all bonds
	M13	10	2	0.69	0.42	As M11 – but C-C, C-N and C-H bonds.

Key points arising from the PLS modelling given in Table 5.1 were;

- No reasonable models could be constructed for the 25 aromatic amines (e.g. M2, M3, and M4)
- Removal of compounds identified as outliers did not significantly improve the results (see M5 and M6).
- Removal of 2-aminoanthracene, whose LogMP value is ~ 1.5 log units higher than all the other compounds in M5, improved the correlation (M5 compared to M7) but the q^2 of 0.32 suggested poor predictive ability. The spread of experimental values is too narrow to be able to build a good model.
- The correlations improved when the compounds with two amino groups were removed (see M8, M9 and M10).
- The diamines produced good correlations when modelled separately, however, there were only 10 compounds in this set (see M11, M12, and M13).

5.3.2.2 1-Methyl-Imidazole-2-Amine (MIA) derivatives

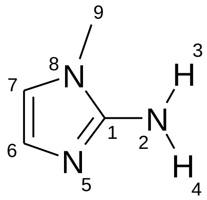
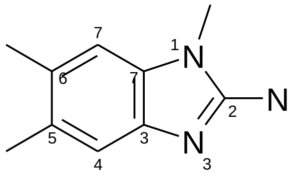


Figure 5.9. The common skeleton of the 1-methyl-imidazole-2-amine derivatives.

The common skeleton of the 1-methyl-imidazole-2-amine (MIA) derivatives is shown in Figure 5.9. The structures of the compounds were examined and subsequently grouped according to the type of ring fused to the C_6-C_7 bond. The ring was either a phenyl ring or a pyridine ring where the position of the nitrogen atom changed in relation to MIA. The different groupings are given in Table 5.2, where the atom numbering refers to Figure 5.10. PLS was used in an attempt to correlate the BCP descriptors to the LogMP. The results are provided in Table 5.3 and an observed versus predicted plot is shown in Figure 5.11 for the best correlation obtained.

Table 5.2. The different MIA groups with the description of the structure.

Group	Description	Colour assigned	No. Of compounds
1	N at the 4 position	Grey	7
2	N at the 7 position	Orange	5
3	No N atom i.e. benzo	Green	6
4	N atom at the 4 position, fused furan ring at the 5,6 position	Purple	1
5	N atom at the 5 position	Pink	3
6	N atom at the 6 position	Blue	3



17

Figure 5.10. The atom numbering scheme used for the MIA derivatives.

Table 5.3. The PLS models generated for the MIA derivatives.

	Model no.	No. of Compounds	No. LVs	r^2	q^2	Comment
All Compounds	M14	25	1	0.51	0.28	All bonds
	M15	25	2	0.73	0.34	C-N and C-C bonds
	M16	25	3	0.54	0.45	C-N bond
Type 1, 2 and 3 MIA	M17	18	2	0.72	0.59	All
	M18	18	1	0.68	0.58	C-N and C-C bonds
	M19	18	4	0.79	0.68	C-N bond
	M20	18	2	0.77	0.71	Bond 6-7
	M21	17	2	0.88	0.85	Bond 6-7, 16 removed

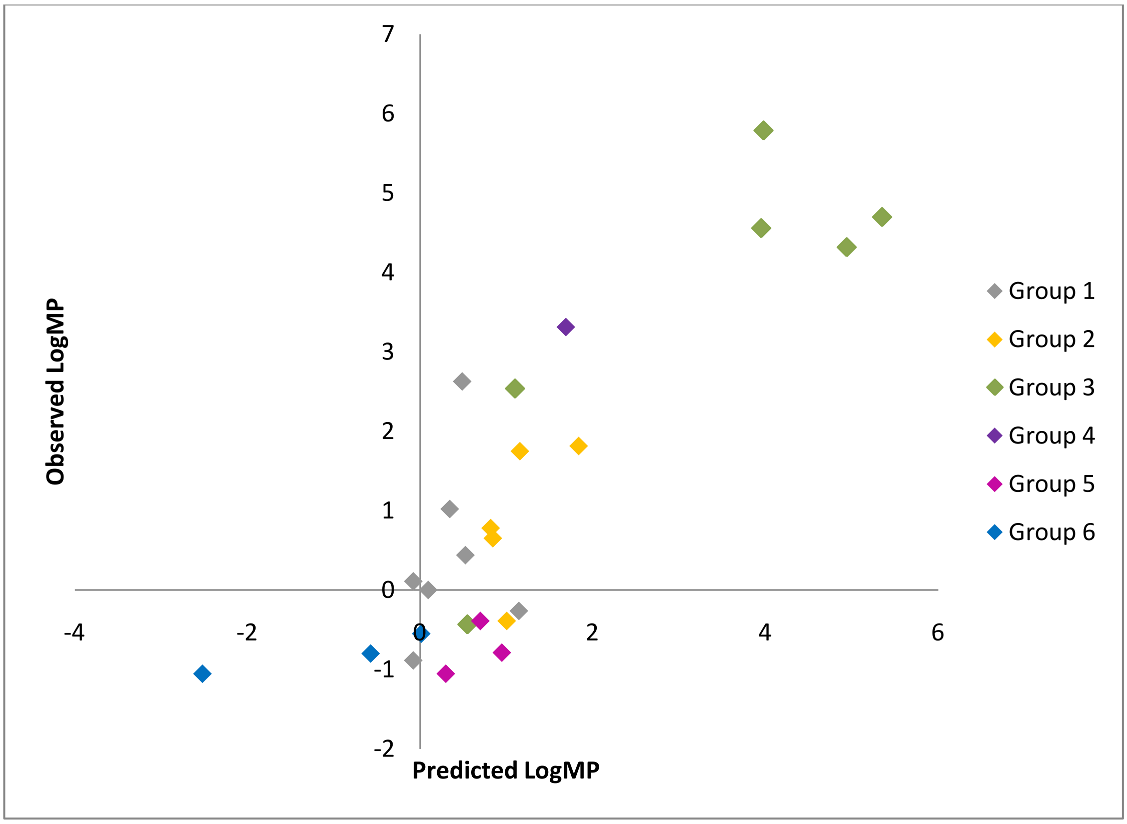


Figure 5.11. The observed versus predicted of model M16 plot for the MIA derivatives.

The number of compounds in groups 1, 2 and 3 and the range of LogMP values allowed us to investigate these three subgroups together. PLS was used to construct models for the 18 compounds belonging to these subgroups (Table 5.3). Models using the BCP descriptors from different bonds gave similar results in terms of r^2 and q^2 (See models M16, M17 and M19). An inspection of the VIP plot for model M16, which was constructed from all the BCP descriptors in the common skeleton (Figure 5.9), highlighted the C_6-C_7 bond was the most important to the model (Figure 5.12).

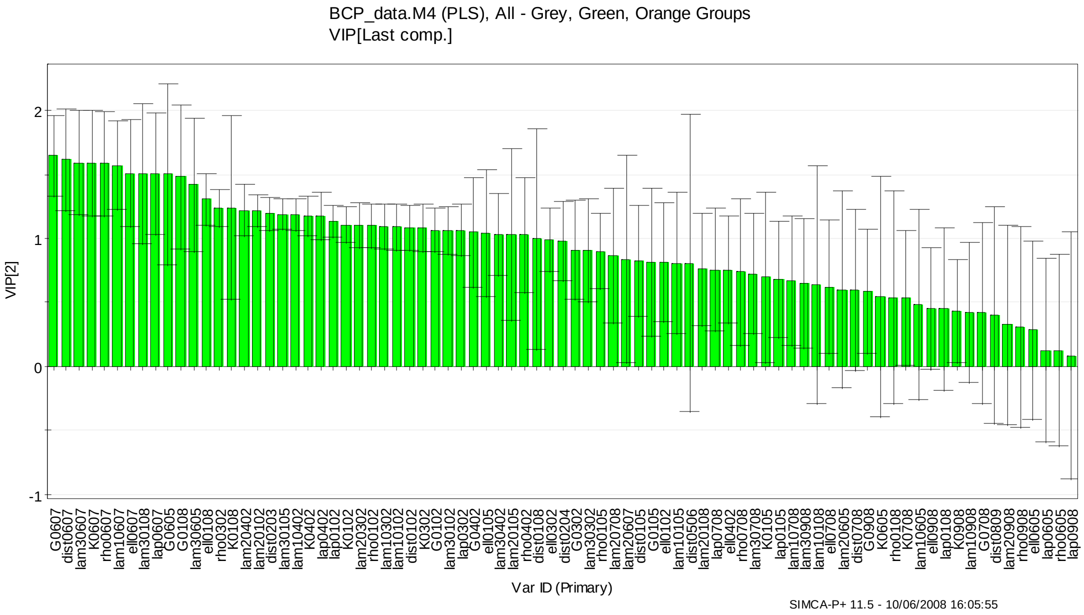
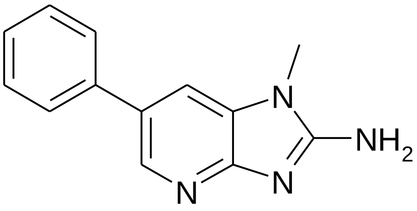


Figure 5.12. The VIP plot for model M16.

This suggested that the C_6-C_7 bond is most sensitive to structural changes in the MIA derivatives. QTMS usually highlights the bonds associated with the active centre, which in this case would be expected to be close to the amine group. However, the C_1-N_2 BCP is separated by least three bonds from a BCP that is directly involved with structural or atomic differences. This raises the issue of the sensitivity of BCP properties, that is how many bonds between a BCP are needed before the properties of the BCP are unaffected by substitutions. This is something which is unresolved.

Visual inspection of the structures showed that substitutions a number of bonds away from the mechanistically important NH_2 group can significantly affect the mutagenic potency of these types of compounds. The importance of the C_6-C_7 bond to the model demonstrates that properties of this BCP capture structural changes.

Eight out of the nine BCP properties of the C_6-C_7 are found in the top ten most important properties to model M16. We constructed a model using only the BCP properties of the C_6-C_7 bond (M20) and saw slight improvements. Compound 16 was identified as an outlier being the only group 1 compound that had a phenyl substituent. It was removed from the modelling and the r^2 and q^2 improved (See M21). Surprisingly, the electron density at the C_6-C_7 BCP produced a linear correlation with LogMP (Figure 5.13).



16

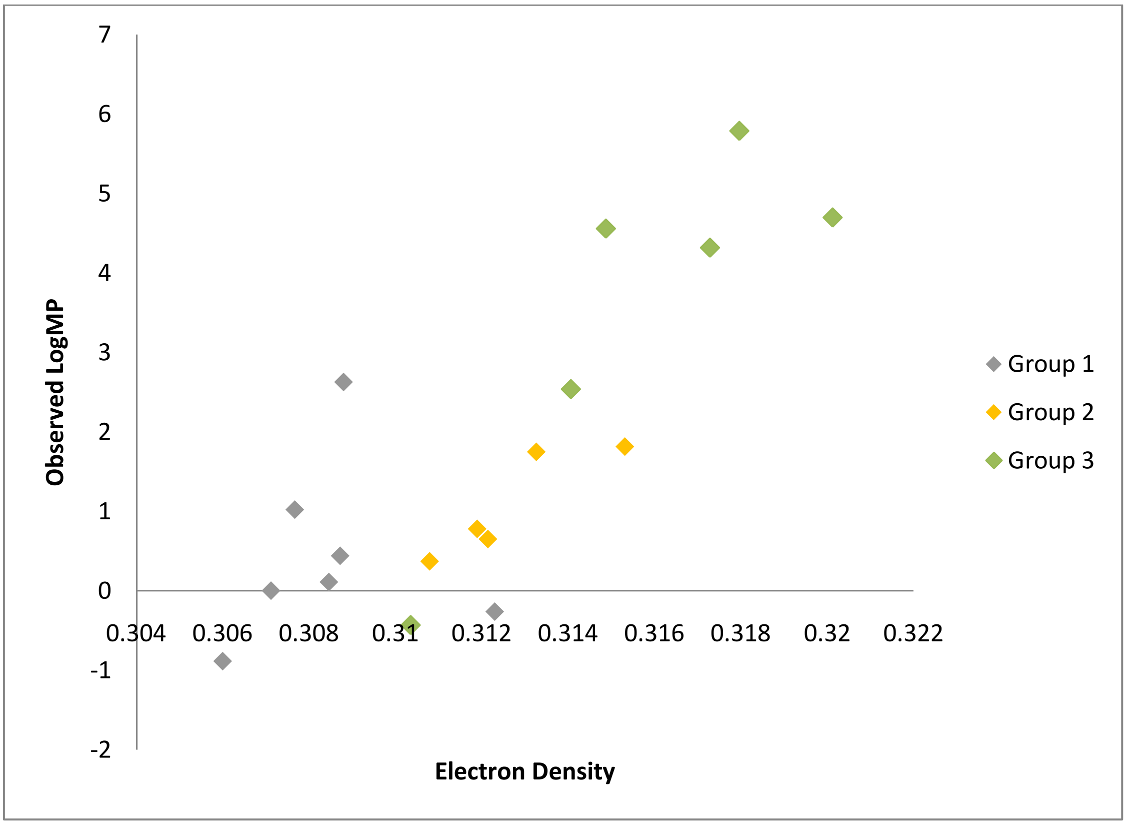


Figure 5.13. A plot of the electron density at the C_6-C_7 BCP versus the logMP.

Correlations of the BCP descriptors with LogMP were found but these required the splitting of the initial data set of 80 compounds. Furthermore, certain compounds had to be excluded to produce reasonable correlations. In an attempt to gain more insight into the use of BCP descriptors for this data set, we employed a number of statistical techniques. Our aim was to improve the models while reducing the numbers of descriptors used. We also wanted to investigate which descriptors and bonds contributed most to explaining the experimental mutagenic potency for both the aromatic amines and the MIA derivatives. Two methods have previously been used in the context of QTMS for this purpose.

Feature selection has been performed by Esteki *et al*.[64, 112] using multi-linear regression. The same technique was performed with our set of aromatic amines and MIA derivatives, separately. All the BCP descriptors from one bond are used to construct a multi-linear regression model. The same was performed for the rest of the bonds in the common skeleton. For example, this involves constructing fourteen models for the aromatic amines (one for each bond in the common skeleton). The models are ranked according to their r^2 value and the bonds that have the highest r^2 are considered most important. The same is performed with the BCP descriptors from the bonds selected as the most important. For example, all the electron density descriptors for the selected bonds are used to construct a model. The descriptors that have the highest r^2 are considered most important. The most important descriptors from the most important bonds are then only considered for further modelling. All the multi-linear regression was performed using TSAR[234] software package.

We also performed a hierarchical PLS approach to select the most important descriptors and bonds. PLS models were constructed for each BCP descriptor from all the bonds in the common skeleton. Subsequently, descriptors with small VIP scores were gradually deleted until a model with only one descriptor remained. A model was then constructed from the top two descriptors in each of the descriptor models and the stepwise deletion of descriptors repeated. The r^2 and q^2 values were then inspected.

Both these techniques can be used to reduce the number of descriptors used to construct models and also highlight which bonds and descriptors are most important. We used both these techniques extensively but no substantial improvement in the models was observed (results not shown).

5.3.3 Changing the Y-Variable

The mutagenic potency of aromatic and heterocyclic amines can be measured in a number of different strains of bacteria. The results presented thus far used the LogMP calculated from the response of the bacterial strain TA98 to the test compounds. Gramatica *et al*.[235] used the Debnath data set to construct separate QSAR equations for mutagenic potency measured in bacterial strains TA98 and TA100. They suggested that steric factors were more important in predicting the mutagenic potency of the compounds measured in TA98 strain, while polarisability, electronic and hydrogen-bonding features were more important in the TA100 strain. However, they highlighted that descriptors were not so easily and singularly interpretable for an understanding of the complex underlying mechanisms. The conclusion may suggest the reason for the mediocre results we obtained for the mutagenic potency of the compounds in the TA98 strain as it is known that QTMS descriptors capture the importance of electronic properties for the property of interest. For this reason, we investigated the use of BCP descriptors to predict the mutagenic potency of aromatic amines measured in the TA100 and the TA98 bacterial strains.

After removal of diamines and highly conformationally flexible compounds, 60 aromatic amines from the Debnath data set[203] remained that had LogMP values measured in TA100 (Appendix D). A plot of the LogMP in TA100 versus LogMP in TA98 (47 aromatic amines LogMP values in both Bacterial strains) revealed that the relative potencies vary widely (Figure 5.14).

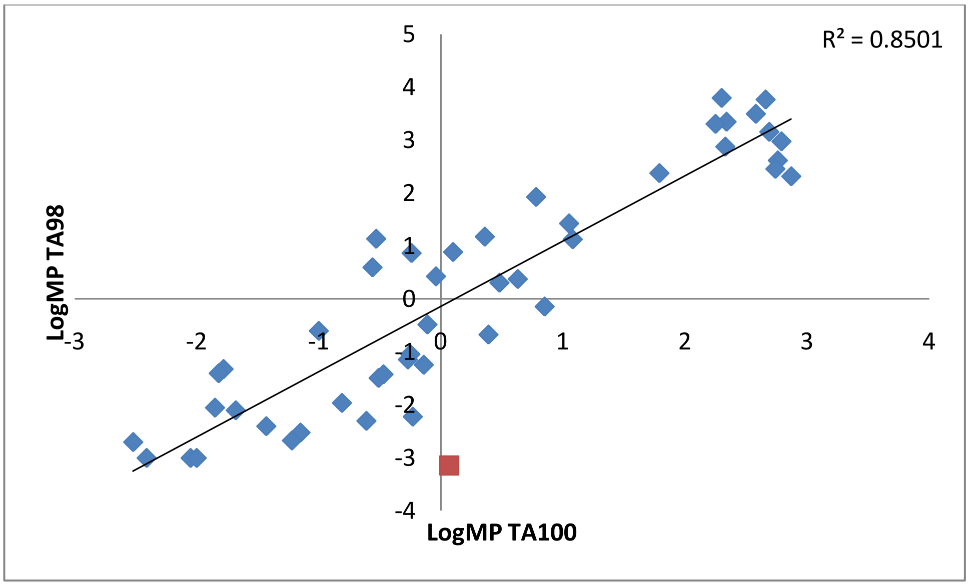
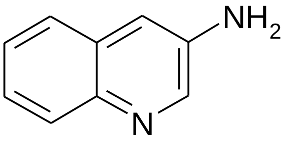


Figure 5.14. A plot of the mutagenic potency of 47 aromatic amines measured in TA100 versus TA98. The red data point represents 3-aminoquinoline and shows the largest difference in mutagenic potency depending on the strain in which it was measured.

3-aminoquinoline is a weak mutagen when measured in TA98 strain (LogMP = -3.14) but its mutagenic potency is stronger when measured in the TA100 strain (LogMP = 0.07).



G_48

Hierarchical PLS was performed on the data set with the LogMP measured in TA100 and TA98. The best model for the prediction of LogMP measured in TA98 contained 38 compounds and had an r^2 and q^2 of 0.69 and 0.61, respectively. The model was constructed from 2 latent variables which contained the information from 13 BCP properties. We tested the model with 15 aromatic amines that had been tested at GSK because ultimately, these are the compounds that we aimed to predict. No LogMP value was provided but a categorisation of positive (assigned 1), negative (assigned -1) or equivocal (assigned 0) was enough to test how the model performed on these compounds. The predictions for the GSK-measured aromatic amines show a spread in potency but the model fails to distinguish the correct activity of the compounds (Figure 5.15).

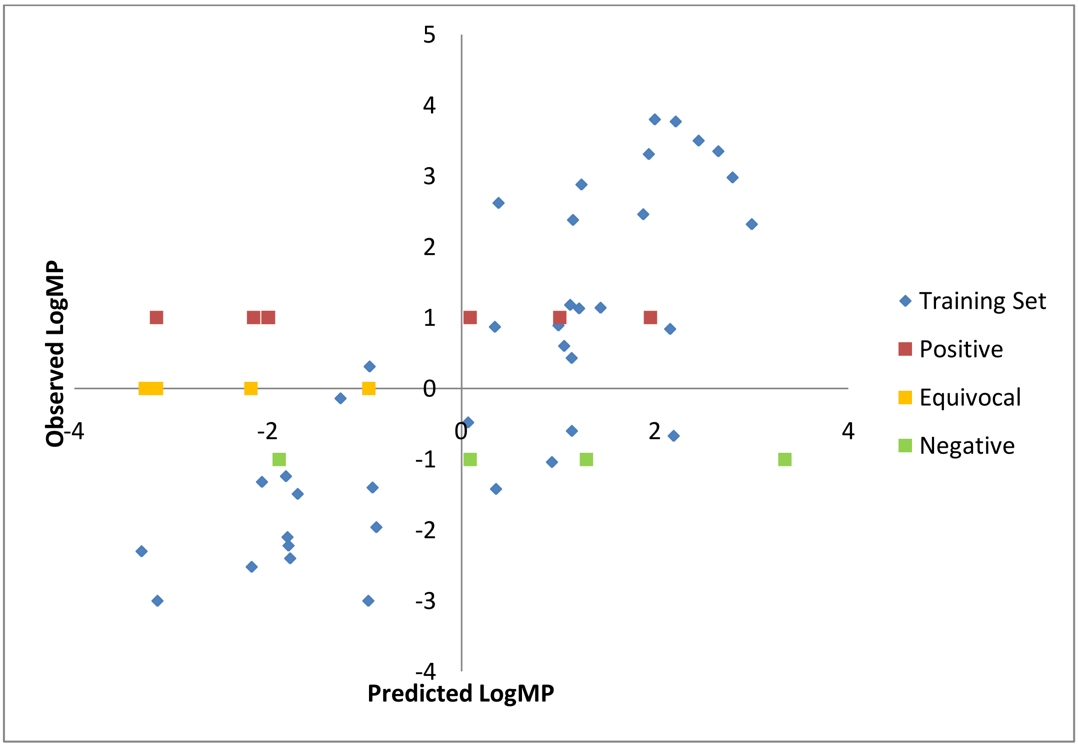


Figure 5.15. Predicted versus observed LogMP measured in bacterial strain TA98 (blue diamonds). Predictions for 15 aromatic amines measured at GSK are also given (red squares = Ames-positive, orange squares = Ames-equivocal and green squares = Ames-negative).

The best model for the prediction of LogMP measured in TA100 contained 50 compounds and had an r^2 and q^2 of 0.57 and 0.54, respectively (Figure 5.16). The model was constructed from only four ellipticity (ε) descriptors with the information contained in one latent variable. A linear relationship is clearly present; however, at the higher end of the mutagenicity scale the relationship plateaus suggesting possibly a non-linear correlation. A further explanation for this observation could be that the 1.5 – 2 LogMP is the upper experimental limit for detecting the mutagenic potency. The spread in LogMP values of the 15 aromatic amines measured at GSK is less than the spread for the TA98 model (Figure 5.15). The strain TA100 is reportedly more influenced by the electronic properties of the compounds tested but also fails to distinguish Ames results for the GSK compounds.

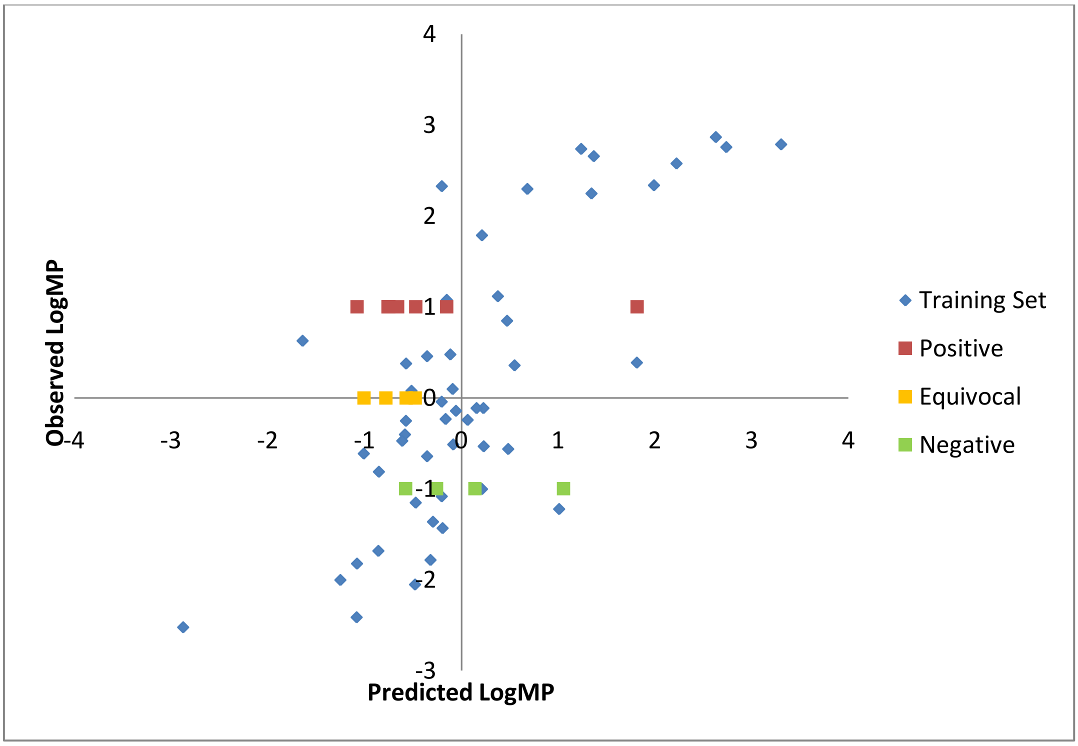
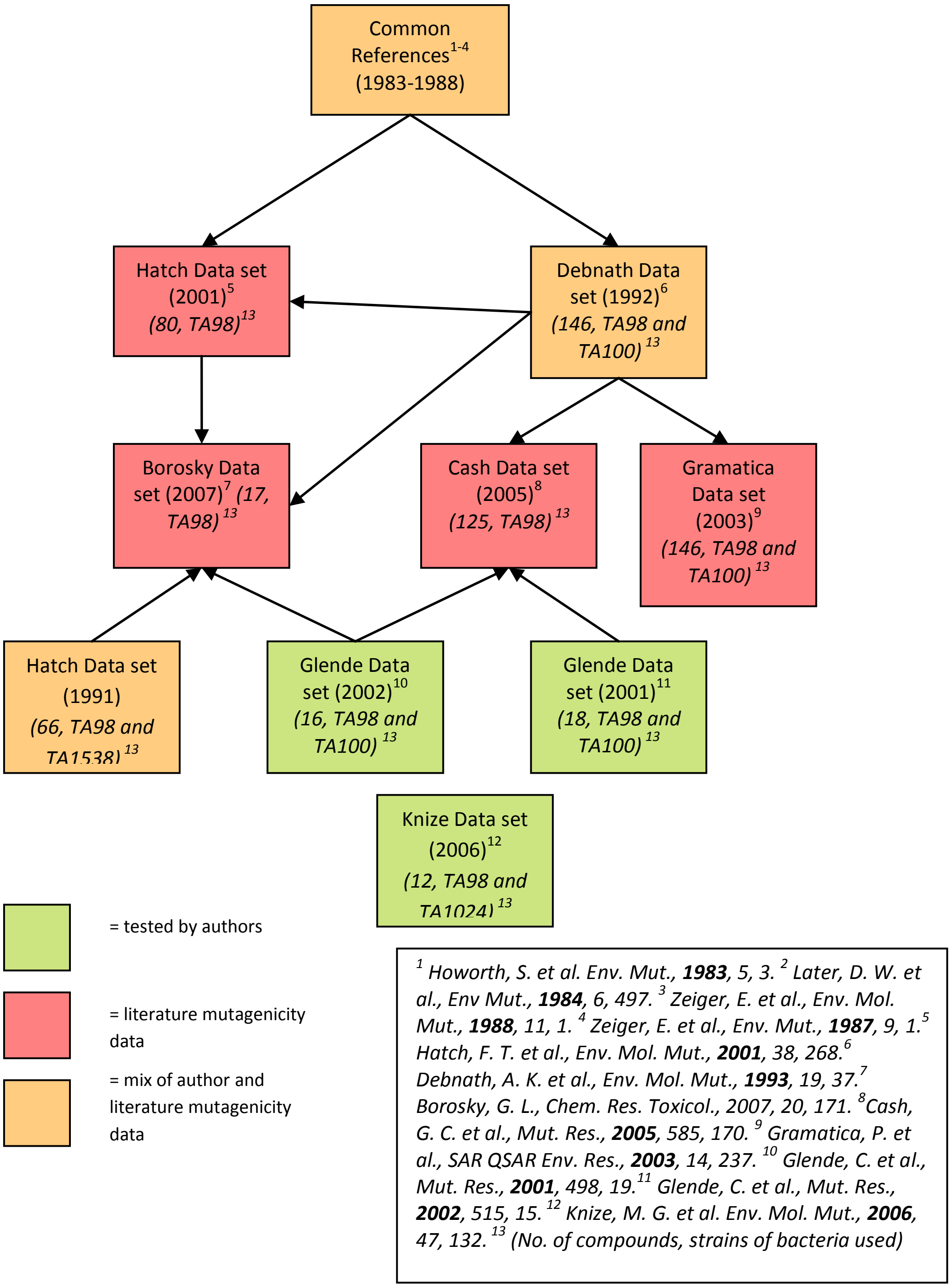


Figure 5.16. . Predicted versus observed LogMP measured in bacterial strain TA100. Predictions for 15 aromatic amines measured by GSK are also given.

5.4 Summary

The statistical analysis of QTMS properties failed to produce a robust, predictive model to predict the toxicity of AAs. We have shown that by using BCP descriptors alone, it is unlikely that a single model for the prediction of these compounds can be successful. Reasonable correlations were obtained when the data sets were split into congeneric series. The use of mutagenic potency of the AAs measured in the bacterial strain TA100 did not improve the results. A reason for mediocre results could be that other factors than electronic effects have a more significant influence on the mutagenic potency of the compounds studied. However, in any QSAR study the integrity of experimental results is vital[236]. Merging Ames assay data from different sources suffers from interlaboratory variations in techniques.

The national Toxicology Program (NTP) determined the average interlaboratory reproducibility of the Ames test data to be 85%[200, 223]. This means that 15% of compounds tested will be either false positive or false negatives. Therefore, any models constructed for the discrimination between positive and negative compounds can only ever have an 85% chance of the prediction being correct if data used is generated in different laboratories. This highlights the difficulty in bringing together data points for a labour-intensive, biological assay in large enough numbers to be able to construct meaningful statistical models. The literature data phylogeny in Figure 5.17 shows that the majority of data used to model the mutagenic potency of AAs are taken from numerous laboratories.


Common References[1-4] (1983-1988)
Hatch Data set (2001)[5] *(80, TA98)*[13]
Debnath Data set (1992)[6] *(146, TA98 and TA100)*[13]
Borosky Data set (2007)[7] *(17, TA98)*[13]
Cash Data set (2005)[8] *(125, TA98)*[13]
Gramatica Data set (2003)[9] *(146, TA98 and TA100)*[13]
Hatch Data set (1991) *(66, TA98 and TA1538)*[13]
Glende Data set (2002)[10] *(16, TA98 and TA100)*[13]
Glende Data set (2001)[11] *(18, TA98 and TA100)*[13]
Knize Data set (2006)[12] *(12, TA98 and TA1024)*[13]
🟩 = tested by authors
🟥 = literature mutagenicity data
🟧 = mix of author and literature mutagenicity data

*[1] Howorth, S. et al. Env. Mut., **1983**, 5, 3. [2] Later, D. W. et al., Env Mut., **1984**, 6, 497. [3] Zeiger, E. et al., Env. Mol. Mut., **1988**, 11, 1. [4] Zeiger, E. et al., Env. Mut., **1987**, 9, 1.[5] Hatch, F. T. et al., Env. Mol. Mut., **2001**, 38, 268.[6] Debnath, A. K. et al., Env. Mol. Mut., **1993**, 19, 37.[7] Borosky, G. L., Chem. Res. Toxicol., 2007, 20, 171. [8]Cash, G. C. et al., Mut. Res., **2005**, 585, 170. [9] Gramatica, P. et al., SAR QSAR Env. Res., **2003**, 14, 237. [10] Glende, C. et al., Mut. Res., **2001**, 498, 19.[11] Glende, C. et al., Mut. Res., **2002**, 515, 15. [12] Knize, M. G. et al. Env. Mol. Mut., **2006**, 47, 132. [13] (No. of compounds, strains of bacteria used)*

Figure 5.17. Literature data phylogeny for Ames data used to model the mutagenic potency of aromatic and heterocyclic amines.

The two largest literature data sets, Debnath *et al*.[203] and Hatch *et al*[212], were compared. Ten compounds were common to both data sets where the mutagenic potency value was taken from different sources. The average difference between the mutagenic potency was 0.70 log units. All LogMP values in the Debnath data set were produced using TA98 +S9 whereas the Hatch data set was produced using TA98 or TA1538 which could be the reason for the discrepancies. Where multiple conflicting data was found when compiling the Hatch data set the authors used the median value, which could be another reason for these differences. The use of LogMP as a dependent variable has recently been questioned because of the influence of experimental noise, variation in environment and differences between laboratories[237]. This confirms the questioning of the reproducibility of the Ames test by Kazius *et al*.[223], who highlighted that the reproducibility is limited by the purity of the tested chemical, inconsistencies in the interpretation of dose-response curves, interferences by toxic side effects (such as cytotoxicity), variations in the methodology employed, and variations in the materials used (bacterial strains and metabolic activation mixtures).

Although CYP1A2 is implicated as the major P450 enzyme involved in the metabolism of AAs, it is not inconceivable to expect AAs to interact and be metabolised by other P450s. To investigate these possible interactions, we selected 73 AAs (34 Ames-positive and 39 Ames-negative according to internal GSK and literature data) based on structural diversity, availability and mutagenic potency and submitted them to the standard P450 enzyme inhibition assays at GSK consisting of the five pharmaceutically important P450 (3A4 (two different assay standards), 1A2, 2C19, 2C9 and 2D6 (Figure 5.2)) at a top concentration that allowed inhibition to be measured above a pIC_{50} of 4.3. These assays are fluorescence intensity-based and in high-throughput format. The results are shown in Table 5.4. The agreement between the two different 3A4 assays was extremely good, so we only report the combined results. Only ca. 20% of the 73 AAs inhibited each of the P450 enzymes but there was little difference between the number of Ames-positive and -negative AAs, even for CYP1A2. We submitted the 73 AAs for testing at higher concentrations in the 1A2 inhibition assay allowing pIC_{50} values to be measured down to 3.4. More of the AAs were found to inhibit 1A2, but once again the differentiation between Ames-positive (13 compounds) and –negative (17 compounds) was poor. These results must be taken into context. The Ames-negative compounds inhibiting CYP1A2 could be enzyme inhibitors but not substrates, hence no genotoxic outcome. The Ames-positive compounds not inhibiting CYP1A2 could acquire their genotoxic potential from conversion by other CYP enzymes, hence exerting their DNA-modifying effect via a mechanism different from the nitrenium ion pathway.

Table 5.4. The results of 73 AAs submitted for P450 profiling in the major P450 enzymes responsible for the metabolism of the majority of drugs.

P450 Enzyme	Minimum pIC_{50} Measurable	Ames-Positive	Ames-Negative	Total
3A4	>4.3	6	10	16
2C19	>4.3	7	8	15
2C9	>4.3	9	9	18
2D6	>4.3	9	9	18
1A2	>4.3	6	10	16
1A2(Higher Concentration)	>3.4	13	17	30

Another complication to consider is the evidence of cooperative binding in the active site of CYP1A2[238]. Heterotropic cooperativity, in which one ligand modifies catalysis of another ligand, has been demonstrated using kinetic studies and computational modelling techniques with 1-isopropoxy-4-nitrobenzene and 1,4-phenylene diisocyanide[239]. In this case the co-occupancy by the two molecules led to enhanced binding but reduced catalytic activity. Homotropic cooperativity with CYP1A2 has also been observed with pyrene and benzo[a]pyrene[240]. The author proposed that co-occupancy of the CYP1A2 active site is a common feature for numerous small substrates (and other ligands) but the nature of cooperative behaviour is highly ligand-dependent. The active site of CYP1A2 (approximately 370 Å^3) is large enough to fit only one α-naphthoflavone[241] molecule but many of the AAs we considered are much smaller. Cooperativity will skew the correlation between a molecular descriptor and the biological test result. Furthermore, cooperativity can lead to enhanced and reduced catalytic activity and therefore, may increase or decrease the mutagenic potency of AAs. While docking studies may be used to predict co-occupancy, the information provided is unlikely to be sufficient to make predictions about the sites of oxidation and so this potential influence on mutagenic potency is difficult to include in any QSAR modelling.

Around 2007, GSK had decided to measure the genotoxicity for a set of 500 synthetic building blocks to generate a list of Ames-negative AA molecules. The work started on two compounds but stalled due to purification issues which resulted in contradictory Ames results for the same compounds. In the next chapter, we report the outcomes of an investigation into the fidelity of the Ames test for a set of AAs.

Chapter 6
Investigating the Reproducibility of the Ames Test with Carbocyclic and Heterocyclic Primary Aromatic Amines for Modelling Ames Test Classification

6.1 Introduction

m-Toluidine is a recognised Ames-negative aniline[242]. It was tested by GSK as the free base in DMSO as a brown solution and was Ames-positive, with and without metabolic activation (+/-S9). The compound was re-purified, stored frozen and protected from light. The free base in dimethyl sulfoxide (DMSO) was a clear liquid and tested Ames-negative. The hydrochloride salt, made from the re-purified m-toluidine was also Ames negative when tested separately in DMSO and water. In other laboratories, 4-aminobenzylamine has also been tested as a brown liquid and produced an Ames positive +S9 at only the two highest concentrations and just over the 2-fold positive criterion[243]. Subsequently, the compound was purified, kept refrigerated and protected from light and turned out to be negative[243]. These examples suggest that degradation could explain the positive Ames outcomes. In contrast, three samples of 4-aminobenzamide from three different suppliers showed 3 very different Ames responses in TA98 +S9 when tested by GSK, whereas without metabolic activation (-S9) in TA 98 and in 4 other bacterial strains (+/-S9) the response remained below the 2-fold increase threshold for Ames-positive classification (Figure 6.1). Both positive batches were confirmed to be 99.8% pure by HPLC-LCMS. This chemical has been reported as Ames negative, which was confirmed in N-methylpyrrolidine (NMP) as a solvent and surprisingly, after standing in DMSO for 1.5 hours[244].

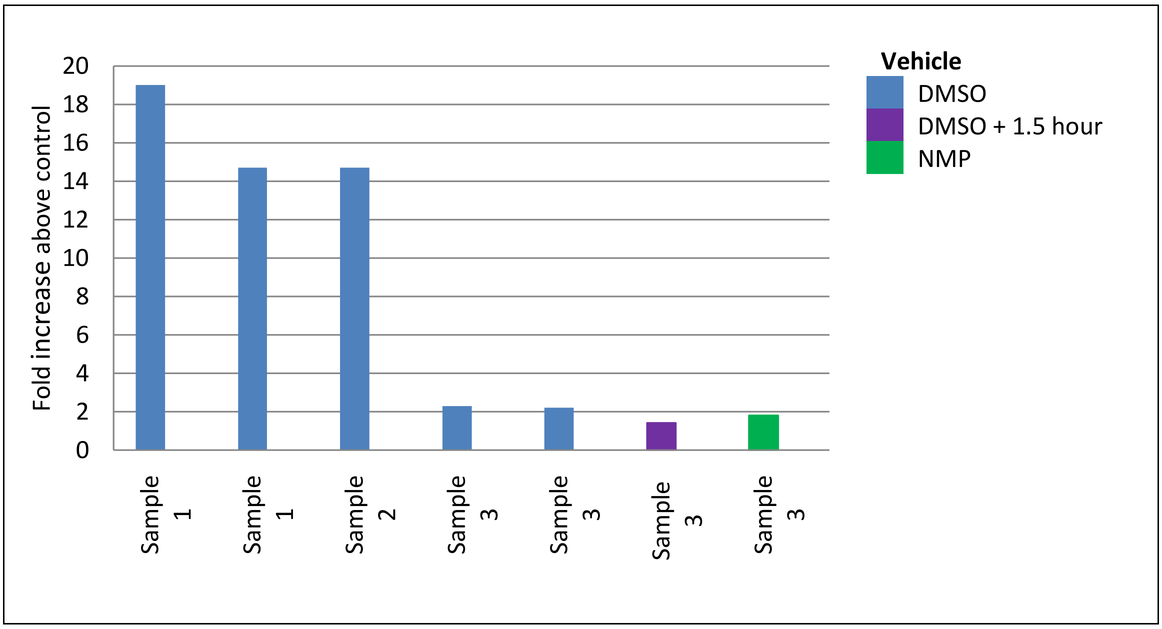


Figure 6.1. The fold increases above the negative controls for the three different batches of 4-aminobenzamide (samples 1 and 3 with retests) screened in TA98 +S9 in DMSO or NMP.

International guidelines have been developed for laboratories to ensure uniformity of testing procedures. These list DMSO as the solvent of choice if the test compound does not dissolve in water. There are three reasons for selecting DMSO: it dissolves a wide range of chemicals, is relatively non-toxic to the bacteria and microsomal S9 enzymes, and it is completely miscible with the molten agar used in the Ames test.

However, DMSO itself is not chemically inert and has been known to affect Ames results[231, 245-247]. Moran *et al*.[248] screened 14 solvents with known mutagens benzo[a]pyrene and 2-aminofluorene for their compatibility with the Ames test. They found 12 solvents to be satisfactory under the conditions specified, including DMSO. They recommended that other solvents be used instead of DMSO when the test compound reacts with DMSO or when DMSO could interfere with the process of metabolic activation. Nestmann and co-workers suggested that repeat tests in a second solvent should be performed to confirm initial findings[246]. Conflicting Ames results for p-phenylenediamine were followed up by Burnett *et al*.[244] with more detailed studies. They found that fresh solutions of either DMSO or water were non-mutagenic whereas DMSO solutions became Ames-positive upon standing, which was not observed for the p-phenylenediamine dissolved in water.

To the best of our knowledge, there are no systematic investigations into the effects of the purity of the samples or the presence of DMSO on the Ames test for AAs. Hence, we performed investigations to establish the robustness of the Ames test for AAs pre- and post-purification and with the hydrochloride salt forms.

6.2 Methodology

6.2.1 Data Sets

We started out with a set of 22 low molecular weight, commercially available AAs that were structurally diverse, known to be Ames-positive and spanning a range of activity. Initially, it was attempted to purify the AAs by distillation or by HPLC. However, boiling points above 220^{o}C for many of the AAs engender the possibility that some of them may decompose during distillation. The time and effort required to produce enough of the purified material also precluded HPLC. It was therefore decided to use recrystallisation techniques. A number of purification and availability issues brought the number of molecules down to the 14 listed in Table 6.1.

6.2.2 Purification and Conversion to the Hydrochloride Salt

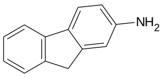
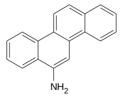
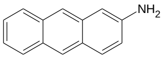
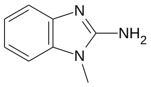
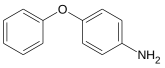
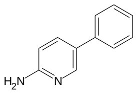
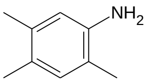
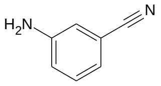
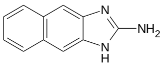
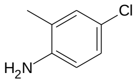
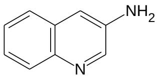
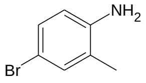
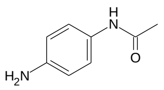
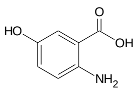
The solvents used to purify the AAs are listed in Table 6.1. Filtration through a PL-Thiol MP SPE+ cartridge (Polymer Laboratories) removed any potential metal impurities and the AAs were recrystallised from a variety of solvents. The crystals were stored at 4°C and protected from light. Missing values in Table 6.1 indicate that no suitable solvent was found from which the AA could be recrystallised.

If the crystallisation failed, the unpurified material was used after being passed through a cartridge. The AAs were dissolved in ethyl acetate, and then hydrogen chloride in diethyl ether was added to give suspensions. The mixtures were stirred for approximately 30 minutes, the solids isolated by filtration and washed with ethyl acetate. The solids were foil-wrapped to exclude light, dried under high vacuum at ambient temperature and then stored at 4°C. In total nine hydrochloride salts were prepared.

The purities of the unpurified, purified and hydrochloride salt materials were determined using HPLC with LC/UV/DS/ELSD detection (Table 6.1)[1]. The percentage purity could not be determined for two compounds (4-aminoacetanilide and 2-amino-5-hydroxybenzoic acid) using this method.

[1] 0.01 ml of a 1 mg/ml solution of each compound was injected in CH3CN/H2O/TFA at a concentration ratio of 10:10:1, respectively. Samples were analysed using a Luna C18(2) 4.6mm i.d. x 150mm column with a gradient of 3% B to 50% B in 30 minutes, holding at 50% B for 20 minutes. A is H2O:TFA, 1000:1 and B is CH3CN:TFA, 1000:2. The flow rate was 1ml/min. A UV detector set at 300nm, 180nm bandwidth with a 550nm, 100nm bandwidth as a reference was used in conjunction with mass spectrometer and an evaporative light scattering detector.

Table 6.1. The names, CAS and structures of the 14 AAs, their purification conditions and levels of purity.

Name (CAS #)	Structure	Recrystallisation Solvent[a]	% Purity		
			Unpurified	Purified	HCl Salt
2-aminofluorene (153-78-6)		ethanol	99.4	99.7	99.3
6-aminochrysene (2642-98-0)		-	99.4	-	-
2-aminoanthracene (613-13-8)		petroleum ether and ethanol	96.0	96.8	100
1-methyl-2-aminobenzimidazole (1622-57-7)		ethanol followed by water	98.6	99.9	-
4-phenoxyaniline (139-59-3)		-	99.2	-	98.6
2-amino-5-phenylpyridine (33421-40-8)		-	97.6	-	-
2,4,5-trimethylaniline (137-17-7)		petroleum ether and ethanol	99.8	100	98.5
3-aminobenzonitrile (2237-30-1)		petroleum ether and cyclohexane	98.2	95.5	-
2-Aminonaphtho(2,3-d)imidazole (102408-31-1)		ethanol	99.8	99.4	100
4-chloro-2-methylaniline (95-69-2)		-	94.6	-	99.5
3-aminoquinoline (580-17-6)		-	99.8	-	100
2-methyl-4-bromoaniline (583-75-5)		cyclohexane	99.9	100	-
4-aminoacetanilide (122-80-5)		ethanol	-	Not determined	-
2-amino-5-hydroxybenzoic acid (394-31-0)		ethanol	-	Not determined	-

[a] The solvent purities were as follows: methanol, 99.9%; cyclohexane, 99.5%; toluene, 99.8%; diethyl ether, 99.0%; dichloromethane, 99.8%; chloroform 99.8%; methyl pentane, 95.0%.

6.2.3 The Ames Test

To evaluate the mutagenic activity of the AAs in relation to the purity, solvent and protomer (i.e. freebase or hydrochloride salt), the bacterial tester strains Salmonella typhimurium TA98 and TA100, with and without metabolic activation (rat liver homogenate S9), were used in the Ames standard plate incorporation assay. The assay was performed in accordance with the procedure described by Maron and Ames[189] and regulatory guidelines[249]. These two strains were chosen, instead of the standard five, as they are often used in cut-down versions of the Ames test[250] because TA98 is capable of detecting frameshift mutations, while TA100 detects base pair substitutions[180]. A heat map of the Ames test results for 100 substituted anilines tested in 6 different bacterial strains (+/-S9) collated from GSK and external sources (e.g. National Toxicology Database in the US) exemplifies the significance of TA98 and TA100 for this chemotype (Figure 6.2). All the compounds in Table 6.1 are known to be Ames positive in either one or both of the strains used. GSK assign a compound equivocal if a clear positive or negative response is not observed or if there is contradictory results for the same compound.

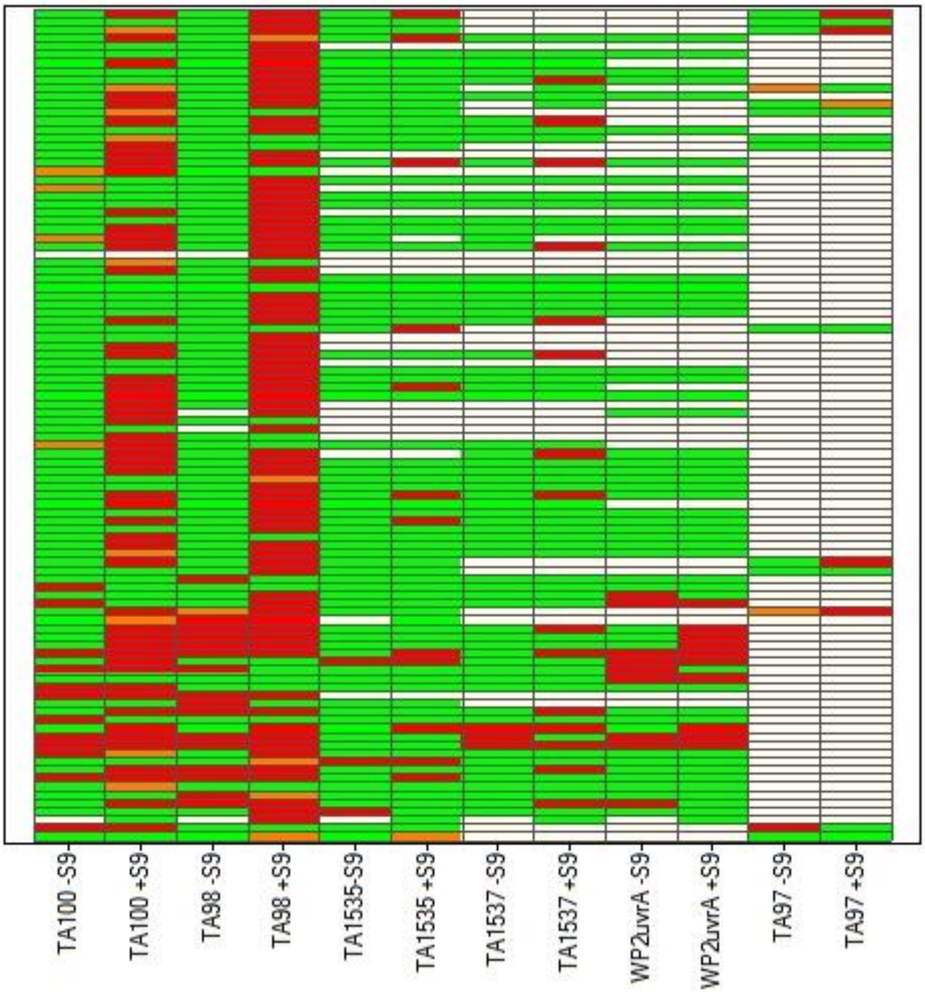


Figure 6.2. Ames test results (red – positive; orange – equivocal; green – negative; white - no data available) for 100 substituted anilines in 6 bacterial strains +/- S9 based on internal and external data sources.

The dilution solvent for compounds was either DMSO (Sigma-Aldrich) or distilled water (GIBCO), and all dosing solutions were prepared immediately prior to testing. All experiments were performed with solvent controls and each run in duplicate with the average value reported. In a few cases where large deviations between runs had been observed, they were repeated. Vehicle controls (DMSO or water) were run in quadruplicates with and without metabolic activation (+/-S9). The positive controls were run in quadruplicate with and without metabolic activation. The positive control for TA98 +S9 was benzo[a]pyrene (B[a]P) (10 µg/plate), 2-nitrofluorene (2NF) (1 µg/plate) for TA98 -S9, 2-aminoanthracene (2-ANN) (5 µg/plate) for TA100 +S9 and sodium azide (2 µg/plate) for TA100 -S9. Positive controls were also run in duplicate. All compounds were generally tested at 6 concentrations (5000, 2500, 1500, 500, 150 and 50 µg/plate) in TA100 and TA98 +/-S9. Some tests were repeated at lower concentrations due to toxicity and/or precipitation observed at low concentrations in the original tests. Where repeats were performed they are highlighted in Table 6.2.

Here the procedure to perform the plate incorporation assay is explained. A cartoon representation of the preparation of one plate in the Ames test is shown in Figure 6.3. A predetermined number of sterile, capped tubes are filled with 2.0 ml of top agar. For the plate incorporation method without activation (-S9), 0.1 ml of the test solution dissolved in the vehicle (water or DMSO) at the required concentration, 0.1 ml of the bacterial strain (TA98 or TA100) and 0.5 ml of sterile buffer are added to the 2.0 ml of top agar in the sterile tube. For the assay with metabolic activation, 0.5 ml of metabolic activation mixture (containing 10% v/v of S9 fraction) replace the 0.5 ml of sterile buffer. The contents of the tube are mixed by rotating between the fingers and poured onto a minimal agar plate. The plate is moved in a circular fashion to spread the agar uniformly. The overlay agar is allowed to solidify before the plate was incubated at 37^{o}C in the dark for 72 hours. After the incubation period, the numbers of revertant colonies on the plate are counted. The Ames test we perform requires the preparation of 60 (6 (concentrations of test compound) x 2 (bacterial strains) x 2 (+/-S9) x 2 (duplicates) + 12 (positive and negative controls) = 60) plates prior to counting.

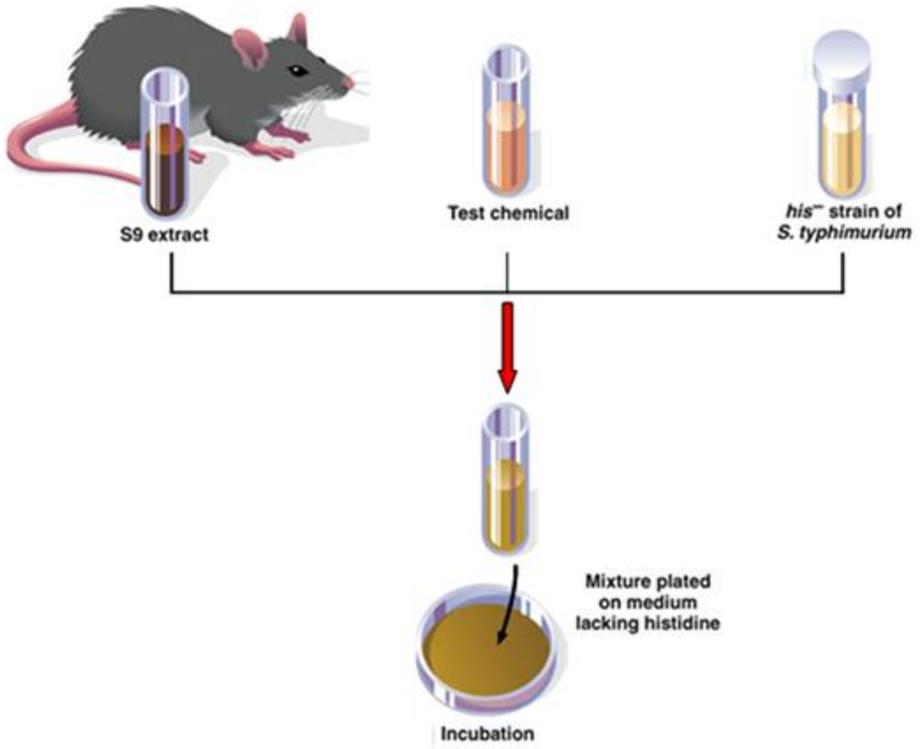


Figure 6.3. Cartoon representation of the plating of a compound at one concentration used to perform an Ames test. The figure represents the use of metabolic activation (+S9), without metabolic activation the S9 extract is replaced with sterile buffer.

Ames Study Manager, Report Generator and Sorcerer Systems (Perceptive Instruments Ltd.) were used to count the number of revertant colonies on plates and curate the raw data. The addition of a small amount of histidine to the top agar allows the plated bacteria to undergo between six and eight cell divisions before the histidine is depleted. If the test compound is mutagenic, the revertant bacteria continue to grow in the absence of the histidine and are visible on the plates. The revertant colonies are easily scored against the hazy looking background lawn which is made up of the histidine-dependent bacteria. Examination of the background lawn reveals if the test compound is toxic to the bacteria. A "thinning" or complete absence of the background lawn compared to the negative controls implies that the compound is toxic. Cell toxicity and test compound precipitation are recorded after being observed with the naked eye and use of a microscope. If precipitation is observed, that plate is counted but no other higher concentrations are considered as the availability of the test compound to the bacteria is unknown. The mutagenic activity is described by an induction factor (IF), which is the quotient of the revertants per plate at each concentration divided by the average number of revertants per plate of the negative controls. In strains TA98 and TA100, substances with an IF of two or greater, in conjunction with a dose-response curve, are considered to be Ames-positive in the test strains used. Dose-response curves can be found in Appendix E for all the Ames tests we performed.

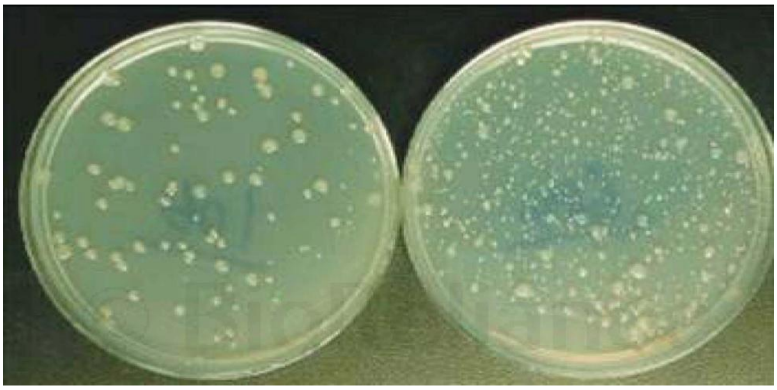


Figure 6.4. A photograph of two plates, after incubation, that have been prepared in the Ames test. The plate on the left is a negative control. The plate on the right is dosed with a chemical that has caused a > two-fold increase compared to the control. The revertant colonies are clearly visible.

6.3 Results and Discussion

The unpurified substances were 94.6 to 99.9% pure according to HPLC-LCMS increasing to 96.8 to 100% after recrystallisation except in four cases. For unknown reasons, the purity deteriorated for the free base of 3-aminobenzonitrile and the hydrochloride salt of 2,4,5-trimethylaniline.

Ames results were generated for all 14 raw compounds, for a subset of 9 purified samples and a slightly different subset of 9 molecules as the hydrochloride salt (Table 6.2). The anticipated transformation from an Ames-positive to an Ames-negative results occurred for only one AA. 4-aminoacetanilide in TA98 +S9 moved from just above the IF two-fold threshold (IF = 2.2) in the unpurified form to just below the threshold (IF = 1.6) as the purified free base. The Ames tests for unpurified and purified free base were repeated using water instead of DMSO (Table 6.3). The Ames results in DMSO were confirmed in water but the IF for the purified compound increased to 1.8. Nevertheless, the hydrochloride salt gave a clear positive response in DMSO (IF = 3.0) and in water (IF = 2.6) therefore, this AA must still be considered Ames-positive.

Small changes in the induction factor in either direction are observed in several cases. The IF (TA100 +S9) for 2,4,5-trimethylaniline diminished from 11.3 to 8.3 from raw (99.8%) to purified (100%) material. Most drastically, aminobenzonitrile's IF (TA98 +S9) raised from 5.0 to 10.9 due to a deterioration in purity from raw (98.2%) to "purified" (95.5%) compound. This amine also undergoes a qualitative shift, i.e. from Ames-negative to Ames-positive (TA98 -S9; IF = 2.3) albeit only by a marginal amount. The IF's for the most genotoxic substance in our dataset, 2-aminofluorene, doubled with and without metabolic activation from raw to purified form despite almost identical purities of 99.4 and 99.7%, respectively. For several unpurified AAs the Ames tests on both strains were repeated in duplicate (the IF values are shown in parentheses in Table 6.2) and the results were remarkably similar.

The 14 AAs can be pooled into 3 groups (Table 6.2): I) those Ames-positive with and without metabolic activation, II) those strongly Ames-positive +S9 and III) those marginally Ames-positive +S9. Remarkably, the IF for the latter group of five amines changed by less than one unit across the Ames tests for all four forms, i.e. unpurified and purified compound, as hydrochloride salt in DMSO and in water. This observation is in sharp contrast to the variations in outcome for some of the AAs mentioned in the Introduction and can probably be attributed to differences in their tendencies for structural deterioration.

Table 6.2. The IF's in the Ames test for 14 AAs (the highest IF observed and independent of the concentration is given); N denotes a negative test.

Compound name	Unpurified				Purified				HCl Salt DMSO				HCl Salt Water			
	TA98		TA100		TA98		TA100		TA98		TA100		TA98		TA100	
	-S9	+S9	-S9	+S9	-S9	+S9	-S9	+S9	-S9	+S9	-S9	+S9	-S9	+S9	-S9	+S9
I 2-aminofluorene	10.8[a]	42.8[a]	N[a]	13.0[a]	23.7[a]	82.0[a]	3.6[a]	19.0[a]	-	-	-	-	-	-	-	-
I 6-aminochrysene	35.5[a] (62.0)[b,c]	8.2[a] (11.6)[b,c]	9.6[a] (13.9)[b,c]	N[a] (10.6)[b,c]	-	-	-	-	-	-	-	-	-	-	-	-
II 2-aminoanthracene	N[a]	27.7[a]	N[a]	4.1[a]	N[a]	29.6[a]	N[a]	5.0[a]	-	-	-	-	-	-	-	-
II 1-methyl-2-aminobenzimidazole	N	24.6 (34.5)[c]	N	14.6 (13.0)[c]	N	34.6	N	15.6	N	30.1	N	13.6	N	27.6	N	15.0
II 4-phenoxyaniline	N[a]	21.2[a]	2.7[a]	16.7[a]	-	-	-	-	-	-	-	-	-	-	-	-
II 2-amino-5-phenylpyridine	N	14.2	N	2.8	-	-	-	-	N[a]	17.2[a]	N[a]	2.4	N[a]	14.5[a]	N[a]	N[a]
II 2,4,5-trimethylaniline	N	4.1 (5.0)[c]	N	11.3 (10.0)[c]	-	4.6	-	8.3	-	-	-	-	-	-	-	-
III 3-aminobenzonitrile	N	5.0 (5.6)[c]	N	N	2.3	10.9	N	N	N	4.7	N	N	N	6.8	N	N
III 2-aminonaphtho(2,3-d)imidazole	N[a]	6.7[a]	N[a]	N[a]	N[a]	7.2[a]	N[a]	N[a]	-	-	-	-	-	-	-	-
III 4-chloro-o-toluidine	N[a]	2.6[a]	N	1.9[a]	-	-	-	-	N[a]	2.4[a]	N[a]	2.8	-	-	-	-
III 3-aminoquinoline	N	1.9	N	3.0	-	-	-	-	N	2.0	N	2.7	N	2.1	N	3.3
III 2-methyl-4-bromoaniline	N[a]	N[a]	N[a]	2.4[a]	N[a]	N[a]	N[a]	2.0[a]	N	N	N	2.6	N[d]	2.1[d]	N[d]	3.0[d]
III 4-aminoacetanilide	N	2.2	N	N	N	N	N	N	N	3.0	N	N	N	2.6	N	N
III 2-amino-5-hydroxybenzoic acid	N	2.1	N	N	N	3.2	N	N	N	2.0	N	N	N[d]	2.1[d]	N[d]	N[d]

[a] The Ames test was limited by toxicity or precipitation and so not all dose levels were counted. [b] Toxicity was observed at all doses +S9 in the initial Ames test, hence the test was repeated at 1.5, 5, 15, 50, 150, 500 µg/plate. [c] Values in brackets correspond to repeat experiments in order to investigate the reproducibility. [d] Compound precipitated out of solution in the initial Ames test so it was repeated at 15, 25, 50, 150, 500, 1500 µg/plate.

Table 6.3. The IF in the Ames test for 4-aminoacetanilide tested in water as the unpurified and purified.

Compound name	Unpurified in Water				Purified in Water			
	TA98		TA100		TA98		TA100	
	-	+	-	+	-	+	-	+
III 4-Aminoacetanilide	N	2.4	N	N	N	N	N	N

These findings can be considered as an endorsement for the reliability of the Ames test for this compound class. However, there are several caveats to this statement:

- The variation in purity remained within a narrow band of 95 – 100% as determined by LCMS
- We did not systematically investigate the relationship between compound purity and the reproducibility of the Ames test, which could be very chemotype-dependent
- We only tested a small set of AAs, whose structures are probably not wholly representative of this compound class
- We employed only two strains out of the five required for regulatory acceptance of the Ames results

Nevertheless, the consistency of the IF values across different assay conditions for the same bacterial strain is remarkable and even more so, considering the variety of carbo- and heterocycles, of mono- and multicycles and of substitution patterns with electron-donating and –withdrawing groups used. Furthermore, the formation of the hydrochloride salt did not appear to influence the magnitude of the Ames response in the T98 and T100 strains which provides a strategy to purifying these types of compounds when other techniques fail.

We subsequently had the stability of 73 low molecular weight AAs determined in DMSO at room temperature for one week. We found that nine molecules degraded to a significant degree (12 – 42%). This emphasises the requirement for this compound class to be kept under appropriate storing conditions and for the Ames test only to be conducted when the purity levels are known in order to prevent spurious outcomes as reported in the Introduction.

Following on from our findings above, a project started at GSK to systematically perform the five-strain Ames test on low molecular weight AAs to assemble a reference set for Safety Assessment and Genetic Toxicology. This collection has currently grown to around 400 molecules and is known from here on as the GSK data set. The 200-strong Ames-negative compound set is of particular interest to chemists as building blocks for chemical synthesis. Preliminary work has been undertaken to investigate QSAR equations to discriminate between positive and negative anilines using the 400 molecule data set. This work is reported in the following section.

6.4 Predictive Toxicology for Aromatic Amines

As part of the data preparation for this investigation, the Ames classifications of the 400 GSK AAs were compared to those published in the literature. The AA subset of the Organon data set[223] consists of 258 molecules[183], after filtering, whereas 674 AAs are listed in the comprehensive Bayer compilation of approximately 6500 chemicals tested in the Ames test[198]. In the overlap of 136 compounds with the GSK data set, only seven were classified differently which could be explained by the inclusion of an equivocal category in the GSK classifications. This was encouraging considering the interlaboratory difference previously discussed. It should be noted that data from multiple Ames tests for the same compound can exist in the literature and can be contradictory. If a judgement is being made on multiple contradictory literature results, then the classification decided can be based upon the individuals' interpretation of the result which can be limited by their experience. Experts within GSK analyse all the available data and make a judgement based on the protocols used for the Ames tests under question (i.e. if regulatory guidelines were adhered to). For a compound to be classified as negative, it must have been tested in at least the five bacterial strains required by the regulators, the test must have been conducted according to regulatory guidelines and the molecule must have displayed an unequivocal negative response in all test strains. For a compound to be classified as positive, it must have shown a clear positive response in at least one of the bacterial strains. If a compound has only been tested in two of the required bacterial strains and was negative, then this does not mean the compound is considered Ames-negative as the data would be considered inadequate because it does not comprise of a full Ames test. If the external data is inadequate, or does not exist, then the compound is tested in-house and categorised according to GSK's protocol. Therefore, the GSK AA data set contained a combination of classifications based on reliable literature/external and in-house data, with a slightly larger proportion based on internal data. In this work we filtered and separated the compounds according to Figure 6.5.

After visual inspection of the 273 substituted anilines we removed 4 large drug-like compounds and one iodine containing compound. We focussed on the subset of 234 substituted anilines (known from here as the aniline data set) with only one primary amine (NH_2) group attached to the aromatic ring to simplify the modelling. The 234 anilines were grouped into seven different categories depending on their Ames classification (Figure 6.5)

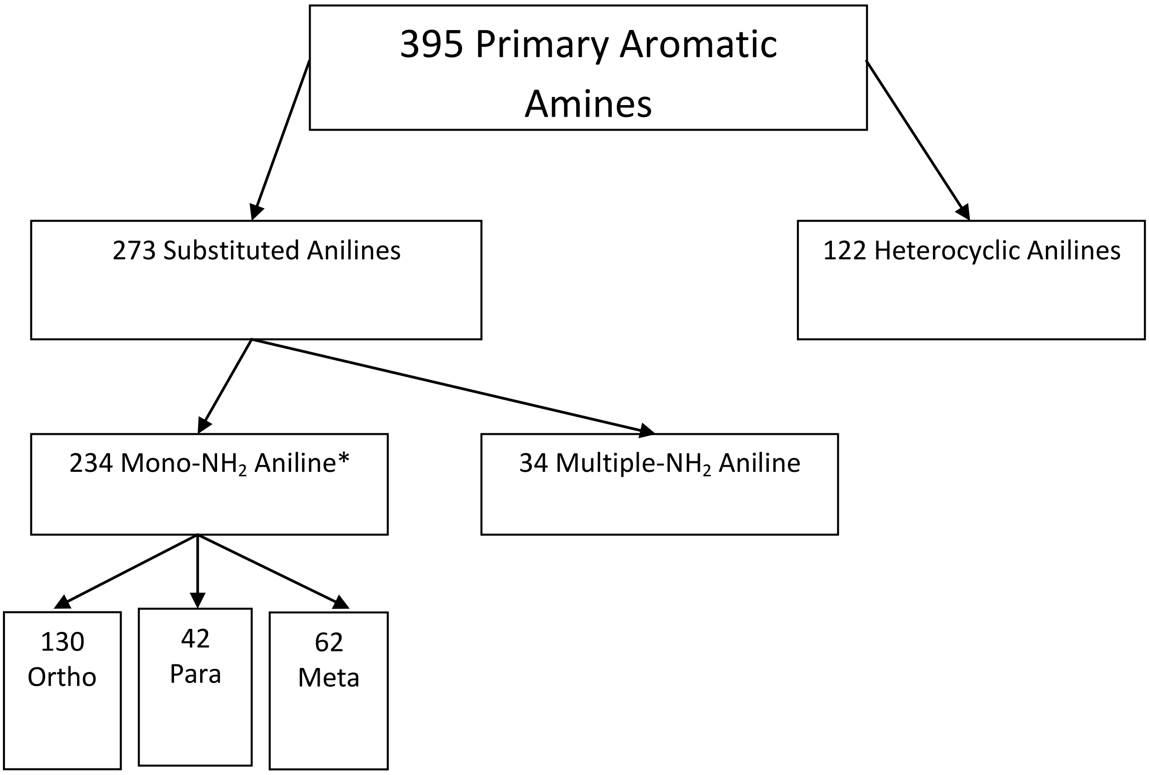

395 Primary Aromatic Amines
273 Substituted Anilines
122 Heterocyclic Anilines
234 Mono-NH$_2$ Aniline*
34 Multiple-NH$_2$ Aniline
130 Ortho
42 Para
62 Meta

Figure 6.5. The structural criteria we applied to separate the GSK data set.* 4 large drug-like compounds and one iodine containing compound removed.

Table 6.4. The classification of the 234 anilines in the GSK data set.

Ames Classification					
+S9	-S9	All	Ortho	Meta	Para
Negative	Negative	105	66	21	18
Positive	Negative	89	39	32	18
Equivocal	Negative	2	2	0	0
Equivocal	Equivocal	3	2	1	0
Positive	Positive	28	16	7	5
Negative	Positive	3	3	0	0
Positive	No Data	4	2	1	1

It can be seen from Table 6.4 that the data set contained an approximate 50:50 split between Ames-positive with metabolic activation (+S9) and Ames-negative with and without metabolic activation (+/-S9). These two categories agree with the recognised mechanism of metabolic activation of AAs that causes them to become mutagenic or not. The category that contained the next greatest number of compounds was Ames-positive with and without metabolic activation (+/-S9). We previously highlighted two compounds (2-aminofluorene and 6-aminochrysene) that also displayed these results for the Ames tests in bacterial strains TA98 and TA100 that we performed. It is known that in addition to DNA adduct formation through the activation of AAs by P450 1A2, intercalation into the DNA without the formation of a covalent bond is another mechanism that can lead to genotoxicity and therefore to an Ames-positive result +/-S9[183]. With the aim to avoid additional mechanisms of genotoxicity, we focussed on the compounds that are thought to act through CYP1A2 activation. The final data set contained 105 ortho-, 53 meta-, and 36-para substituted anilines.

Our motivation for this work stemmed from the encouraging results obtained for the discrimination of Ames classifications for AAs using descriptors generated from quantum mechanics[183, 221]. For the 194 anilines, we calculated the reaction energies for the equations that display the largest discrimination for AAs according to Leach *et al.*[221] discussed in Chapter 5 and represented in Equation 6.1 and Equation 6.2. We also calculated the relative stability of the nitrenium ion, using aniline as the reference, according to the equation utilised by Bentzien *et al.*[183] (Equation 6.3). To make a fair comparison, the geometries were optimised at the (U)B3LYP/6-31G* level of theory. Bentzien *et al*. suggested that the AM1 semiempirical level was a good compromise between accuracy and speed, however, as Leach *et al*. obtained their results at the (U)B3LYP/6-31G* level we opted for the higher level of theory. A single conformation was generated for all structures using the LigPrep[251] application in Schrödinger's Maestro[252] modelling platform. The structures were visually inspected and changed if they were obviously not the lowest energy conformers. Geometry optimisation was then performed using GAUSSIAN03[48]. All the equations require the syn and anti conformations of the nitrenium ion to be considered. We optimised both conformers and the lowest energy conformations were used in the calculations, which is suggested by the authors from both publications.

$$ArNHOH + H_3O^+ \rightarrow ArNH^+ + 2H_2O \qquad \text{Equation 6.1}$$

$$ArNHOAc \rightarrow ArNH^+ + AcO^- \qquad \text{Equation 6.2}$$

$$ArNH_2 + PhNH^+ \rightarrow ArNH^+ + PhNH_2 \qquad \text{Equation 6.3}$$

Equation 6.1 is considered as a surrogate reaction to represent the energy changed for the combined process of protonation and deprotonation of the hydroxylamine to react with DNA (Figure 5.3). Equation 6.2 was used by Leach *et al*. to model the formation of the nitrenium ion from the N-acetoxy or N-sulphate esters, formed in second phase metabolism of the hydroxylamine by N-acetyltransferase and sulphotransferase enzymes (Figure 5.3)[253]. The origins of Equation 6.3 were discussed in Chapter 5. The relationship between the reaction energies calculated from the three equations was investigated. The results are shown in the correlation matrix in Table 6.5.

Table 6.5. Correlation matrix of the energies calculated from Equation 6.1, Equation 6.2 and Equation 6.3.

	Equation 6.1	Equation 6.2	Equation 6.3
Equation 6.1	1.000	0.995	0.995
Equation 6.2	-	1.000	0.989
Equation 6.3	-	-	1.000

The correlation matrix showed the energies were highly correlated and so would discriminate between the anilines virtually identically. Because of this reason we only considered the results obtained from Equation 6.3 from here.

We discussed in Chapter 5 that the classification of Ames-positive and negative correlates with the energy calculated from Equation 6.3 which uses aniline as a reference. Bentzien *et al*.[183] classified an AA as Ames-negative if the ΔΔ*E,* calculated from Equation 6.4, was positive i.e. the nitrenium ion was less stable than that of aniline and vice versa. This method makes the assumption that aniline, as the parent compound, is the most stable of the Ames-negative AAs. Bentzien *et al*. observed a significant improvement in correct predictions when an uncertainty interval of $\Delta\Delta E$ = ±21 kJmol^{-1} around the reference energy (0 kJmol^{-1}) was introduced, where any AA with a calculated energy between this interval was not considered.

$$\Delta\Delta E = \Delta E_{ArNH^{+}} + \Delta E_{PhNH_2} - \Delta E_{ArNH_2} + \Delta E_{PhNH^{+}} \qquad \text{Equation 6.4}$$

We employed a slightly different approach to discriminate between the anilines. We only considered the meta- and para-substituted anilines (89 anilines in total) that were classified Ames negative +/- S9 (39 anilines) and Ames positive +S9 only (50 aniline). The 89 meta-/para-substituted anilines had a variety of electron-withdrawing and electron donating substitutions and included fused rings. The Ames-positive anilines were assigned a value of 1 and the Ames-negative a value of -1. Using SIMCA-P[54] we constructed a discrimination model using ΔΔ*E*. The model assigned a value between -1 and 1 to the anilines. Subsequently, any aniline with a negative value is considered to be predicted as Ames-negative and any aniline with a positive value Ames-positive. We did not apply an uncertainty interval. We report the results in the 2 x 2 confusion matrix, also known as a truth table (Table 6.6). The term sensitivity is defined as the ratio of true positives divided by all experimental positives. Similarly, specificity is the ratio of true negatives divided by all experimental negatives and accuracy is defined as true positives summed with true negatives divided by the sample size.

Table 6.6. Truth table for Ames Prediction of the 89 meta- and para-substituted anilines in the GSK data set using the nitrenium ion stability hypothesis.

	Predicted Ames Positive	Predicted Ames Negative
Ames Positive Experimental	40 (50)	10
Ames Negative Experimental	14	25 (39)
Accuracy 0.73, Sensitivity 0.80, Specificity 0.64		

The accuracy of 0.73 is slightly lower compared to that (0.85) reported by Bentzien *et al.*[183] . However, the sensitivity and specificity were similar. We subsequently used these results as a benchmark to compare our model discussed below.

In Chapters 3 and 4, we reported the use of BCP properties in predicting pK_a values for phenols, carboxylic acids and anilines. Compound pK_a values are dependent on the stability of the related ion in solution. The stability of the nitrenium ion of parent anilines has been determined as important in discriminating between Ames-positive and Ames-negative compounds. In Chapter 4, we found that the C-N bond length can be used to predict the pK_a of anilines. In Chapter 5 we found that the electron density at the C-N BCP can be linked to the difference in mutagenic potency of AAs. Accordingly, we investigated if the electron density at the C-N BCP in the data set of 89 meta- and para-substituted anilines could discriminate between the Ames-positive and Ames-negative compounds. Our work and findings are reported below.

The electron densities at the C-N BCPs of the anilines were calculated from the previously (U)B3LYP/6-31G* optimised structures using MORPHY98[49]. In SIMCA-P, $\Delta\Delta E$ from the model discussed above, was replaced with the corresponding electron density. The results from the reconstructed model are shown in Table 6.7 and graphically in Figure 6.6 .

Table 6.7. Truth table for Ames Prediction of the 89 meta- and para-substituted anilines in the GSK data set using the electron density at the C-N BCP.

	Predicted Ames Positive	Predicted Ames Negative
Ames Positive Experimental	41 (50)	9
Ames Negative Experimental	17	22 (39)
Accuracy 0.70, Sensitivity 0.82, Specificity 0.56		

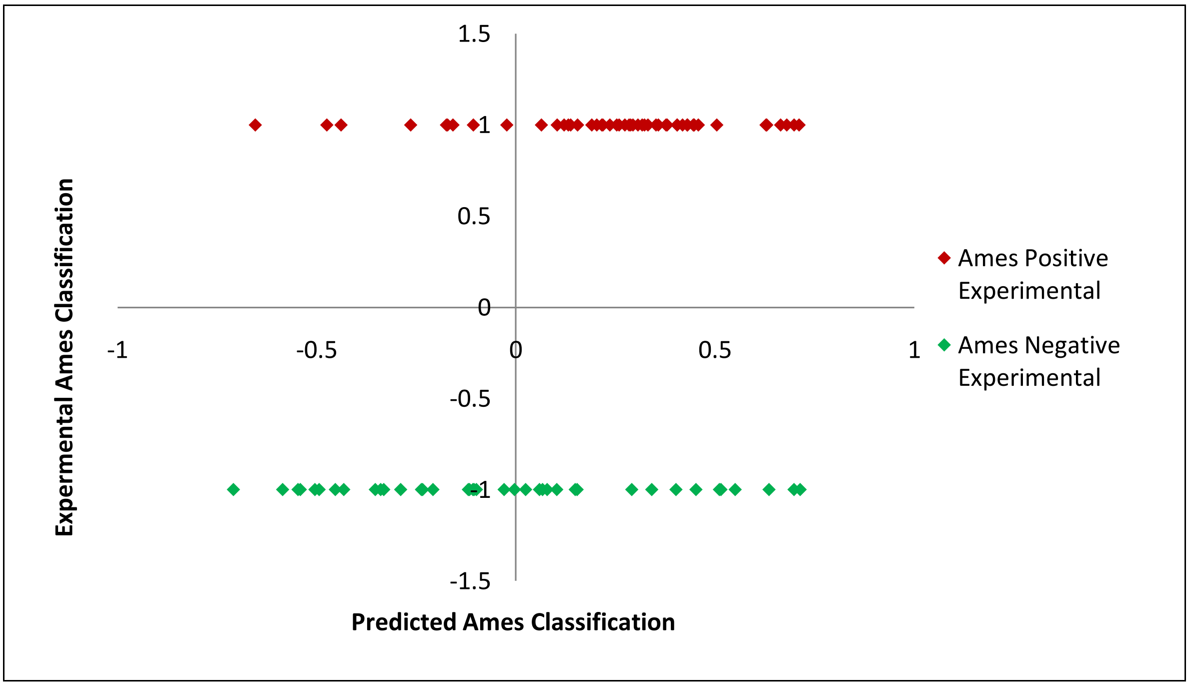


Figure 6.6. A graphical representation of the discrimination of the 89 meta- and para-substituted Ames-positive and Ames-negative anilines using the electron density at the C-N BCP.

According to the model, a relatively lower electron density at the C-N BCP corresponds to Ames – positive anilines and conversely a relative higher electron density at the C-N BCP corresponds to Ames-negative anilines. The accuracy of 0.70 and sensitivity of 0.82 is comparable to that obtained using Δ*ΔE*, where an accuracy and sensitivity of 0.73 and 0.80 were obtained, respectively. However, the specificity of 0.56 is lower than that obtained using Δ*ΔE* (0.64). The specificity can be increased by changing the 0 threshold (X-axis in Figure 6.6) used to discriminate between Ames-positive and Ames-negative, although it would be at the expense of the sensitivity. In relation to the predictions of Ames classification for AAs, it is a difficult choice whether improved sensitivity or specificity is desirable. Increasing sensitivity means that more Ames-positive compounds are predicted correctly, therefore there are fewer predicted false-negatives. However, this is at the expense of the specificity and so there will be an increase in number of false-positive predictions. A false-negative prediction (wrongly identified as an Ames-negative compound) could lead to a compound being used in drug design that when tested later in the Ames test is found to be positive. A false-positive prediction (wrongly identified as Ames-positive compound) could lead to a useful compound being disregarded. The desired levels of sensitivity and specificity can depend on what stage in drug design the prediction is being made and how it is applied[250]. The advantage of using one descriptor (i.e. the electron density at the C-N BCP) to discriminate between the anilines is that it allows chemists to target this property when considering using meta- or para-substituted aniline building blocks. A relatively low electron density at the C-N BCP should be avoided. Furthermore, using just the electron density at the C-N bond requires one geometry optimisation compared to the two required to calculate Δ*ΔE*.

The mediocre accuracy of the model created using the electron density at the C-N bond was disappointing considering meta- and para-substituted anilines had only been used. To understand false predictions, clearly visible in Figure 6.6, we investigated the individual data points. We start with the nine false-negative predictions in turn (Figure 6.7).

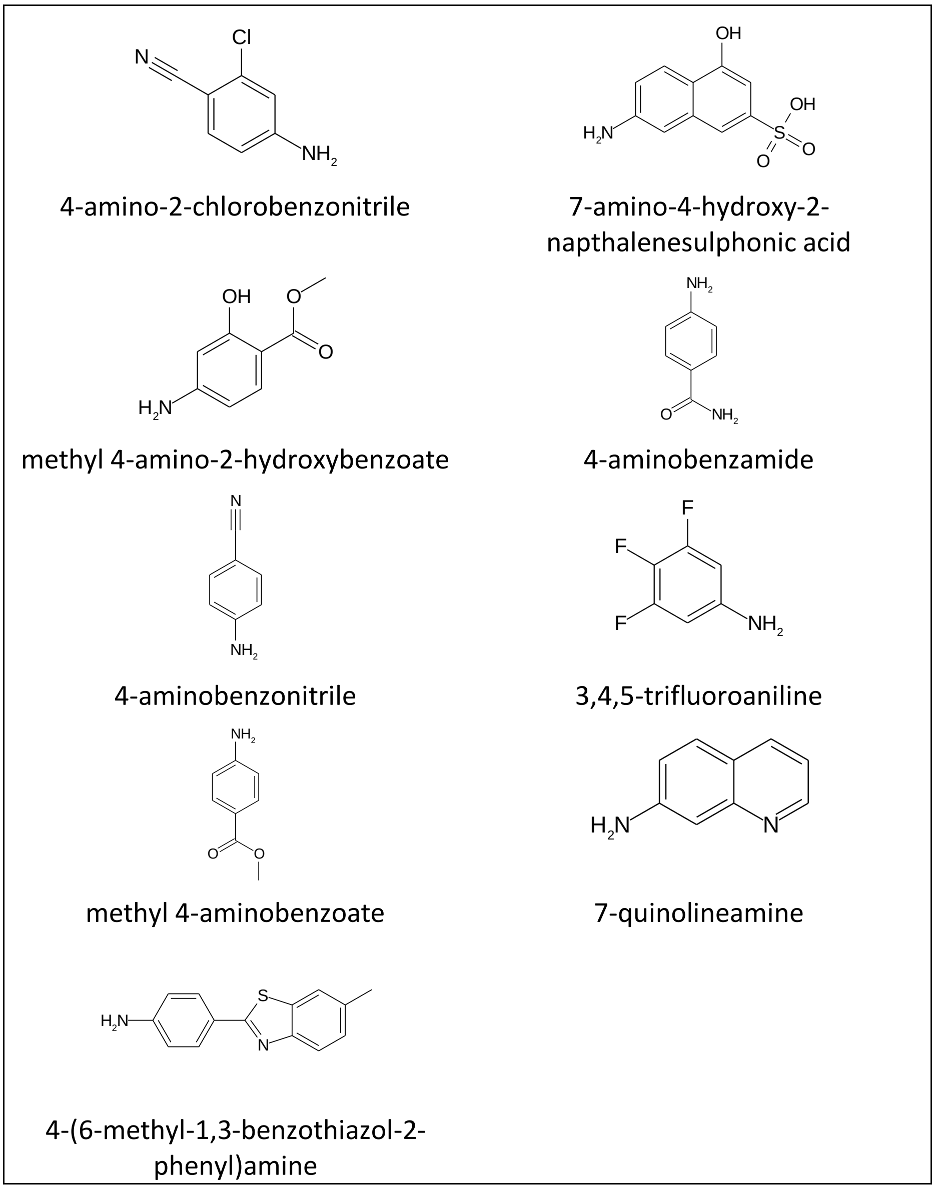
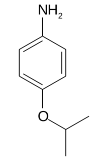
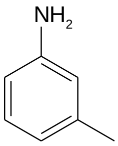
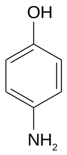
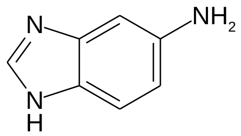
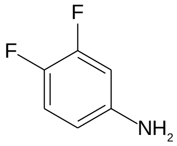
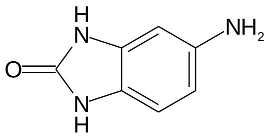
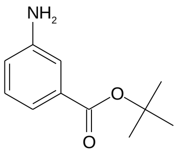
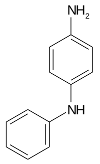
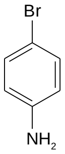
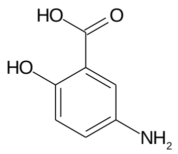
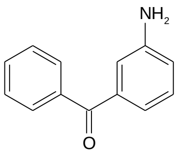
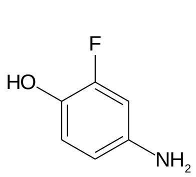
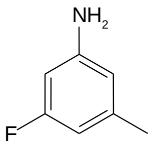
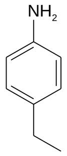

4-amino-2-chlorobenzonitrile
7-amino-4-hydroxy-2-napthalenesulphonic acid
methyl 4-amino-2-hydroxybenzoate
4-aminobenzamide
4-aminobenzonitrile
3,4,5-trifluoroaniline
methyl 4-aminobenzoate
7-quinolineamine
4-(6-methyl-1,3-benzothiazol-2-phenyl)amine

Figure 6.7. The structures and chemical names of the anilines classified as false-negatives using the electron density at the C-N BCP (Figure 6.6).

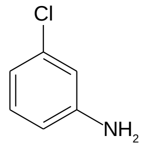
According to Figure 6.2 the vast majority of positive Ames results are detected in the bacterial strains TA98 and TA100, which is the reason we restricted our Ames testing to these two strains. There is a slim chance that any of nine compounds are Ames-positive in one of the other bacterial strains. The classification for 4-amino-2-chlorobenzonitrile was based on GSK data and was found to be positive in TA98 and TA100 +S9. The classification for methyl 4-amino-2-hydroxybenzoate was based on a GSK five-strain Ames test where a clear positive response (IF = 6.8 at 5000 µg/plate) was observed in TA98 +S9. 4-aminobenzonitrile was positive in TA98 +S9 (IF = 3.1 at 1600 µg/plate) based on GSK data. A different source[242] classified this compound as negative. However, the Ames test was not performed to the maximum concentration of 5000 µg of test compound/plate, therefore the former classification must stand. The classification for methyl 4-aminobenzoate was based on a positive response (IF = 3.8 at 5000 µg/plate) observed in TA98 +S9. 4-(6-methyl-1,3-benzothiazol-2-phenyl)amine was tested under the National Toxicology Programme in the US and positive responses were observed in both TA98 and TA100 +S9. Mixed external data[254] exists for 7-amino-4-hydroxy-2-napthalenesulphonic acid ranging from negative +S9 to positive +S9 in both TA98 and TA100. The mixed Ames results for 4-aminobenzamide have been discussed in the introduction and are shown in Figure 6.1. No explanation can yet be offered to explain these results so this compound must still be considered Ames positive. A clear positive response (IF = 6 at 1600 µg/plate) for 3,4,5-trifluoroaniline was observed in TA98 +S9. The origins of this compound displaying a positive response are unclear as similar structures, for example 2,3-difluoroaniline and 2,5-difluoro-4-bromoaniline are Ames-negative. 7-Quinolineamine was classified as positive +S9 however, there is literature evidence[255] that suggests this compound is Ames-positive in TA98 and TA100 +/- S9. Considering this result, 7-quinolineamine would have been filtered from our data set of 89 meta-/para-substituted anilines. Furthermore, quinoline itself is Ames-positive and so the origins of the positive result for 7-quinolineamine cannot be specifically attributed to the amine function[256]. Based on evidence presented above, we believe we have justification to remove at least half the predicted false-negative anilines from the data set.

Next we discuss the 17 false-positive predictions (Figure 6.8). In contrast to Ames-positive anilines, for a compound to be classified as negative it must have been tested in at least the five bacterial strains required by the regulatory and up to a maximum concentration of 5000 µg/plate, providing the test is not limited by toxicity or solubility. Explaining false-positive predictions is more difficult because in many of the Ames tests, no response is observed at all. 4-[(1-methylethyl)oxy]aniline has been tested by GSK in the standard Ames test and an adapted version (i.e. the preincubation Ames assay[180]) , which can be used when a dose-response is observed in the standard test but does not reach the two-fold threshold required for an Ames-positive classification. In the standard Ames test with 4-[(1-isopropyl)oxy]aniline an IF of 1.7 was observed at 5000 µg/plate. In the adapted test the number of revertants still did not reach the two-fold threshold but the test was limited by toxicity to 3000 µg/plate. 4-[(1-isopropyl)oxy]aniline has a structure similar to other compounds in the aniline data set that are correctly predicted as Ames-positive, for example 4-methoxyaniline and 4-ethoxyaniline. However, according to the criteria for determining Ames classification, this compound must be considered negative. The false-positive prediction for 4-amino-2-fluorophenol may be explained by the maximum concentration counted in the Ames test being limited by toxicity. For the remaining compounds the Ames test results (mainly based on GSK and NTP data) all displayed clear negative responses.

 4-[(1-isopropyl)oxy]aniline	 3-methylaniline
 4-aminophenol	- Proprietary Structure
 1H-benzimidazol-5-amine	 3,4-difluoroaniline
 5-amino-1,3-dihydro-2H-benzimidazol-2-one	 1,1-dimethylethyl 3-aminobenzoate
 (4-aminophenyl)phenlamine	 4-bromoaniline
 2-amino-2-hydroxybenzoic acid	 (3-aminophenyl)(phenyl)methanone
 4-amino-2-fluorophenol	 3-fluoro-5-methylaniline
	-

4-ethylaniline

Proprietary Structure



3-chloroaniline

Figure 6.8. The structures and chemical names of the anilines classified as false-positives using the electron density at the C-N BCP.

The origin of the experimental data for ten of the most extreme outliers in either classification was analysed and three predicted false-negative compounds (i.e. 4-aminobenzonitrile, methyl 4-aminobenzoate, and 3,4,5-trifluoroaniline) were selected for Ames retesting in the most indicative bacterial strain (TA98) for AAs (Figure 6.2). These three compounds had previously tested negative in TA100 +/-S9 and positive in TA98 +S9 but the purity was unknown. Their purities were confirmed to be 100% by LCMS. 4-aminobenzonitrile displayed a clear positive response but methyl 4-aminobenzoate and 3,4,5-trifluoroaniline were negative in TA98 +S9 (Figure 6.9). The experimental results used to construct our model were corrected for these two compounds and 4-aminobenzamide and 7-quinolineamine were removed based on the discussions above. The improved statistics for the new model are shown in Table 6.8.

Table 6.8. Updated truth table for Ames Prediction of the 87 meta- and para-substituted anilines in the GSK data set using the electron density at the C-N BCP.

	Predicted Ames Positive	Predicted Ames Negative
Ames Positive Experimental	40 (46)	6
Ames Negative Experimental	13	28 (41)
Accuracy 0.78, Sensitivity 0.87, Specificity 0.68		

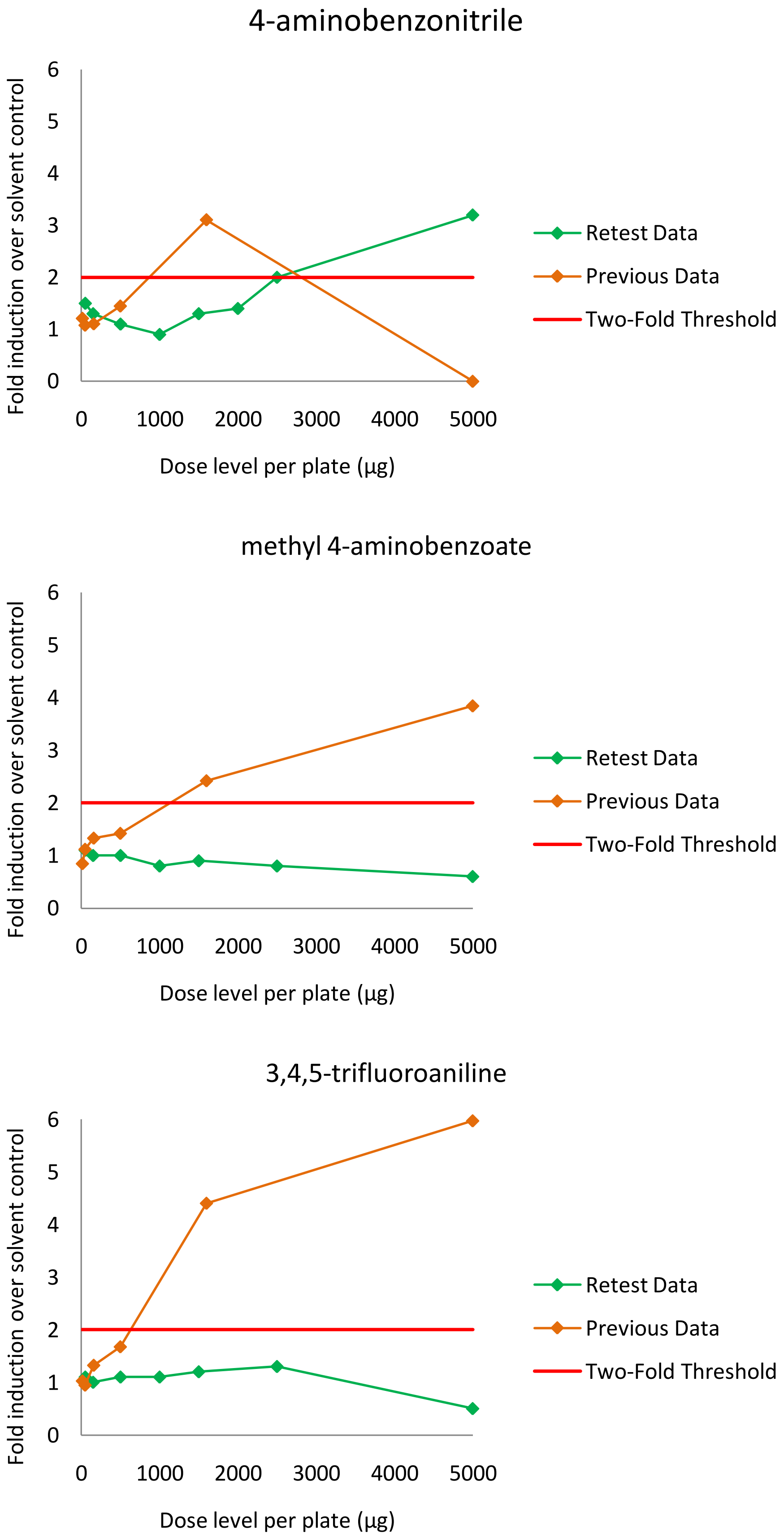


Figure 6.9. Does-response curves for the original Ames data and the retest data for 4-aminobenzonitrile, methyl 4-aminobenzoate and 3,4,5-trifluoroaniline in bacterial strain TA98 +S9.

It is worth mentioning here a note of caution when collecting Ames results for AAs from the literature, although this did not apply to the meta- and para-substituted anilines we investigated. Many AAs were tested under the National Toxicology programme (NTP) in the US. Their criteria for determining Ames classification is based on statistically significant increases in the number of revertants compared to the negative controls and not an increase over a specified threshold. This means that for some compounds the classifications can differ, obviously affecting modelling attempts and results. Furthermore, many of the NTP Ames tests employed both rat and hamster S9 to mimic metabolic activation. It is known that the use of hamster S9 produces higher responses and therefore can be considered more sensitive[257]. Some AAs have been classified as Ames-positive based on a positive response being observed with hamster S9 only. However, current guidelines only require the use of rat S9. To produce models for the prediction of Ames results for AAs, it is therefore important to be careful when mixing results obtained with S9 extracted from different animal species where the classification differs.

6.5 Summary

We investigated the reproducibility of the Ames test for a data set of 14 AAs. We found that the small changes in purity, the protomer (i.e. free base and hydrochloride salt) tested and the solvent (i.e. DMSO or water) had only marginal influences on the IF's and did not result in a change of any Ames classification. These results formed the basis for the systematic Ames testing of synthetic building block AAs leading to a data set of 400 molecules. We subsequently extracted a set of 89 meta- and para-substituted anilines and constructed predictive models, using the electron density at the BCP or $\Delta\Delta E$ as the only descriptors, to discriminate between Ames-positive and Ames-negative compounds in this class. The statistics for the model compare favourably to similar methods published in the literature[183]. We were able to explain some of the false-negative and false-positive-predictions. These are currently being considered for repeat Ames tests. Three compounds have been retested and the model correctly predicted two of the compounds as Ames-negative. The considerations and difficulties in modelling AAs have been highlighted. We believe this method can be applied to different AAs subsets for the prediction of Ames test classifications.

Chapter 7
Conclusions and Future Work

The use of QCT descriptors has been extended to predict two different properties of interest to the industrial sponsor and the wider scientific community. We have constructed models for one of the largest data sets so far used for QTMS applications to predict pK_a values for carboxylic acids. It was not surprising that the SVM statistical learning method performed the best but it was unexpected that the linear PLS method would provide very similar CV statistics. It was advantageous that the less CPU intensive HF/6-31G(d) level of theory provided comparable results to the more demanding B3LYP/6-311G(2d,p) level of theory. The predicted pK_a values from the models constructed from the 228 carboxylic acids compared favourably to commonly used pK_a prediction tools. The ortho-substituted benzoic acids were the least well predicted because QCT descriptors mainly account for the electronic contributions to the predicted property and do not fully capture steric effects.

To improve pK_a predictions for ortho-substituted compounds, we used data sets containing benzoic acids, phenols and anilines. We considered the use of *ab initi*o bond lengths exclusively as descriptors to predict pK_a. The aim was to investigate their effectiveness in pK_a prediction but we also focused on comparing single-bond-length models and all-bond-length models. *Ab intio* bond lengths can be extracted directly from the optimised geometries and do not require a further programme, such as MORPHY, to calculate the BCP properties. The results indicate that single-bond-length, compound-class specific models can be used to predict the pK_a of meta-/para-substituted compounds but this is not the case for ortho-substituted compounds. However, we identified high-correlation subsets that were able to accommodate the steric effects specific only to ortho-substituted compounds. These high-quality models provided us with the confidence to successfully challenge the assigned experimental pK_a values of compounds that were outliers. It is remarkable that models constructed from a single bond length are able to accurately predict the pK_a for a set of drug compounds used as a test set.

Extending the use of QCT descriptors for biological property prediction, we investigated their application to predict the mutagenicity of carbocyclic and heterocyclic aromatic amines. This class of compounds is synthetically very useful to medicinal chemists involved in drug design but they are often used with trepidation because of the large percentage of genotoxic compounds found in this group of chemicals. Surprisingly, the comprehensive literature on the prediction of the mutagenic potency, measured in the Ames test, for these compounds is divided over which properties are important.

Comprehensive investigations into the prediction of mutagenic potencies taken from the literature produce a number of models but these failed to classify a test set of GSK Ames data. It became apparent that contradictory Ames test results existed for several compounds; something that is problematic to modelling methods. The systematic experimental investigation of AAs in the Ames test confirmed the reproducibility of this test but did not account for interlaboratory variations. We questioned whether the prediction of mutagenic potency using literature data of variable quality is possible or even worthwhile in relation to the use of such a model at GSK. A recent, carefully constructed list at GSK provides Ames test results from external and internal data for approximately 400 aromatic and heterocyclic aromatic amines. The majority of the Ames-positive compounds are only positive with metabolic activation. This agrees with the widely accepted mechanism involving metabolic activation of these compounds by CYP1A2 to reactive metabolites that invoke a genotoxic response. Still, for a considerable number of compounds, the Ames results are inconsistent with this mechanism. Furthermore, our P450 enzyme inhibition assay results suggest that other cytochrome P450s could further complicate the mechanism leading to genotoxicity.

In consideration of the above, a model was constructed using the electron density at the common Carbon-Nitrogen BCP of meta-/para-substituted anilines, which had Ames test results consistent with the accepted mechanism (i.e. negative without and positive with metabolic activation). The model correctly classified 70% of the anilines. Three outliers in the model were retested after establishing their purity levels. It was encouraging that two of the compounds turned out to be negative as opposed to the original positive outcome for molecules of unknown purity. When the classification for these compounds was updated and two compounds with a high potential for genotoxic mechanisms were removed, the model correctly classified 78% of the anilines.

The two properties predicted in this work required very different considerations. However, there were examples in both cases where models constructed using QCT descriptors were able to correctly challenge experimentally determined pK_a values and Ames classifications. QCT descriptors, calculated from solutions to the Schrödinger equation, have a strong physical basis but their calculations take longer than other popular descriptors. This work has demonstrated their successful extension to larger data sets and in conjunction with the ever increasing speed of modern computation their routine use in drug discovery is an exciting opportunity.

The suggestion and opportunities for future work are listed below.

- In Chapter 3 we discussed the prediction of pK_a for carboxylic acids using QCT descriptors. Beyond extending this work to other classes of compounds, QTMS needs to match experiment by taking into account the nature of different tautomers present.
- The QTMS method can be extended to predict pK_a values for multiprotic compounds. Multiple ionisation centres have complex effects on the ionisation of a particular group. Hence for many of the published methods, results are limited to a series of monoprotic structures. However, as QCT descriptors are based on *ab initio* calculations, these effects will automatically be encapsulated in the wave function and thus the descriptors and therefore the final prediction. A similar method to that of Jelfs *et al.*[79], who developed an algorithm that applies multiple predictive models in a stepwise manner and reproduces the correct ionisation order for different groups within a compound, should be used.
- In Chapter 4 we demonstrated that a single *ab initio* bond length can be used to predict pK_a. In the case of ortho-substituted compounds, the consideration of conformation was important to model construction. The dependence of conformation on QCT descriptors should be fully investigated. This should involve returning to previous data sets used in QTMS analysis and investigating if unexplained outliers were caused by conformational differences to the rest of the data set. This analysis may also be able to explain the so called "active centre contamination" in VIP plots.
- Presently high-correlation subsets are visually identified. Different clustering methods should be explored to investigate if they choose the same compounds for the same high-correlation subsets or if less chemically intuitive, higher-correlation subsets are found.
- A direct comparison using single-bond-length models to other QCT descriptor models should be performed to understand the added value of using QCT descriptors.
- In Chapter 5 and Chapter 6 we discussed the collaborative work involving screeners, medicinal chemists, toxicologists, analytical and computational chemists, performed to purify a set of aromatic amines, screen them in the Ames test and use the data to generate predictive models. With the knowledge gained from this work, carefully designed studies should be carried out to answer the questions that remain open. These are discussed below.
- The models constructed to discriminate Ames outcomes for meta-/para-substituted anilines should be extended to account for all AAs. As more compounds are added to the GSK data set, these models can be continually tested and updated. Investigations into how the compounds in the data set need to be split into subsets and whether better models are constructed using high-correlation subsets should be performed.

- Direct comparisons between external and internal Ames test results should be made to fully appreciate the limits of using data from mixed sources for this class of compound. Models constructed using only data generated at GSK will avoid the problems associated with interlaboratory variations.
- P450 enzyme inhibition data should be collected for all compounds considered in modelling to exclude mechanisms other than DNA modifications via the nitrenium ion formation. Ames-positive +S9 compounds that do not interact with 1A2 are either inherently genotoxic or are metabolised by another P450 enzyme. Outliers identified in computational models should be excluded if strong interactions with P450 enzymes other than CYP1A2 are observed.
- To fully understand the influence of DMSO in the Ames test, the Ames-positive compounds that were found to degrade in DMSO should be systematically Ames-tested in another solvent. The degradation products should be identified to understand the mechanism involved and explain the Ames test outcomes.
- Recently, workers have proposed chemical models for the activation of aromatic amines[258, 259] to avoid the use of the S9 metabolic activating mix, which itself is toxic to the bacterial strains. The Ames results generated from such a model should be compared to the standard Ames test results. While such a system is not accepted by the regulators, it simplifies the mechanisms involved and may be useful in explaining the current difficulties in modelling Ames data for this class of compound.
- Ultimately, the generated models should be made available to the medicinal chemists so they can predict the likelihood of genotoxicity for an untested AA that they want to employ in synthesis or that appears as an impurity or metabolic product.

References

1. Bader, R. F. W., *Atoms in Molecules. A Quantum Theory.* Oxford Univ. Press: Oxford, Great Britain, 1990.

2. Popelier, P. L. A., *Atoms in Molecules. An Introduction.* Pearson Education: London, Great Britain, 2000.

3. Popelier, P. L. A., Quantum molecular similarity. 1. BCP space. *J.Phys.Chem.A* **1999,** 103, (15), 2883-2890.

4. WIKIPEDIA, List of Pharmaceutical Companies. *Available at <http://en.wikipedia.org/wiki/List_of_pharmaceutical_companies>, Accessed 21st August 2010*.

5. GSK, GSK Annual Review. **2009**.

6. GSK, GSK Annual Review. **2005**.

7. Adams, C. P.; Brantner, V. V., Estimating The Cost Of New Drug Development: Is It Really $802 Million? *Health Affairs* **2006,** 25, 420-428.

8. Lombardino, J. G.; Lowe III, A. J., The Role Of the Medicinal Chemist in Drug Discovery - Then and Now. *Nature Reviews* **2004,** 3, 853-862.

9. Brown, D.; Superti-Furga, G., Rediscovering the Sweet Spot in Drug Discovery. *Drug Discovery Today* **2003,** 8, 1067-1077.

10. Macarron, R., Critical Review of the Role of HTS in Drug Discovery. *Drug Discovery Today* **2006,** 11, 277-279.

11. Hajduk, P. J.; Greer, J., A decade of Fragment-Based Drug Design: Strategic Advances and Lesson Learned. *Nature Reviews* **2007,** 6, 211-219.

12. Astex Therapeutics, Astex Press Release. *Available at <http://www.astex-therapeutics.com/investorsandmedia/pressdetail.php?uid=111>, Accessed 22th August 2010*.

13. Jorgensen, W. L., The Many Roles of Computers in Drug Discovery. *Science* **2004,** 303, (1813-1818).

14. GSK, GSK Press Release. *Available at <http://www.gsk.com/media/pressreleases/2010/2010_pressrelease_10046.htm>, Accessed 20th August 2010*.

15. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development. *Adv. Drug. Deliv. Rev.* **2001,** 46, 3-26.

16. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002,** 45, 2615-2623.

17. Clark, D. E.; Picket, S. D., Computational Methods for the Prediction of Drug-Likeness. *Drug Discovery Today* **2000,** 5, 49-56.

18. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discovery Today* **2003,** 8, 876-877.

19. Gleeson, M. P., Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008,** 51, 817-834.

20. Sheridan, R. P., The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002,** 42, 103-108.

21. Ujvary, I., BIOSTER-A Database of Structurally Analogous Compounds. *Pesticide Science* **1997**, (51), 92-95.

22. Valler, M. J.; Green, D., Diversity Screening Versus Focussed Screening in Drug Discovery. *Drug Discovery Today* **2000,** 5, 286-293.

23. O'Brien, S. E.; Popelier, P. L. A., Quantum Molecular Similarity. Part 3: QTMS descriptors. *J.Chem. Inf. Comp. Sc.* **2001,** 41, 764-775.

24. Hohenberg, P.; Kohn, W., *Phys. Rev. B* **1964,** 136, B864.

25. Slater, J. C., *Int. J. Quantum Chem. Symp.* **1975,** 9, 7.

26. Painter, G. S., *Phys. Rev. B.* **1981,** 24, 4264.

27. Becke, A. D., *J. Chem. Phys.* **1996,** 104, 1040.

28. Perdew, J. P., *Phys. Rev. B* **1986,** 33, 8822.
29. Becke, A. D., *J. Chem. Phys.* **1993,** 98, 1372.
30. Perdew, J. P.; Wang, Y. W., *Phys. Rev. B* **1992,** 45, 13244.
31. Howard, S. T.; Lamarche, O., Description of covalent bond orders using the charge density topology. *J.Phys.Org.Chem.* **2003,** 16, 133-141.
32. Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E., Description of Conjugation and Hyperconjugation in Terms of Electron Distributions. *J.Am.Chem.Soc.* **1983,** 105, (15), 5061-5068.
33. Bader, R. F. W.; Preston, H. J. T., The Kinetic Energy of Molecular Charge Distributions and Molecular Stability. *Int.J.Quant.Chem.* **1969,** 3, 327-347.
34. Livingstone, D., *Data Analysis for Chemists*. 1 ed.; Oxford University Press: New York, 1995.
35. Hansch, C.; Fujita, T., A method for the correlation of biological activity and chemical structure. *J.Am.Chem.Soc.* **1964,** 86, 1616-1626.
36. Miller, B., *Advanced Organic Chemistry. Reactions and Mechanisms.* Prentice-Hall, New Jersey, USA.: 1998.
37. Johnson, C. D., *The Hammett Equation*. Cambridge University Press, Cambridge, GB, 1973.
38. Jaffe, H. H., *Chem.Rev.* **1953,** 53, 191.
39. Hansch, C.; Hoekman, D.; Leo, A.; Zhang, L.; Li, P., The expanding role of quantity structure-activity relationships (QSAR) in toxicology. *Toxicology letters* **1995,** 79, 45-53.
40. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical Reviews* **1996,** 96, 1027-1043.
41. Doweyko, A. M., QSAR: dead or alive? *J.Comp.Aided Molec.Des.* **2008,** 22, 81-89.
42. O'Brien, S. E. Quantum Molecular Similarity, an Atoms in Molecules Approach. PhD Thesis, Dept. of Chemistry, UMIST, Manchester, Great Britain, 2000.
43. Young, D.; Martin, T.; Venkatapathy, R.; Harten, P., Are the Chemical Structures in Your QSAR Correct? *QSAR & Comb.Sci.* **2008,** 27, 1337-1345.
44. Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A., Critical Assessment of QSAR Models of Environmental Toxicity Against Tetrahmena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Comput. Sci.* **2008,** 48, (1733-1746).
45. Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V., Combinatorial QSAR Modelling of Chemical Toxicants Tested Against Tetrahymena Pyriformis. *J. Chem. Inf. Comput. Sci.* **2008,** 48, 766-784.
46. Fourches, D.; Muratov, E.; Tropsha, A., Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Comput. Sci.* **2010,** 50, 1189-1204.
47. Schaftenaar, G.; Noordik, J. H., Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Design* **2000,** 14, 123-134.
48. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02,*, Revision C.02,
; Gaussian, Inc.,: Wallingford CT, , 2004.
49. Popelier, P. L. A. *MORPHY98*, MORPHY98 - a program written by P.L.A. Popelier with a contribution from R.G.A. Bone. UMIST, Manchester, England, 1998.

50. Popelier, P. L. A., A Robust Algorithm to Locate Automatically All Types of Critical- Points in the Charge-Density and Its Laplacian. *Chem.Phys.Lett.* **1994,** 228, (1-3), 160-164.
51. Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J., Quantum Topological Molecular Similarity. Part 5: Further Development with an Application to the toxicity of Polychlorinated dibenzo-p-dioxins (PCDDs). *J.Chem.Soc., Perkin II* **2002**, 1231-1237.
52. Wold, S.; Sjostrom, M.; Eriksson, L., Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encycl.of Comp.Chem.*, Schleyer, P. v. R., Ed. Wiley, Chichester, GB, 1998; Vol. 3, pp 2006-2021.
53. Wold, S., *Technometrics* **1978,** 20, 397-405.
54. UMETRICS *SIMCA-P 10.0*, info@umetrics.com: www.umetrics.com, Umeå, Sweden, 2002.
55. Golbraikh, A.; Tropsha, A., Beware of q2! *J.Molec.Graph and Modell.* **2002,** 20, 269-276.
56. Hawkins, D. M.; Basak, S. C.; Mills, D., Assessing model fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003,** 43, 579-586.
57. Chaudry, U. A.; Popelier, P. L. A., Estimation of pKa using Quantum Topological Molecular Similarity (QTMS) Descriptors: Application to Carboxylic Acids, Anilines and Phenols. *J.Org.Chem.* **2004,** 69, 233-241.
58. Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J., Quantitative Structure-Activity Relationships of mutagenic activity from quantum topological descriptors: triazenes and halogenated hydroxyfuranone (mutagen-X) derivatives. *J. Comp.-Aided Molec.Design* **2004,** 18, 709-718.
59. Popelier, P. L. A.; Smith, P. J., QSAR models based on Quantum Topological Molecular Similarity. *Eur.J.Med.Chem.* **2006,** 41, 862-873.
60. Alsberg, B. K.; Marchand-Geneste, N.; King, R. D., A New 3D Molecular Structure Representation using Quantum Topology with application to structure-property relationships. *Chemo. & Intell. Lab. Syst.* **2000,** 54, 75-91.
61. Alsberg, B. K.; Marchand-Geneste, N.; King, R. D., Modeling quantitative structure-property relationships in calculated reaction pathways using a new 3D quantum topological representation. *Analyt.Chim.Acta* **2001,** 446, 3-13.
62. Buttingsrud, B.; Ryeng, E.; King, R. D.; Alsberg, B. K., Representation of Molecular Structure Using Topology with Inductive Logic Programming in Structure-Activity Relationships. *J. Comput.-Aided Mol. Design* **2006,** 20, 361-373.
63. O'Brien, S. E.; Popelier, P. L. A., Quantum Topological Molecular Similarity. Part 4 : A QSAR study of Cell Growth Inhibitory Properties of Substituted (E)-1-Phenylbut-1-en-3-ones. *J.Chem.Soc., Perkin Trans. 2* **2002**, 478-483.
64. Mohajeri, A.; Hemmateenejad, B.; Mehipour, A.; Miri, R., Modelling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analysed by GA-PLS and PC-GA-PLS. *Journal of Molecular Graphics and Modelling* **2008,** 26, 1057-1065.
65. Roy, K.; Popelier, P. L. A., Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors. *Bioorg.Med.Chem.Lett.* **2008,** 18, 2604-2609.
66. Roy, K.; Popelier, P. L. A., Exploring Predictive QSAR Models Using Quantum Topological Molecular Similarity (QTMS) Descriptors for Toxicity of Nitroaromatics to Saccharomyces cerevisiae
QSAR & Comb.Sci. **2008,** 27, 1006-1012.
67. Chaudry, U. A.; Popelier, P. L. A., Ester Hydrolysis Rate Constant Prediction from Quantum Topological Molecular Similarity Descriptors. *J. Phys. Chem. A* **2003,** 107, 4578-4582.
68. Hemmateenejad, B.; Mohajeri, A., Application of Quantum Topological Molecular Similarity Descriptors in QSPR Study of the O-Methylation of Substituted Phenols. *J.Comp.Chem.* **2008,** 29, 266-274.
69. Kar, S.; Harding, A. P.; Roy, K.; Popelier, P. L. A., QSAR with Quantum Topological Molecular Similarity Indices: Toxicity of Aromatic Aldehydes to *Tetrahymena Pyriformis*. *SAR and QSAR in Environmental Research* **2010,** 21, 149-168.
70. Loader, R. J.; Singh, N. K.; O'Malley, P. J.; Popelier, P. L. A., The Cytotoxicity of ortho alkyl substituted 4-X-phenols: A QSAR based on theoretical bond lengths and electron densities *Bioorg.&Med.Chem.Lett.* **2006,** 16, 1249-1254

71. Selassie, C. D.; Verma, R. P.; Kapur, S.; Shusterman, A. J.; Hansch, C., QSAR for the cytotoxicity of the radical reaction. *J.Chem.Soc., Perkin II* **2002**, 1112-1117.

72. Avdeef, A., *Absorption and Drug Development: Solubility, Permeability and Charge State*. Wiley-Interscience: New Jersey, USA, 2003.

73. da Silva, C. O.; da Silva, J. B.; Nascimento, M. A. C., Ab initio calculations of absolute pK_a values in aqueous solution. 1. carboxylic acids. *J.Phys.Chem.A* **1999,** 103, 11194-11199.

74. Wells, J. I., *Pharmaceutical Preformulation*. Ellis Horwood Ltd: London, 1998, p. 25.

75. Comer, J.; Tam, K., *In Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*. Wiley-VCH: Weinheim: New York, 2001; pp 275-304.

76. Cookson, R. F., The determination of acidity constants. *Chem. Rev.* **1974,** 74, 5-28.

77. Namazian, M.; Halvani, S., Calculations of pK_a values of carboxylic acids in aqueous solution using density function theory. *J.Chem.Thermodynamics* **2006,** 38, 1495-1502.

78. Prankerd, D. D., *In Profiles of Drug Substances, Excipients, and Related Methodology.* Elsevier Academic Press: San Diego, CA, 2007; Vol. 1.

79. Jelfs, S.; Ertl, P.; Selzer, P., Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J.Chem. Inf. Model.* **2007,** 47, 450-459.

80. Lee, A. C.; Crippen, G. M., Predicting pK_a. *J. Chem. Inf. Model.* **2009,** 49, 2013-2033.

81. Lee, A. C.; Yu, J. Y.; Crippen, G. M., pK_a Prediction of Monoprotic Small Molecules the SMART Way. *J. Chem. Inf. Model.* **2008,** 2008, 2042-2053.

82. Milletti, F.; Storchi, L.; Goracci, L.; Bendels, S.; Wagner, B.; Kansy, M.; Cruciani, G., Extending pK_a Prediction Accuracy: High-Throughput pK_a Measurements to Understand pK_a Modulation of New Chemical Series. *European Journal of Medicinal Chemistry* **2010,** 45, 4270-4279.

83. Xing, L.; Glen, R. C., Novel Methods for the Prediction of logP, pK_a and logD. *J. Chem. Inf. Comput. Sci.* **2002,** 42, 796-805.

84. Xing, L.; Glen, R. C.; Clark, R. D., Predicting pK_a by Molecular Tree Structure Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003,** 43, 870-879.

85. Zhang, J.; Kleinoder, T.; Gasteiger, J., Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model* **2006,** 46, 2256-2266.

86. Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G., New and original pK_a prediction method using grid molecular interaction fields. *J. Chem. Inf. Model* **2007,** 47, 2172-2181.

87. da Silva, C. O.; da Silva, E. C.; M. A. C. Nascimento, Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution II. Aliphatic Alcohols, Thiols, and Halogenated Carboxylic Acids. *J. Phys. Chem. A* **2000,** 104, 2402-2409.

88. Eckert, F.; Diedenhofen, M.; Klamt, A., Towards a First Principles Prediction of pK_a: COSMO-RS and the Cluster-Continuum Approach. *Molecular Physics* **2010,** 108, 229-241.

89. Liptak, M. D.; Shields, G. C., Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian Models Combined with CPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* **2001,** 123, 7314-7319.

90. Liptak, M. D.; Shields, G. C., Experimentation with Different Thermodynamic Cycles Used for pK_a Calculations on Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *International Journal of Quantum Chemistry* **2001,** 85, 727-741.

91. Lu, H.; Chen, X.; Zhan, C. G., First-principles calculation of pK_a for cocaine, nicotine, neurotransmitters, and anilines in aqueous solution. *J. Phys. Chem. B.* **2007,** 111, 10599-10605.

92. Namazian, M.; Zakery, M.; Noorbala, M. R.; Coote, M. L., Accurate Calculation of the pKa of Trifluoroacetic Acid Using High-Level ab initio Calculation. *Chem. Phys. Lett.* **2008,** 451, 163-168.

93. Gruber, C.; Buss, V., Quantum-Mechanically Calculated Properties for the Development of Quantitative Structure-Activity Relationships (QSAR's). pK_a-Values of Phenols and Aromatic and Aliphatic Carboxlic Acids. *Chemosphere* **1989,** 19, 1595-1609.

94. Citra, M. J., Estimating the pK_a of Phenols, Carboxylic Acids and Alcohols From Semi-Empirical Quantum Chemical Methods. *Chemosphere* **1999,** 38, 191-206.

95. Gross, K. C.; Seybold, P. G., Substituent Effects on the Physical Properties and pK_a of Phenols. *Int. J. Quant. Chem.* **2001,** 85, 569-579.

96. Tehan, B. G.; Lloyd, E. J.; G., W. M.; Pitt, W. R.; Gancia, G.; Manallack, D. T., Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.-Act. Relat.,* **2002,** 21, (5), 473-485.

97. Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E., Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.,* **2002,** 21, 457-471.

98. Kogej, T.; Muresan, S., Database Mining for pK_a Prediction. *Curr. Drug Discovery Technol.* **2005,** 4, 221-229.

99. Soriano, E.; Cerdan, S.; Ballesteros, P., Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J.Molec.Struct (Theochem)* **2004,** 684, 121-128.

100. Ho, M.; Schmider, H.; Edgecombe, K. E.; Smith, V. H. J., Topological Analysis of Valence Electron Charge Distributions from Semi-empirical and ab initio methods. *Int.J. Quant.Chem., Quant. Chem. Symp.* **1994,** 28, 215-226.

101. Zhang, S.; Baker, J.; Pulay, P., A Reliable and Efficient First Principles-Based Method for Predicting pK_a Values. 1. Methodology. *J. Phys. Chem. A* **2010,** 114, 425-431.

102. Zhang, S.; Baker, J.; Pulay, P., A Reliable and Efficient First Principles-Based Method for Predicting pK_a Values. 2. Organic Acids. *J. Phys. Chem. A* **2010,** 114, 432-442.

103. Klamt, A.; Schuurmann, G. J., COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *J. Chem. Soc., Perkin Trans 2* **1993**, 799.

104. Jaguar version 6.0; Schrodinger, LLC: New York, NY, 2005.

105. Namazian, M.; Halvani, S., Calculations of pK_a Values of Carboxylic Acids in Aqueous Solution Using Moller-Plesset Perturbation Theory. *Journal of the Iranian Chemical Society* **2005,** 2, (1), 65-70.

106. Namazian, M.; Halvani, S.; Noorbala, M. R., Density Functional Theory Response to the Calculations of pK_a Values of some Carboxylic Acids in Aqueous Solution. *Journal of Molecular Structure* **2004,** 711, 13-18.

107. Namazian, M.; Heidary, H., Ab Initio Calculations of pK_a Values of some Organic Acids in Aqueous Solution. *Journal of Molecular Structure* **2003,** 620, 257-263.

108. Namazian, M.; Kalantary-Fotooh, F.; Noorbala, M. R.; Searles, D. J.; Coote, M. L., Moller-Plesset Perturbation Theory Calculations of the pK_a Values for a Range of Carboxylic Acids. *Journal of Molecular Structure: THEOCHEM* **2006,** 758, 275-278.

109. Roy, K.; Popelier, P. L. A., Predictive QSPR modelling of the acidic dissociation constant (pK_a) of phenols in different solvents. *J.Phys.Org.Chem.* **2009,** 22, 186-196.

110. Adam, K. R., New Density Functional and Atoms in Molecules Method of Computing Relative pK_a Values in Solution. *J. Phys.Chem.A* **2002,** 106, 11963-11972.

111. Bader, R. F. W.; Beddall, P. M., The Spatial Partitioning and Transferability of Molecular Energies. *Chem.Phys.Lett.* **1971,** 8, 29-36.

112. Esteki, M.; Hemmateenejad, B.; Khayamian, T.; Mohajeri, A., Multi-way analysis of Quantum Topological Molecular Similarity descriptors for modelling acidity constants of some phenolic compounds. *Chem Biol Drug Des* **2007,** 70, 413-423.

113. Dearden, J. C.; Cronin, M. T. D.; Lappin, D. C., A Comparison of Commercially avaliable software from the prediction of pK_a values. *J. Pharm. Pharmacol.* **2007,** 59, A7.

114. ADME Boxes <http://pharma-algorithms.com/webboxes/>.

115. VCCLAB <http://www.vcclab.org/>.

116. ADMET Predictor <http://www.simulations-plus.com/>.

117. Pipeline Pilot <http://accelrys.com/>.

118. SPARC <http://sparc.chem.uga.edu/sparc/>.

119. Marvin <http://www.chemaxon.com/>.

120. QikProp <http://www.schrodinger.com/>.

121. ACD/Labs <http://www.acdlabs.com/home/>.

122. PALLAS <http://www.compudrug.com/>.

123. CSpK$_a$ <http://www.chemsilico.com/>.

124. Meloun, M.; Bordovska, S., Benchmarking and Validating algorithms that estimate pK$_a$ values of drugs based on their molecular structure. *Anal. Bioanal. Chem* **2007,** 389, 1267-1281.

125. Balogh, G. T.; Gyarmati, B.; Nagy, B.; Molnar, L.; Keseru, G. M., Comparative Evaluation of in Silico pK$_a$ Prediction Tools on the Gold Standard Dataset. *QSAR & Combinatorial Science* **2009,** 28, 1148-1155.

126. Epik <http://www.schrodinger.com/>.

127. Liao, C.; Nicklaus, M. C., Comparison of Nine Programs Predicting pK$_a$ Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009,** 49, 2801-2812.

128. Jaguar <http://www.schrodinger.com/>.

129. Livingstone, D. J., *Data Analysis for Chemists*. 1st Ed ed.; Oxford University Press: Oxford, Great Britain, 1995.

130. Vapnik, V., The Nature of Statistical Learning Theory. 2nd Edition. In Springer-Verlag, NY, USA: 1995.

131. Burges, C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining & Knowl.Disc.* **1998,** 2, 121-167.

132. Smola, A. J.; Schoelkopf, B., A tutorial on support vector regression. *Statistics and Computing* **2004,** 14, (3), 199--222.

133. Ivanciuc, O., Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*, Lipkowitz, K. B.; Cundari, T. R., Eds. Wiley-VCH: 2007; Vol. 23, pp 291-400.

134. Jover, J.; Basque, R.; Sales, J., QSPR prediction of pK$_a$ for benzoic acids in different solvents *QSAR & Combinatorial Science* **2008,** 27, (5), 563-581.

135. E. Dimitriadou; K. Hornik; F. Leisch; D. Meyer; A. Weingessel *;e1071: Misc Functions of the Department of Statistics (e1071)*, R package version 1.5-16; TU Wien, 2006.

136. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. **2007**.

137. Haykin, S., *Neural Networks: A Comprehensive Foundation, 2nd Ed.* Prentice-Hall, Upper Saddle River, New Jersey, USA.: 1999.

138. Gurney, K., *An Introduction to Neural Networks*. Routledge: London, Great Britain, 1997.

139. Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Lui, M. C.; Hu, Z. D.; Fan, B. T., Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemem. Intell. Lab. Syst.* **2002,** 62, 217-225.

140. Chen, S.; Cowan, C.; Grant, P., Orthogonal least squares learning for radial basis function networks. *IEEE Transactions on Neural Networks* **1991,** 2, 302–309

141. http://www.chemicaldictionary.org/dic/F/Flurenol_2074.html (15th Dec. 2008),

142. ACD/Labs, *version 3, ACD Labs, Toronto, ON, Canada*.

143. Kresge, A. J.; Pojarlieff, I. G.; Rubinstein, E. M., The acidity constant of fluorene-9-carboylic acid in aqueous solution. Determination of the pK$_a$, of a sparingly soluble substance. *Can. J. Chem* **1993,** 71, 227.

144. Jumppanen, J. H.; Siren, H.; Riekkola, M. L., Correlation of Resolution with Frictional Coefficients and pK$_a$, Values in Capillary Electrophoresis of Four Diuretics: Determination of Electric Field Strength and Electro-Osmotic Velocity. *J. Microcolumn Seperation* **1993,** 5, (5), 451.

145. Leonard, J. T.; Roy, K., On selection of training and test sets for the development of predictive QSAR models. . *QSAR Comb. Sci.* **2006,** 25, 235-251.

146. Perrin, D. D.; Dempsey, B.; Serjean, E. P., *pKa Prediction for Organic Acids and Bases*. Chapman and Hall, London, GB, 1981.

147. Hilal, S. H.; Karickoff, S. W.; Carreira, L. A., A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK$_a$'s *Quant.Struc.Act.Rel.* **1995,** 14, 348-355.

148. VCCLAB, *Virtual Computational Chemistry Laboratory, http://www.vcclab.org, 2005.*

149. ChemAxon, *Calculator Plugins were used for structure property prediction and calculation, Marvin 2.0.4, 2006, ChemAxon (http://www.chemaxon.com)*.

150. Weininger, D., SMILES, a Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model* **1988,** 28, 1267-1281.

151. Popelier, P. L. A., Quantum Chemical Topology: on Bonds and Potentials. In *Structure and Bonding. Intermolecular Forces and Clusters, Ed, D.J.Wales*, Springer: Heidelberg, Germany, 2005; Vol. 115, pp 1-56.

152. Harding, A. P.; Wedge, D. C.; Popelier, P. L. A., pKa Prediction from “Quantum Chemical Topology” Descriptors. *J.Chem.Inf.Mod.* **2009,** 49, 1914–1924.

153. Hawe, G. I.; Alkorta, I.; Popelier, P. L. A., Prediction of the Basicities of Pyridines in the Gas Phase and in Aqueous Solution. *J. Chem. Inf. Model.* **2010,** 50, 87-96.

154. Mitra, I.; Roy, P. P.; Kar, S.; Ojha, R.; Roy, K., On Further Application of the r^{2m} as a Metric for Validation of QSAR Models. *Chemometrics* **2009,** 24, 22-33.

155. Zhou, T.; Huang, D.; Caflisch, A., Quantum Mechanical Methods for Drug Design. *Current Topics in Medicinal Chemistry* **2010,** 10, 33-45.

156. Han, J.; Deming, R. L.; Tao, F. M., Theoretical Study of Hydrogen-Bonded Complexes of Chlorophenols with Water or Ammonia: Correlations and Predictions of pK_a Values. *J. Phys. Chem. A* **2005,** 109, 1159-1167.

157. Han, J.; Tao, F. M., Correlations and Predictions of pK_a Values of Fluorophenols and Bromophenols Using Hydrogen-Bonded Complexes with Ammonia. *J. Phys. Chem. A* **2006,** 110, 257-263.

158. Zhang, J. D.; Zhu, Q. Z.; S.J., L.; Tao, F. M., Prediction of Aqueous pK_a Values of Hydroxybenzoic Acids Using Hydrogen-Bonded Complexes with Ammonia. *Chem. Phys. Lett.* **2009,** 475, 15-18.

159. Han, J.; Deming, R. L.; Tao, F. M., Theoretical Study of Molecular Structures and Properties of the Complete Series of Chlorophenols. *J. Phys. Chem. A* **2004,** 108, 7736-7743.

160. Han, J.; Lee, I.; Tao, F. M., Molecular Structures and Properties of the Complete Series of Bromophenols: Density Functional Theory Calculations. *J. Phys. Chem. A* **2005,** 109, 5186-5192.

161. Tao, L.; Han, J.; Tao, F. M., Correlations and Predictions of Carboxylic Acid pK_a Values Using Intermolecular Structure and Properties of Hydrogen-Bonded Complexes. *J. Phys. Chem. A* **2008,** 112, 775-782.

162. Yu, H.; Kuhne, R.; Ebert, R. U.; Schuurmann, G. J., Comparative Analysis of QSAR Models for Predicting pK_a of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure. *J. Chem. Inf. Model* **2010,** 50, 1949-1960.

163. Chis, V., Molecular and Vibrational Structure of 2,4-Dinitrophenol: FT-IR, FT-Ramen and Quantum Chemical Calculations. *Chemical Physics* **2004,** 300, 1-11.

164. Ragnar, M.; Lindgren, C. T.; Nilverbrant, N. O., pK_a Values of Guaiacyl and Syringyl Phenols Related to Lignin. *J. Wood. Chem. Tech.* **2000,** 20, 277-305.

165. Chapman, E.; Bryan, M. C.; Wong, C. H., Mechanistic Studies of *B*-Artlsulfotransferase IV. *PNAS* **2003,** 100, 910-915.

166. Schafer, B.; Engwald, W., Enrichment of Nitrophenols from Water by Means of Solid-Phase Microextraction. *Frenenius' Journal of Analytical Chemistry* **1995,** 352, 535-536.

167. Vaughan, W. R., Effects of Alkyl Groups on 4-Nitro- and 4-Nitroso-Phenols. *J. Org. Chem.* **1956,** 21, 1201-1210.

168. Hurwitz, A. R.; Liu, S. T., Determination of Aqueous Solubility and pK_a Values of Estrogens. *Journal of Pharmaceutical Sciences* **1977,** 66, 624-627.

169. Cronin, D. M., Opportunities for Computer-aided Prediction of Toxicity in Drug Discovery. *Liverpool John Moores University* **2003**, 50-52.

170. Prentis, R. A.; Lis, Y.; Walker, S. R., Pharmaceutical innovation by Seven UK-owned Pharmaceutical Companies. *British Journal of Clinical Pharmacology* **1988,** 25, 387-396.

171. Kola, I.; Landis, J., Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **2004,** 3, 711-715.

172. Kalgutkar, A. S.; al., e., A comprehensive Listing of Bioactivation Pathways of Organic Functional Groups. *Current Drug Metabolism* **2005,** 6, 161-225.

173. Guengerich, F. P., Cytochrome P450 and Chemical Toxicology. *Chemical Research in Toxicology* **2008,** 21, 70-83.

174. Rendic, S.; Di Carlo, F. J. D., Human Cytochrome P450 Enzymes: A Status Report Summarizing Their Reactions, Substrates, Inducers, and Inhibitors. *Drug Metabolism Reviews* **1997,** 387, 1-16.

175. Masimirembwa, C. M.; Thompson, R.; Andersson, T. B., In Vitro High Throughput Screening of Compounds for Favourable Metabolic Properties in Drug Discovery. *Comb. Chem. High. Throughput Screening* **2001,** 4, (245-263).

176. Shimada, T.; Yamazaki, H.; Mimura, M.; Inui, Y.; Guengerich, F. P., Interindividual Variations in Human Liver Cytochrome P-450 Enzymes Involved in the Oxidation of Drugs, Carcinogens and Toxic Chemicals: Studies with Liver Microsomes of 30 Japanese and 30 Caucasians. *J. Pharmacol. Exp. Ther.* **1994,** 270, 414-423.

177. Lewis, D. F. V., Structural Characteristics of Human P450s Involved in Drug Metabolism: QSARs and Lipophilicity Profiles. *Toxicology* **2000,** 144, 197-203.

178. Borosky, G. L., Ultimate Carcinogenic Metabolites from Aromatic and Heterocyclic Aromatic Amines: A Computational Study in Relation to Their Mutagenic Potency. *Chem. Res. Toxicol.* **2007,** 20, 171-180.

179. *International Agencies for Research on Cancer, IARC Monographs on the Evaluation of the Carcinogenicity Risk to Humans, Polycyclic Aromatic Compounds, Part 1, Chemical Environmental and Experimental Data*. IARC: Lyon, 1983.

180. Mortelmans, K.; Zeiger, E., The Ames Salmonella/microsome mutagenicity assay. *Mutation Research* **2000,** 455, 29-60.

181. Kirkland, D.; Aardema, M.; Henderson, L.; Muller, L., Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutation Research* **2005,** 584, 1-256.

182. Matthews, E. J.; Kruhlak, N. L.; Cimino, M. C.; Benz, R. D.; Contrera, J. F., An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints. *Regulatory Toxicology and Pharmacology* **2006,** 44, 83-96.

183. Bentzien, J.; Hickey, E. R.; Kemper, R. A.; Brewer, M. L.; Dyekjaer, J. D.; East, S. P.; Whittiker, M., An in Silico Method for Predicting Ames Activities of Primary Aromatic Amines by Calculating the Stabilities of the Nitrenium Ion. *J. Chem. Inf. Model* **2010,** 50, 274-297.

184. European Medicines Agency; *Non-Clinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals ICH M3 (R2)*; London, UK, 2008.

185. United States Food and Drug Administration *Guidance for Industry: Nonclinical Safety Studies for the Conduct of Human Clinical Trails for Pharmaceuticals ICH M3*; Rockville, MD, 1997.

186. United States Food and Drug Administration *Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use ICH S2 (R1)*; Rockville, MD, 1997.

187. Atrakchi, A. H., Interpretation and Consideration on the Safety Evaluation of Human Drug Metabolites. *Chem. Res. Toxicol.* **2009,** 22, 1217-1220.

188. Ames, B. N.; McCann, J.; Yamasaki, E., Methods for Detecting Carcinogens and Mutagens with the Salmonella/Mammalian-Microsome Mutagenicity Test. *Mutation Research* **1975,** 31, 347-364.

189. Maron, D. M.; Ames, B. N., Revised Methods for the Salmonella Mutagenicity Test. *Mutation Research* **1983,** 113, 173-215.

190. Nigsch, F.; Macaluso, M.; Mitchell, J. B. O.; Zmuidinavicius, D., Computational Toxicology: An Overview of the Sources of Data and of Modelling Methods. *Expert. Opin. Drug. Metab. Toxicol.* **2009,** 5, 1-14.

191. Walmsely, R. M., Genotoxicity screening: the slow march to the future. *Expert Opinion in Drug Metabolism and Toxicology* **2005,** 1, (2), 261-268.

192. Fluckiger-Isler, S.; Baumeister, M.; Braunk, K., Assessment of the Performance of the Ames II assay: a Collaborative Study with 19 Coded Compounds. *Mutation Research* **2004,** 558, 181-197.

193. Hastwell, P. W.; Chai, L.; Roberts, K. J.; Webster, T. W.; Harvey, J. S.; Rees, R. W.; Walmsley, R. M., High-Specificity and High-Sensitivity Genotoxicity Assessment in a Human Cell Line: Validation of the GreenScreen HC GADD45a-GFP Genotoxicity Assay. *Mutation Research* **2006,** 607, 160-175.

194. Guengerich, F. P.; MacDonald, J. S., Applying Mechanisms of Chemical Toxicology to Predict Drug Safety. *Chemical Research in Toxicology* **2007,** 20, 344-369.

195. European Medicines Agency; *Guideline on the Limits of Genotoxic Impurities*; London, UK, 2007.

196. Benigni, R.; Tatiana, I. N.; Benfenati, E.; Bossa, C.; Franke, R.; Helma, C.; Hulzebos, E.; Marchant, C.; Richard, A.; Woo, Y. T.; Yang, C., The Expanding Role of Predictive Toxicology: An Update on the Q(SAR) Models for Mutagens and Carcinogens. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **2007,** 25, 53-97.

197. Carlsson, L.; Ahlberg, A.; Boyer, S., Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model* **2009,** 49, 2551-2558.

198. Hansen, K.; Mika, S.; Schroeter, T.; Sutter, T.; ter LaaK, A.; Steger-Hartmann, T.; Heinrick, N.; Muller, K. R., Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model* **2009,** 49, 2077-2081.

199. Langham, J. J.; Jain, A. N., Accurate and Interpretable Computational Modeling of Chemical Mutagenicity. *J. Chem. Inf. Model* **2008,** 48, 1833-1839.

200. Mazzatorta, P.; Tran, L. A.; Schilter, B.; Grigorov, M., Integration of Structure-Activity Relationship and Artificial Intelligence Systems To Improve in Silico Prediction of Ames Test Mutagenicity. *J. Chem. Inf. Model* **2007,** 47, 34-38.

201. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W., Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004,** 19, 365-377.

202. Benigni, R.; Bossa, C., Predictivity in QSAR. *J. Chem. Inf. Model.* **2008,** 48, 971-980.

203. Debnath, A. K.; Debnath, D.; Shusterman, A. J.; Hansch, C., A QSAR Investigating the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA98 and TA100. *Environmental and Molecular Mutagenesis* **1993,** 19, 37-52.

204. Torres-Cartas, S.; Martin-Biosca, Y.; Villanueva-Camanas, R. M.; Sagrado, S.; Medina-Hernandez, M. J., Biopartitioning Micellar Chromatography to Predict Mutagenicity of Aromatic Amines. *Eur.J.Med.Chem.* **2007,** 42, 1396-1402.

205. Basak, S. C.; Mills, D.; Balaban, A. T.; Gute, B. D., Prediction of Mutagenicity of Aromatic and Heteroaromtic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001,** 41, 671-678.

206. Maran, U.; Kareleson, M.; Katritzky, A. R. A., A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.,* **1999,** 18, 3-10.

207. Cash, G. G., Prediction of the Genotoxicity of Aromatic and Heteroaromatic Amines using Electrotopological State Indices. *Mutation Research* **2001,** 491, 31-37.

208. Vracko, M.; Mills, D.; Basak, S. C., Structure-Mutagenicity Modelling Using Counter Propagation Neural Networks. *Environmental Toxicology and Pharmacology* **2004,** 16, 25-36.

209. Valkova, I.; Varako, M.; Basak, S. C., Modeling of Structure-Mutagenicity Relationships: Counter Propagation Neural Network Approach Using Calculated Structural Descriptors. *Analytica Chemica Acta* **2004,** 509, 179-186.

210. Cash, G. C.; Anderson, B.; Mayo, K.; Bogaczyk, S.; Tunkel, J., Predictive Genotoxicity of Aromatic and Heteroaromatic Amines Using Electrotopological State Indices. *Mutation Research* **2005,** 585, 170-183.

211. Bhat, K. L.; Hayik, S.; Sztandera, L.; Bock, C. W., Mutagenicity of Aromatic and Heteroaromatic Amines and Related Compounds: A QSAR Investigation. *QSAR & Comb.Sci.* **2005,** 24, 831-843.

212. Hatch, F. T.; Knize, M. G.; Colvin, M. E., Extended Quantitative Structure-Activity Relationships for 80 Aromatic and Heterocyclic Amines; Structural, Electronic, and Hydropathic Factors Affecting Mutagenic Potency. *Environmental and Molecular Mutagenesis* **2001,** 38, 268-291.

213. Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A., Quantitative Structure-Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. *Chem. Rev.* **2000,** 100, 3697-3714.

214. Felton, J. S.; Knize, M. G.; Wu, R. W.; Colvin, M. E.; Hatch, F. T.; Malfatti, M. A., Mutagenic Potency of Food-Derived Heterocyclic Amines. *Mutation Research* **2007,** 616, 90-94.

215. Ford, G. P.; Herman, P. S., Relative Stabilities of Nitrenium Ions Derived From Polycyclic Aromatic Amines: Relationship to Mutagenicity. *Chem.-Biol. Interact.* **1992,** 81, 1-18.

216. Ford, G. P.; Griffin, G. R., Relative Stabilities of Nitrenium Ions Derived From Heterocyclic Amine Food Carcinogens: Relationship to Mutagenicity. *Chem.-Biol. Interact.* **1992,** 81, 19-33.

217. Novak, M.; Rajagopal, S., Correlations of Nitrenium Ion Selectivities with Quantitative Mutagenicity and Carcinogenicity of the Corresponding Amines. *Chem. Res. Toxicol.* **2002,** 15, 1495-1503.

218. Borosky, G. L., Carcinogenic Carbocyclic and Heterocyclic Aromatic Amines: A DFT Study Concerning Their Mutagenic Potency. *J. Mol. Graph. Mod.* **2008,** 27, 459-465.

219. Benigni, R., Structure-Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chem. Rev.* **2005,** 2005, 1767-1800.

220. Benigni, R.; Andreoli, C.; Giuliani, A., QSAR Models for Both Mutagenic Potency and Activity: Application to Nitroarenes and Aromatic Amines. *Environmental and Molecular Mutagenesis* **1994,** 24, 208-219.

221. Leach, A. G.; Cann, R.; Tomasi, S., Reaction Energies Computed with Density Functional Theory Corresponds With a Whole Organism Effect' Modelling the Ames Test for Mutagenicity. *Chem. Commun.* **2009**, 1095-1096.

222. Ashby, J.; Tennant, R. W., Chemical Structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis Among 222 Chemicals Tested by the U.S.NCI/NTP. *Mutation Research* **1988,** 204, 17-115.

223. Kazius, J.; McGuire, R.; Bursi, R., Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005,** 48, 312-320.

224. Casalegno, M.; Benfenati, E.; Sello, G., Application of a Fragment-Based Model to the Prediction of the Genotoxicity of Aromatic Amines. *Internet Electronic Journal of Molecular Design* **2006,** 5, 431-446.

225. DEREK for Windows (Deductive Estimation of Risk from Existing Knowledge), Lhasa Ltd, available at <www.lhasalimited.org/index.php/derek/>, accessed 27th May 2010.

226. TOPKAT (Toxicity Prediction by Komputer Assisted Technology), Accelrys Software Inc., available at <www.accelrys.com/products/topkat>, accessed 27th May 2010.

227. MultiCASE (Multiple Computer Automated Structure Evaluation), MultiCASE Inc., available at <www.multicase.com> , accessed 27th May 2010.

228. Snyder, R. D.; Smith, M. D., Computational Prediction of Genotoxicity: Room for Improvement. *Drug Discovery Today* **2005,** 10, 1119-1124.

229. Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J.; Burlinson, B.; Gombar, V., Comparison of the Computer Programmes DEREK and TOPKAT to Predict Bacterial Mutagenicity. *Mutagenesis* **2002,** 17, 321-329.

230. Sasaki, J. C.; Fellers, R. S.; Colvin, M. E., Metabolic oxidation of the carcinogenic arylamines by p450 monooxygenases: theoretical support for the one-electron transfer mechanism. *Mutation Reserach* **2002,** 506-507, 79-89.

231. Hammons, G. J.; Guengerich, F. P.; C.Weis, C.; Beland, F. A.; Kadlubar, F. F., Metabolic oxidation of carcinogenic arylamines by rat, dog, and human hepatic microsomes and by purified flavin-containing and cytochrome P450 monooxygenases. *Cancer Research* **1985,** 45, 3578-3585.

232. Connor, T. H.; Ramanujam, V. M. S.; Rinkus, S. J.; Legator, M. S.; Trieff, N. M., The evaluation of mutagenicities of 19 structurally related aromatic amines and acetamides in Salmonella typhimurium TA98 and TA100. *Mutation Research* **1983,** 118, 49-59.

233. Later, D. W.; Pelroy, R. A.; Stewart, D. L.; McFall, T.; Booth, G. M.; Lee, M. L.; Tedjamulia, M.; Castle, R. N., Microbial mutagenicity of isomeric two-, three-, and four-ring amino polycyclic aromatic hydrocarbons. *Environmental Mutagenesis* **1984,** 6, 497-515.

234. *TSAR, Oxford Molecular Ltd*, Version 3.3; 2000.

235. Gramatica, P.; Consonni, V.; Pavan, M., Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. *SAR and QSAR in Environmental Research* **2003,** 14, (4), 237-250.

236. Cronin, M. T. D.; Schultz, T. W., Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM* **2003,** 622, 39-51.

237. Holder, A. J.; Ye, L., Quantum Mechanical Quantitative Structure-Activity Relationships to Avoid Mutagenicity. *Dental Material* **2009,** 25, 20-25.

238. Davydov, R. D.; Halpert, J. R., Allosteric P450 Mechanisms: Multiple Binding Sites, Multiple Conformers or Both? *Expert Opinion in Drug Metabolism and Toxicology* **2008,** 4, 1523-1535.

239. Isin, E. M.; Sohl, C. D.; Eoff, R. L.; Guengerich, F. P., Cooperativity of Cytochrome P450 1A2: Interactions of 1,4-Phenylene Diisocyanide and 1-Isopropoxy-4-Nitrobenzene *Archives of Biochemistry and Biophysics* **2008,** 473, 69-75.

240. Sohl, C. D.; Isin, E. M.; Eoff, R. L.; Marsch, G. A.; Stec, D. F.; Guengerich, F. P., Cooperativity on Oxidation Reactions Catalysed by Cytochrome P450 1A2. *Journal of Biological Chemistry* **2008,** 283, 7293-7308.

241. Sansen, S.; Yano, J. K.; Reynald, G.; Schoch, G.; Griffin, K.; Stout, C. D.; Johnson, E. F., Adaptations for the Oxidation of Polycyclic Aromatic Hydrocarbons Exhibited by the Structure of Human P450 1A2. *Journal Biological Chemistry* **2007,** 282, 14348-14355.

242. Thompson, C. Z.; Hill, L. E.; Epp, J. K.; Probst, G. S., The induction of Bacterial Mutation and Hepatocyte Unscheduled DNA Synthesis by Monosubstituted Anilines. *Environmental Mutagenesis* **1983,** 5, 803-811.

243. Nicolette, J., Identifying and Controlling GTI: A Toxicology Perspective. *IIRUSA Conference* **December 2008**.

244. Burnett, C.; Fuchs, C.; Corbett, J.; Menkart, J., The Effect of DMSO on the Mutagenicity of the Hair Dye p-Phenylenediamine. *Mutation Research* **1982,** 103, 1-4.

245. Chauret, N.; Gauthier, A.; Nicoll-Griffith, D. A., Effect of Common Organic Solvents on in vitro Cytochrome P450-Mediated Metabolic Activities in Human Liver Microsomes. *Drug Metabolism and Disposition* **1997,** 26, 1-4.

246. Nestmann, E. R.; Douglas, G. R.; Kowbel, D. J.; Harrington, T. R., Solvent Interactions with Test Compounds and Recommendations for Testing to Avoid Artefacts. *Environmental and Molecular Mutagenesis* **1985,** 7, 163.

247. Yahagi, T.; Nagao, M.; Seino, Y.; Matsushima, T.; Sugimura, T.; Okada, M., Mutagenicities of N-Nitrosamines on Salmonella. *Mutation Research* **1977,** 48, 121-130.

248. Maron, D.; Katzenellenbogen, J.; Ames, B. N., Compatibility of Organic Solvents with the Salmonella/Microsome Test. *Mutation Research* **1981,** 88, 343-350.

249. Organisation for the Economic Cooperative and Development (OECD); *Guidelines for the Testing of Chemicals: Bacteria Reverse Mutation Test, Guideline 471*; 1997.

250. Walmsley, R. M., Genotoxicity screening: the slow march to the future. *Expert Opinion in Drug Metabolism and Toxicology* **2005,** 1, 261-268.

251. LigPrep, version 2.3, Schrodinger, LLC, New York, NY, 2009.

252. Maestro, version 9.0, Schrodinger, LLC, New York, NY, 2009.

253. Alaejos, M. S.; Pino, V.; Afonso, A. M., Metabolism and Toxicology of Heterocyclic Aromatic Amines When Consumed in Diet: Influence of the Genetic Susceptibility to Develop Human Cancer. A Review. *Food Research International* **2008,** 41, 327-340.

254. http://www.bgchemie.de/files/95/ToxBew226-E.pdf and http://dra4.nihs.go.jp/mhlw_data/home/paper/paper87-02-5e.html Accessed 15th October 2010.

255. Takahashi, K.; Kaiya, T.; Kawazoe, Y., Strucuture-Mutagenicity Relationship Among Aminoquinolines, Aza-Analogues of Naphylamine, and Their Derivatives. *Mutation Research* **1987,** 187, 191-197.

256. Willems, M. I.; Dubois, G.; Boyd, D. R.; Davis, R. J. H.; Hamilton, L.; McCullough, J. J.; van Bladen, P. J., Compariosn of the Mutagenicity of Quinoline and all Monohydroxyquinolines with a Series of Arene Oxide, Trans-Dihyrodiol, Diol Epoxide, N-Oxide and Arene Hydrate Derivatives of Quinoline in the Ames/Salmonella Microsome Test. *Mutation Research* **1992,** 278, 227-236.

257. Dunkel, V. C.; Zeiger, E.; Brusick, D.; McCoy, E.; McGregor, D.; Mortelmans, K.; Rosenkranz, H. S.; Simmon, V. F., Reproducibility of Microbial Mutagenicity Assay: II. Testing of Carcinogens and Noncarcinogens in Salmonella Typhimurium and Escherichia Coli. *Environmental Mutagenesis* **1985,** 7, 1-248.

258. Inami, K.; Nagao, M.; Ishikawa, S.; Mochizuki, M., Mutagenicity of Heterocyclic Amines by Biomimetic Chemical Models for Cytochrome P450 in Ames Assay. *Genes and Environment* **2010,** 32, 7-13.

259. Inami, K.; Okazawa, M.; Mochizuki, M., Mutagenicity of Aromatic Amines and Amides with Chemical Models for Cytochrome P450 in Ames Assay. *Toxicology in Vitro* **2009,** 23, 986-991.

Appendix A

pK_a Prediction from "Quantum Chemical Topology" Descriptors

A. P. Harding, D. C. Wedge, and P. L. A. Popelier*

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain, and
School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

Received May 12, 2009

Knowing the pK_a of a compound gives insight into many properties relevant to many industries, in particular the pharmaceutical industry during drug development processes. In light of this, we have used the theory of Quantum Chemical Topology (QCT), to provide *ab initio* descriptors that are able to accurately predict pK_a values for 228 carboxylic acids. This Quantum Topological Molecular Similarity (QTMS) study involved the comparison of 5 increasingly more expensive levels of theory to conclude that HF/6-31G(d) and B3LYP/6-311+G(2d,p) provided an accurate representation of the compounds studies. We created global and subset models for the carboxylic acids using Partial Least Square (PLS), Support Vector Machines (SVM), and Radial Basis Function Neural Networks (RBFNN). The models were extensively validated using 4-, 7-, and 10-fold cross-validation, with the validation sets selected based on systematic and random sampling. HF/6-31G(d) in conjunction with SVM provided the best statistics when taking into account the large increase in CPU time required to optimize the geometries at the B3LYP/6-311+G(2d,p) level. The SVM models provided an average q^2 value of 0.886 and an RMSE value of 0.293 for all the carboxylic acids, a q^2 of 0.825 and RMSE of 0.378 for the ortho-substituted acids, a q^2 of 0.923 and RMSE of 0.112 for the para- and meta-substituted acids, and a q^2 of 0.906 and RMSE of 0.268 for the aliphatic acids. Our method compares favorably to ACD/Laboratories, VCCLAB, SPARC, and ChemAxon's pK_a prediction software based of the RMSE calculated by the leave-one-out method.

1. INTRODUCTION

The pK_a of a compound is an important property in both life sciences and chemistry since the propensity of a compound to donate or accept a proton is fundamental to understanding chemical and biological processes. As the pK_a value of a molecule also determines the amount of protonated and deprotonated species at a specific pH, for example at physiological pH, knowing the pK_a of a molecule gives insight into pharmacokinetic properties. The latter includes the rate at which a molecule will diffuse across membranes and other physiological barriers, such as the blood brain barrier. More often than not, phospholipid membranes easily absorb neutral molecules, while ionized molecules tend to remain in the plasma or the gut[1] before being excreted. Many biological systems also use proton-transfer reactions to communicate between the intra- and extracellular media, and the rate of the proton-transfer reaction depends, in-part, on the pK_a values of the species involved.[2]

There are a number of well established experimental techniques,[3] such as spectroscopy, potentiometry, conductimetry, competitive reactions, and titrimetry that can accurately determine pK_a values for a molecule. However, experimental determination of acidity of a specific part of a large biological molecule such as a protein is not a straightforward task[4] and is often associated with large uncertainties in the results. The benefits of a technique that accurately predicts the dissociation constant without the need for "wet" experiments are clear. The chemical industry, in particular the pharmaceutical and agrochemical sectors, screen thousands of compounds during the discovery process for many properties simultaneously, including the dissociation constant. More efficient techniques are required because of the logistics of measuring the pK_a values of these compounds are coupled with known problems associated with certain techniques. For example, high-throughput UV absorption measurements can often miss groups not in close proximity to a UV-chromophore.[5] It is clear that *in silico* techniques can increase efficiency and reduce costs at a challenging time for the pharmaceutical industry as they can be consulted before expensive synthetic work is undertaken.

The need for accurate pK_a prediction has led to a large number of publications tackling the issue. While the precise implementation of methodology varies, the publications generally fall into three categories: (i) predictive models, using a range of descriptors and statistical methods, (ii) *ab initio* quantum chemical methods based on different thermodynamic cycles, and (iii) fragment-based approaches.

The first category of models relies on choosing the right descriptors to model the pK_a of a particular data set. Structural, physiochemical, topological, geometrical, constitutional, electrostatic, quantum-chemical, and thermodynamic descriptors have all been used to predict pK_a with varying success. Gruber and Buss[6] performed semiempirical calculations on some 190 phenols and carboxylic acids. They used multilinear regression, with descriptors such as heats of formation, molecular orbital energies, and charge densities, to produce a three-term equation for the benzoic acid derivatives and a four-term equation for the aliphatic acids. Citra[7] criticized this linear free energy relationship based approach for lacking scientific basis and vast use of correc-

* Corresponding author e-mail: pla@manchester.ac.uk.

10.1021/ci900172h CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/24/2009

tion factors favoring quantum mechanical methods. A three-term equation for benzoic acids was reported using a method similar to that of Gruber and Buss. Gross and Seybold[8] rejected the use of semiempirical methods, instead using density functional theory, after a set of survey calculations demonstrated it performed significantly better for the descriptors they employed. After studying phenols they found that atomic charge, and the HOMO and LUMO energies, correlated with pK_a. After rejecting the use of *ab initio* methods as too computationally demanding, Tehan[1] and co-workers produced QSARs for aliphatic acids and substituted benzoic acids. Xing and Glen[9] fashioned a novel structure tree representation of atoms to align molecules. Atom types and group types that were of biological interest were used in conjunction with PLS to produce a QSAR for a large set of acids and bases. Following this procedure, Xing[10] and co-workers reduced the number of atom types and increased the group types used, also noting that splitting the data set improved their results. The approach introduced by Xing has been taken up by numerous authors.[11,12] For example, Jelfs et al.[5] utilized the tree fingerprint method to develop a prediction method using semiempirical chemical properties, such as partial charge and electrophilic superdelocalizability of atoms undergoing protonation or deprotonation, to produce an online prediction Web-tool for pK_a prediction at Novartis. In the second category of methods, *ab initio* quantum chemical methods based on different thermodynamic cycles[13,14] have started to receive more attention.[15,16]

The method involves calculating the standard change in Gibbs energy related to the dissociation of a proton from the compound under study in water. The method utilizes gas phase and aqueous phase *ab initio* calculations but, depending on the thermodynamic cycle used, involves at least four separate geometry optimizations for each prediction. Out of all the commercial packages available for pK_a prediction, Schrödinger's Jaguar[17] application is the only tool that employs this method. Although these methods do not rely upon a predictive model, the results can be associated with large uncertainties because different thermodynamic cycles, levels of theory and models of solvation produce different pK_a values. The Jaguar package, for example, uses empirical correction terms, where calculated values are fitted to experimental values stored in a database, to repair deficiencies in both the *ab initio* calculations and solvation models. Namazian[4,18−21] and co-workers used an equation[2,22] that relates the standard change of Gibbs energy to the pK_a of various carboxylic acids.

In the third and final category of methods, the fragment-based approach supported by a back end database is implemented in the LogD suite of ACD/Laboratories.[23] It uses an internal database of around 16,000 structures with over 31,000 experimental values. Problems associated with such methods occur when fragments present in a molecule are absent from the database.

Quantum Topological molecular similarity (QTMS), pioneered in our group,[24,25] is a novel approach to solving QSAR/QSPR problems using descriptors defined by Quantum Chemical Topology (QCT).[26,27] QTMS uses properties evaluated at special points in space as a means of summarizing the electronic structure of a set of molecules and correlating them with some given activity. These special points are called bond critical points (BCPs) and are briefly reviewed in the next Section. Over the last several years QTMS methods have produced excellent results of relevance to biological,[28,29] medicinal,[30−34] environmental,[35] industrial, and physical organic chemistry.[36−39] In all cases the models had excellent validation statistics and also provide information about the active center or region of the compounds thought to be important to the activity. Here we evaluate QTMS descriptors in an extension to previous QTMS studies, which have shown good predictive ability for pK_a.[40,41]

Regarding pK_a, Adam[42] obtained impressive results by incorporating QCT into his study. Using transferability between similar molecules, an idea at the origin of QCT,[43] he obtained an r^2 for aliphatic and benzoic acids greater than 0.84, in most cases, using the energy of the dissociating proton in solution as the only descriptor. On the other hand, QTMS descriptors have been successfully employed for carboxylic acids and anilines[41] and phenols in aqueous[44] and polar solvents.[40]

From the aforementioned QTMS publications it has become clear that QTMS descriptors are effective at capturing electronic effects. Therefore we deduce that when QTMS fails electronic effects are not as important to the predicted property or activity as, for example, solubility or steric effects. This was the case for the predictive QSAR models of phenols[32] where logP was included to describe the importance of hydrophobicity for hepatocyte toxicity prediction in conjunction with QTMS descriptors capturing the important electronic effects. Another example is the QTMS study[45] on a remarkable and unusual set of *ortho* alkyl substituted phenols, known for their cytotoxicity and previously investigated by the Hansch group.[46] The QTMS results do not support their proposal that a steric factor is important in the determination of the cytotoxicity. In fact, the QTMS results suggest no steric contribution whatsoever.

In the past, we tended to utilize the interpretative ability of PLS at the chemometric stage. The program SIMCA-P[47] contains a module that calculates the relative importance of each descriptor to the model; these are known as the Variable Importance in the Projection (VIP). The most important descriptors for prediction related to individual bonds can be examined and the so-called active center of the compounds under study detected. While this is a compelling feature of QTMS, we concentrate here on highly predictive models for pK_a estimation. For this reason we have used other statistical methods, such as support vector machines (SVM) and radial basis function neural networks (RBFNN), which yield models that are not so interpretable but possibly more accurate. We have also extensively cross-validated our models and moved away from relying on SIMCA-P for validation. The 228 carboxylic acid compounds included in this study constitute the largest set of compounds investigated with QTMS. This large set of diverse carboxylic acids facilitates the aim of extending the domain of applicability and producing more predictive models.

2. METHODS AND COMPUTATIONAL DETAILS

2.1. Data Set. We seek to predict the pK_a of molecules or fragments with pharmaceutical relevance. We therefore used the data set of Tehan et al.,[1] who had previously applied a variety of filters in order to remove non-druglike molecules from their data set. We selected carboxylic acids as compounds of interest because we want to apply the QTMS methodology to a set of large and diverse compounds further developing previous applications of QTMS. After iodine containing molecules were removed, since the basis sets were not readily available, our data set contained 228 carboxylic acids with a pK_a range of 0.51−6.20. This included 44 meta- and para-substituted benzoic acids, 50 ortho-substituted benzoic acids, and 134 aliphatic carboxylic acids. The observed pK_a values for all 228 carboxylic acids are listed in Table S1 of the Supporting Information.

2.2. Bond Critical Points. Broadly speaking, a BCP forms when the gradient of the electron density, ρ, vanishes ($\nabla\rho=\mathbf{0}$) at a point in space between two bonded nuclei. The electron density at the BCP, ρ_b, is the first QTMS descriptor, from which bond orders can be derived.[48] It also displays strong correlations with bond energy.[49] At the BCP, the Hessian of the electron density has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive one ($\lambda_3 > 0$). The eigenvector associated with λ_3 is tangential to the bond, and so λ_3 describes the curvature along the bond. The eigenvectors corresponding to λ_1 and λ_2 are orthogonal to the bond, and so λ_1 and λ_2 describe the curvatures perpendicular to the bond. The sum of the three eigenvalues is the Laplacian of the electron density, denoted by $\nabla^2\rho$, which gives a measure of the local charge concentration or depletion at the BCP. If the negative eigenvalues λ_1 and λ_2 dominate, then an accumulation of charge takes place in the plane perpendicular to the bond. This is common for shared interactions, such as covalent bonds. This results in a negative value for the Laplacian. If the positive eigenvalue λ_3 dominates, then the electron density accumulates along the bond toward the nuclei. This is common for closed-shell interactions such as ionic, hydrogen, and van der Waals bonds. This results in a positive value for the Laplacian. The ellipticity of ρ, denoted by ε and defined as $\varepsilon = \lambda_1/\lambda_2 - 1$, provides a further useful property associated with BCPs. If ρ protrudes more in one of two directions perpendicular to a bond, then an oval pattern appears, such as in a pure double bond. This ovality is measured by the ellipticity. Single bonds are characterized by λ_1 and λ_2 being nearly identical, and hence ε is near zero. Since $\nabla^2\rho$ and ε are calculated using the three eigenvalues, we have chosen to exclude $\nabla^2\rho$ and ε in this study and instead work with $\lambda_{1,}$ λ_2, and $\lambda_{3.}$ Two types of kinetic energy, denoted by K($\mathbf{r}$) and the more 'classical' kinetic energy G($\mathbf{r}$)[50] can also be used as QTMS descriptors. We also include the equilibrium bond length, R_e, as a QTMS descriptor. In this QTMS study we therefore have seven QTMS descriptors to represent each bond used, namely ρ_b, $\lambda_{1,}$ $\lambda_{2,}$ λ_3, K($\mathbf{r}$), G($\mathbf{r}$), and R_e.

2.3. Data Generation. The data generation process for QTMS can be found in previous publications.[24] In short, an approximation of the geometry of each molecule is provided by MOLDEN.[51] Using the program GAUSSIAN03[52] geometries were optimized successively at five different levels of theory: AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31+G(d,p), and B3LYP/6-311+G(2d,p). The levels are denoted by letters *A, B, C, D,* and *E,* respectively, for consistency with previous publications. The optimization steps are by far the most computationally expensive stage in the QTMS process. The electronic wave function calculated by GAUSSIAN is then passed on to a local version of the program MORPHY98,[53] which locates the BCPs using a robust algorithm.[54] This yields a property vector for each BCP, providing a discrete "quantum fingerprint" of molecules, if all BCPs appearing in a molecule are combined. In this study the electron density (ρ), the three eigenvalues of the Hessian of the electron density (λ_1, λ_2, and λ_3), the two types of kinetic energy (K($\mathbf{r}$) and G($\mathbf{r}$)), and the equilibrium bond lengths (R_e) have been used to describe each BCP. The numbering scheme given to atoms in the common skeleton of each molecule is shown in Figure 1. This allows the location of descriptors important to the statistical analysis to be identified in each molecule. In addition, the scheme allows BCPs in one molecule to be mapped onto corresponding BCPs in other molecules. This is not a fundamental requirement of the method, and the constraint of a common skeleton can be relaxed.[30] We end up with five data matrices (one for each level of theory) consisting of 228 observations (i.e., measured pK_a values) and 21 descriptors for each observation, that is, 7 descriptors obtained for each of the three bonds in the common skeleton.

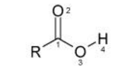


Figure 1. Numbering scheme of the common skeleton of the carboxylic acids.

2.4. Machine Learning and Chemometric Analysis. *2.4.1. Partial Least Squares.* Partial least-squares (PLS)[55] analysis was carried out to fit the BCP descriptor variables to the experimental pK_a values. The program SIMCA-P[47] was used along with its predefined criterion for determining the significant number of Latent Variables (LVs) to appear in the PLS equation. If the value of q^2 of the newly constructed LV is less than 0.097, then that LV is considered not significant, and no more LVs are computed; the PLS regression is then deemed complete.

Two statistics provided by SIMCA-P are the squared correlation coefficient, r^2, and the cross-validated r^2, denoted by q^2. The generated q^2 is based on 'leave-one-seventh' of the data out rather than the popular 'leave-one-out', which is not recommended because of its known pitfalls.[55,56] First, the initial model is constructed, involving all descriptors at each level of theory. Then the VIP plots are examined because they offer a concise summary of the importance of each of the descriptors. Descriptor variables with a VIP value less than unity are considered unimportant to the model and hence discarded.[55] The models are then reconstructed with the reduced set of variables. We also built models using just R_e to demonstrate that using BCP properties as descriptors provides more information than R_e alone. As well as producing global models for the carboxylic acids, we repeated the PLS analysis after splitting the data set into aliphatic and benzoic acids, which were further split into meta/para-substituted and ortho-substituted sets. Altogether there are three subsets.

2.4.2. Support Vector Machines. Support vector machines, originally proposed by Vapnik[57] to solve pattern recognition problems,[58] were extended in 1996 for linear and nonlinear support vector regression (SVR).[59] Due to their remarkable generalization performance, SVM have found numerous applications in chemistry including drug design, QSAR/QSPR, chemometrics, sensors, chemical engineering, and text mining.[60]

Table 1. Summary of the Initial PLS Analysis To Determine the Level of Theory To Use for the Comparison of Learning Methods[b]

		all acids			ortho subset			para/meta subset			aliphatic subset		
level	descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.554	0.537	1	0.733	0.717	1	0.768	0.756	2	0.795	0.767
B	bond lengths	2	0.506	0.484	1	0.741	0.717	1	0.664	0.573	2	0.728	0.693
	BCP properties	1	0.600	0.595	1	0.761	0.716	1	0.683	0.601	2	0.771	0.731
	BCP properties (VIP > 1)	3	0.638	0.616	1	0.757	0.718	1	0.708	0.651	2	0.726	0.713
C	bond lengths	2	0.593	0.581	1	0.770	0.754	1	**0.783**	**0.758**	2	0.729	0.696
	BCP properties	2	0.660	0.637	1	0.770	0.747	1	0.779	0.750	3	0.805	0.744
	BCP properties (VIP > 1)	2	0.661	0.648	1	0.783	0.763	1	0.779	0.752	4	**0.823**	**0.791**
D	bond lengths	2	0.448	0.431	1	0.745	0.728	1	0.720	0.672	2	0.704	0.674
	BCP properties	6	**0.783**	**0.742**	1	0.769	0.752	1	0.759	0.724	5	0.815	0.768
	BCP properties (VIP > 1)	7	0.696	0.646	1	0.794	0.775	1	0.772	0.742	2	0.788	0.758
E	bond lengths	2	0.462	0.446	1	0.767	0.757	1	0.737	0.700	2	0.700	0.672
	BCP properties	6	0.766	0.731	1	0.767	0.754	1	0.749	0.710	5	0.813	0.763
	BCP properties (VIP > 1)	7	0.728	0.693	1	**0.792**	**0.782**	1	0.750	0.712	4	0.808	0.778

[a] Number of latent variables. [b] The bold text highlights the best models for each set (i.e all, ortho, para/meta and aliphatic) based on the highest q^2.

As with other multivariant statistical methods, the performance of SVM for regression depends on the combination of several parameters. We employed a Gaussian Radial Basis Function kernel for SVR because of its effectiveness and speed in the training process.[61] This function contains an extra parameter γ (a constant) that controls the amplitude of the Gaussian function, thereby controlling the generalization ability of the SVM to some extent. The user-prescribed parameters (i.e γ, ε of the ε-insensitive loss function and capacity parameter C) were chosen based on the lowest root mean squared error (RMSE) of the training data. The SVR programs were written by former group member C. X. Xue in an R-file, based on a script written in the R language for SVM, which utilized the e1071 package.[62] The scripts were compiled using the R 2.5.1 compiler[63] and run on a Pentium D PC with 1GB RAM.

2.4.3. Radial Basis Function Neural Networks (RBFNN). The subject of neural networks is covered in depth in the work by Haykin[64] and Gurney.[65] The theory of RBFNNs as applied to QSARs has been extensively described in the paper of Yao et al.[66] The training procedure involved the forward subset selection routine, which selected the centers for the RBF one at a time and adjusted the weights between the hidden layer and the output layer after the addition of each center, using a least-squares[67] solution. One third of the training set was randomly selected and 'held back' as test data, and training was terminated when the error on the test data showed no further improvement. RBFNN training was carried out using a range of RBF widths between 0.2 and 5.0, and the width yielding the lowest error on the test set was selected.

2.4.4. Comparison of the Methods. To compare these three machine learning methods we used r-fold cross-validation (CV), where the data sets were divided in 4, 7 (as implemented in SIMCA-P), and 10 CV groups. The division of the data sets was carried out using systematic sampling where the compounds were ordered according to their pK_a values and assigned to a group accordingly. For example, for 4-fold CV the first compound was grouped with the fifth, ninth, thirteenth, etc. In the random sampling method, the compounds were ordered by random numbers and divided into groups of different sizes depending on the r-fold CV being used (e.g., for 10-fold CV for the 50 ortho-substituted benzoic acids the first five compounds were group one, the next five group 2, etc.). Each CV group was excluded in turn so that each compound was excluded from the training data exactly once, and the RMSE of prediction and q^2 were calculated for the CV group according to the following equations

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_{obs,i} - \hat{y}_{pred,i})^2}{n}} \tag{1}$$

$$q2 = 1 - \frac{\sum_{i=1}^{n}(y_{obs,i} - \hat{y}_{pred,i})^2}{\sum_{i=1}^{n}(y_{obs,i} - \bar{y})^2} \tag{2}$$

where n is the number of observations in the CV set, $y_{obs,i}$ is the observed pK_a value for the molecule i in the CV set, $\hat{y}_{pred,i}$ is the associated predicted pK_a, and $\bar{y}$ is the mean pK_a value of the entire data set. The RMSE and q^2 values obtained for each CV group were then averaged with the values obtained from the corresponding groups to give final values for the global and subset models, for each CV group size, selected based on systematic and random sampling (see Table S2 in the Supporting Information for a complete table of the averaged RMSE and q^2 values). The method that produced the lowest RMSE in conjunction with the highest q^2 was considered to be the most accurate.

3. RESULTS AND DISCUSSION

3.1. Choice of Level of Theory. Table 1 shows a summary of the initial PLS analysis at the five different levels of theory for the data set and subsets. At each level, three different models were generated: a bond length only model, a model with bond length involving all BCP descriptors (amounting to 21, which derives from 3 bonds and 7 descriptors per bond), and a model including only those descriptors from the bond length and BCP descriptor model having a VIP score greater than one. At level A, only a bond length model can be generated. This is because semiempirical (AM1) wave functions do not contain core densities, which corrupts the topology by affecting the position or even appearance of BCPs.[25,68] The results in Table 1 are rather disappointing compared to the previous QTMS study of carboxylic acids,[41] where an r^2 and q^2 of 0.920 and 0.891 were obtained, respectively, and compared to the results in the literature discussed in the introduction earlier. Outlier detection was undertaken using the subset models as they were more easily distinguishable from correct predictions than in the models containing all the carboxylic acids. In all the para- and meta-substituted models, compound 37 (3,4-diaminobenzoic acid) was always an outlier. This compound was one of three zwitterions in the subset including compound 15 (3-aminobenzoic acid) and 21 (4-aminobenzoic acid). These compounds were predicted reasonably well (observed pK_a values of 4.74 and 4.85 and predicted pK_a values of 4.14 and 4.82, respectively) according to the best para- and meta-substituted model marked in bold in Table 1. However, compound 37 was predicted consistently poorly (observed pK_a of 3.49 and a predicted pK_a of 4.62 according to the same model). The zwitterions were all modeled in their neutral form, which sufficed for the monoamino-benzoic acids but was not appropriate for diaminobenzoic acid, which was predicted to have a larger pK_a value due to the fact that the increased stability in its zwitterionic form is not encapsulated in the BCP descriptors. As in this work, Tehan[1] and co-workers struggled to model this effect and had few problems with the monoaminobenzoic acids. In line with Tehan we omitted 37 as an outlier. In the ortho-substituted models, two compounds were identified as outliers and removed from the models, namely 72 (2,6-dihydroxybenzoic acid) and 79 (2-hydroxy-3,5-dinitrobenzoic acid) whose pK_a values were predicted to be 2.60 and 1.86 compared to observed values of 1.05 and 0.70, respectively. The hydroxyl group at the ortho position(s) in these compounds could be held responsible for their overprediction because internal hydrogen bonding in the anionic form could increase the stability of the ion therefore decreasing their pK_a values. This effect is not encapsulated in the BCP descriptors and therefore absent from the models. The issue with this reasoning is that there are a further 11 compounds in the subset that are hydroxyl-substituted at the 2 position and they are predicted well.

Four further compounds were identified as outliers from the aliphatic carboxylic acid models and removed. They were 124 (4-[(4-chloro-2-methylphenyl)oxy]butanoic acid), 150 (cyanoacetic acid), 155 (9-hydroxy-9H-fluorene-9-carboxylic acid/flurenol), and 228 (4-(cyclopropylcarbonyl)-3,5-dioxo-cyclohexanecarboxylic acid)). Tehan[1] and co-workers brought into question the reliability of the observed pK_a value of compound 124, with which we concur. Compound 155 (Figure 2a) was excluded from their model on the basis that the proximity of the carboxyl group to the two aromatic rings and the presence of an α-hydroxy group make its pK_a difficult to predict. Alternatively, the observed pK_a is incorrect. We can confirm that it is the *observed* pK_a value that is the most likely cause for discrepancy. Our best model predicts the pK_a to be 2.87, while the experimental value given by Tehan (our source data set) is 1.09. A different source[69] gives the observed pK_a value of 2.96, which is close to our predicted value and the value predicted by ACD/Laboratories[23] as 3.04. Furthermore, there is a similar structure (104, hydroxy-(diphenyl)acetic acid) in the data set that has an observed pK_a of 3.05 (Figure 2b). A literature value for the observed pK_a values of these compounds (155 and 104) with their respective hydroxyl group removed was found to be 3.61[70] for 155 (Figure 2c) and 3.9[71] for 104 (Figure 2d), a difference of only 0.3 log units. The difference of 1.96 log units (=3.05−1.09) between compounds 155 and 104 generated by the addition of a hydroxyl group at the same position in each is unlikely when considering the difference is 0.3 log units between the analogous compounds, thus further supporting a wrong observed pK_a. No reason for compounds 150 and 228 being outliers can be offered, but they were both excluded.

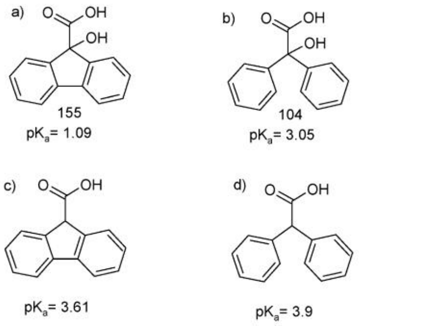


Figure 2. Structures and observed pK_a values for compounds 155 and 104 and their analogues fluorene-9-carboylic acid and diphenylacetic acid.

Table 2 shows the results of the PLS analysis after the outliers were removed. Here, one of the BCP property models always outperforms the bond length model at each level of theory in terms of both r^2 and q^2. Generally, the models improved when the VIP < 1 "cut-off" was used. Level *E* gave the best results out of the all acid models. After observing the r^2 and q^2 for all of the models we found that level *C* gave the best results for the subsets. However, level *E* gave the best results for a model involving *all* carboxylic acids, although this model is constructed from 6 LVs. Figure 3 shows a comparison of the CPU time needed for optimization for nine of the compounds (three from each subset) that converged at each level of theory without any restarts versus the highest q^2 from each level of theory. This demonstrates that the less computationally expensive level *C* provides better results for the subsets and only worse results for the all acid models. Level *A* and *B* are inferior in all cases. At this stage we are in a position to explore different statistical learning methods to predict pK_a values based of BCP properties. To test the suitability of Level *C* in pK_a prediction and to examine whether the initial results still hold we carried out the analysis with level *C* and the more computationally expensive level *E*.

3.2. The Data for all the Comparisons in Section 3.2. to 3.5. can be Found in Table S2 in the Supporting Information. Comparison of the Statistical Learning Methods. For each learning method, there are four groups of compounds to test: all carboxylic acids, aliphatic acids, meta/para- and ortho-substituted acids. There are three different sizes of CV sets for both the systematic and random sampling method. This gave 24 values (= 4 groups of compounds × 3 CV set sizes × 2 sampling methods) of q^2 and RMSE to compare for each learning method employed. At level *C*, SVM gave the lowest RMSE value 16 times out

Table 2. Summary of the Initial PLS Analysis after the Outliers Had Been Removed[b]

		all acids			ortho subset			para/meta subset			aliphatic subset		
level	descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.630	0.612	2	0.797	0.787	1	0.900	0.896	2	0.795	0.748
B	bond lengths	2	0.601	0.581	2	0.838	0.801	1	0.905	0.902	2	0.837	0.828
	BCP properties	4	0.769	0.730	1	0.837	0.810	1	0.912	0.911	2	0.863	0.852
	BCP properties (VIP > 1)	4	0.764	0.749	1	0.835	0.815	1	0.909	0.908	6	0.895	0.861
C	bond lengths	2	0.696	0.688	1	0.841	0.832	1	0.931	0.929	1	0.875	0.875
	BCP properties	3	0.802	0.779	1	0.837	0.819	1	0.931	0.928	3	0.897	0.873
	BCP properties (VIP > 1)	2	0.787	0.782	1	**0.848**	**0.836**	1	**0.933**	**0.932**	4	**0.911**	**0.901**
D	bond lengths	2	0.486	0.455	1	0.777	0.762	1	0.912	0.906	3	0.807	0.784
	BCP properties	6	0.860	0.839	2	0.851	0.813	1	0.923	0.919	3	0.882	0.869
	BCP properties (VIP > 1)	4	0.738	0.689	1	0.842	0.833	1	0.927	0.925	2	0.890	0.881
E	bond lengths	2	0.535	0.498	1	0.803	0.794	2	0.917	0.907	2	0.810	0.788
	BCP properties	6	**0.879**	**0.844**	2	0.866	0.818	1	0.917	0.913	2	0.875	0.858
	BCP properties (VIP > 1)	7	0.845	0.818	1	0.844	0.835	1	0.918	0.913	3	0.897	0.886

[a] Number of latent variables. [b] The bold text highlights the best models for each group based on having the highest q^2.

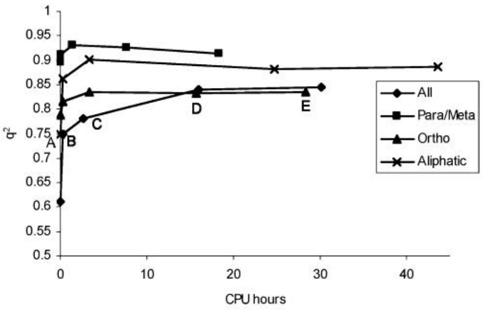


Figure 3. Comparison of the CPU times needed to optimize the compounds versus the highest q^2 value at each level of theory (*A*, *B*, *C*, *D*, and *E*).

of 24 comparisons, PLS 7 out of 24 comparisons, and RBFNN 2 times out of 24 comparisons (the RMSE for the random sampling para/meta 10-fold CV models were identical for PLS and SVM). Again out of 24 comparisons but at level *E* this time, SVM gave the lowest RMSE 21 times, PLS 3 times, and RBFNN 2 times (the RMSE for the systematic sampling para/meta 4-fold CV models and the random sampling 7-fold CV models were identical for PLS and SVM). This clearly demonstrates that SVM is superior to both PLS and RBFNN, at both levels of theory.

3.3. Comparison of the Level of Theory. Comparing the results of each q^2 and RMSE values obtained at both levels of theory, level *C* produces the highest q^2 46 times out of 72 (=3 × 24) comparisons and the lowest RMSE 45 times out of the 72 comparisons, with one value being the same for level *C* and *E*. This clearly demonstrates that level *C* is superior to level *E*.

3.4. Comparison of the Validation Set Selection Methods. At each level of theory, there are 36 RMSE and q^2 values to compare because there are 3 machine learning methods, 3 values for *r* in r-fold CV, and 4 compound groups (36=3 × 3 × 4). At level *C*, there are 24 higher q^2 values for systematic sampling compared to 11 for random sampling and 1 q^2 value that is the same (36=24 + 11 + 1). There are 22 RMSE values lower for systematic sampling compared to 13 lower RMSE values for random sampling and 1 RMSE value that is the same (36=22 + 13 + 1). At level *E*, there are 19 higher q^2 values for systematic sampling compared to 16 for random sampling and 1 q^2 value that is the same (36=19 + 16 + 1). There are 21 RMSE values that are lower for systematic sampling compared to 15 values for random sampling (36=21 + 15). The better results gained from systematic sampling is not surprising because this method ensures the maximum value for the denominator in the q^2 equation. Using systematic sampling means that the most dissimilar compounds in relation to their pK_a values are excluded, and so it is more likely that the training set will contain similar compounds to the CV set therefore leading to better predictions.[72] When random sampling is used, one cannot ascertain that the CV sets contain all similar compounds in terms of their pK_a values. Hence, the models are possibly trained using compounds dissimilar to the CV set, which can lead to poor predictions. If the CV sets contain similar compounds in terms of their pK_a values, then this can lead to a small denominator in eq 1, thus reducing the q^2 value.

3.5. Comparison of Validation Set Size. With regards to the r-fold validation, 10-fold validation generally provided the highest q^2 and the lowest RMSE values, although this was not always the case. When the difference is calculated between the highest and lowest q^2 values for the r-fold CV set size for each subset and each method, the mean difference is 0.057 (standard deviation (SD) of 0.127) and 0.039 (SD=0.042) for levels *C* and *E*, respectively. When the same is carried out for the RMSE values, then the mean difference is 0.041 (SD=0.040) and 0.042 (SD=0.042) for levels *C* and *E*, respectively. When the poor result for 4-fold-para/meta-random sampling using RBFNN at level *C* is omitted, then the mean difference for q^2 is 0.032 (SD=0.025), and the mean value for RMSE is 0.035 (SD=0.024). The r-fold CV does give different results, but the difference is small. It is not surprising that 10-fold CV generally gives the best validation statistics because the smaller the CV groups, the more compounds there are to train the models. What these results do suggest is that the models still provide good predictions even when 25% of the compounds are omitted in training when using 4-fold CV.

3.6. Confirmation of Findings Based on Averages. Table 3 and Figure 4 show the results of all the subsets and learning methods when the r-fold CV results have been averaged and the results of the subsets and learning methods when the sampling methods have been averaged (from Table S2 in

Table 3. Average Values of the Results from Table S2

		PLS		SVM		RBFNN	
	set	q^2	RMSE	q^2	RMSE	q^2	RMSE
Level *C*							
systematic sampling	all	0.780	0.422	0.893	0.291	0.897	0.293
	ortho	0.834	0.375	0.821	0.380	0.784	0.427
	para/meta	0.930	0.109	0.924	0.113	0.896	0.128
	aliphatic	0.906	0.296	0.903	0.271	0.891	0.294
random sampling	all	0.777	0.430	0.879	0.295	0.875	0.322
	ortho	0.821	0.377	0.829	0.375	0.774	0.426
	para/meta	0.928	0.110	0.922	0.110	0.634	0.212
	aliphatic	0.904	0.270	0.909	0.265	0.885	0.303
average of systematic and random sampling	all	0.778	0.426	0.886	0.293	0.886	0.307
	ortho	0.828	0.376	0.825	0.378	0.779	0.427
	para/meta	0.929	0.109	0.923	0.112	0.765	0.170
	aliphatic	0.905	0.283	0.906	0.268	0.888	0.298
Level *E*							
systematic sampling	all	0.817	0.381	0.880	0.291	0.872	0.327
	ortho	0.825	0.386	0.851	0.356	0.763	0.431
	para/meta	0.913	0.129	0.916	0.125	0.893	0.144
	aliphatic	0.884	0.299	0.905	0.267	0.852	0.336
random sampling	all	0.805	0.392	0.887	0.295	0.881	0.316
	ortho	0.861	0.389	0.822	0.380	0.686	0.508
	para/meta	0.896	0.136	0.912	0.127	0.820	0.176
	aliphatic	0.881	0.300	0.903	0.273	0.886	0.301
average of systematic and random sampling	all	0.811	0.387	0.883	0.293	0.876	0.322
	ortho	0.843	0.388	0.837	0.368	0.724	0.469
	para/meta	0.905	0.132	0.914	0.126	0.857	0.160
	aliphatic	0.883	0.300	0.904	0.270	0.869	0.319

the Supporting Information). These results confirm what had previously been suggested. There is little difference between the CV statistics for the random and systematic sampling methods. Excluding the RBFNN para/meta results at level *C*, the largest difference in q^2 and RMSE values is 0.021 and 0.029, respectively, for the RBFNN all acid models. At level *E*, the largest difference in q^2 and RMSE is 0.078 and 0.077, respectively, for the RBFNN ortho models. Based on

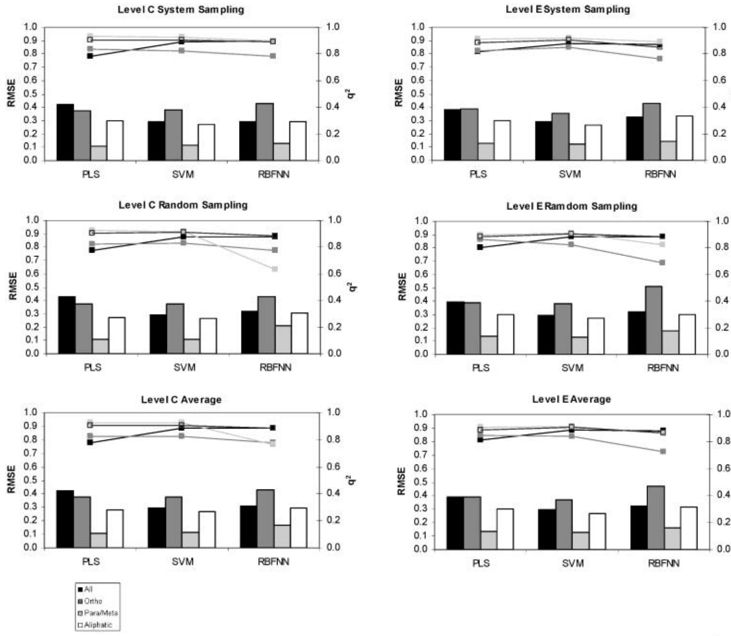


Figure 4. A graphical representation of Table 3. The bar charts represent the RMSE, and the lines represent the q^2 values obtained.

Table 4. RMSE for the Three Learning Methods Based on Leave-One-out[a]

	RMSE		
method	PLS	SVM	RBFNN
Level *C*			
C QTMS (all)	0.427	0.293	0.363
C QTMS (subset avg.)	0.285	0.276	0.321
C QTMS ortho subset	0.388	0.407	0.477
C QTMS para/meta subset	0.121	0.123	0.138
C QTMS aliphatic subset	0.278	0.252	0.291
Level *E*			
E QTMS (all)	0.396	0.301	0.323
E QTMS (subset avg.)	0.311	0.278	0.323
E QTMS ortho subset	0.407	0.367	0.486
E QTMS para/meta subset	0.136	0.131	0.168
E QTMS aliphatic subset	0.313	0.276	0.285
compared software/tools	RMSE		
ACD/Laboratories	0.263		
VCCLAB	0.279		
SPARC	0.356		
ChemAxon	0.398		

[a] The RMSE for the commercial computer programs are given at the bottom of the table.

the final average results of the r-fold and sampling methods, it is clear that SVM generally gives the best CV statistics. However, at level *C* the PLS statistics for the para/meta and ortho models provide the highest q^2 value and lowest RMSE. However, these are close to the q^2 value and RMSE value provided by SVM. Comparing the "all acid" models and the aliphatic models the q^2 value and RMSE for SVM are much better than the q^2 and RMSE for PLS. At level *E*, SVM provides the highest q^2 and lowest RMSE in all cases. Although PLS was better than SVM, in two cases at level *C* we chose SVM as the best learning method. This decision is based on the fact that, when PLS was superior, the difference between the statistics was small and, when SVM was better, then the difference between the statistics was large. Comparing the averages of systematic and random sampling for SVM level *E* only provides better CV statistics for the ortho data set. The difference between the statistics is small, but this may suggest that the more expensive level *E* accounts more for the steric effects than level *C* in some way.

Improved models were created when the data set was split into aliphatic and aromatic subsets, which were further split into meta/para and ortho-substituted carboxylic acids. The subset models provided significant improvements on the "all acid" model. The para/meta model was the most accurate in prediction, followed by the aliphatic and then the ortho model. The excellent CV results of the para/meta models are not surprising because the pK_a difference caused by substituent changes can be accurately predicted by the Hammett equation and BCP properties display a strong correlation with Hammett's sigma parameter.[25] The lower q^2 value and higher RMSE for the ortho model can be explained in terms of steric effects.[73] Whereas the pK_a of meta- and para-substituted carboxylic acids is affected mainly by inductive and resonance contributions, the pK_a of ortho-substituted carboxylic acids is highly sensitive to steric contributions. Primary steric hindrance to deprotonation is important where there are bulky groups around the acidic center. Secondary steric effects may either be acid-weakening (if there is steric hindrance to solvation) or acid-strengthening (if there is steric inhibition of resonance in the neutral molecule). We have already stated that when QTMS fails one can be certain that steric effects are very important. Since the BCP properties do not account for steric effects, we can be confident that this is the reason for the poorer results. Since the results of the aliphatic subset are an improvement on the ortho subset, the steric effects must have less importance than electronic contributions.

3.7. Comparison to pK_a Prediction Software. To compare pK_a prediction by the QTMS method with available (commercial) software, we have used the methods provided by a number of organizations, namely ACD/Laboratories' pK_a DB,[23] the SPARC[74] online calculator (SPARC Performs Automated Reasoning in Chemistry), VCCLAB's[75] Web-based ALOGPS 2.1 program, and ChemAxon's pK_a plugin[76] for their Marvin Beans applications. Recently, it was stated that the Web version of the SPARC performs 50,000−100,000 calculations per month.[77] Each software package enables the user to input the structures in SMILES[78] format.

We removed each of the compounds in turn from the global and subset carboxylic acid models built using PLS, SVM, and RBFNN, for both levels *C* and *E* and rebuilt the models using the new model to predict the pK_a of the compound omitted.[79] Using this method we acknowledge that the compounds are not an external test set (e.g. they have been used for initial variable and parameter selection in some cases), but nor can we be sure that they have not been used to train the packages we compare to. Table 4 gives the RMSE for the methods based on leave-one-out. The RMSE obtained from testing the alternative computer programs are also given in Table 4. These results confirm that SVM provides the best models to predict pK_a. The SVM models have the lowest RMSE in all the LOO cases apart from the level *C* ortho and para/meta carboxylic acid models, where the PLS models have the lowest RMSE of 0.388 for the ortho model and 0.121 for the para/meta model, compared to 0.407 and 0.123 for the SVM models, respectively. Comparing levels of theory confirms that level *C* is the best as it generally provides the lowest RMSE for the models. There are some exceptions to this. For example, the ortho

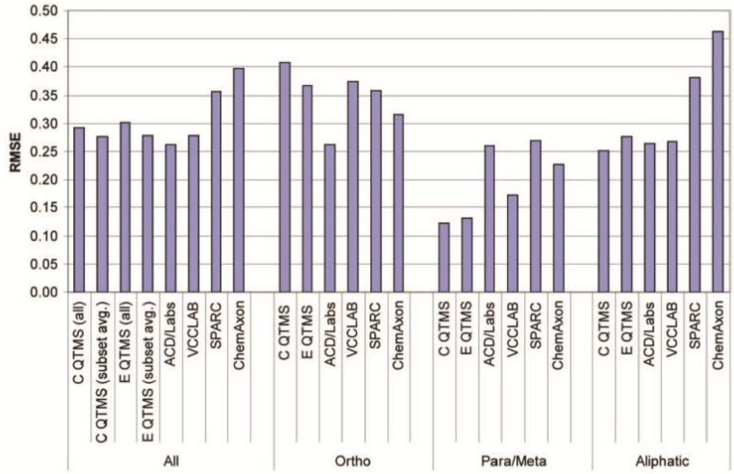


Figure 5. Comparison between QTMS and other pK_a prediction software, based on the RMSE.

model at level *E* using SVM has an RMSE of 0.367 compared to 0.407 for level *C*. Where level *E* provides a lower RMSE the largest difference observed in RMSE between levels *C* and *E* is 0.04 for the SVM ortho models and the PLS all carboxylic acid models. In fact, the difference between the models at the different levels judged by RMSE (based on LOO) is negligible when considering the large increase in CPU time needed to optimize the compounds at level *E* (see Figure 3).

Figure 5 graphically compares our QTMS SVM models to the results obtained from the commercial predictions. Recently, Meloun and Bordovská[77] have rigorously compared the same packages using 64 drugs and other organic molecules with complex and diverse structural patterns. Although we only base our ranking on the RMSE, we too found ACD/pK_a to be the most accurate method. This conclusion contradicts the findings of Dearden et al.[80] who compared ten prediction tools using a test set of 653 compounds. They found that ACD/pK_a was the *least* accurate out of the four programs we compared with, even when tautomeric compounds were excluded. The other three programs were ordered in the same way as our results: VCCLAB being the most accurate, followed by SPARC, and then ChemAxon. Apart from ACD/Laboratories, which is consistent across all the subsets, the different methods vary in their prediction ability. Out of all the methods, QTMS has the lowest RMSE for the para/meta substituted benzoic acids and aliphatic carboxylic acids but has the highest RMSE for the ortho-substituted benzoic acids. As has been previously pointed out,[33,45] QTMS fails when steric effects are important, which is the case for the ortho substituted benzoic acids.

4. CONCLUSION

The results presented in this systematic study indicate that BCP descriptors are effective in predicting the pK_a of small to large sized carboxylic acids of pharmaceutical relevance. Furthermore, extensive cross-validation shows that there is no need to use the computationally more expensive level *E* when level *C* provides similar, if not superior, CV statistics. More predictive models were gained from splitting the data set. Generally, SVM provides the best learning method although the lack of interpretability may mean it is not necessarily the most suitable method when mechanistic understanding is important. Finally, we have also demonstrated that predictions from our QTMS method compete with contemporary and frequently used pK_a prediction tools.

ACKNOWLEDGMENT

The work was funded by the EPSRC with sponsoring from GlaxoSmithKline (GSK), Stevenage, Great Britain. Thanks to Dr. M. Kranz at GSK for helpful discussions.

Supporting Information Available: Experimental pK_a values for the carboxylic acids (Table S1) and results for the three machine learning methods (PLS, SVM, RBFNN), using r-fold CV ($r = 4$, 7, and 10), and the systematic and random sampling method, at levels *C* and *E* (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pKa Using Semiemperical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457–471.

(2) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution I. Carboxylic Acids. *J. Phys. Chem. A* **1999**, *103*, 11194–11199.

(3) Cookson, R. F. The determination of acidity constants. *Chem. Rev.* **1974**, *74*, 5–28.

(4) Namazian, M.; Halvani, S. Calculations of pK_a values of carboxylic acids in aqueous solution using density function theory. *J. Chem. Thermodyn.* **2006**, *38*, 1495–1502.

(5) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.

(6) Gruber, C.; Buss, V. Quantum-Mechanically Calculated Properties for the Development of Quantitative-Structure Activity Relationships. *Chemosphere* **1989**, *19*, 1595–1609.

(7) Citra, M. J. Estimating the pKa of Phenols, Carboxylic Acids and Alcohols from Semi-Empirical Quantum Chemical Methods. *Chemosphere* **1999**, *38*, 191–206.

(8) Gross, K. C.; Seybold, P. G. Substituent Effects on the Physical Properties and pKa of Phenols. *Int. J. Quantum Chem.* **2001**, *85*, 569–579.

(9) Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK_a and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(10) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by Molecular Tree Structure Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(11) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pK_a prediction method using grid molecular interaction firlds. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(12) Kogej, T.; Muresan, S. Database Mining for pK_a Prediction. *Curr. Drug Discovery Technol.* **2005**, *4*, 221–229.

(13) Liptak, M. D.; Shields, G. C. Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* **2001**, *123*, 7314–7319.

(14) Liptak, M. D.; Shields, G. C. Experimentation with Different Thermodynamic Cycles Used for pK_a Calculations on Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *Intern. J. Quant. Chem.* **2001**, *85*, 727–741.

(15) Lu, H.; Chen, X.; Zhan, C. G. First-principles calculation of pK_a for cocaine, nicotine, neurotransmitters, and anilines in aqueous solution. *J. Phys. Chem. B* **2007**, *111*, 10599–10605.

(16) Soriano, E.; Cerdan, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theochem)* **2004**, *684*, 121–128.

(17) *Jaguar version 6.0*; Schrodinger, LLC: New York, USA, 2005.

(18) Namazian, M.; Halvani, S. Calculations of pK_a Values of Carboxylic Acids in Aqueous Solution Using Moller-Plesset Perturbation Theory. *J. Iran. Chem. Soc.* **2005**, *2* (1), 65–70.

(19) Namazian, M.; Halvani, S.; Noorbala, M. R. Density Functional Theory Response to the Calculations of pK_a Values of some Carboxylic Acids in Aqueous Solution. *J. Mol. Struct.* **2004**, *711*, 13–18.

(20) Namazian, M.; Heidary, H. Ab Initio Calculations of pK_a Values of some Organic Acids in Aqueous Solution. *J. Mol. Struct.* **2003**, *620*, 257–263.

(21) Namazian, M.; Kalantary-Fotooh, F.; Noorbala, M. R.; Searles, D. J.; Coote, M. L. Moller-Plesset Perturbation Theory Calculations of the pK_a Values for a Range of Carboxylic Acids. *J. Mol. Struct. THEOCHEM* **2006**, *758*, 275–278.

(22) da Silva, C. O.; da Silva, E. C.; M, A. C. Nascimento, Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution II. Aliphatic Alcohols, Thiols, and Halogenated Carboxylic Acids. *J. Phys. Chem. A* **2000**, *104*, 2402–2409.

(23) ACD/Labs, version 3;ACD Labs, Toronto, ON, Canada.

(24) O'Brien, S. E.; Popelier, P. L. A. Quantum Molecular Similarity. Part 3: QTMS descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764–775.

(25) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103* (15), 2883–2890.

(26) Bader, R. F. W. *Atom in Molecules. A Quantum Theory*; Oxford Univ. Press: 1990.

(27) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education: London, Great Britain, 2000.

(28) O'Brien, S. E.; Popelier, P. L. A. Quantum Topological Molecular Similarity. Part 4: A QSAR study of Cell Growth Inhibitory Properties of Substituted (E)-1-Phenylbut-1-en-3-ones. *J. Chem. Soc., Perkin Trans. 2* **2002**, 478–483.

(29) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. Quantitative Structure-Activity Relationships of Mutagenic Acitvity from Quantum Topological Descriptors: Triazenes and Halogenated Hydroxfuranones (mutagen-X) Derivatives. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 709–718.

(30) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. Quantum Topological Molecular Similarity. Part 5: Further Development with an Application to the toxicity of Polychlorinated dibenzo-p-dioxins (PCDDs). *J. Chem. Soc., Perkin II* **2002**, 1231–1237.

(31) Mohajeri, A.; Hemmateenejad, B.; Mehdipour, A.; Miri, R. Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS. *J. Mol. Graphics Modell.* **2008**, *2008*, 1057–1065.

(32) Roy, K.; Popelier, P. L. A. Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2604–2609.

(33) Popelier, P. L. A.; Smith, P. J. QSAR models based on Quantum Topological Molecular Similarity. *Eur. J. Med. Chem.* **2006**, *41*, 862–873.

(34) Hemmateenejad, B.; Mehdipour, A. R.; Popelier, P. L. A. Quantum Topological QSAR Models based on the MOLMAP Approach. *Chem. Biol. Drug Des.* **2008**, *72*, 551–563.

(35) Roy, K.; Popelier, P. L. A. Exploring Predictive QSAR Models Using Quantum Topological Molecular Similarity (QTMS) Descriptors for Toxicity of Nitroaromatics to Saccharomyces cerevisiae. *QSAR Comb. Sci.* **2008**, *27*, 1006–1012.

(36) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. A New 3D Molecular Structure Representation using Quantum Topology with application to structure-property relationships. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 75–91.

(37) Chaudry, U. A.; Popelier, P. L. A. Ester hydrolysis rate constant prediction from quantum topological molecular similarity (QTMS) descriptors. *J. Phys. Chem. A* **2003**, *107*, 4578–4582.

(38) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. Modeling quantitative structure-property relationships in calculated reaction pathways using a new 3D quantum topological representation. *Anal. Chim. Acta* **2001**, *446*, 3–13.

(39) Hemmateenejad, B.; Mohajeri, A. Application of quantum topological molecular similarity descriptors in QSPR study of the O-methylation of substituted phenols. *J. Comput. Chem.* **2007**, *29*, 266–274.

(40) Roy, K.; Popelier, P. L. A. Predictive QSPR modeling of the acidic dissociation constant (pKa) of phenols in different solvents. *J. Phys. Org. Chem.* **2009**, *22*, 186–196.

(41) Chaudry, U. A.; Popelier, P. L. A. Estimation of pKa using Quantum Topological Molecular Similarity (QTMS) descriptors: Application to Carboxylic Acids, Anilines and Phenols. *J. Org. Chem.* **2004**, *69*, 233–241.

(42) Adam, K. R. New Density Functional and Atoms in Molecules Method of Computing Relative pKa Values in Solution. *J. Phys. Chem. A* **2002**, *106*, 11963–11972.

(43) Bader, R. F. W.; Beddall, P. M. The Spatial Partitioning and Transferability of Molecular Energies. *Chem. Phys. Lett.* **1971**, *8*, 29–36.

(44) Esteki, M.; Hemmateenejad, B.; Khayamian, T.; Mohajeri, A. Multi-way analysis of Quantum Topological Molecular Similarity descriptors for modelling acidity constants of some phenolic compounds. *Chem. Biol. Drug Des.* **2007**, *70*, 413–423.

(45) Loader, R. J.; Singh, N. K.; O'Malley, P. J.; Popelier, P. L. A. The Cytotoxicity of ortho alkyl substituted 4-X-phenols: A QSAR based on theoretical bond lengths and electron densities. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1249–1254.

(46) Selassie, C. D.; Verma, R. P.; Kapur, S.; Shusterman, A. J.; Hansch, C. QSAR for the cytotoxicity of the radical reaction. *J. Chem. Soc., Perkin Trans. II* **2002**, 1112–1117.

(47) UMETRICS. *SIMCA-P 10.0*; Umeå, Sweden, 2002. info@umetrics.com, www.umetrics.com (accessed month day, year).

(48) Howard, S. T.; Lamarche, O. Description of covalent bond orders using the charge density topology. *J. Phys. Org. Chem.* **2003**, *16*, 133–141.

(49) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. Description of Conjugation and Hyperconjugation in Terms of Electron Distributions. *J. Am. Chem. Soc.* **1983**, *105* (15), 5061–5068.

(50) Bader, R. F. W.; Preston, H. J. T. The Kinetic Energy of Molecular Charge Distributions and Molecular Stability. *Int. J. Quantum Chem.* **1969**, *3*, 327–347.

(51) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123–134.

(52) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, J., T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian; 2003.

(53) Popelier, P. L. A. *MORPHY98*; MORPHY98 - a program written by P. L. A. Popelier with a contribution from R. G. A. Bone. UMIST: Manchester, England, 1998.

(54) Popelier, P. L. A. A Robust Algorithm to Locate Automatically All Types of Critical- Points in the Charge-Density and Its Laplacian. *Chem. Phys. Lett.* **1994**, *228* (1−3), 160–164.

(55) Wold, S.; Sjostrom, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R., Ed.; Wiley: Chichester, GB, 1998; Vol. 3, pp 2006−2021.

(56) Livingstone, D. J. *Data Analysis for Chemists*, 1st ed; Oxford University Press: Oxford, Great Britain, 1995.

(57) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer-Verlag: New York, USA, 1995.

(58) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowl. Disc.* **1998**, *2*, 121–167.

(59) Smola, A. J.; Schoelkopf, B. A tutorial on support vector regression. *Stat. Comp.* **2004**, *14* (3), 199–222.

(60) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH: 2007; Vol. *23*, pp 291−400.

(61) Jover, J.; Basque, R.; Sales, J. QSPR prediction of pK_a for benzoic acids in different solvents. *QSAR Comb. Sci.* **2008**, *27* (5), 563–581.

(62) Dimitriadou, E.; Hornik, K.; Leisch, D. M. F.; Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071)*; R package version 1.5−16; TU Wien: 2006.

(63) Team, R. D. C. R: A language and environment for statistical computing; R Foundation for Statistical Computing: Vienna, Austria, 2007. ISBN 3-900051-07-0. URL: http://www.R-project.org (accessed month day, year).

(64) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice-Hall: Upper Saddle River, New Jersey, USA, 1999.

(65) Gurney, K. *An Introduction to Neural Networks*; Routledge: London, Great Britain, 1997.

(66) Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Lui, M. C.; Hu, Z. D.; Fan, B. T. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217–225.

(67) Chen, S.; Cowan, C.; Grant, P. Orthogonal least squares learning for radial basis function networks. *IEEE Trans. Neural Networks* **1991**, *2*, 302–309.

(68) Ho, M.; Schmider, H.; Edgecombe, K. E.; Smith, V. H. J. Topological Analysis of Valence Electron Charge Distributions from Semi-empirical and ab initio methods. *Int. J. Quant. Chem., Quant. Chem. Symp.* **1994**, *28*, 215–226.

(69) http://www.chemicaldictionary.org/dic/F/Flurenol_2074.html (accessed month day, year).

(70) Kresge, A. J.; Pojarlieff, I. G.; Rubinstein, E. M. The acidity constant of fluorene-9-carboylic acid in aqueous solution. Determination of the pKa, of a sparingly soluble substance. *Can. J. Chem.* **1993**, *71*, 227.

(71) Jumppanen, J. H.; Siren, H.; Riekkola, M. L. Correlation of Resolution with Frictional Coefficients and pK, Values in Capillary Electrophoresis of Four Diuretics: Determination of Electric Field Strength and Electro-Osmotic Velocity. *J. Microcolumn Separation* **1993**, *5* (5), 451.

(72) Leonard, J. T.; Roy, K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.* **2006**, *25*, 235–251.

(73) Perrin, D. D.; Dempsey, B.; Serjean, E. P. *pKa Prediction for Organic Acids and Bases*; Chapman and Hall: London, GB, 1981.

(74) Hilal, S. H.; Karickoff, S. W.; Carreira, L. A. A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK_a's. *Quant. Struc. Act. Relat.* **1995**, *14* (348), .

(75) VCCLAB. *Virtual Computational Chemistry Laboratory*; 2005. http://www.vcclab.org (accessed month day, year).

(76) ChemAxon. *Calculator Plugins were used for structure property prediction and calculation, Marvin 2.0.4, 2006*; ChemAxon. http://www.chemaxon.com (accessed month day, year).

(77) Meloun, M.; Bordovska, S. Benchmarking and Validating algorithms that estimate pK_a values of drugs based on their molecular structure. *Anal. Bioanal. Chem.* **2007**, *389*, 1267–1281.

(78) Weininger, D. SMILES, a Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(79) Zhang, J.; Kleinoder, T.; Gasteiger, J. Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.

(80) Dearden, J. C.; Cronin, T. D.; Lappin, D. C. A comparison of commercially avaliable software for the prediction of pK_a values (poster). At UK-QSAR, 24 April 2007, Astrazeneca, Alderley Park, UK, 2007.

CI900172H

SAR and QSAR in Environmental Research
Vol. 21, Nos. 1–2, January–March 2010, 149–168

QSAR with quantum topological molecular similarity indices: toxicity of aromatic aldehydes to *Tetrahymena pyriformis*

S. Kar[a], A.P. Harding[b], K. Roy[a]* and P.L.A. Popelier[b]*

[a]*Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India;* [b]*Manchester Interdisciplinary Biocenter (MIB), Manchester, UK and School of Chemistry, University of Manchester, Manchester, UK*

(*Received 7 August 2009; in final form 2 December 2009*)

Extensive production and utilization of aromatic aldehydes and their derivatives without proper certification is alarming with regard to environmental safety. This concern motivated our construction of predictive quantitative structure–activity relationship (QSAR) models for the toxicity of aldehydes to the ecologically important species *Tetrahymena pyriformis*. Quantum topological molecular similarity (QTMS) descriptors, along with the lipid-water partition coefficient (log $K_{o/w}$), were used as predictor variables. The QTMS descriptors were calculated at different levels of theory including AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31 + G(d,p), B3LYP/6-311 + G(2d,p) and MP2/6-311+ G(2d,p). The data set of 77 aromatic aldehydes was divided into a training set ($n = 58$) and a test ($n = 19$) set, and 58 models were developed using partial least squares (PLS) and genetic partial least squares (G/PLS). We evaluated the overall predictive capacity of the models based on leave-one-out predictions for the training set compounds and model derived predictions for the test set compounds. For both PLS and G/PLS, the models built at the HF/6-31G(d) level show better predictivity (based on overall prediction) than the models developed at any of the other five levels. Further validation was also performed utilizing (process and model) randomization tests. We show that improved predictive QSAR models for aldehydic toxicity to *Tetrahymena pyriformis* can be generated using QTMS descriptors along with log $K_{o/w}$.

Keywords: QTMS; QSAR; *ab initio*; aldehydes; QSPR; external validation; electron density; atoms in molecules; quantum chemical topology

1. Introduction

In the twenty-first century, we are exposed daily to numerous environmental toxins and pollutants [1]. Multiple chemical sensitivity can cause asthma, chest pains, and hives among other symptoms. Each year, industry adds 10,000 newly synthesized chemicals to the over 1 million already in existence [1 4]. However, most of the commercially used chemicals do not have proper documentation for their toxicologically relevant properties [1]. Thus, it is vital to have methods to assess the effects of these compounds on the environment and on human health.

There is a growing need to use *in silico* methods to provide information about the physicochemical properties of chemicals, their environmental fate as well as their human

*Corresponding authors. Emails: pla@manchester.ac.uk; kunalroy_in@yahoo.com

ISSN 1062 936X print/ISSN 1029 046X online

DOI: 10.1080/10629360903568697
http://www.informaworld.com

health effects. Computer-aided toxicity prediction makes use of the relationship between chemical structure and biological activity to compute such properties. Quantitative structure activity relationships (QSARs) enable predictions of properties or effects to be made directly from chemical structure [5]. Such approaches have proven capabilities when applied to well-defined toxicity endpoints and/or regions of chemical space. In practice, the ways in which these approaches are used depends on the requirements of the specific legislation and the possibilities offered by regulatory authorities.

Aldehydes are an important class of environmental and industrial chemicals of wide-spread use causing tissue damage, cytotoxicity, mutagenicity and carcinogenicity leading to various health complications [6]. They are important intermediates in the production of a variety of industrial processes in the agrochemicals and pharmaceuticals industries and are key in the flavour and fragrance industry [7]. Their inherent reactivity means that they are able to interact with the electron-rich biological macromolecules, including proteins and nucleic acids, and therefore have the potential to cause a number of adverse effects. Due to the many origins and magnitude of uses, aldehydic compounds are widespread in ecosystems and consequently have a high potential for environmental pollution. It is therefore useful to develop theoretical models to predict the toxicity of aldehydes against different ecologically relevant organisms [8].

The unicellular ciliated protozoan, *Tetrahymena pyriformis*, displays fast growth rates and inexpensive assays [8]. Hence it is an attractive candidate for the assessment of the environmental impact of toxicants. In addition to the environmental safety, toxicity to this organism has proven useful in estimating the toxic potencies of chemicals [9 11]. For the development of toxicity models, *T. pyriformis* has been extensively used in the last three decades. Cronin and Schultz developed QSAR models for the prediction of the toxicity of phenols to *T. pyriformis* using 166 phenols [12], and then later refined the models using 250 phenolic compounds [13]. Roy and coworkers constructed QSAR models for six different groups of aliphatic compounds such as alcohols, esters, acids, aldehydes, ketones and amines to provide a reasonably good prediction of aliphatic toxicity against the *T. pyriformis* [14]. Netzeva and Schultz [15] developed QSAR for the toxicity of aldehydes to *T. pyriformis* using the maximum acceptor superdelocalizability in a molecule and the octanol water partition coefficient ($\log K_{o/w}$) as predictor variables. However, their statistical analysis was confined to cross-validation without true external validation. In the present paper, we have used the same data set of toxicity of aldehydes to *T. pyriformis* and developed predictive models using quantum topological molecular similarity (QTMS) descriptors along with $\log K_{o/w}$. In a previous study on this data set [16], we developed significant models using Mulliken charges along with $\log K_{o/w}$ values suggesting the importance of electronic descriptors along with hydrophobicity for the toxicity against *T. pyriformis*. QTMS descriptors are known to be powerful in modelling electronic properties and activities for which the electronic factor is important [17 25]. This knowledge prompted us to attempt to develop predictive models for the toxicity of aromatic aldehydes against *T. pyriformis*. Sufficient validation strategies (internal and external validation, randomization) have been applied to check the predictability of the developed models.

2. Materials and methods

The toxicity of aldehydes against *T. pyriformis* has been used as the model data set (see Table 1 below) for the present QSAR study. QTMS descriptors were used as the predictor

Table 1. Observed and calculated inhibition toxicity of aldehydic compounds to protozoa *T. pyriformis*.

			Toxicity to T. pyriformis *(log 1/C)*		
				Calculated models	
Sl. No.	*Compounds*	*log* $K^a_{o/w}$	*Observed*[a]	*Model 19 (PLS)*	*Model 55 (G/PLS)*
Training set					
1	4-Nitrobenzaldehyde	1.56	0.203	0.036	0.016
2	1-Naphthaldehyde	2.67	0.423	0.365	0.444
3	4-Biphenylcarboxaldehyde	3.38	1.119	1.048	1.017
4	4-Bromobenzaldehyde	2.48	0.587	0.554	0.542
5	4-Cyanobenzaldehyde	1.21	0.043	−0.145	−0.167
6	Benzaldehyde	1.48	−0.196	0.009	−0.039
7	*p*-Tolualdehyde	1.99	−0.057	0.317	0.249
8	4-Fluorobenzaldehyde	1.54	−0.127	0.052	0.0185
9	4-Chlorobenzaldehyde	2.13	0.4	0.367	0.350
10	4-Ethylbenzaldehyde	2.52	0.291	0.587	0.534
12	4-Anisaldehyde	1.65	−0.047	0.081	0.0145
13	4-Ethoxybenzaldehyde	2.31	0.073	0.441	0.385
14	4-Acetamidobenzaldehyde	1.25	−0.224	−0.088	−0.186
15	2-Tolualdehyde	2.26	0.011	0.333	0.427
17	2-Chlorobenzaldehyde	2.33	0.487	0.651	0.618
18	3-Chlorobenzaldehyde	2.26	0.406	0.460	0.522
20	3-Nitrobenzaldehyde	1.47	0.178	0.071	0.043
21	Phenyl-1,3-dialdehyde	1.36	0.183	−0.036	−0.114
22	2-Anisaldehyde	1.72	0.148	0.275	0.294
23	3-Anisaldehyde	1.71	0.232	0.216	0.293
24	3-Bromobenzaldehyde	2.48	0.506	0.569	0.610
25	3-Fluorobenzaldehyde	1.76	0.154	0.161	0.118
26	2,4-Dichlorobenzaldehyde	3.08	1.036	1.085	1.061
27	2,4-Dimethoxybenzaldehyde	1.79	−0.056	0.364	0.347
28	2,4,5-Trimethoxybenzaldehyde	1.19	−0.101	0.035	0.040
29	4-(Dimethylamino)benzaldehyde	1.81	0.231	0.247	0.137
30	4-Phenoxybenzaldehyde	3.96	1.257	1.416	1.387
31	2-Bromobenzaldehyde	2.48	0.477	0.716	0.680
32	2-Fluorobenzaldehyde	1.76	0.079	0.597	0.337
35	4-Isopropylbenzaldehyde	2.92	0.67	0.787	0.739
36	Pentafluorobenzaldehyde	2.39	0.815	0.892	0.984
37	2-Chloro-5-nitrobenzaldehyde	2.25	0.527	0.624	0.525
38	2-Chloro-6-fluorobenzaldehyde	2.51	0.155	0.654	0.591
39	3-Cyanobenzaldehyde	1.18	−0.02	−0.103	−0.102
41	6-Chloro-2-fluoro-3-methylbenzaldehyde	3.01	1.238	1.135	1.079
42	3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde	3.5	1.723	1.565	1.609
43	2,3,5-Trichlorobenzaldehyde	3.69	1.499	1.351	1.583
44	2-Fluorenecarboxaldehyde	3.43	1.499	1.088	1.105
45	2-Methyl-1-naphthaldehyde	3.17	1.231	0.983	1.247
46	4-Methyl-1-naphthaldehyde	3.17	1.123	1.156	1.132
48	5-Hydroxy-2-nitrobenzaldehyde	1.75	0.329	0.107	0.214
50	3-Hydroxybenzaldehyde	1.38	0.085	0.007	0.091
51	3-Hydroxy-4-methoxybenzaldehyde	0.97	−0.142	−0.284	−0.218

(*Continued*)

Table 1. Continued.

			Toxicity to T. pyriformis *(log 1/C)*		
				Calculated models	
Sl. No.	*Compounds*	*log* $K^a_{\mathrm{o/w}}$	*Observed*^{a}	*Model 19 (PLS)*	*Model 55 (G/PLS)*
52	3,4-Dimethoxy-5-hydroxycarboxaldehyde	0.69	−0.39	−0.420	−0.259
53	2,3-Dihydroxybenzaldehyde	1.03	0.111	0.164	0.177
54	2,5-Dihydroxybenzaldehyde	1.33	0.277	0.043	0.124
55	3,4-Dihydroxybenzaldehyde	1.03	0.107	−0.235	−0.160
57	2,3,4-Trihydroxybenzaldehyde	0.79	0.001	0.047	0.031
60	3-Ethoxy-2-hydroxycarboxaldehyde	2.17	0.85	0.833	0.831
62	3,5-Dibromosalicylaldehyde	3.42	1.648	1.362	1.413
63	4,6-Dimethoxy-2-hydroxybenzaldehyde	1.86	0.617	0.619	0.493
64	2-Hydroxy-3-nitrocarboxaldehyde	1.84	0.87	0.546	0.551
67	4-Hydroxybenzaldehyde	1.35	0.266	−0.058	−0.119
69	5-Bromovanillin	1.92	0.617	0.257	0.400
71	5-Bromosalicylaldehyde	2.8	1.107	0.971	0.926
72	5-Chlorosalicylaldehyde	2.65	1.009	0.881	0.845
74	3-Bromo-4-hydroxycarboxaldehyde	2.15	0.61	0.509	0.468
76	3,5-Dibromo-4-hydroxycarboxaldehyde	2.77	0.89	0.769	0.764
Test set					
11	Terephthaldicarboxaldehyde	1.36	−0.086	−0.089	−0.124
16	3-Tolualdehyde	1.99	0.081	0.270	0.268
19	2-Nitrobenzaldehyde	1.74	0.167	0.254	0.220
33	4-Butoxybenzaldehyde	3.37	0.716	1.015	0.980
34	4-(Pentyloxy)benzaldehyde	3.89	1.179	1.298	1.272
40	2-Chloro-3-hydroxy-4-methoxybenzaldehyde	1.72	0.204	0.271	0.501
47	Phenanthrene-9-carboxaldehyde	3.84	1.708	1.736	1.679
49	3-Hydroxy-4-nitrobenzaldehyde	1.47	0.273	0.150	0.232
56	3,4,5-Trihydroxybenzaldehyde	0.42	−0.196	−0.503	−0.401
58	2,4,6-Trihydroxybenzaldehyde	0.72	0.128	−0.015	−0.159
59	2,4-Dihydroxybenzaldehyde	1.33	0.515	0.316	0.135
61	3-Methoxysalicylaldehyde	1.37	0.377	0.401	0.381
65	2-Chloro-4-hydroxycarboxaldehyde	2.28	0.89	0.728	0.629
66	4-Hydroxy-3-nitrobenzaldehyde	1.48	0.614	0.199	0.086
68	2-Hydroxy-1-naphtaldehyde	2.99	1.32	1.422	1.283
70	4-Hydroxy-1-naphtaldehyde	2.62	1.05	0.865	0.865
73	2-Hydroxybenzaldehyde	1.81	0.424	0.467	0.364
75	3-Methoxy-4-hydroxybenzaldehyde	1.21	−0.03	−0.132	−0.028
77	3-Ethoxy-4-hydroxybenzaldehyde	1.58	0.015	0.065	0.178

Notes: [a]Taken from Netzeva and Schultz [15].

variables along with log $K_{\mathrm{o/w}}$, whose values were taken from the literature [15]. The QTMS descriptors were calculated at different levels of theory including AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31 + G(d,p), B3LYP/6-311 + G(2d,p) and MP2/6-311 + G(2d,p). QTMS descriptors focus on bond critical points (BCPs), which occur when the gradient of the electron density ρ vanishes ($\nabla\rho = \mathbf{0}$) at some point between two bonded nuclei. The electron density at a BCP, denoted by ρ_{b}, can be related to the bond order via an exponential relationship. At a BCP, the Hessian of ρ has two negative eigenvalues

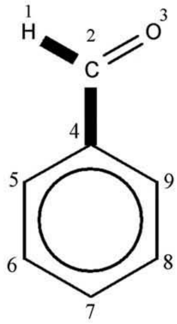


Figure 1. Arbitrary numbering of common atoms in the aldehyde skeleton (bonds identified by model 59 as important for the toxicity are shown in bold).

($\lambda_1 < \lambda_2 < 0$) and one positive one ($\lambda_3 > 0$). The eigenvalues express local curvature of ρ in a point: negative eigenvalues correspond to curvatures perpendicular to the bond, while the positive eigenvalue correspond to the curvature along the bond. The sum of the eigenvalues is the Laplacian, denoted by $\nabla^2\rho$, which is a measure of how much ρ is concentrated ($\nabla^2\rho < 0$) or depleted ($\nabla^2\rho > 0$) in a point. The descriptors ρ_b and λ_3 can be interpreted as measures of σ character whilst $\lambda_1 + \lambda_2$ measures the degree of π character [26]. Another measure of π character for homopolar bond is ellipticity, which is defined as $\varepsilon = \lambda_1/\lambda_2 - 1$ and is always positive because $\lambda_1 < \lambda_2 < 0$ at the BCP. In the QTMS bond descriptor vector, there are two more components, the kinetic energy density $K(\mathbf{r})$ and a more classical kinetic energy $G(\mathbf{r})$, as defined earlier [27]. Additionally, the equilibrium bond length (R_e) has also been used as one of the descriptors along with other QTMS descriptors [18].

First, an estimated geometry was obtained using the program *GaussView* [28], which was then passed on to the *ab initio* program *GAUSSIAN03* [29]. We used, in succession, AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31 + G(d,p), B3LYP/6-311 + G(2d,p) and MP2/6-311 + G(2d,p) levels of theory passing on the optimized geometry of each level as a starting geometry for the next. Since the AM1 level is unable to produce a sensible topology, only bond lengths were retrieved from it. Secondly, the wave function was read by a local version of the program *MORPHY98* [30], which located the BCPs using an automatic and robust algorithm [31]. The BCP descriptors of nine common bonds of the aldehydic compounds (six C C aromatic bonds and the C C bond, C=O bond and O H bond of the aldehyde functional group) were considered as variables (along with log $K_{o/w}$ values) for the statistical model development. Figure 1 shows the nine atoms shared by all 77 aldehydic derivatives, collectively referred to as the common skeleton.

Thirdly, the program *SIMCA* [32] was used for partial least squares (PLS) analysis of the data set. PLS is a generalization of regression, which can handle data with numerous independent variables, possibly strongly correlated and/or noisy [33]. It gives a reduced solution, which is statistically more robust than multiple linear regression (MLR). The linear PLS model finds 'new variables' (latent variables (LV)) that are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive LV is necessary in which no new LVs are added when they become non-significant. Cross-validation is a practical and reliable method of testing this significance [34]. PLS models were constructed for each type of descriptor, that is, ρ, ∇^2_ρ, ε, λ_1, λ_2, λ_3, K, G and R_e apart from the log $K_{o/w}$ values. Because there are nine bonds in the common skeleton there are nine descriptors in every type of descriptor. At the outset, models were built with all available descriptors, but, subsequently, descriptors with smaller variable importance for the projection (VIP) values were gradually deleted until a model with the best leave-one-seventh-out cross-validation correlation coefficient, $Q^2_{(1/7)}$, was obtained. Then, the PLS model for the combined set of descriptors was also built using a similar stepwise method of variable reduction based on VIP values (i.e., variables with smaller VIP values were successively removed from the model). All final PLS models (after variable deletion) were also run by the program *MINITAB* [35], which calculates the leave-one-out correlation coefficient, LOO-Q^2.

We have also constructed G/PLS models with 1000 iterations and scaled variables for combined set of descriptors along with log $K_{o/w}$ for all six levels of theory separately by *Cerius 2* (version 4.10) software [36]. The genetic partial least squares (G/PLS) method has been derived from two methods: the genetic function approximation (GFA) and partial least squares regression (PLS). G/PLS uses the GFA to select appropriate basis functions (i.e., descriptors) and uses the PLS technique to determine the relative contributions of the basis functions in the final model. The PLS technique generates LVs from the original descriptors and thus helps to analyse data with strongly collinear, noisy and numerous predictor variables, and simultaneously models several response (Y) variables. The G/PLS technique thus helps to construct QSAR models and avoids overfitting of variables. Figure 2 shows the flow chart of the main computational modules and programs used here in this work.

The main target of any QSAR modelling is that the built model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds [37 39]. So, QSAR models derived from a training set should be validated using new chemical entities for checking the predictive capacity of the constructed models.

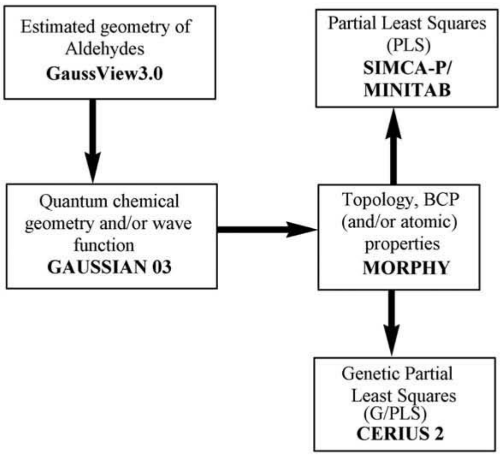


Figure 2. Chart representing the main computational modules involved in a QTMS analysis. The bold text represents the names of the programs used in this work.

The validation strategies check the reliability of the models for their possible application on a new data set, and so confidence in the prediction can be judged [40]. For the division of the data set into training and test sets, the compounds were ranked according to the toxicity values and every fourth compound was assigned to the test set. All the models were cross-validated by $Q^2_{(1/7)}$, as default in *SIMCA*. The model quality was characterized by the determination coefficient R^2 (note that capital R represents the multiple correlation coefficient of a model involving more than one descriptor), $Q^2_{(1/7)}$ and LOO-Q^2. The latter coefficient is calculated according to

$$Q^2 = 1 - \frac{\sum (Y_{\text{obs(training)}} - Y_{\text{pred(training)}})^2}{\sum (Y_{\text{obs(training)}} - \overline{Y}_{\text{training}})^2} \tag{1}$$

where $\overline{Y}_{\text{training}}$ represents the average activity value of the training set while $Y_{\text{obs(training)}}$ and $Y_{\text{pred(training)}}$ represent observed and predicted activity values of the training set compounds, respectively. Often, a high Q^2 value ($Q^2 > 0.5$) is considered as proof of the predictive ability of the model [41].

In order to evaluate the prediction potential of the models, the quantity R^2_{pred} was calculated from

$$R^2_{\text{pred}} = 1 - \frac{\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{\sum (Y_{\text{obs(test)}} - \overline{Y}_{\text{training}})^2} \tag{2}$$

where $Y_{\text{pred(test)}}$ and $Y_{\text{obs(test)}}$ indicate, respectively, the predicted and observed activity values of the test set compounds and $\overline{Y}_{\text{training}}$ indicates the mean activity value of the training set compounds. The value of R^2_{pred} for an acceptable model should be more than 0.5. It has been previously shown [41] that R^2_{pred} may not be sufficient to indicate external predictivity of a model. The value of R^2_{pred} is mainly controlled by $\sum (Y_{\text{obs(test)}} - \overline{Y}_{\text{training}})^2$, i.e., the sum of the squared differences between observed values of the test set compounds and the mean observed activity value of the training data set. Hence, R^2_{pred} depends on the training set mean. Thus, it may not truly reflect the predictive capability with regard to a new data set. Moreover, one cannot infer from the squared correlation coefficient (r^2) between observed and predicted values of the test set compounds that the predicted values are very near to the observed activity. Indeed, there may be considerable numerical difference between the observed and predicted values in spite of a good overall intercorrelation being maintained. So, for a better external predictive potential of the model, a modified r^2 ($r^2_{\text{m(test)}}$) was introduced [42],

$$r^2_{\text{m(test)}} = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \tag{3}$$

where r_0^2 represents the squared correlation coefficient between the observed and predicted values of the test set compounds when the intercept is set to 0. Note that r^2 is always larger than r_0^2. In the case of good external prediction, predicted values will be very close to the observed activity values. Hence, the r^2 value will be very near to the r_0^2 value. In the best possible case, r^2_{m} will be equal to r^2. The value of $r^2_{\text{m(test)}}$ should be greater than 0.5 for an acceptable model.

Initially, the concept leading to r^2_{m} was applied only to the test set prediction [42], but it can also be applied for the training set if one considers the correlation between observed and LOO-predicted values of the training set compounds [43,44]. More interestingly, this can be used for *the whole set* considering LOO-predicted values for the training set and predicted values of the test set compounds. There are two advantages associated with this consideration. First, unlike external validation parameters (R^2_{pred} etc.), the $r^2_{\mathrm{m(overall)}}$ statistic is not based only on a limited number of test set compounds. It includes a prediction for both test set and training set (using LOO predictions) compounds. Thus, this statistic is based on the prediction of a comparably larger number of compounds. In many cases, the test set size is considerably small and regression-based external validation parameters may be less reliable and highly dependent on individual test set observations. In such cases, the $r^2_{\mathrm{m(overall)}}$ statistic may be advantageous. Secondly, in many cases, comparable models are obtained where some models show comparatively better internal validation parameters and some other models show relatively superior external validation parameters. This may create a problem in selecting the final model. The $r^2_{\mathrm{m(overall)}}$ statistic may be used for selection of the best predictive models from among comparable models. For the present QSAR study, we have determined r^2_{m} values for both training (based on LOO-predicted values) and test sets, and also for the whole set. It may be noted that the $r^2_{\mathrm{m(overall)}}$ statistic is not used for the development of the models but for the selection of the best model from among comparable models.

3. Results and discussion

Table 1 gives the observed and calculated toxicity data of the 77 compounds, divided into the training set ($n = 58$) and the test set ($n = 19$), along with the values of $\log K_{\mathrm{o/w}}$. All PLS models developed with QTMS descriptors along with $\log K_{\mathrm{o/w}}$ are summarized in Table 2. A model was also generated (model 1) using only $\log K_{\mathrm{o/w}}$ as independent variable for setting a reference to which all models containing QTMS descriptors are compared. This model showed an R^2 value of 0.754, a $Q^2_{\mathrm{(LOO)}}$ value of 0.736 and a $Q^2_{(1/7)}$ value of 0.741. To consider the predictability of a model, we lay more emphasis on external predictability parameters. The predicted R^2 value of model 1 for the test set compounds is 0.642. However, the r^2 value differs significantly from the r_0^2 value leading to the considerably lower $r^2_{\mathrm{m(test)}}$ value of 0.514. In case of the training set, the value of $r^2_{\mathrm{m(LOO)}}$ is 0.695, which is moderately good. The $r^2_{\mathrm{m(overall)}}$ value of model 1 is 0.637, which includes a prediction for both test set and training set compounds. In order to obtain models with better external as well as overall predictability, we have used QTMS descriptors along with $\log K_{\mathrm{o/w}}$.

At the AM1 level, only bond distances were considered as BCP descriptors. At the start, all nine bond lengths along with $\log K_{\mathrm{o/w}}$ were used, followed by sequential deletion of less significant descriptors based on VIP values, eventually resulting in model 2. Clearly, model 2 is superior to model 1. Next we proceed to the higher levels of theory. Note that all models contain two LVs.

At the HF/3-21G(d) level we obtain ten models (models 3 12). Model 3 was built from the collection of all types of descriptors. As we mentioned earlier, we put more emphasis on external predictability parameters for selection of the best model. Based on both external and overall validation characteristics, the best model of this level was derived from ellipticity descriptors (model 6). Model 11 shows the internal validation parameters superior to model 6, but model 6 is superior to model 11 based on external and overall validation parameters.

At the next level, HF/6-31G(d), again we derived ten models (models 13 22), as is the case for all higher levels. The best model at this level came from the λ_3 descriptor (model 19). Model 19 is better than model 2 and model 6 not only in internal and external

Table 2. Statistical quality of PLS models obtained from variable selection based on VIP values (*SIMCA*)*.

Level of theory	*Model no.*	*Type of descriptors in addition to log $K_{o/w}$*	*No. of descriptors*	*LV*	R^2	$Q^2_{(LOO)}$	$Q^2_{(1/7)}$	$r^2_{m(LOO)}$	R^2_{pred}	$r^2_{m(test)}$	$r^2_{m(overall)}$
–	**1**	–	**1**	–	**0.754**	**0.736**	**0.741**	**0.695**	**0.642**	**0.514**	**0.637**
AM1	**2**	**Distance**	**6**	**2**	**0.824**	**0.788**	**0.783**	**0.786**	**0.766**	**0.637**	**0.740**
HF/3-21G(d)	3	All	6	2	0.777	0.753	0.720	0.745	0.762	0.634	0.706
	4	ρ	3	2	0.774	0.741	0.730	0.731	0.740	0.612	0.690
	5	$\nabla^2\rho$	4	2	0.790	0.756	0.737	0.749	0.669	0.547	0.681
	6	ε	**2**	**2**	**0.793**	**0.773**	**0.772**	**0.769**	**0.856**	**0.759**	**0.764**
	7	λ_1	4	2	0.797	0.766	0.746	0.757	0.731	0.597	0.701
	8	λ_2	3	2	0.778	0.745	0.740	0.736	0.777	0.654	0.707
	9	λ_3	3	2	0.800	0.769	0.756	0.759	0.735	0.613	0.707
	10	K	3	2	0.771	0.740	0.728	0.730	0.737	0.608	0.688
	11	G	3	2	0.817	0.794	0.787	0.784	0.842	0.71	0.757
	12	Distance	4	2	0.797	0.765	0.756	0.762	0.730	0.596	0.708
HF/6-31G(d)	13	All	6	2	0.794	0.767	0.747	0.758	0.790	0.663	0.725
	14	ρ	3	2	0.798	0.767	0.763	0.757	0.780	0.650	0.721
	15	$\nabla^2\rho$	4	2	0.817	0.786	0.773	0.777	0.805	0.676	0.742
	16	ε	3	2	0.814	0.787	0.789	0.781	0.827	0.699	0.754
	17	λ_1	3	2	0.801	0.767	0.764	0.756	0.770	0.642	0.717
	18	λ_2	5	2	0.809	0.780	0.786	0.778	0.785	0.645	0.729
	19	$\boldsymbol{\lambda_3}$	**3**	**2**	**0.829**	**0.804**	0.802	**0.796**	**0.886**	**0.780**	**0.788**
	20	K	4	2	0.811	0.782	0.768	0.772	0.775	0.644	0.728
	21	G	4	2	0.825	0.793	0.787	0.778	0.812	0.678	0.742
	22	Distance	4	2	0.790	0.761	0.744	0.753	0.792	0.663	0.724
B3LYP/6-31 + G(d,p)	23	All	4	2	0.788	0.767	0.770	0.760	0.714	0.586	0.704
	24	ρ	4	2	0.800	0.765	0.766	0.755	0.824	0.716	0.737
	25	$\nabla^2\rho$	3	2	0.808	0.779	0.775	0.768	0.804	0.680	0.738
	26	$\boldsymbol{\varepsilon}$	**5**	**2**	**0.826**	**0.794**	**0.789**	**0.792**	**0.813**	**0.679**	**0.754**
	27	λ_1	4	2	0.803	0.770	0.768	0.758	0.807	0.704	0.733
	28	λ_2	6	2	0.819	0.785	0.800	0.785	0.823	0.692	0.750
	29	λ_3	4	2	0.815	0.781	0.771	0.774	0.823	0.716	0.752
	30	K	4	2	0.797	0.764	0.765	0.752	0.820	0.713	0.734

(*Continued*)

Table 2. Continued.

Level of theory	*Model no.*	*Type of descriptors in addition to log $K_{o/w}$*	*No. of descriptors*	*LV*	R^2	$Q^2_{(LOO)}$	$Q^2_{(1/7)}$	$r^2_{m(LOO)}$	R^2_{pred}	$r^2_{m(test)}$	$r^2_{m(overall)}$
	31	G	3	2	0.777	0.753	0.741	0.746	0.653	0.513	0.667
	32	Distance	3	2	0.787	0.762	0.755	0.754	0.717	0.582	0.700
B3LYP/6-311 + G(2d,p)	33	All	6	2	0.794	0.770	0.751	0.761	0.790	0.660	0.726
	34	ρ	3	2	0.801	0.772	0.767	0.762	0.781	0.650	0.724
	35	$\nabla^2\rho$	3	2	0.804	0.775	0.770	0.765	0.793	0.663	0.731
	36	$\boldsymbol{\varepsilon}$	**3**	**2**	**0.816**	**0.791**	**0.792**	**0.782**	**0.820**	**0.702**	**0.758**
	37	λ_1	3	2	0.805	0.774	0.767	0.762	0.773	0.644	0.721
	38	λ_2	6	2	0.823	0.786	0.785	0.783	0.810	0.684	0.743
	39	λ_3	3	2	0.795	0.760	0.731	0.754	0.686	0.540	0.675
	40	K	4	2	0.797	0.765	0.755	0.755	0.753	0.621	0.710
	41	G	3	2	0.775	0.751	0.741	0.744	0.661	0.520	0.668
	42	Distance	3	2	0.787	0.761	0.755	0.754	0.721	0.587	0.700
MP2//6-311 + G(2d,p)	43	All	6	2	0.783	0.760	0.769	0.752	0.685	0.550	0.686
	44	ρ	3	2	0.808	0.777	0.774	0.765	0.621	0.444	0.648
	45	$\nabla^2\rho$	3	2	0.809	0.780	0.776	0.768	0.791	0.662	0.732
	46	ε	3	2	0.810	0.772	0.772	0.763	0.700	0.551	0.692
	47	λ_1	3	2	0.810	0.776	0.771	0.764	0.773	0.640	0.721
	48	$\boldsymbol{\lambda_2}$	**4**	**2**	**0.817**	**0.792**	**0.795**	**0.788**	**0.792**	**0.668**	**0.740**
	49	λ_3	3	2	0.782	0.741	0.720	0.724	0.667	0.522	0.656
	50	K	3	2	0.783	0.759	0.754	0.749	0.776	0.646	0.713
	51	G	3	2	0.785	0.757	0.749	0.747	0.669	0.528	0.674
	52	Distance	3	2	0.783	0.756	0.754	0.750	0.691	0.555	0.686

Note: *The best model of each level based on $r^2_{m(overall)}$ values is shown in bold face.

validation parameters but also on the basis of overall validation parameter ($r^2_{\mathrm{m(overall)}} = 0.788$).

At the B3LYP/6-31 + G(d,p) level, ten models were obtained (models 23 32). Based on the overall validation criteria, model 26 is the best one. This model is better than model 2 but inferior to models 6 and 19 based on the overall validation criteria. These values are marginally lower than the corresponding values of model 19 obtained at the HF/6-31G(d) level. Again, the predicted R^2 value of 0.813 and the $r^2_{\mathrm{m(test)}}$ value of 0.679 for model 26 are quite lower than the corresponding values of model 19. Another model (model 29) with four λ_3 descriptors shows R^2_{pred} and $r^2_{\mathrm{m(test)}}$ values higher than model 26. Again, the $r^2_{\mathrm{m(overall)}}$ value (0.752) of model 29 is marginally lower than that of model 26.

At the next level, B3LYP/6-311 + G(2d,p), models 33 42 were obtained. Based on internal validation, models 36 and 38 are pretty comparable. However, considering the external validation parameters model 36 is marginally superior to model 38. Based on the overall validation characteristics, model 36 is better than models 1, 2 and 26 and poorer than models 6 and 19.

At the last level, MP2/6-311 + G(2d,p), models 43 52 were obtained. Based on the internal, external and overall validation parameters, model 48 appeared to be the best model for this level. Model 48 shows very good internal validation ($Q^2_{\mathrm{(LOO)}} = 0.792$ and $Q^2_{(1/7)} = 0.795$). The external validation parameters of model 48 are better than those of model 1 and model 2, but these values are lower than the corresponding values of models 6, 19, 26 and 36.

According to the VIP values, $\log K_{\mathrm{o/w}}$ appeared to be the most significant descriptor for all models. Thus we can infer that the partition coefficient ($\log K_{\mathrm{o/w}}$) is a very important descriptor for the toxicity of aromatic aldehydes. Among all the levels, the PLS model 19 at the HF/6-31G(d) level with three λ_3 descriptors and two LVs outperformed all other models with regard to internal, external as well as overall validation parameters. In this work the overall validation parameter is taken as the selection criterion for the choice of the best model, so model 19 is the best model among the 52 PLS models derived from *SIMCA*. As QTMS descriptors increase predictability of the models when used along with $\log K_{\mathrm{o/w,}}$ the electronic factor is also found to be important for the toxicity along with hydrophobicity. It is interesting to note that model 19 shows better internal and external validation characteristics than those of the models developed earlier [16] using Mulliken charge parameters as electronic descriptors. Furthermore, all QTMS models listed in Table 2 have two LVs, one of which possibly signifies the hydrophobicity component and the other being electronic.

Table 3 summarizes the results obtained from the developed G/PLS models with 1000 iterations and scaled variables for the combined set of descriptors along with $\log K_{\mathrm{o/w}}$ for all six levels of theory. At the AM1 level, nine bond distances along with $\log K_{\mathrm{o/w}}$ were considered as descriptors. In the case of higher levels of theory (HF/3-21G(d), HF/6-31G(d), B3LYP/6-31 + G(d,p), B3LYP/6-311 + G(2d,p) and MP2/6-311 + G(2d,p)), five models (models 54 58) were developed from the pool of all QTMS descriptors along with $\log K_{\mathrm{o/w}}$.

In the case of G/PLS, all six models gave excellent internal predictive values. The various statistics adopt the following ranges: $R^2 = 0.858$ 0.881, $Q^2 = 0.827$ 0.856, $Q^2_{(1/7)} =$ 0.797 0.825 and $r^2_{\mathrm{m(LOO)}} = 0.817$ 0.849, which are satisfactory. In the case of external validation the predicted R^2 values (0.745 0.820) are encouraging for all the levels but the values for $r^2_{\mathrm{m(test)}}$ are somewhat lower than the best PLS models developed from *SIMCA*. The range of $r^2_{\mathrm{m(test)}}$ values of the G/PLS models is 0.631 0.696. However, the

Table 3. Statistical quality of G/PLS models obtained using combined set of descriptors (*Cerius 2*)*.

Level of theory	*Model no.*	*Type of descriptors in addition to log $K_{o/w}$*	*No. of descriptors*	*LV*	R^2	$Q^2_{(\mathrm{LOO})}$	$Q^2_{(1/7)}$	$r^2_{\mathrm{m(LOO)}}$	R^2_{pred}	$r^2_{\mathrm{m(test)}}$	$r^2_{\mathrm{m(overall)}}$
AM1	53	Distance	5	3	0.861	0.827	0.802	0.817	0.745	0.631	0.760
HF/3-21G(d)	54	All	5	3	0.862	0.828	0.798	0.819	0.807	0.687	0.778
HF/6-31G(d)	**55**	**All**	**5**	**3**	**0.870**	**0.844**	**0.825**	**0.835**	**0.807**	**0.696**	**0.794**
B3LYP/6-31 + G(d,p)	56	All	4	3	0.881	0.856	0.813	0.849	0.778	0.652	0.790
B3LYP/6-311 + G(2d,p)	57	All	7	4	0.870	0.834	0.797	0.825	0.820	0.686	0.782
MP2//6-311 + G(2d,p)	58	All	4	3	0.858	0.831	0.800	0.825	0.781	0.663	0.775

Note: *The best model based on $r^2_{\mathrm{m(overall)}}$ values is shown in bold face.

$r^2_{\mathrm{m(overall)}}$ value of model 55 developed by G/PLS is higher than that of the best PLS model (model 19) developed from *SIMCA*. The $r^2_{\mathrm{m(overall)}}$ value of model 55 is 0.794. So, based on overall validation parameters, model 55 at the HF/6-31G(d) level, is the best model among the models developed by both PLS and G/PLS. In our study, all PLS and G/PLS models show better predictability than model 1 (the reference model), which is derived with only $\log K_{\mathrm{o/w}}$. This suggests that QTMS descriptors have helped to increase internal, external as well as overall predictability. It is interesting that model 55 based on QTMS descriptors along with $\log K_{\mathrm{o/w}}$ outperforms (in both internal and external validation parameters) our previously reported models [16] based on Mulliken charges along with $\log K_{\mathrm{o/w}}$.

Now we discuss *process randomization* in detail. The robustness of model 55 was judged by Y-randomization of the model building process at the 99% confidence level using *Cerius 2* (version 4.10) software. The test was done by repeatedly (99 times) scrambling the toxicity values to generate QSAR models from the whole pool of descriptors and then comparing the resulting scores with the score of the original QSAR model generated from non-randomized activity values. The process randomization is different from *model randomization*. In the former, the descriptor selection process is repeated from the whole pool of descriptors, while in the latter only those descriptors present in the model are used. For an acceptable QSAR model, the average correlation coefficient (R_{r}) of a randomized model should be less than the correlation coefficient (R) of a non-randomized model. We find (see Table 5 below) that the value of R_{r} (0.498) is significantly less than the value of R (0.933). We propose the use of a new parameter R^2_{p} in the present paper, which penalizes the model R^2 for a small difference between the squared mean correlation coefficient (R^2_{r}) of randomized models and the squared correlation coefficient (R^2) of non-randomized models. The above-mentioned parameter can be calculated as

$$R^2_{\mathrm{p}} = R^2 \times \sqrt{R^2 - R^2_r}. \tag{4}$$

For an acceptable QSAR model, the value of R^2_{p} should be greater than 0.5. The value of R^2_{p} (0.686) for model 55 is well above the permissible limit indicating that the model is not obtained by chance. The results of process randomization of model 55 are shown in Table 4.

Next we discuss model randomization. The best PLS model (model 19) and the best G/PLS model (model 55) also underwent a randomization test with 100 permutations (default [45] is 20) using *SIMCA*. In this test, the toxicity data (Y) are randomly permuted keeping the descriptor matrix intact, followed by a PLS run. Each randomization and

Table 4. Results of process randomization test of the best model for toxicity of aldehydes to *T. pyriformis*.

Statistical method	*G/PLS*
Model No.	55
Confidence level	99%
R from non-random trial	0.933
R^2 from non-random trial	0.870
R_{r} from random trial	0.498
R^2_{r} from random trial	0.248
R^2_{p}	0.686
Mean value of R from random trial ± SD	0.498 ± 0.065

Table 5. List of R^2_{int} and Q^2_{int} values from model randomization test of selected models for toxicity of aldehydes to *T. pyriformis*.

Statistical method	*Level*	*Model No.*	R^2_{int}	Q^2_{int}
PLS	HF/6-31G(d)	19	0.076	−0.291
G/PLS	HF/6-31G(d)	55	−0.006	−0.254

subsequent PLS analysis generates a new set of R^2 and Q^2 values, which are plotted against the correlation coefficient between the original Y values and the permuted Y values. The intercepts for the R^2 and Q^2 lines in this plot are a measure of the overfit. A model is considered valid if $R^2_{\mathrm{int}} < 0.4$ and $Q^2_{\mathrm{int}} < 0.05$. Both the models fulfil the required criteria. This indicates that the models 19 and 55 are not obtained by chance. The results of model randomization are shown in Table 5.

The applicability domain (AD) of a QSAR is the physico-chemical, structural or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds [45]. The purpose of the AD is to state whether the model's assumptions are met. In general, this is the case for interpolation rather than extrapolation. To investigate the AD of a training set one can directly analyse properties of the multivariate descriptor space of the training compounds or, more indirectly, analyse distance (or similarity) metrics. This can be achieved by different means of feature selection and successive principle component analyses. When a compound is highly dissimilar to all compounds of the modelling set, reliable prediction of its activity is unlikely. The concept of AD [46] was used to avoid such an unjustified extrapolation of activity predictions.

The residuals of Y and X are of diagnostic value for the quality of the model [33]. Since there are many X-residuals one needs a summary for each observation (compound). This is accomplished by the residual standard deviation (SD) of the X-residuals of the corresponding row of the residual matrix E. Because this SD is proportional to the distance between the data point and the model plane in X-space, it is also often called DModX (distance to the model in X-space). Here, X is the $(N \times K)$ matrix of predictor variables, Y is the $(N \times M)$ matrix of response variables and E is the $(N \times K)$ matrix of X-residuals, N is number of objects (cases, observations), k is the index of X-variables $(k = 1, 2, \ldots, K)$ and m is the index of Y-variables $(m = 1, 2, \ldots, M)$. A DModX larger than around 2.5 times the overall SD of the X-residuals (corresponding to an F-value of 6.25) indicates that the observation is outside the AD of the model [33].

Figure 3 represents the residual SD of X-residuals (DModX) of test set compounds for models 55 and 19.

For model 55, the DModX values of all the test compounds are below the critical value of 2.425 at the 99% confidence level. So, none of the compounds are outside the AD and predictions for the 19 test compounds are acceptable. Similarly for model 19, DModX values of all 19 test compounds are below the critical point at 99% level (3.47). In this case also, all test set compounds are within the AD and predictions for the 19 test compounds are reliable. The calculated (training set) and predicted (test set) values of the toxicity according to models 19 and 55 are given in Table 1.

Finally, different BCP descriptors of each common bond calculated at the HF/6-31G(d) level were separately subjected to factor analysis [47,48] using the *MINITAB* software [35]. One of the differences between factor analysis and principal

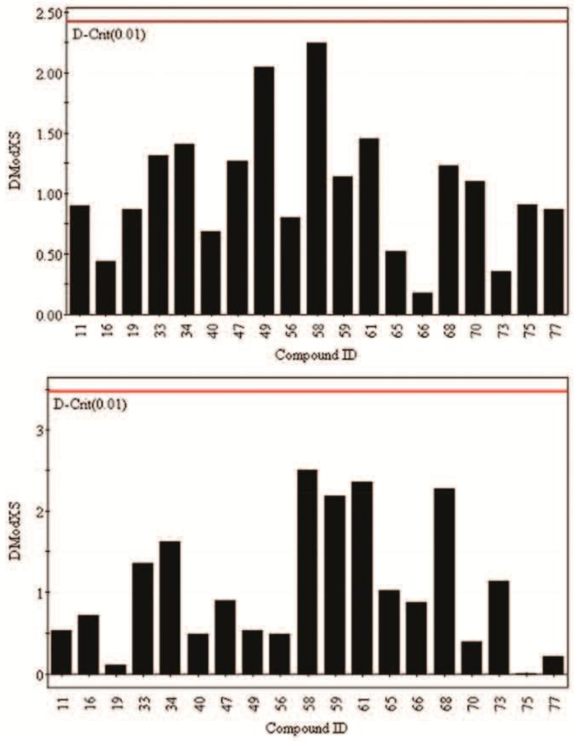


Figure 3. DmodX values of the 19 test set compounds at the 99% confidence level for model 55 (top) and model 19 (bottom). The lines at the top of each plot signify the critical DModX values (2.425 and 3.470) at the 99% confidence level.

component analysis is that in the former case the original variables are defined as linear combinations of the factors, while in the latter case the components are calculated as linear combinations of the original variables. The principal objectives of factor analysis are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining greater than 95% of the variance of the data matrix) and to extract the basic features behind the data with the ultimate goal of interpretation and/or prediction. The factors were extracted by principal component analysis and then rotated by VARIMAX rotation (a kind of rotation such that the axes are rotated to a position in which the sum of the variances of the loadings is the maximum possible) to obtain Thurston's simple structure. The simple structure is characterized by the property that as many variables as possible fall on the coordinate axes when presented in common factor space, so that the largest possible number of factor loadings becomes zero. This is done to obtain a numerically comprehensive picture of the relatedness of the variables. The principal component scores obtained for different bonds were then used as input variables and subjected to regression using *SIMCA* (with successive variable deletion based on VIP scores as detailed previously) for modelling the toxicity response. The details of the model (model 59) obtained from principal component scores are given in Table 6. From the VIP

Table 6. Statistical quality of the PLS model developed from principal component scores.

Model no.	*Level of theory*	*Statistical method*	*Type of descriptors in addition to log $K_{o/w}$*	*No. of descriptors*	*LV*	R^2	$Q^2_{(LOO)}$	$Q^2_{(1/7)}$	$r^2_{m(LOO)}$	R^2_{pred}	$r^2_{m(test)}$	$r^2_{m(overall)}$
59	HF/6-31G(d)	PLS	pc2_0102, pc1_0204, pc1_0904, pc2_0203, pc2_0607, pc1_0405, pc2_0506	8	2	0.814	0.767	0.778	0.761	0.744	0.711	0.758

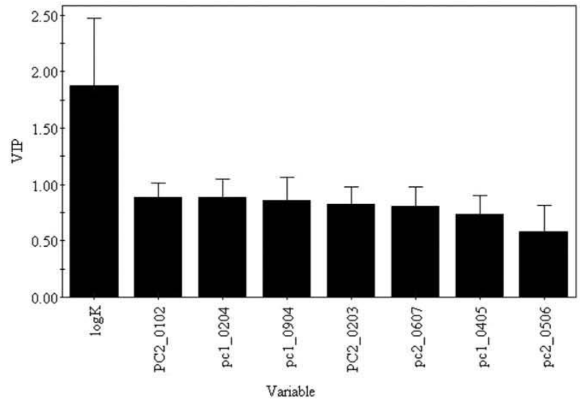


Figure 4. VIP plot for model 59.

plot (Figure 4) of this model, it can be found that the most important variable is log $K_{o/w}$ and the next two important variables are the second principal component of the BCP descriptors of the 0102 bond and the first principal component of the 0204 bond. From Figure 1, it can be found that 0102 and 0204 bonds (shown in bold lines) correspond to the aldehyde functionality and thus the model indicates the importance of the aldehyde moiety for the toxicity. Netzeva and Schultz [15] reported that the toxicity of aldehydes can be typically explained by Schiff-base formation and the interaction with amino nucleophiles [49]. Studies performed with α,β-unsaturated aldehydes such as cinnamaldehyde [50] generally support the hypothesis that toxicity is due to Michael-type addition. In the present work, QTMS descriptors were able to identify the importance of the aldehyde moiety for the toxicity of the compounds, which is consistent with the reports in the literature.

4. Conclusion

Predictive QSAR models for the toxicity of aldehydes against *T. pyriformis* can be constructed from QTMS descriptors along with log $K_{o/w}$. The results also suggest that lipophilicity plays a dominant role in describing the toxicity. However, electronic factors are also important for the predictability of the models. The models built in the present study can be used for the prediction of toxicity of untested aldehydes against *T. pyriformis*. The model based on factors scores of QTMS descriptors of different bonds was able to identify the important fragment (aldehyde functionality) responsible for the mediation of toxicity of aromatic aldehydes as corroborated by their established mechanism of toxicity.

Acknowledgements

Financial assistance from the Ministry of Human Resource Development, Government of India, New Delhi in the form of a scholarship to SK is gratefully acknowledged. We also thank the EPSRC (GB) and GlaxoSmithKline for financial support for APH.

References

[1] S.C. Basak, G.D. Grunwald, B.D. Gute, K. Balasubramanian, and D. Opitz, *Use of statistical and neural net approaches in predicting toxicity of chemicals*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 885–890.

[2] M. Mwense, X.Z. Wang, F.V. Buontempo, N. Horan, A. Young, and D. Osborn, *Prediction of noninteractive mixture toxicity of organic compounds based on a fuzzy set method*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 1763–1773.

[3] D.A. Smith, *Assessing toxicology quickly and efficiently*, Chem. Eng. 107 (2000), pp. 125–126.

[4] B.I. Escher and M.J.L. Hermens, *Modes of action in ecotoxicology: their role in body burdens, species sensitivity, QSARs, and mixture effects*, Environ. Sci. Technol. 36 (2002), pp. 4201–4217.

[5] I. Lessigiarska, A.P. Worth, T.I Netzeva, J.C. Dearden, and M.T.D. Cronin, *Quantitative structure–activity–activity and quantitative structure–activity investigations of human and rodent toxicity*, Chemosphere 65 (2006), pp. 1878–1887.

[6] P.J. O'Brien, A.G. Siraki, and N. Shangari, *Theme: industrial dust & chemical toxicology*, Crit. Rev. Tox. 35 (2005), pp. 609–662.

[7] R.C. Prince and D.E. Gunson, *Just plain vanilla?*, Trends Biochem. Sci. 19 (1994), p. 521.

[8] I. Kahn, S. Sild, and U.J. Maran, *Modeling the toxicity of chemicals to tetrahymena pyriformis using heuristic multi-linear regression and heuristic back-propagation in neural networks*, Chem. Inf. Model. 47 (2007), pp. 2271–2279.

[9] S.D. Dimitrov, O.G. Mekenyan, G.D. Sinks, and T.W. Schultz, *Global modeling of narcotic chemicals: Ciliate and fish toxicity*, J. Mol. Struct. (THEOCHEM) 622 (2003), pp. 63–70.

[10] J.R. Seward, E.L. Hamblen, and T.W. Schultz, *Regression comparison of* Tetrahymena pyriformis *and* Poecilia reticulata *toxicity*, Chemosphere 47 (2002), pp. 93–101.

[11] I. Kahn, U. Maran, E. Benfenati, T.I. Netzeva, T.W. Schultz, and M.T.D. Cronin, *Comparative quantitative structure–activity–activity relationships for toxicity to* Tetrahymena pyriformis *and* Pimephales promelas, ATLA-Altern. Laborat. Anim. 35 (2007), pp. 15–24.

[12] M.T.D. Cronin and T.W. Schultz, *Structure–toxicity relationships for phenols to* Tetrahymena pyriformis, Chemosphere 32 (1996), pp. 1453–1468.

[13] M.T.D. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V. Valkova, and T.W. Schultz, *Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to* Tetrahymena pyriformis, Chemosphere 49 (2002), pp. 1201–1221.

[14] D.R. Roy, R. Parthasarathi, B. Maiti, V. Subramanianb, and P.K. Chattaraja, *Electrophilicity as a possible descriptor for toxicity prediction*, Bioorg. Med. Chem. 13 (2005), pp. 3405–3412.

[15] T.I. Netzeva and T.W. Schultz, *QSARs for the aquatic toxicity of aromatic aldehydes from* Tetrahymena *data*, Chemosphere 61 (2005), pp. 1632–1643.

[16] S. Kar and K. Roy, *Exploring QSAR of toxicity of diverse aromatic aldehydes to* Tetrahymena pyriformis *using lipophilicity and quantum chemical parameters*, Proceeding of International Conference on Open Source for Computer Aided Drug Discovery, 22–26th March 2009, Institute of Microbial Technology, Chandigarh (India).

[17] P.L.A. Popelier, *Quantum molecular similarity. Part 1: BCP space*, J. Phys. Chem. A, 103 (1999), pp. 2883–2890.

[18] S.E. O'Brien and P.L.A. Popelier, *Quantum molecular similarity. Part 3: QTMS descriptors*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 764–775.

[19] P.L.A. Popelier, U.A. Chaudry, and P.J. Smith, *Quantum topological molecular similarity. Part 5: Further development with an application to the toxicity of polychlorinated dibenzo-p-dioxins (PCDDs)*, J. Chem. Soc., Perkin Trans. 2 (2002), pp. 1231–1237.

[20] S.E. O'Brien and P.L.A. Popelier, *Quantum molecular similarity. Part 4: anti-tumour activity of phenylbutenones*, J. Chem. Soc., Perkin Trans. 2 (2002), pp. 478–483.

[21] U.A. Chaudry and P.L.A. Popelier, *Ester hydrolysis rate constant prediction from quantum topological molecular similarity descriptors*, J. Phys. Chem. A 107 (2003), pp. 4578–4582.

[22] P.L.A. Popelier and P.J. Smith, *QSAR models based on quantum topological molecular similarity*, Eur. J. Med. Chem. 41 (2006), pp. 862–873.

[23] K. Roy and P.L.A. Popelier, *Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors*, Bioorg. Med. Chem. Lett. 18 (2008), pp. 2604–2609.

[24] A. Mohajeri, B. Hemmateenejad, A. Mehdipour, and R. Miri, *Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS*, J. Mol. Graph. Mod. 26 (2008), pp. 1057–1065.

[25] A. Mohajeri and M.H. Dinpajooh, *Structure–toxicity relationship for aliphatic compounds using quantum topological descriptors*, J. Mol. Struct. (THEOCHEM) 855 (2008), pp. 1–5.

[26] S.T. Howard and O. Lamarche, *Description of covalent bond orders using the charge density topology*, J. Phys. Org. Chem. 16 (2003), pp. 133–141.

[27] R.F.W. Bader and H.J.T. Preston, *The kinetic energy of molecular charge distributions and molecular stability*, Int. J. Quantum Chem. 3 (1969), pp. 327–347.

[28] *GaussView3.0*, Semichem Inc., Gaussian Inc., Pittsburgh, PA, USA, 2003. Available at http://www.gaussian.com/g-prod/gv5.htm (accessed 29 January 2010).

[29] *GAUSSIAN03*, M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A.J. Montgomery, J.T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Ba boul, S. Clifford, J. Cioslowski, B.B.Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, and J.A. Pople, In Gaussian, Inc., Pittsburgh, PA, 2003.

[30] *MORPHY98*, a program written by P.L.A. Popelier, with a contribution from R.G.A. Bone, UMIST, Manchester, England, EU, 1998.

[31] P.L.A. Popelier, *A robust algorithm to locate automatically all types of critical points in the charge density and its Laplacian*, Chem. Phys. Lett. 228 (1994), pp. 160–164.

[32] *UMETRICS SIMCA-P 10.0*, Umea, Sweden, 2002; software available at infoumetrics.com, www.umetrics.com.

[33] S. Wold, M. Sjostrom, and L. Eriksson, *PLS-regression: a basic tool of chemometrics*, Chemo. Intell. Lab. Sys. 58 (2001), pp. 109–130.

[34] S. Wold and L. Eriksson, *Statistical validation of QSAR results*, in *Chemometric Methods in Molecular Design*, H.E. van de Waterbeemd, ed., VCH, Weinheim, Germany, 1995, pp. 195–218.

[35] *MINITAB*, Minitab Inc., USA; software available at http://www.minitab.com

[36] *Cerius2 Version 4.10*, Accelrys Inc., San Diego, CA.

[37] K. Roy, *On some aspects of validation of predictive quantitative structure–activity relationship models*, Expert Opin. Drug Discov. 2 (2007), pp. 1567–1577.

[38] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.

[39] P.P. Roy, J.T. Leonard, and K. Roy, *Exploring the impact of the size of training sets for the development of predictive QSAR models*, Chemom. Intell. Lab. Syst. 90 (2008), pp. 31–42.

[40] K. Roy and A.S. Mandal, *Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones*, J. Enzyme Inhib. Med. Chem. 23(6) (2008), pp. 980–995.

[41] H. Kubinyi, F.A. Hamprecht, and T. Mietzner, *Three-dimensional quantitative similarity–activity relationships (3D QSiAR) from SEAL similarity matrices*, J. Med. Chem. 41 (1998), pp. 2553–2564.

[42] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27(3) (2008), pp. 302–313.

[43] P.P. Roy, S. Paul, I. Mitra, and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.

[44] I. Mitra, P.P. Roy, S. Kar, P.K. Ojha and K. Roy, *On further application of r^2_m as a metric for validation of QSAR models,* J. Chemometrics, 24 (2010), pp. 22–33.

[45] http://en.wikipedia.org/wiki/Applicability_Domain (accessed 9 November 2009).

[46] L. Zhang, H. Zhu, T. Oprea, A. Golbraikh, and A. Tropsha, *QSAR modeling of the blood–brain barrier permeability for diverse organic compounds*, Pharm. Res. 25 (2008), pp. 1902–1914.

[47] R. Franke, *Theoretical Drug Design Methods*, Elsevier, Amsterdam, 1984, pp. 184–195.

[48] R. Franke and A. Gruska, *Principal component and factor analysis*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 113–163.

[49] S. Karabunarliev, O.G. Mekenyan, W. Karcher, C.I. Russom, and S.P. Bradbury, *Quantum-chemical descriptors for estimating the acute toxicity of electrophiles to fathead minnow (*Pimephales promelas*): an analysis based on molecular mechanism*, Quant. Struct.–Act. Relat. 4 (1996), pp. 302–310.

[50] H. Niknahad, A. Shuhendler, G. Galati, A.G. Siraki, E. Easson, R. Poon, and P.J. O'Brien, *Modulating carbonyl cytotoxicity in intact rat hepatocytes by inhibiting carbonyl metabolizing enzymes. II. Aromatic aldehydes*, Chem.–Biol. Interact. 143–144 (2003), pp. 119–128.

Appendix B

The carboxylic acids and associated experimental pK_a values.

ID	CAS no.	Chemical Name	Exp. pK_a
meta- and para-substituted benzoic acids			
1	000051-44-5	3,4-dichlorobenzoic acid	3.64
2	000051-36-5	3,5-dichlorobenzoic acid	3.54
3	000057-66-9	4-[(dipropylamino)sulfonyl]benzoic acid/probenecid	3.40
4	000062-23-7	4-nitrobenzoic acid	3.44
5	000065-85-0	benzoic acid	4.19
6	000074-11-3	4-chlorobenzoic acid	3.98
7	000093-09-4	2-naphthalenecarboxylic acid	4.17
8	000093-07-2	3,4-dimethoxybenzoic acid	4.36
9	000098-73-7	4-(1,1-dimethylethyl)benzoic acid	4.40
10	000099-96-7	4-hydroxybenzoic acid	4.54
11	000099-94-5	4-methylbenzoic acid	4.37
12	000099-50-3	3,4-dihydroxybenzoic acid	4.48
13	000099-34-3	3,5-dinitrobenzoic acid	2.82
14	000099-10-5	3,5-dihydroxybenzoic acid	4.04
15	000099-05-8	3-aminobenzoic acid	4.74
16	000099-04-7	3-methylbenzoic acid	4.27
17	000100-09-4	4-(methyloxy)benzoic acid	4.47
18	000121-92-6	3-nitrobenzoic acid	3.46
19	000121-34-6	4-hydroxy-3-(methyloxy)benzoic acid	4.51
20	000149-91-7	3,4,5-trihydroxybenzoic acid	4.21
21	000150-13-0	4-aminobenzoic acid	4.85
22	000455-38-9	3-fluorobenzoic acid	3.86
23	000456-22-4	4-fluorobenzoic acid	4.14
24	000528-45-0	3,4-dinitrobenzoic acid	2.82
25	000530-57-4	4-hydroxy-3,5-bis(methyloxy)benzoic acid	4.34
26	000535-80-8	3-chlorobenzoic acid	2.81

27	000536-66-3	4-(1-methylethyl)benzoic acid	4.35
28	000585-76-2	3-bromobenzoic acid	3.81
29	000586-89-0	4-acetylbenzoic acid	3.70
30	000586-76-5	4-bromobenzoic acid	4.00
31	000586-38-9	3-(methyloxy)benzoic acid	4.09
32	000619-86-3	4-(ethyloxy)benzoic acid	4.45
33	000619-66-9	4-formylbenzoic acid	3.77
34	000619-65-8	4-cyanobenzoic acid	3.55
35	000619-64-7	4-ethylbenzoic acid	4.35
36	000619-21-6	3-formylbenzoic acid	3.84
37	000619-05-6	3,4-diaminobenzoic acid	3.49
38	001132-21-4	3,5-dimethoxybenzoic acid	3.97
39	001877-72-1	3-cyanobenzoic acid	3.60
40	002215-77-2	4-(phenyloxy)benzoic acid	4.52
41	003739-38-6	3-(phenyloxy)benzoic acid	3.92
42	004052-30-6	4-(methylsulfonyl)benzoic acid	3.64
43	005438-19-7	4-(propyloxy)benzoic acid	4.46
44	007496-53-9	4-(glycylamino)benzoic acid	4.20

ortho-substituted benzoic acids

45	000050-85-1	2-hydroxy-4-methylbenzoic acid	3.40
46	000050-84-0	2,4-dichlorobenzoic acid	2.68
47	000050-79-3	2,5-dichlorobenzoic acid	2.47
48	000050-78-2	2-(acetyloxy)benzoic acid	3.49
49	000050-31-7	2,3,6-trichlorobenzoic acid	1.50
50	000050-30-6	2,6-dichlorobenzoic acid	1.59
51	000059-07-4	4-amino-2-(ethyloxy)benzoic acid	5.09
52	000061-68-7	2-[(2,3-dimethylphenyl)amino]benzoic acid	4.20
53	000065-49-6	4-amino-2-hydroxybenzoic acid	3.66
54	000069-72-7	2-hydroxybenzoic acid	2.97
55	000083-40-9	2-hydroxy-3-methylbenzoic acid	2.95
56	000088-65-3	2-bromobenzoic acid	2.88
57	000089-86-1	2,4-dihydroxybenzoic acid	3.11

58	000089-56-5	2-hydroxy-5-methylbenzoic acid	3.15
59	000089-55-4	5-bromo-2-hydroxybenzoic acid	2.66
60	000089-52-1	2-(acetylamino)benzoic acid	3.40
61	000091-52-1	2,4-dimethyloxybenzoic acid	4.36
62	000091-40-7	2-(phenylamino)benzoic acid	3.99
63	000092-70-6	3-hydroxy-2-naphthalenecarboxylic acid	2.79
64	000096-97-9	2-hydroxy-5-nitrobenzoic acid	2.12
65	000099-60-5	2-chloro-4-nitrobenzoic acid	2.14
66	000118-92-3	2-aminobenzoic acid	4.95
67	000118-91-2	2-chlorobenzoic acid	2.89
68	000119-90-1	2-methylbenzoic acid	3.98
69	000129-66-8	2,4,6-trinitrobenzoic acid	0.65
70	000133-90-4	3-amino-2,5-dichlorobenzoic acid	3.40
71	000303-38-8	2,3-dihydroxybenzoic acid	2.91
72	000303-07-1	2,6-dihydroxybenzoic acid	1.05
73	000321-14-2	5-chloro-2-hydroxybenzoic acid	2.65
74	000445-29-4	2-fluorobenzoic acid	3.27
75	000490-79-9	2,5-dihydroxybenzoic acid	2.95
76	000552-16-9	2-nitrobenzoic acid	2.17
77	000577-56-0	2-acetylbenzoic acid	4.13
78	000579-75-9	2-(methyloxy)benzoic acid	3.90
79	000609-99-4	2-hydroxy-3,5-dinitrobenzoic acid	0.70
80	000610-30-0	2,4-dinitrobenzoic acid	1.42
81	000632-46-2	2,6-dimethylbenzoic acid	3.35
82	000652-32-4	2,3,5,6-tetrafluoro-4-methylbenzoic acid	2.00
83	000947-84-2	2-biphenylcarboxylic acid	3.46
84	001466-76-8	2,6-bis(methyloxy)benzoic acid	3.44
85	001521-38-6	2,3-dimethyloxybenzoic acid	3.98
86	001918-00-9	3,6-dichloro-2-(methyloxy)benzoic acid	1.97
87	002243-42-7	2-(phenyloxy)benzoic acid	3.53
88	002438-04-2	2-(1-methylethyl)benzoic acid	3.63
89	002516-96-3	2-chloro-5-nitrobenzoic acid	2.17

90	003970-35-2	2-chloro-3-nitrobenzoic acid	2.02
91	004727-29-1	2-[(phenylamino)carbonyl]benzoic acid	2.50
92	005344-49-0	2-chloro-6-nitrobenzoic acid	1.34
93	021327-86-6	2-chloro-6-methylbenzoic acid	2.75
94	025784-02-5	2-[(2-amino-2-oxoethyl)amino]benzoic acid	4.20

aliphatic carboxylic acids

95	000050-21-5	2-hydroxypropanoic acid	3.86
96	000053-86-1	[1-[(4-chlorophenyl)carbonyl]-2-methyl-5-(methyloxy)-1H-indol-3-yl]acetic acid/indomethacin	4.50
97	000061-78-9	N-[(4-aminophenyl)carbonyl]glycine	3.80
98	000061-33-6	3,3-dimethyl-7-oxo-6-[(phenylacetyl)amino]-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid/benzylpenicillin	2.74
99	000061-32-5	6-({[2,6-bis(methyloxy)phenyl]carbonyl}amino)-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acidmethicillin	2.77
100	000064-19-7	acetic acid	4.76
101	000068-11-1	mercaptoacetic acid	3.55
102	000075-99-0	2,2-dichloropropanoic acid	1.79
103	000075-98-9	2,2-dimethylpropanoic acid	5.03
104	000076-93-7	hydroxy(diphenyl)acetic acid	3.05
105	000076-05-1	trifluoroacetic acid	0.52
106	000076-03-9	trichloroacetic acid	0.51
107	000077-06-5	gibberellic acid	4.00
108	000079-43-6	dichloroacetic acid	1.26
109	000079-31-2	2-methylpropanoic acid	4.84
110	000079-14-1	hydroxyacetic acid	3.83
111	000079-11-8	chloroacetic acid	2.87
112	000079-09-4	propanoic acid	4.88
113	000079-08-3	bromoacetic acid	2.89
114	000081-25-4	3,7,12-trihydroxycholan-24-oic acid	4.98
115	000085-34-7	(2,3,6-trichlorophenyl)acetic acid	3.70
116	000086-87-3	1-naphthalenylacetic acid	4.23
117	000087-51-4	1H-indol-3-ylacetic acid	4.75
118	000087-08-1	3,3-dimethyl-7-oxo-6-{[(phenyloxy)acetyl]amino}-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid/phenoxymethylpenicillin	2.79
119	000088-09-5	2-ethylbutanoic acid	4.71
120	000090-64-2	hydroxy(phenyl)acetic acid	3.41

121	000093-76-5	[(2,4,5-trichlorophenyl)oxy]acetic acid	2.83
122	000093-72-1	2-[(2,4,5-trichlorophenyl)oxy]propanoic acid	2.84
123	000094-82-6	4-[(2,4-dichlorophenyl)oxy]butanoic acid	4.95
124	000094-81-5	4-[(4-chloro-2-methylphenyl)oxy]butanoic acid	6.20
125	000094-75-7	[(2,4-dichlorophenyl)oxy]acetic acid	2.73
126	000094-74-6	[(4-chloro-2-methylphenyl)oxy]acetic acid	3.13
127	000097-61-0	2-methylpentanoic acid	4.79
128	000098-89-5	cyclohexanecarboxylic acid	4.90
129	000099-66-1	2-propylpentanoic acid	4.60
130	000102-32-9	(3,4-dihydroxyphenyl)acetic acid	4.25
131	000103-82-2	phenylacetic acid	4.31
132	000104-03-0	(4-nitrophenyl)acetic acid	3.85
133	000104-01-8	[4-(methyloxy)phenyl]acetic acid	4.36
134	000107-94-8	3-chloropropanoic acid	3.99
135	000107-92-6	butanoic acid	4.82
136	000116-53-0	2-methylbutanoic acid	4.81
137	000117-34-0	diphenylacetic acid	3.94
138	000120-36-5	2-[(2,4-dichlorophenyl)oxy]propanoic acid	3.10
139	000122-88-3	[(4-chlorophenyl)oxy]acetic acid	3.10
140	000122-59-8	(phenyloxy)acetic acid	3.17
141	000123-76-2	4-oxopentanoic acid	4.64
142	000141-82-2	propanedioic acid	2.85
143	000144-49-0	fluoroacetic acid	2.59
144	000300-85-6	3-hydroxybutanoic acid	4.41
145	000305-03-3	4-{4-[bis(2-chloroethyl)amino]phenyl}butanoic acid/chlorambucil	5.75
146	000306-08-1	[4-hydroxy-3-(methyloxy)phenyl]acetic acid	4.41
147	000327-97-9	3-{[(2E)-3-(3,4-dihydroxyphenyl)-2-propenoyl]oxy}-1,4,5-trihydroxycyclohexanecarboxylic acid/chlorogenic acid	2.66
148	000331-25-9	(3-fluorophenyl)acetic acid	4.13
149	000348-10-7	[(2-fluorophenyl)oxy]acetic acid	3.08
150	000372-09-8	cyanoacetic acid	2.45
151	000404-98-8	[(3-fluorophenyl)oxy]acetic acid	3.13
152	000405-79-8	[(4-fluorophenyl)oxy]acetic acid	3.13

153	000405-50-5	(4-fluorophenyl)acetic acid	4.24
154	000462-60-2	N-(aminocarbonyl)glycine	3.89
155	000467-69-6	9-hydroxy-9H-fluorene-9-carboxylic acid/flurenol	1.09
156	000473-81-4	2,3-dihydroxypropanoic acid	3.55
157	000501-52-0	3-phenylpropanoic acid	4.66
158	000503-74-2	3-methylbutanoic acid	4.77
159	000503-66-2	3-hydroxypropanoic acid	4.51
160	000515-30-0	2-hydroxy-2-phenylpropanoic acid	3.53
161	000516-05-2	methylpropanedioic acid	3.12
162	000539-35-5	6-(4-oxo-1,3-thiazolidin-2-yl)hexanoic acid/mycobacidin	5.10
163	000581-96-4	2-naphthalenylacetic acid	4.25
164	000588-32-9	[(3-chlorophenyl)oxy]acetic acid	3.07
165	000588-22-7	[(3,4-dichlorophenyl)oxy]acetic acid	2.92
166	000594-61-6	2-hydroxy-2-methylpropanoic acid	3.61
167	000595-46-0	dimethylpropanedioic acid	3.15
168	000595-37-9	2,2-dimethylbutanoic acid	5.03
169	000598-78-7	2-chloropropanoic acid	2.80
170	000598-72-1	2-bromopropanoic acid	2.97
171	000601-75-2	ethylpropanedioic acid	2.96
172	000614-61-9	[(2-chlorophenyl)oxy]acetic acid	3.05
173	000616-62-6	propylpropanedioic acid	2.99
174	000617-31-2	2-hydroxypentanoic acid	2.89
175	000622-47-9	(4-methylphenyl)acetic acid	4.37
176	000646-07-1	4-methylpentanoic acid	4.84
177	000689-13-4	N-formyl-N-hydroxyglycine	3.50
178	000940-64-7	[(4-methylphenyl)oxy]acetic acid	3.21
179	001643-15-8	[(3-methylphenyl)oxy]acetic acid	3.20
180	001759-53-1	cyclopropanecarboxylic acid	4.83
181	001798-99-8	[(3-bromophenyl)oxy]acetic acid	3.09
182	001798-11-4	[(4-nitrophenyl)oxy]acetic acid	2.89
183	001821-12-1	4-phenylbutanoic acid	4.76
184	001877-75-4	{[4-(methyloxy)phenyl]oxy}acetic acid	2.31

185	001877-73-2	(3-nitrophenyl)acetic acid	3.97
186	001878-91-7	[(4-bromophenyl)oxy]acetic acid	3.13
187	001878-88-2	[(3-nitrophenyl)oxy]acetic acid	2.95
188	001878-87-1	[(2-nitrophenyl)oxy]acetic acid	2.90
189	001878-85-9	{[2-(methyloxy)phenyl]oxy}acetic acid	3.23
190	001878-82-6	[(4-cyanophenyl)oxy]acetic acid	2.93
191	001878-68-8	(4-bromophenyl)acetic acid	4.19
192	001878-66-6	(4-chlorophenyl)acetic acid	4.19
193	001878-65-5	(3-chlorophenyl)acetic acid	4.14
194	001878-49-5	[(2-methylphenyl)oxy]acetic acid	3.23
195	001879-58-9	[(3-cyanophenyl)oxy]acetic acid	3.03
196	001879-56-7	[(2-bromophenyl)oxy]acetic acid	3.13
197	002088-24-6	{[3-(methyloxy)phenyl]oxy}acetic acid	3.14
198	002270-20-4	5-phenylpentanoic acid	4.88
199	002976-75-2	(1-naphthalenyloxy)acetic acid	3.20
200	003813-05-6	(4-chloro-2-oxo-1,3-benzothiazol-3(2H)-yl)acetic acid	3.04
201	005292-21-7	cyclohexylacetic acid	4.80
202	006324-11-4	[(2-hydroxyphenyl)oxy]acetic acid	3.02
203	010502-44-0	hydroxy[4-(methyloxy)phenyl]acetic acid	3.42
204	014387-10-1	(4-ethylphenyl)acetic acid	4.37
205	015307-86-5	{2-[(2,6-dichlorophenyl)amino]phenyl}acetic acid/ diclofenac	4.15
206	015687-27-1	2-[4-(2-methylpropyl)phenyl]propanoic acid/ibuprofen	4.45
207	016484-77-8	2-[(4-chloro-2-methylphenyl)oxy]propanoic acid	3.68
208	016563-41-0	3-(1-naphthalenyloxy)propanoic acid	4.00
209	018046-21-4	[4-(4-chlorophenyl)-2-phenyl-1,3-thiazol-5-yl]acetic acid/fentiazac	3.60
210	020225-24-5	2-ethylpentanoic acid	4.71
211	022071-15-4	2-[3-(phenylcarbonyl)phenyl]propanoic acid/ketoprofen	4.45
212	022131-79-9	[3-chloro-4-(2-propen-1-yloxy)phenyl]acetic acid/alcofenac	4.29
213	022204-53-1	2-[6-(methyloxy)-2-naphthalenyl]propanoic acid/naprosyn	4.15
214	029679-58-1	2-[3-(phenyloxy)phenyl]propanoic acid/fenoprofen	4.50
215	032857-63-9	[4-(1,1-dimethylethyl)phenyl]acetic acid	4.42
216	036330-85-5	4-(4-biphenylyl)-4-oxobutanoic acid/fenbufen	4.51

217	038194-50-2	((1E)-5-fluoro-2-methyl-1-{[4-(methylsulfinyl)phenyl]methylidene}-1H-inden-3-yl)acetic acid/sulindac	4.70
218	040828-46-4	2-[4-(2-thienylcarbonyl)phenyl]propanoic acid/suprofen	3.91
219	040843-25-2	2-({4-[(2,4-dichlorophenyl)oxy]phenyl}oxy)propanoic acid	3.43
220	053808-88-1	[3-(4-chlorophenyl)-1-phenyl-1H-pyrazol-4-yl]acetic acid/ionazolac	4.30
221	055335-06-3	[(3,5,6-trichloro-2-pyridinyl)oxy]acetic acid	2.68
222	055863-26-8	(11-oxo-6,11-dihydrodibenzo[b,e]thiepin-2-yl)acetic acid/tiopinac	3.71
223	058667-63-3	N-(3-chloro-4-fluorophenyl)-N-(phenylcarbonyl)alanine	3.72
224	069335-91-7	2-[(4-{[5-(trifluoromethyl)-2-pyridinyl]oxy}phenyl)oxy]propanoic acid	3.12
225	069806-34-4	2-[(4-{[3-chloro-5-(trifluoromethyl)-2-pyridinyl]oxy}phenyl)oxy]propanoic acid	2.90
226	074103-06-3	5-(phenylcarbonyl)-2,3-dihydro-1H-pyrrolo[1,2-a]pyrrole-1-carboxylic acid/ketorolac	3.49
227	089894-13-3	[(4-chloro-3-nitrophenyl)oxy]acetic acid	2.96
228	104273-73-6	4-(cyclopropylcarbonyl)-3,5-dioxocyclohexanecarboxylic acid	5.32

Appendix C

Tehan					
ID	ID	SMILES	Chemical Name	Exp pK_a	Ref.
Meta/Para-Substituted Phenols					
1	1	NCCc1ccc(O)cc1	4-(2-aminoethyl)phenol	9.77	1
2	2	Cc1c(Cl)ccc(O)c1	3-methyl-4-chlorophenol	9.20	1
3	3	CC(Cc1ccc(O)cc1)(C)C	4-tert-amylphenol	10.43	1
4	4	c1(cc(O)ccc1[N+](=O)[O-])C(F)(F)F	3-trifluoromethyl-4-nitrophenol	6.07	1
5	5	c1(Cl)c(C)cc(cc1C)O	4-chloro-3,5-dimethylphenol	9.70	1
6	6	c1(ccc(cc1)O)c2ccccc2	4-phenylphenol	9.55	1
7	7	C(OCCCC)(=O)c1ccc(cc1)O	4-hydroxy nutyl benzoate	8.47	1
8	8	c1(Cl)c(Cl)ccc(O)c1	3,4-dichlorophenol	8.63	1
9	9	c1(C)cc(O)ccc1C	3,4-dimethylphenol	10.36	1
10	10	C(C)(C)(C)c1ccc(cc1)O	4-t-butylyphenol	10.39	1
11	11	C(F)(F)(F)c1cccc(O)c1	3-trifluoromethylphenol	8.95	1
12	12	C(C)(=O)c1ccc(cc1)O	4-hydroxyacetophenone	8.05	1
13	13	c1(C(C)C)ccc(cc1)O	4-isopropylphenol	10.24	1
14	14	O=Cc1cccc(O)c1	3-hydroxybenzaldehyde	8.98	1
15	15	[N+](=O)([O-])c1ccc(cc1)O	4-nitrophenol	7.15	1
16	16	C(C)(=O)Nc1ccc(cc1)O	n-(4-hydroxyphenyl)acetamide	9.38	1
17	17	c1(ccc(cc1)O)Cl	4-chlorophenol	9.41	1
18	18	c1(O)ccc(cc1)C	4-cresol	10.26	1
19	19	c1(ccc(cc1)O)Br	4-bromophenol	9.17	1
20	20	c1(ccccc1)O	phenol	9.99	1
21	21	c1(O)cc(C)cc(C)c1	3,5-dimethylphenol	10.19	1
22	22	c1c(cccc1Cl)O	3-chlorophenol	9.12	1
23	23	c1c(C)cccc1O	3-cresol	10.09	1

24	24	C(OCC)(=O)c1ccc(cc1)O	4-hydroxybenzoic acid, ethyl ester	8.34	1
25	25	c1(cccc(O)c1)C(C)=O	3-hydroxyacetophenone	9.25	1
26	26	c1(ccc(cc1)O)N	4-aminophenol	10.45	1
27	27	O=Cc1ccc(cc1)O	4-hydroxybenzaldehyde	7.61	1
28	28	c1(O)ccc(cc1)CC	4-ethylphenol	10.00	1
29	29	c1(O)ccc(cc1)OC	4-methoxyphenol	10.10	1
30	30	c1c(cccc1O)OC	3-methoxyphenol	9.65	1
31	31	c1(ccc(cc1)O)F	4-fluorophenol	9.91	1
32	32	c1c(cccc1F)O	3-fluorophenol	9.21	1
33	33	C(F)(F)(F)c1ccc(cc1)O	4-trifluoromethylphenol	8.68	1
34	34	c1(O)cc(OC)cc(OC)c1	3,5-dimethoxyphenol	9.34	1
35	35	c1(ccc(cc1)O)I	4-iodophenol	9.21	1
36	36	[N+](=O)([O-])c1cccc(O)c1	3-nitrophenol	8.36	1
37	37	c1(cc(O)ccc1[N+](=O)[O-])[N+](=O)[O-]	3,4-dinitrophenol	5.42	1
38	38	c1(cccc(O)c1)c2ccccc2	3-phenylphenol	9.64	1
39	39	c1(cccc(O)c1)C(C)(C)C	3-(1,1-dimethylethyl)-phenol	10.12	1
40	40	[N+](=O)([O-])c1cc(O)cc(c1)[N+](=O)[O-]	3,5-dinitrophenol	6.69	1
41	41	c1(Cl)cc(O)cc(Cl)c1	3,5-dichlorophenol	8.18	1
42	42	c1c(cccc1N)O	3-aminophenol	9.86	1
43	43	c1c(cccc1Br)O	3-bromophenol	9.03	1
44	44	c1(Cl)c(Cl)cc(cc1Cl)O	3,4,5-trichlorophenol	7.84	1
45	45	c1(cccc(O)c1)C(C)C	3-isopropylphenol	10.16	1
46	46	c1c(CC)cccc1O	3-ethylphenol	9.90	1
47	47	c1c(cccc1O)OCC	3-ethoxyphenol	9.65	1
48	48	c1(O)ccc(cc1)OCC	4-ethoxyphenol	10.13	1
49	49	c1(Br)cc(O)cc(Br)c1	3,5-dibromophenol	8.06	1
50	50	c1c(cccc1I)O	3-iodophenol	9.03	1

51	51	c1(O)ccc(cc1)CCC	4-propylphenol	10.34	1
52	52	c1(O)cc(CC)cc(C)c1	3-ethyl-5-methylphenol	10.10	1
53	53	N#Cc1ccc(cc1)O	4-cyanophenol	7.97	1
54	54	N#Cc1cccc(O)c1	3-cyanophenol	8.61	1
55	55	c1(O)ccc(cc1)CS	4-methiophenol	9.53	1
56	56	c12CCCc1ccc(O)c2	5-indanol	10.32	1
57	57	[N+](=O)([O-])c1c(C)ccc(O)c1	3-nitro-4-cresol	8.62	1
58	58	C(C)(=O)c1ccc(cc1)O	hydroxyacetophenone	8.05	1
Ortho-Substituted Phenols (Capable of forming internal hydrogen bonds)					
1	59	c1(ccccc1O)C(N)=O	2-hydroxybenzamide	8.89	1
2	60	C(=O)(Nc1ccccc1)c2ccccc2O	salicylanilide	7.40	1
3	61	c1(ccccc1O)C(OC)=O	methyl salicylate	9.87	1
4	62	c1(O)c(cccc1C=O)OC	2-vanillin	7.91	1
5	63	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccc(cc2)Cl	3,5,4'-trichloro salicylanilide	4.70	1
6	64	C(=O)(Nc1ccccc1Cl)c2ccccc2O	2'-chloro salicylanilide	7.31	1
7	65	c1(c(O)ccc(c1)N(=O)=O)C(=O)Nc2ccccc2	5-nitro salicylanilide	3.03	1
8	66	C(=O)(Nc1ccc(cc1)Br)c2ccccc2O	4'-bromo salicylanilide	7.31	1
9	67	c1(cc(ccc1O)Br)C(=O)Nc2ccc(cc2)Cl	4'-chloro-5-bromo salicylanilide	6.00	1
10	68	C(=O)(Nc1ccc(cc1)Cl)c2ccccc2O	4'-chloro salicylanilide	7.30	1
11	69	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccccc2	3,5-dichloro salicylanilide	4.70	1
12	70	c1(cc(ccc1O)Cl)C(=O)Nc2ccccc2	5-chlorosalicylanilide	6.17	1
13	71	c1(cc(ccc1O)Cl)C(=O)Nc2ccccc2C	5-chloro-2'-methyl salicylanilide	6.60	1
14	72	C(=O)(Nc1ccc(Cl)cc1Cl)c2ccccc2O	2',4'-dichloro salicylanilide	7.14	1
15	73	N(=O)(=O)c1ccccc1NC(=O)c2ccccc2O	2'-nitro salicylanilide	6.91	1
16	74	c1(cc(ccc1NC(=O)c2ccccc2O)Cl)N(=O)=O	2'-nitro-4'-chloro salicylanilide	6.74	1
17	75	c1(ccc(Br)c(C)c1O)C(NC)=O	5-bromo-2-hydroxy-n,3-dimethyl-benzamide	7.52	1
18	76	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccc(cc2)F	3,5-dichloro-4'-fluoro salicylanilide	4.80	1

19	77	c1(cc(cc(Br)c1O)Br)C(=O)Nc2ccc(Cl)cc2N(=O)=O	3,5-dibromo-2'-nitro-4'-chloro salicylanilide	4.11	1
20	78	C(=O)(Nc1ccc(Cl)cc1C)c2ccccc2O	2'-methyl-4'-chloro salicylanilide	7.43	1
21	79	c1(cc(ccc1O)F)C(=O)Nc2ccc(Br)cc2C	5-fluoro-2'-methyl-4'-bromo salicylanilide	7.10	1
22	80	c1(cc(ccc1O)F)C(=O)Nc2ccc(Cl)cc2C	5-fluoro-2'-methyl-4'-chloro salicylanilide	7.30	1
23	81	c1(cc(cc(Br)c1O)Br)C(=O)Nc2ccc(F)cc2F	3,5-dibromo-2',4'-difluoro salicylanilide	4.77	1
24	82	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccc(F)cc2F	3,5-dichloro-2',4'-difluoro salicylanilide	4.77	1
25	83	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccc(Cl)cc2N(=O)=O	3,5,-4'-trichloro-2'-nitro salicylanilide	4.11	1
26	84	c1(cc(cc(Cl)c1O)Cl)C(=O)Nc2ccc(cc2C)N(=O)=O	3,5-dichloro-2'-methyl-4'-nitro salicylanilide	4.41	1
Ortho-Substituted Phenols					
1	85	c1(c(O)ccc(c1)[N+](=O)[O-])[N+](=O)[O-]	2,4-dinitrophenol	4.09	1
2	86	c1(Cl)c(O)c(Cl)cc(Cl)c1Cl	2,3,4,6-tetrachlorophenol	5.22	1
3	87	c1(c(cccc1N(=O)=O)O)N(=O)=O	2,3-dinitrophenol	4.96	1
4	88	c1(Cl)c(Cl)c(O)c(c(Cl)c1Cl)Cl	pentachlorophenol	4.70	1
5	89	c1(O)c(cccc1Cl)Cl	2,6-dichlorophenol	6.79	1
6	90	c1(O)c(C)cccc1Cl	2-methyl-6-chlorophenol	8.69	1
7	91	c1(cc(cc([N+](=O)[O-])c1O)[N+](=O)[O-])[N+](=O)[O-]	2,4,6-trinitrophenol	0.38	1
8	92	c1(cc(Cl)cc([N+](=O)[O-])c1O)[N+](=O)[O-]	4-chloro-2,6-dinitrophenol	2.96	1
9	93	c1(cc(cc(C(C)CC)c1O)[N+](=O)[O-])[N+](=O)[O-]	2-sec-butyl-4,6-dinitrophenol	4.62	1
10	94	[N+](=O)([O-])c1ccccc1O	2-nitrophenol	7.23	1
11	95	c1(ccccc1O)C(C)C	2-isopropylphenol	10.47	1
12	96	c1(ccccc1O)C(C)(C)C	2-t-butylphenol	10.28	1
13	97	c1(O)c(Cl)cc(cc1Cl)Cl	2,4,6-trichlorophenol	6.23	1
14	98	c1(O)cc(C)ccc1C(C)C	thymol	10.62	1
15	99	c1(C(C)C)c(O)cc(c(Cl)c1)C	chlorothymol	9.98	1
16	100	[N+](=O)([O-])c1c(O)ccc(Cl)c1	4-chloro-2-nitrophenol	6.46	1
17	101	c1(ccccc1O)c2ccccc2	2-phenylphenol	9.92	1
18	102	c1(ccccc1OC)O	2-methoxyphenol	9.98	1

19	103	c1(ccccc1CC)O	2-ethylphenol	10.20	1
20	104	c1(OC)cc(C)ccc1O	4-methyl-2-methoxyphenol	10.28	1
21	105	c1(ccccc1OCC)O	2-ethoxyphenol	10.11	1
22	106	c1c(Cl)c(O)cc(Cl)c1Cl	2,4,5-trichlorophenol	7.40	1
23	107	c1(O)cc(C)ccc1C	2,5-dimethylphenol	10.41	1
24	108	c1(ccccc1O)Cl	2-chlorophenol	8.56	1
25	109	c1(ccccc1O)Br	2-bromophenol	8.45	1
26	110	c1(ccccc1O)N	2-aminophenol	9.75	1
27	111	c1(ccccc1C)O	2-cresol	10.28	1
28	112	c1(c(O)ccc(C(C)(C)C)c1)C(C)(C)C	2,4-di-t-butylphenol	11.72	1
29	113	c1(OC)cc(ccc1O)C=CC	2-methoxy-4-(1-propenyl)phenol	9.88	1
30	114	c1(OC)cc(CC=C)ccc1O	eugenol	10.19	1
31	115	c1(ccc(c(Cl)c1)O)C(C)(C)C	4-(tert-butyl)-2-chlorophenol	8.58	1
32	116	c1(ccc(c(C)c1)O)C(C)(C)C	4-(t-butyl)-2-cresol	10.59	1
33	117	[N+](=O)([O-])c1ccc(c(N)c1)O	2-amino-4-nitrophenol	7.60	1
34	118	[N+](=O)([O-])c1cc(Br)c(c(Br)c1)O	2,6-dibromo-4-nitrophenol	3.39	1
35	119	c1(C)cc(C)ccc1O	2,4-dimethylphenol	10.60	1
36	120	c1(O)c(Br)cc(cc1Br)Br	2,4,6-tribromophenol	6.10	1
37	121	N(=O)(=O)c1c(O)ccc(N)c1	phenol, 4-amino-2-nitro-	7.81	1
38	122	[N+](=O)([O-])c1cc(C)ccc1O	4-methyl-2-nitrophenol	7.40	1
39	123	c1(Cl)c(O)ccc(Cl)c1	2,4-dichlorophenol	7.89	1
40	124	c1(OC)cc(ccc1O)C=O	vanillin	7.40	1
41	125	c1(O)c(cccc1C(C)(C)C)C(C)(C)C	2,6-di-t-butylphenol	11.70	1
42	126	c1(O)c(cc(cc1C(C)(C)C)C)C(C)(C)C	2,6-di-t-butyl-4-methylphenol (bht)	12.23	1
43	127	c2(cc(cc(C1CCCCC1)c2O)[N+](=O)[O-])[N+](=O)[O-]	2-cyclohexyl-4,6-dinitrophenol	4.52	1
44	128	[N+](=O)([O-])c1ccc(cc1O)[N+](=O)[O-]	2,5-dinitrophenol	5.21	1
45	129	c1(ccccc1O)F	2-fluorophenol	8.70	1

46	130	[N+](=O)([O-])c1ccc(cc1O)F	5-fluoro-2-nitrophenol	6.07	1
47	131	c1(O)c(C)cc(c(C)c1)C	2,4,5-trimethylphenol	10.57	1
48	132	c1(C)c(C)cccc1O	2,3-dimethylphenol	10.54	1
49	133	c1(O)c(C)cc(cc1C)C	2,4,6-trimethylphenol	10.86	1
50	134	c1(ccccc1O)I	2-iodophenol	8.51	1
51	135	c1(c(O)c(C)cc(c1)[N+](=O)[O-])[N+](=O)[O-]	4,6-dinitro-o-cresol	4.31	1
52	136	c1(O)c(O)ccc(CCN)c1C	2-methyldopamine=	9.54	1
53	137	[N+](=O)([O-])c1cccc(c1O)[N+](=O)[O-]	2,6-dinitrophenol	3.97	1
54	138	c1(O)c(C)cccc1C	2,6-dimethylphenol	10.62	1
55	139	c1(Cl)c(cccc1Cl)O	2,3-dichlorophenol	7.70	1
56	140	c1(O)c(Cl)ccc(Cl)c1	2,5-dichlorophenol	7.51	1
57	141	[N+](=O)([O-])c1cccc(Cl)c1O	6-chloro-2-nitrophenol	5.48	1
58	142	c1(Br)c(Br)c(O)c(c(Br)c1Br)Br	pentabromophenol	4.62	1
59	143	c1(O)c(cccc1Br)Br	2,6-dibromophenol	6.67	1
60	144	c1(cc(C)cc([N+](=O)[O-])c1O)[N+](=O)[O-]	2,6-dinitro-p-cresol	4.23	1
61	145	N#Cc1ccccc1O	2-cyanophenol	6.86	1
62	146	[N+](=O)([O-])c1ccc(cc1O)Cl	5-chloro-2-nitrophenol	6.05	1
63	147	c1(Br)c(O)ccc(Br)c1	2,4-dibromophenol	7.79	1
64	148	[N+](=O)([O-])c1cc(Cl)c(c(Cl)c1)O	2,6-dichloro-4-nitrophenol	3.55	1
65	149	[N+](=O)([O-])c1ccc(c(Cl)c1)O	2-chloro-4-nitrophenol	5.45	1
66	150	c1(O)cc(ccc1OC)C=O	isovanillin	8.89	1
67	151	c1(ccccc1CCC)O	2-propylphenol	10.47	1
68	152	c1(O)cc(C)cc(C)c1C	2,3,5-trimethylphenol	10.67	1
69	153	c1(C)cc(c([N+]([O-])=O)cc1)O	5-methyl-2-nitrophenol	7.41	1
70	154	c2(cc(cc(c1ccccc1)c2O)[N+](=O)[O-])[N+](=O)[O-]	2,4-dinitro-6-phenylphenol	3.85	1
71	155	c1(C(C)(C)C)cc(C(C)(C)C)cc(C(C)(C)C)c1O	2,4,6-tri(tert-butyl)phenol	12.19	1
72	156	c1(F)c(F)c(O)c(c(F)c1F)F	pentafluorophenol	5.53	1

73	157	[N+](=O)([O-])c2c(O)ccc(c1ccccc1)c2	4-phenyl-2-nitrophenol	6.73	1
74	158	c1(Cl)c(Cl)ccc(Cl)c1O	2,3,6-trichlorophenol	5.80	1
75	159	c1(O)c(Cl)c(Cl)cc(Cl)c1Cl	2,3,5,6-tetrachlorophenol	5.14	1
76	160	c1(cc(cc(Cl)c1O)[N+](=O)[O-])[N+](=O)[O-]	6-chloro-2,4-dinitrophenol	2.10	1
77	161	[N+](=O)([O-])c1c(O)ccc(OC)c1	4-methoxy-2-nitrophenol	7.31	1
78	162	c1(C)cc(Cl)ccc1O	2-methyl-4-chlorophenol	9.71	1
79	163	c1(O)c(Br)cc(cc1Br)C#N	bromoxynil	3.86	1
80	164	c1(cc(C)cc(C)c1O)C(C)(C)C	2-(1,1-dimethylethyl)-4,6-dimethylphenol	12.04	1
81	165	c1(O)c(cccc1C(C)C)C(C)C	phenol, 2,6-bis(1-methylethyl)-	11.10	1
82	166	c1(cc(C)ccc1O)C(C)(C)C	2-(tert-butyl)-4-methylphenol	11.72	1
83	167	[N+](=O)([O-])c1cc(C)c(c(C)c1)O	2,6-dimethyl-4-nitrophenol	7.07	1
84	168	c1(O)c(Cl)cc(cc1Cl)C	4-methyl-2,6-dichlorophenol	7.19	1
85	169	c1(O)c(Cl)cc(cc1Cl)Br	4-bromo-2,6-dichlorophenol	6.21	1
86	170	[N+](=O)([O-])c1c(O)ccc(C(C)CC)c1	4-(sec-butyl)-2-nitrophenol	7.59	1
87	171	C(O)(=O)c1ccc(O)c(Cl)c1	3-chloro-4-hydroxybenzoic acid	7.52	1
88	172	c1(Cl)c(Cl)c(O)cc(Cl)c1Cl	2,3,4,5-tetrachlorophenol	6.35	1
89	173	N(=O)(=O)c1ccc(c(O)c1)C	phenol, 2-methyl-5-nitro-	8.59	1
90	174	Cl-c(cc(c1)C)c(c1)O	2-chloro-4-methylphenol	8.74	1
91	175	c2(c(ON=Cc1cc(Br)c(c(Br)c1)O)ccc(c2)N(=O)=O)N(=O)=O	bromofenoxim	5.46	1
Meta/Para-Substituted Benzoic Acids					
1	176	C(O)(=O)c1ccc(c(Cl)c1)Cl	3,4-dichlorobenzoic acid	3.64	1
2	177	c1(cc(Cl)cc(Cl)c1)C(O)=O	3,5-dichlorobenzoic acid	3.54	1
3	178	S(=O)(=O)(N(CCC)CCC)c1ccc(cc1)C(O)=O	probenecid	3.40	1
4	179	[N+](=O)([O-])c1ccc(cc1)C(O)=O	p-nitrobenzoicacid	3.44	1
5	180	C(O)(c1ccccc1)=O	benzoic acid	4.19	1
6	181	C(O)(=O)c1ccc(cc1)Cl	4-chlorobenzoic acid	3.98	1
7	182	C(O)(=O)c2ccc1c(cccc1)c2	2-naphthoic acid	4.17	1

8	183	C(O)(=O)c1ccc(c(OC)c1)OC	3,4-dimethoxybenzoic acid	4.36	1
9	184	C(O)(=O)c1ccc(cc1)C(C)(C)C	4-(tert-butyl)-benzoic acid	4.40	1
10	185	C(O)(=O)c1ccc(cc1)O	p-hydroxybenzoic acid	4.54	1
11	186	C(O)(=O)c1ccc(cc1)C	p-toluic acid	4.37	1
12	187	C(O)(=O)c1ccc(c(O)c1)O	3,4-dihydroxybenzoic acid	4.48	1
13	188	[N+](=O)([O-])c1cc(cc(c1)[N+](=O)[O-])C(O)=O	3,5-dinitrobenzoic acid	2.82	1
14	189	c1(cc(O)cc(O)c1)C(O)=O	3,5-dihydroxybenzoic acid	4.04	1
15	190	C(O)(=O)c1cccc(N)c1	3-aminobenzoic acid	4.74	1
16	191	C(O)(=O)c1cccc(C)c1	m-toluic acid	4.27	1
17	192	C(O)(=O)c1ccc(cc1)OC	p-methoxybenzoic acid	4.47	1
18	193	[N+](=O)([O-])c1cccc(c1)C(O)=O	m-nitrobenzoic acid	3.46	1
19	194	C(O)(=O)c1ccc(c(OC)c1)O	4-hydroxy-3-methoxybenzoic acid	4.51	1
20	195	c1(O)c(O)cc(cc1O)C(O)=O	3,4,5-trihydroxybenzoic acid	4.21	1
21	196	C(O)(=O)c1ccc(cc1)N	4-aminobenzoic acid	4.85	1
22	197	C(O)(=O)c1cccc(F)c1	m-fluorobenzoic acid	3.86	1
23	198	C(O)(=O)c1ccc(cc1)F	p-fluorobenzoic acid	4.14	1
24	199	c1(cc(ccc1[N+](=O)[O-])C(O)=O)[N+](=O)[O-]	3,4-dinitrobenzoic acid	2.82	1
25	200	c1(O)c(OC)cc(cc1OC)C(O)=O	4-hydroxy-3,5-dimethoxybenzioc acid	4.34	1
26	201	C(O)(=O)c1cccc(Cl)c1	m-chlorobenzoic acid	3.81	1
27	202	C(O)(=O)c1ccc(cc1)C(C)C	cumic acid	4.35	1
28	203	C(O)(=O)c1cccc(Br)c1	m-bromobenzoic acid	3.81	1
29	204	C(O)(=O)c1ccc(cc1)C(C)=O	p-acetylbenzoic acid	3.70	1
30	205	C(O)(=O)c1ccc(cc1)Br	p-bromobenzoic acid	4.00	1
31	206	C(O)(=O)c1cccc(OC)c1	m-methoxybenzoic acid	4.09	1
32	207	C(O)(=O)c1cccc(I)c1	3-iodobenzoic acid	3.85	1
33	208	C(O)(=O)c1ccc(cc1)OCC	p-ethoxybenzoic acid	4.45	1
34	209	c(cc(c1)C(=O)O)c(c1)C=O	4-formylbenzoic acid	3.77	1

35	210	N#Cc1ccc(cc1)C(O)=O	p-cyanobenzoic acid	3.55	1
36	211	C(O)(=O)c1ccc(cc1)CC	4-ethylbenzoic acid	4.35	1
37	212	C(O)(=O)c1ccc(cc1)I	4-iodobenzoic acid	4.00	1
38	213	C(O)(=O)c1cccc(c1)C=O	3-formylbenzoic acid	3.84	1
39	214	Nc1ccc(cc1N)C(O)=O	3,4-diamino-benzoic acid	3.49	1
40	215	c1(cc(OC)cc(OC)c1)C(O)=O	3,5-dimethoxybenzoic acid	3.97	1
41	216	C(O)(=O)c1cccc(c1)C#N	m-cyanobenzoic acid	3.60	1
42	217	C(O)(=O)c1ccc(cc1)Oc2ccccc2	p-phenoxybenzoic acid	4.52	1
43	218	C(O)(=O)c2cccc(Oc1ccccc1)c2	m-phenoxybenzoic acid	3.92	1
44	219	S(C)(=O)(=O)c1ccc(cc1)C(O)=O	p-methylsulfonylbenzoic acid	3.64	1
45	220	C(O)(=O)c1ccc(cc1)OCCC	4-propoxybenzoic acid	4.46	1
46	221	O=C(O)c(cccc1O)c1	4-[(acetylamino)amino]-benzoic acid	4.20	1
Ortho-Substituted Benzoic Acids					
1	222	c1(ccc(cc1O)C)C(O)=O	4-methylsalicylic acid	3.40	1
2	223	c1(ccc(cc1Cl)Cl)C(O)=O	2,4-dichlorobenzoic acid	2.68	1
3	224	c1(cc(Cl)ccc1Cl)C(O)=O	2,5-dichlorobenzoic acid	2.47	1
4	225	c1(ccccc1OC(C)=O)C(O)=O	acetylsalicylic acid	3.49	1
5	226	c1(c(Cl)ccc(Cl)c1Cl)C(O)=O	2,3,6-trichlorobenzoic acid	1.50	1
6	227	c1(c(cccc1Cl)Cl)C(O)=O	2,6-dichlorobenzoic acid	1.59	1
7	228	c1(ccc(cc1OCC)N)C(O)=O	2-ethoxy-4-aminobenzoic acid	5.09	1
8	229	c1(ccccc1Nc2cccc(C)c2C)C(O)=O	mefenamic acid	4.20	1
9	230	c1(ccc(cc1O)N)C(O)=O	p-aminosalicylic acid	3.66	1
10	231	c1(ccccc1O)C(O)=O	salicylic acid	2.97	1
11	232	c1(cccc(C)c1O)C(O)=O	3-methylsalicylic acid	2.95	1
12	233	c1(ccccc1I)C(O)=O	2-iodobenzoic acid	2.93	1
13	234	c1(ccccc1Br)C(O)=O	o-bromobenzoic acid	2.88	1
14	235	c1(ccc(cc1O)O)C(O)=O	2,4-dihydroxybenzoic acid	3.11	1

15	236	c1(cc(C)ccc1O)C(O)=O	5-methylsalicylic acid	3.15	1
16	237	O=C(O)c(c(O)ccc1Br)c1	5-bromosalicylic acid	2.66	1
17	238	c1(ccccc1NC(C)=O)C(O)=O	n-acetyl-o-aminobenzoic acid	3.40	1
18	239	c1(ccc(cc1OC)OC)C(O)=O	2,4-dimethoxybenzoic acid	4.36	1
19	240	c1(ccccc1Nc2ccccc2)C(O)=O	n-phenyl-o-aminobenzoic acid	3.99	1
20	241	c2(C(O)=O)c(O)cc1ccccc1c2	2-naphthalenecarboxylic acid, 3-hydroxy-	2.79	1
21	242	[N+](=O)([O-])c1ccc(c(C(O)=O)c1)O	5-nitrosalicylic acid	2.12	1
22	243	[N+](=O)([O-])c1ccc(c(Cl)c1)C(O)=O	2-chloro-4-nitro-benzoic acid	2.14	1
23	244	c1(ccccc1N)C(O)=O	2-aminobenzoic acid	4.95	1
24	245	c1(ccccc1Cl)C(O)=O	2-chlorobenzoic acid	2.89	1
25	246	C(O)(=O)c1ccccc1C	o-toluic acid	3.98	1
26	247	c1(cc(I)ccc1O)C(O)=O	2-hydroxy-5-iodo-benzoic acid	2.62	1
27	248	c1(C(O)=O)c(cc(cc1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]	2,4,6-trinitrobenzoic acid	0.65	1
28	249	c1(cc(I)cc(I)c1O)C(O)=O	3,5-diiodosalicylic acid	2.30	1
29	250	c1(cc(Cl)cc(N)c1Cl)C(O)=O	3-amino-2,5-dichlorobenzoic acid	3.40	1
30	251	c1(cccc(O)c1O)C(O)=O	2,3-dihydroxybenzoic acid	2.91	1
31	252	c1(c(cccc1O)O)C(O)=O	2,6-dihydroxybenzoic acid	1.05	1
32	253	c1(cc(Cl)ccc1O)C(O)=O	5-chlorosalicylic acid	2.65	1
33	254	c1(ccccc1F)C(O)=O	2-fluorobenzoic acid	3.27	1
34	255	c1(cc(O)ccc1O)C(O)=O	2,5-dihydroxybenzoic acid	2.95	1
35	256	[N+](=O)([O-])c1ccccc1C(O)=O	2-nitrobenzoic acid	2.17	1
36	257	c1(ccccc1C(C)=O)C(O)=O	o-acetylbenzoic acid	4.13	1
37	258	c1(ccccc1OC)C(O)=O	o-methoxybenzoic acid	3.90	1
38	259	c1(cc(cc(C(O)=O)c1O)N(=O)=O)N(=O)=O	2-hydroxy-3,5-dinitro-benzoic acid	0.70	1
39	260	c1(cc(ccc1C(O)=O)[N+](=O)[O-])[N+](=O)[O-]	2,4-dinitrobenzoic acid	1.42	1
40	261	C(O)(=O)c1c(C)cccc1C	2,6-dimethylbenzoic acid	3.35	1
41	262	c1(C(O)=O)c(F)c(F)c(c(F)c1F)C	2,3,5,6-tetrafluoro-4-methyl-benzoic acid	2.00	1

42	263	c1(ccccc1c2ccccc2)C(O)=O	[1,1-biphenyl]-2-carboxylic acid	3.46	1
43	264	c1(c(cccc1OC)OC)C(O)=O	2,6-dimethoxybenzoic acid	3.44	1
44	265	c1(cccc(OC)c1OC)C(O)=O	2,3-dimethoxybenzoic acid	3.98	1
45	266	c1(c(Cl)ccc(Cl)c1OC)C(O)=O	3,6-dichloro-2-methoxybenzoic acid	1.97	1
46	267	c1(ccccc1Oc2ccccc2)C(O)=O	o-phenoxybenzoic acid	3.53	1
47	268	c1(ccccc1C(C)C)C(O)=O	o-isopropylbenzoic acid	3.63	1
48	269	[N+](=O)([O-])c1ccc(c(C(O)=O)c1)Cl	2-chloro-5-nitrobenzoic acid	2.17	1
49	270	[N+](=O)([O-])c1cccc(C(O)=O)c1Cl	2-chloro-3-nitrobenzoic acid	2.02	1
50	271	c2c(NC(c1c(C(O)=O)cccc1)=O)cccc2	n-phenylphthalamic acid	2.50	1
51	272	N(=O)(=O)c1cccc(Cl)c1C(O)=O	2-chloro-6-nitro-benzoic acid	1.34	1
52	273	c1(c(C)cccc1Cl)C(O)=O	2-chloro-6-methyl-benzoic acid	2.75	1
53	274	c1(ccccc1NCC(N)=O)C(O)=O	2-[(acetylamino)amino]-benzoic acid	4.20	1
Anilines					
1	275	Nc1cc(C(O)=O)ccc1	3-aminobenzoic acid	4.53	3
2	276	Nc1ccc(C(O)=O)cc1	4-aminobenzoic acid	2.38	2
3	277	Nc1cc(O)c(C(O)=O)cc1	p-aminosalicylic acid	2.05	2
4	278	Nc1ccc(O)cc1	4-amino-phenol	5.48	2
5	279	Nc1cc(O)ccc1	3-amino-phenol	4.37	2
6	280	Nc1cc(N(=O)=O)cc(N(=O)=O)c1	3,5-dinitroaniline	0.30	2
7	281	Nc1cc(Cl)cc(Cl)c1	3,5-dichloroaniline	2.51	2
8	282	Nc2ccc(c1ccccc1)cc2	4-aminobiphenyl	4.35	2
9	283	Cc1c(N(=O)=O)ccc(N)c1	3-methyl-4-nitroaniline	1.64	2
10	284	C(c1ccccc1)(=O)c2ccc(cc2)N	4-benzoylaniline	2.24	2
11	285	[N+](=O)([O-])c1c(C)cc(cc1C)N	3,5-dimethyl-4-nitrobenzenamine	2.54	2
12	286	[N+](=O)([O-])c1c(C)ccc(N)c1	2-nitro-p-toluidine	0.40	2
13	287	S(C)(=O)(=O)c1ccc(cc1)N	4-methylsulfonylaniline	1.35	2
14	288	N(=O)(=O)c1c(Cl)ccc(N)c1	4-chloro-3-nitro-benzenamine	1.90	2

15	289	C(F)(F)(F)c1ccc(cc1)N	p-trifluoromethylaniline	2.45	2
16	290	C(OC)(=O)c1ccc(cc1)N	methyl-p-aminobenzoate	2.47	2
17	291	C(OCCCC)(=O)c1ccc(cc1)N	butyl-p-aminobenzoate	2.47	2
18	292	C(OCCC)(=O)c1ccc(cc1)N	propyl-p-aminobenzoate	2.49	2
19	293	C(OCC)(=O)c1ccc(cc1)N	p-aminobenzoic acid, ethyl ester	2.51	2
20	294	c1(Cl)c(Cl)ccc(N)c1	3,4-dichloroaniline	2.97	2
21	295	C(F)(F)(F)c1cccc(N)c1	3-trifluoromethylaniline	3.49	2
22	296	c1c(cccc1N)Br	m-bromoaniline	3.58	2
23	297	c1c(cccc1N)I	3-iodo-benzenamine	3.61	2
24	298	c1(ccc(cc1)I)N	4-iodo-benzenamine	3.78	2
25	299	c1(ccc(cc1)Br)N	p-bromoaniline	3.86	2
26	300	c1(C)cc(ccc1Br)N	3-methyl-4-bromoaniline	4.05	2
27	301	c1(ccccc1N)C(O)=O	2-aminobenzoic acid	2.14	2
28	302	c1(ccccc1O)N	o-aminophenol	4.84	2
29	303	[N+](=O)([O-])c1ccc(c(N)c1)O	2-amino-4-nitrophenol	3.10	2
30	304	c1(Br)c(N)ccc(Br)c1	2,4-dibromoaniline	2.30	2
31	305	[N+](=O)([O-])c1cc(C)ccc1N	3-nitro-4-toluidine	3.03	2
32	306	c1(ccccc1N)c2ccccc2	2-aminobiphenyl	3.83	2
33	307	[N+](=O)([O-])c1cc(Cl)c(c(Cl)c1)N	2,6-dichloro-4-nitroaniline	-2.55	2
34	308	[N+](=O)([O-])c1cc(C)c(c(C)c1)N	2,6-dimethyl-4-nitrobenzenamine	0.98	2
35	309	c1(N)c(Cl)ccc(Cl)c1	2,5-dichloroaniline	2.05	2
36	310	c1(c(N)ccc(c1)[N+](=O)[O-])[N+](=O)[O-]	2,4-dinitroaniline	-4.25	2
37	311	[N+](=O)([O-])c1c(N)ccc(Cl)c1	4-chloro-2-nitroaniline	-1.02	2
38	312	[N+](=O)([O-])c1ccc(c(Cl)c1)N	2-chloro-4-nitroaniline	-0.94	2
39	313	c1(N)c(F)c(F)c(c(F)c1F)F	2,3,4,5,6-pentafluoroaniline	-0.28	2
40	314	c1(N)c(cccc1Cl)Cl	2,6-dichloroaniline	0.42	2
41	315	N(=O)(=O)c1c(N)ccc(OC)c1	4-methoxy-2-nitro-benzenamine	0.77	2

42	316	[N+](=O)([O-])c1ccc(c(C)c1)N	4-nitro-2-toluidine	1.04	2
43	317	c1(Cl)c(cccc1N)Cl	2,3-dichloroaniline	1.76	2
44	318	c1(Cl)c(N)ccc(Cl)c1	2,4-dichloroaniline	2.00	2
45	319	c1(ccccc1N)C(OCC)=O	o-aminobenzoic acid, ethyl ester	2.18	2
46	320	c1(ccccc1N)C(OC)=O	methyl anthranilate	2.23	2
47	321	[N+](=O)([O-])c1ccc(c(N)c1)C	5-nitro-2-toluidine	2.35	2
48	322	c1(c(C)c(C)c(c(C)c1C)N)[N+](=O)[O-]	2,3,5,6-tetramethyl-4-nitrobenzenamine	2.36	2
49	323	[N+](=O)([O-])c1ccc(c(N)c1)OC	2-methoxy-5-nitroaniline	2.49	2
50	324	c1(ccccc1Br)N	o-bromoaniline	2.53	2
51	325	c1(ccccc1I)N	2-iodoaniline	2.60	2
52	326	N(=O)(=O)c1cccc(c1N)N(=O)=O	2,6-dinitroaniline	-5.00	2
53	327	c1(N)c(Cl)cc(cc1Cl)Cl	2,4,6-trichloroaniline	-0.03	2
54	328	c1c(Cl)c(Cl)cc(Cl)c1N	2,4,5-trichloroaniline	1.09	2
55	329	c1(N)c(OC)ccc(OC)c1	2,5-dimethoxyaniline	3.93	2
Han 2006 - Fluorophenols					
#	330	Oc1ccc(F)cc1F	2,4-difluorophenol	8.58	4
#	331	Oc1c(F)cccc1F	2,6-difluorophenol	7.51	4
#	332	Oc1c(F)c(F)cc(F)c1F	2,3,5,6-tetrafluorophenol	6	4
Regnar 2000					
R_3	333	COc1cc(CO)ccc1O	4-(hydroxymethyl)-2-(methoxy)phenols	9.78	5
R_4	334	COCc1ccc(O)c(OC)c1	2-(methyloxy)-4-[(methyloxy)methyl]phenol	9.79	5
R_7	335	COC(=O)c1ccc(O)c(OC)c1	methyl 4-hydroxy-3-(methyloxy)benzoate	8.3	5
R_8	336	COc1cc(CCO)ccc1O	4-(2-hydroxyethyl)-2-(methoxy)phenol	10.09	5
R_10	337	COc1cc(ccc1O)C(C)O	4-(1-hydroxyethyl)-2-(methoxloxy)phenol	9.83	5
R_11	338	COC(C)c1ccc(O)c(OC)c1	2-(methyloxy)-4-[1-(methyloxy)ethyl]phenol	9.75	5
R_12	339	COc1cc(ccc1O)C(O)CO	1-[4-hydroxy-3-(methyloxy)phenyl]-1,2-ethanediol	9.5	5
	340	COc1cc(ccc1O)C(C)=O	1-{4-hydroxy-3-(methyloxy)phenyl]ethanone	7.81	5

R_14					
R_16	341	CCCc1ccc(O)c(OC)c1	2-(methyloxy)-4-propylphenol	9.85	5
R_20	342	COc1cc(\C=C\CO)ccc1O	4-[(1E)-3-hydroxy-1-propen-1-yl]-2-(methyloxy)phenol	9.54	5
R_21	343	COc1cc(\C=C\C=O)ccc1O	(2E)-3-[4-hydroxy-3-(methyloxy)phenyl]-2-propenal	7.94	5
R_23	344	CCC(O)c1ccc(O)c(OC)c1	4-(1-hydroxypropyl)-2-(methyloxy)phenol	9.83	5
R_25	345	CCC(=O)c1ccc(O)c(OC)c1	1-[4-hydroxy-3-(methyloxy)phenyl]-1-propanone	7.98	5
R_26	346	COc1cc(ccc1O)C(=O)C(C)O	2-hydroxy-1-[4-hydroxy-3-(methyloxy)phenyl]-1-propanone	7.32	5
R_27	347	COc1cc(ccc1O)C(O)C(CO)Oc1ccccc1OC	1-[4-hydroxy-3-(methyloxy)phenyl]-2-{[2-(methyloxy)phenl]oxy}-1,3-propanediol	9.88	5
R_33	348	COCc1cc(\C=C/Oc2ccccc2OC)ccc1O	2-[methyloxy)methyl]-4-((Z)-2-{[2-methyloxy)phenyl]oxy}ethenyl)phenol	9.49	5
meta/para nitrophenols added					
#	349	Oc1ccc(c(F)c1)[N+]([O-])=O	3-fluoro-4-nitrophenol	5.3	6
#	350	Oc1cc(F)c(c(F)c1)[N+]([O-])=O	3,5-difluoro-4-nitrophenol	4.4	6
#	351	Cc1cc(O)ccc1[N+]([O-])=O	3-methyl-4-nitrophenol	7.29	7
#	352	Cc1cc(O)cc(C)c1[N+]([O-])=O	3,5-dimetyl-4-nitrophenol	8.25	8
#	353	Oc1ccc(c(Cl)c1)[N+]([O-])=O	3-chloro-4-nitrophenol	6.49	8
Meta-Anilines					
	354	Nc1cccc(N)c1	3-aminoaniline	4.88	9
	355	Nc1cccc(Cl)c1	3-chloroaniline	3.34	9
	356	Nc1cccc(c1)C#N	3-cyanoaniline	2.76	9
	357	Nc1cccc(F)c1	3-fluoroaniline	3.59	9
	358	COc1cccc(N)c1	3-methoxyaniline	4.2	9
	359	Cc1cccc(N)c1	3-methylaniline	4.69	9
	360	Nc1cccc(c1)[N+]([O-])=O	3-nitroaniline	2.5	9

361	Cc1cc(C)cc(N)c1	3,4-dimethylaniline	5.17	9
362	Nc1ccc(O)c(N)c1	3-amino-4-hydroxyaniline	5.7	9
363	COc1ccc(N)cc1Br	3-bromo-4-methoxyaniline	4.08	9
364	Cc1ccc(N)cc1Br	3-bromo-4-methylaniline	3.98	9
365	Cc1ccc(N)cc1Cl	3-chloro-4-methylaniline	4.05	9
366	Nc1cc(Br)cc(Br)c1	3,5-dibromoaniline	2.34	9
367	COc1cc(N)cc(OC)c1	3,5-dimethoxyaniline	3.82	9
368	Cc1cc(C)cc(N)c1	3,5-dimethylaniline	4.91	9
369	COc1cc(N)cc(Cl)c1	3-chloro-5-methoxyaniline	3.1	9
370	COc1cc(N)cc(c1)[N+]([O-])=O	3-methoxy-5-nitroaniline	2.11	9
371	Nc1cc(Br)c(O)c(Br)c1	3,5-dibromo-4-hydroxyaniline	3.2	9
372	COc1c(Br)cc(N)cc1Br	3,5-dibromo-4-methoxyaniline	2.98	9
373	Cc1c(Br)cc(N)cc1Br	3,5-dibromo-4-methylaniline	2.87	9

Zhang

373	OC(=O)c1ccccc1O	2-hydroxybenzoic acid	2.98	10
374	OC(=O)c1cccc(O)c1	3-hydroxybenzoic acid	4.08	10
375	OC(=O)c1ccc(O)cc1	4-hydroxybenzoic acid	4.58	10
376	OC(=O)c1cccc(O)c1O	2,3-hydroxybenzoic acid	2.98	10
377	OC(=O)c1ccc(O)cc1O	2,4-hydroxybenzoic acid	3.29	10
378	OC(=O)c1cc(O)ccc1O	2,5-hydroxybenzoic acid	2.97	10
379	OC(=O)c1c(O)cccc1O	2,6-hydroxybenzoic acid	1.3	10
380	OC(=O)c1ccc(O)c(O)c1	3,4-hydroxybenzoic acid	4.48	10
381	OC(=O)c1cc(O)cc(O)c1	3,5-hydroxybenzoic acid	4.04	10
382	OC(=O)c1c(O)cc(O)cc1O	2,4,6-hydroxybenzoic acid	1.68	10
383	OC(=O)c1cc(O)c(O)c(O)c1	3,4,5-hydroxybenzoic acid	4.19	10
384	OC(=O)c1ccccc1	hydroxybenzoic acid	4.2	10

Drug Molecules Liao

385	CC(=O)Nc1ccc(O)cc1	Acteaminophen	9.63	11
386	CN(C)[C@H]1[C@@H]2C[C@H]3C(C(=O)c4c(O)ccc(Cl)c4[C@@]3(C)O)=C(O)[C@]2(O)C(=O)C(C(N)=O)=C1O	Chlortetracycline	9.3	11
387	CN(C)[C@H]1[C@@H]2C(O)[C@H]3C(C(=O)c4c(O)cccc4[C@@]3(C)O)=C(O)[C@]2(O)C(=O)C(C(N)=O)=C1O	Oxytetracycline	9.11	11
388	CN(C)[C@H]1[C@@H]2C[C@H]3C(C(=O)c4c(O)cccc4[C@@]3(C)O)=C(O)[C@]2(O)C(=O)C(C(N)=O)=C1O	Tetracycline	9.69	11
389	Oc1c(I)cc(Cl)c2cccnc12	Clioquinol	8.16	11
390	C[C@]12CC[C@H]3C(=CCc4cc(O)ccc34)[C@@H]1CC[C@H]2O	17a-Dihydroequiline	10.29	11
391	C[C@]12CC[C@H]3C(=CCc4cc(O)ccc34)[C@@H]1CC[C@@H]2O	17b-Dihydroequilin	9.77	11
392	C[C@]12CCc3c(ccc4cc(O)ccc34)C1CC[C@@H]2O	Equilenin	9.77	***
393	CSCC[C@H](NC(=O)[C@@H](Cc1ccccc1)NC(=O)CNC(=O)CNC(=O)[C@H](N)Cc1ccc(O)cc1)C(O)=O	Enkephalin	9.89	11
394	C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CC[C@@H]2O	17B-Estradiol	10.71	11
395	C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CCC2=O	Estrone	10.34	11
396	C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CC[C@@]2(O)C#C	Ethinyl estradiol	10.4	11
397	CC(C)NCC(O)c1ccc(O)c(O)c1	Isoproterenol	10.07	11
398	C[C@H](CCc1ccccc1)NC[C@@H](O)c1ccc(O)c(c1)C(N)=O	Labetalol	7.41	11
399	CN1CC[C@@]23[C@H]4Oc5c2c(CC1C3C=C[C@@H]4O)ccc5O	Morphine	9.4	11
400	CN1CC[C@@]23[C@H]4Oc5c2c(CC1C3C=C[C@@H]4O)ccc5O	Normoorphine	9.8	11
401	CCCN(C)[C@H]1C=C(OC2OC(C(O)[C@H](O)[C@H]2O)C(O)=O)[C@@H]2Oc3c4C2C1CCc4ccc3O	Morphine-6-glucuronide	9.36	11
402	N[C@@H](Cc1ccc(O)cc1)C(O)=O	Tyrosine	10.27	11
403	COc1cc(C=O)ccc1O	Vanillin	7.4	11
404	COc1ccc(C=O)cc1O	iso-Vanillin	8.89	11
405	COc1cccc(C=O)c1O	ortho-Vanillin	7.91	11
406	Nc1ccc(cc1)C(O)=O	4-Aminobenzoic acid	2.5	11
407	CCOC(=O)c1ccc(N)cc1	Benzocaine	2.52	11
408	CCN(CC)CCOC(=O)c1ccc(N)cc1	Procaine	2.29	11
409	CC(=O)NS(=O)(=O)c1ccc(N)cc1	Sulfacetamide	1.76	11

1) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E., Estimation of pKa Using Semiemperical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.,* **2002,** 21, 457-471.

2) Tehan, B. G.; Lloyd, E. J.; G., W. M.; Pitt, W. R.; Gancia, G.; Manallack, D. T., Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.-Act. Relat.,* **2002,** 21, (5), 473-485.

3) Avdeef, A., *Absorption and Drug Development: Solubility, Permeability and Charge State*. Wiley-Interscience: New Jersey, USA, 2003.

4) Han, J.; Tao, F. M., Correlations and Predictions of pK_a Values of Fluorophenols and Bromophenols Using Hydrogen-Bonded Complexes with Ammonia. *J. Phys. Chem. A* **2006,** 110, 257-263.

5) Ragnar, M.; Lindgren, C. T.; Nilverbrant, N. O., pK_a Values of Guaiacyl and Syringyl Phenols Related to Lignin. *J. Wood. Chem. Tech.* **2000,** 20, 277-305.

6) Chapman, E.; Bryan, M. C.; Wong, C. H., Mechanistic Studies of *B*-Artlsulfotransferase IV. *PNAS* **2003,** 100, 910-915.

7) Schafer, B.; Engwald, W., Enrichment of Nitrophenols from Water by Means of Solid-Phase Microextraction. *Frenenius' Journal of Analytical Chemistry* **1995,** 352, 535-536.

8) Vaughan, W. R., Effects of Alkyl Groups on 4-Nitro- and 4-Nitroso-Phenols. *J. Org. Chem.* **1956,** 21, 1201-1210.

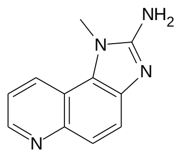
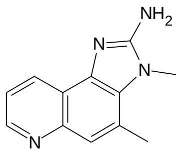
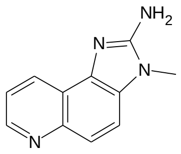
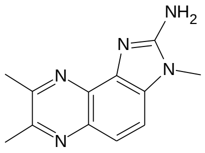
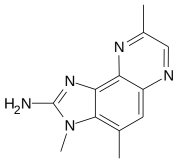
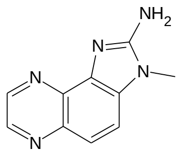
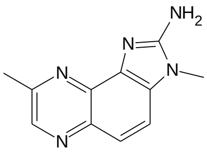
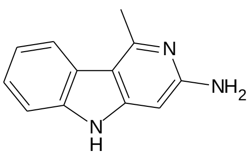
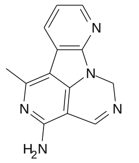
9) Chaudry, U. A.; Popelier, P. L. A., Estimation of pKa using Quantum Topological Molecular Similarity (QTMS) Descriptors: Application to Carboxylic Acids, Anilines and Phenols. *J.Org.Chem.* **2004,** 69, 233-241.

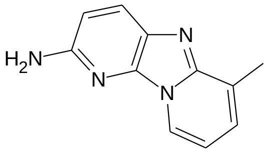
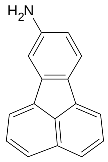
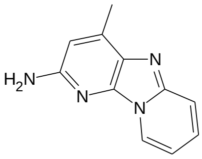
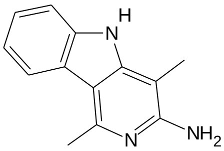
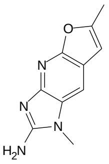
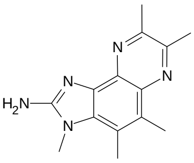
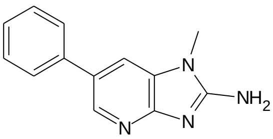
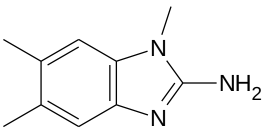
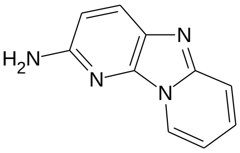
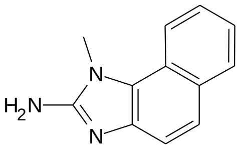
10) 10, J. D.; Zhu, Q. Z.; S.J., L.; Tao, F. M., Prediction of Aqueous pK_a Values of Hydroxybenzoic Acids Using Hydrogen-Bonded Complexes with Ammonia. *Chem. Phys. Lett.* **2009,** 475, 15-18.

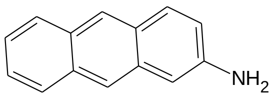
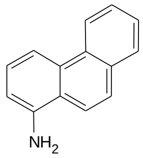
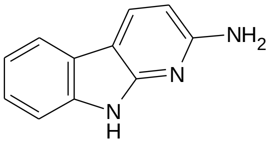
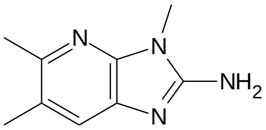
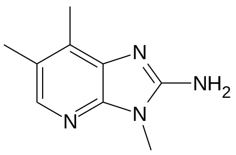
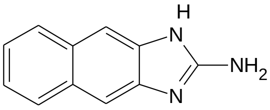
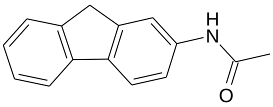
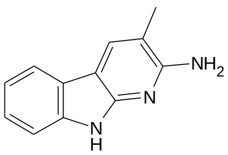
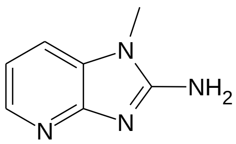
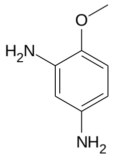
11) Liao, C.; Nicklaus, M. C., Comparison of Nine Programs Predicting pK_a Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009,** 49, 2801-2812.

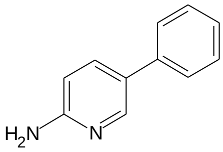
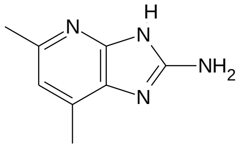
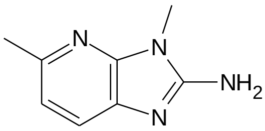
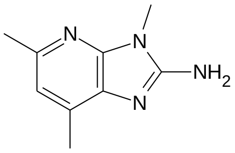
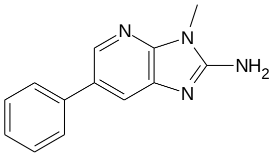
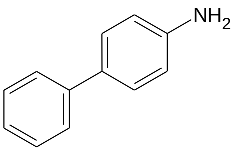
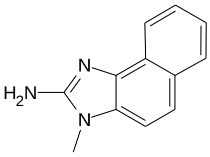
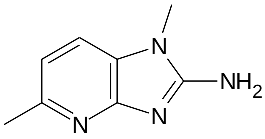
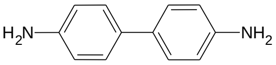
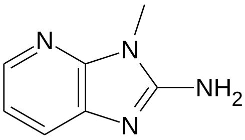
Appendix D

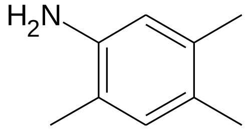
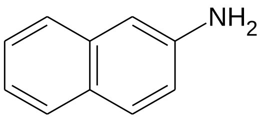
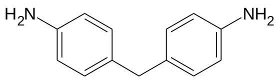
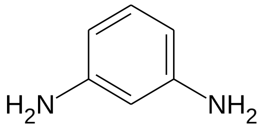
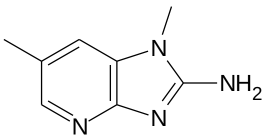
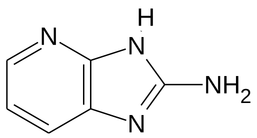
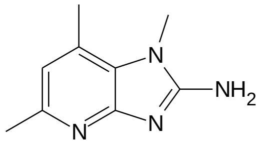
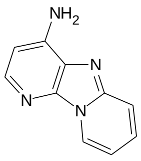
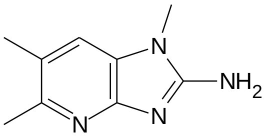
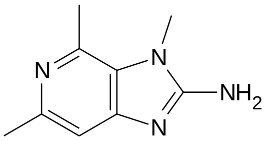
Hatch Data set

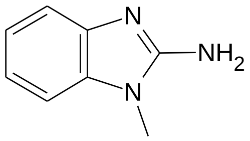
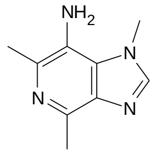
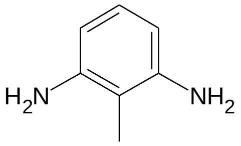
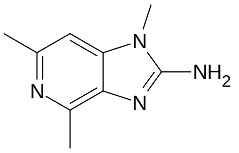
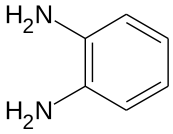
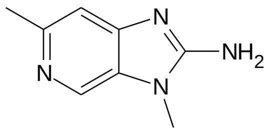
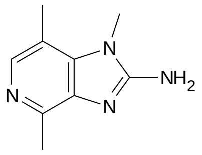
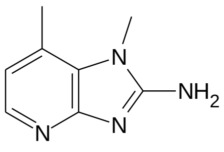
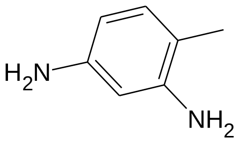
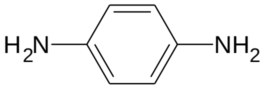
CAS	ID	Structure	LogMP TA98	AA/MIA
102408-25-3	1		5.79	MIA
77094-11-2	2		5.15	
76180-96-6	3		4.7	MIA
92180-79-5	4		4.66	
95896-78-5	5		4.56	MIA
108354-47-8	6		4.32	MIA
77500-04-0	7		4.25	
72254-58-1	8		4.17	
78859-36-6	9		4.17	

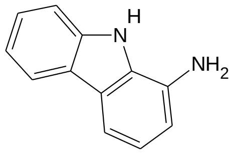
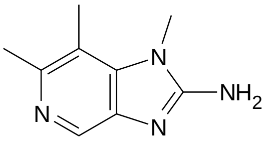
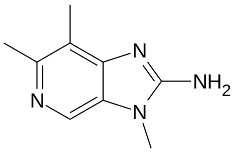
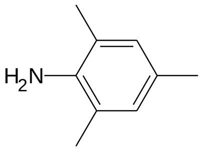
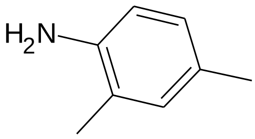
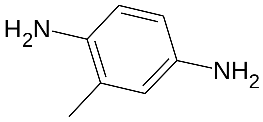
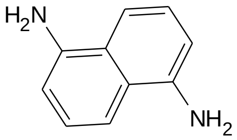
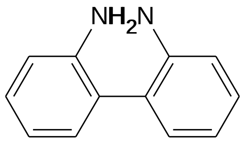
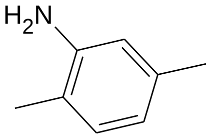
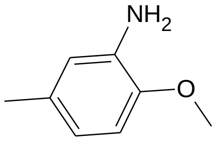
67730-11-4	10		3.98	
5869-25-0	11		3.8	
75679-01-5	12		3.4	
75104-43-7	13		3.39	
	14		3.31	MIA
146177-62-0	15		2.99	
105650-23-5	16		2.63	MIA
15777-02-3	17		2.54	MIA
67730-10-3	18		2.36	
	19		2.29	

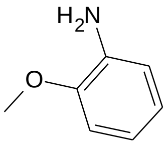
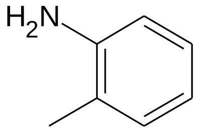
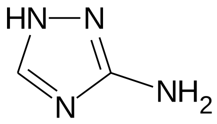
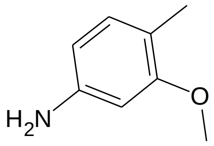
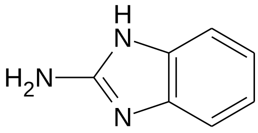
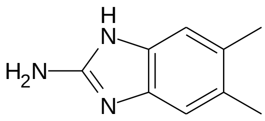
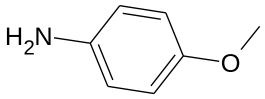
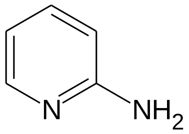
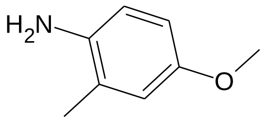
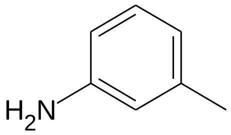
613-13-8	20		2.17	AA
4176-53-8	21		2.11	AA
26148-68-5	22		1.9	
57667-51-3	23		1.813	MIA
401560-72-3	24		1.747	MIA
102408-31-1	25		1.54	
53-96-3	26		1.186	
68006-83-7	27		1.11	
178885-60-4	28		1.022	MIA
39156-41-7	29		0.913	

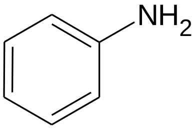
33421-40-8	30		0.83	
	31		0.81	
155789-83-6	32		0.778	MIA
132898-06-7	33		0.693	
17351-87-1	34		0.65	MIA
92-67-1	35		0.626	AA
35199-58-7	36		0.59	
155789-81-4	37		0.44	MIA
92-87-5	38		0.396	AA
30458-69-6	39		0.37	MIA

137-17-7	40		0.361	
91-59-8	41		0.332	AA
101-77-9	42		0.258	AA
108-45-2	43		0.15	AA
132898-04-5	44		0.11	MIA
	45		0.03	
401560-75-6	46		0.001	MIA
81810-23-3	47		-0.126	
161091-55-0	48		-0.263	MIA
193690-74-3	49		-0.392	MIA

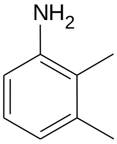
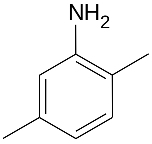
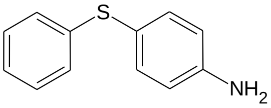
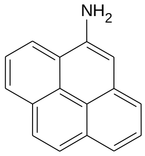
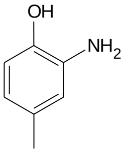
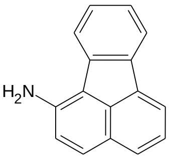
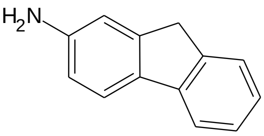
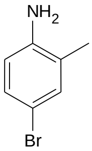
1622-57-7	50		-0.43	MIA
193690-81-2	51		-0.453	
823-40-5	52		-0.496	AA
401560-75-4	53		-0.55	MIA
95-54-5	54		-0.77	AA
155789-84-7	55		-0.79	MIA
193690-65-2	56		-0.8	MIA
155789-82-5	57		-0.887	MIA
95-80-7	58		-0.894	AA
106-50-3	59		-0.983	AA

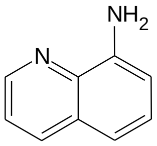
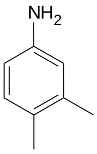
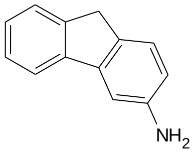
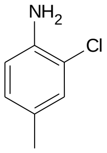
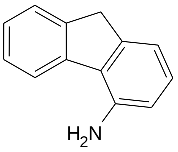
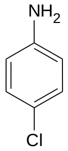
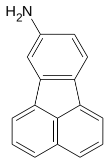
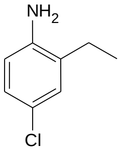
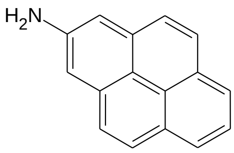
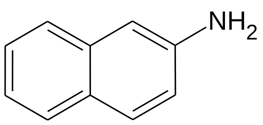
18992-86-4	60		-1.04	AA
193690-71-0	61		-1.055	MIA
401560-74-5	62		-1.055	MIA
88-05-1	63		-1.11	AA
95-68-1	64		-1.16	AA
95-70-5	65		-1.28	AA
2243-62-1	66		-1.35	AA
1454-80-4	67		-1.52	
95-78-3	68		-1.64	AA
120-71-8	69		-1.66	AA

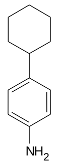
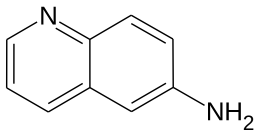
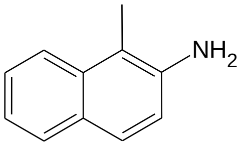
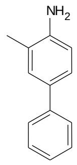
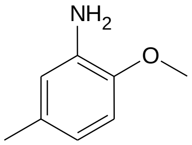
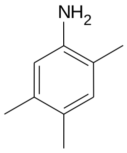
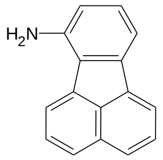
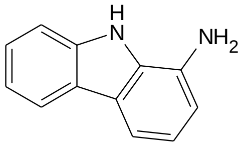
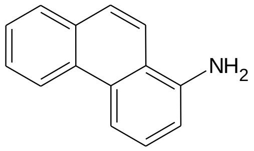
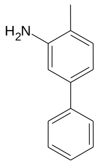
90-04-0	70		-1.672	AA
95-53-4	71		-1.8	AA
61-82-5	72		-1.9	
16452-01-0	73		-1.96	AA
934-32-7	74		-1.97	
29096-75-1	75		-2	
104-94-9	76		-2.305	AA
504-29-0	77		-2.41	
102-50-1	78		-2.58	AA
108-44-1	79		-2.92	AA

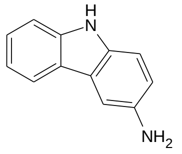
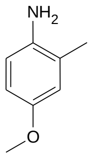
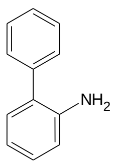
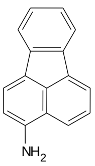
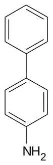
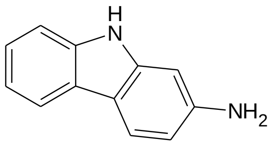
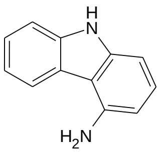
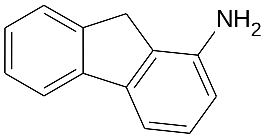
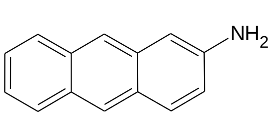
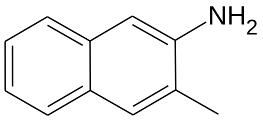
62-53-3	80		-3.39	AA

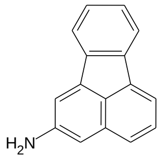
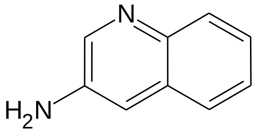
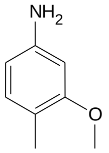
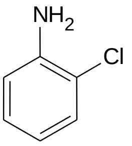
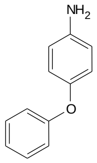
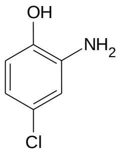
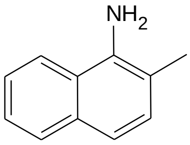
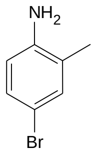
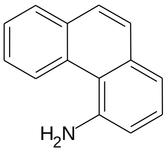
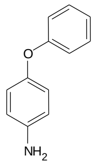
Gramatica data set

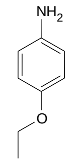
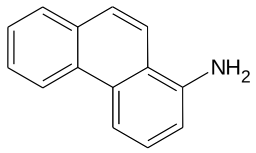
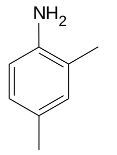
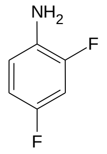
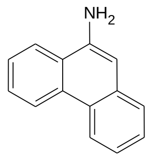
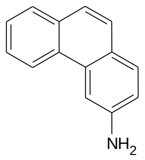
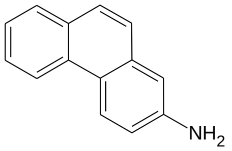
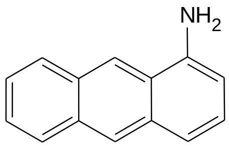
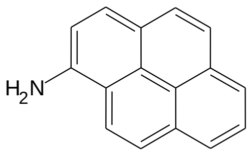
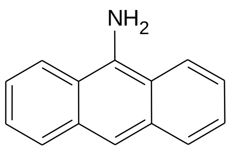
ID	Structure	LogMP TA98	LogMP TA100
1		-1.36	
2		-1.43	-2.4
4		0.48	0.31
5		2.69	3.16
6		-1.68	-2.1
7		2.34	3.35
8		0.78	1.93
10		-0.64	

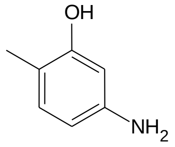
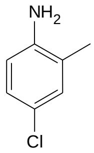
11		-0.53	1.14
12		-1.08	
13		0.1	0.89
14		-0.4	
15		1.08	1.13
16		-1.15	-2.52
17		2.3	3.8
18		0.08	
19		2.58	3.5
20		0.39	-0.67

21		-0.14	-1.24
22		-1.22	-2.67
23			0.84
24		1.12	
26		-1.85	-2.05
27		-1.78	-1.32
29		2.33	2.88
30		-0.25	-1.04
31		1.79	2.38
32		0.09	

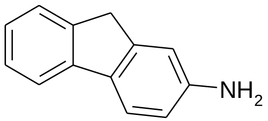
33		-0.11	-0.48
34		-2.41	-3
35		-0.51	-1.49
36		2.25	3.31
37		0.85	-0.14
42		-0.56	0.6
43		-0.47	-1.42
44		-0.04	0.43
45		2.76	2.62
46		1.09	

47		2.87	2.32
48		0.07	-3.14
49		-0.81	-1.96
50		-2.05	-3
51		0.63	0.38
52		-2	-3
53		-0.37	
55		0.46	
56		-0.11	
57		-0.27	-1.14

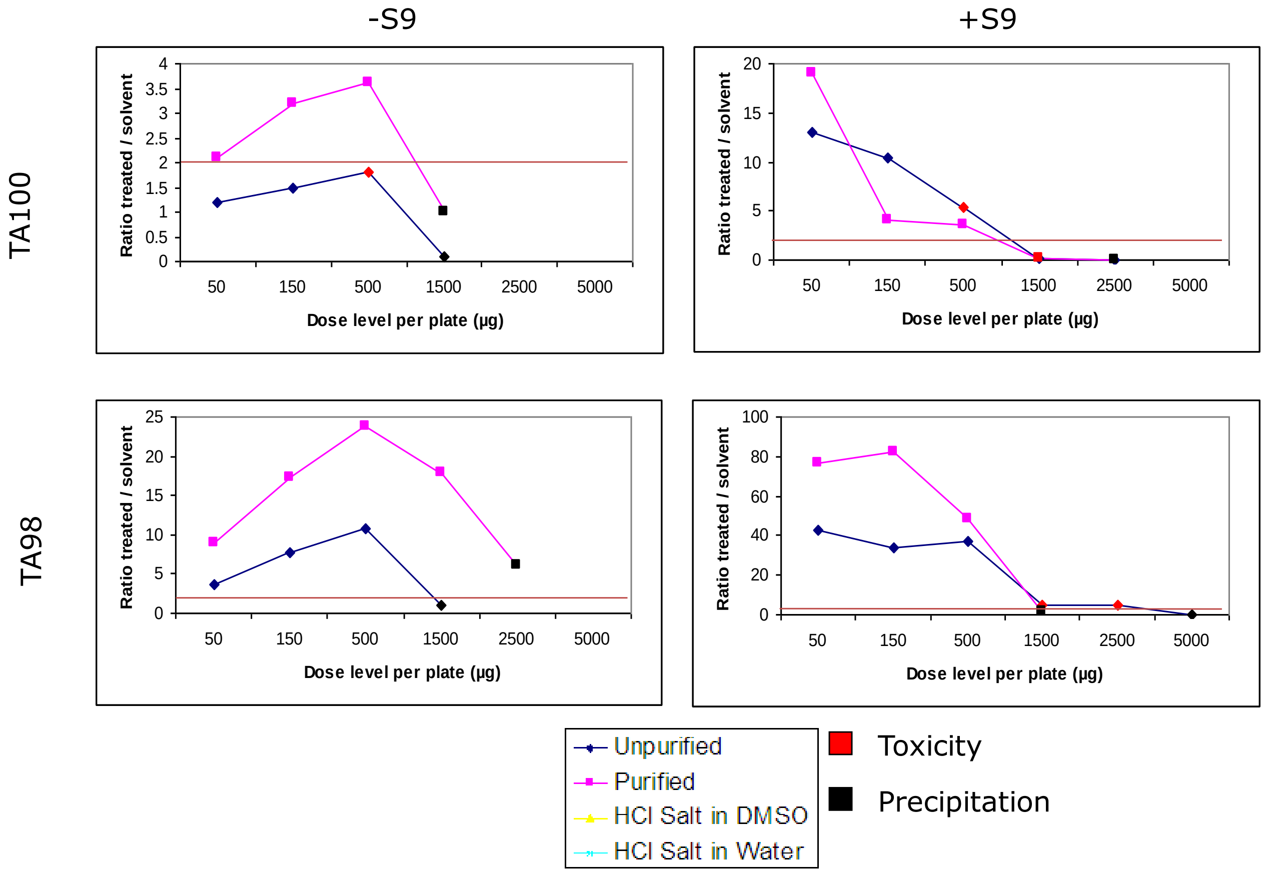
58		-0.61	-2.3
59		-1	-0.6
60		-0.23	-2.22
61		-2.52	-2.7
63		2.79	2.98
65		2.66	3.77
66		2.74	2.46
67		0.36	1.18
68		1.05	1.43
69		-0.24	0.87

75		-1.82	-1.4
79		0.38	

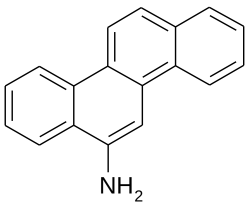
Appendix E

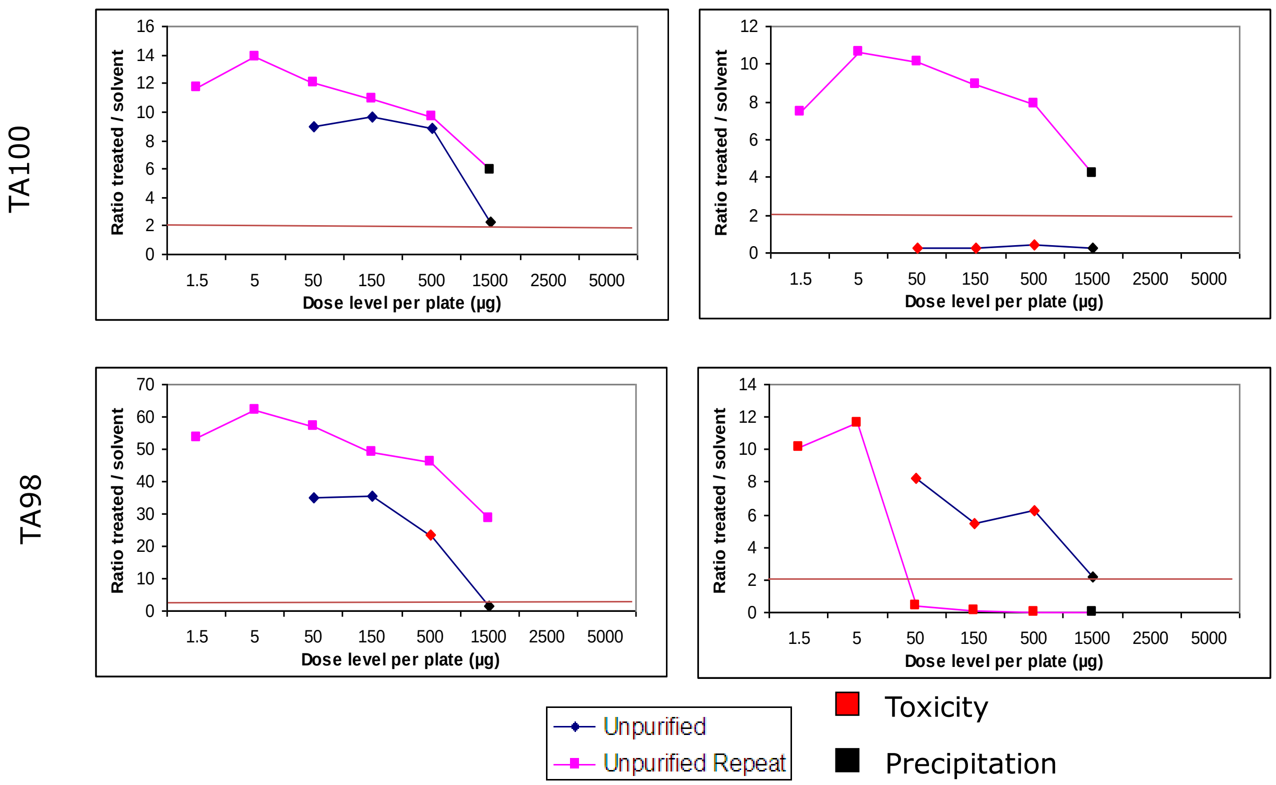


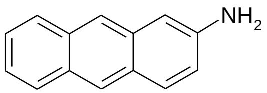
Ames Results
2-aminofluorene



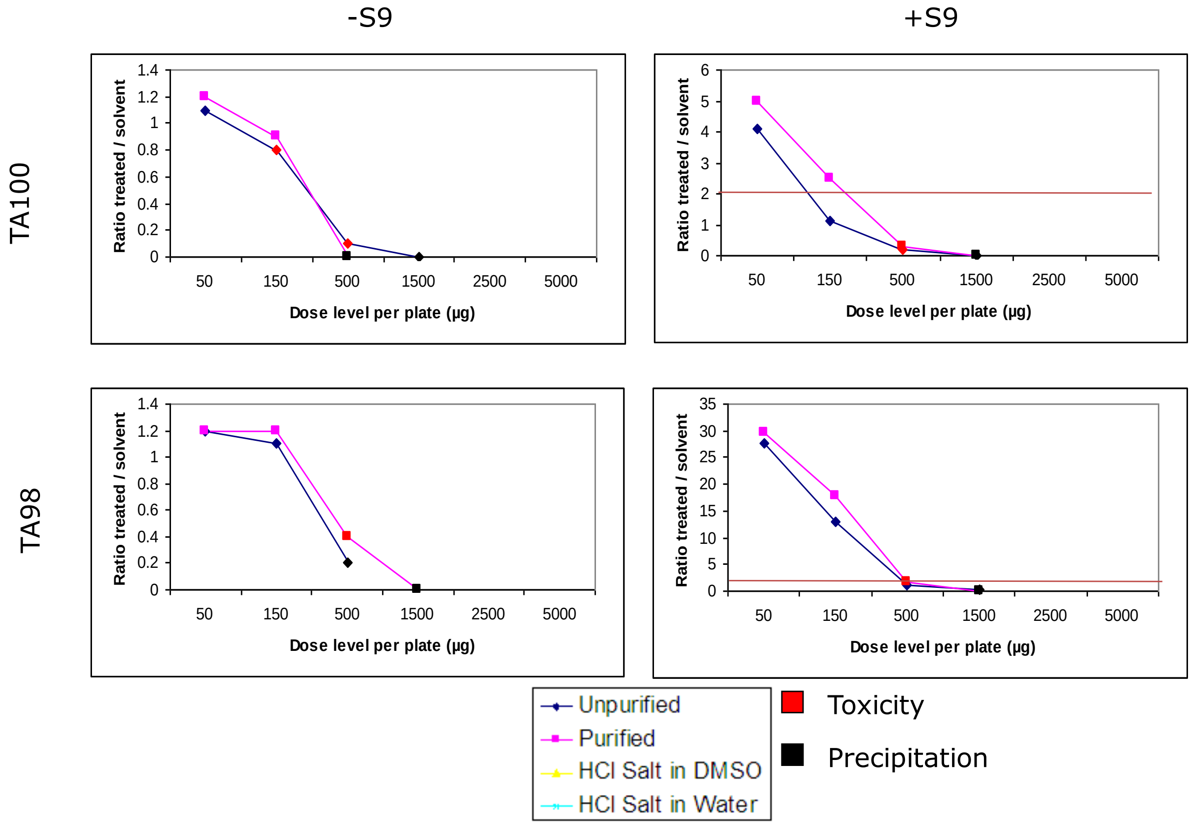
Ames Results
6-chrysenylamine

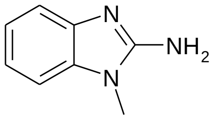




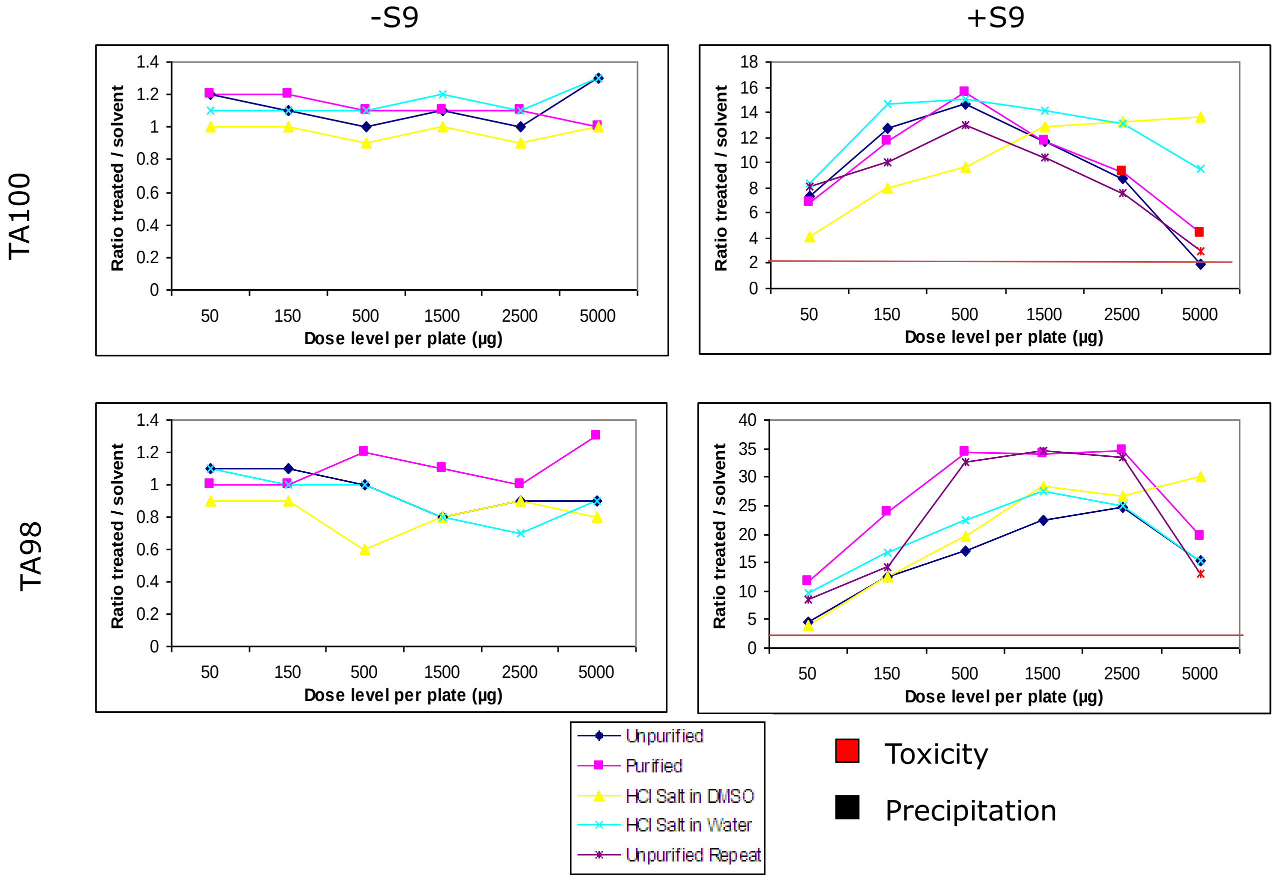


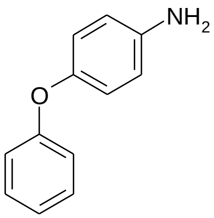
Ames Results
2-aminoanthracene



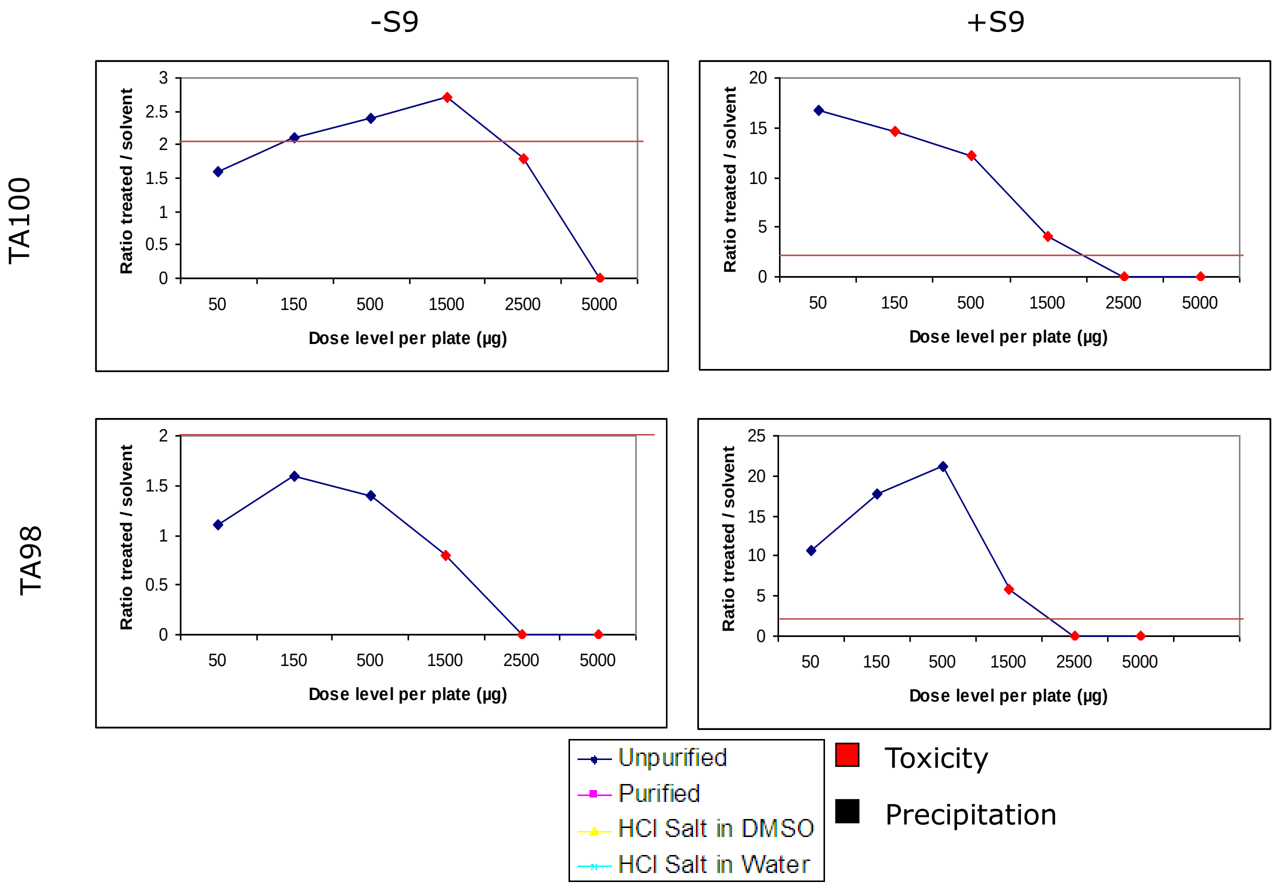


Ames Results
1-methyl-2-aminobenzimidazole

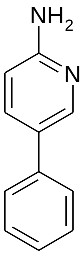


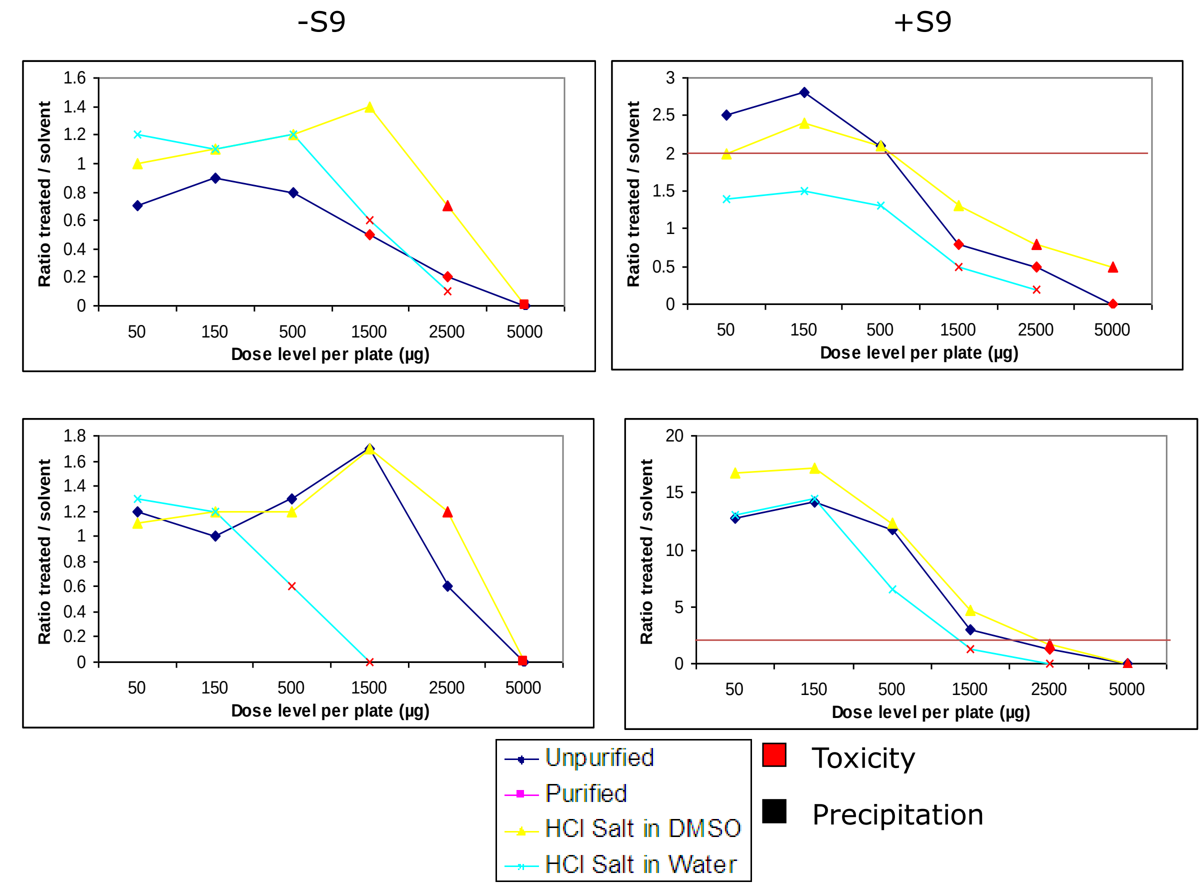


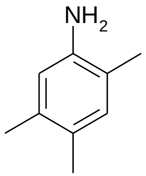
Ames Results
4-phenoxyaniline



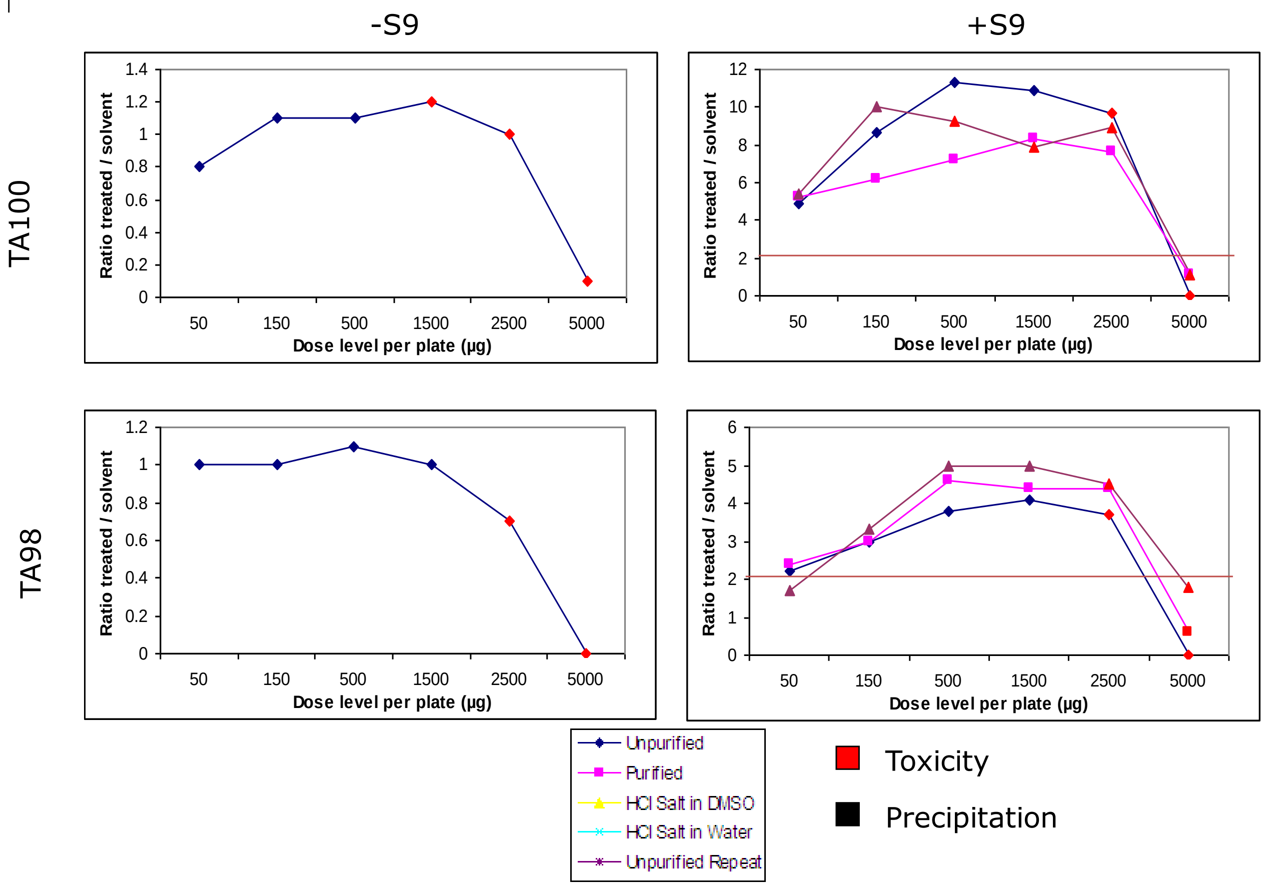
Ames Results
2-amino-5-phenylpyridine

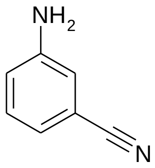






Ames Results
2,4,5-trimethylaniline

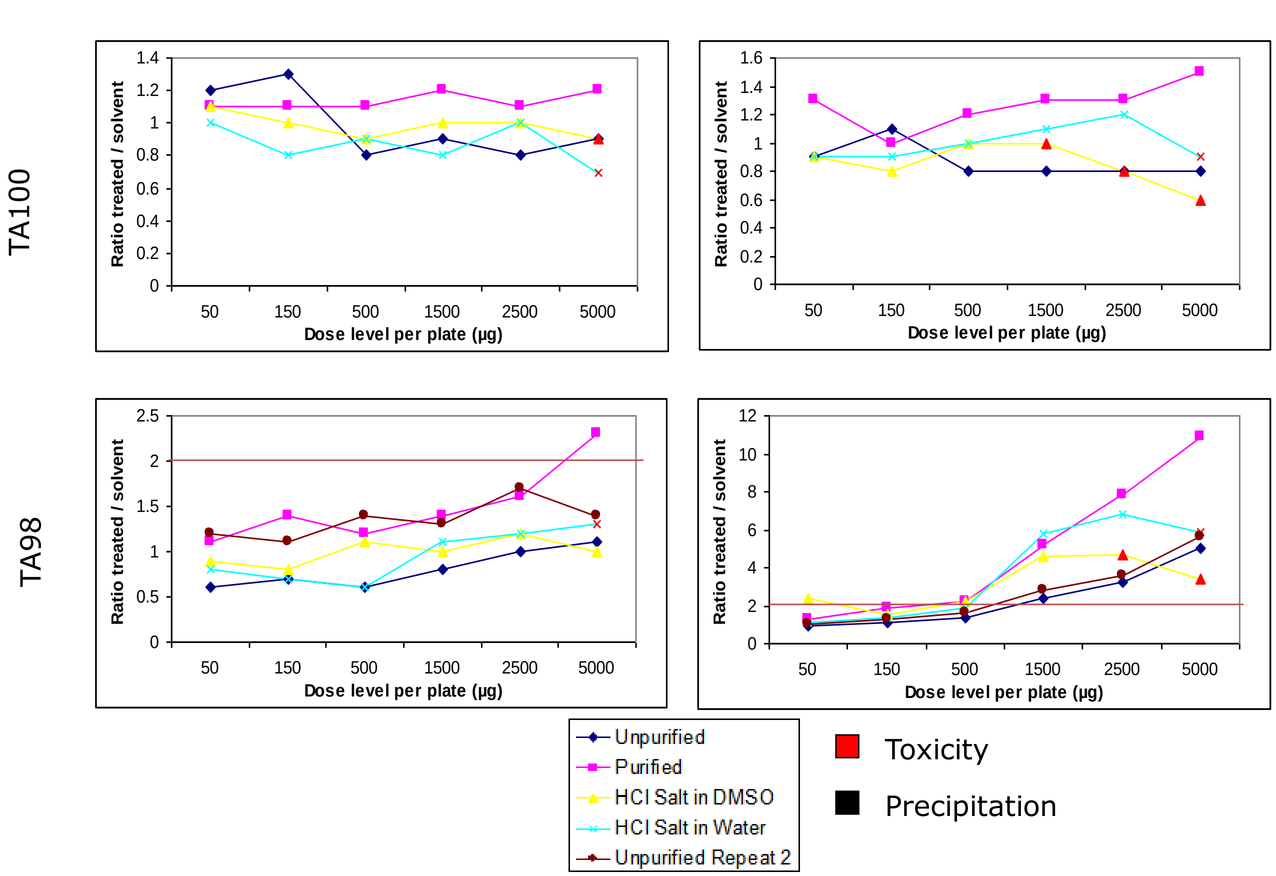




Ames Results
3-Aminobenzonitrile

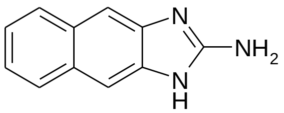
-S9

+S9

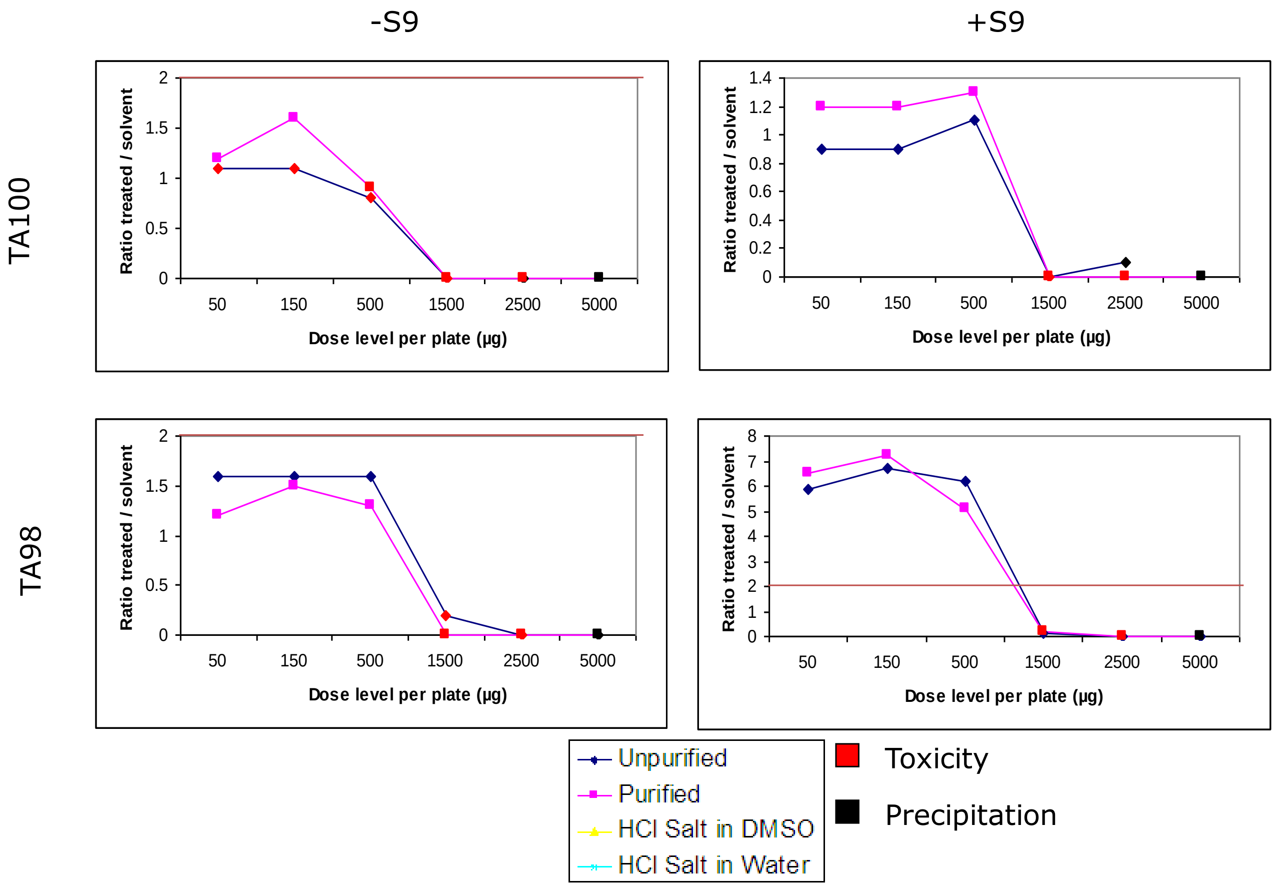


🟥 Toxicity

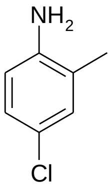
⬛ Precipitation

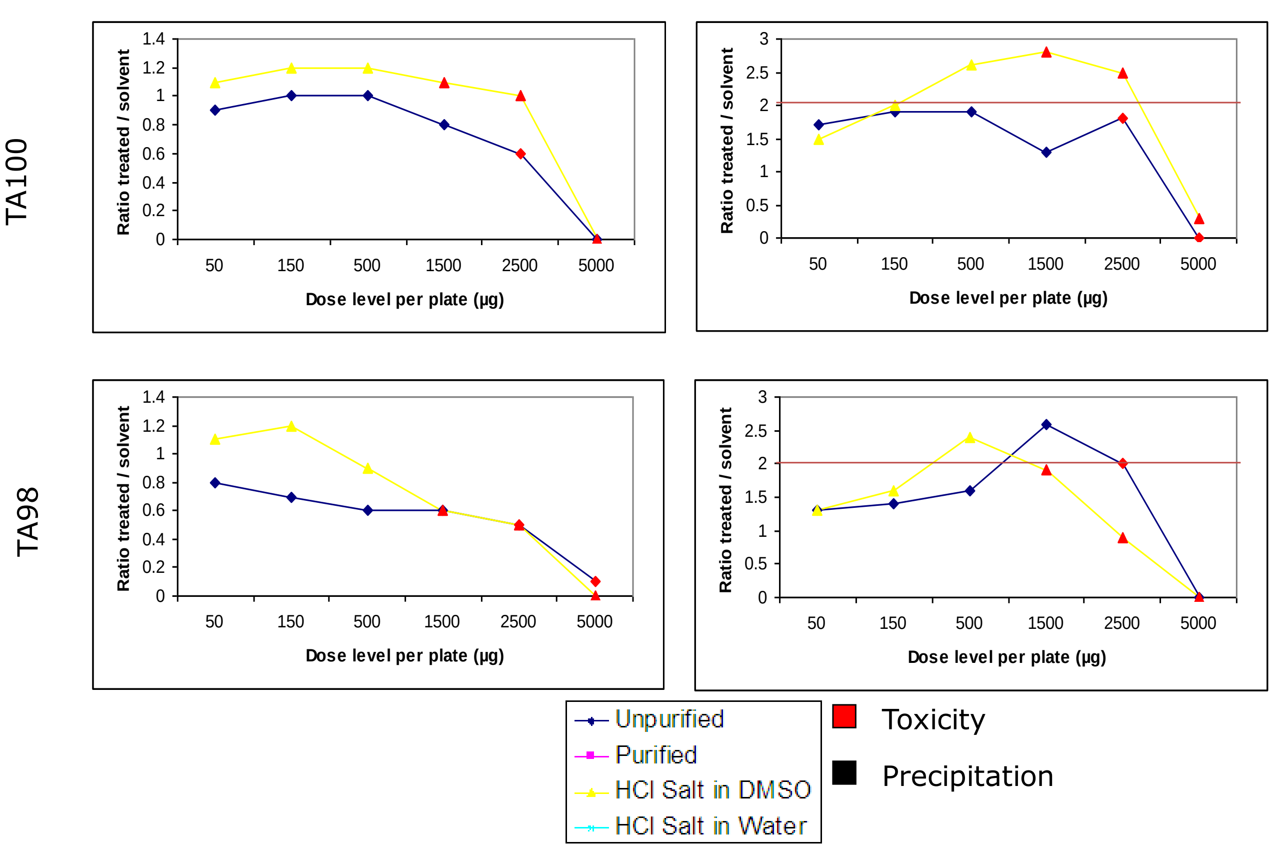


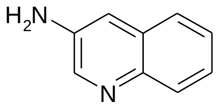
Ames Results
2-aminonaphtho(2,3-d)imidazole



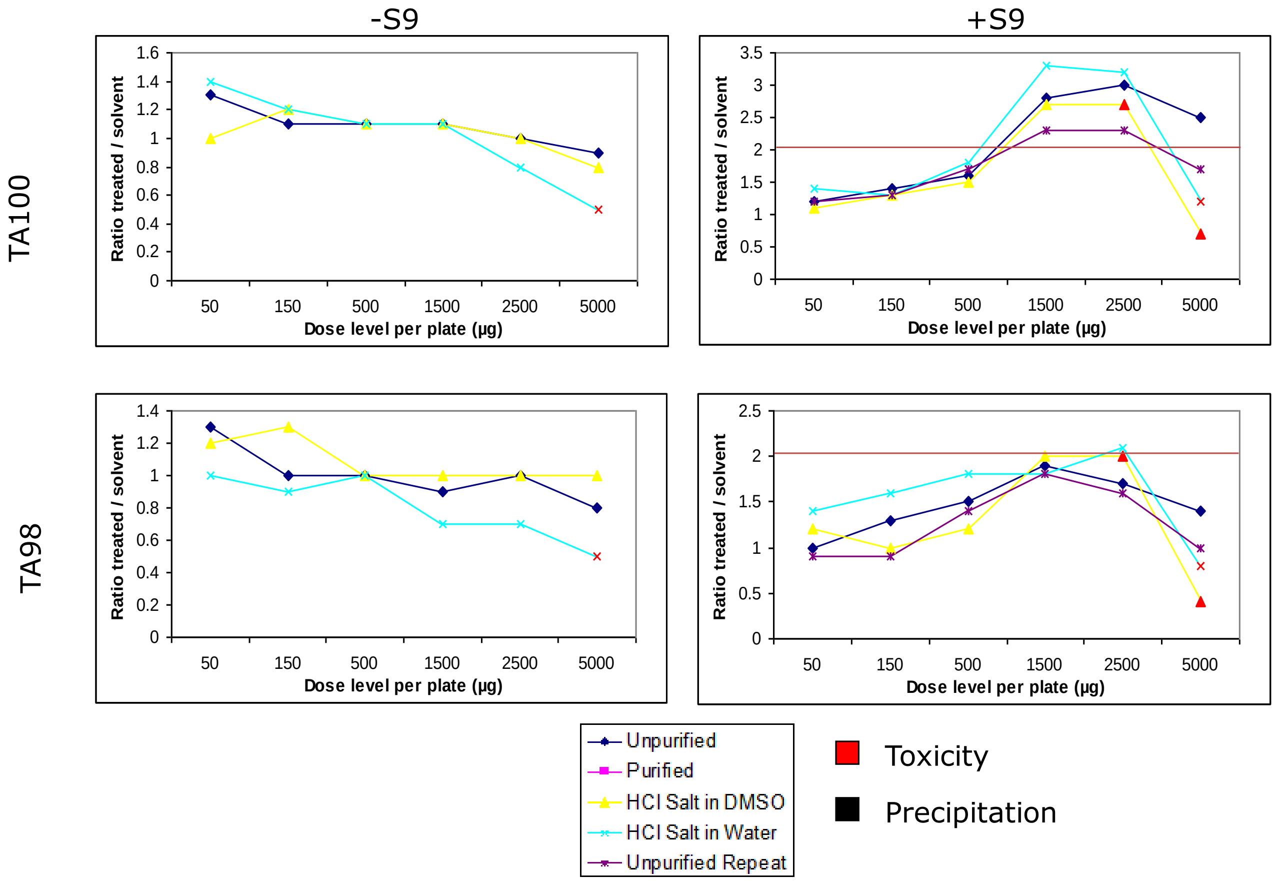
Ames
4-Chloro-2-methylaniline



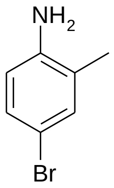


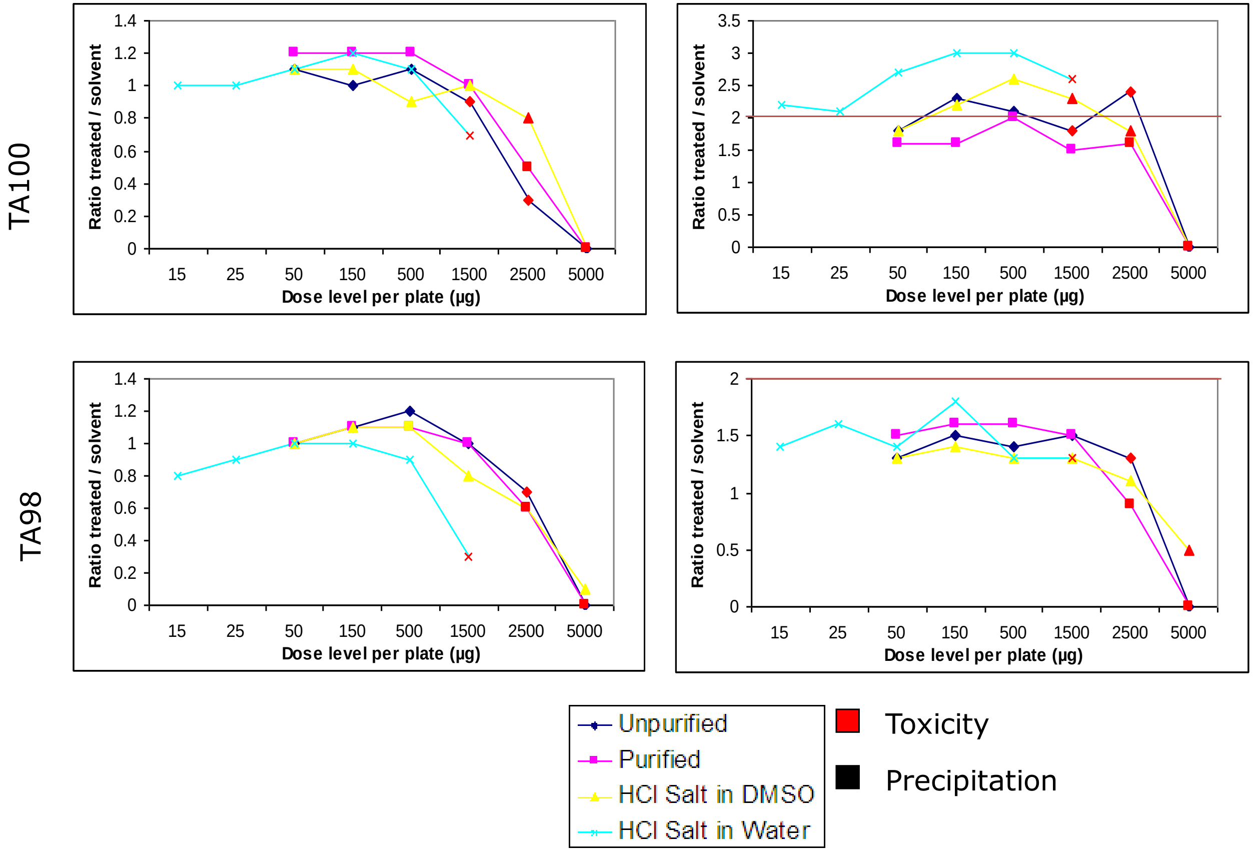


Ames Results
3-aminoquinoline

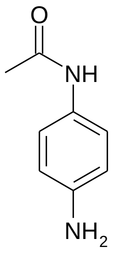


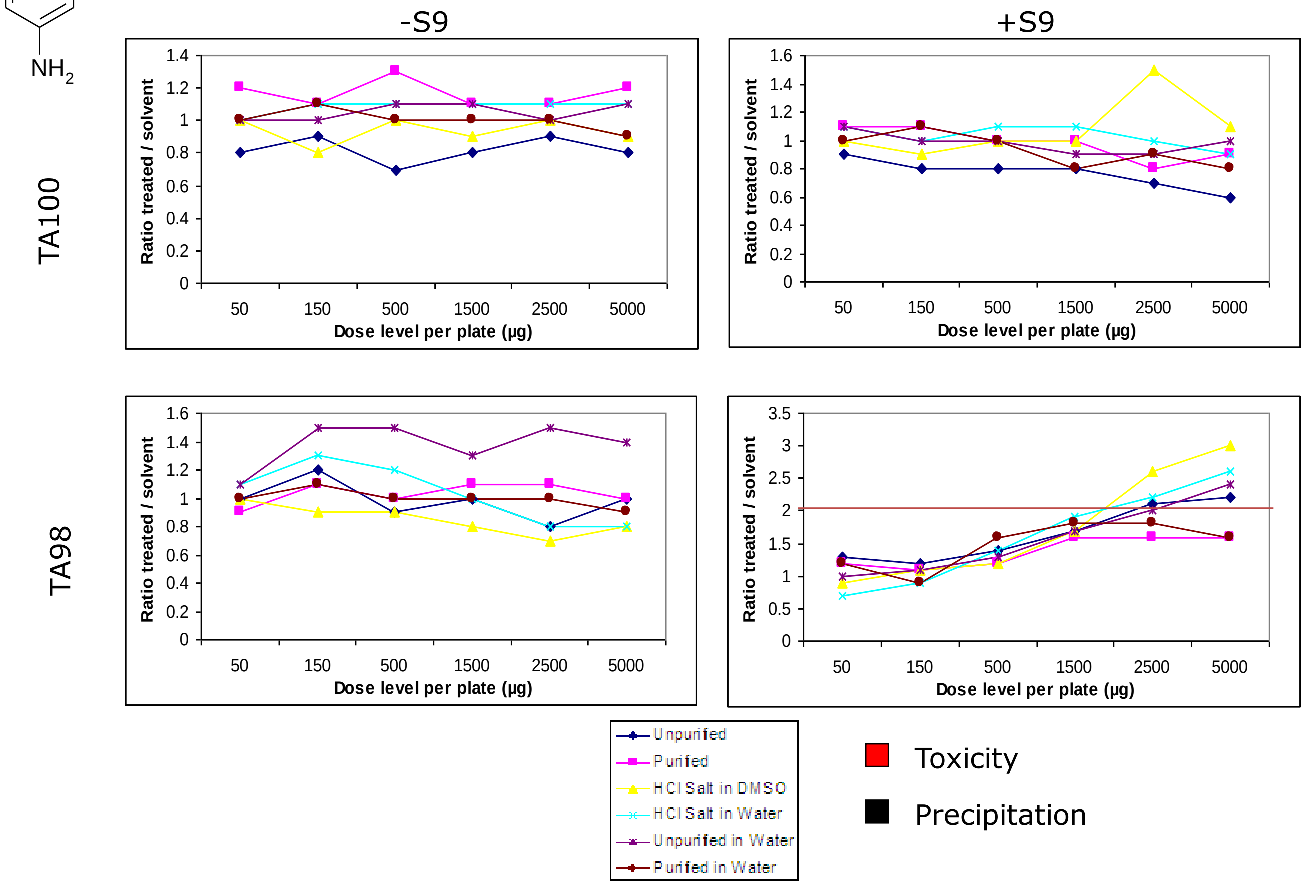
Ames Results
4-bromo-2-methylaniline



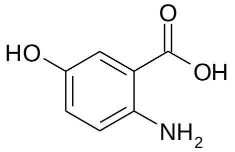


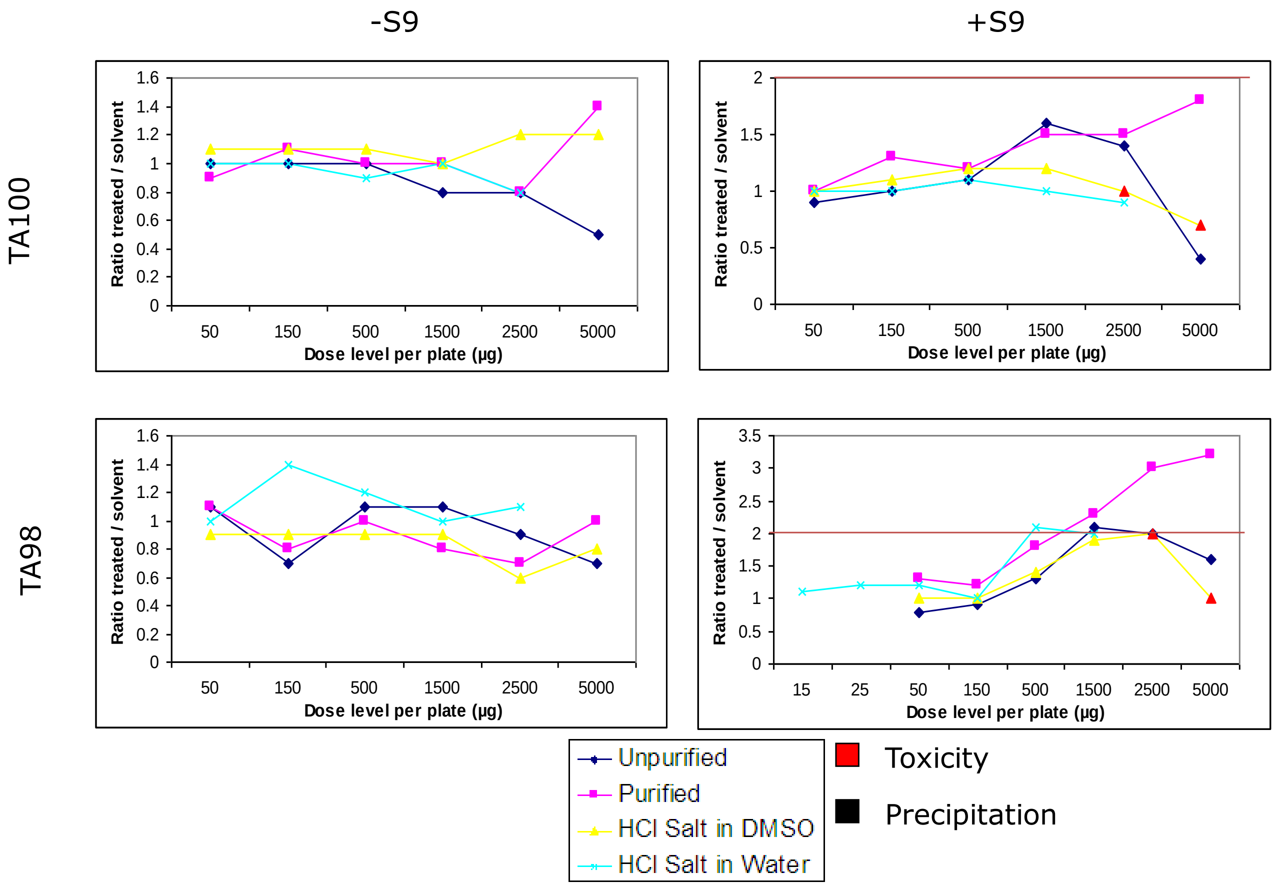
Ames Results
4-Aminoacetanilide

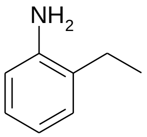




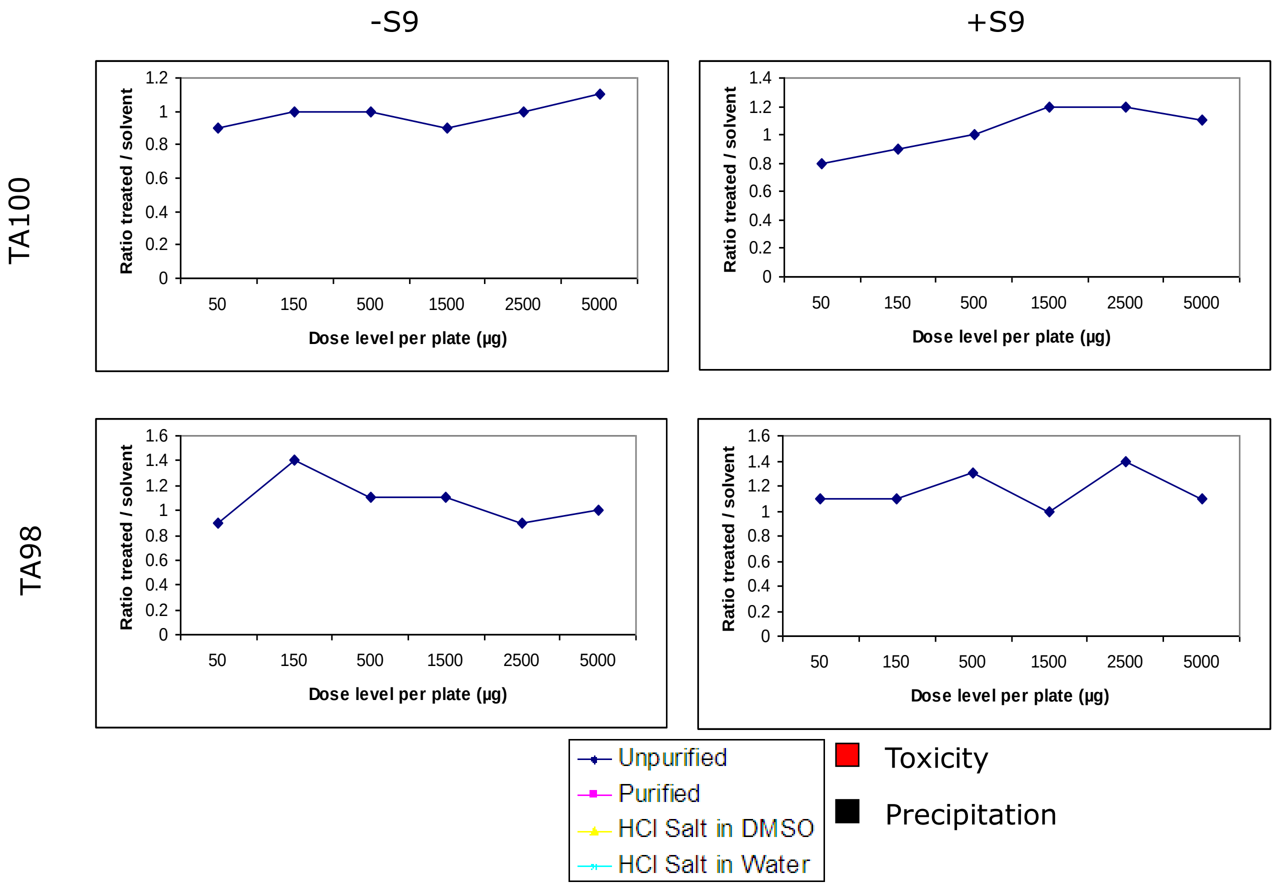
Ames Results
2-amino-5-hydroxybenzoic acid







Ames Results
2-ethylaniline



-S9
+S9
TA100
TA98
Ratio treated / solvent
Dose level per plate (µg)
Unpurified
Purified
HCl Salt in DMSO
HCl Salt in Water
🟥 Toxicity
■ Precipitation