UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Chemistry

The Connected Lab: Digital Synergies from Data to Models

by

Nicola Knight

Thesis for the degree of Doctor of Philosophy

June 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Chemistry

Thesis for the degree of Doctor of Philosophy

The Connected Lab: Digital Synergies from Data to Models

by

Nicola Knight

Many areas of research centre around data; how it is collected, stored, analysed and connected with existing data and ideas. In the changing technological landscape and infrastructure these data interactions are becoming increasingly digital. This thesis considers aspects of the full scope of the research environment from experimental creation and extraction of data through the human and technological interaction with the lab, to the analysis of the resulting data. The work described begins to make clear the synergies enabled by the digitalization of the research pathway.

In the first example, literature sources were carefully analysed and uncertainties exposed and explored to allow creation of high quality anion transporter datasets from which QSAR analysis could be carried out to obtain a predictive model. Within these datasets the effect of compound classification and substituent changes were investigated. In parallel work a closer collaboration between modelling and experimental groups took this area further. While no model was developed covering the whole dataset, transformations in the position of the optimum log P and peak activity were discovered dependent on substituent.

In the second example, interactions with the laboratory were investigated through two different aspects. Firstly though remote experiments that were created for undergraduate teaching, which provided a teaching resource for the physical chemistry practical course. The second developed novel interaction methods for application in the lab environment through the use of smart-assistants and the creation of Talk2Lab. This work sets the scene for a framework that could bring the 21st century technology into the research lab to create the connected lab of the future.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Nicola Knight, declare that the thesis entitled *The Connected Lab: Digital Synergies from Data to Models* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:
  Knight, N. J. et al. QSAR analysis of substituent effects on tambjamine anion transporters. Chem. Sci. 7, 16001608 (2016).

Signed:.....

Date:.....

Acknowledgements

This work and thesis would never have been completed without a number of people who have provided their help and support throughout the entire process. I would like to thank a few people specifically:

My sister Pippa, parents Chris and Linda, and Neil and Thalia for keeping me well-fed and helping to proof-read, my boyfriend Philip for all his support throughout my entire PhD. My friends, and members of SUBLDS, particularly Aneesa, James, Fergus & Kirsty for always keeping my spirits up but also everyone who has offered me their words of encouragement.

Thank you to all my colleagues for their help when completing my research, particularly to the Gale, Quesada and Frey/Brocklesby groups for providing access to their data and labs which were central to my work.

Finally, a particular thank you to my Supervisor, Jeremy Frey, for providing me with the opportunity to do this PhD and for his unending encouragement and support throughout.

In memory of my grandfathers, Phil and Geoff.

List of Abbreviations

AICc Corrected Akaikes Information Criterion

AVS Alexa Voice Services

BIC Bayesian Information Criterion

BLL Beer Lambert Law

CAS Chemical Abstracts Service

CV Cross Validation

D3 Data Driven Documents

DFT Density functional theory

DOI Digital Object Identifier

ELN Electronic Lab Notebook

ESI Electronic Supplementary Information

HPLC High-performance Liquid Chromatography

HTTPS Secure HyperText Transfer Protocol

IoT Internet of Things

InChI International Chemical Identifier

IUPAC International Union of Pure and Applied Chemistry

LOO Leave-one-out

LMO Leave-many-out

LWT Last Will and Testament

MLR Multi linear regression

MQTT MQ telemetry transport

MW Molecular Weight

NLP Natural Language Processing

NLR Non linear regression

NLU Natural Language Understanding

PCA Principal components analysis

PLS Partial least squares

POPC 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine - see glossary

PRESS Predicted residual error sum of squares

QoS Quality of Service

QSAR Quantitative structure-activity relationship

RDF Resource Description Framework

RMSE Root-mean-square error

RSS Residual sum of squares

RT Retention Time

SS Sum of squares

TSS Total sum of squares

VPN Virtual Private Network

WHIM Weighted Holistic Invariant Molecular (descriptors)

Glossary of Terms

Terms

Anion Transporter A Chemical entity which aids the transport of anions in biological systems

Beer Lambert Law (BLL) Law defining the relationship between absorbance and concentration of a solution, see Equation 6.1 on pg.137

CAS Number Unique numerical identifier assigned by the chemical abstracts service (CAS) to every chemical in their registry

Cytotoxic Toxic to living cells

DOI Digital Object Identifer - unique alphanumeric string which provides a persistent identifier for objects

Descriptor Numerical representation of chemical information

EC_{50} Half maximal concentration - concentration of a 'drug' which induces half of the maximum response effect, see Section 2.1.3 on pg.24

GRID Computational procedure for determining favourable binding sites on a molecule [1]

Hill Number Measure of co-operativity in the binding process

InChI IUPAC International Chemical Identifier - textual identifier for chemical substances, a standard format for encoding molecular information

Ionophore A chemical species that can bind ions

Lipophilicity The ability of a compound to dissolve in non-polar solvents or lipids, measured through logP - octanol-water partition coefficient

Pharmacokinetics Study of drug absorption, distribution, metabolism, and excretion

POPC Type of phosphatidylcholine lipid, used to mimic cell membranes in experiments

Retention Time (RT) measure of the hydrophobic nature of a compound, measured by liquid chromatography

R^2 Statistical measure of regression fit, see Equation 1.2 on pg.7

SMILES Line notation for describing chemical structures

Tambjamines Group of natural products investigated for anion transport properties

Vesicle A small structure comprised of fluid enclosed by a lipid bilayer

Computer Software

Chemical Drawing and Descriptor generation software

ACD/I-Lab see ref.2

ALOGPS2.1 see ref.3

ChemDraw see ref.4

Chemicalize.org see ref.5

Chemspider see ref.6

Daylight see ref.7

DRAGON see ref.8

E-DRAGON 1.0 see ref.9

MarvinSketch see ref.10

Open Babel see ref.11

Spartan see ref.12

Torchlite see ref.13

Molecular Modelling and Computational software

AMBER see ref.14

Gaussian see ref.15

MATLAB see ref.16

Javascript software and libraries

D3.js see ref.17

Node-RED see ref.18

Node.js see ref.19

Statistical software

JMP see ref.20 and 21

RSudio & R see ref.22

Notetaking software

Blog³ see ref.23

LabTrove see ref.24

OneNote see ref.25

RNotebooks see ref.26

Other software

Classyfire Web based classification software - see ref.27

SciFinder Database of chemical and bibliographic information - see ref.28

Slack Team communication & collaboration tools - see ref.29

File Formats

CSV Comma separated values

HTML Hyper Text Markup Language

JSON Javascript Object Notation

PDF Portable Document Format

SVG Scalable vector graphics

XML Extensible Markup Language

Introduction

Data is the thread that weaves its way through the research; from the experiments that create it, through the analyses which build knowledge from it, to the dissemination of this data and knowledge. The age of computing has fundamentally changed the way in which we carry out research and interact with data. No field has been left unaltered by the advances in computing technology.

The age of computers has ushered in entirely new areas of scientific research ranging from in silico drug testing to bioinformatics to big data analytics. But computing also fundamentally transformed many existing areas of research including areas such as chemical modelling and the use of chemoinformatics.

The use of chemical information encompasses a wide range of topics within itself, including data analysis using traditional techniques and complex computer modelling. In the past modelling was a smaller research area due to the lack of chemical data and computing power available. However, now there is a significantly larger amount of data produced from practical experiments, theoretical calculations and associated metadata. Manipulation of data to produce information has become much more widespread, chemical modelling is now present in areas related to chemical safety, materials, human health and ecological systems.

This thesis considers aspects of the full scope of the research environment from experimental creation and computational extraction of data through the human and technological interaction with the lab, to the analysis of the resulting data. The work described begins to make clear the synergies enabled by the digitalization of the research pathway.

QSAR

Quantitative Structure Activity Relationship (QSAR) is an important aspect of research in many areas of modelling. In particular, the area of drug discovery makes use of QSAR methods in researching new drugs. QSAR methods can be used to screen compounds for their potential drug activity prior to synthesising the compounds. However, there are also a wide range of applications beyond this.

In this thesis a number of literature sources were examined to provide the data for use in QSAR analysis of anion transport. On top of the analysis of the data a number of areas were investigated in which data curation and visualisation could be improved to aid future collaboration and investigation.

Connected Science

In recent years, connected technology has become an integral part of everyday life, but has yet to make a significant impact on interaction within the lab. Technologies can be incorporated to provide better access to data and more efficient working environments.

This work investigated the integration of commodity technologies into the lab environment for both teaching and research. The incorporation of these technologies aided the creation of a framework for the connected lab of the future.

Structure of this thesis

The content of this thesis is separated in two parts. Part I presents research on QSAR analysis of Anion Transporters. The topics in this part cover QSAR analysis of anion transporters from the Gale group (Chapter 2) and tambjamine compounds (Chapter 3) along with discussion surrounding data handling and the development of data visualisation for QSAR (Chapter 4). Part II presents research for connected science. This begins with the creation of remote experiments (Chapter 6) and is followed by the development of a smart lab interface (Chapter 7) focusing on voice interaction with the lab. A glossary can be found on page xxvii containing an explanation of some terminology used in this thesis. An appendix is also included which contains some additional plots and datatables. Further documents and files accompany this thesis in the electronic supplementary information (ESI) available through ePrints. [30][1]

[1] http://dx.doi.org/10.5258/SOTON/D0563

Part I

QSAR

Chapter 1

QSAR Theory

1.1 Introduction

The concept behind Quantitative Structure Activity Relationship (QSAR) is the construction of a form of model which relates an observed activity or property of a molecule to its molecular structure. [31] Using these QSAR models the activity of new and untested molecules can be predicted, often without need for synthesis. The basic assumption for these model hypotheses is that structurally similar molecules exhibit similar activities, but with the complexity of chemical interaction this is not always the case. [32, 33]

The foundation of modern QSAR is associated by many to the publication from Hansch et al in 1962 [34] working on the Hammett substituent effects, followed by investigation of the effects of lipophilicity on biological potency [35] and using octanol-water partition coefficients [36] as the measure of lipophilicity. However, around this time research using QSAR style analysis was also being carried out by other researchers in many areas such as; partition coefficients and their effect on drug absorption [37], the relationship between ether-water distribution and antimalarial potency [38] and the permeability of plant protoplasts to non-electrolytes. [39]

QSAR modelling methods have expanded significantly from their original origins, although they are still most widely applied by pharmaceutical industry for drug discovery they are also widely used in many different areas such as cosmetics, toxicology, agriculture and environmental research. [40–42] Advances in computing software, increased computing power and reduced cost, coupled with combinatorial synthesis and high throughput screening have meant that it is possible for huge numbers of compounds to be synthesised more easily and screened for drug testing. [43] However, this is still small compared to the scale of possible compounds that exist within chemical space. [44] QSAR methods can be used to model processes and molecules and predict the activity

or 'drug-ability' of new molecules prior to synthesis, eliminating the time-consuming and costly stage of synthesising a large number of compounds which turn out to be unsuitable.

The application of QSAR models extends beyond the use in prediction of activities; other areas of benefit include gaining insight into mechanisms, identification of key structural characteristics in compounds, identification of chemical activity deviations and the generation of hypotheses for further research. [45]

The standard practice of QSAR model development can usually be split into 3 sections: preparation of data, analysis of data and model validation. [46] The first stage involves obtaining the molecular dataset for use in the QSAR study, including calculation of necessary descriptors and the selection of the QSAR method to be applied. The second stage of the process applies the QSAR method for model development. This could involve a wide variety of different statistical approaches; however, the most common involve the use of linear (or multi-linear) regression. The last stage of the model development is the validation of the developed model, where the reliability of the model is tested for its particular purpose.

1.2 QSAR Methods

QSAR models are provided by mathematical equations which quantify a relationship between the activities of compounds and numerical representations of their structural and physiochemical information provided by descriptors. [47, 48] In its simplest form a QSAR model can be shown as:

$$\text{Biological activity} = \text{F(compound properties)}$$

The conventional methods for employing QSAR on numerical data use regression techniques. These include (multi)linear regression and non-linear regression, depending on the type of relationship. Regressions are also often used in conjunction with dimensionality reduction techniques such as PCA or feature selection. [45, 49]

Many methods are also available when the response variable has graded values or can be split into different categories. These include: linear discriminant analysis, cluster analysis and logistic regression. [50] However, the methods outlined below focus on methods which can be applied to continuous numerical data.

1.2.1 Linear Regression

The most widely used method in QSAR models are linear regressions due to their simplicity and easy interpretability, with a common method being the multi-linear regression, which takes the following form: [51]

$$\hat{Y} = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \cdots + \beta_n\chi_n \tag{1.1}$$

Where $\hat{Y}$ is the compound property being modelled (dependent variable), β_0 is the model constant, $\chi_1 \ldots \chi_n$ are molecular descriptors (independent variables) with their corresponding coefficients $\beta_1 \ldots \beta_n$ these coefficients are obtained through the use of estimators like least-squares method which minimizes the residual sum of squares (RSS) [52].

The quality of the model fit can be measured in a number of ways, one of which is the calculation of the Squared Correlation coefficient, R^2 (Eq.1.2).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1.2}$$

The components of Eq.1.2 are calculated as shown in Eq.1.3 and Eq.1.4 where:

$y_{calc,i}$ are the property values predicted by the equation using the relevant independent variables

y_i are the corresponding experimental observations

$\bar{y}$ is the mean of the observed values

Residual sum of squares:

$$RSS = \sum_{i=1}^{N}(y_i - y_{calc,i})^2 \tag{1.3}$$

Total sum of squares:

$$TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{1.4}$$

R^2 has a value between zero and 1, and describes the percentage of the variation in the data which can be explained by the regression equation [31]. An R^2 value of zero indicates that none of the variation in the data is explained by the regression equation, a value of 1 indicates complete explanation. This is a useful statistical tool; however, taken in isolation it can be misleading description of the actual fit of the data. [31, 53]

R^2 values may indicate good explanation of the fit of modelled data; however, they do not give an indication of how well the model will predict values for new datapoints. This is a crucial aspect of QSAR models as they need to be able to reliably predict properties for unmeasured compounds.

1.2.2 PLS/ PCA

Partial least squares (PLS) and Principal component analysis (PCA) are two methods of carrying out dimensionality reduction on data for regression. [54, 55] The aim of dimensionality reduction is to reduce the number of descriptors in a dataset whilst maintaining an explanation of the maximum variation. [56] This can be useful when handling large datasets with a large number of possible variables.

Both PLS and PCA reduce the number of dimensions by transforming the variables to produce new linear combinations of the original variables. In PCA the first principal component (PC1) is the linear combination of the predictor variables which yields the highest variance. Once this is constructed the next principal component (PC2, PC3, etc) is selected as a linear combination of the variables which accounts for the maximum variance which is not already explained. Each PC created is uncorrelated, or orthogonal, to the previously created principal components.

PLS is similar to PCA; however, it also uses the response variable when identifying the linear combinations of variables, referred to as latent variables. PLS attempts to explain the variance in both the response and predictor variables in its creation of new linear combinations. As in PCA the latent variables produced by PLS are orthogonal.

PLS is considered a supervised method as it uses correlation with the response variable when determining the new variables, whereas PCA only accounts for the variation in the independent variables so is referred to as an unsupervised method. [50]

These are not feature selection methods. Although they reduce the number of variables, each variable is a combination of all the original features. It is often found that much of the variance in the dataset can be explained by a small number of PCs or latent variables. [31] However, these methods reduce the ability to interpret the model in relation to the descriptors used to model the response.

1.2.3 Non Linear Regression

Non linear regressions (NLR) are equations where the function is non-linear with respect to the unknown parameters. An example of a non-linear model is:

$$y = \theta_1 e^{\theta_2 x} + \epsilon \tag{1.5}$$

where the equation is not linear in the unknown parameters θ_1 and θ_2. [57]

These types of regression are often used where there is an empirical or theoretical relationship established between the predictor and the response variables, this theory frequently involves the solution to differential equations. Key areas of use for these regressions are in growth models and modelling rates of change, such as in reaction rates. [57]

1.3 Model Validation

The ultimate aim of QSAR modelling is the creation of statistically robust models which can be used to predict accurate values for activities of unseen molecules. As such the validation of the models is a crucial stage of QSAR as it tests the robustness and predictive ability of the model. Validation methods are based around a comparison of observed values against model predictions to check that models are reliable and not due to chance correlations. [58, 59]

There are a vast number of different methods that have been developed for model validation; however, these can largely be split into internal and external validation. [60] Internal validation methods check the fit and predictive ability of the model using compound data that has been used in the model building process, also referred to as the training set of molecules. Examples of internal validation include: least squares fit (R^2), model fit statistics, leave-one-out cross validation (LOO CV), leave-many-out cross validation (LMO CV), bootstrapping and y-randomization. [48] External methods perform statistical tests using data that was not incorporated at all in the model building process, referred to as an external test set. [60]

1.3.1 Internal Validation

There are a number of statistics that can be calculated on a developed model alongside the previously mentioned calculation of the R^2 value. The tests in this section can all be carried out internally using the data of the training set that was used to develop the model.

1.3.1.1 Model fit statistics

Model fit statistics are used to determine the goodness-of-fit and robustness of the model with respect to the training set data.

Squared correlation coefficient (R^2) - measure of the model fit, see equation 1.2. This parameter increases as extra descriptors are added, therefore the R^2_{adj} statistic is often a better measure in models with many parameters.

Adjusted R^2 value (R^2_{adj}) - this is an adjusted version of R^2 which takes into account the number of variables included in the model, where n is the number of data points and k is the the number of variables in the model. These values also range from -1 to 1 and the closer the value is to 1/-1 the more variation can be explained by the regression equation.

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n-1)}{n-k-1} \right) \tag{1.6}$$

Root-mean-square error (RMSE) - quantifies the deviation of the errors between the predicted values and the observed values, where $y_{calc,i}$ is the predicted value for i, y_i is its observed value and n is the number of data points. [58] Lower values of RMSE indicate a smaller spread in the data and therefore a stronger fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_{calc,i})^2}{n}} \tag{1.7}$$

Although these statistics can be useful indicators of a poor quality fit and poor prediction, good values of these statistics are not sufficient indicators of model validity and should be accompanied by additional validation tests.

1.3.1.2 Leave-x-out Cross Validation

A popular validation criteria is leave-one-out cross-validated R^2 (LOO q^2). The training set is modified by the removal of a single compound, following this the model is rebuilt using the same descriptors and the remaining (n-1) molecules from the training set. The new model equation is then used to calculate the activity of the removed compound. [47]

The process is repeated until each compound has been left out in turn and the predicted values can be used to calculate internal validation parameters, and the cross-validated R^2 (q^2), calculated by the equation below.

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_{calc,i,-i} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \tag{1.8}$$

where $y_{calc,i,-i}$ is the predicted value for the ith compound, y_i its observed value, $\bar{y}$ is the mean of the observed values and n the number of compounds in the training set.

This is similar to the equation used for the calculation of R^2 (Eq 1.2); however, PRESS differs from RSS as the predicted values in PRESS ($y_{calc,i,-i}$) are for observations that were not used in the calculation of the model parameters. In the case of LOO CV the model is built n times, each time using n-1 observations to fit the model and using that model to predict the left-out observation.

An expansion of the cross-validation method that is considered stronger is leave-many-out cross-validation (LMO-CV) where the number of compounds that are left out at a single time is increased. LMO can be used as a method to counteract the slight overoptimism of the LOO-CV values. [60] The same modelling recalculation is carried out as with the leave-one-out cross validation, leading to the calculation of LMO-q^2. With both of these methods a threshold value of $q^2 = 0.5$ is considered the cutoff, below which a model is not considered predictive. [47]

1.3.1.3 Bootstrapping

A further validation method similar to the cross-validation methods is the bootstrap method. [50] This method carries out repeated analysis on random samples from the dataset, where after each sampling the objects are replaced in the dataset. This may result in some compounds being selected multiple times and others not being selected; however, these selections are smoothed out over many repeats. [48, 60]

With each sample statistics for the model fit are calculated and the bootstrap sampling is repeated many times to create a distribution of the statistics. These distributions can be used to determine the robustness of the statistics. [61] The sampling is usually repeated at least 100 times; however, with the continual increase in computing power the repetitions can often be increased to over 1000 with negligible impact on computation time.

1.3.1.4 Randomization tests

Testing the robustness of a model against chance correlations can be carried out through use of y-randomization. In this test the response variables are randomised across the dataset. With this new dataset modelling and validation procedures are repeated to produce statistics for the model fit.

If the resultant statistics generated in the random models are similar to those of the original model, showing a good fit or indicating high predictability, then it shows that the model is not robust and that the correlation in the model was obtained through chance only. [47, 48]

1.3.2 External validation

The methods of internal validation assess the robustness of a model using compounds that belong to the training set used to develop the model. When the goal of QSAR is to provide accurate prediction for new compounds it is highly desirable to validate the predictive ability against compounds that are 'unseen' by the model development. [47, 59, 62]

The external test set for this validation can either be created through subsetting the data prior to model building, or the generation of new compound data after the initial model build. Following the generation of a model using the training set alone, external validation statistics for the model are calculated on the test set. A common statistic is the calculation of the predictive squared correlation coefficient (q^2), also sometimes referred to as R^2_{pred}.

There are multiple different methods proposed for the calculation of q^2 value for external validation. [58,63,64] Two of these methods are given by equation 1.9 and equation 1.10, these equations are expanded from the calculation of the q^2 value for cross validation.

$$q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_{calc,i} - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{TR})} \tag{1.9}$$

$$q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_{calc,i} - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{EXT})} \tag{1.10}$$

where $y_{calc,i}$ is the predicted value for the ith compound in the external test set (calculated using a model built from the training set), y_i its observed value, $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ are the means of the observed values of the training set and external test set respectively

Equation 1.9 (q^2_{F1}) is the recommended equation in the OECD QSAR guidelines [65] which uses the predicted values for the test set compounds alongside the response mean for the training set. Equation 1.10 (q^2_{F2}) differs from the first equation by the use of the external test set mean in the denominator.

It is argued [63,64] that q^2_{F1} produces overestimated prediction values when compounds are on the boundary of the response domain, and underestimates when they are in the center of the domain, and is highly dependent on the selection and distribution of the test set.

q^2_{F2} was proposed [63] as an improved measure of predictive ability using the test set activity mean instead. q^2_{F2} removes some of the potential overestimation of q^2_{F1}; however, it is not necessarily an improvement on the equation as q^2_{F2} does not contain information about the reference model [58]. Additionally q^2_{F2} cannot be calculated if the test set only contains one object.

When employing a training/set split for external validation, care must be taken to ensure that the test set is a good representation of the range of the whole dataset. [47,60] Small datasets may not be suited to splitting in this fashion as this would likely produce a test set that is too small and as such give poor, and random, estimates of the predictive power of the model. For small datasets <50 compounds it can be beneficial to use internal validation methods instead such as bootstrapping or LMO-CV. [58, 66, 67]

1.3.3 Comments on validation

In the literature there are a significant number of different formulas available for validation methods and calculations of model statistics, with different notation and derivations. It is often difficult to distinguish between the different equations as the notation and terminology are often used interchangeably. Some of these equations are the same, but not always. It is important to try and identify the methods used to calculate statistics so that the model statistics are unambiguous and models can be fully reproduced. [65]

The use of internal or external validation is an area of debate with QSAR research. [59, 60, 66, 67] Some researchers insist that models can only properly be validated through use of external test sets; however, others argue that in smaller datasets it is wasteful to exclude data from the modelling process and validation can be sufficiently carried out through the application of internal cross-validation methods.

Many people consider a high q^2 value to be proof that the model is highly predictive; however, many researchers argue this is not really the case. Although a low q^2 value using cross validation in the training set is usually an indicator of a low predictive ability, researchers argue that a high q^2 value does not necessarily imply high predicative ability and the models must still be tested against external test sets. [59] When possible it seems that external validation should always be carried out in addition to internal validation, to provide a more rigorous assessment of a models predictive ability. However, in some cases this may not be possible due to the size of the dataset. [62]

Within the selection of a test set and the general application of a model it is important to ensure that the compounds selected fall within the applicability domain of the model. This domain is a theoretical region of chemical space which is defined by the nature of the chemicals in the training set. It is unfeasible to reliably predict the whole universe of compounds using a single QSAR model [47], model predictions for any compounds

outside of the domain are extrapolations of the model and cannot be considered as reliable. [60]

1.4 Molecular Descriptors

Molecular descriptors are a crucial aspect of QSAR modeling as they provide the independent variables used in the creation of the statistical models. Before a QSAR model can be developed, molecular descriptors must be generated and selected for use as parameters within the model.

> "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."
>
> — R. Todeschini and V. Consonni [68]

Molecular descriptors are numeric values that characterise the properties of molecules. A large number of descriptors have been developed, which vary in the complexity of information they encode and the time required to compute them. [31] The generation of descriptors involves the application of theory from many different areas such as: algebra, computational chemistry, quantum-chemistry, information theory, organic chemistry, physical chemistry and graph theory.

Examples of descriptors range from the simple counts, such as number of atoms or molecular weight to highly complex quantum descriptors based upon the surfaces and interactions of the molecules. Molecular descriptors can be classified into many different types based upon what they are derived from. [68] A major method of describing them is:

- 0D - atom list (e.g. bonds counts, mol weight, atom counts)
- 1D - substructure list (fragment counts, fingerprints)
- 2D - molecular graph - topological descriptors (topostructural, topochemical)
- 3D - molecular geometry derived (surface, shape, volume, WHIM descriptors)
- 4D - 3D co-ordinates + conformations (GRID based techniques)

A good theoretical descriptor must have; invariance to atom labelling, invariance to roto-translation and an unambiguous computational method. They should also have; a structural interpretation, good correlation with at least one property, no inclusion

of experimental properties in its definition and not be restricted to a small set of molecules. [69]

The descriptors can be determined from experimental measurements or calculated from the application of theory. In general experimentally determined descriptors are not favoured as they require the synthesis of the molecule and can be time-consuming and expensive to obtain, especially when working with large datasets. Some calculated descriptors may have an experimental equivalent e.g. partition coefficients; however, others are purely computational constructs e.g. fingerprints. [70]

Molecular descriptors can be calculated from a wide variety of programs. These range from drawing/visualisation programs which calculate simple descriptors such as molecular weight to molecular simulation programs calculating complex quantum descriptors and force fields. Due to the extremely large number of descriptors that can now be calculated, in excess of 3000 descriptors are listed in Todeschini and Consonni's Handbook [71], care must be taken when selecting descriptors for use in models. In particular, the variance of descriptors should be considered, along with the correlation between different descriptors.

Although there has been a significant increase in the number of descriptors available it is important not to use too many descriptors in the model to try and achieve the 'perfect' fit. This does not produce a good model for prediction, simply a model that is over-fitted to the dataset provided. [67] Additionally descriptors with little to no variance across the dataset should be removed, and multiple descriptors with high correlation should not be modelled together. [47]

1.4.1 Molecular Representation

Within all aspects of chemistry it is important to be able to depict and refer to chemicals in a way that allows molecules to be distinguished and, most importantly for QSAR, allows chemical descriptors to be calculated. The following section outlines some of the methods available for representation of molecules. For the computational calculation of descriptors these representations must be machine readable and in an ideal representation each structure only has one 'code' and each code only converts to a single structure. [72]

Nomenclature
Compounds often have multiple names which include common names, drug names and systematic names. The most widely accepted format for naming compounds is the IUPAC naming system. [73]. Compound names are not machine readable and for complex molecules the conversion of names to structures often generates errors.

Formulas

Molecular formulas are those which only give an elemental composition, whereas structural formulas show which atoms are bonded together. Molecular formulas are not unique and a single formula can refer to many different molecules.

Line Notations

Line Notations are linear representations of a structure; encoding the connectivity of a molecule in a line of text, but not including 2D or 3D co-ordinates. They are widely used as they are human readable (to a certain extent) and easy to input into software. [74] The most popular forms are:

SMILES Simplified Molecular Input Line Entry System [75] - this is the most popular line notation, but there are many methods to generate SMILES strings, so they are not necessarily unique.

InChI IUPAC International Chemical Identifier [76] - This is designed to provide a unique string for depicting a chemical substance, it has a layered structure where different layers can handle molecular connectivity, charge, stereochemistry etc.

Connection Tables

A listing of atoms and bonds in a tabular form which can easily be interpreted by machine. These can be used for structure and substructure searching, through application of graph theory. [72]

MDL Molfile (.mol) This connection table is the most commonly used format for storing connection data. Multiple structures can also be depicted in a structure-data file (.sdf)

Proprietary formats Many proprietary formats exist for encoding structural information; however, the conversion and interoperability of these formats is often limited as they cannot always be read by other programs.

1.4.2 Common Descriptors

Below are some of the descriptors that were encountered in the QSAR work. More detailed descriptions of many descriptors and their origins can be found in Todeschini's Handbook [71]

Lipophilicity descriptors

Lipophilicity is a key concept in many areas of drug prediction and many descriptors exist that characterise it. Log P is a measure of the partition coefficient, a measure of a compounds solubility between two solvents; usually octanol/water.

ALOGP Ghose-Crippen octanol-water partition coeff. (DRAGON)

ALOGPs Calculated octanol/water log P - based off PHYSPROP database [77] (ALOGPS2.1)

CLogP Calculated octanol/water log P - fragment method (Daylight)

MLOGP Moriguchi octanol-water partition coefficient (DRAGON)

LogD(pH7.2)-sp blood Distribution coefficient - pH dependant measure of the solubility between octanol and water. (ACDiLabs2.0)

Other descriptors

AMW average molecular weight

MW molecular weight

Mv mean atomic van der Waals volume

nCIC number of rings (cyclomatic number)

nH number of hydrogens

pKa acid dissociation constant

RT retention time

SPAN span radius of the molecule

TSA total surface area

$\mathbf{V_{s,max}}$ electrostatic potential surface maxima (calculated via DFT)

Chapter 2

Modelling Anion Transport in Vesicles

2.1 Background

The movement of anions within biological systems in an important area of research for the treatment of many serious diseases [78], but also for the potential development of anti-cancer agents. Development of new synthetic compounds which function as potent anion transporters is an area of great interest.

Within this chapter research was undertaken on the curation of anion transporter data and the creation of a dataset of anion transporter compounds covering a variety of compound types. Collation of all the data across a group of papers allows investigation into the behaviour of a wider spectrum of compounds. This dataset was then processed and QSAR analysis carried out, with the focus on creating a model for the anion transport ability of the compounds.

2.1.1 Anion Transport

Cellular membranes define the boundaries of cells, forming the shape of the cell and separating the intracellular contents from extracellular components. The cell membrane is selectively permeable to allow control of the flow of molecules into and out of a cell; composed of a lipid bilayer and embedded membrane proteins these membranes are permeable to nonpolar compounds but impermeable to most polar or charged entities.

The lipid bilayer is formed mainly by a mixture of glycerophospholipids, sphingolipids and sterols depending on the type of membrane. The bilayer is formed with the hydrophobic tail groups in the centre of the bilayer and the hydrophilic heads interacting with

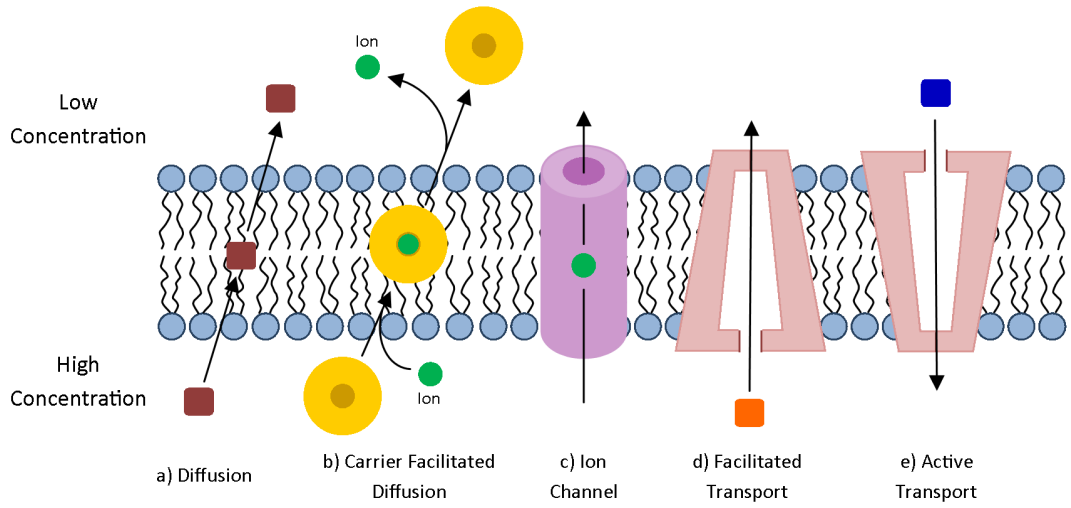


Figure 2.1: Mechanisms of anion transport through the membrane

the surrounding water. The hydrophobic core of this bilayer makes the membrane highly impermeable to polar solutes and ions. As the transport of these is crucial to many biological processes, the passage of the solutes across the membrane must be facilitated by transporters.

2.1.1.1 Transport Mechanisms

Transport across the membrane can occur in many different ways (see Figure 2.1); non-polar compounds can pass through the membrane via simple non-mediated diffusion; however, ions must be facilitated by an anion transporter either through passive or active transport. Passive transport is where the movement of the molecule is aided from an area of higher concentration to one of lower concentration, active transport is where the movement of a molecule occurs against the concentration gradient, from low to high concentration. Active transport requires an energy input and the process can either be coupled directly to ATP hydrolysis or coupled to the transport of a second solute down its concentrations gradient, which supplies energy to drive the other solute against its concentration gradient. [79]

Passive transport functions via a number of mechanisms, one mechanism is ionophore faciliated diffusion, ionophores increase the permeability of the membrane to ions. This can be in the form of a mobile carrier (Fig 2.1b) which binds the ion and the complex diffuses through the membrane, releasing the ion on the other side, or as a channel (Fig 2.1c) where the ionophore forms a membrane-spanning channel through which ions can diffuse. [80] Anions can also be transported via the relay mechanism, which uses modified phospholipids to pass the anion through the membrane.

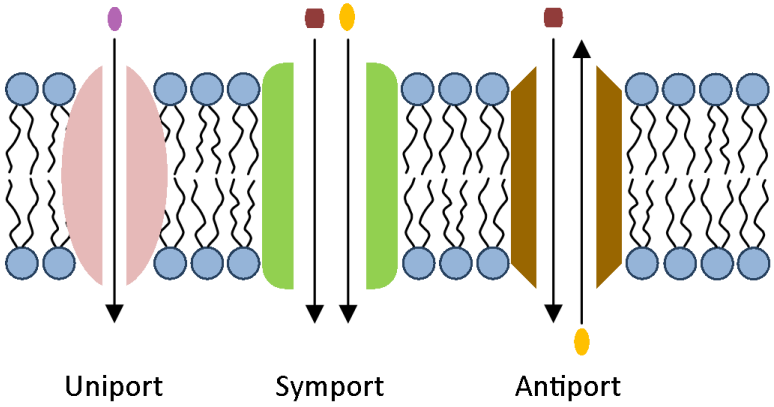


Figure 2.2: Types of action within membrane transport systems

The transport of anions through the membrane can happen through a variety of different methods shown in Figure 2.2. Either a single substrate can be transported through the membrane with a uniport action, or two different substrates can be transported through the membrane at the same time. The two substrates can be transported in the same direction through a symport action or in opposite directions through the antiport action.

2.1.1.2 Anion transport in nature

Since anions play a vital role throughout biological systems, their movement across the cell membrane is fundamental to many biological processes. They maintain osmotic balance, regulating cell pH and chemical potentials across the bilayer. These are key to driving metabolic processes and cell signalling, an example of which is triggering the death of cells through apoptosis. [81, 82]

The transport of these anions in nature is mostly controlled by specialised proteins that span the cell membrane and transport the anions, both through passive and active transport mechanisms. Chloride channels aid transport through either a uniport or cation coupled transport, many of these channels are gated and open in response to a variety of stimuli including: changes in membrane potential, cell volume and intracellular calcium concentration. [83] Defects in these proteins and channels can result in a number of diseases, collectively referred to as "channelopathies" [84].

In addition to ion channels there are small-molecule anion transporters; however, there are only a few natural products which are able to facilitate the transmembrane transport of anions. [85] The most widely studied family are prodigiosins [86], but there are other examples of spingolipids [87] and monoacylglycerols [88].

Development of synthetic molecules which can mimic the action of these anion transporting proteins is a wide area of current research in supramolecular chemistry. These

synthetic molecules have potential to be used as treatment for many of the diseases caused by faulty anion transport in addition to the possibility of anti-cancer applications through the triggering of apoptosis in cells.

2.1.2 Synthetic Anion Transporters

The challenge of creating synthetic therapeutics for anion transport has stimulated the creation of synthetic membrane spanning channels. [89] Although research into synthetic channels and relay-style systems is an area where much research has been carried out, the channel-like compounds are too large to be considered drug-like. [85] This has prompted interest in the development of small molecules which are capable of transmembrane anion transport.

Research into supra-molecular chemistry, principally lipid bilayer transport of anionic species and molecular recognition [90] has been the focus research area of Phil Gale's group within the Chemistry Department.[1] A number of their papers have focused specifically on developing small molecules for use as anion transporters, in particular they have examined chloride transport which has potential applications in treatment for "channelopathies" such as Bartter's syndrome, Cystic Fibrosis and inducing apoptosis in cancer cells. [85]

They have taken inspiration from natural anion transporters as well as drawing on work carried out in anion receptor chemistry, synthesising and investigating the efficiency of chloride ion transporters from a diverse chemical range. Papers published by the Gale group include studies on ureas, thioureas, squaramides and phenylureas along with myriad of other compound functional groups. [91–100] Figure 2.3 shows a selection of compounds that were contained in the papers.

The Gale group research mainly focused on the synthesis and characterization of anion transporter compounds through transport studies in vesicles and cell-based assays; however, some of their papers [91–93, 95] explored a small amount of QSAR analysis, including investigation of correlations between log P (partition co-efficient) and EC_{50} values. The QSAR carried out in most of their papers did not result in the production of a QSAR model, but simply commented on some trends in the data across a small subset of compounds. They looked at identifying trends within a small group of similar molecules, such as the effect of changing a substituent at a single position [91].

[1] formerly: Chemistry, University of Southampton, Southampton, SO17 1BJ, UK
currently: School of Chemistry, University of Sydney, NSW 2006 Australia
Email: philip.gale@sydney.edu

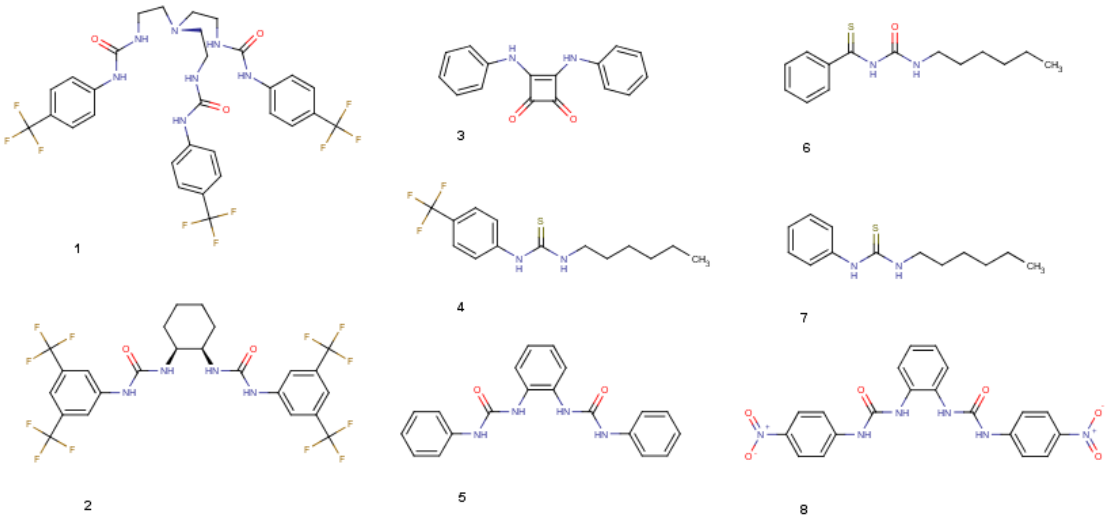


Figure 2.3: Selection of compounds contained within Gale group papers

2.1.2.1 Mechanisms of examined transporters

The compounds synthesised by the Gale group have focused on the creation of transporters which are mobile carriers and mediate carrier facilitated diffusion across a lipid bilayer. (Figure 2.1) These carriers must bind the anion and effectively shield it from the lipids whilst the complex passes through the membrane. [101]

These carriers are often designed with the antiport movement of anions in mind (Figure 2.2). This is where two different anions are transported through the membrane in opposite directions. In general the antiport mechanism allows a single binding site to be used for the transport of anions in both directions [101], as only a single anion is moved in each direction at a time. Many of the compounds synthesised by the Gale group (Chapter 2) and Quesada group (Chapter 3) will be protonated under the experimental conditions (which approximately mimicks physiological conditions at pH7.2) - Figure 3.3, this promotes anion binding and the complex can pass through the membrane as a neutral complex.

Two processes of primary interest in the research are the Cl^-/ NO_3^- antiport and Cl^-/HCO_3^- antiport processes. In the Cl^-/ NO_3^- antiport the chloride ions are bound and transported in one direction, with nitrate ions being bound and transported in the opposite direction. It is also possible for the unbound (free) transporter to cross the membrane; however, this may not be efficient if the molecule is in a protonated state. The transport experiment for this process is explained in Section 2.1.3.

The chloride could alternatively be transported by a symport mechanism (Figure 2.2), where the charge of Cl^- is balanced by Na^+ or H^+ being transported across the bilayer with the chloride ion. To examine this experiments were undertaken with other cations (K^+, Cs^+), which showed little difference in transport activity compared to Na^+ [91,102]

and experiments with different external anions (SO_4^{2-}), which showed little transport activity at all. [93, 96, 97] These observations suggest the predominant mechanism is chloride/nitrate antiport rather than either symport mechanism. However, development of novel transporters may produce compounds that also operate via the symport mechanism.

2.1.3 Experimental Transporter Studies

To investigate the potential potency of compounds as anion transporters a series of physical experiments were carried out. Transport experiments were carried out to determine the compounds' ability to transport anions across the lipid bilayer of a POPC vesicle and binding studies were carried out to determine anion binding affinities. These experiments are designed as extreme simplifications of a biological membrane due to the complexity of biological systems.

All of the syntheses and transport experiments carried out on the chloride ion transporters examined were performed by the same research group, following the same processes and methodology. The precise methodology used can be found in any of their papers. [91–100]; however, a summary of the Cl^-/ NO_3^- antiport experiment can be found below.

Anion Transport Study

The transport studies were designed to measure anion transport ability by monitoring the efflux of chloride ions from a vesicle over time. The vesicles were comprised of a POPC bilayer encapsulating a sodium chloride (NaCl) solution. The vesicles were suspended in a solution of sodium nitrate ($NaNO_3$) for the Cl^-/ NO_3^- antiport experiment.

This gave the initial state for the system; where all the chloride ions were inside the vesicle and none in the surrounding solution, and all the nitrate ions were external to the vesicle. The transporter compound was then added to the system to initiate ion transport and the experiment was started. The concentration gradients present for both anions meant the chloride ions were diffused out of the vesicle and the nitrate ions into the vesicle, via the antiport mechanism.

Throughout the experiment chloride efflux was monitored using a chloride sensitive electrode which measured the amount of chloride present outside of the vesicles. The chloride efflux was monitored over time and after 5min (300 seconds) the vesicles were lysed to break up the lipid bilayer and release the remaining chloride ions. A final chloride reading was taken at 7 min, which was used as the 100% chloride efflux reading. The chloride efflux at the point of transporter addition was used as the 0% reading and all other readings were calibrated to these readings.

These transport experiments gave values for chloride efflux over time for a specific concentration of transporter molecule. They were then repeated a number of times with different concentrations of transporter molecule, which allowed analysis of how potent the compound was for anion transport through calculation of the transport ability (EC_{50}).

2.1.3.1 Experimental Data in Papers

The experiments summarised above allowed the calculation of $EC_{50,270s}$ values (concentration - mol% carrier w.r.t lipid, required to obtain 50% efflux of anion after 270 seconds) which were presented in the papers as a measure of transport ability of the transporter molecule. Some papers also included the Hill number (n) which is produced in the calculation of the EC_{50} and the initial rate of chloride efflux (k_{ini}) which could be calculated from a plot of chloride efflux over time. Neither the Hill number nor k_{ini} were used in the following QSAR analysis.

EC_{50} is the 'half maximal effective concentration', often used as a measure of the potency of a 'drug' molecule. It is the concentration of a molecule necessary to induce a response at a given time which is 50% of the maximal possible response. [103] While EC_{50} is predominantly used in bioassays, anion transporter research uses this terminology for anion transport ability determined through vesicle experiments. When discussing chloride ion transport the EC_{50} is calculated using chloride efflux as the response.

EC_{50} can be calculated from the dose-response curve of a compound, as shown in Figure 2.4, where the response (chloride efflux at a given time) is measured as a function of concentration of transporter molecule. The process of determining EC_{50} from the fitted curve is called Hill analysis.[2]

A low EC_{50} indicates an effective transporter as it only takes a low concentration to achieve 50% efflux, whereas a high EC_{50} values indicates that more transporter is required to achieve the same efflux. When these values are logged for QSAR analysis, $\log(1/EC_{50})$, this is inverted and high values indicate potent transport.

K_a (stability constant) values were also present in a number of papers. These are not a measure of anion transport, but instead a form of equilibrium constant, which signify a compounds' ability to bind an anion. They are determined through analysis of chemical shifts in ^1H NMR in response to the presence of anions. [104]

In some experiments the EC_{50} measurements were taken w.r.t both the Cl^-/ NO_3^- and Cl^-/HCO_3^- antiport processes and the K_a measurements w.r.t chloride (Cl^-), bicarbonate (HCO_3^-) and nitrate (NO_3^-). However, the curation and modelling carried out in this research only focused on those K_a values relating to the chloride ion binding ($K_{a\text{-}Cl^-}$) and the EC_{50} values relating to the chloride/nitrate antiport process.

[2] Discussed further in Section 3.7.1

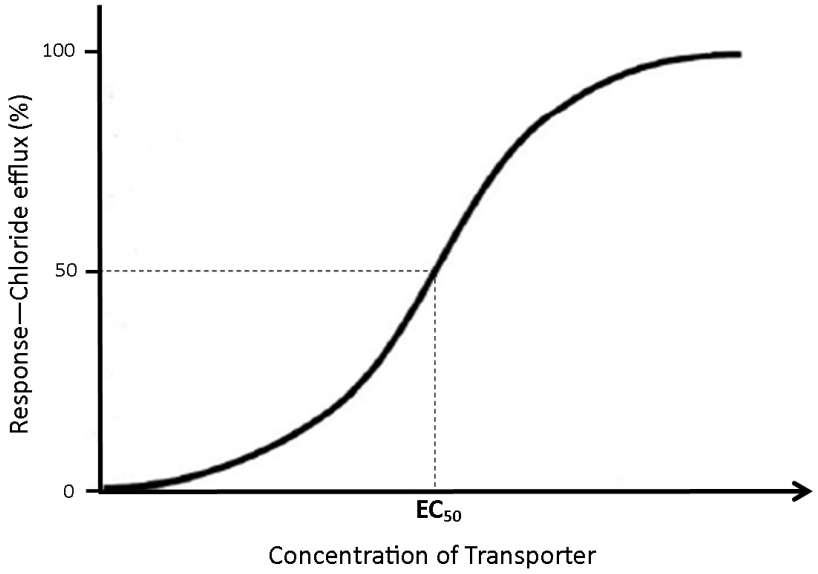


Figure 2.4: EC_{50} for chloride ion transport demonstrated on dose-response curve

The experimental data was provided in tables in their published papers alongside fitted plots and analyses in image format in the respective ESI.

2.2 Data Extraction

In order to carry out QSAR analysis on on the anion transporters it was necessary to have a dataset which contained all the relevant compounds and the data associated with them. Where QSAR analysis is carried out it is often on data that has been provided by other researchers or obtained from multiple sources. When this is the case is it necessary to curate the data to ensure that structures and data extracted from sources are as accurate as possible. [105] In data analysis well curated data are the building blocks for good analysis, if you put garbage in you'll get garbage out.

For ease of data extraction the ideal format of the compound data would have been all compound data contained within a single database or multiple databases with clearly defined structures. Previously a degree of modelling had been carried out on small sets of data. However, no modelling had been attempted across a larger dataset, as such no single database existed for the compounds, nor were any data stored in an easily accessible manner.

Therefore the first step in obtaining a curated set of data was to 'extract' the data for the compounds from the papers in which they were published [91–100] and create a database to collate them in.

Data extraction would not have been an arduous process if the information had been presented in the papers in an identical or similar fashion with easy access to the compound structure files and underlying data. However the papers were each presented in different ways to highlight the important aspects related to the topic being discussed, and the data in associated files were largely in image format. As such all data required had to be manually extracted from each paper.

2.2.1 Compound Structures

The first problem encountered with data extraction was that no compound structures were provided in the supplementary information. The structures presented in the papers were only given in image form (Figure 2.5) and also depicted as Markush structures to minimise the space required. This format meant they could not be extracted automatically from the papers. Nor could the compounds be converted from names as not all compounds were named in the papers. The image form given in the papers was presented in a human readable format but this was not computer readable. Research has grown in the area of optical character recognition and paper-mining; however, these are not yet widely utilised techniques. [106, 107]

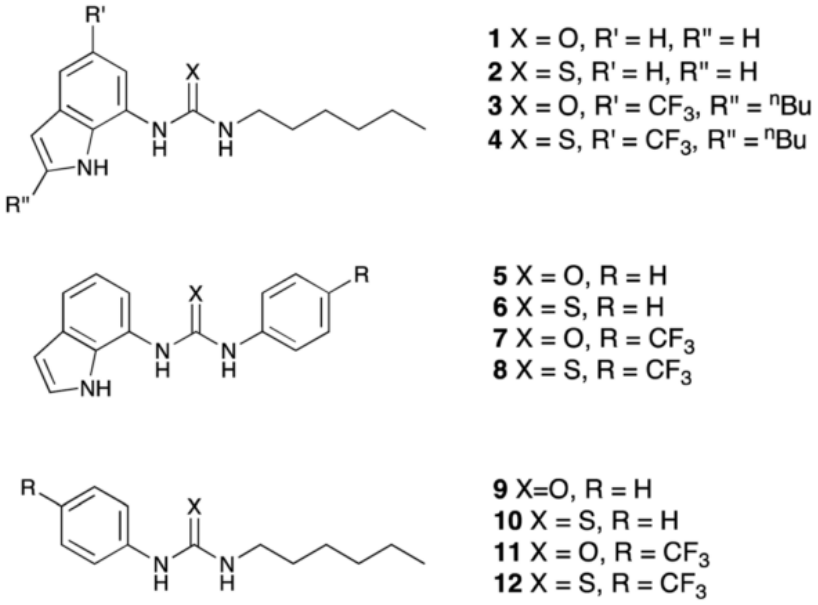


Figure 2.5: Example of structures in a paper

In order for a computer to process the structure and calculate properties and descriptors it was necessary to create a computer readable molecular representation for each compound. Many programs can take inputs in the form of a compound name or representation in the form of a SMILES string or InChI string alongside a structure file. However,

compound names are prone to typographical errors and difficulties in conversion and SMILES and InChI both require a name or structure to generate them from. The most suitable option for molecular representation was the creation of a 2D structure.

The molecular structure for each compound was extracted from the paper and drawn 'by-hand' in ChemDraw [4] and saved as .cdx (ChemDraw format) and the more widely used .mol file formats. Within this process the structures were checked to ensure they were accurate compared to the depictions in the paper, that they obeyed the standard bonding rules and that, where given, they corresponded to the name in the paper. [105]

Each compound extracted from a paper was given a unique identifying reference, comprised of the DOI of the containing paper and compound number of the compound within that paper. Examples include: '10.1039/c3sc51023a:1', '10.1021/ja205884y:7' and '10.1002/anie.201200729:1'. The structures drawn in ChemDraw were saved with their identifying reference (substituting illegal filename characters - removed '.', replaced '/' with '_' and replaced ':' with '-').[3]

This encoded information about the atoms present in the molecule, their co-ordinates and the connectivity between the atoms in a computer readable format, suitable for use as an input to a variety of software, such as Chemicalize.org [5] and DRAGON 6.0 [8, 108] to calculate chemical descriptors for the molecules.

2.2.2 Experimental Values

Alongside the extraction of the compound structures, it was also necessary to obtain the experimental data for each compound. This was extremely time consuming as each paper presented the data in different ways, preventing the automation of the data extraction. Some information was tabulated, whilst other values were included the text, even within tables the formats varied wildly and differing terminology was used.

Careful examination and manual transcription of the data was required to extract the data from the papers. If a database or spreadsheet had been created initially when the compounds had been tested then it would have been a significantly easier process to simply add new compounds to the database as they were synthesised, with data from the spreadsheet being easily convertible to other usable formats.

Across the selection of papers the most frequently present experimental values were EC_{50} values for Cl^-/NO_3^- transport and K_a values for Cl^- binding obtained from the physical experiments. Not all publications contained both of these values and others publications also contained additional experimental values such as; Hill parameter values (n), melting points, K_{ini} values (initial chloride efflux) and additionally values for bicarbonate transport (EC_{50} and K_{ini}, which were not used in our database).

[3] All compound structures can be found in ESI - both in .mol file format and as images in a pdf.

Throughout the extraction of the experimental data the following steps were followed to minimise the risk of errors being introduced to the dataset [109], often this required examination of the ESI to check full details that were not covered in the paper.

- Careful checking of values to eliminate transcription errors
- Checking for discrepencies in the data, if the values were available in more than one location
- Ensuring the units of all measurements were exactly the same
- Carefully check all measurements were using the same protocol and conditions

Developments in open data initiatives and guidelines by funding bodies are pushing researchers to publish data associated with publications to eliminate obstacles such as those encountered when trying to validate results presented in papers, or carry out analysis using larger datasets created using multiple sources of research. However, these are not widely implemented and many papers still do not contain access to datasets. [110, 111]

2.2.3 Compound Database

The compounds were initially collated in an access database where the data extracted from papers was combined with other information generated on Chemicalize.org. Information populated in the database included an ID, structure references, compound names, experimental values from papers (where present), InChI descriptors and molecular formulas. See Table 2.1 for the described structure of the access database.

Compound type and subtype were also extracted from the papers for each compound. It is worth noting that the values in this database were based upon the terminology used in the papers, rather than a robust classification process and as such they are not the most reliable classifier. The compound types are based around the urea/thiourea/squaramide motif; however, the subtypes potentially have multiple methods of classification.

In total 131 structures and their associated data were extracted from the 10 papers that were examined, some compounds were also included from unpublished work provided by the researchers. The total number of extracted compounds included a number of duplicate compounds which were present in more than one paper. The full database can be found in the ESI.[4]

Throughout the process of data extraction from the papers a number of difficulties had to be overcome which included:

[4] Experimental_Chemical_database.xlsx [30]

Field Name	**Type**	**Description**
ID	Autonumber	Primary key, unique identifier in database
Compound_name	Text	Name of compound (from ChemDraw)
IUPAC_name	Text	Name of compound (from Chemicalize.org)
Compound_type	Text	compound type e.g. urea (from paper, if present)
Compound_subtype	Text	compound subtype e.g. bis-urea (from paper, if present)
EC50_Cl-NO3-	Number	EC_{50} measurement for Cl-/NO3- antiport (mol% w.r.t lipid)
Ka_Cl-	Number	Ka measurement for Cl- binding (M^{-1})
CAS Number	Text	CAS Registry number from SciFinder [28] or similar database - if found
Paper_structure_reference	Text	Location of structure within paper. DOI: compound number (unique identifier)
Contained_in_paper	Text	DOI of paper in which the compound is contained
ChemspiderID	Number	Chemspider Identifier, would be generated in searching the chemspider database - if found
InChiKey	Text	Shorter form of the InChI descriptor, can more readily be used for searching
InChi_String	Memo	InChI descriptor, describes the structure of the compound
Found_Chemspider	Yes/No	Indicates whether the compound was resolved in Chemspider
Molecular_Formula	Text	Molecular Formula
Equivalent_Structure	Text	If the compound is a duplicate, which compound(s) is it equivalent to
Notes	Memo	
Melting_Point	Text	melting point (if recorded)
Hill_Coefficient	Number	Hill coefficient (if included in paper)

Table 2.1: Access database structure for Gale Group Data

- Many values were blank - either no measurement was taken or the value could not be determined
- Some values were given as <x, >x or a range
- Values had comments attached to them, these are difficult to assign in a database.
- Some compounds had been published in more than one paper, in some cases the values were exactly the same, in other cases further studies had been done and average values had been taken, it was time consuming to identify which compounds were duplicated and whether the values were the same or not.
- Melting point frequently had 'decomposition' as its value.

Values in the papers which were blank or specified as a range were not incorporated into the database as only numerical values can be easily utilised in the regression models. Comments were also unable to be carried across with the record as they could not be assigned to a specific value's measurement due to the design constraints of the database.

Duplicate identification was carried out through comparison of the InChI strings. This molecular representation was preferred over SMILES strings, as despite its lack of human readability it produces unique identifiers, whereas multiple SMILES strings can exist for the same molecule. [76]

Where two or more records existed for the same compound the measured values were compared. If one record had a more complete set of values then that compound was selected and the other compound removed from the dataset. If the records were identical then any of the compound records could be selected. If compounds had differing values then the source information was examined further to identify which values were erroneous.

Once the compound data had been extracted and examined from all papers the database contained 131 compound records, of these records 17 were duplicate records. From the 114 non-duplicated records, 29 compounds were missing the EC_{50} value and an additional 22 records were missing the K_a value. This resulted in 85 unique compounds with EC_{50} values and 63 records containing both measured parameters of interest (EC_{50} and K_a).

2.3 Generation of Descriptors

Following the population of the database with compounds it contained a number of properties for each compound record, but these were all either experimentally determined values (e.g. EC_{50}, K_a etc.) or non-numeric identifiers, such as compound name or InChI string which cannot be used in models. In order to perform analysis on the data it was necessary to calculate additional descriptors for all the molecules which can be used as independent variables in QSAR models.

2.3.1 Cheminformatics software

Many different chemical programs can be used to generate the same descriptors such as DRAGON [8, 108], Daylight [7], Chemicalize [5] and ChemDraw [4]. Although all chemical descriptors should be calculated through the use of a well defined algorithm many of the programs are proprietary software and therefore may use different algorithms that are not published and produce results that should not be used interchangeably.

Program	Mass	Exact Mass	LogP	ClogP	tPSA	nHAcc
DRAGON	503.67	N/A	N/A	N/A	126.63	10
ChemDraw	503.61	503.26	2.06	4.3122	126.63	N/A
ChemDraw 3D Pro	503.607	503.2645	2.23082	4.312199	126.63	4
Chemicalize.org	503.596	503.2645	2.98	N/A	126.63	4

Table 2.2: Comparison of descriptors generated by various programs for a single molecule - 101021_ja205884y-1.mol

In a comparison of different software programs it was discovered that even for the simple calculations e.g. Atomic Weight, which is a straightforward calculation, different programs can produce differing values from their calculations. This may occur from a program using previously published IUPAC atomic weight data [112,113] or from errors arising through handling number rounding. Simple calculations like that do not even use a complex algorithm but can still result in differing values for what may initially appear to be the same calculation.

Different versions of programs from the same provider may also contain adjustments to the underlying algorithms used in calculating descriptors, as well as different programs containing the same, or very similar, names for calculations which are not identical. Table 2.2 shows an example comparison of the calculations made by different programs for the same molecule. The tPSA (total Polar Surface Area) results are the same across all programs, but the number of hydrogen bond acceptors varies significantly and lipophilicity (log P) calculations vary from 2.06 to 2.98. Log P in particular has a number of different methods for calculating it, resulting in multiple descriptors such as LogP, ClogP, ALOGP, MLOGP, ALOGPs and many others.

It is very important to try and make sure that program details including version are included when talking about methodology for obtaining chemical descriptors in the same way that a synthetic chemist should record the parameters of their experiments. Ideally enough description should be included that would allow another scientist to repeat the analysis and obtain the same results. [114]

2.3.2 Processing compounds in DRAGON

Although a number of programs were available for calculating descriptors, DRAGON 6.0 was selected to generate the majority of descriptors. This program was able to calculate chemical descriptors in bulk batches, eliminating a significant effort that would have been required to manually process individual compounds or descriptors. It also minimised the risk of errors arising due to incorrect molecule selection or due to transcription errors in collating the data.

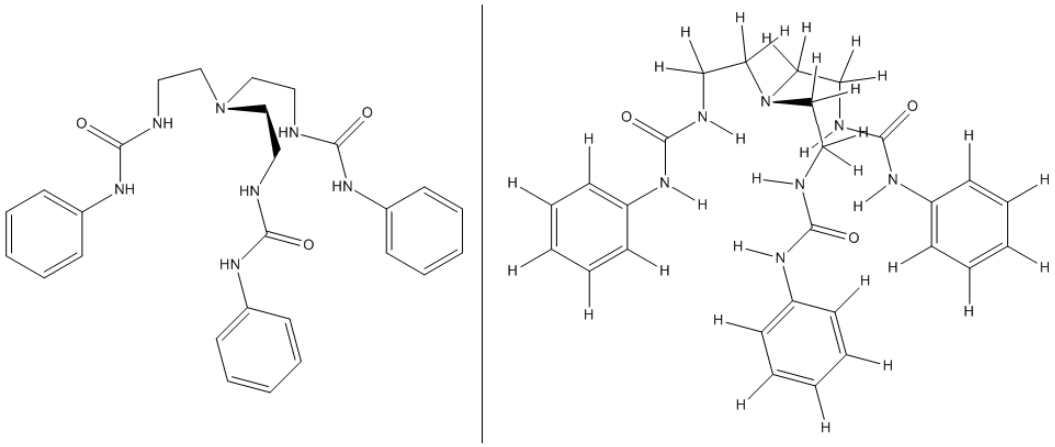


Figure 2.6: Implicit and explicit hydrogens in structure representation

The inputs used for descriptor generation were the 2D structures produced during the data extraction process, in the form of .mol files. These molecular representations allow the calculation of the simpler forms of descriptor, such as atom counts, fragment based descriptors and connectivity based descriptors.

It was also possible to calculate 3D descriptors in DRAGON; however, calculation of these descriptors relies on the input of a molecular representation containing 3D co-ordinates. These structures were not available, as such these descriptors were unable to be calculated.

In the process of descriptor generation it was discovered that in order for the chemical software programs to correctly process the molecules, it was necessary to expand any groups (e.g. CH_3, SF_5) and explicitly add hydrogens. This required the individual editing of each chemical structure, as the chemical descriptors were calculated incorrectly or failed to calculate if the structures were not fully described. Although the structures in Figure 2.6 are interpreted the same by a human, a computer has more strict rules on the interpretation of structures.

DRAGON can calculate a total of 4885 descriptors across 2D and 3D structures, of these many were not possible to calculate for 2D only. Other categories were excluded as they would not be beneficial to a linear regression QSAR model. Descriptors could be recalculated if a significantly different approach was taken for the QSAR model. The excluded descriptors included Atom pairs (1632 descriptors), Functional group counts and all 3D descriptors.

DRAGON returned 821 descriptors for each molecule in the raw form. This set of data was then cleaned to remove any descriptors which were not calculated properly (resulting

in NaN values) or which had negligible variance. Cleaning reduced the dataset to 702 descriptors for each compound.[5]

This number of descriptors exceeded the maximum number of fields (255) allowed in Access, so the database was moved to the statistical analysis program JMP 11 [20]. This removed the ability to have linked fields between different data tables; however, that functionality had not been widely used. DRAGON descriptors for all compounds can be found in ESI. [30]

2.3.3 Additional Descriptors

In one of the papers examined [91] there were also descriptors present that were calculated in other programs including Chemicalize.org, ChemDraw and Daylight. The calculation of descriptors like these was investigated for the full dataset. Chemicalize.org and ChemDraw do not allow batch processing of molecules and as such these were excluded, due to the time and transcription required to generate the descriptors. Many of which were already covered by those generated in DRAGON, such as molecular weight, log P, hydrogen bond donors etc. Descriptors from Daylight could not be investigated as this was not a program that was licensed for our use.

2.3.4 Quantum Descriptors

Whereas many simpler descriptors are concerned only with the connectivity or counts within a molecule, more complex descriptors exist involving quantum mechanics theory. These descriptors are derived from the Schrödinger equation. These methods can either be ab-initio (from first principles) or semi-empirical methods (using an approximated theory).

Quantum DFT (density functional theory) calculations were utilised in two of the papers examined [91, 93] to obtain values for $V_{s,max}$ (electrostatic potential surface maxima). The process of obtaining this descriptor required conformational analysis optimization (molecular mechanics) followed by geometry optimisation (DFT theory) and then calculations of the electrostatic potential. These were carried out through use of AMBER12 [14], Gaussian09 [15] and Wavefunction [115] programs.

Gaussian09 was used to investigate the feasibility of molecular structure optimisation with the aim of calculating $V_{s,max}$ for the additional compounds in the dataset. Testing of Gaussian was carried out using the IRIDIS supercomputer[6] with a single test compound. Compound $10.1039.c0sc00503g-3$ was selected for testing as it was the smallest compound in the dataset.

[5] Cleaning was carried out prior to elimination of duplicates

[6] IRIDIS High Performance Computing Facility, University of Southampton

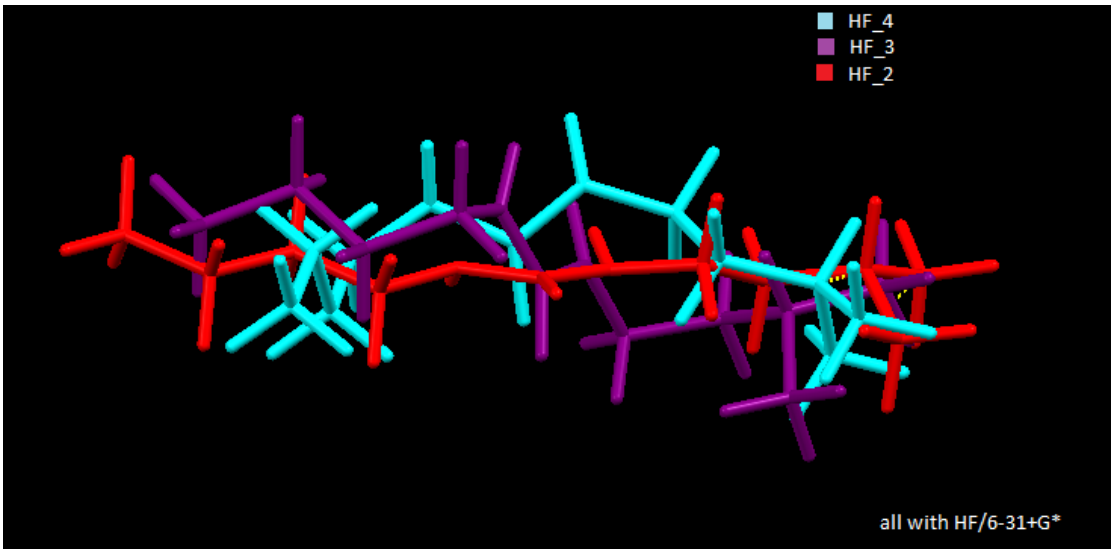


Figure 2.7: Output structures from Gaussian produced from 3 different input structures, molecules were aligned as much as possible

Geometry optimisations were run in Gaussian with a theory level of B3LYP/6-311+G** (the same theory level used in the previous papers). Compound $10.1039.c0sc00503g-3$ was optimised using 3 different algorithms to provide the input for Gaussian (2D optimisation - ChemDraw, 3D optimisation - Marvinsketch [10], 3D optimisation - OpenBabel [11]). Use of these inputs resulted in 2 runs timing out, at a runtime greater than 5hours and one run successfully completing, with a runtime of 4hours 50mins.

Use of a simpler theory level (HF/6-31+G*) resulted in runtimes from 13mins to 30mins for the same 3 input structures; however, the output structures obtained from the different input structures varied quite significantly, as can be seen in Figure 2.7.

The energy minimisation algorithm in Gaussian was identifying the local energy minimum rather than a global energy minimum. This means that the output structure is heavily reliant on the 'roughly cleaned' input structure. In the previous papers the DFT minimisation was preceded by molecular mechanics optimisation which analysed over 10000 structures. That method likely provided a better input structure; however, access to AMBER was not available.

Although quantum descriptors may provide a more accurate representation of the molecule, they are much more time-consuming to compute. [31] From the test of geometry optimisation it was decided that the amount of time and effort (including computational time on IRIDIS) required for the generation of the 'minimum energy' structures could not be justified. Especially considering this did not include the time required to calculate descriptors and the whole process was only providing a single descriptor for each compound.

2.4 QSAR Analysis

In the development of any QSAR model a response variable is required which the model is trying to predict. In many models this is a measure of biological activity of the compounds. For these molecules the experiments were carried out in vesicles in the lab and as such these measurements are physical measurements rather than biological. In the examined data sets there were 2 main properties measured for the compounds, these were the EC_{50} and K_a values.

The EC_{50} value was selected as the primary response variable for the QSAR analysis as it is an indicator of the transport ability rather than the binding ability. This is the variable which models will be predicting from inputs of calculated molecular descriptors. It would also be possible to build models for the K_a value; however, this was not the focus. Throughout the modelling EC_{50} values were converted to $\log(1/EC_{50})$ values in a similar fashion to how IC_{50} (inhibitory concentration) values are handled in QSAR analysis of drug responses.

The aim was to examine the data that had been extracted for the sets of chloride ion transporter molecules and build a model for chloride ion transport ability which could be applied to the set as a whole, or to a large portion of the set. Creation of a good model would allow a broad range of potential transporters to be investigated, testing them within the model to predict the transport ability prior to synthesis, only taking those molecules which have values above an acceptable level on to synthesis and physical testing.

The QSAR analysis was carried out in a number of sections: Analysis carried out on the data set as a whole, analysis carried out on subsets of the data and classification of the compounds for analysis.

2.4.1 Full Dataset Analysis

The collated dataset contained compounds that were all analysed for their anion transport ability, and have had the same experimental variables measured for them. However, within the dataset there was a large diversity in the chemical structures and functional groups of the compounds. Across the 114 compounds there were 6 different compound types and 22 subtypes (as extracted from the papers) with molecular weights that ranged from 186.34g mol^{-1} to 1067.76g mol^{-1}.

Due to the complexity of chemical interactions it was unlikely that modelling the dataset with a simple multi-linear regression (MLR) would produce a model that successfully fits the predicted variable, in this case $\log(1/EC_{50})$, unless the descriptors can describe all the underlying causes of changes in efficiency of the compounds for chloride ion

No. of Param.	Descriptors			R^2	RMSE
1	VE2_H2	-	-	0.1526	0.9027
1	DLS_04	-	-	0.1413	0.9087
1	VE2_A	-	-	0.1380	0.9104
1	VE2_X	-	-	0.1372	0.9108
1	EE_L	-	-	0.1354	0.9118
2	SpDiam_Dz.e.	ECC	-	0.2975	0.8269
2	SpDiam_Dz.i.	ECC	-	0.2882	0.8324
2	SpPosA_D	WiA_Dz.e.	-	0.2701	0.8429
2	SpMaxA_D	WiA_Dz.e.	-	0.2701	0.8429
2	SpMAD_D	WiA_Dz.e.	-	0.2698	0.8430
3	CSI	TI1_L	WiA_Dz.v.	0.3820	0.7803
3	SpPos_D	VE3_X	AVS_Dz.e.	0.3819	0.7804
3	SpMax_D	VE3_X	AVS_Dz.e.	0.3819	0.7804
3	SpAD_D	VE3_X	AVS_Dz.e.	0.3819	0.7804
3	CSI	TI1_L	SpMAD_Dz.v.	0.3770	0.7835

Table 2.3: Top models for $\log(1/EC_{50})$ produced through fit-all - up to 3 parameters ranked by R^2 for the full Anion transporter dataset

transport. However, exploratory QSAR analysis will provide insight into the data and how to proceed with modelling.

2.4.1.1 Fit all Models

The full dataset of 114 compounds alongside their descriptors from DRAGON was examined, in the statistical program JMP 11,[7] to investigate fitting a single MLR model for all compounds. To try and identify descriptors that could produce a good model, the 'Fit all models' feature was used to generate the best fitting models (ranked by R^2) out of all possible combinations of the descriptors, with a maximum of 3 terms per model. 3 terms were selected as the maximum number due to the memory constraints of the program, even limiting this to 3 generated 57,658,653 possible models. Increasing the number of terms above 3 exceeded the total number of models that JMP can generate at one time; however, when building a model it would not be beneficial to include too many terms as this could lead to over-fitting the model. [67]

The fit-all method was run using all the descriptors for the 85 compounds with EC_{50} values to build the models for the dataset, Table 2.3 shows the top models produced through 'fit-all'.

The top single parameter model gave an R^2 value of just over 0.15 with VE_H2 being the top modelled descriptor and DLS_04 being the second - this is a modified drug-like

[7] Analysis was also carried out in JMP 12 and JMP 13 as the licenses changed annually

score with 7 rules, similar to Lipinski's rule of 5. Moving up to 3 parameters increased the R^2 value to 0.38 for the best 3 term model, this was a reasonable increase on the single parameter model but still does not produce a good fit for the compounds. The plots for the top models are given in the appendix (Figure A.1 - A.5).

Neither the model statistics nor the plots suggested that these were good fits for the full dataset, the single model using DLS_04 showed a very poor inverse correlation to the $\log(1/EC_{50})$. This was the opposite direction to that which might be expected, as DLS_04 is a drug-like score. The rules surrounding lipophilicity and size of molecule would be expected to still apply to the anion transport; however, there are obviously other factors which may be more influential.

The 2 and 3 parameter models did not produce a strong linear fit either, with a number of outliers and scattered points. It was interesting to note that 2 of the furthest outlier points were the same across the two plots, both being bis-squaramide compounds, suggesting that this compound group may not model well with the other groups.

As anticipated a model built using the whole dataset with minimal parameters exhibited little correlation to the activity and would not be useful to accurately predict the activity, $\log(1/EC_{50})$, of a molecule for chloride ion transport. None of the models suggested here are strong enough to be taken forward for further testing. Further models should be investigated using additional parameters or by modelling different compound groups separately.

2.4.1.2 Stepwise descriptor selection

An alternative method to the 'fit-all' models for descriptor selection is to take a stepwise approach; in contrast to the multiple models produced by using the fit-all models feature the stepwise fit produces a single model for the dataset. The algorithm can be operated in either a forward, backwards or bidirectional fashion; in the forward direction the algorithm selects the most significant variable from the variables available and in each subsequent step selects the most significant variable which has not already been included. In the backward direction the algorithm starts by including all variables and at each step removes the least significant one. The bidirectional method is a combination of forward and backward testing at each step for inclusion and elimination of variables. These methods all proceed until they reach the stopping criterion specified by the user. [116]

The stepwise method was carried out on the full dataset with all DRAGON calculated descriptors available as variables under a number of different conditions outlined below.

Running the stepwise method in the forward direction with a minimum BIC value (Bayesian Information Criterion) as the selected criteria for stopping the algorithm produced the steps shown in Figure 2.8. The best selected model in this stepwise algorithm

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	DLS_04	Entered	0.0002	12.13237	0.1613	.	2	217.489	224.402
2	ICR	Entered	0.0252	3.90051	0.2132	.	3	214.466	223.574
3	LOC	Entered	0.0007	8.22388	0.3226	.	4	204.465	215.709
4	VE2_X	Entered	0.0027	5.671815	0.3980	.	5	197.117	210.437
5	ChiA_Dz(i)	Entered	0.0008	6.316452	0.4820	.	6	187.189	202.522
6	SpMax_B(p)	Entered	0.0005	5.837439	0.5596	.	7	176.336	193.617
7	VE2_H2	Entered	0.0009	4.580808	0.6205	.	8	166.657	185.817
8	P_VSA_MR_2	Entered	0.0008	4.069488	0.6746	.	9	156.64	177.608
9	RBF	Entered	0.0004	3.907976	0.7266	.	10	145.043	167.746
10	EE_B(m)	Entered	0.0004	3.38403	0.7716	.	11	133.048	157.406
11	SpMax_B(s)	Entered	0.0141	1.426253	0.7906	.	12	128.77	154.705
12	AVS_H2	Entered	0.0112	1.411858	0.8093	.	13	123.985	151.411
13	SpMaxA_L	Entered	0.0226	1.063486	0.8235	.	14	120.67	149.498
14	Psi_e_0	Entered	0.1528	0.401893	0.8288	.	15	121.246	151.384
15	S0K	Entered	0.1241	0.456477	0.8349	.	16	121.478	152.83
16	Hy	Entered	0.0827	0.566298	0.8424	.	17	120.945	153.409
17	J_Dz(p)	Entered	0.0618	0.633558	0.8508	.	18	119.841	153.31
18	ON0V	Entered	0.0732	0.56146	0.8583	.	19	119.142	153.506
19	TPSA(NO)	Entered	0.2035	0.276507	0.8620	.	20	120.616	155.757
20	J_D/Dt	Entered	0.2599	0.215494	0.8649	.	21	122.648	158.443
21	Mp	Entered	0.1896	0.289703	0.8687	.	22	124.158	160.478
22	P_VSA_s_3	Entered	0.1578	0.331081	0.8731	.	23	125.38	162.088
23	MAXDN	Entered	0.0971	0.445923	0.8790	.	24	125.617	162.571
24	Best	Specific	.	.	0.8235	.	14	120.67	149.498

Figure 2.8: Step history for forward stepwise algorithm with minimum BIC criteria

had 14 parameters, including the intercept. The R^2 value calculated in the stepwise process was 0.8235; however, the method of processing the data is different in the stepwise procedure to the Run Model procedure which can lead to less data being used in the stepwise method giving R^2 values which are more of an approximation. When fitting the 14 parameter model through the Least Squares fit it gave an R^2 value of 0.7655 and an R^2_{adj} value of 0.7226. However, 2 of the parameters (ICR and SpMax_B(s) indicate statistical insignificance and a further two parameters (EE_B(m) and SpMaxA_L) are also close to the threshold.

Using the AICc (Corrected Akaikes Information Criterion) as the stopping criteria instead gave a similar model to the BIC stopping criteria; however, the model contains 19 parameters, including the intercept. Fitting the 19 parameter model through Least Squares fit gave an R^2 value of 0.8139 and an R^2_{adj} value of 0.7631. This also has parameters which are not statistically significant (EE_B(m), AVS_H2, SpMaxA_L, ON0V) and a number which are close to the threshold (SpMax_B(s), S0K, J_Dz(p))

The plots produced for these models (see appendix, Figure A.6 & A.7) showed much stronger correlations; however, the number of parameters included is considered too high to produce a good model. Including too many parameters in a model can result in over-fitting. It is suggested that the ratio of observations to variables should be in the region of 10 to 20 [117,118] depending on the model application to avoid overfitting. In these cases the observations per variable range from 6 in the BIC model to 4.5 in the

AICc model. As there are only 85 compounds with EC_{50} observations in the dataset the model should contain significantly fewer variables. In addition a number of variables in each model were not statistically significant and should not be used in the model.

The stepwise method is also considered to have a number of deficiencies in its methods [119,120], resulting in potential overestimation of R^2 values and its inability to cope with collinear variables. Collinear variables are variables which have a linear relationship between them. In these models some of the variables that have been selected by the stepwise regression are the same descriptor weighted in different ways, for example the SpMax_B(p) and SpMax_B(s) are both the leading eigenvalue from Burden matrix, weighted by polarizability and I-state respectively. Although the two variables in this case are not well correlated to each other ($R^2 = 0.11$) it is important to check for cross correlation in the variables.

Rerunning the linear fit after removing the four variables (ICR, SpMax_B(s), EE_B(m), SpMaxA_L) from the BIC fit gives a 10 parameter model, ($R^2 = 0.683$, $R^2_{adj} = 0.644$). Looking for cross correlated variables showed a number of high correlations, in particular VE2_X, VE2_H2 and ChiA_Dz(i) with approx 0.895 cross correlation between them all. Variables VE2_X and ChiA_Dz(i) were removed as they had higher correlation with the other variables. (See appendix for plots - Figure A.8 - A.11)

This resulted in an 8 parameter model. However, the removal of the two highly correlated variables reduced the fit significantly ($R^2 = 0.455$, $R^2_{adj} = 0.406$) with the variable AVS_H2 no longer being statistically significant.

Although the higher parameter models gave much stronger correlations to $\log(1/EC_{50})$ than the fit-all models they had a number of issues: firstly the number of parameters included in the model were higher than desirable, secondly some of the variables were not statistically significant and thirdly some of the variables were correlated to each other. This resulted in no model that could be taken further in validation testing.

To obtain models with an acceptable number of parameters it is highly probable that the data will have to be categorised to model them slightly differently for the presence of different chemical features. To achieve this it was possible to either start with the whole dataset and try to partition it, or start with a smaller subset and try to expand it. As no key terms were identified in the initial model building stage the first step was to start with a small subset and try to increase the size of the set.

2.4.1.3 log P fits

In many papers researchers have commented on the relationship between the anion transport and lipophilicity. In some of the Gale group papers the relationship of these compounds was examined, using ClogP from Spartan [12]. Lipophilicity is a key factor

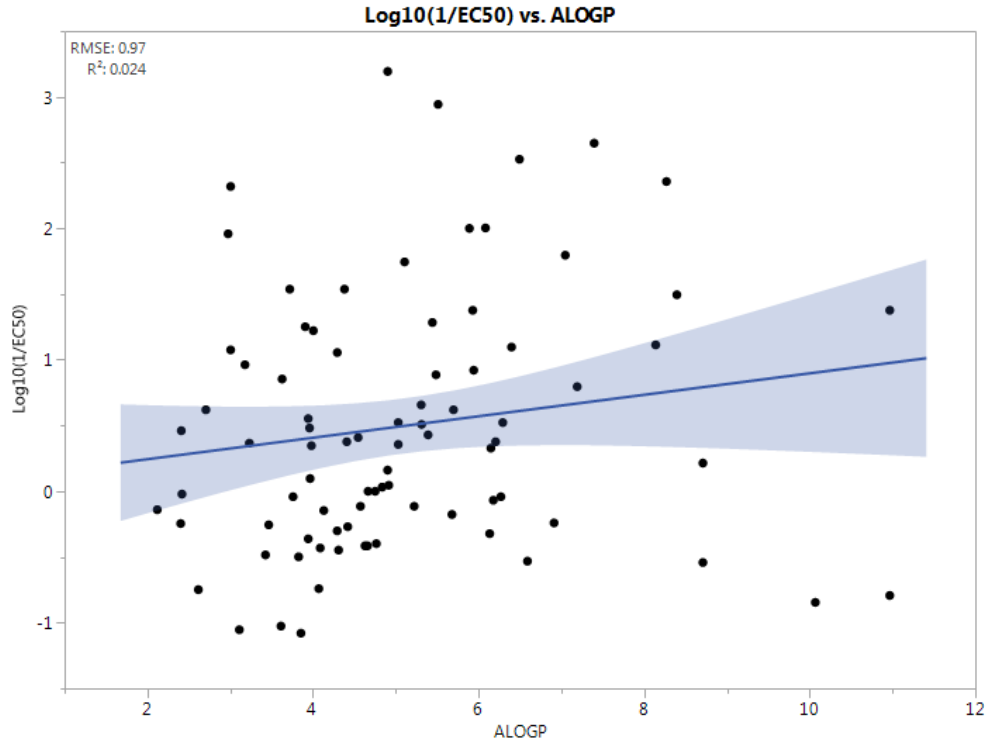


Figure 2.9: Linear fit of $\log(1/EC_{50})$ vs ALOGP for full dataset

in the movement across a lipid bilayer, it appeared that the less lipophilic a compound is, the worse it functions as an anion transporter. However, most of these relationships were not modelled mathematically but qualitatively examined.

The relationship of log P to $\log(1/EC_{50})$ was examined for the full dataset. This used the ALOGP descriptor from DRAGON rather than the ClogP value from Spartan as this was not available. Figure 2.9 shows the linear fit of $\log(1/EC_{50})$ with ALOGP. This exhibits almost no correlation, with the points being very scattered.

It is possible that some compounds are too lipophilic to pass through the membrane easily. This is referred to as lipophilic balance, where a compound needs to have a certain level of lipophilicity, but not too much, to easily move through a lipid bilayer. This can be modelled through a quadratic equation, Figure 2.10 shows the equivalent quadratic fit to Figure 2.9. This shows a fractionally better fit (from R^2); however, no obvious correlation exists across the whole dataset.

2.4.1.4 Implication of Experimental Error

When examining experimental data it is important to consider the scale of errors that are present in the data, as large uncertainty in the data can lead to poor quality models. In order to examine the effect of errors in this dataset the error values for transport

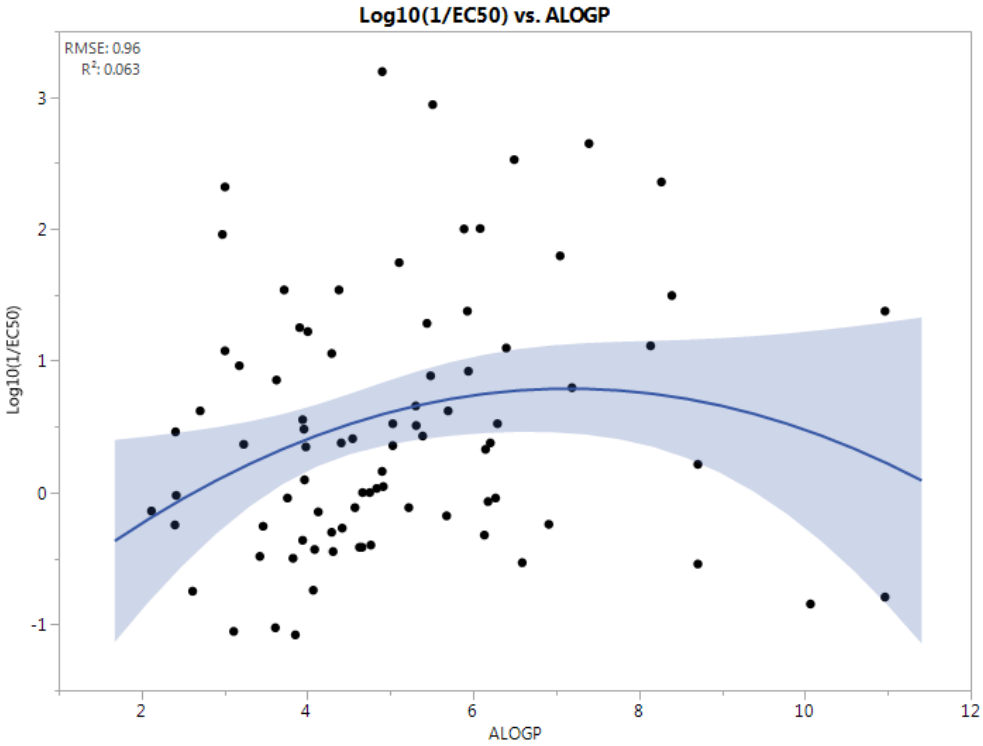


Figure 2.10: Quadratic fit of $\log(1/EC_{50})$ vs ALOGP for full dataset

ability (EC_{50}) were examined. These error values were extracted from the papers in which the EC_{50} values were originally provided [91–100].

Very few error values were included in the main body of the papers examined. Instead the errors had to be extracted from graphs within supplementary information files. Many of the papers did not include error measurements even in the ESI, only providing an image plot without parameters. Of the 85 compounds that had EC_{50} values only 44 compounds had error values provided.[8]

Once extracted the EC_{50} errors were propagated through to give an error for the $\log(1/EC_{50})$ value and plotted to examine their magnitude. Figure 2.11 shows a simple scatter plot of $\log(1/EC_{50})$ against ALOGP, with the incorporation of error bars for $\log(1/EC_{50})$. It can been seen that the magnitude of the error bars is small for the majority of the compounds, with only a few compounds exhibiting error bars that might require further examination.

No strong correlation existed across the whole dataset so the presence of error bars does not support or refute any correlation. However, across the majority of the dataset it shows very little error.

Only 5 compounds exhibited $\log(1/EC_{50})$ errors greater than 0.1 log units, (101039_c3sc 51023a-16, 101039_c3sc51023a-18, 101039_c3sc51023a-19, 101039_c3sc51023a-20, 101039

[8]Spreadsheet containing $\log(1/EC_{50})$ errors can be found in ESI

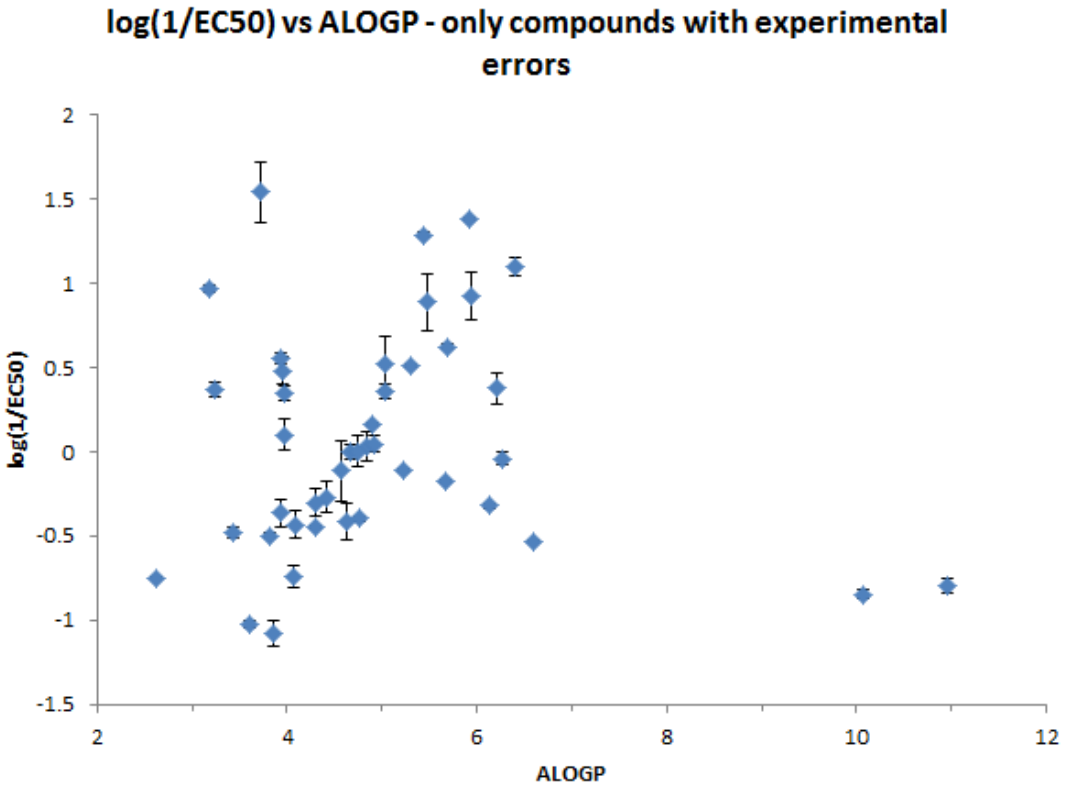
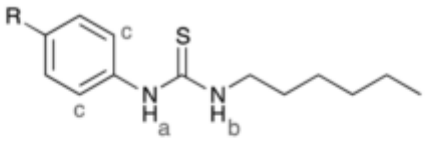


Figure 2.11: Plot of $\log(1/EC_{50})$ against ALOGP showing experimental error bars (limited to compounds that gave error values)

_c3sc51023a-21). It could be expected that the largest errors would be present at the extreme of the transport ability and lipophilicity, due the the difficulties in measuring low transport compounds and partitioning problems. However, these compounds are mainly in the centre of the range, with all 5 compounds coming from the same paper and the same compound type. Their structures can be seen in Figure 2.12 in Section 2.4.2. These compounds all have similar structures which may be the cause, but other compounds also have similar chemical features (thiourea, medium length alkyl chains, aromatic rings) but smaller errors. Further investigation of the errors would be required to determine the likely cause; however, the errors are not large enough to cause significant concern.

The errors given in the EC_{50} are, however, only those errors arising in the calculation of the EC_{50}. Errors in the experiment are included through repetitions rather than inclusion of error values for the experimental readings used in calculation. Further investigation would be beneficial to assess the error present in the underlying data of the chloride efflux over time. This would require access to the errors in the raw transport experiment readings as well; however, none of this data was available for analysis.

Due to the lack of error values across the whole dataset, and the relatively small magnitude of the given error values, errors were not incorporated further into the exploratory



1	R = Br	**12**	R = O(CO)Me
2	R = CF_3	**13**	R = OCF_3
3	R = Cl	**14**	R = OEt
4	R = CN	**15**	R = OMe
5	R = $COCF_3$	**16**	R = SMe
6	R = COMe	**17**	R = SO_2Me
7	R = COOMe	**18**	R = CH_3
8	R = F	**19**	R = CH_2CH_3
9	R = H	**20**	R = $(CH_2)_2CH_3$
10	R = I	**21**	R = $(CH_2)_3CH_3$
11	R = NO_2	**22**	R = $(CH_2)_4CH_3$

Figure 2.12: Structures of Compounds included in paper 10.1039/c3sc51023a

QSAR analysis carried out. However, the construction of robust models for future prediction would ideally include additional commentary on error measurements.

2.4.2 Subset Analysis from 10.1039/c3sc51023a

2.4.2.1 Summary of previous modelling

Previous attempts were made to perform QSAR on a small subset of the data in an existing Gale group paper, 10.1039/c3sc51023a [91] which contained 22 compounds within the thiourea chemical group. This subset of compounds was used as the starting point for examining the data and models for these molecules and to expand the data subset from that point.

The dataset presented in paper 10.1039/c3sc51023a was made up of 22 1-hexyl-3-phenyl-thioureas with a variety of substituents at the para-position of the phenyl ring (see Figure 2.12). In the paper these compounds were split into a test set and a training set. The test set contained compounds 1, 6, 14 & 20 and the training set the remaining 18 compounds. The test set was randomly selected from the dataset and not used to build the QSAR model. The modelling carried out looked at both the anion binding ability, through K_a, and the anion transport ability, through $\log(1/EC_{50})$.

Anion binding

The models created for anion binding mainly examined the relationship between K_a and the Hammett constants. This model would not be possible to apply directly to the full dataset as the K_a values are not present for many molecules and Hammett constants

are specific to benzene substituents and cannot easily be calculated directly from the structures.

A relationship was also investigated between $V_{s,max}$ and $\log K_a(Cl^-)$. This would be more generally applicable in comparison to the model containing Hammett parameters as it can take into account multiple substituents of more complex molecules. $V_{s,max}$ is a quantum parameter which could be calculated for any type of receptor; however, it requires minimised structures, which take a significant time to compute and were not calculated for the full dataset.

Anion Transport

Modelling for the anion transport ability showed a good correlation between $\log(1/EC_{50})$ and both log P values and HPLC retention times (RT - experimental). A high correlation also existed between log P and RT. ClogP values calculated through Daylight gave an R^2 correlation value of 0.95 when modelled against RT. Therefore, either retention times or log P values could be used as a descriptor to build models for predicting anion transport; however, retention times were not available for all the compounds in our wider dataset so would not have been applicable for expanding the model significantly. Additionally they are experimental parameters which means they cannot be calculated without the synthesis of the compound.

ClogP values modelled against $\log(1/EC_{50})$ showed a good correlation ($R^2 = 0.79$) for the phenylthioureas and highlighted the importance of lipophilicity as a factor in the anion transport. When modelled against retention time a stronger correlation of $R^2 = 0.84$ was produced.

An additional 286 molecular descriptors were calculated for the molecules in a variety of programs, and stepwise MLR was carried out in JMP 9.0.0. From their findings the best two parameter models contained one term describing lipophilicity and one describing molecular size/ shape, and the best three parameter models contained a term describing lipophilicity (e.g. RT), an electronic term (e.g. σ_p) and term describing molecular size (e.g. SPAN). However lipophilicity was the most important factor in the models built.

This gave models which had an easily interpretable physical meaning, a useful feature when trying to understand models and their chemical theory behind them. The models obtained showed a reasonable ability to predict the $\log(1/EC_{50})$ values for the test set compounds; however, no validation statistics were obtained. Additionally the small size of the test set (4 compounds) would make it difficult to obtain a reliable estimate of the predictive ability.

2.4.3 Expanded Subset Analysis

The dataset analyses observed in 10.1039/c3sc51023a [91] were re-run for the same subset, using descriptors obtained from ESI. The statistics obtained from the re-runs aligned with those published and their papers and showed good reproducibility for the correlations and models previously obtained by Busschaert et al. [91]

To examine if these correlations and models still held for other compounds outside of the original data subset of phenylthioureas, the subset was expanded to include more similar compounds, and the correlations re-examined.

The Open Source Chemistry Toolbox, OpenBabel [11] was used to compare molecular structures. In this process FP2 fingerprints were generated for each compounds molecular representation and a Tanimoto coefficient was calculated for the similarity of each molecule to a single selected molecule.

Fingerprints are binary vectors which encode structural features (or fragments) contained in a molecule, with a bit set for the presence of a feature. Their similarity is calculated through a comparison of the number of bits encoded in the two vectors as shown by the Tanimoto equation below. [121]

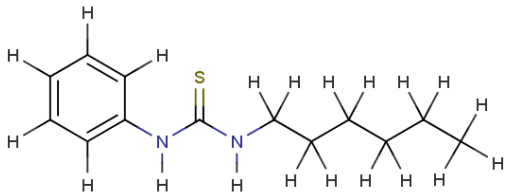
$$SIM_{AB} = \frac{c}{a + b - c} \tag{2.1}$$

where c is the bits set in common between A and B, and a and b are the bits set in fingerprints A and B respectively.

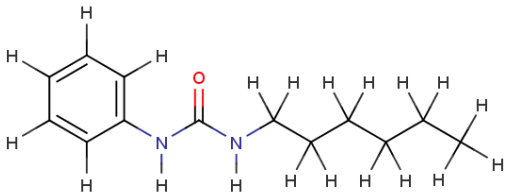
Compounds were selected if they met the following criteria; either a Tanimoto similarity coefficient value ≥ 0.8, or 'possible superstructure' to the comparison structure, thiourea compound 101039.c3sc51023a-9, see Figure 2.13a. This compound was chosen for the comparison structure as it was the smallest compound in the dataset, closest to a backbone structure.

This expanded the dataset to 27 compounds, with all of the selected compounds being thiourea compounds, this only showed an expansion of 5 compounds beyond the original subset. To expand the dataset further the corresponding urea analogue, 101039.-c2sc20551c-9 (see Figure 2.13b), was also used in OpenBabel to find similar compounds. this was selected as the overall backbone is similar but it contains a S atom instead of an O at one position.

The second selection expanded the dataset further to contain 41 compounds in total; however, 5 of urea compounds were missing EC_{50} values. This provided 36 compounds which could be used to examine to correlations with respect to anion transport. 33 compounds also had K_a values; however, correlations for this variable were not widely



(a) Thiourea molecule - 101039.c3sc51023a-9



(b) Urea molecule - 101039.c2sc20551c-9

Figure 2.13: Molecules used for structure similarity search

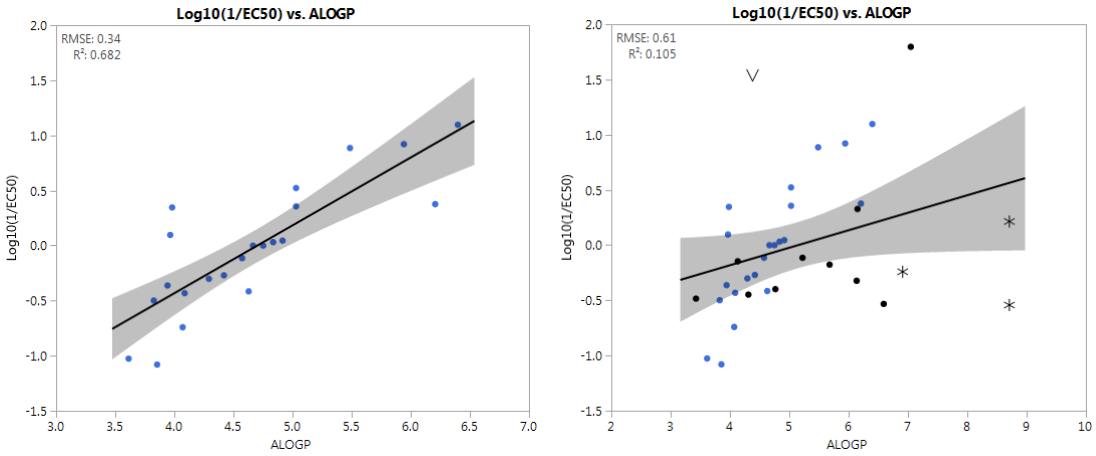
investigated. A list of compounds in the expanded subset with EC_{50} values and structures of the compounds can be found in the appendix - Section A.2.

In the original thiourea dataset (Figure 2.12) additional parameters were available compared to the descriptors generated through DRAGON. Additional descriptors included Hammett constants (for substituents), $V_{s,max}$ values, ClogP values and HPLC retention times, with a total of 286 molecular descriptors calculated through a variety of programs. It would be possible to calculate the majority of these descriptors for the expanded subset, with the exception of HPLC retention times as these were obtained through experiments. However, access to a number of the programs (Daylight, ACD/I-labs and ChemAxon) was not available and DFT calculations were not performed due to time requirements.

This limited the number of descriptors available to carry out a direct comparison of the models. Correlations could still be directly compared if models used descriptors from DRAGON, and where possible descriptors that were not available could be replaced by their corresponding or similar DRAGON descriptor.

The SPAN descriptor used in the initial model is a geometrical descriptor calculated from the 3D structure of a molecule and as such was not available. This could be replaced by another simpler metric of size such as atom count, bond count or weight. The lipophilicity measure CLogP could be replaced with MLOGP or ALOGP from DRAGON.

Due to the small size of the dataset it was not split into a training and test set for the purposes of this analysis.



(a) Initial thiourea subset (b) Expanded subset

Figure 2.14: Linear fits of lipophilicity for initial and expanded subsets

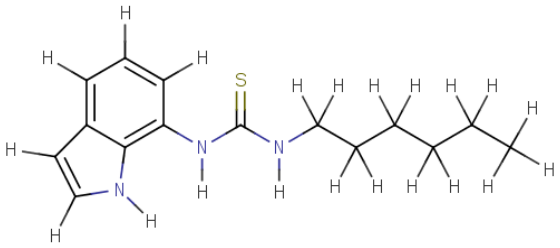


Figure 2.15: Structure of outlier compound 101039_c2sc20551c-2 marked by ∨

2.4.3.1 Modelling Lipophilicity

Linear modelling

As lipophilicity was identified as the key component of the previous models, the expanded subset was modelled against lipophilicity. ALOGP values were used instead of CLogP values as a measure of lipophilicity as CLogP could not be obtained for all compounds. Figure 2.14 shows linear fits for the two subsets, in the expanded subset the points marked by * show the bis-urea and bis-thioureas, the point marked by ∨ was a potential outlier. (Compound 101039_c2sc20551c-2 - structure shown in Figure 2.15)

The initial thiourea subset had an R^2 value of 0.682 when linearly modelling ALOGP against $\log(1/EC_{50})$. This was reduced compared to the correlation of CLogP, due to different calculation methods for the lipophilicity descriptors. Modelling the expanded subset gave a significant decrease in R^2 down to 0.105 when thioureas and ureas were modelled in a single linear model.

Splitting the linear plot by compound type, giving separate fits for the thiourea and urea compounds, gave an R^2 value of 0.16 for the thioureas and 0.01 for the ureas. This fit included the bis-ureas and bis-thioureas, which may be best modelled in separate groups

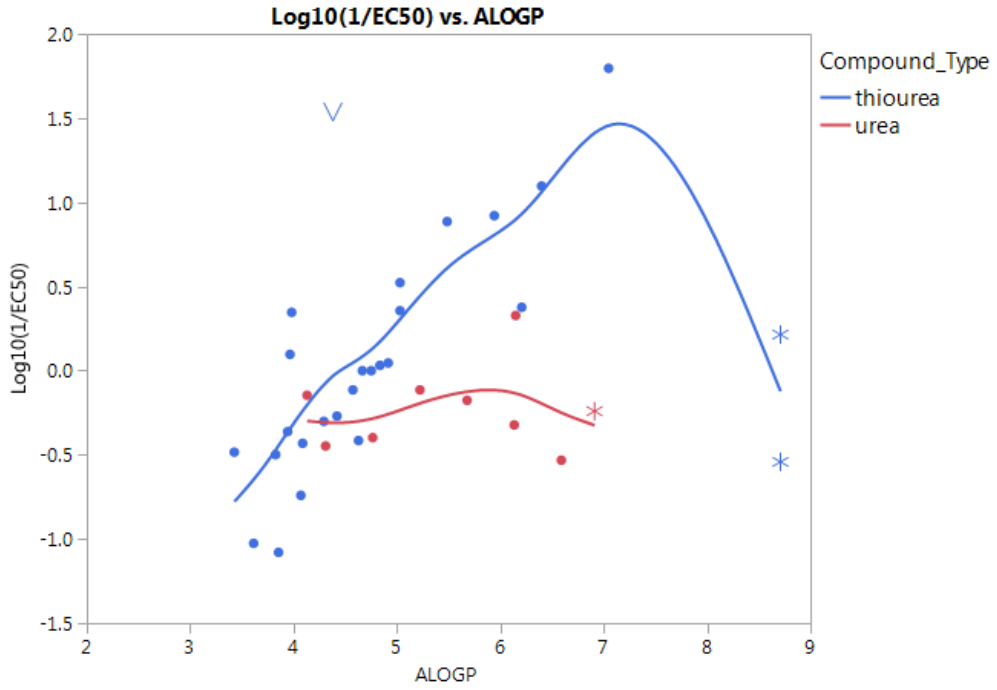


Figure 2.16: Plot of $\log(1/EC_{50})$ against ALOGP with smooth curve for the expanded subset, split by compound type

to the ureas/thioureas as they contained slight differences to the functional group and the bis-thioureas exhibit much higher log P values than the rest of the dataset. (See appendix for plots - Figures A.16 - A.17)

Excluding these bis compounds gave R^2 values of 0.59 for the thioureas, which showed that a correlation is somewhat continued with an expansion of the thiourea group through similarity, albeit a less strong one. The urea group does not exhibit good correlation with ALOGP with an R^2 value of only 0.02. The very low R^2 value for the urea group may be exacerbated by the low number of compounds in this group; however, across the urea group there is very little variation in the value of $\log(1/EC_{50})$.

The point marked by $\vee$ still appeared to be an outlier in the thiourea group; however, there were no observations within the paper that obviously suggested an erroneous result. This compound did contain a indole group though rather than a phenyl group present in the initial thiourea subset (Figure 2.15 & 2.12) which could account for the variation in activities. This was another occasion in which better data visualisation would have greatly aided the investigation - see Section 4.3.

Quadratic Modelling

Fitting the plot of $\log(1/EC_{50})$ against lipophilicity with a smooth cubic spline curve in both the split (Figure 2.16) and non split plots indicated that the compounds may potentially be better modelled by a curve rather than a linear fit to ALOGP.

In the linear fit the bis-thiourea and bis-urea compounds appeared to require a separate subset from the ureas and thioureas; however, it was possible that the series spans an optimal log P value for the transport activity. At this point the transport would be most efficient and as the log P moves away from optimum the transport ability decreases, due to inability to move between the aqueous and lipid layers. It could be overcome by the addition of a squared term to the model giving a parabolic relationship.

In the thiourea group the compounds exhibited a wide range of values across both the log P and EC_{50} values with a range of -1.1 to 1.8 in the $\log(1/EC_{50})$ and ALOGP values from 3.4 to 8.7. The EC_{50} values of the urea compounds on the other hand were much closer together, with all compounds except one having a $\log(1/EC_{50})$ value in the range -0.53 to -0. 11. The ALOGP values for the ureas were also a narrower range, from 4.1 to 6.9. It is possible that all compounds in the urea compound group are not particularly effective anion transporters, as they exhibited low $\log(1/EC_{50})$ and little variation in their transport efficiency. However, the small size of the urea subset limits the confidence with which this can be concluded.

Modelling the expanded thiourea and urea datasets (including bis compounds) with a quadratic fit rather than a linear fit gave an R^2 value of 0.51 for the thioureas and 0.09 for the ureas, see Figure 2.17. The urea group still showed little correlation between ALOGP and $\log(1/EC_{50})$; however, the thioureas including the bis thio-ureas did indicate a possible quadratic relationship to ALOG. The value for the thioureas increased to R^2=0.64 on exclusion of the potential outlier ($\vee$). Excluding the bis-thioureas appears to give a very weak parabola but with a minimum $\log(1/EC_{50})$ value rather than the maximum that would be expected. (See appendix for plots - Figure A.18 & A.19) There was a lack of data points in the high log P range, which has a large influence on the parabola.

Although the models for lipophilicity for the expanded subset are a large improvement on the full dataset they still do not exhibit correlations that would allow predictions to be made from the model equations. The best fits were exhibited when the subset was split into thiourea and urea groups, although this only gave a maximum R^2 value of 0.11 when considering the whole group. The expanded subset for this section was created through similarity of compounds to the backbone structure; however, the correlations should be examined in the whole group of ureas across the wider dataset.

2.4.3.2 2 Parameter Models

The expanded subset model was extended beyond lipophilicity to include a term for molecular size/shape. The SPAN descriptor that was used previously was unavailable so a number of simple descriptors that were a measure for molecular size/shape were

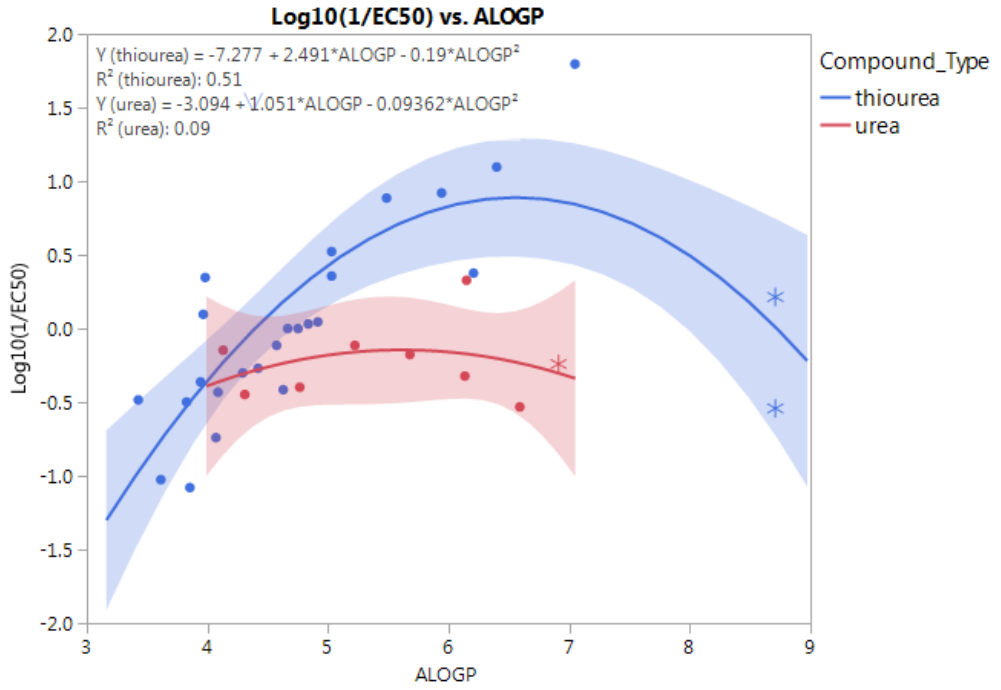


Figure 2.17: Quadratic fit of $\log(1/EC_{50})$ against ALOGP for the expanded subset, split by compound type

selected. In a fit-all models the highest fitting combination was ALOGP with MW (molecular weight) as a size descriptor.

Modelling ALOGP and MW against $\log(1/EC_{50})$ for the whole expanded subset gave a fit as shown in Figure 2.18, (R^2=0.447, R^2_{adj} = 0.414). In this plot the point marked by $\vee$ (Compound 101039_c2sc20551c-2 - see Figure 2.15) still appeared to be an outlier.

This 2 parameter fit for the whole subset produces a much stronger fit than ALOGP by itself, ($R^2_{ALOGP,MW}$=0.447, R^2_{ALOGP}=0.105) but it still was not good enough for predictions. Removal of the outlier improved the fit slightly (R^2=0.54, R^2_{adj} = 0.511).

Modelling separately for the thioureas and ureas produced the following fits and equations (excluding outlier) - see Figure 2.19:

$$\text{Thiourea: } R^2\text{=0.646, } R^2_{adj} = 0.616$$
$$log(1/EC_{50}) = 0.7226 * ALOGP - 0.0076 * MW - 1.167 \tag{2.2}$$

$$\text{Urea: } R^2\text{=0.448, } R^2_{adj} = 0.264$$
$$log(1/EC_{50}) = 0.193 * ALOGP - 0.0026 * MW - 0.112 \tag{2.3}$$

The fits given by ALOGP and MW for the split sets were the most promising fits generated for the dataset to this point; however, they still did not exhibit a strong enough fit to take through to validation testing.

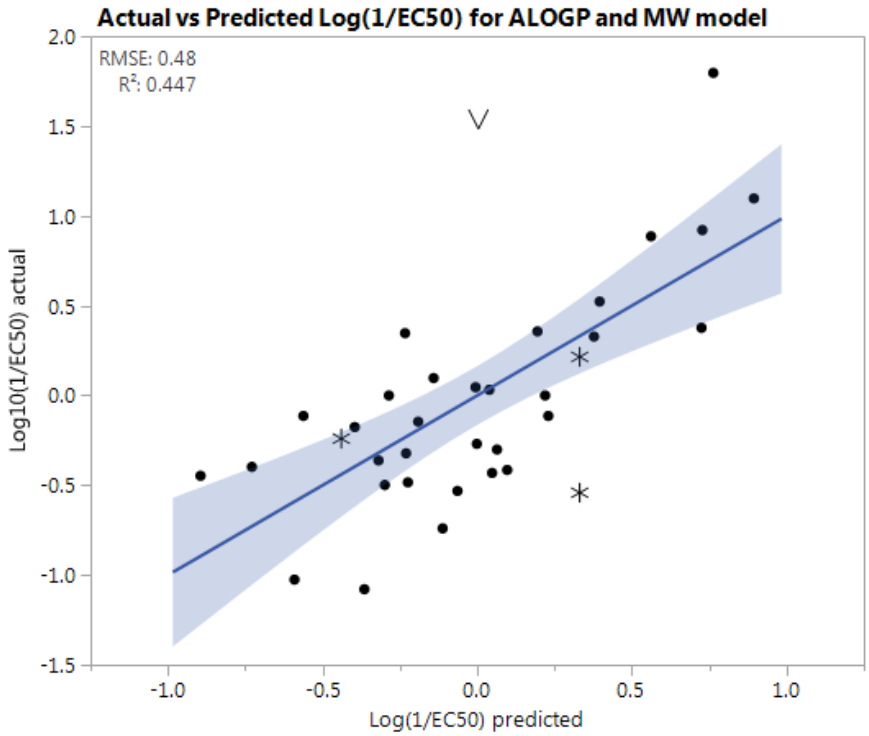
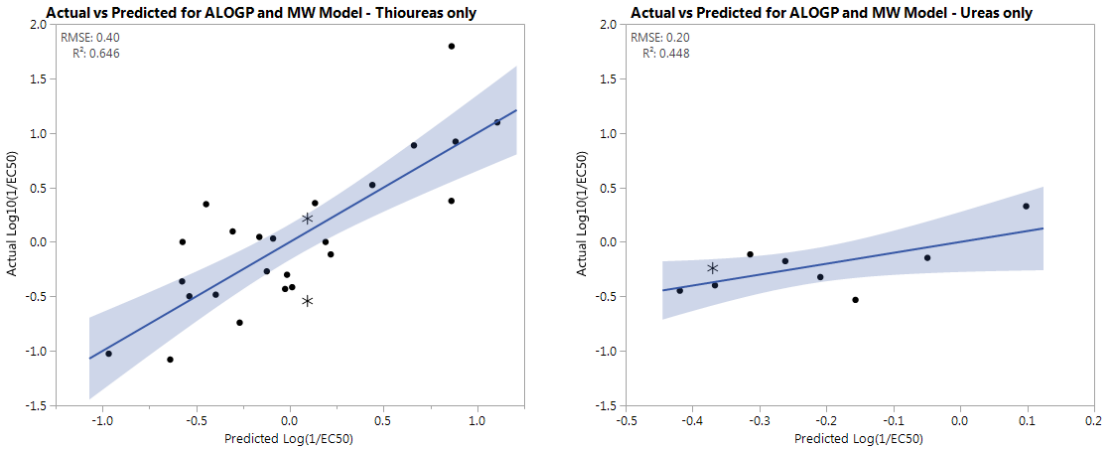


Figure 2.18: Actual vs Predicted for 2 parameter fit using ALOGP and MW for expanded subset



(a) Thioureas only - excluding outlier

(b) Ureas only

Figure 2.19: Actual vs Predicted values for $\log(1/EC_{50})$ modelling ALOGP and MW, compound types modelled separately

The urea analogue had been included in the expansion the subset due to its structural similarity to the thioureas; however, the presence of the O instead of S may have a significant effect on the anion transport, potentially due to a difference in hydrogen bonding affecting the anion binding. Modelling the thioureas and ureas separately always resulted in stronger model fits than the combined group. This separation should be applied to modelling across the whole dataset.

2.4.4 Grouping of Compounds

Within the data obtained from the Gale group the compounds exhibited a wide diversity in structures and functional groups. This increased the complexity when trying to model the data and find correlations. As mentioned in earlier sections the variation in chemical structures may require different groups of compounds to be modelled with separate models or parameters within a model.

Classification of the compounds could be carried out in a number of different ways, the concept is to group compounds containing similar features together. One method is to classify based upon chemical features present in the structure of the molecules, such as functional groups, separating the different types into groups, either through manual or automatic methods. An alternative is to carry out classification based upon the similarity of compounds through their physical action or chemical attributes.

2.4.4.1 Manual Classification

Manual classification involves each compound being assigned descriptors or a group through examination rather than processing by an algorithm.

Grouping by compound type

Classification by chemical features was carried out on the dataset during the subset analysis discussed earlier in this section. This was done manually with each compound being assigned a compound type and compound sub-type. The compound groups assigned were based upon the terminology used by the researchers in the initial papers alongside application of chemical knowledge.

Through manual classification the compound shown in Figure 2.20 was classified as compound type thiourea and compound subtype phenylthiourea. In the case of this compound it has identified the most significant functional groups in the compound; however, the compound contains alkyl groups, phenyl ring, an ether group and thiourea group in total.

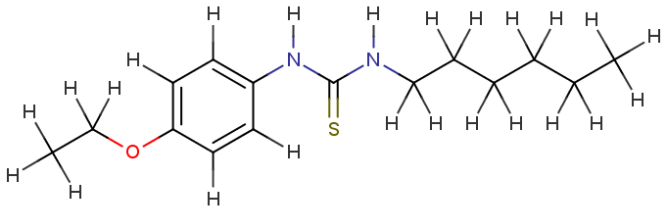


Figure 2.20: Example of classification - Compound 101039_c3sc51023a-14

A disadvantage with the application of this type of classification was that each compound was only assigned to a single group. Compounds frequently contain multiple functional groups, with more complex compounds containing more functional groups. The influential functional group within a molecule may change depending on the biological property of interest so the necessary splitting may not be available if only a single group is assigned to a molecule. It would be possible to manually 'tag' a compound with all of the functional groups that it contained, but this would be very time consuming and generate a large number of possible groupings.

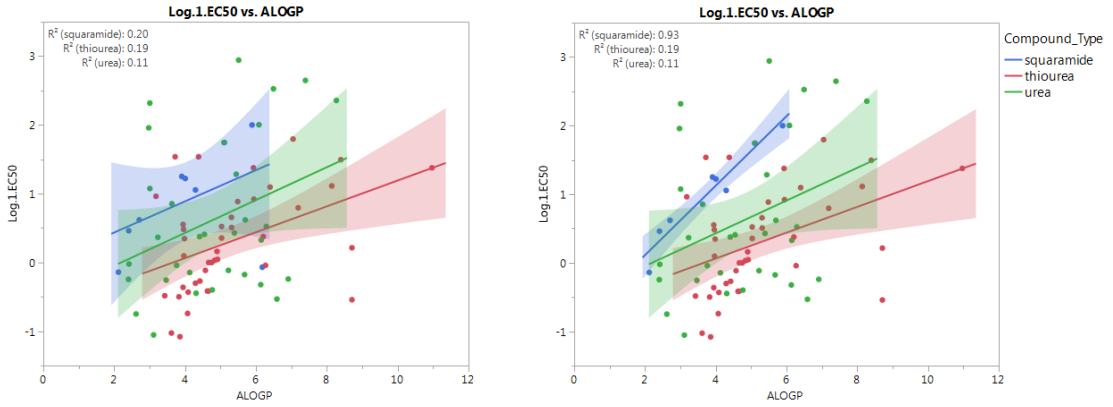
Modelling full dataset with compound types

When used in the expanded subset, the grouping by compound type gave stronger fits than modelling the whole subset in lipophilicity plots. This grouping was also applied to modelling the full dataset of 114 compounds.

The 85 compounds with EC_{50} measurements were split across 5 compound types (pyrrole, squaramide, thiourea, triamide, urea). Of these, the pyrrole and triamide groups only contained 1 and 2 compounds respectively and therefore were excluded.

The remaining 3 groups of compounds were modelled linearly against lipophilicity and splitting was done by compound type (Figure 2.21). From Figure 2.21a, compound NB_quarterly_report_7-13 appeared to be a significant outlier within the squaramide group. The data for this compound was re-examined. Unfortunately there were very few experimental details for this compound as it was not from a published paper but from a report. The text accompanying the measurements suggested this may be a transcription error in the report as it stated *'compounds 11-13 exhibited significant transport ability'* but the value of 13 showed much lower transport than 11 and 12.

This compound was excluded from the linear fit, giving the plot shown in Figure 2.21b, the thiourea and urea groups were unchanged between the two plots. The exclusion of the probable erroneous point gave a significant increase in R^2 of the squaramide group, from 0.2 to 0.93, giving the strongest correlation found so far. The thiourea and urea groups still exhibited low correlation. But these compound groups contained a much



(a) all thioureas, ureas and squaramides (b) compound NB_quarterly_report_7:13 excluded

Figure 2.21: Linear fit of ALOGP vs $\log(1/EC_{50})$ split by compound type - with and without outlier.

wider variety of chemical structures than the squaramide group, and likely required further splitting beyond the thiourea/urea group.

Quadratic fits were also examined as the wide range of log P values would be expected to span an optimum log P value, after which activity decreases, see Figure 2.22. In these fits the squaramide group showed a strong fit, but thiourea and urea were still very poor fits, with the urea group even exhibiting an inverted parabola. Table 2.4 shows the fit statistics. Although the squaramide group showed a strong fit it was lacking compounds with high log P values, making it impossible to determine whether a quadratic or linear fit was more appropriate.

	Linear fit for ALOGP				Quadratic fit for ALOGP			
	all compounds		with exclusion[a]		all compounds		with exclusion[a]	
	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Squaramide	0.2	0.085	0.93	0.919	0.536	0.381	0.941	0.917
Thiourea	0.189	0.168	-	-	0.204	0.162	-	-
Urea	0.112	0.082	-	-	0.145[b]	0.086	-	-

[a] Excluding compound NB_quarterly_report_7-13, only affects Squaramide fits
[b] Exhibits inverted parabola

Table 2.4: Fit statistics for full dataset, split by compound

Two parameter models (ALOGP and MW) and three parameter models ($ALOGP^2$, ALOGP and MW) were also investigated for the full dataset with compound splitting. These fits were largely the same as the ALOGP linear/ ALOGP quadratic fits respectively and offered no significant increase in R^2_{adj}. (See ESI [30])

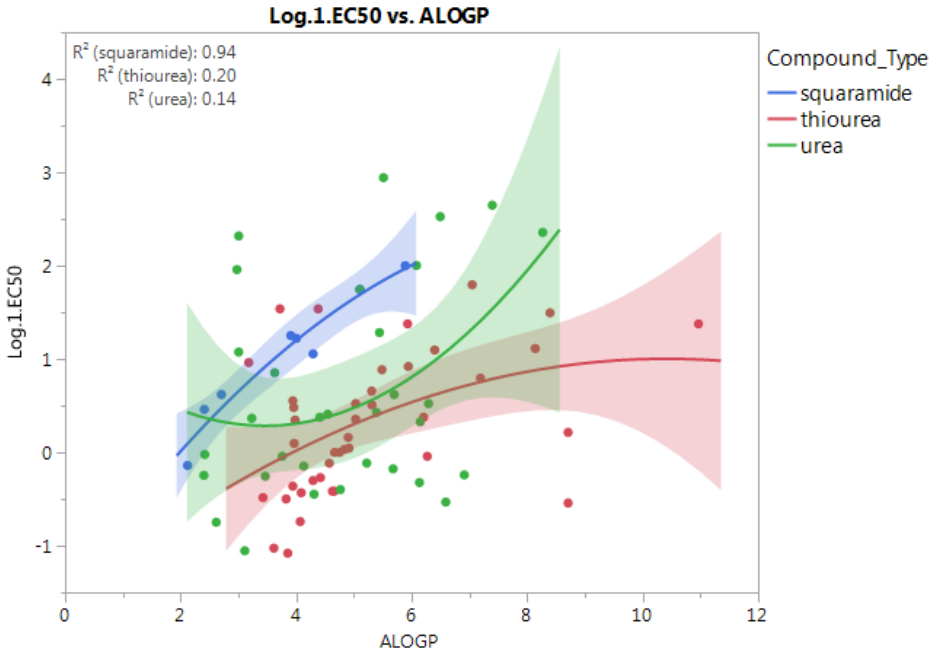


Figure 2.22: Quadratic fit of ALOGP vs log(1/EC_{50} split by compound type, excluding outlier

The compound subtype was also investigated but was not beneficial for splitting the dataset in these models as it created too many groups which only contain a few compounds. From the 22 compound subtype groups only 7 contained more than 3 compounds with EC_{50} values. Additionally because of the manual method of classification some compounds may have been assigned to a group which only identifies one of its key structural features.

When examining the distribution and fits of compounds, particular with compound type and subtype, easy access to the structures of the compounds would have been very useful to allow comparison or structural features and application of chemistry knowledge. Section 4.3 shows some investigation carried out into the use of visualisation in analysis.

Grouping by physical action

The action of chemical entities in biological systems are very complex and are controlled by a myriad of interrelated interactions. Even within the simplified experiments they have interactions with solvents, ions and the different parts of the lipid bilayer. The complexity of these interactions are not covered by the calculated descriptors, but some aspects of these interactions have been investigated in the transport experiments.

In addition to information relating to the magnitude of anion binding and anion transport the papers also contain further experiments relating to the methods and types of transport and binding. All papers were re-examined and additional information was extracted for potential use in classifications. Although the information included in each

paper varied, the following information was extracted: Evidence of Cl^-/NO_3^- antiport, Evidence of H^+/Cl^- symport, Evidence of additional transports, Type of mechanism (and method to justify), Binding Mode (H-NMR) and Binding Mode (crystal).

Following this additional data collection it could be seen that while most compounds exhibited the same features (Cl^-/NO_3^- antiport mechanism, acts as a mobile carrier, 1:1 binding mode) there were some compounds which differed from the norm. These included: presence of additional transport mechanisms, inconclusive results to show mobile carrier mechanisms and 1:2, 2:1 or unclear binding modes.

The disadvantage of a method like this was that it required access to significant amounts of experimental data, which must be determined for each compound and cannot be directly computed. Additionally it was time-consuming to extract the extra data and not all papers contained the same experiments.

This information may not be able to be directly included in a model for classification; however, it would be useful in conjunction with the model building to flag potential outliers or compounds with behaviour that may exclude it from a model, for example: if a compound has a 1:2 binding mode it should require half the concentration of transporter to move the same amount of Cl^- as a 1:1 binding molecule.

2.4.4.2 Automatic Classification

Another avenue of interest for classification of the compounds is utilising automatic classification of the compounds, to avoid the necessity of manual classification. This would minimise the amount of time taken as well as reducing the risk of human error in classifying compounds.

Two methods were found that could potentially achieve this, one method which uses the counts of functional groups in a molecule, and a second which assigns a compound group based on a hierarchy of structural features.

The first method utilises a class of descriptors called 'Functional group counts' within DRAGON. These include 153 counts for the presence and number of various functional groups, 'Ring descriptors' could also be used giving another 14 descriptors. These counts could be used directly in a model by themselves e.g. number of 6-membered rings or grouped together to form another descriptor e.g. no. of primary, secondary and tertiary amides could be combined to number of amides. Although DRAGON calculates lots of functional group counts it does not contain separate groups for all possible features, e.g. thioureas and ureas are grouped together in a single count and no count exists for a squaramide.

The second method involves assigning the molecules to a group using a newly developed program Classyfire [27]. Classyfire is a web-based application which uses a rule-based approach to carry out automated structural classification for chemical entities. Hierarchical classification is carried out through use of a computable chemical taxonomy called ChemOnt [122].

As a comparison to the manual classification the compound c3sc51023a-14 (Figure 2.20) was assigned to the following categories through ClassyFire: N-phenylthioureas, Phenoxy compounds, Phenol ethers, Alkyl aryl ethers, Thioureas, Organonitrogen compounds, Hydrocarbon derivatives. This identifies the thiourea and phenylthiourea groups from the manual classification but also assigns classifications for other structural features which may be of use when grouping diverse compound sets.

Although the prospect of carrying out automated chemical classification was appealing, when the anion transporter dataset was processed by Classyfire just under half of the compounds did not return a classification, making it somewhat useless for the purposes of model building. This may have been a consequence of the program still being in development at the time of investigation. Improvements in the functionality would be expected as the program use is expanded.

The classification that was carried out in ClassyFire produced a much larger number of categories than a manual classification. The compounds that were correctly processed were assigned to 61 different classes, but from these classes 15 of them only contained a single compound and 32 contained 3 or fewer compounds. Access to the full hierarchy would be wanted to enable selection of a suitable level of classification to get the right balance between number of groups and the population of the group. With improvements to the compound processing this could be a promising method of classifying compounds for analysis.

2.4.5 Dimensionality reduction

PCA

Due to the large number of descriptors present in the dataset and the amount of processing power required to model greater than three descriptors from all combinations dimensionality reduction was investigated. This was carried out using PCA which reduces the number of variables by producing new linear combinations of the variables.

PCA was carried out with both 2D descriptors and 3D descriptors from DRAGON to examine if further investigation was needed into creation of 3D structures. The use of descriptors generated from 3D structures had previously been excluded due to the time required for computation of an accurate conformer.

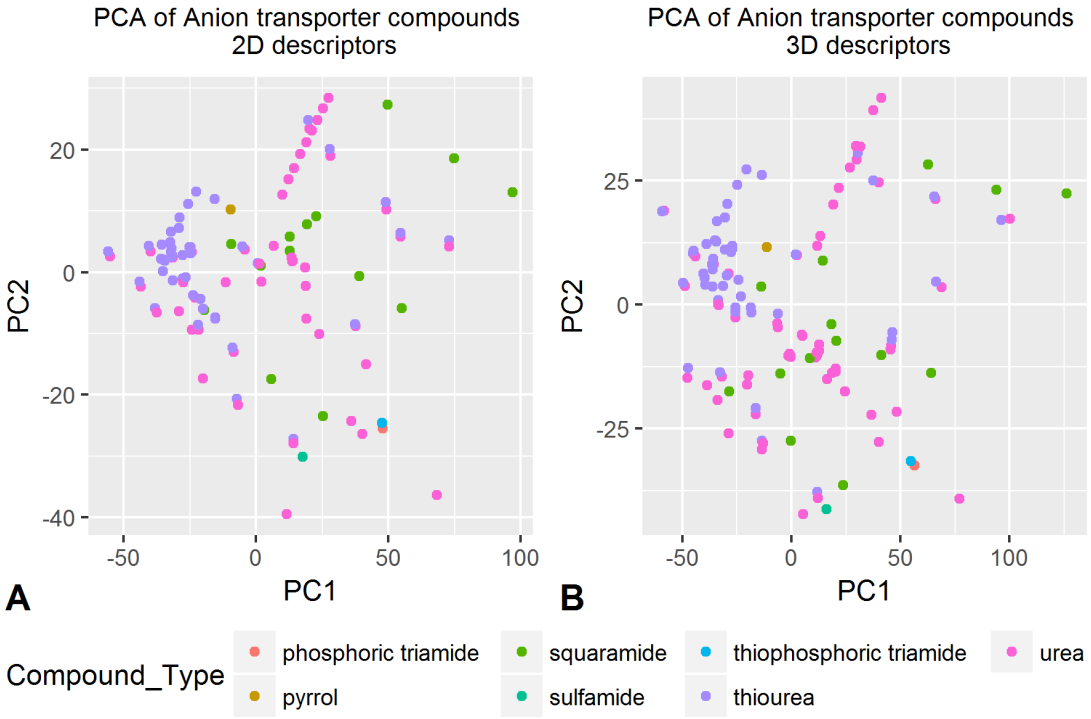


Figure 2.23: PC2 vs PC1 for 2D and 3D descriptors - coloured by compound type

3D structures were obtained through use of gen3d in OpenBabel [123,124] which carries out a geometry optimization for a single conformer. This is not as powerful and accurate as other molecular dynamics simulations; however, it is significantly quicker and provides a reasonable approximation of the 3D structure. If further investigation was required with 3D conformers then additional simulation should be carried out.

Descriptors were generated in DRAGON to give a set of 2D descriptors and a set of 2D & 3D descriptors. All descriptors excluding charge descriptors were generated. Subsequent cleaning excluded descriptors if they were constant, near constant or contained missing values. For 2D structures 1914 descriptors were exported, for 3D structures 2996 descriptors were exported.

PCA was carried out using prcomp in R [125] with the full cleaned descriptor set. Using 3D descriptors the cumulative explained variance by PCs 1 & 2 was 58.7% , and 65.9% for PCs 1, 2 & 3. Using 2D descriptors it gave a cumulative explained variance of 60.1% for PCs 1 & 2, and 68.1% for PCs 1, 2 & 3. This showed that a large proportion of the variance could be explained by a small number of PCs.

A comparison of the plots generated from the first two PCs for the 2D and 3D descriptors can be seen in Figure 2.23. This plot was coloured by compound type to see if this identified any features or clusters in the data. The two plots show largely similar distributions, although the 3D plot shows a slightly wider distribution.

The generated PCs were used in the construction of a linear regression, modelling for $\log(1/EC_{50})$. While over 100 PCs were generated, only the top 3 were selected for use in the regression in the form: $log(1/EC_{50}) = aPC1 + bPC2 + cPC3 + d$

The regression fits obtained show almost no correlation to $\log(1/EC_{50})$. For 2D descriptors it gave R^2=0.1498 and R^2_{adj}=0.1224, for 3D descriptors R^2=0.1578 and R^2_{adj}=0.1307. This showed no improvement for correlation by utilising the 3D descriptors. Additionally neither of these models would be used as only the PC1 variable was considered statistically significant. (See ESI [30] - PCA)

As PCA is an unsupervised method it was not completely unsurprising to find little correlation in the models. An alternative method that could be employed is PLS, which uses correlation with the response variable in the selection of the new descriptor variables. PCA could also be repeated with pre-PCA variable reduction using domain specific knowledge to reduce the number of descriptors in the input. Additionally, the PCA was performed on compounds which did not contain EC_{50} values, these could be removed from the dataset.

While PLS may potentially produce a better fit for the data both of these dimensionality reduction techniques remove interpretability from the models by creation of the new variables. This reduces the amount of insight that can be made directly from the models. As the 3D descriptors did not provide a large improvement to the fit the calculation of more accurate 3D structures for descriptors was not re-visited.

2.5 Discussion & Future work

2.5.1 Data Extraction

From examination of the Gale group data a number of important lessons have been learnt about data, in particular about the collection and storage of data so it can be re-used at a later stage. It is especially important to store data in an easily accessible format, ideally computer readable, and with sufficient information associated with it. Such that another researcher or collaborator would be able, with the correct software, to attempt to reproduce the analyses. Mining data from publications is 'lossy' as so much raw data doesn't get into the final paper or is presented in an unusable format such as images.

It is often difficult to include enough supporting data to ensure that results will be unambiguous at a later stage, or when examined by someone else. Important things to include are; clear numbering of compounds with unique and unambiguous identification, where the descriptors were obtained and details of which programs and/or models were used to obtain results.

The initial part of the investigation focused on the extraction and collation of data for synthetic chloride ion transporter molecules. Data was extracted for 131 compounds and compiled to a database to allow further analysis to take place on it. Following removal of duplicates this produced a high quality and well-curated dataset containing 114 anion transporter compounds. From these compounds 85 had EC_{50} values.

As new compounds are synthesised it would be beneficial to incorporate them into the database giving more datapoints to use in analysis. If manual extraction was required this could be a time consuming process; however, working with the scientists synthesising the compounds may allow the data to be formatted in a more computer friendly format at the point of creation.

2.5.2 QSAR Analysis

QSAR analysis was carried out on the dataset, following the generation of 2D descriptors, attempting to model the anion transport ability of the full dataset of transporters. Although a number of attempts were made through different methods no linear regression model could be created which produced a statistically robust model for the entire dataset.

Simple QSAR analysis of the full dataset using MLR generated through a fit-all process did not produce strong fits for the EC_{50} observations, with the best 3-parameter MLR only generating a model with an R^2 value of 0.38.

Following a stepwise approach for model selection a number of models were generated with higher R^2 values. However, these models contained parameters which were not statistically significant, multiple parameters that were highly correlated and too many parameters in relation to the number of observations contained in the dataset. Removing the cross-correlated and non significant variables gave an 8 parameter model with an R^2 value similar to the 3 parameter MLR and also produced new non-significant variables. The difficulty in obtaining a statistically valid model for the whole dataset via a linear regression suggested that the dataset may be better modelled in subsets.

Splitting the dataset into groups was initially examined using the dataset from a single paper [91]. This subset was expanded through molecular similarity from 22 thioureas to 36 compounds (27 thioureas and 9 ureas). However, expansion of the subset produced weaker fits than the original thiourea set.

All of the subset models focused around lipophilicity which is a major component of anion transport; however, no significant correlation was found for the expanded subset either via a linear or quadratic fit to ALOGP. In all models for the subset, splitting the dataset by compound type produced stronger fits for the thiourea and urea groups

separately. But the urea compounds didn't fit well to any model as they exhibited a very narrow range of $\log(1/EC_{50})$ values.

2.5.3 Classification methods

Multiple methods for grouping compounds were examined. Classification through compound group was the most straightforward to carry out; however, this relied on the manual classification by a scientist. Using a rule based method would be a more rigorous approach; however, these methods require further investigation before they can be implemented.

When modelling with compound group across the full dataset the squaramide group gave a very strong correlation to ALOGP although this group was limited to low log P values only. (Figure 2.22) The other compound type groups were largely uncorrelated with ALOGP. This was likely due to the presence of more uniformity in the squaramide group compared to the full thiourea/urea groups as these parent groups also included bis-compounds and a wide range of aromatic and non-aromatic groups.

The other classification methods need further investigation. Classification of physical action from underlying data would be useful in flagging any compounds that have an unusual action which may distort their activity response; however, it required time-consuming extraction of additional data. Automatic classification methods through ClassyFire were very promising as a rule based classification system but the program was not functioning optimally at the time of investigation. Classyfire compound assignments should be revisited once the program is fully released. These classifications could be used in conjunction with expert knowledge to create new categories which replace the compound subtype assignments used previously.

2.5.4 Further Expansion

The knowledge that was gained from studying the tambjamine dataset (Chapter 3) could be used to explore new avenues with the Gale dataset. The Gale set was a larger dataset than the tambjamines which provided more datapoints; however, it also had a more diverse range of chemical compounds and structures which increased the complexity of the data. These datasets have been combined into a large transporter dataset (Section 3.9) which will hopefully provide a base for further modelling.

The mixed effect model methods (lmer) that were utilised with the tambjamine data could have potential applications within the larger set of anion transporter data. These models used the whole dataset to fit the parabola and substituent grouping to adjust the parameters. Although the Gale dataset does not have the uniformity that was seen in

the tambjamines, as there are many more potential substituent positions, the different chemical groups assigned through classification could potentially be used to split the lmer model instead of the substituent types.

Although development of these classification models may not facilitate the accurate prediction of unknown compounds they should help to provide further insight into the development of potent anion transporters.

Chapter 3

Tambjamine Anion Transporters

3.1 Background

The work shown in this chapter centered around the analysis of an additional set of anion transporter molecules, separate from those examined in Chapter 2. The work carried out in this chapter was in collaboration with Roberto Quesada[1] and members of his group, who synthesised the molecules examined.

Quesada and his group have been researching the 'underexamined' class of molecules, tambjamines, for the purpose of developing synthetic molecules with good anion transport ability and potential cytotoxicity. These molecules have potential applications in similar areas to those compounds researched by the Gale group due to their ability to effectively transport chloride ions, see Section 2.1

Tambjamines were discovered as marine natural products, isolated from bryzoans, nudibranchs and ascidians. [126–128] They are characterised as having a 4-methoxy-2,2'-bipyrrolenamine structure (Figure 3.1). This is structurally similar to Prodigiosine and Prodigionines, compounds known to exhibit antimicrobial and cytotoxic properties [129–131]. The properties of naturally occurring tambjamines have been investigated, with some compounds also exhibiting cytotoxic effects. [127, 132]

Prior to the development of synthetic methods for the production of the bipyrrolic aldehyde precursor [133], little research had been carried out on synthetic tambjamines as they were not easily synthesised. Following the discovery of an accessible synthetic pathway many more tambjamines could be synthesised via the acid catalyzed condensation of the 4-alkoxy-2,2′-bipyrrole aldehyde and the corresponding amine; shown in Figure 3.2 [134]. This allowed further studies into the synthesis and analysis of tambjamines, which showed promising results for chloride anion transport. [135, 136]

[1]Departamento de Química, Facultad de Ciencias, Universidad de Burgos, 09001 Burgos, Spain Email:rquesada@ubu.es

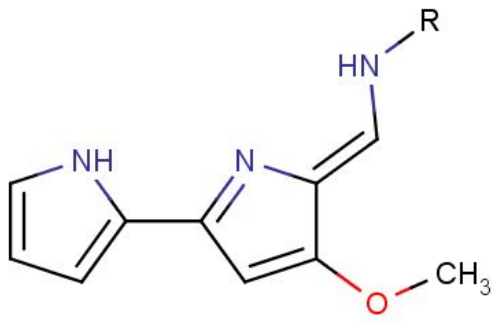


Figure 3.1: Backbone structure of the naturally occurring Tambjamines

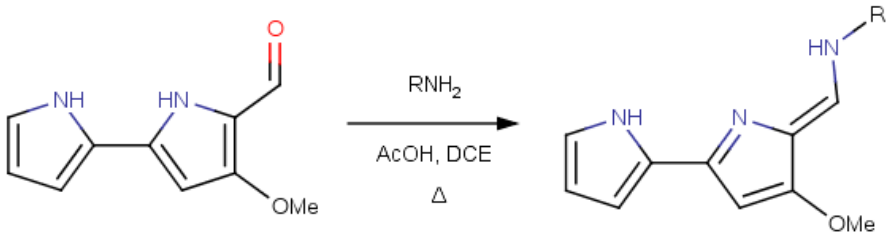


Figure 3.2: Reaction for synthesis of tambjamine analogues

Quesada's group created a range of tambjamine derivatives with varying substitution patterns through use of these synthetic pathways [137–139]. The naturally occurring tambjamines all possess alkyl substituents in the R position (Figure 3.1); however, in their previous studies [137,139] Quesada and colleagues discovered that placing aromatic substituents in the R position neighbouring the nitrogen can produce compounds that outperform natural tambjamine derivatives, exhibiting higher chloride transport rates.

Natural tambjamines also all possess an OMe substituent on the pyrrol ring; however, in the studies [138, 139] an additional series of synthetic molecules were created with a benzyloxy group (OBn) in that position, creating a more diverse collection of molecules for analysis.

In particular, Quesada's group has been examining the chloride ion transport abilities of the molecules in vesicles and the cytotoxicity of the molecules in cancer cell lines. The transport activities were examined in relation to the initial rate of chloride efflux k_{ini}. Outside of these papers some attempts have also been made to mathematically model the chloride ion transport ability of the molecules.

The compounds included in Figure 3.3 were the compounds initially synthesised by Quesada's group (details of the synthetic methods can be found in their papers [137, 139]) and a series of transport experiments were carried out in vesicles following the same procedures as the Gale group experiments, see section 2.1.3. The experimental variables collected and calculated were; EC_{50} (NO_3^-/Cl^-), Hill parameter (n), initial rate of chloride release (k_{ini}) and Retention Time (RT).

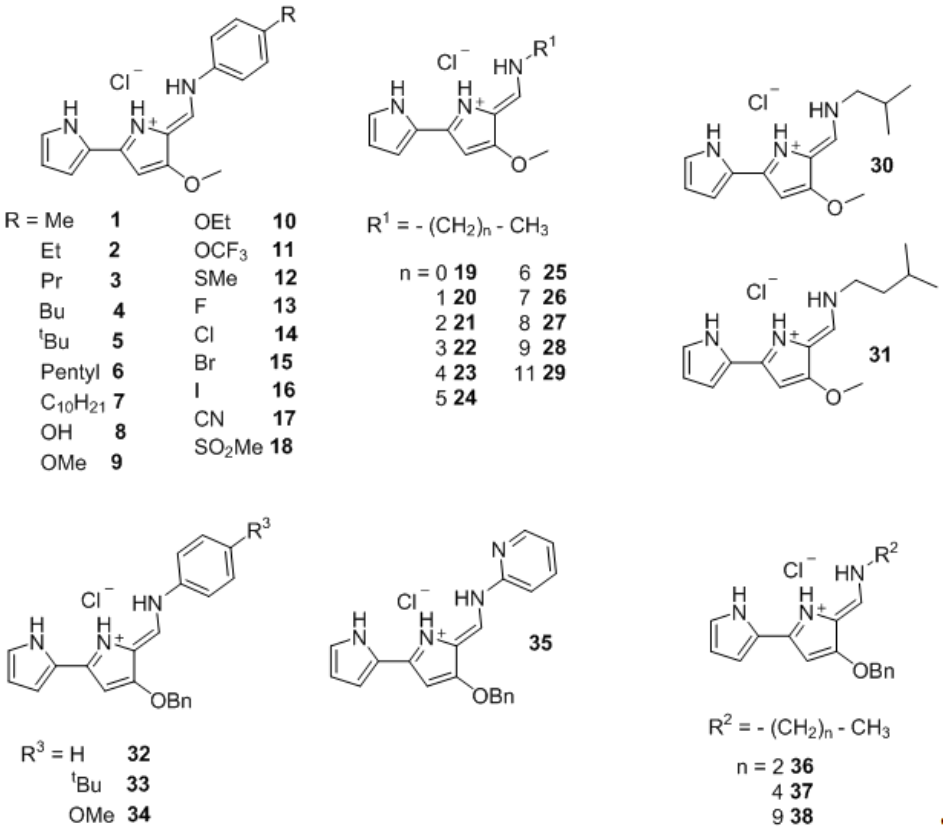


Figure 3.3: Structures of previously synthesised tambjamine derivatives

In addition to the measured variables, 507 chemical descriptors were calculated for the compounds from their molecular structures; descriptors were obtained through ALOGPS2.1 [3], E-DRAGON 1.0 [9], Chemicalize.org [5], ACDiLabs 2.0 [2], TorchV10lite [13] and ChemDraw 12.0 ultra [4] software. This range of descriptors were cleaned, removing any descriptors containing no variance, non-numeric values or missing values. Following cleaning, their dataset contained 380 descriptors which were used to create a selection of models.[2]

Throughout this work the aim was to produce models that could help identify the features in a compound that make a good anion transporter.

[2]Full descriptor sets for the initial data and their sources can be found in the ESI under Initial Models [30]

3.2 Evaluation of existing models

3.2.1 Outline of initial models

The tambjamine compounds (See Figure 3.3) synthesised by the Quesada group were used in QSAR modelling to build models predicting the $\log(1/EC_{50})$ values for the compounds as a measure of anion transport efficiency. The process chosen for model building and validation by the Quesada group employed a training/test set split. The initial dataset containing 38 compounds was split into a training set (32 compounds) and a test set (6 compounds - 6, 8, 20, 26, 32, 36). Multiple regression analysis using the fit-all method was performed on the data for the training set compounds using JMP 9.0.0 [21], modelling $\log(1/EC_{50})$ against the 380 descriptors previously obtained, to generate the best possible models for the dataset.[3]

The models produced in the initial fit-all process contained a maximum of 3 terms and were ranked according to best fit (determined by R^2 value), the highest ranked models included a variety of different descriptors, including the following; ALOGPs, ALOGPs-sq, nH, LogD(pH7.2)-sp blood, Fraction unbound in plasma, AMW and pKa.[4]

Table 3.1 contains a list of linear regression models for $\log(1/EC_{50})$ that were considered for further testing by Quesada's group. These models were selected using the results from the fit-all analysis along with input from the scientists. The models favoured descriptors which were readily understandable, as well as models containing RT to compare this to the ALOGPs models. RT is often used as an indirect measure of the lipophilicity (log P - ALOGPs). From the selected models Model 4 and Model 6 were excluded from further analysis, as they were not statistically valid. Statistical validity was determined through p-values, with a threshold of 0.05.

Model Name	**Parameters used**
Model 1 / Eq 4	ALOGPs-sq, ALOGPs, nH
Model 2	ALOGPs, ALOGPs-sq, nH, TSA
Model 3	ALOGPs, ALOGPs-sq, pKa(enamin a)
Model 4	ALOGPs, ALOGPs-sq, pKa(enamin a), TSA [a]
Model 5 / Eq 5	RT, RT-sq, nH
Model 6	RT, RT-sq, nH, TSA [a]
Eq 6	ALOGPs-sq, ALOGPs, AMW
Eq 7	RT-sq, RT, AMW
Eq 8	LogD(pH7.2)-sq, Fraction Unbound in Plasma, AMW

[a]Marked as not statistically valid

Table 3.1: Models selected for further testing in initial evaluation

[3]This dataset can be accessed in the ESI - under Initial Models
[4]Fit-all results can be found in ESI - under Initial Models

3.2.2 Discussion of model process

The process that had been carried out by the Quesada group to obtain these models was examined, evaluating the reproducibility of the models. In addition to this the underlying data which was obtained from the tambjamine compound structures were investigated.

3.2.2.1 Model selection

In the selection of models a lipophilicity descriptor was always included as it has been shown to be of key importance [138]. This follows similar findings for other anion transporters [91]. Due to the forced inclusion of a lipophilicity parameter only Model 1 and Eq 8 were selected from the fit-all models[5] where they were among the highest ranked 3 parameter models. The other top models from the fit-all were not selected as the preference was to include parameters which were easily understandable.

The additional models were selected using the scientists knowledge, and selection of parameters that were readily understandable. This was a reasonable approach to the selection of descriptors and gave models that could be interpreted in terms of the chemical/physical interactions.

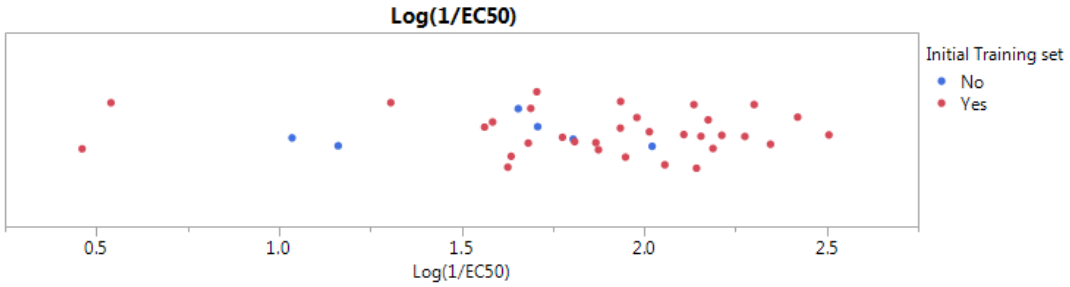
Model 5/eq 5, Model 6 and Eq 7 were not preferable for a final model to predict the transport efficiency as they utilised RT as a parameter. RT is an experimentally measured value, rather than a calculated value and, therefore, the activity in the model could not be predicted without synthesis of the compound. However, it was included as a parameter to allow comparison of the results to the ALOGPs model with the same additional parameters.

The number of parameters included in the models were an acceptable level. There were 32 compounds in the training set and a maximum of 4 parameters in the models. This gave a minimum ratio of 8 observations per variable. Increasing the number of parameters beyond this would lead to potential overfitting.

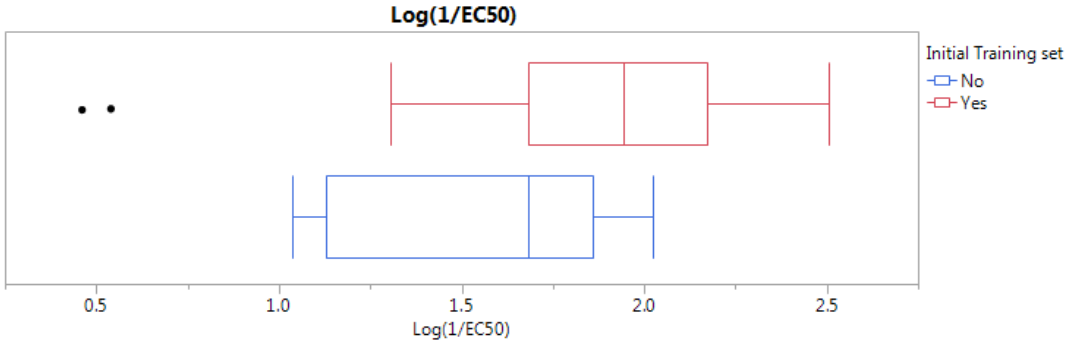
3.2.2.2 Test set distribution

During the creation of the models the initial tambjamine dataset was split into a training set and test set, which is a common technique used in QSAR for validating models. [46] Although the 32-6 split between the training and test set was a suitable split for a dataset of this size, more attention should have been paid to the distribution of the test set within the main dataset, to give a good representation of the dataset.

[5] Available in ESI



(a) Scatter plot



(b) Box Plot

Figure 3.4: Distribution of the $\log(1/EC_{50})$ values for the initial tambjamine compounds

The distribution of the $\log(1/EC_{50})$ values for the compounds can be seen in Figure 3.4. This plot shows that the spread of $\log(1/EC_{50})$ values is quite large; however, the test set compounds, marked in blue, are concentrated in the center half of the $\log(1/EC_{50})$ range. None of the test set compounds have a $\log(1/EC_{50})$ value in the low or high ends of the distribution. The distribution of this test set was not very representative of the dataset distribution.

If a test set/training set split is selected as the method implemented then a stratified selection method should be utilised as this would ensure that a proportionate number of compounds were selected from the low and high parts of the $\log(1/EC_{50})$ range. Due to the presence of ALOGPs and ALOGPs-sq descriptors in the models developed so far, the lack of points in the test set at the ends of the $\log(1/EC_{50})$ range could cause highly leveraging points to distort the statistics in the evaluation of the models. Figure 3.7 on page 75 shows an example of how important it could be within this dataset. However, other methods could be employed instead of the training set/test set method, such as k-fold crossvalidation or bootstrapping, some methods are outlined in Section 1.3.

3.2.2.3 Validation

Validation was carried out though a variety of different methods; LS statistics, LMO-CV, external validation and randomisation tests. Model 1, 2, 3 & 5 were tested through

	Model	Model 1 /Eq.4	Model 2	Model 3	Model 5 /Eq.5	Eq.6	Eq.7	Eq.8
	no of param.	3	4	3	3	3	3	3
Int Valid.	R^2	0.86	0.88	0.8	0.8	0.84	0.78	0.84
	R^2_{adj}	0.84	0.86	0.78	0.78	0.82	0.75	0.83
CV_{LMO}	Q^2(CV)	0.7	0.78	0.74	0.65	0.5	0.49	-1.21
Y_{Rand}	R^2	0.14	0.16	-	0.16	-	-	-
	R^2_{adj}	0.05	0.04	-	0.12	-	-	-
Ext - a	Q^2	0.7	0.71	0.62	0.69	0.61	0.71	0.5
Ext - b	Q^2	0.42	-	-	0.88	0.02	0.45	0.31
Ext - c	Q^2	0.45	-	-	-	-0.1	-	0.27
Ext - d	Q^2	-0.27	-	-	0.14	-0.28	-0.1	0.22
Ext - e	Q^2	0.54	-	-	-	0.55	-	0.12
Ext - f	Q^2	-0.2	-	-	0.15	-0.27	-0.08	0.21

Table 3.2: Validation Statistics for initial tambjamine models - obtained from multiple Quesada group reports.

internal, external and randomisation tests, while Eq 4, 5, 6, 7, 8 were tested through internal validation and external validation with multiple test sets. The external test sets were made up as follows: Ext-a (test set from the initial database), Ext-b (test set + 3 additional tambjamine derivatives), Ext-c (Ext-b + prodigiosine), Ext-d(thioureas from Gale paper [91]) and Ext-e (all above compounds).

The work appears to have been carried out by different people as there were many differences between the two reports received. The models were numbered differently with different validation methods selected and the statistics reported used differing terminology. For example; when referring to use of the training set vs. test set mean in the calculation of q^2. With all of these aspects combined it was quite difficult to follow the validation process and reproduce the statistics independently. The validation statistics extracted and combined from the reports are shown in Table 3.2.

From the validation statistics in Table 3.2 a number of observations could be made about the models and the procedure.

- The increase of parameters to 4 in model 2 (ALOGP, ALOGPs-sq, nH, TSA) did not provide much increase in the predictive ability of the model relative to the 3 parameter model 1 (ALOGPs, ALOGPs-sq, nH), suggesting the inclusion of a 4th parameter was unnecessary.

- The results of randomisation tests showed that the fits were not generated due to chance correlations.
- The selection of thioureas for use in an external test set was inappropriate as the compounds had wildly different structures and activities to the tambjamine compounds. These compounds were outside of domain that the models had been built for. Therefore it was not surprising that the q^2 values obtained for this set were very low
- The LMO CV was not repeated enough times to be a true representation of the dataset. These values should be treated as estimates.

3.2.2.4 Difficulties in analysis

Sections of the work had been carried out by multiple different people who each had their own way of recording the information and processes, along with different terminology. As such it was often difficult to determine exactly which datasets, descriptors and methods had been used for the analysis.

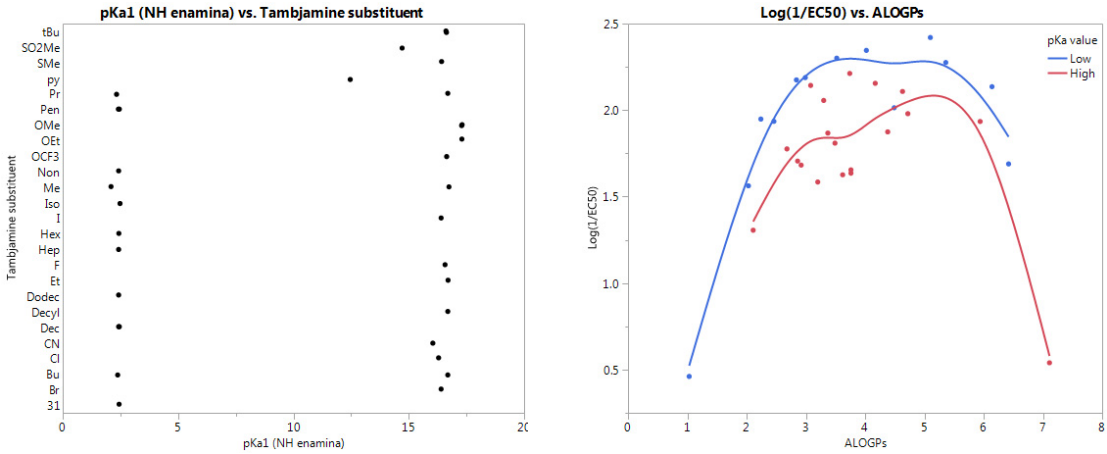
Conflicting reports and datasets existed containing different splits for training and test set along with varying models. It was difficult to try and determine which statistics referred to which models and the compounds that these were modelled on. A number of these errors were generated through human error or accidental transcription.

In addition to a number of errors that were encountered within the reports many of the programs used were in Spanish, as the researchers were from Spain. Spanish terminology was sometimes used in the names and descriptions of items. This caused added difficulty in trying to identify descriptors, or establish if two items were actually the same but with the English and Spanish names.

3.2.2.5 pKa distribution

Alongside the investigation of the model process, a number of descriptors used in the models were examined, including their correlations and distributions. One variable of interest from Quesada's models was pKa(enamine a), which was a calculated pKa value generated through Chemicalize.org [5]. Initial examination showed that almost all of the calculated pKa values were very clearly split between a high pKa region (16-18) and a low pKa region (2-3) with only 2 compounds lying outside these ranges, see Figure 3.5. When plotting the ALOGPs against the $\log(1/EC_{50})$ it also appeared to give a split to the distribution.

However, it was discovered whilst calculating the additional pKa(enamine a) values for the test set (through Chemicalize.org) that the pKa values stated in the Quesada



(a) Split in calculated pKa

(b) Split in correlation

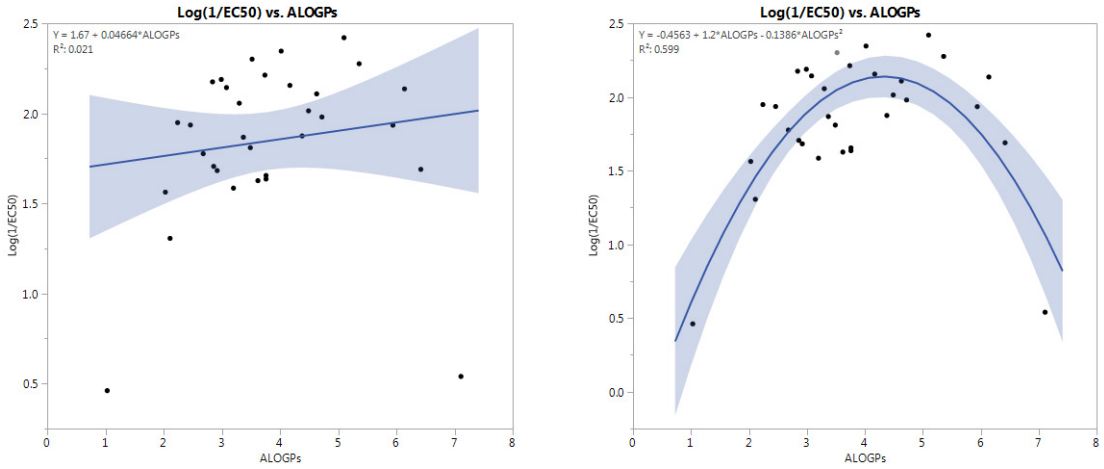
Figure 3.5: previous pKa (enamine) distribution for the initial tambjamine test set

dataset were incorrect. Many of the compounds did not give a calculated pKa value for the position specified (NH enamine) and other compounds which did return a value had differing values to those in the dataset. After recalculating and correcting the pKa values the distribution showed no significant splitting, with only 2 compounds having a pKa value below 16. Since many of the compounds did not return a value for the pKa it would not be a suitable descriptor for use in a model. Models using pKa as a parameter were excluded from further consideration. This also highlighted the need to have access to data and methodology to allow results to be reproduced and checked by other researchers.

Prior to the error being discovered in the pKa values a split was found in the ALOGPs vs $\log(1/EC_{50})$ plot for the two "sets" of pKa enamine values (Figure 3.5). Although it is unknown how the error in the pKa values arose, when possible causes were examined it was noted that the two groups had differing structures. The 'high' group was solely comprised of compounds with a aromatic group neighbouring the enamine. The correlation between the pKa values and the structural differences inspired another avenue of investigation which, despite the initial incorrect prompt, produced good results. This is discussed in Section 3.6 - 'Classification of Compounds'.

3.3 Synthesis of new tambjamines

Following preliminary evaluation of the Quesada anaylsis, the dataset and the distributions of data were re-examined. In the models that Quesada's group had produced so far, the most frequently selected descriptor was ALOGPs, a lipophilicity descriptor.



(a) Linear fit (b) Quadratic fit

Figure 3.6: Comparison of linear and quadratic fits for initial tambjamine training set no exclusions

This was not particularly surprising as it has often previously been noted that biological activity is related to log P [34, 35].

The lipophilicity descriptor selected by Quesada in the models was ALOGPs, as it exhibited the best correlation to the experimentally determined RT measurement, which is frequently used as an indirect measure of the lipophilicity.[6] Examining the correlation of $\log(1/EC_{50})$ and lipophilicity (ALOGPs) suggested the presence of a parabolic rather than linear relationship, Figure 3.6 shows the comparison, which agreed with previous findings [138] and the presence of ALOGPs and ALOGP-sq in the models.

However, modelling the parabolic fit of ALOGPs against $\log(1/EC_{50})$ indicated the presence of a number leveraging points in the fit - Compounds 7 and 19 from Figure 3.3. The points are marked on Figure 3.7 with an 'X'. These points are highly leveraging as inclusion, exclusion or error in these points could have had a large effect on a model fit. This was a factor which could have exacerbated difficulties in validating the models.

To reinforce whether the relationship between ALOGPs and $\log(1/EC_{50})$ was parabolic or linear it was deemed necessary to obtain more datapoints with ALOGPs values in the regions lacking in data. These regions were an ALOGPs value of 1-2.5 and an ALOGPs value of 6-7.5.

A number of tambjamine compounds, with similar structures to those already synthesised, were 'created' in ChemDraw and their ALOGPs values calculated using ALOGPS2.1.[7] In particular there was a gap in the structures between compound 6 (R = Pentyl) and compound 7 ($R = C_{10}H_{21}$) where there were a number of possible compounds that could be synthesised. In total 95 compounds were created 'in silico'. The

[6] Additional lipophilicity correlations can be found in the appendix - Figure B.1
[7] Work carried out in collaboration with summer student Ziyan Zhao

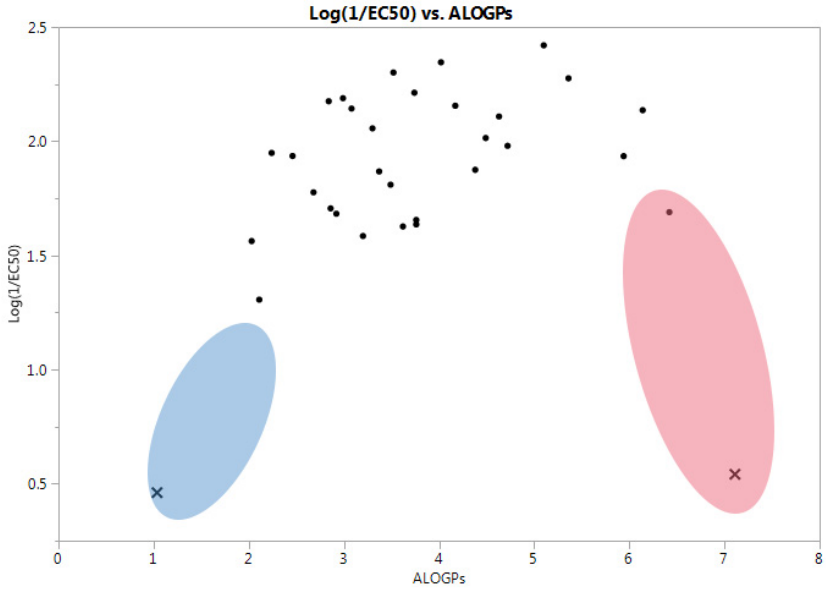


Figure 3.7: Plot of $\log(1/EC_{50})$ vs ALOGPs for initial tambjamine training set - highlighted regions lack datapoints

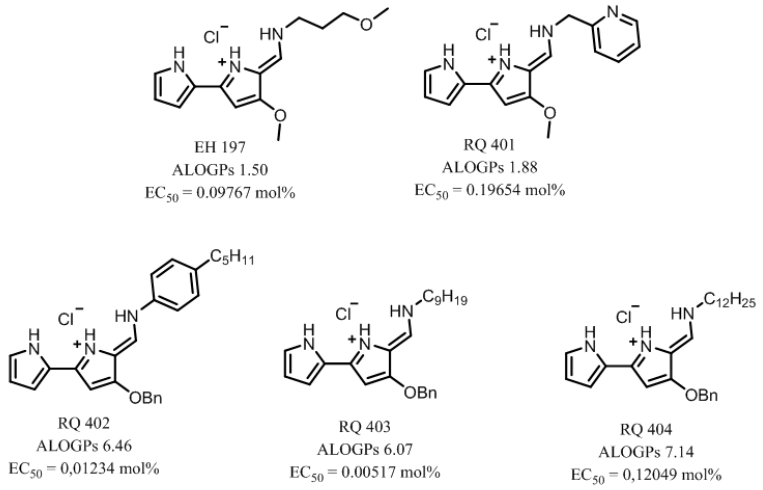
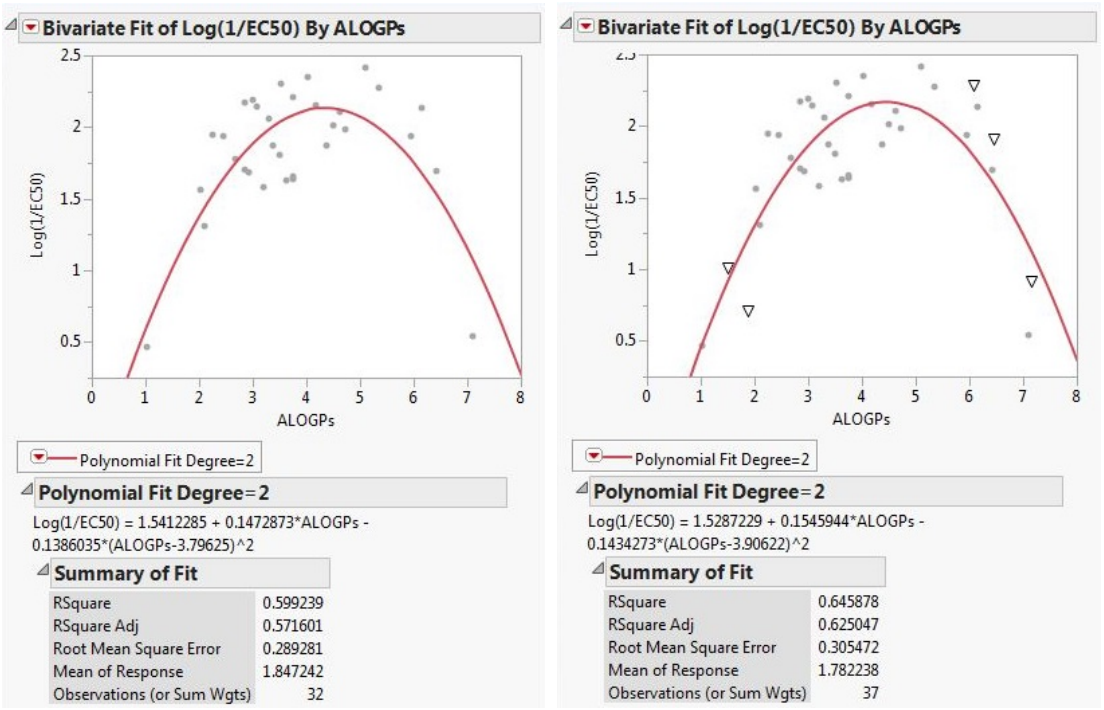


Figure 3.8: Additionally synthesised tambjamine derivatives

ALOGPs values were calculated and filtered down to the range of desired ALOGPs values. A total of 24 possible compounds were sent to Quesada's group with the suggestion that 5 additional compounds should be synthesised, split across the ranges of ALOGPs. The 5 compounds that were synthesised are shown in Figure 3.8 along with their measured EC_{50} values and calculated ALOGPs values.

The plots in Figure 3.9 show the parabolic fits of the initial dataset (training set) and the new dataset (training set + newly synthesised compounds). As predicted the measured EC_{50} values for these new compounds reinforced the observation that the relationship



(a) Initial training set only (b) Training set including new compounds

Figure 3.9: Fit Models using ALOGPs and ALOGPs-sq

between the ALOGPs values and $\log(1/EC_{50})$ is a parabolic one. If further synthesis were possible it would be beneficial for additional compounds to be made with ALOGPs values above 6.5 as this area was still sparsely populated.

Following the synthesis of these new compounds the full set of tambjamines contained 43 compounds. The structures of these can be seen in Figure 3.10. Note the numbering was updated by the Quesada group to include the new compounds in sequence of their structures.

3.4 Generation of descriptors

Additional descriptors were required following the synthesis of the new compounds. The generation of descriptors was carried out in the same fashion as for the Gale group compounds (Section 2.3). This included the creation of molecular representations, generation of descriptors in DRAGON and removal of constant and non complete descriptors.

In addition to the descriptors generated in DRAGON, ALOGPs values were calculated using ALOGPs2.1 and a selection of descriptors were provided by the Quesada group from ACDiLabs2.0 and TorchLite. In total 331 descriptors were selected following cleaning.[8]

[8]Dataset can be found in ESI - Tambjamines_dataset_cleaned.csv

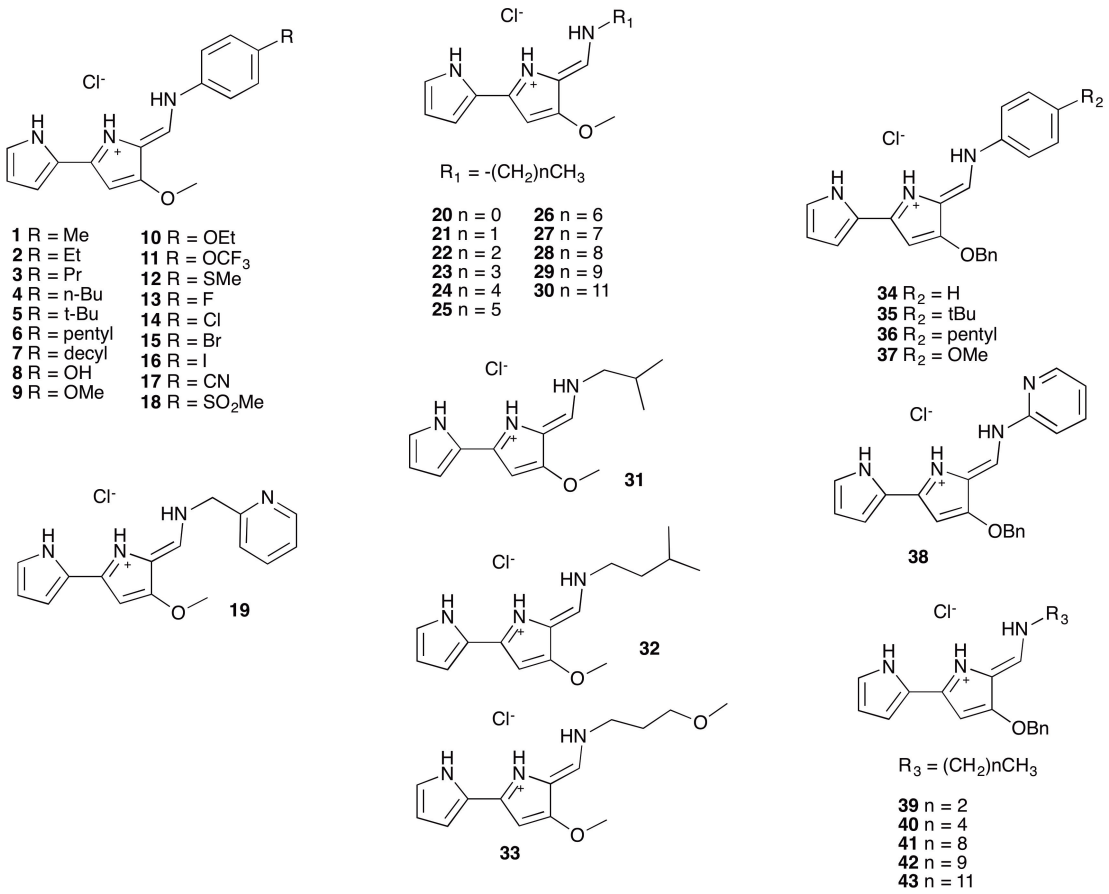


Figure 3.10: Structures of tambjamines including newly synthesised compounds - reordered numbering

3.5 Modelling whole dataset

Following the synthesis of the new compounds, models were re-investigated for the dataset. These models built upon the knowledge gained from examining the previous models; however, due to the difficulties encountered in following the procedures and reproducing the statistics, new models were created.

In the initial models, the original dataset (38 compounds) had been examined using training/test set methods; however, the parabolic distribution of the dataset relative to lipophilicity introduces high leverage when test sets are selected. Even with the addition of 5 compounds to the dataset the size of the dataset and the sparsity of molecules present in the high/low regions would not have allowed much flexibility in the selection of the test set and the selection of the test set would still have had a large influence on validation statistics. To minimise the possibility of test set selection bias and maximize the information from all the molecules in the dataset the entire dataset was used in the selection of models. Validation of the model fits was carried out using internal validation only at this stage; through the use of a bootstrap method.

3.5.1 Fit-all Models

The first avenue explored was fitting the whole dataset to one model. The full descriptor set for the 43 compounds, containing 331 descriptors, was examined in JMP. A 'fit all' method was utilised, linearly modelling the $\log(1/EC_{50})$ against all descriptors, with a maximum of three parameters for the model. Due to the number of available descriptors four parameters could not be selected as it generated too many possible models for the available computing power, four parameter models were generated with a 'fit all models' running using a subset of 30 descriptors[9]. These descriptors were selected as a mixture of interpretable descriptors and descriptors that performed well in the 3 parameter fit-all models.

The lipophilicity descriptors used in the models were ALOGPs and ALOGP-sq, as the ALOGPs descriptor was identified as the best log P descriptor through correlation with Retention Times. (For full lipophilicity correlations see Appendix B) The additional log P descriptors were removed from the dataset.

3.5.1.1 Two parameter model

The best 2 parameter model produced from the fit-all was the model with ALOGPs and ALOGPs-sq. Modelling the data with this two parameter model generated the following equation (3.1) with an R^2 value of 0.63 ($R^2_{adj} = 0.61$). The fit is shown in Figure 3.11.

$$log(1/EC_{50}) = -0.579 + 1.203 * ALOGPs - 0.133 * ALOGPs^2 \tag{3.1}$$

3.5.1.2 3 & 4 Parameter models

Increasing the number of parameters to 3 increased the R^2 value to approximately 0.79 for the top models. All of the top 20 models had an R^2 value above 0.74. Summary information about the 10 best three-parameter models for the whole dataset is shown in Table 3.3, ranked by R^2 values (additional models can be seen in ESI of our paper [140]) Running the 'fit all models' with a maximum of 4 parameters (a subset selected from the full descriptor set) slightly increased the R^2 value, with the top models having an R^2 value of approximately 0.815.

The predicted vs. actual plots can be found in the ESI [30] (model_fit_plots). These showed a fairly uniform distribution of the residuals with little in the way of outliers or skewing of the residuals and fairly similar appearance for all of the models. The

[9] the subset of descriptors can be found in the ESI

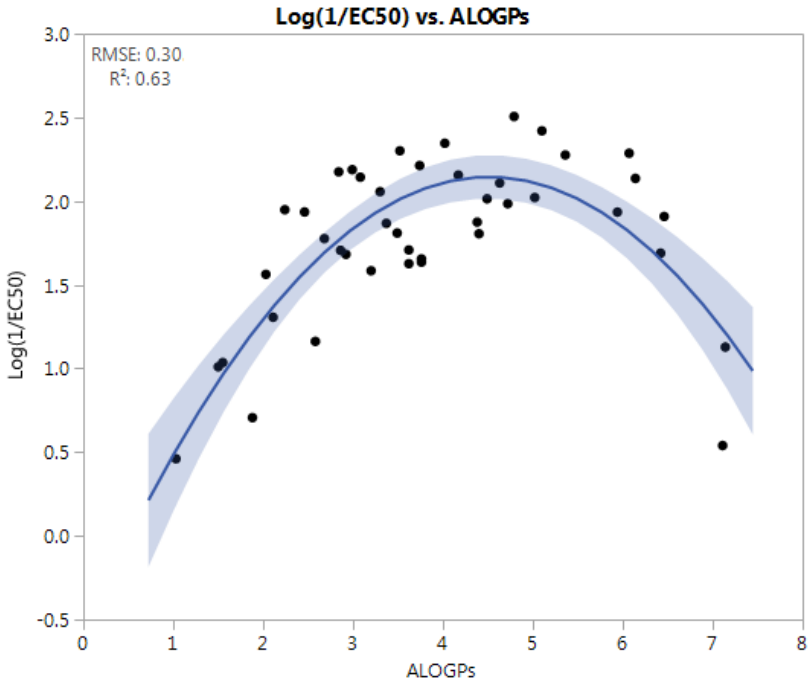


Figure 3.11: Log(1/EC_{50}) fit for ALOGPs and ALOGPs-sq model

No. of Par.	**Descriptors**				**R^2**	**R^2_{adj}**
3	ALOGPs	ALOGPs-sq	Mv	-	0.791	0.776
3	ALOGPs	ALOGPs-sq	J3D	-	0.790	0.775
3	ALOGPs	ALOGPs-sq	Mp		0.786	0.770
3	ALOGPs	ALOGPs-sq	nH	-	0.782	0.766
3	ALOGPs	ALOGPs-sq	AMW	-	0.777	0.762
3	ALOGPs	ALOGPs-sq	J	-	0.769	0.753
3	ALOGPs	ALOGPs-sq	E3u	-	0.768	0.754
3	ALOGPs	ALOGPs-sq	ARR	-	0.765	0.749
3	ALOGPs	ALOGPs-sq	Density (g/cm3)	-	0.762	0.746
3	ALOGPs	ALOGPs-sq	Surface tension (dyne/cm)	-	0.759	0.741
4	ALOGPs	ALOGPs-sq	nCIC	J3D	0.816	0.799
4	ALOGPs	ALOGPs-sq	nH	J	0.815	0.797
4	ALOGPs	ALOGPs-sq	AMW	J	0.815	0.797
4	ALOGPs	ALOGPs-sq	AMW	J3D	0.814	0.796
4	ALOGPs	ALOGPs-sq	J3D	Ui	0.814	0.796
4	ALOGPs	ALOGPs-sq	Density (g/cm3)	J3D	0.814	0.795
4	ALOGPs	ALOGPs-sq	Density (g/cm3)	J	0.812	0.794
4	ALOGPs	ALOGPs-sq	Parachor (cm3)	nH	0.810	0.791
4	ALOGPs	ALOGPs-sq	Molar refractivity (cm3)	nH	0.809	0.789
4	ALOGPs	ALOGPs-sq	Polarizability (cm3)	nH	0.809	0.789

Table 3.3: Best fitted 3 and 4 parameter models, ranked by R^2 values.
4 parameter models are fitted with a small subset.

addition of an extra parameter increased the R^2 value as could be expected, but showed no particular affinity for a particular single descriptor over others.

3.5.1.3 Bootstrap Validation

After running the 'fit all models' fit, 95% confidence intervals were obtained for a selected number of models from the linear fit least-squares analysis. These included the ALOGPs, ALOGPs-sq model, 5 three parameter models and 4 four parameter models.

These models were also run through a bootstrap method in R to obtain confidence intervals a as method of validation.[10] Due to the distribution of the data being heavily biased towards the middle of the ALOGPs range, a stratified selection was utilised within the bootstrap function to ensure that the selection always included a range of points from the upper and lower ends. Using the bootstrap package, boot, in R, [22, 141, 142] the data were sampled from the full dataset and the fit model statistics calculated, and this was repeated using a resampling of the dataset 999 times. Comparing the confidence intervals for the bootstrap fit and the linear least squares prediction was a method to examine how robust the fits are.

The coefficients and confidence intervals for the best 2, 3 and 4 parameters models are shown in Table 3.4 and it can be seen that the confidence intervals obtained using the bootstrap function were well aligned with the confidence intervals obtained directly from the linear fit.[11] This suggests that the fits are quite robust. The most variation comes in the coefficient for the intercept with a much narrower range in the confidence intervals for the ALOGPs and ALOGPs-sq coefficients.

As shown by the models described in Table 3.3, there were a large number of calculated descriptors that seemed to offer potentially useful additional descriptive power to the fits, but without any clear advantage of one descriptor over the others (apart from the clear importance of log P).

Another avenue of investigation for selection of descriptors was principal components analysis (PCA) or partial least squares (PLS) as these create new descriptors with a combination of the original descriptors. PCA was carried out with 4 PCs calculated; however, the resultant fit had almost no correlation with $\log(1/EC_{50})$ and obliterated the interpretability of the models in terms of contributions of the terms to the models.

Returning to the initial investigation into pKa distribution, Section 3.2.2.5 suggested a different avenue of investigation. This was through a classification approach, modelling subsets of the compounds based on the structural features of the molecules.

[10] R Code can be found in the ESI [30] - C3_TambAnionTransport/RCode

[11] Further confidence intervals for other models can be found in the ESI - model_coefficients_CI

		Model Parameters	**ALOGPs ALOGPs-sq**	**ALOGPs ALOGPs-sq Mv**	**ALOGPs ALOGPs-sq nCIC J3D**
		R^2	0.629	0.791	0.816
COEFFICIENTS		Intercept	-0.579	3.362	-5.105
	Linear fit	2.5% C.I.	-1.165	1.838	-7.579
		97.5% C.I.	0.008	4.887	-2.632
	Bootstrap	2.5% C.I.	-1.108	2.159	-7.681
		97.5% C.I.	-0.086	4.419	-2.694
		ALOGPs	1.203	1.372	1.284
	Linear fit	2.5% C.I.	0.903	1.135	1.056
		97.5% C.I.	1.504	1.610	1.511
	Bootstrap	2.5% C.I.	0.904	1.126	1.087
		97.5% C.I.	1.470	1.579	1.493
		ALOGPs-sq	-0.133	-0.158	-0.146
	Linear fit	2.5% C.I.	-0.168	-0.186	-0.172
		97.5% C.I.	-0.098	-0.129	-0.120
	Bootstrap	2.5% C.I.	-0.166	-0.190	-0.173
		97.5% C.I.	-0.093	-0.123	-0.116
		3rd Parameter		-6.616	0.411
	Linear fit	2.5% C.I.		-9.063	0.057
		97.5% C.I.		-4.168	0.764
	Bootstrap	2.5% C.I.		-8.432	0.064
		97.5% C.I.		-4.473	0.796
		4th Parameter			1.587
	Linear fit	2.5% C.I.			0.808
		97.5% C.I.			2.367
	Bootstrap	2.5% C.I.			0.796
		97.5% C.I.			2.330

Table 3.4: Coefficients and confidence intervals for the best two, three and four parameter models

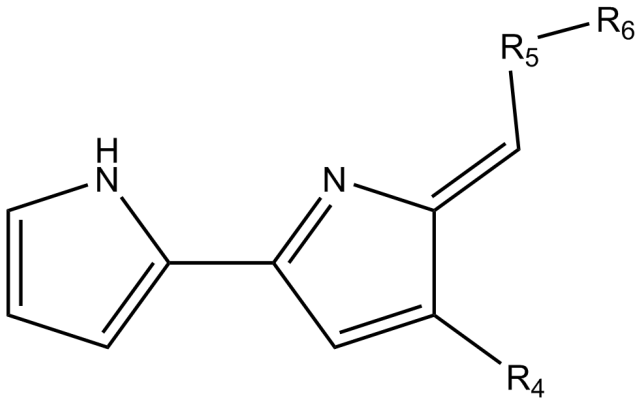


Figure 3.12: Backbone structure of Tambjamine molecule derivatives

3.6 Classification of Compounds

3.6.1 Dataset splitting by substituent

All of the compounds in this dataset had a structure containing the same bipyrrole core, see Figure 3.12. From this backbone structure, there were 3 main positions (R4, R5, R6) in which the structures differed, these could be used to classify the compounds. The substituent at the base of the ring (marked by R4) was either OMe or OBn. The top substituent at the enamine position (marked by R5) was either non aromatic (signified by NH) or aromatic where the aromatic ring was benzene (signified by NH-Ar) with a few exceptions. There was also an 'R' substituent attached to either the NH or NH-Ar group (marked by R6). This 'R' group varied quite significantly and included alkyl groups, halogens, alkyl halogens etc.

In the case of the incorrect pKa grouping that was discovered previously, the grouping of datapoints appeared to correspond to structural differences at the R5 position. All of the 'high-pKa' value compounds had an aromatic (NH-Ar) group in the R5 position and all of the 'low-pKa' value compounds had a non-aromatic (NH) group in the R5 position. As the 'pKa values' showed some splitting of the ALOGPs vs. $\log(1/EC_{50})$ plot previously (Fig.3.5), this led to investigation of how changes in all substituents affected the correlation in the dataset.

Due to the presence of multiple positions at which structural differences occurred each split of the dataset investigated was an isolation of a single type of structural change. For each comparison it was necessary to create a subset of compounds which contained a sufficient number of compounds to enable identification of a trend. Many of the compound subsets for the 'R' group (R6) only contained 2 or 3 compounds which was not sufficient for analysis.

In the ring position (R4), out of 43 compounds, 10 had OBn as the substituent and 33 had OMe. In the enamine position (R5), 19 compounds had an NH substituent and 22 compounds had the NH-Ar substituent, with two compounds that did not fit into either group; one compound had pyridine (py) and the other compound CH_2-py. In the R-type position (R6) the compounds were split across 7 groups, the most populated group was the alkyl group, containing 28 compounds, the remaining 15 compounds fit into six other groups. Compound structures can be seen in Figure 3.10 on page 77.

After the compounds were classified into their groups for the 3 substitution positions it was possible to plot the data to identify the presence of any splitting. The data plotted was ALOGPs vs $\log(1/EC_{50})$ assuming a parabolic fit. The plots generated by this are shown in Figure 3.13.

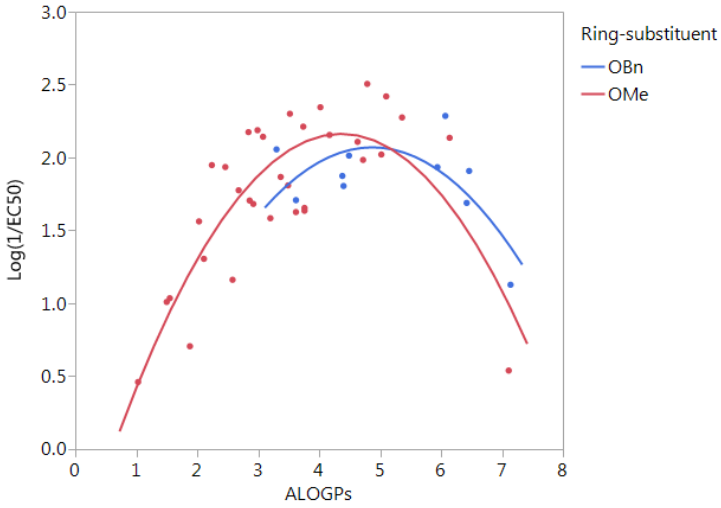
Plot (a) shows the dataset split by Ring substituent (R4), this indicated the possibility of a slight shift in the position of the optimum ALOGPs; however, there were not sufficient datapoints on the low end of the ALOGPs scale for the OBn group to confirm this theory.

Plot (b) shows the dataset split by Enamine substituent (R5) this showed a more significant split between the non-aromatic (NH) and aromatic (NH-Ar), with the NH group having a higher peak $\log(1/EC_{50})$ value. The optimum ALOGPs value appeared to be the same between the groups. This may be due to the electronic effect of the aromatic ring on the anion binding site of the enamine and pyrrole ring. Although the effect of changing OMe to OBn was found to have little affect on chloride binding (K_a) [139] no binding values were available for comparison of the NH/NH-Ar substituent change.

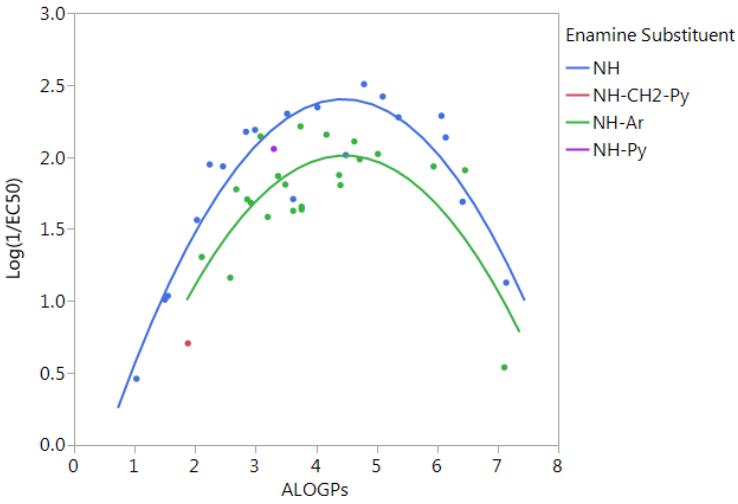
Plot (c) shows the 'R-type' split (R6), the alkyl group showed a strong parabolic relationship; however, the other groups were not highly populated so little correlation could be made within those groups. The reason for this was that in the non-aromatic group the main substitution possible is an alkyl chain, but in the aromatic group a wider variety of substitutions can be made on the benzene ring. Since the substitutions were only present in the para position it limits the number of compounds that will have the same R-type substituent. More compounds would be required to determine if any splitting occurs due to the R-type substituent.

3.6.2 Mixed effect models

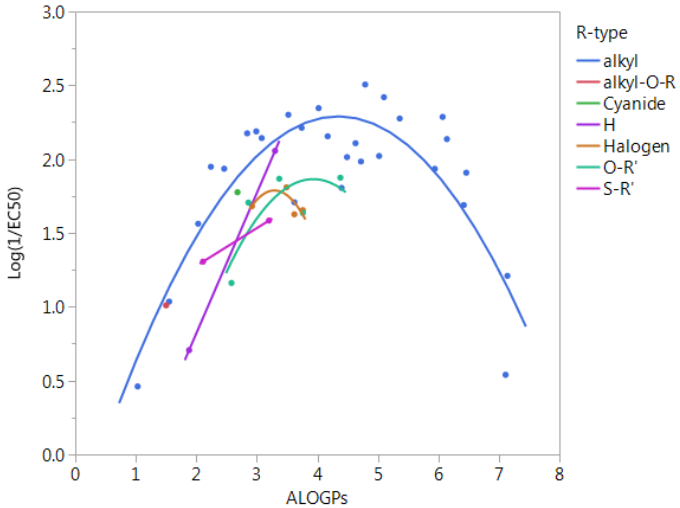
Further modelling that was carried out utilised the lme4 package in R [143]. This package allows the use of an entire dataset to fit the curve of a parabola, whilst allowing subsets of data to adjust the positioning of the curve by changing the intercept. Due to the size of 'R-type' groupings, only the alkyl R-group compounds were selected for further analysis.



(a) Ring Substituent



(b) Enamine Substituent



(c) R-type Substituent

Figure 3.13: Parabolic fits of ALOGPs vs $\log(1/EC_{50})$ - splitting compounds by substituents

A linear mixed effect model (lmer) was run for the subset of the compounds containing an alkyl R-Type, modelling the dataset to the form:

$$log(1/EC_{50}) = a + b * ALOGPs + c * ALOGPs^2$$

with further splitting being done by the enamine substituent (R5). The resulting plot for all compounds with an alkyl R-type is shown in Figure 3.14, and the coefficients are shown in Table 3.5.

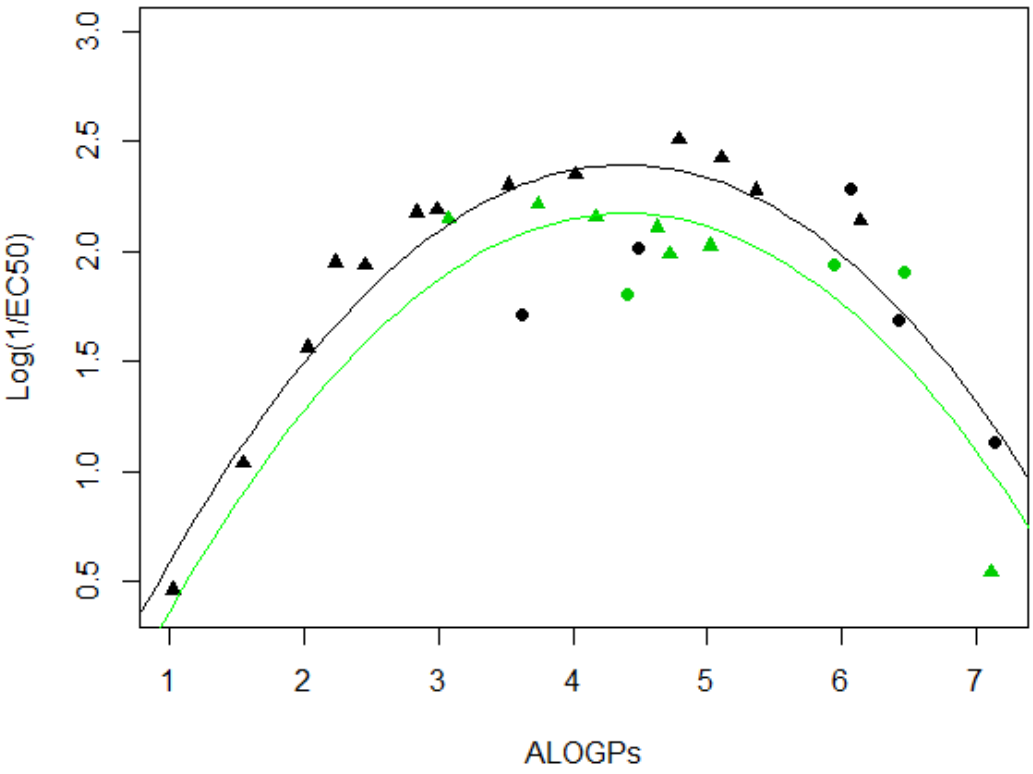


Figure 3.14: lmer fit for alkyl R-type for OMe and OBn substituents
Points coloured by enamine-substituent: black - NH, green - NH-Ar.
Shape by ring-substituent: circle OBn, triangle OMe

Enamine substituent	Intercept	ALOGPs	ALOGPs-sq
NH	-0.6255	1.3728	-0.1560
NH-Ar	-0.8531		

Table 3.5: Coefficients for lmer model for alkyl R-type, both OMe and OBn substituents

Narrowing the subset further by taking only the OMe ring substituted compounds (20 of the 28 alkyl compounds) and carrying out the lmer model produces the plot in Figure 3.15 and the coefficients in Table 3.6.

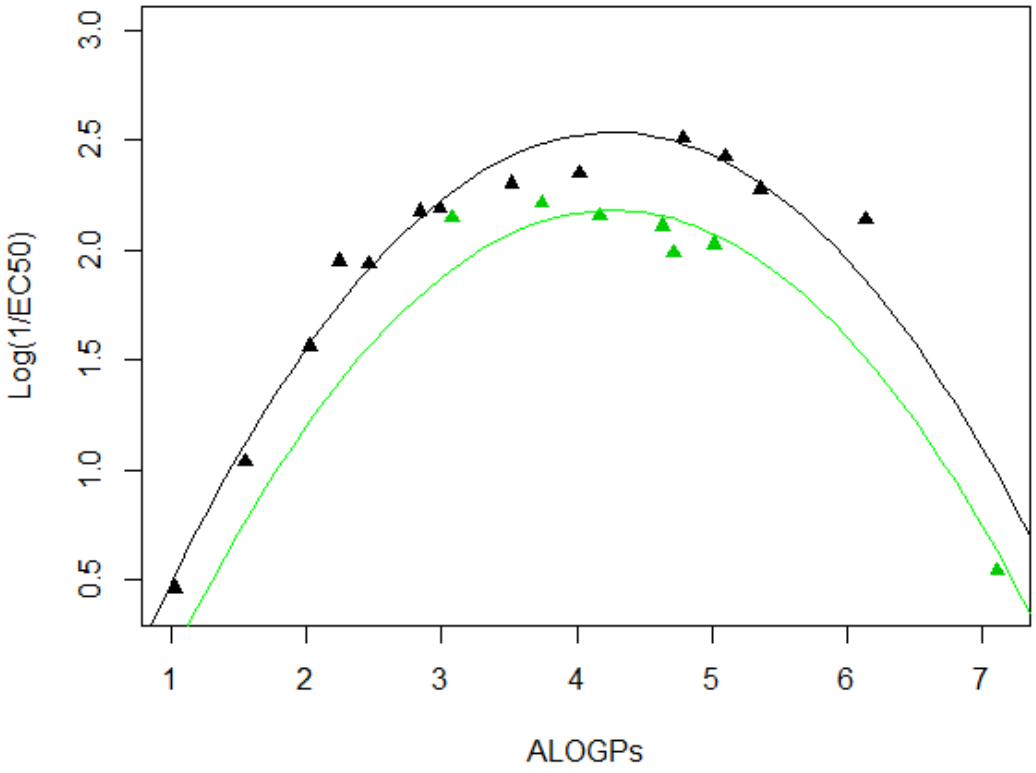


Figure 3.15: lmer fit and model for the alkyl R-type, OMe ring substituent
Points coloured by enamine-substituent: black - NH, green - NH-Ar

Enamine substituent	Intercept	ALOGPs	ALOGPs-sq
NH	-0.9557	1.6373	-0.1920
NH-Ar	-1.3088		

Table 3.6: Coefficients for lmer model for alkyl R-type,
OMe ring substituent only

The use of the lmer model allowed for adjustment in the vertical positioning of the parabola, which fitted the split present for the enamine substituents. It is possible that the other substituents may require horizontal translations instead to provide a change in the optimum log P. This was indicated as a possibility by the OBn/OMe split plot (Figure 3.13a) but there were not enough points with low log P to make a robust conclusion. Although a shift in peak activity could be expected to be affected by the substituents the optimal lipophilicity may be affected more by the type of membrane through which the compounds are moving.

3.6.3 Structural changes due to substituents

Figure 3.16 shows the overlaid 3D structures of 3 tambjamine molecules (optimised in Marvinsketch) with attempted alignment of the structures. These three structures are representative of the variety of substitution patterns, to see how the substituent changes affect the 3D molecule. Although the structures were 3D 'optimised' this is not the conditions that they would be under in a cellular environment. Molecular simulation would be required to get more accurate conformations.

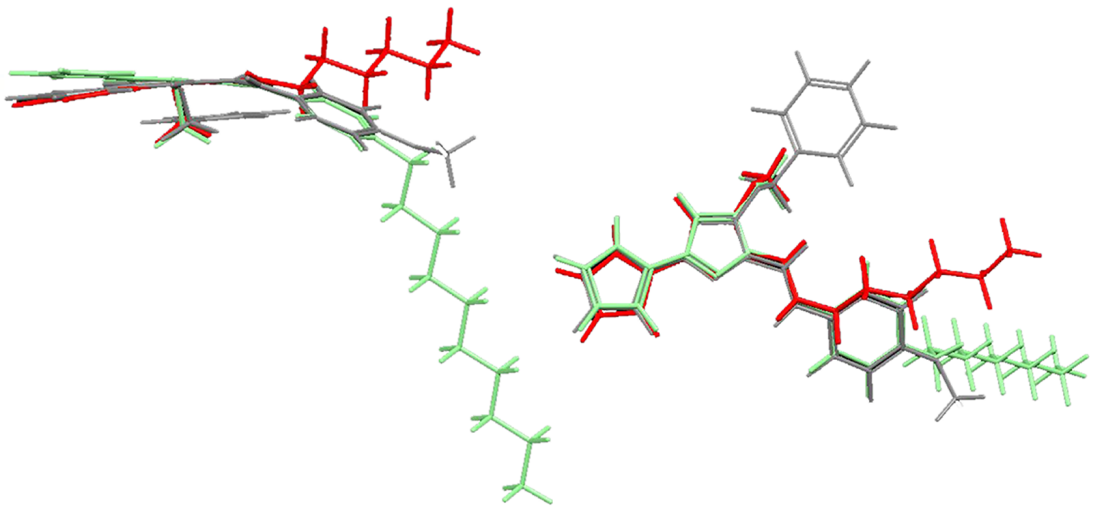


Figure 3.16: Overlaid 3D structures for tambjamines with different substitution patterns - side on and top down views
light green - Tambjamine 7, red - Tambjamine 25, grey - Tambjamine 37
see Figure 3.10 for 2D structures

This shows the mainly planar core of the two rings in the backbone. The OMe or OBn group which is present directly neighbouring the ring does not appear to significantly change the 'depth' of the molecule as the OBn lies in a parallel plane to the core rings (indicated in the grey molecule); however, the OBn does extend a fair distance, increasing the 'width' of the compound in comparison to the OMe molecules.

The presence or absence of the aromatic group neighbouring the NH appears to have more of an effect on the overall size/shape of the molecule. It also increases the 'width' of the compound compared to the backbone, additionally the long side chain (on the green molecule) significantly increases the 'depth' of the molecule in this overlay. However, this side chain would be relatively flexible to undergo twisting changes to form the lowest energy structures.

The changes in 3D conformation may be an underlying cause for the changes in the transport activity, as the steric bulk affects how efficiently a molecule can move through the lipid membrane. This would require additional molecular simulations structural

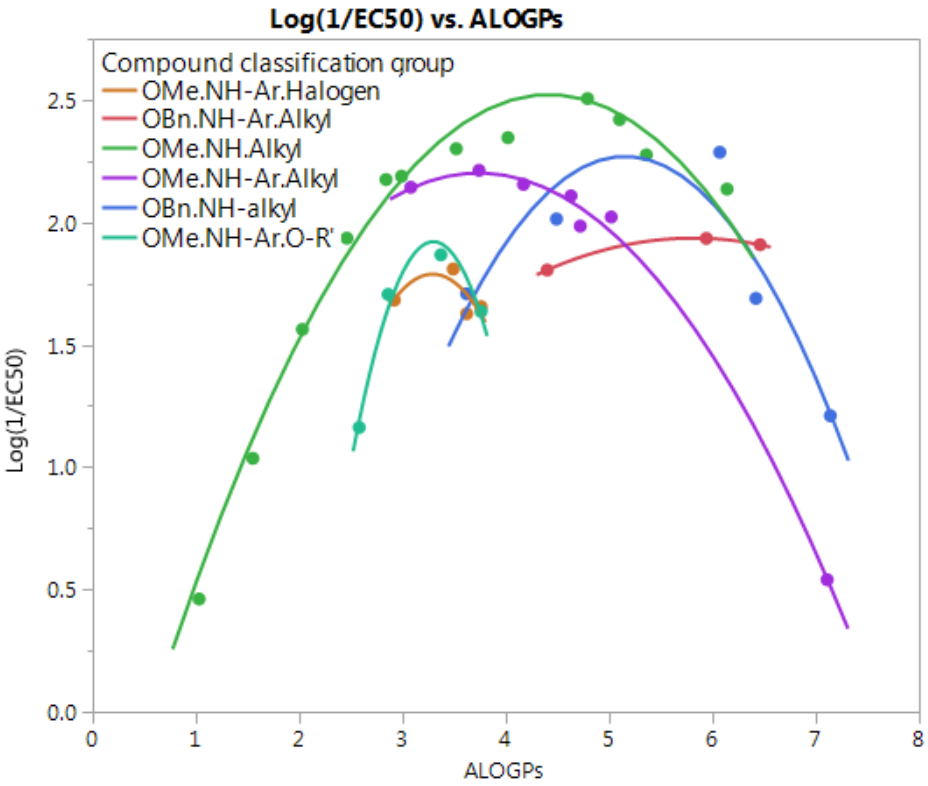


Figure 3.17: Splitting the Dataset into compound subsets, excludes sets with < 3 datapoints

conformations of the molecules before further investigation could be carried out. A measurement of how wide the molecule is, or a ratio of maximum height to width could be plotted against the $\log(1/EC_{50})$ values to look for correlations.

3.6.4 Combined substituent grouping

In addition to splitting the data by a single substituent the compounds can be classified using all three substituents to fully classify the molecule substituent pattern of the molecule. This resulted in 12 classification groups, with 6 of the groups containing 2 compounds or less which could not be fitted. The remaining 6 classification groups were fitted to a parabolic curve modelling ALOGPs against $\log(1/EC_{50})$. The resultant curves are shown in Figure 3.17 with their equations in Table 3.7.

These parabolas produced strong fits for most of the groups; however, the small size of many of the groups mean that these fits may not properly reflect the behaviour of that compound group and they would not be reliable for fitting new compounds. Additional compounds would be required to provide better justification of the fits for the smaller sets. The only compound group which contained a sufficient distribution of compounds was the OMe.NH.Alkyl group which produced a very strong fit with ALOGPs. The

Sub Group	Equation	R^2
OBn.NH.Alkyl	$Y = -4.783 + 2.737 * ALOGPs - 0.2656 * ALOGPs^2$	0.84
OBn.NH-Ar.Alkyl	$Y = -0.2663 + 0.7575 * ALOGPs - 0.06513 * ALOGPs^2$	0.999
OMe.NH.Alkyl	$Y = -0.8097 + 1.509 * ALOGPs - 0.1707 * ALOGPs^2$	0.97
OMe.NH-Ar.Alkyl	$Y = 0.1699 + 1.088 * ALOGPs - 0.1456 * ALOGPs^2$	0.999
OMe.NH-Ar.Halogen	$Y = -6.332 + 4.936 * ALOGPs - 0.7501 * ALOGPs^2$	0.48
OMe.NH-Ar.O-R	$Y = -13.52 + 9.364 * ALOGPs - 1.42 * ALOGPs^2$	0.98

Table 3.7: Model equations and R^2 values for quadratic fits of compound grouping shown in Fig.3.17 modelling for $\log(1/EC_{50})$

OMe.NH-Ar.alkyl group also produced a very strong fit; however, it only contained datapoints on half of the parabola.

The equations for the parabolas were produced in the form:

$$log(1/EC_{50}) = a + b(ALOGPs) + c(ALOGPs)^2 \tag{3.2}$$

However, these equations could be transformed into the following form giving a definition in terms of an overall maximum and a parameter that describes the shifts away from this point, such as:

$$y = ymax + k(x - xmax)^2 \tag{3.3}$$

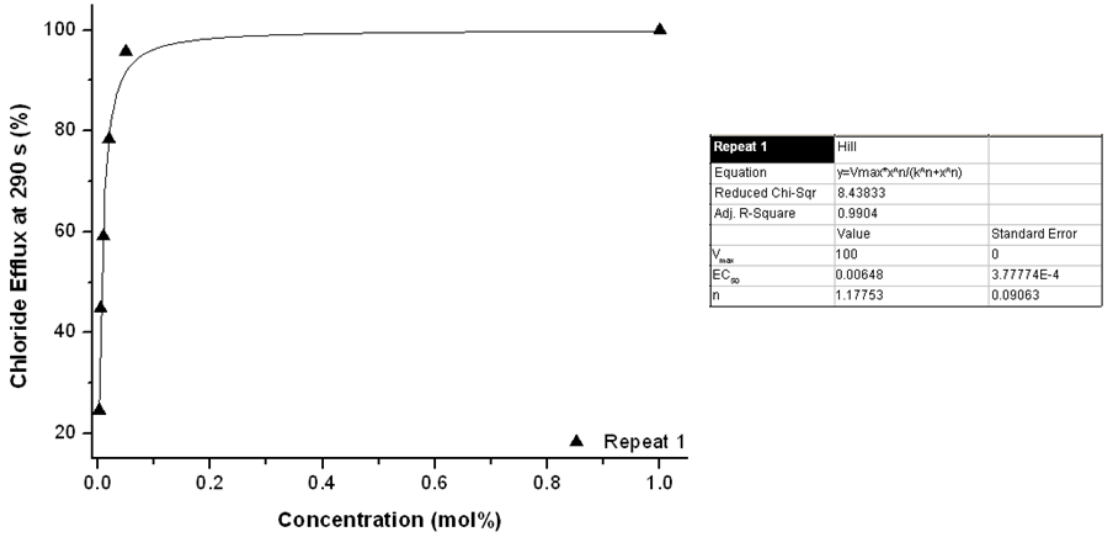
Where $y = log(1/EC_{50})$, ymax is the maximum $log(1/EC_{50})$ value, k is the slope, x is ALOGPs, and xmax is the optimum ALOGPs value, which produces the highest $1/log(EC_{50})$ value.

3.7 Underlying data

Throughout all of the analysis discussed so far the data used was obtained from processed values. These values all have a large amount of raw data which is used in their calculation. Some of this data was incorporated into the ESI of our paper [144] in the form of chloride efflux plots and Hill plots.

Chloride efflux plots are a plot of chloride efflux over time and the gradient of the curve is used in the calculation of initial rates of chloride efflux (k_{ini}). QSAR models can use k_{ini} as an alternative measure of anion transport ability. Hill plots are plots of chloride efflux at 290s against concentration of transporter and are used in the determination of the EC_{50} value through the Hill equation, see Section 3.7.1.

The use of processed data can often mask behaviour as it only produces a single value, so examination of the raw data is a useful tool to check for unusual behaviours or mechanisms that may be present.



Repeat 1	Hill	
Equation	y=Vmax*x^n/(k^n+x^n)	
Reduced Chi-Sqr	8.43833	
Adj. R-Square	0.9904	
	Value	Standard Error
V_{max}	100	0
EC_{50}	0.00648	3.77774E-4
n	1.17753	0.09063

Figure 3.18: Example of a well fitted Hill plot - compound 24

3.7.1 Hill plots

The Hill equation is used in anion transport to fit plots of response against transporter concentration and allow calculation of the transport ability through EC_{50}. The Hill equation is as follows:

$$V = \frac{V_{max} * [T]^n}{(K_{0.5})^n + [T]^n} \tag{3.4}$$

Where V is the reaction velocity (in this case chloride efflux), V_{max} is the maximum possible velocity for the reaction, $[T]$ is the concentration of transporter, n is the Hill coefficient and $K_{0.5}$ is the concentration which gives rise to V that is half of V_{max}. When V_{max} is 100% chloride efflux the value of $K_{0.5}$ is equal to the EC_{50} value which has been used throughout this research. An example of a plot fitted using the Hill equation can been seen in Figure 3.18. The parameters that are fitted are V_{max}, $K_{0.5}$ (given as EC_{50}) and n.

This equation is closely related to the Michaelis-Menten equation for enzyme kinetics. When $n = 1$ the equation is the same as the Michaelis-Menten equation given below:

$$V = \frac{V_{max} * [S]}{K_m + [S]} \tag{3.5}$$

Where V, V_{max} are the same as Eq. 3.4, $[S]$ is the concentration of substrate and K_m is the Michaelis constant (substrate concentration at which the reaction rate is half of V_{max}) which is the same as $K_{0.5}$.

These equations are both used to model the kinetics of reactions by enzymes or transport by transporters and their dependence on concentration of substrate. The Michaelis-Menten equation is used to model simple systems or enzymes where a single ligand is

bound. In more complex systems that allow binding of more than one ligand with co-operativity (where the binding of a ligand affects the subsequent binding of ligands) the Hill equation is applied. [145]

These curves indicate saturation of the transporter through the asymptotic behaviour of the curve, where it reaches V_{max}. In these transport experiments this is the transporter reaching 100% efflux by a given time. Further increase in concentration of transporter cannot increase the amount of chloride transported above this amount.

The value of $K_{0.5}$ or K_m is the inverse of the apparent efficacy of the transporter. A low value of $K_{0.5}$ indicates that a low concentration of transporter is required to reach 50% efflux (high efficacy), whereas a high value of $K_{0.5}$ indicates that a large concentration of transporter is required to reach 50% efflux (low efficacy).

The value of n (Hill coefficient) provides a measure of the binding co-operativity. Where $n > 1$ it indicates positive co-operativity between ligands, where $n = 1$ it indicates no co-operativity and where $n < 1$ it indicates negative co-operativity. The Hill coefficient is sometimes taken as the number of ligand binding sites present on the molecule, however this is an erroneous conclusion. [146]

In anion transport chemistry the analysis of Hill coefficients suggest that the co-operative binding is present for the transporter molecules rather than additional anion molecules, with Hill coefficients $n > 1$ typically being considered evidence of molecular systems containing more than one monomer. [147] However, there are many molecular systems which are known to self assemble but give low Hill coefficients and compounds that give $n > 1$ when they only exist as a momomer. Due to this the values of Hill coefficients should not be considered strong evidence for a specific type of system. The plots provide the transport ability through EC_{50} and may provide some indication of the binding methods but they should be verified through other experiments.

3.7.1.1 Examination of Hill plots

Examination of the Hill plots showed that the majority of the compounds had a good fit for the Hill equation with little deviation. However, low activity compounds required multiple additional measurements at higher concentrations to allow the curves to be adequately fit. Although the measurements for low activity compounds were more difficult to obtain, the measurement of 'bad' compounds were necessary to allow analysis and fitting of the dataset as a whole.

A number of compounds had only a single repeat of the transport experiment carried out. These could have benefited from repeat experiments, especially if the points deviated slightly from the Hill equation curve. In particular a compound with a poor fitting Hill

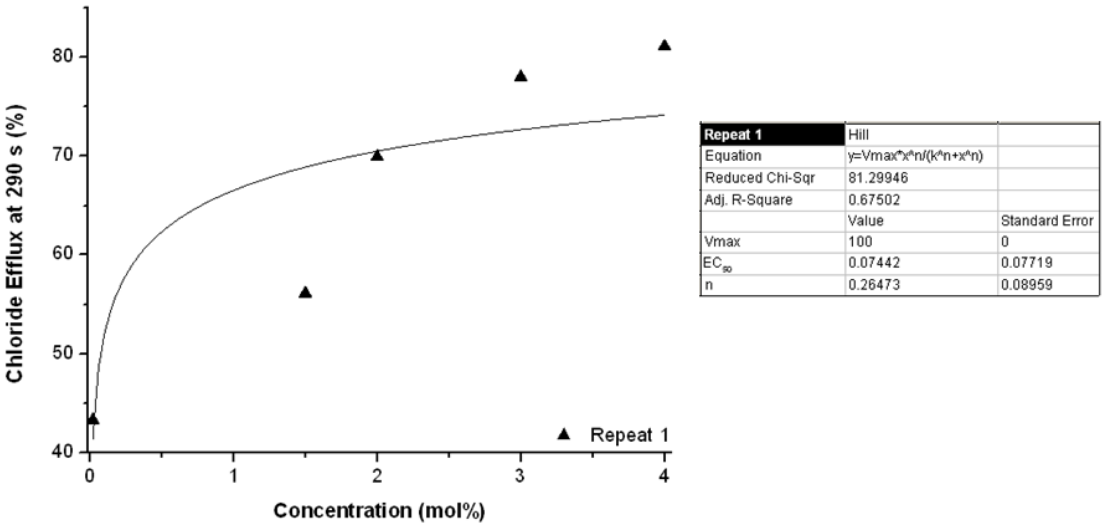


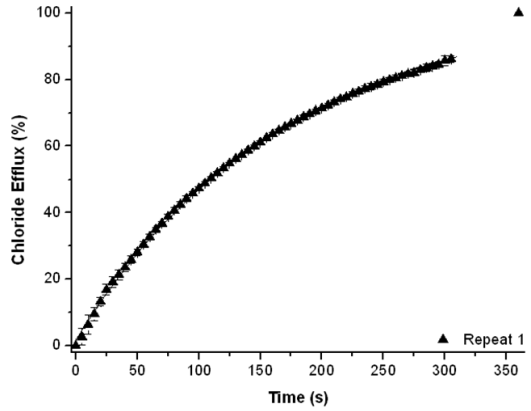
Figure 3.19: Hill plot for compound 43 showing unusual behaviour

plot was compound 43, which can be seen in Figure 3.19. The data points indicated the potential presence of a sigmoidal curve, although the curve did not fit to this shape.

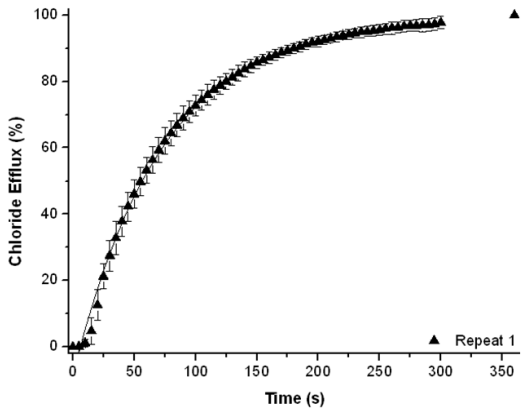
The presence of a sigmoidal Hill curve is usually accompanied by a large Hill coefficient and is considered evidence for co-operative or multisite binding [146]; however, as x-ray structures for some tambjamine compounds showed only a single binding site [137], multisite binding would be an unlikely occurrence. Further repetition should be carried out to identify if the previous measurements were erroneous, and if not the binding of the compound should be investigated further.

Due to the unusual behaviour of the Hill plot for compound 43 the EC_{50} value for this compound was validated through an alternative method. For this the correlation of $\log(1/EC_{50})$ vs $\log(1/k_{ini})$-(R^2=0.913, RMSE=0.14) was used in the prediction of the value of EC_{50} [144]. But repetition of the transport experiments would be a preferable route of action.

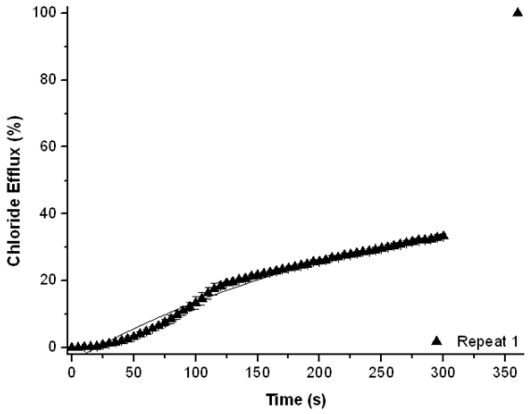
Across the tambjamine group the Hill coefficients were approximately n=1 for the majority of compounds which aligns with the suggestion of a single binding site on the molecule. If the assumption is made that no co-operative binding is occurring, due to the single binding site, then the data could be modelled using the Michaelis-Menten equation (Eq. 3.5) instead of the Hill equation. If further investigation were carried out then the data could be examined to see how the use of the Michaelis-Menten equation affects the calculated EC_{50} values.



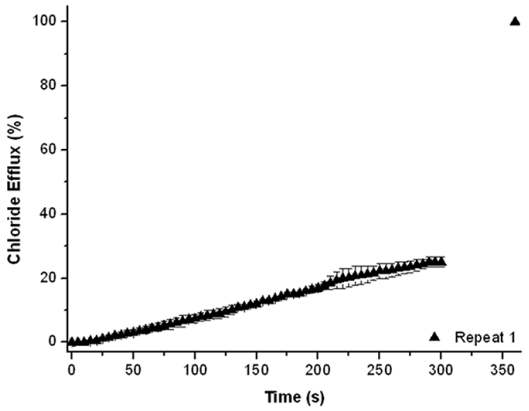
(a) Compound 37 - No induction period



(b) Compound 25 - Short induction period



(c) Compound 21 - Long induction period



(d) Compound 19 - Undetermined induction period

Figure 3.20: Chloride efflux plots for Tambjamine compounds showing different induction periods

3.7.2 Induction Periods

The chloride efflux plots in the ESI of our paper [144] (image form) were examined and induction periods were observed in a number of the plots. An induction period is an initial phase where the 'reaction' proceeds at a slower rate which later accelerates. [148] In the case of anion transport it could indicate an interaction with the membrane.

The behaviour of the chloride efflux curve was categorised into 5 groups with respect to induction periods: none - no evidence of induction period, short - slight evidence of induction <15 secs, medium - 15-50 secs, long >50secs, unknown - compounds with no obvious induction, but fitted to a linear regression. Examples of the curves can be seen in Figure 3.20. For two of the unknown compounds the chloride efflux was not measured for the full 360 seconds, this may have contributed to the inability to fit the compound to the curve. Compound 7 was only measured for 50 seconds and compound 20 for 90 seconds.

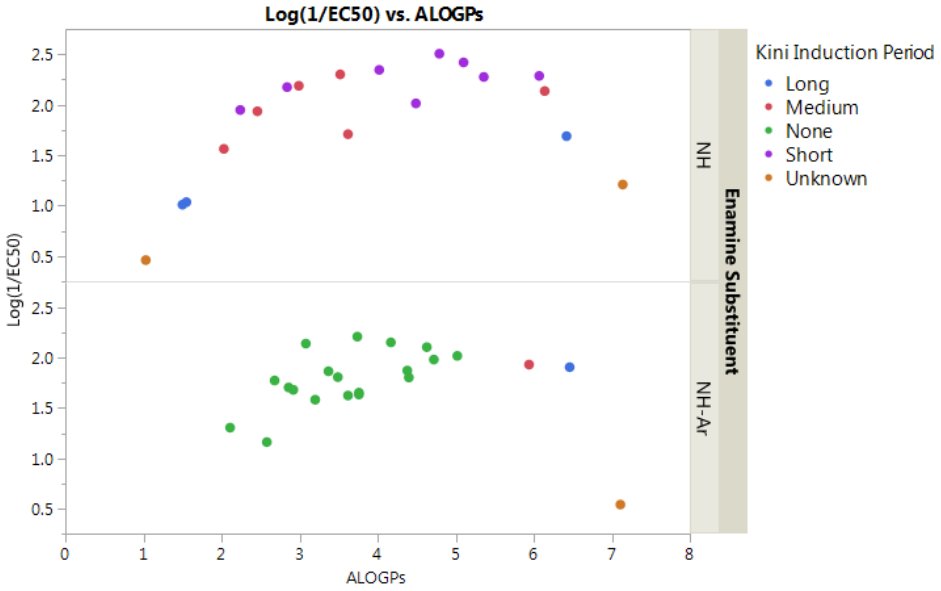


Figure 3.21: Log(1/EC_{50}) vs ALOGPs - split by enamine substituent and coloured by induction period length, only NH-alkyl and NH-Ar-R groups

This length of induction period was then used in examination of the data to see if any trends or grouping could be identified. A number of observations can be made about the plots, demonstrated in Figure 3.21.

- All of the compounds with no induction period belonged to the NH-Ar group.
- Compounds with an unknown induction period are the furthest outliers on the log P range.
- Compounds with a long induction period are next furthest out in log P.
- The short and none groups occupy the central portion of the log P range.

While the presence of the induction period for compounds on the ends of the log P scale is likely due to the transporter's interaction, or lack of, with the lipid bilayer the possibility of an alternative method to mobile carrier has not been conclusively ruled out in experiments.

3.7.3 Fitting chloride efflux curves

The chloride efflux plots shown in Figure 3.20 were used in the calculation of k_{ini}. The method used by the Quesada group for determination of k_{ini} involved the measurement of chloride efflux over time and fitting the data to a non linear regression with 3 parameters shown in equation 3.6a, where y is chloride efflux (%) and x is time (s), k_{ini} was then calculated through equation 3.6b. Some compounds fits were alternatively fit to a linear

regression in the form $y = m * x + c$ where m was the k_{ini}.

$$y = a - b * c^x \tag{3.6a}$$

$$k_{ini} = -b * ln(c) \tag{3.6b}$$

This equation of fit was re-examined as a number of compounds did not appear to fit this equation very well. Examining the kinetics of the chloride transfer this reaction ($Cl_{in} \rightarrow Cl_{out}$) can be modelled through a first order rate equation (eqn. 3.7), where y is chloride efflux (%) and x is time (s).

$$y = a(1 - e^{-kt}) \tag{3.7}$$

The raw data used to calculate the initial fits were obtained from the researchers. Re-examination of the raw data was carried out for a selection of compounds, a subset was selected as the process of extracting the data was lengthy. The data consisted of many folders each with multiple spreadsheets containing transport experiments and multiple repeats for each compound.

The values for chloride efflux over time were extracted for compounds 1, 8, 11, 21, 25, 30 and 37 and the values from multiple repeats were averaged. This raw data was refitted in R using a non-linear regression fitting to equation 3.7. No fit could be produced for compound 21 (Figure 3.20c), which had initially been fit with a linear equation.

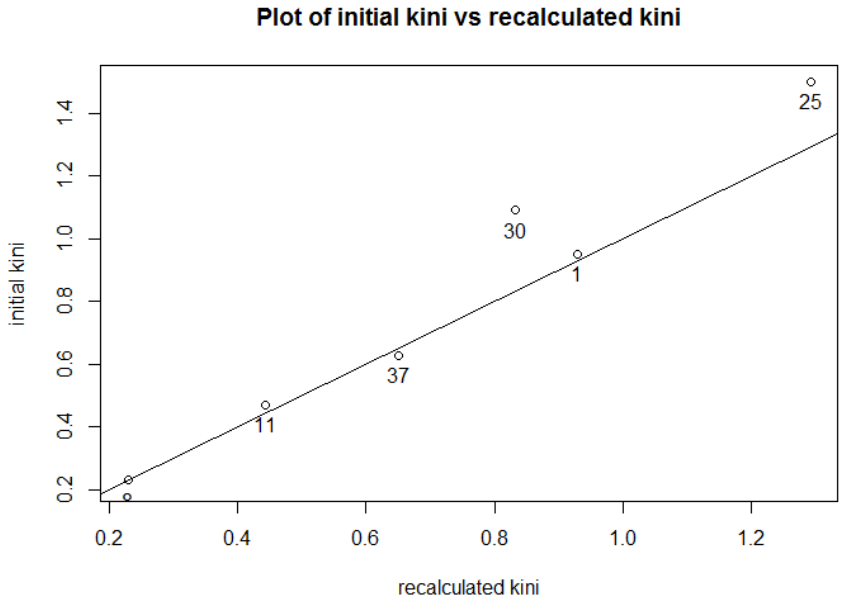
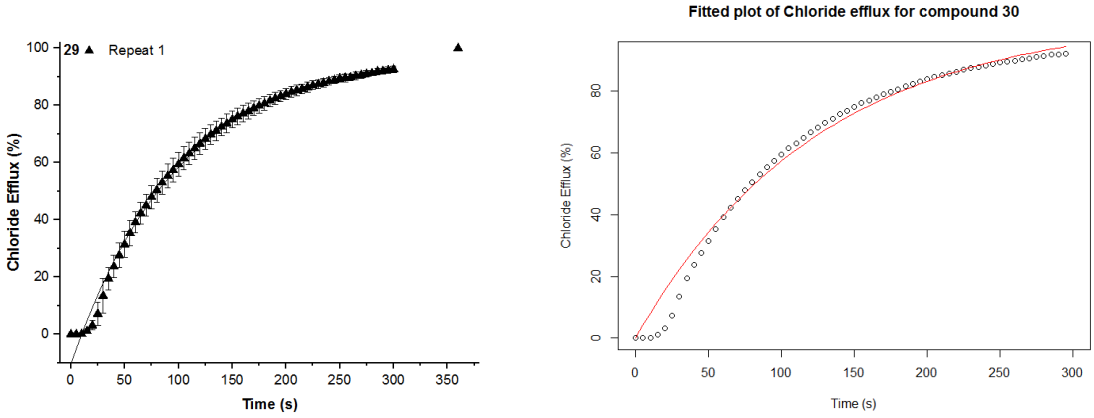


Figure 3.22: Plot of initial k_{ini} (eqn.3.6a) vs. recalculated k_{ini} (eqn.3.7) for 6 compounds, with $y = x$ line



(a) Initial fit (eqn. 3.6a) (b) Refitted (eqn. 3.7)

Figure 3.23: Comparison of chloride efflux fits for compound 30

A strong correlation was shown between the two equations for most compounds, except compounds 25 and 30 which gave a sizeable discrepancy. Figure 3.22 shows the plot of the values from the two equations with a line of $y = x$. Both compound 25 and 30 exhibited a short induction period which likely is a factor in the difference. A comparison of the two fits for compound 30 can be seen in Figure 3.23.

While the initial equation fitted the main section of the data better, it produces a negative intercept which is not a physical possibility; however, the refitted equation always has a 0,0 intercept. The initial fit equation leads to a potential overestimation of the initial rate of k_{ini} as it largely excludes the induction period. The presence of an induction period in the transport is an important aspect of the action. If k_{ini} was to be used as the measure of activity in models further investigation would be required as to which calculation is the more appropriate measure of the k_{ini}, and if the 3 parameter fit (eqn. 3.6a) is used whether it should also include some measure of the induction period present.

3.8 Implication of Experimental Error

As in Chapter 2 - Section 2.4.1.4 the experimental error of the tambjamine compounds was investigated through examination of the error values for transport ability (EC_{50}), determined from a Hill plot. These errors were obtained from the ESI of our paper [144], with the errors being propagated through to give error values for the $\log(1/EC_{50})$ values.[12]

[12]Spreadsheet containing $\log(1/EC_{50})$ errors can be found in ESI

Once obtained these errors were used in a plot of $\log(1/EC_{50})$ against ALOGPs (Figure 3.24) to see their magnitude and the implication of their presence with respect to the quadratic fit.

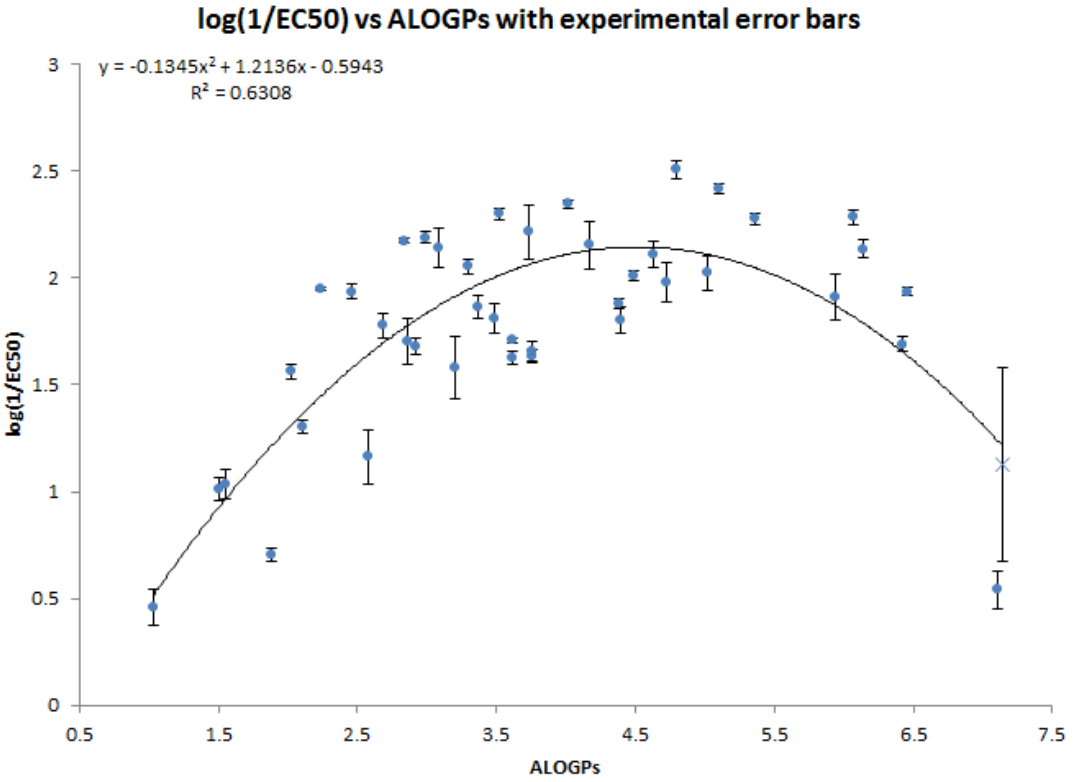


Figure 3.24: Plot of $\log(1/EC_{50})$ against ALOGPs showing error bars, compound 43 marked by $\times$

The errors within the tambjamine dataset were on the whole slightly larger than those of the Gale group dataset (Chapter 2), with 7 compounds having an error greater than 0.1 log units and 12 compounds having an error between 0.05 and 0.1 log units. Only one compound contained an error of concerning magnitude. This was compound 43, marked on the plot by $\times$. The error in $\log(1/EC_{50})$ for this compound was 0.45 log units.

The error for compound 43, calculated from the Hill plot shown in Figure 3.19, was very large compared to all other compounds. The Hill plot curve did not exhibit a good fit to the data; however, only a single repeat was available so it could not be determined if there was an erroneous measurement. As compound 43 was one of only a few high lipophilicity compounds in the dataset the $\log(1/EC_{50})$ value was validated through the correlation of $\log(1/EC_{50})$ and $\log(1/k_{ini})$, see appendix - Figure B.2. The predicted and Hill calculated $\log(1/EC_{50})$ values were well aligned with a difference of only 0.08 log units and as such compound 43 was allowed in further modelling. Additional repeats would be recommended for this compound if further modelling was undertaken on this dataset, or an expansion of it, as this datapoint is highly leveraging.

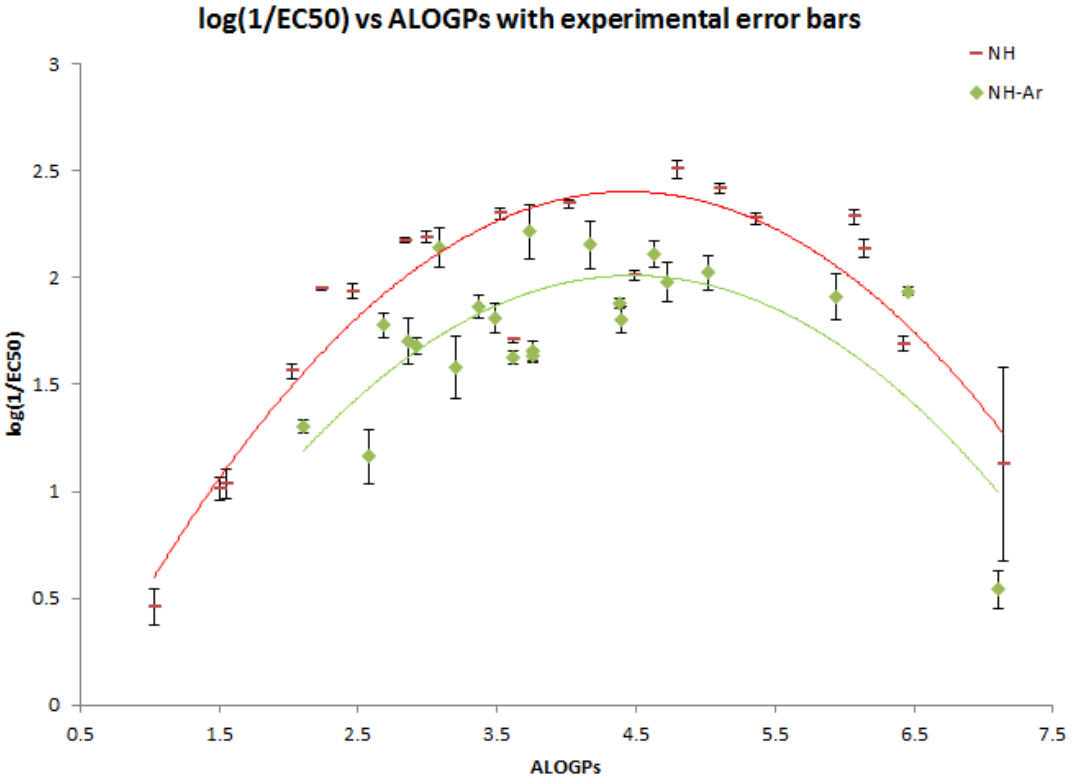


Figure 3.25: Plot of $\log(1/EC_{50})$ against ALOGPs showing error bars, split by enamine substituent

Considering the rest of the dataset, the errors would not have a large effect on the type of fit with the data still supporting the quadratic fit rather than a linear relationship. The magnitude of the errors also did not appear to be skewed to the extremes of either $\log(1/EC_{50})$ or ALOGPs.

The splitting of compounds through the enamine substituent, examined in 'Classification of Compounds', was carried out, shown in Figure 3.25. It indicates that most of the compounds with larger errors belong to the the NH-Ar group (excluding compound 43 from consideration). The errors in the NH group strongly support the quadratic fit; however, the errors in NH-Ar group show more variation around the curve in addition to the weaker fit. Further investigation would be advisable for this group to ensure these errors were correctly calculated from the Hill curves and investigate possible causes for the larger error within this group.

As with the errors discussed in Section 2.4.1.4 these errors are only from the calculation of EC_{50} and do not fully propagate the errors from the individual experiments. In this dataset some compounds only contained a single repeat of the Hill plot experiment (example curves in Figure 3.18 & 3.19) and as such the reliability of the results cannot be confirmed with a high degree of certainty. Further repeats of the experiments would

be advised for the production of robust models; however, care must be taken with the propagation of errors from multiple repeats to ensure appropriate errors are calculated.

Additional investigation should also be done into the errors in the underlying data. However, as discussed in Section 3.7, the process of examining the raw data is extremely time-consuming. Due to this, closer examination of the experimental error was not undertaken at this time.

3.9 Combined Anion Transporter Dataset

To allow expansion of the classifications examined in this chapter, the anion transporter dataset was expanded by combining this dataset with the anion transporters studied in Chapter 2. The transport experiments for these compounds were all carried out with the same procedures so their transport abilities (EC_{50}) can be compared. The combination of these two sets gave a database of 157 compounds.

The database was further expanded through the extraction and addition of new compounds from the Gale group which had been published since the initial data extraction, as well a number of tambjamine compounds that were not previously published.

In new research [149] a different assay was developed to measure the chloride ion transport rather than the Cl^-/NO_3^- exchange method used in the initial papers. The change was due to finding that the nitrate transport in the Cl^-/NO_3^- exchange was rate-limiting for some compounds. [150] This change in methods meant it would not be possible to directly compare the data using the new method to the previous EC_{50} values.

There were 9 additional papers which used the existing method of determining chloride transport. [151–158] From these papers 52 compounds were extracted following the same procedure as the initial anion transport dataset (see Section 2.2).

These were combined with the initial compounds extracted from the Gale group papers (Section 2.2) and the tambjamines examined in this Chapter. Following exclusion of duplicates the dataset contained 199 compounds, of these 160 had EC_{50} measurements.

All compounds were categorised by compound type; based upon the paper and existing compound groups. Most of these new compounds belonged to the urea, thiourea and squaramide compound types, but there were also completely new compound groups including isophthalamides [155] and perenosins [151]

Figure 3.26 shows the resultant plot of $\log(1/EC_{50})$ vs ALOGPs for the combined anion dataset coloured by compound group. Groups with < 3 compounds were excluded from the plot for clarity. It can be seen that the compounds cover a diverse range of anion transport efficiencies and lipophilicity values.

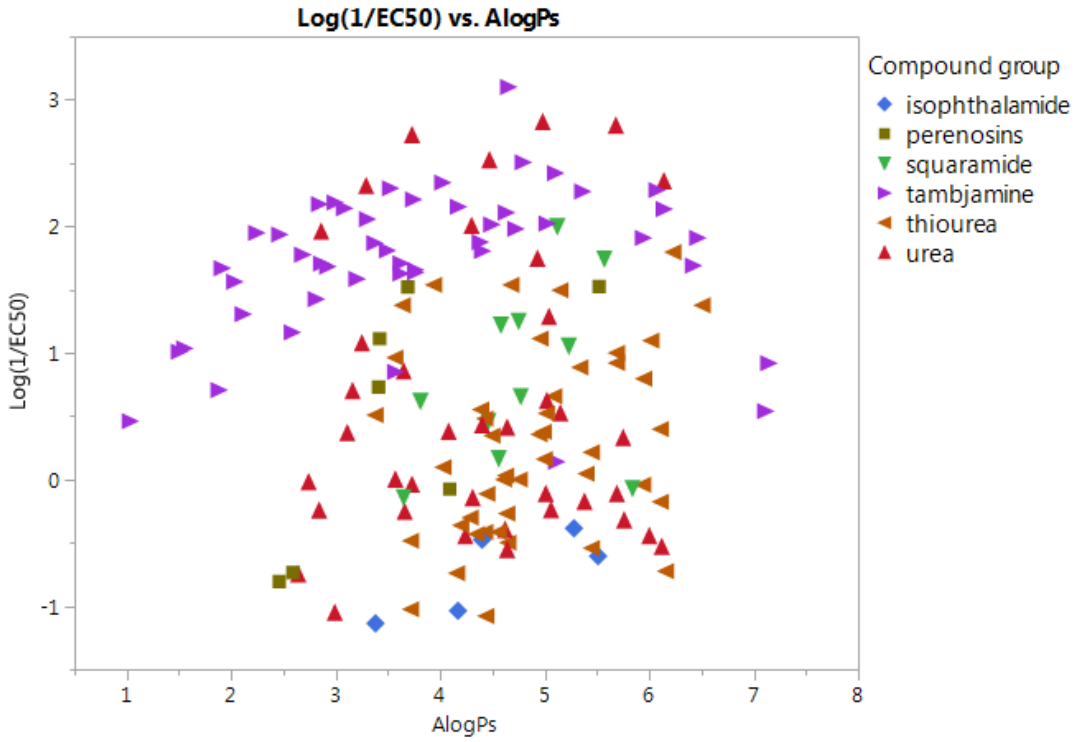


Figure 3.26: Combined dataset (initial Gale, tambjamines and new Gale compound) - plot of ALOGPs vs $\log(1/EC_{50})$ coloured by compound group

Significant further categorisation would be required before modelling of the parabolas could be attempted. Particularly in the thiourea and urea parent groups which encompass a significant variety of compound subtypes. However, as discovered in the research in Chapter 2 the use of compound subtype as a classifier requires more investigation to ensure that the same method of classification is used across all compounds.

In addition to classification by the chemical functional groups, the physical action of the compounds should be investigated as discussed in Section 2.4.4.1. There is evidence in some papers that some compounds may not act via a mobile carrier mechanism which could impact significantly on the model. [102, 155]

3.10 Discussion & Future Work

This area of research successfully examined a set of tambjamine compounds with a common structural backbone, using QSAR analysis to gain additional insight into the behaviour of these molecules with a focus on the affect of lipophilicity on the anion transport ability.

Examination of existing models highlighted a number of important issues in the model process:

- Researchers must endeavour to always include sufficient and unambiguous information on their processes and files to allow another researcher to reproduce their work.
- Test/training set splitting were not optimal for this data due to leveraging datapoints at the extremes of the dataset.
- Regions of the lipophilicity range were sparsely populated - this was mitigated by the synthesis of additional compounds.

Modelling the full dataset highlighted the importance of the lipophilicity descriptor; however, the models were not significantly improved by the by the addition of a 3rd or 4th parameter with multiple parameters all giving similar minor improvement to the model. The two-parameter model of ALOGPs and ALOGPs-sq provided a strong position for further investigation.

Splitting in the lipophilicity plot was initially identified by the pKa enamine descriptor; however, these values were determined to have been calculated incorrectly. This prompted investigation into underlying structural features in the molecules that could cause splitting of the parabola.

Splitting was carried out by the individual substituents at 3 points on the core backbone. This gave the most pronounced effect for the R5 (enamine substituent) where the non-aromatic (NH)/ aromatic (NH-Ar) substituent change indicated a shift in the position of the peak effectiveness, $\log(1/EC_{50})$ with the same optimum log P value for the two groups, this aligned with previous observations of an optimum log P of 4.2 [138]. The shift in peak $\log(1/EC_{50})$ was modelled through the use of mixed effect models, fitting the parabola curve with the whole dataset and using the subsets to determine the vertical shift. The R4 (ring substituent) change from OMe to OBn indicated the possibility of a shift in the position of the optimum log P value; however, this was not supported by enough datapoints.

Classifying the substitution pattern fully by the combination of the three substituents gave a number of multiple promising fits; however, many of the subsets did not contain sufficient datapoints for robust conclusions to be made.

The equations generated from full classification could be transformed into a form defined in terms of a maximum and a parameter that describes the shifts away from this point, such as:

$$y = ymax + k(x - xmax)^2$$

Where $y = log(1/EC_{50})$, ymax is the maximum $log(1/EC_{50})$ value, k is the slope, x is ALOGPs, and xmax is the optimum ALOGPs value, which produces the highest $1/log(EC_{50})$ value.

For carrying out this further research it would be beneficial to synthesise additional compounds with specific substitutions to give a larger number of compounds within a structural subset. This research did not begin with the intent of splitting the compounds into subsets, as such the dataset had a few groups that were well populated and a large number of groups that only contained 1 or 2 compounds. Synthesising additional compounds with specific alterations, in particular non-alkyl substituents, could further populate the smaller subsets. This would allow further identification and comparison of the parameters for different subsets and whether or not these parameters could be modelled to predict the values for a compound not present in an existing subset. Expansion would also allow use of external validation methods, as traditional external validation was not possible on the majority of the sets as it would have reduced the number of datapoints available for model building to an unacceptable level.

Additional compounds that are synthesised should not only be selected from compounds that are expected to perform well. It is necessary to also create compounds that are bad anion transporters as this provides a more complete picture of the transporter behaviours.

Continued examination should be made into the raw data obtained in the calculation of the EC_{50} and k_{ini} values as there may be more complex interactions occurring which provide further information to explain variation in the fits. In particular, additional investigation is needed into the chloride efflux plots and whether additional parameters from the fits should be incorporated if k_{ini} is used in models.

Other avenues of investigation include: examination of chloride binding and molecular shape. Values for chloride binding were not available for these compounds. However, measuring these could provide important insight into the overall transport ability as the process of transporting an anion relies on both the compounds ability to bind an anion and its ability to pass through the membrane. Molecular size/shape is something that was affected by the changes in substituent, use of molecular simulations for a selection of models would give further insight into how the substituent changes affect the overall size of the molecule and any impact they have on the anion binding site.

The tambjamine data obtained here were also combined with the anion transporter data collected in Chapter 2 and newly synthesised compounds to create a database of 199

anion transporter compounds. This combination of datasets gives a large amount of information from which further insight could be gained.

Although the analysis of the Gale group dataset did not provide models that were as promising as the tambjamine dataset, the insights from the tambjamine dataset can hopefully be applied to the expanded dataset to provide a better understanding of the whole dataset and identify if the effect of the presence of 'substituents' or chemical groups can be generalised when the backbone structure is not uniform.

For the application of models across the wider dataset there are a number of areas that should be investigated more closely. These relate firstly to the underlying data behind all compounds. The presence of many possible binding modes, methods of actions and additional interactions in the wider datasets makes it important to identify the behaviour which the model should apply to and flag any compounds that lie outside of this behaviour. Including compounds which operate via different actions as though they were part of a single group will distort the models.

Secondly the determination of groupings will need further attention. In the tambjamine dataset this was carried out through changes on a structural backbone; however, the expanded dataset does not contain a backbone across all structures. Compound subtypes were investigated but these require a more rigorous approach to their assignment. Classification using a taxonomy such as ClassyFire [27] should be investigated further to see if this can provide adequate classification.

Chapter 4

Data Handling and Visualisation

In any work, irrespective of the field of research; the creation, storage and use of data is central to all work. Since the data are of key importance to the research it is necessary to ensure that the data is handled adequately from the point of creation to the point of dissemination [159].

Throughout this research observations have been made about data storage and access, note-keeping and data visualisation which are important aspects in the data lifecycle.

4.1 Data Storage and Access

The proper storage of data is a crucial requirement to allow subsequent retrieval and analysis, which in itself is a crucial aspect of research. The need for well-curated data applies not only to the scientist creating the data but also to any researcher wanting to use the data at a later stage.

As computer systems have progressed in their development, the quantity of chemical data produced has massively increased [160]. The expansion in data means it is more important to ensure that the data produced are adequately stored, otherwise the data may be lost entirely or become meaningless strings of numbers.

Although the quantity of data has exploded, the methods through which these data are handled, stored and analysed has not evolved at the same pace. In chemistry the adoption of data sharing and open access has been slow in uptake [161], mainly driven by policies from institutions and funding bodies.

As discovered in the QSAR sections of this research it was not always easy to access the data underpinning scientific research nor simple to extract the data in a fashion that allowed further examination or verification. Data in this respect can refer not only to

the actual measurements but also chemical structures and in the case of processed data the source code.

To facilitate good storage of data and allow verification and reproduction of work by an external party a number of observations have been made about principles that should be adhered to as much as possible, some of these are general principles and some apply mainly to publication or dissemination of data: [114, 161, 162]

- Always save data files in an openly accessible format rather than a proprietary software file, this allows easy access and maximises future-proofing
- Ensure files are backed up to minimise risk of loss, many institutions have hosted file systems designed with this in mind
- Store files in a logical system, preferably with an index or key
- Include units - if these cannot be included unambiguously in a datatable ensure that the format of the data is fully described alongside the datatable
- Always provide structures in machine readable form, in a commonly used format (SDF or InChI) rather than, or in addition to image/pdf format
- Include data in a data format rather than pasting tables to pdf
- Where possible include script files or parameters for processing data in readily available software
- When adding data to a repository include metadata which describes the data

4.2 Note-keeping

Alongside the production and storage of datafiles, it is also necessary to record the notes of the researcher as they carry out their work. This not only provides context to the data but also parameters under which the data was obtained and frequently additional information that does not get included in the publication of the research. [163]

In the case of scientific experiments thorough note-keeping is often a legal and safety requirement to ensure that sufficient documentation is kept in case of an accident or incident. They are also used in cases to support intellectual property claims or in cases of scientific misconduct. But even in areas where the legal aspects are not as thorough, the scientific notes serve as a record of the thinking and process of the researcher. The content of these notes, along with notes of failures and success, can aid further research by that researcher and others.

Notation not only consists of the comments that researchers make or the steps they took in a procedure but can also encompass a wide variety of other items including; equations, tables, diagrams, photos and spectra. In the pursuit of the ‘perfect’ laboratory notebook

both historically and in the modern age, entries should be; clear and easy to understand, dated, unambiguous in their content and unable to be secretly edited. [164]

Many companies in industry have already made the move from paper based systems to electronic lab notebooks (ELNs). [165] But in research environments the recording of notes is very fragmented with many different recording options being used within a university or department. Many researchers still opt to record their work in paper notebooks [166] but this is changing as the shift towards digital interaction continues. [167]

Throughout this research a number of different recording mechanisms were utilised ranging from paper, to electronic and web based options. The use of multiple methods allowed discussion on their benefits for note-taking, here follows commentary on a number of the different systems that were used. The methods included paper notes, two blog based systems, an electronic note system and a markup system for code. This did not include a 'traditional' ELN, such as those offered by IDBS[1] and PerkinElmer[2] as they were not available in the department. However, they are designed with practical chemistry in mind and are not particularly suited to non-experimental research.

4.2.1 Paper Notebooks

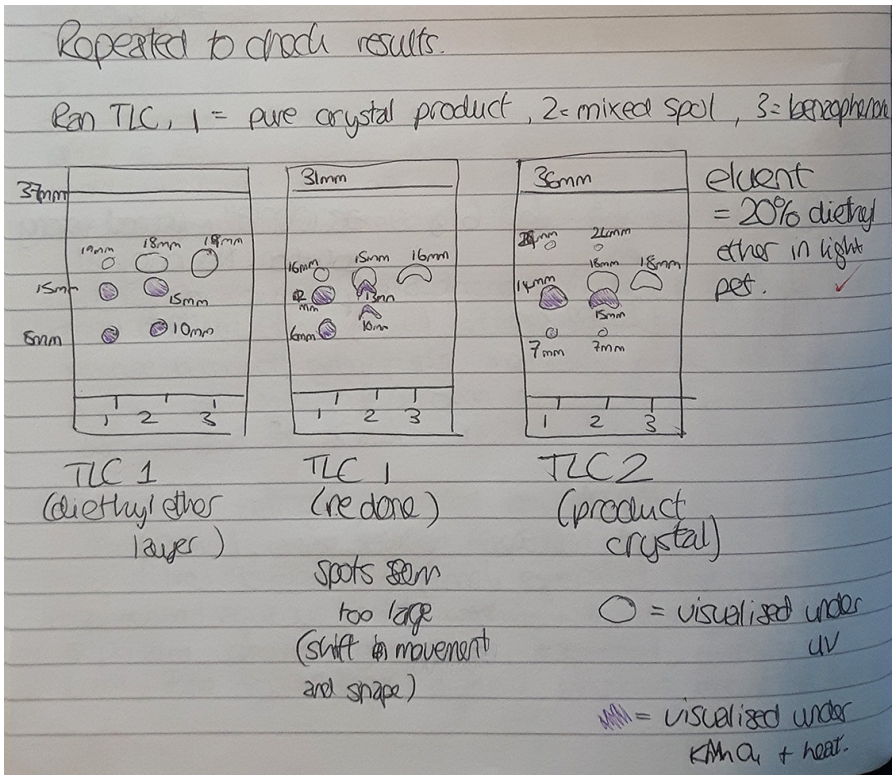
The traditional method for note-keeping is paper with notes being recorded in paper lab notebooks by hand. This was how scientists recorded their notes for centuries and still continues to be a widely used method today. [168, 169]

The advantages of paper notes include the flexibility for the user to record notes in whatever fashion they wish. Many different types of notation can be made in a paper notebook, Figure 4.1 and Figure 4.2 show examples of records. Alongside procedural notes and comments, researchers can record chemical structures, sketches, equations and data tables. Once entries are made it is obvious if the values are edited at a later time.

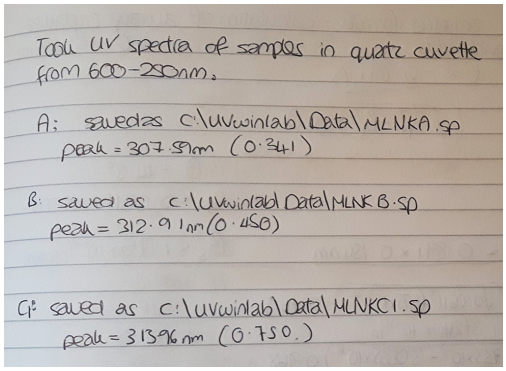
Although the paper notebook is probably the most familiar method for a researcher there are a number of drawbacks to the system that may prompt users to move away from them. [170]

Backing up a paper notebook usually requires every page to be scanned in to create a digital copy. Without backup the researcher risks losing their notes as paper can easily be damaged, misplaced or degraded over time. Paper notes are also heavily reliant on good practices by the researchers, including the quality of handwriting and keeping the notebook in good condition.

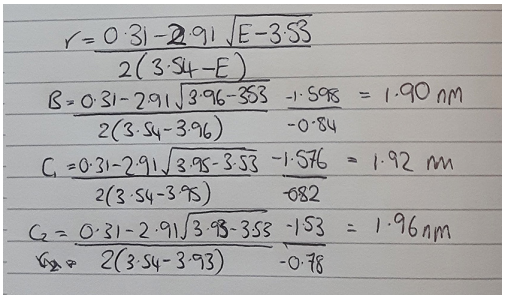
[1] E-Workbook ELN from IDBS
[2] E-Notebook for Chemistry or Signals by Perkin Elmer



(a) Sketches



(b) Filenames for spectra



(c) Equations

Figure 4.1: Examples of different types of notes taken in paper notebooks

Single measurements or small amounts of data can be recorded directly in a paper notebook; however, larger datafiles such as IR or NMR spectra cannot be stored in the notebook. These datafiles must be printed out which reduces the data resolution, or saved on a computer and referenced in the notebook. As the nature of the data continues to change towards more complex measurements and larger datasets the proportion of data that can be recorded in paper notebooks decreases.

A major feature absent in paper notebooks is the ability to search notes for specific words or phrases. Indexes can be created for the notebooks but this must be done by hand and is time consuming. Electronic notes can be indexed automatically making

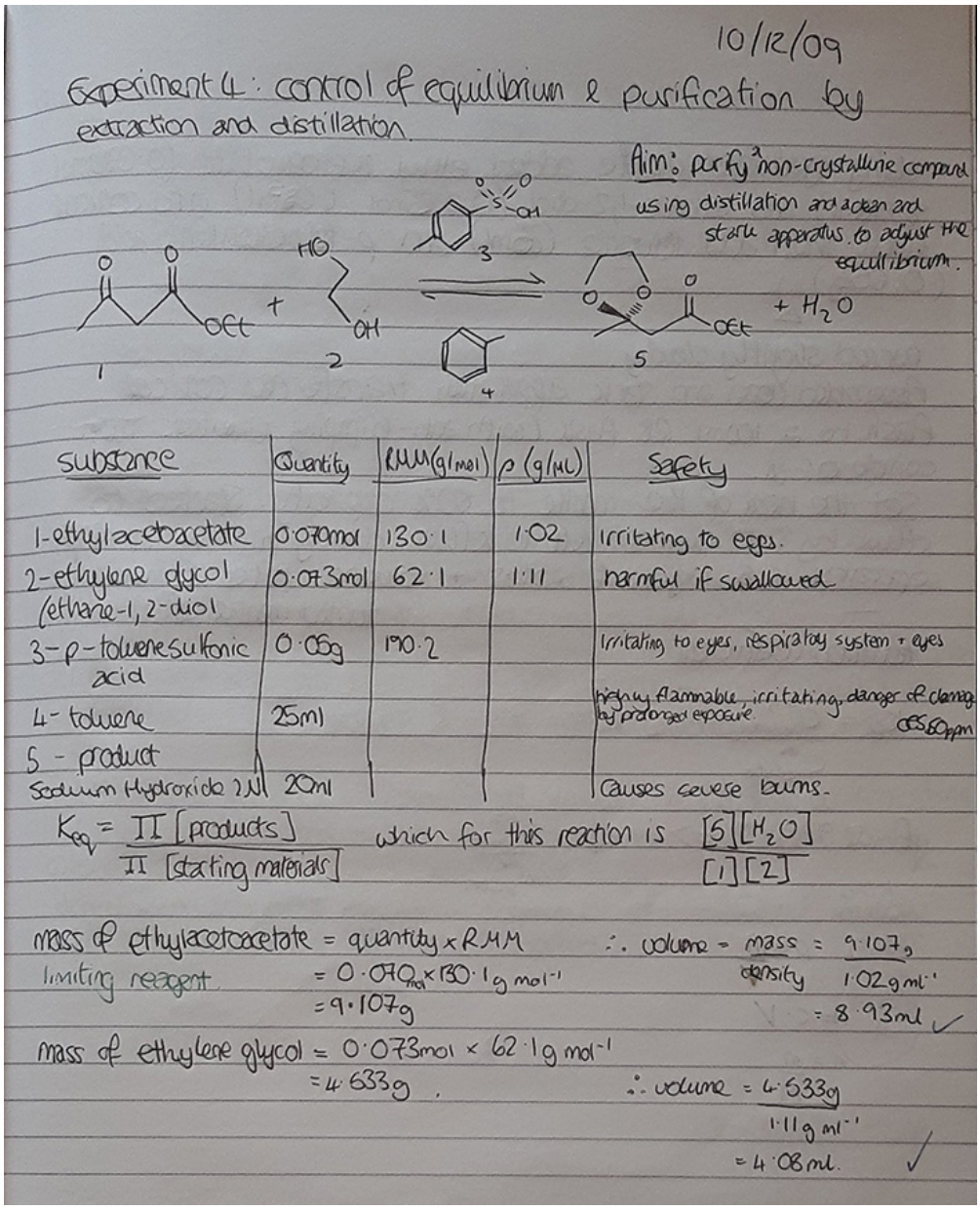


Figure 4.2: Example of an experimental record in a paper notebook

them searchable as soon as they are created. Additionally, the entire content of the notes can be searched, rather than just the items which have been selected for the paper index.

4.2.2 Electronic Notes

The use of electronic notetaking is not a solution that will fix all problems instantly. However, there are a number of functionalities that are added due to the digital nature of the notes. The extra functionality gained on top of this is very dependant on the particular ELN or software chosen.

- Can be searched
- Can be easily backed up (although they can still be corrupted)
- Can be shared
- Easier to track data and link between items
- Copy and paste functionality or use of templates to record repeated events and methodologies.

4.2.3 LabTrove

LabTrove was one of the electronic note-keeping methods used throughout this research. LabTrove is an Open Source blogging platform developed by the University of Southampton [24, 171, 172] The system is a Smart Research Framework (SRF) which has created infrastructure to support the collaborative research environment with a focus on making data the centre of the system.

The LabTrove system is implemented in a wide range of research environments, noteable uses being the Open source drug discovery project Open Source Malaria which is centred in Sydney, Australia [173, 174] but contributed to from across the world, and research blogs within the University of Southampton in multiple departments. [175]

The LabTrove system was designed to allow a single researcher or a group of researchers to electronically record their notes alongside metadata, attachments and direct data upload. The interface can be accessed via the web and has a similar style to blogging systems whereby users create posts in specific sections. These posts can include many aspects of formatting as well as images, tables and links to data attachments. The web-based nature of the LabTrove system always required internet access, in most research environments it is possible to have continuous internet access but this may be unsuitable for some researchers.

A key feature of LabTrove was the ability to add metadata to a post. Within each post metadata could be added in Key/Value pairs, Keys can be chosen from a list existing Keys or new ones can be created. Once a Key has been selected the Value can then be selected from existing value options, or a new one can be created, see Figure 4.3. This allows the researcher to tag the posts with metadata relevant to their work, this could include information about compounds used, techniques employed or which project/sub-project it was from. It also allows automated posts to be tagged with the relevant metadata.

Figure 4.4 shows an instance of LabTrove being utilised in the Talk2Lab project. This shows an example of a post including slight text formatting, links and custom tags for the posts. In this example some of the metadata for the posts can be seen in the right

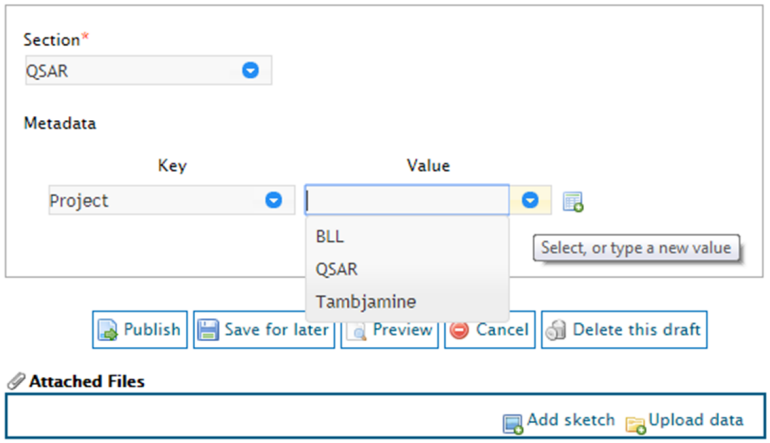


Figure 4.3: Addition of custom metadata to a LabTrove post

hand menu and these can be used to filter the posts in the blog. The archives, authors and sections tags are standard across the LabTrove blogs but the Sensor Type tag is custom metadata for this blog.

Although the system was powerful for its integration with data and ability to capture more information about experiments and procedures, the process was often time consuming to navigate and more complex than a paper system. Linking items such as tables and images into posts was difficult and may discourage inclusion if the user thinks it will take too long.

4.2.4 Blog[3]

Blog[3] is a blog system designed by Mark Borkum[3] to enable researchers to create and share blogs containing research notes. [23] It was designed to provide a system which improved the user experience from that of LabTrove which, although powerful, was not always easy to use. The resultant site had a appearance more similar to a 'traditional' blog site, allowing users to not only create blog posts but also pages which could be used to collate posts or write about the project.

The interface of Blog[3] was found to be more user friendly than that of the LabTrove system, with better integration for images, chemical structures, tables and equations. Although some of these features were subsequently improved within the LabTrove system [176], they were still not as straightforward to use as within the Blog[3] interface.

The web interface for Blog[3] can be seen in Figure 4.5 which shows the main framework of the blog site along with an example post. This example post shows use of formatting in addition to the ability to embed fullsize images in the posts, which LabTrove did not allow.

[3]Formerly a Post Doctoral researcher at University of Southampton

Figure 4.4: Labtrove web interface showing an example post and functionality

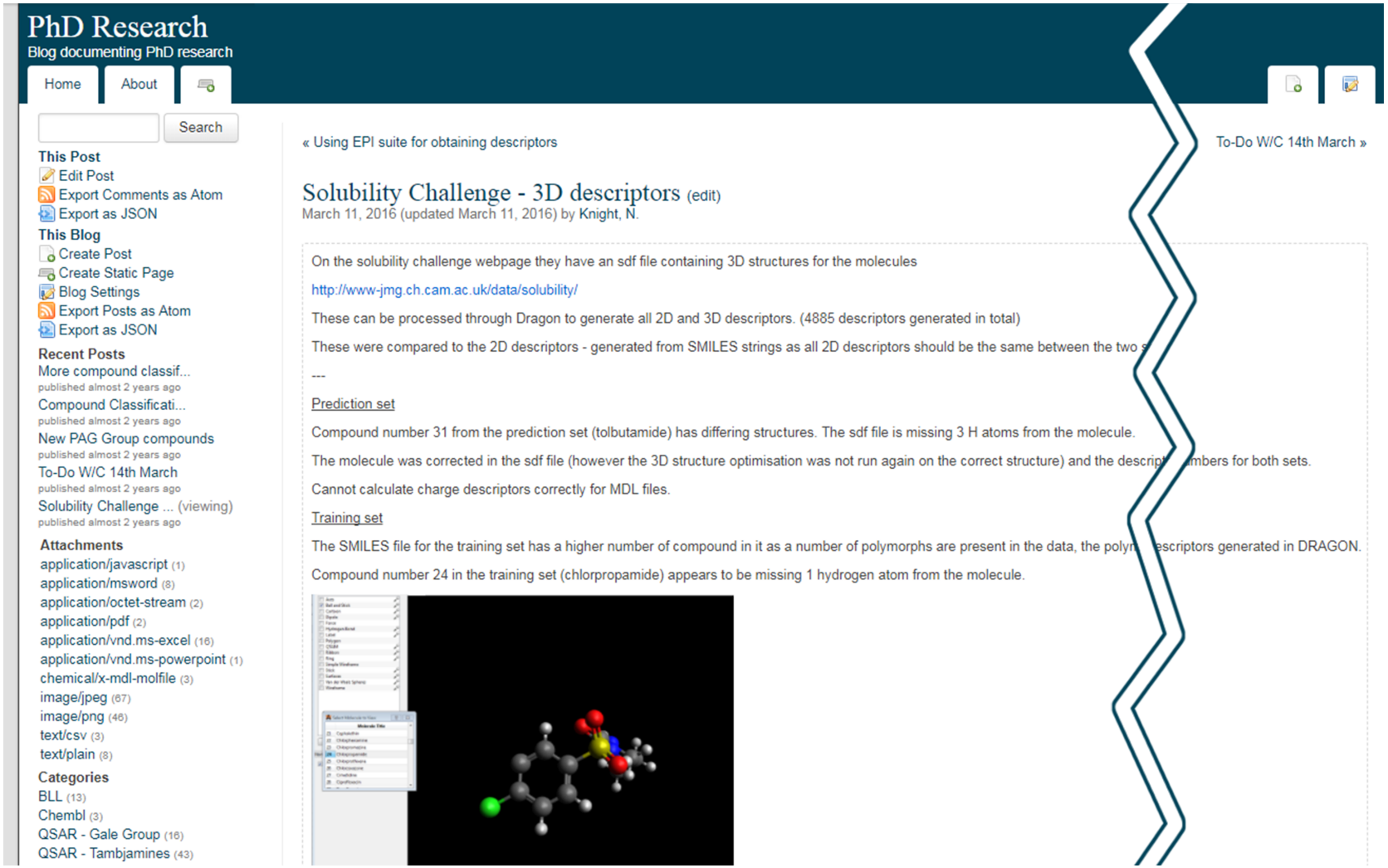


Figure 4.5: Blog[3] web interface showing an example post and functionality

Both the Blog[3] and LabTrove sites automatically dated entries as they were made. If a post was subsequently edited it would be recorded and old versions of the post could be viewed as well. This means corrections can be made but there is a record of it.

While the Blog[3] system showed great promise and was easier to interact with, in particular for including pictures and attachments, the system did not contain tools for tagging the posts in a custom manner like the LabTrove software allowed with the custom metadata. Posts could be labelled with tags but these did not contain the Key Value pairs like the LabTrove system.

It would have been possible to make technical modifications to the Blog[3] system to make it more useful to a researcher; however, the project did not have the resources to develop it significantly and following the departure of the developer from the university the system maintenance was not continued. This lack of upkeep limited the scope in which this system could be used as over time bugs developed which could not be fixed. Although the blog site is currently still active and posts that were created are still accessible they will be unlikely to remain there indefinitely.

The uncertainty of whether systems will continue to be available in the future highlights the need to be able to back up and extract data and records in a format that will allow them to be accessed through a different system in the future. Both the LabTrove and Blog[3] systems allowed export of the posts in bulk, but these exports could not easily be imported into an alternative system if a user wished to migrate their notes.

4.2.5 OneNote

OneNote [25] is a digital note-taking computer program from Microsoft designed to allow users to gather their notes (typed and handwritten), screen clippings, audio files and drawings together in a fashion that can be shared and accessed from many locations. Although the functionality of the programs vary slightly; OneNote shares many similarities with other note-taking programs such as Evernote[4] or BoxNotes[5].

Using the OneNote interface was beneficial for brainstorming and collating ideas as it was freeform and allowed easy screen clipping to pull ideas from multiple places whilst adding comments and notes to the pages. Figure 4.6 shows an example of notetaking in a OneNote Notebook. This shows some of the functionality that was available in OneNote. In this example it includes To-Do lists, text formatting, sections within a Notebook, Notebook pages and subpages.

The freeform nature of the note-taking often led to pages which appeared quite messy but templates could also be created for pages to help in structuring notes. These could

[4]https://evernote.com/
[5]https://www.box.com/en-gb/notes

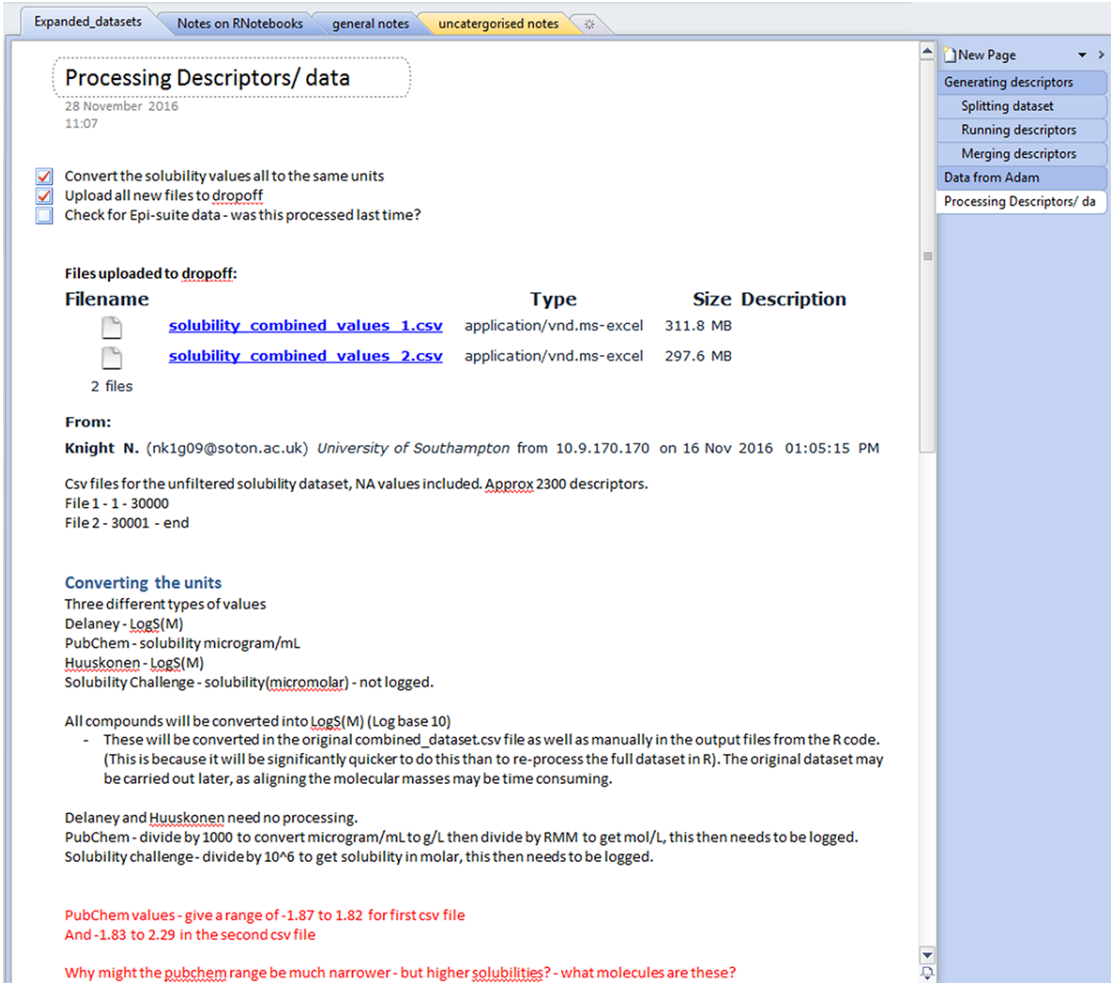


Figure 4.6: Example of notetaking in a OneNote Notebook

be used for taking day-to-day notes and experimental notes, particularly when repeating experiments of a similar nature. It also had a powerful search functionality which allowed location of notes even if they had not been well structured.

To keep track of notes pages were also automatically dated when they are created, along with screen clippings and copied text. However, these dates could be easily removed or edited, which will not always be sufficient for date recording. Notebooks could also be backed up to keep previous versions of pages and the databases can be shared to allow collaboration or storage on a department server.

The functionality of this program was impressive; but it was not designed with scientific recording in mind so adding and linking data was not particularly easy. As such it was mostly used for general purpose notes and brainstorming. However, it became quite difficult to manage the notebook when it contained a significant number of pages.

4.2.6 RNotebooks

RNotebooks were a recent addition to the R Markdown notebook interface from RStudio [26,177] which allowed users to bring together text and code to produce high quality reports and documents in a number of different formats including .html and .pdf. RNotebooks allow users to closely associate their code and output, and intersperse this code with commentary and notes. This is a different style of notebook to the other solutions, with a heavy focus on the integration of code, which is of benefit to theoretical scientists.

Figure 4.7 shows a segment of output that was created through an RNotebook. This was .html format displayed in a browser and shows the presence of text, R code and the output produced by R code all together in a single document. A full RNotebook file and the code used to produce it can be found in the ESI. [30]

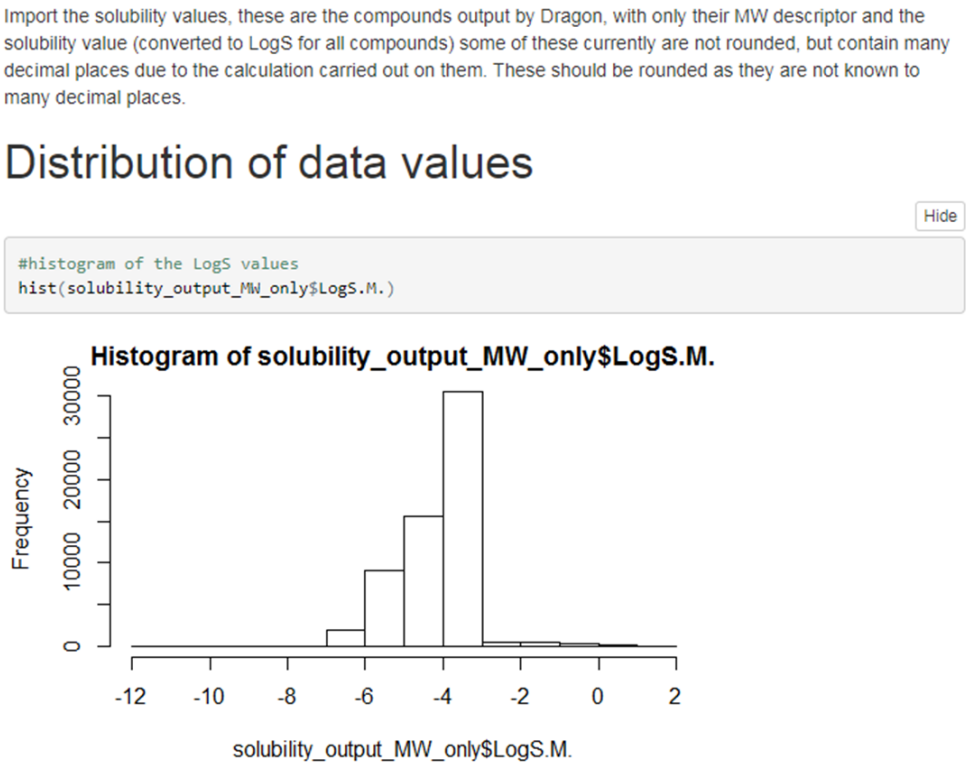


Figure 4.7: Example segment of HTML output from an RNotebook

RNotebooks were exceptionally useful addition to RStudio and opened up the possibility to create a full analysis workflow in R from input of data to the production of reports with the use of the many available R packages. What in principle could be done; however, was not always straightforward to achieve. The use of R on a Windows Enterprise system presented many problems when installing and using packages, frequently requiring many hours of debugging to ensure that all the integrations were functioning correctly.

Once functioning the system was no more complex than using R normally and provided much easier annotation to analysis than the use of simple .R scripts. In addition it easily generated graphs and plots inline with the code. While the use of RNotebook was straightforward, in depth knowledge of R and its packages would be required to use RNotebook for all notes. Inclusion of other data formats and linking to data are more complex but still possible.

Unlike the other systems used, each RNotebook file is created separately so must be stored in a sensible file hierarchy otherwise it could easily be lost; it also lacked in-built version control and backup, but works well with repository systems like Git. Additionally, although searches could be carried out within a single file they cannot be carried out across all RNotebooks.

4.2.7 Discussion

After use of multiple systems there was no single system which stood out as being the best option. The selection of a system is very dependant on user preference and application, with a huge number of technological solutions available. If a user is comfortable with a system then they are more likely to interact with it to the best of their ability. No one system was perfect, but the continuous use of a 'sub-optimal' system is better than the continual search for the perfect solution.

Although paper was very familiar and convenient it was frustrating to have to manually search through to locate notes every time. In this respect the electronic systems were far superior. However, with the electronic systems the benefits are only fully realised after the system has been populated with a reasonable quantity of data, these include the searching and filtering functionalities which can be used to find previous work or similar experiments/notes.

Functionality in the LabTrove system such as the ability to create custom metadata and the possibility of integration with other aspects of the lab were powerful features that were very useful in the creation and storage of notes. Unfortunately the system was let down by the interface as it was not smooth to work with. The ideal notetaking solution would simplify the users experience whilst adding to the usability. If it requires more effort on the behalf of the user then the system must generate even more benefits for it to seem worth it.

RNotebooks were the most suitable solution for much of research carried out as it pertained to analysis and the creation of code. But the program was rendered almost unusable by the many errors that were generated. In the future the combination of RNotebooks+Git for work involving code and another electronic solution which could

handle chemical structures and images would be a beneficial solution for the style of research carried out throughout this PhD.

Due to the time required to learn the nuances of a new system researchers will likely continue to use a wide variety of different systems until they are forced to use specific software by their department or a standout solution establishes itself in the note-taking matter. It is also unlikely that paper notebooks will disappear entirely from the research environment as they are simple and familiar.

4.3 Data Visualisation with d3.js

Visualisation is the graphical or pictorial representation of information, designed to provide the viewer with a qualitative understanding of the contents. It has been used in scientific research for many centuries with scientists using graphs and charts to plot their results; however, as the amount and type of data available has expanded so have the methods of visualisation.

Data visualisation is often used to allow viewers to digest larger quantities of data by presenting it in a fashion that allows easy identification of patterns and trends. Due to how the brain processes images, it is often easier to process the information from a graphical rather than numeric representation. [178]

Data visualisation is frequently used in businesses to identify emerging market trends, predict sales and identify areas that require attention. [179] However data visualisation has moved from being solely used in academia and business to become a mainstream tool that people are exposed to on a daily basis. Examples of everyday visualisations include use in news articles, graphical depictions in weather forecasts, and infographics which combine data and text to provide the reader with information. [180, 181]

Why do we want to do data visualisation?

- help the brain process data and gain insights, for larger datasets it can help with the comprehension of large amounts of data
- help identify relationships and patterns
- can contain so much more information in a dynamic visualisation compared to a static plot

During the completion of the QSAR modelling it was often difficult and time consuming to identify which compound was signified by a specific point on a statistical plot. Within the statistics program JMP it was possible to identify the line number of a point, checking in the data table would then allow the determination of the compound number from the line number. To then identify the structure the compound number must be looked

up in a directory to open the .mol file in a viewer. This was a laborious task for simply wanting to identify the structure of a compound, which is a common occurrence in QSAR analysis.

This prompted investigation into solutions that would create smooth data visualisation processes, producing simple visualisations which presented the data in a clear manner and allowed the user to gain more insight into the data. In particular code was developed to produce 'chemically-aware' visualisations that could plot data whilst making use of the chemical descriptors and identifiers such as compound numbers and InChI strings. The visualisation developed within the QSAR analysis research, focused mainly on the tambjamine dataset, but was also applied to the larger set of anion transporter molecules which included the Gale group compounds.

4.3.1 D3 Javascript Library

D3 (Data driven documents), also known as d3.js, is an Open-Source JavaScript library [17] which was developed to allow the creation of HTML-embedded interactive data visualisations. D3 visualisations are built using javascript code, making use of the widely implemented web standards (HTML5, CSS, SVG) to create animatable HTML pages.

In particular SVG images are used for the implementation of interactivity. SVG is Scalable Vector Graphics, which define vector based graphics in an XML format. Every element that is contained within an SVG can be animated. D3 is a particularly powerful visualisation tool as it allows dynamic manipulation of graphics and documents based on underlying data. In D3 there is no standard form of visualisation; it provides the building blocks to create any visualisation that the user can dream of, from simple bar charts[6] to interactive weather forecasts[7].

Although it is possible to create exceptionally complex visualisations through D3, a simple scatter plot was selected as the first visualisation for development as this mirrored the regression analysis. The concept behind the visualisation was the creation of 'chemically-aware data visualisation' which joined the graphical depiction of the data with descriptors and identifiers that were of use to the scientist.

This combination of data and chemical identifiers aimed to address some of the barriers that were encountered in visualisation during QSAR analysis, in particular, where insight into data was hindered by the process of identifying datapoints. D3 could also be used to develop more complex visualisations which probe further into the data relationships; however, these were not the focus of this work.

[6] example at: `https://bl.ocks.org/mbostock/3885304`
[7] can be viewed at: `https://www.ventusky.com/`

The tambjamine dataset from Chapter 3 was used in the data visualisation development. This was chosen as the data visualisation was designed to aid QSAR analysis and the tambjamine dataset was a manageable size for initial development.

4.3.2 Interactive plots

The D3 scripts that were created used the tambjamine data spreadsheet in conjunction with custom code to produce an HTML-embedded SVG. The tambjamine examples focused on the relationship between lipophilicity and transport efficiency as this was the key relationship of interest for the tambjamine compounds. The initial D3 visualisation for the tambjamine dataset is shown in Figure 4.8.

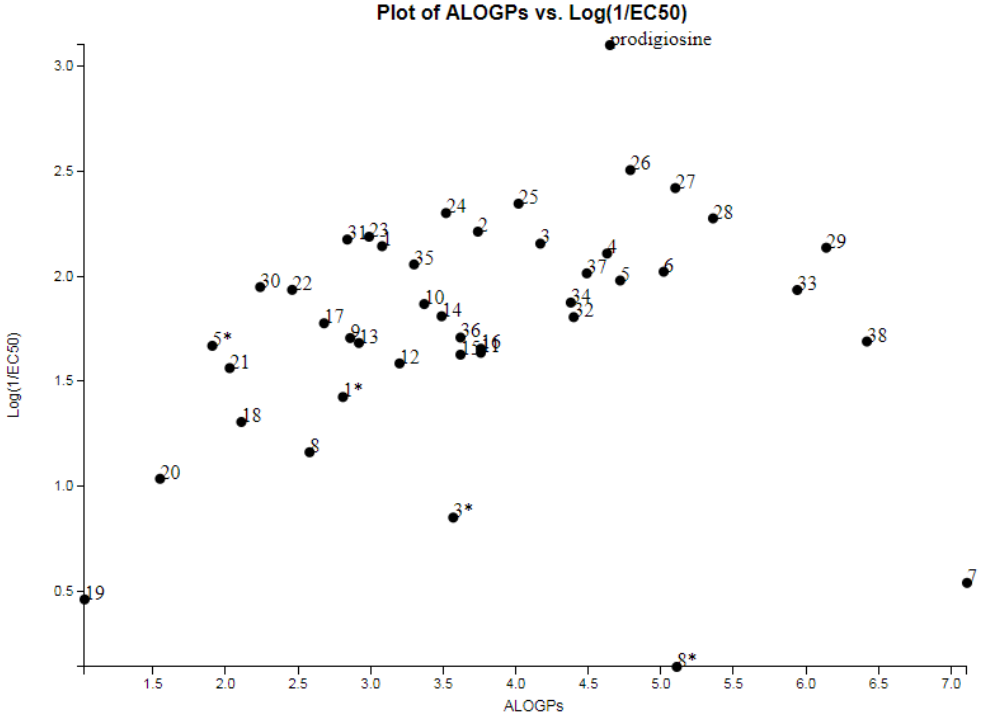


Figure 4.8: Initial D3 static plot of tambjamine compounds with compound numbers

This first example was a static plot of the data which was generated on loading the html page, but did not include any interactivity. It was configured to display axes which scaled to the range of values in the data and created elements on the graph for each data point in the dataset, with positions that corresponded to their values.

This initial plot did not provide much benefit to the viewer compared to similar excel plot. However, code was included which displayed the property 'Compound Name' next to each plotted point. The inclusion of this did provide additional insight to the data, which would not have been available from a plot in Excel or JMP.

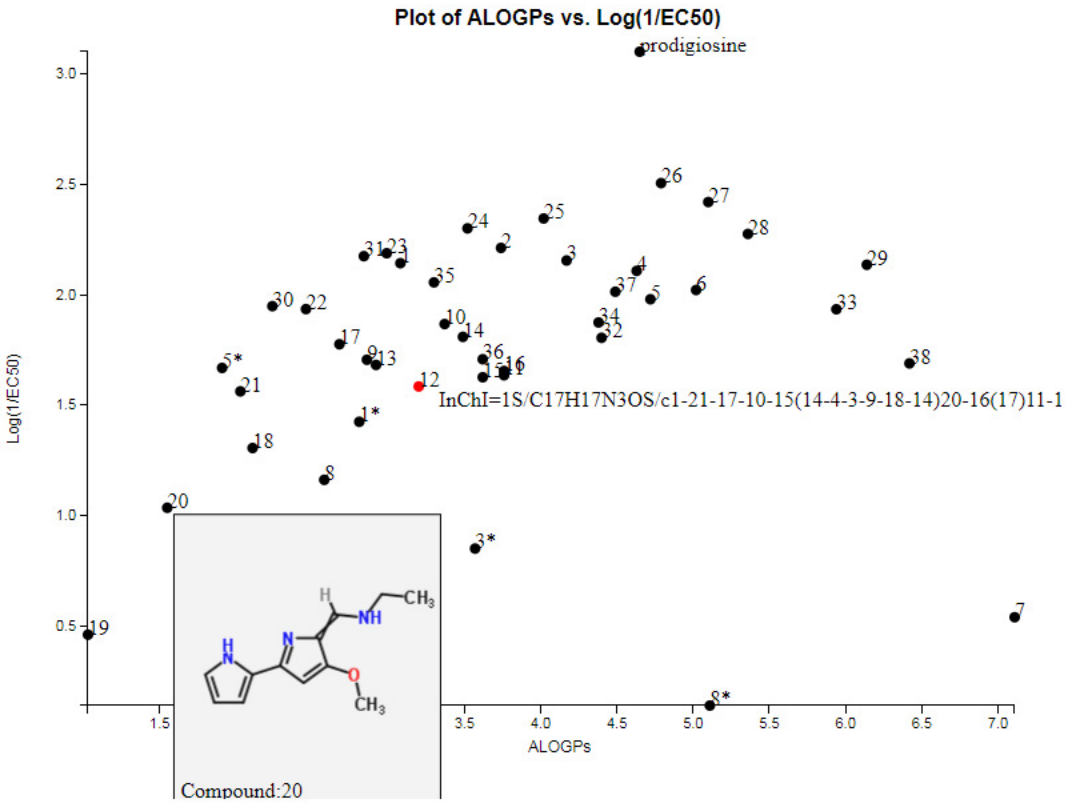


Figure 4.9: Interactive functionality of the d3.js data visualisation - on.hover InChI display, on.click molecular structure display

To create 'chemically-aware' visualisations that would allow further insight, the code was expanded to include interactive functionality. Many of the actions in D3 are triggered by events in the web page environment, the list of events that can be used is lengthy [182] but these visualisations focused on the mouse interaction events.

A key concept in the creation of the D3 visualisation was the ability to assign elements of the SVG with attributes derived from the tambjamine datafile. These attributes can then be used in conjunction with events to dynamically change the content of the SVG. The first plot with interactive functionality can be seen in Figure 4.9. A screenshot must be used as it is not possible to embed dynamic images in pdfs. Interactive versions of the HTML pages can be viewed in the ESI [30].[8]

The 'on.hover' mouse action was used to display the InChI string for the compound when its data point was scrolled over, this was combined with a colour change to highlight the datapoint. Once the mouse focus left the point the it was reset. Further information was added through the 'on.click' mouse action. Clicking a datapoint would display the compound number and its chemical structure. This structure would remain on the plot until it was clicked to remove it or another point was clicked. The chemical structure for

[8]The html files require use of server-side javascript - this can be implemented through use of a locally hosted server such as wamp

each compound was generated from the InChI string through the use of the ChemSpider Structure API [6].

The addition of interactivity was more useful to a scientist than the static plots as they could see further information about the compounds. In particular they could look at the structures of outlier points to gain insight. The structure representation generated wasn't always perfect, double bonds were sometimes depicted as crossed, but they are understandable to a chemist.

Both the 'on.hover' and 'on.click' actions positioned their information relative to the corresponding datapoint (diagonally down from them), this caused information for some points to be truncated by the edge of the SVG environment, as can been seen in Figure 4.9. InChI strings were long text strings which often extended beyond the edge of the frame. The compound structures were better positioned, but some datapoints generated structures positioned below the SVG boundary.

4.3.3 Responsive axes selection

To allow insight into other variable correlations the visualisation was expanded beyond lipophilicity to include access to a fuller descriptor set. This code was designed to allow the user to select the axes prior to the generation of the SVG image. The selected variables dictated the composition of the plot as the information and attributes were pulled from the underlying data.

The new script determined the possible variables by reading the underlying .csv file which contained compound data, these variables were offered as options for the x and y axes. On the web page dropdown lists were created, allowing the user to select x and y axes from the available variables. Once selected, the script generated an interactive plot showing the relationship of the two selected variables. Figure 4.10 shows the plot generated for nH vs. ALOGPs-sq, with the axis selection boxes visible above the graph area. The plot contained automatically scaled axes, axis titles and plot title that were generated from the selected x and y variables.

The interactivity of the plot contained the same hover and click functionality that was incorporated in the ALOGPs vs Log($1/EC_{50}$) plot. However, a number of modifications were made to the code controlling the display of compound information. These adjustments addressed the previous bugs, ensuring that compound information was always displayed within the plot area. InChI strings were wrapped so they did not trail off the side of the image, and the position variables for the compound structure were modified to account for how close a point was to the image edges.

Incorporating responsive axes highlighted a number of issues surrounding data inputs. For this style of visualisation the script could only use numeric values as the input. In

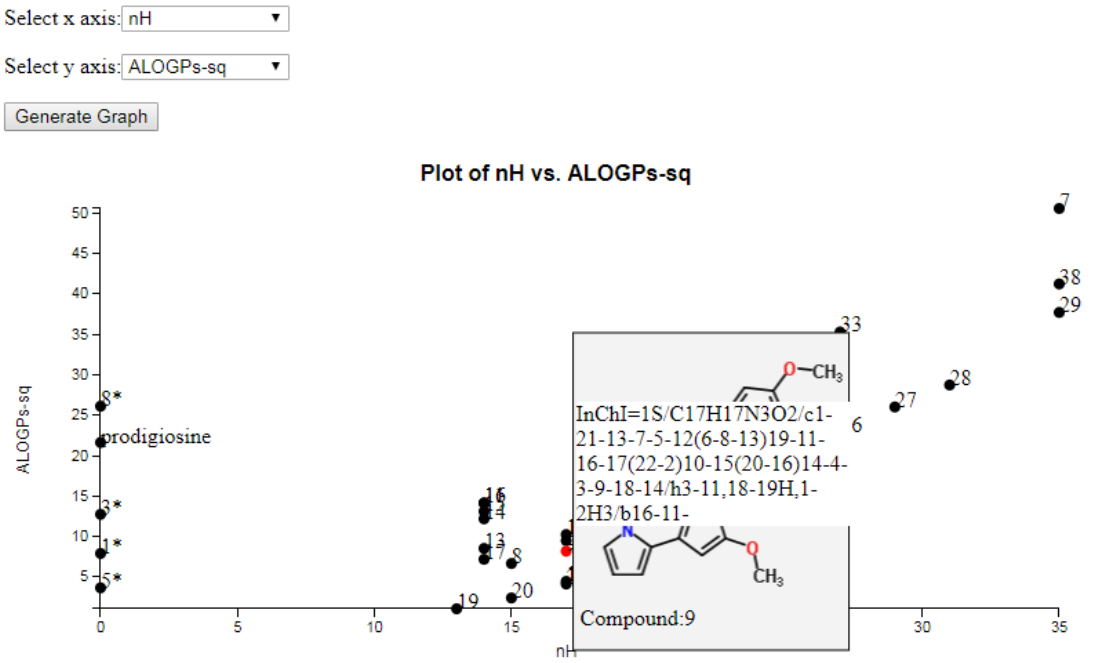


Figure 4.10: Expanded Functionality of data visualisation in D3 with interactive axes selection

the initial ALogPs vs Log($1/EC_{50}$) plot the variables had been checked to only contain numeric values and no null entries. However, when the data source was expanded to the full dataset a number of variables were non-numeric (such as compound group, InChI string and molecular formula) and some had absent values (such as the nH variable in Figure 4.10) These would need to be resolved to provide a higher quality visualisation.

4.3.4 Plotting compound similarities

With a view to gaining further insight into the data the visualisation was extended beyond single correlations. A key aspect of QSAR is the assumption that similar molecules have similar activities and behaviour. While this is not always the case [33], viewing similarity would be beneficial in analysis of datasets. While similarity can be measured in a number of different ways [183], structural similarity was investigated in the next visualisation.

Using similarities in visualisations required the creation of a similarity matrix. This matrix contained the pairwise similarities between all of the compounds in the dataset. It was created by looping through all of the paired combinations of molecules and calculating the Tanimoto similarity coefficient using fingerprints generated in OpenBabel.[9]

[9]Similar to the work in Section 2.4.3

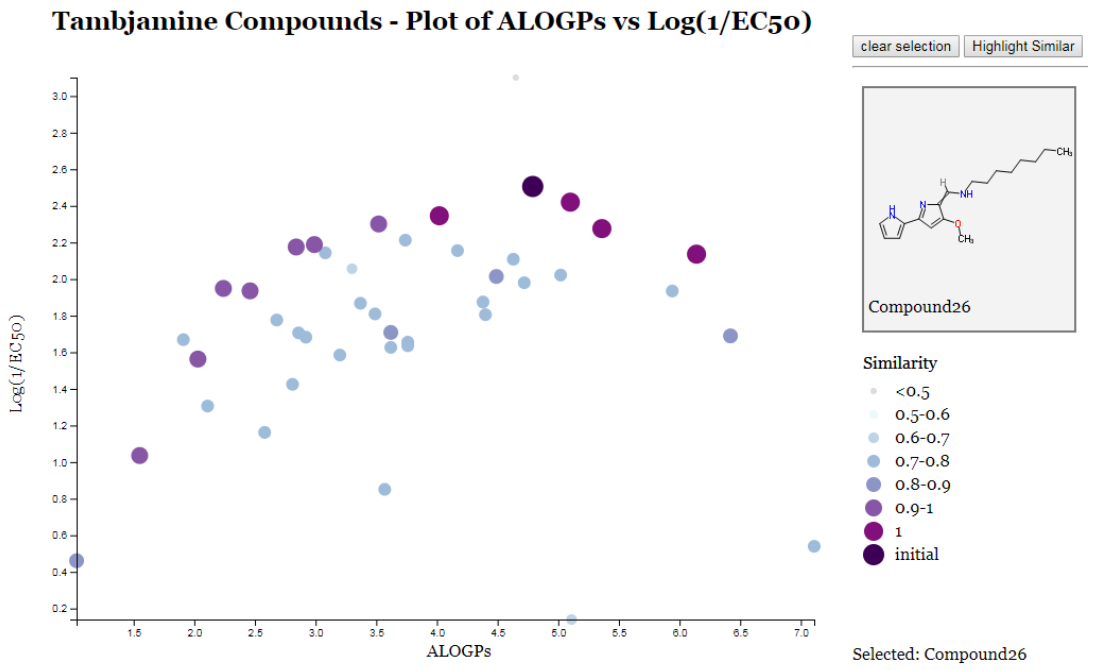


Figure 4.11: Data Visualisation for Tambjamine compounds showing similarity highlighting

4.3.4.1 Scatter plot

Figure 4.11 shows a plot which highlights molecular similarity to the selected compound. For this visualisation the 'on.click' display of compound details was moved to a panel on the side of the display, this displayed compound numbers and structures for the selected compound, but this could be expanded to show other pertinent descriptors from the datafile. The 'on.hover' display was limited to compound number, as the InChI strings were found to not be very useful in the visualisations.

When a compound was selected and 'Highlight similar' was clicked the visualisation changed the appearance of all the datapoints to indicate their similarity to the selected molecule. The graph maintained its overall appearance but the size and fill colour of the points were adjusted, those compounds that were most similar to the initial compound became darker in colour and larger.

The inclusion of similarities was beneficial to aid identification of potential patterns and grouping in the data based upon similiar molecular structures. Although this plot used structural fingerprints the similarity could be expanded to include similarity based upon conformational descriptors or one dimensional properties.

4.3.4.2 Force directed graph

Until now all the visualisations used a scatter plot as their basis, which examined correlations between variables and, in the case of the similarity plot, the similarity between a single molecules and all others. But it would be useful to examine the similarity of all the molecules in the dataset and whether these begin to produce clusters. Initial work for this was carried out using force-directed graphs and structural similarity measures.

A force directed graph is a type of network graph, displaying graph connections through use of 'nodes' and 'links'. Nodes are distributed in space based upon attractive and repulsive forces between the nodes depending on the presence or absence of links. [184]

Figure 4.12 shows a force directed graph for the tambjamine dataset based upon structural similarity. Every compound is a node on the graph and a link was created between two compounds if the similarity was above the threshold (> 0.85). The force directed graph then distributed the points for the visualisation.

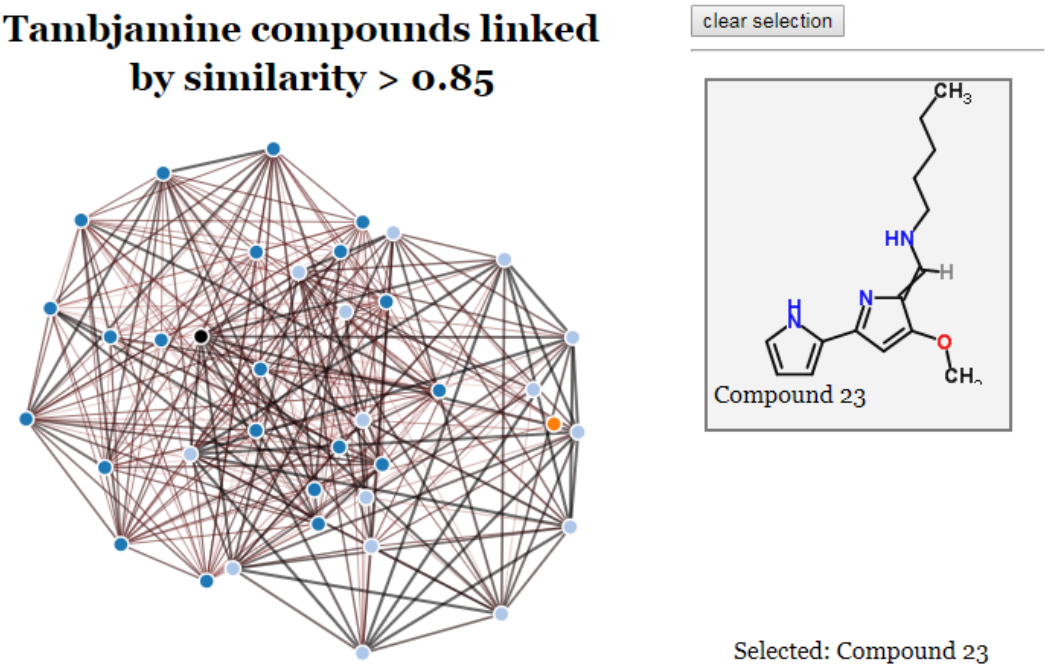


Figure 4.12: Visualisation of tambjamines showing connection by similarity, points coloured by NH substituent

In this version the nodes were coloured by NH-substituent with the colour of the link determined by the 'strength' of the similarity. This layout indicated the beginnings of two groups in the data, from the colours it seems these are the NH-Ph-R and NH-R groupings. Increasing the threshold to 0.9 showed distinct separated groups in the layout, with three clusters being formed. Two containing NH-R compounds and the other

the NH-Ph-R compounds, with the additional splitting aligning with the OMe/OBn presence.

While the observations from the structural similarity layouts are not particularly surprising given the structural classification carried out, the use of the force directed layout showed promise. Further insight into the data may be achieved through the use of similarity measures based on descriptors. Additionally the introduction of dynamic threshold selection would allow observations to made into how the grouping changes with respect to the similarity threshold.

4.3.5 Visualisation in other areas

Developing the 'proof of concept' for data visualisation using d3.js gave promising visualisations and aided in the examination of our tambjamine dataset. It allowed identification of the different points to be carried out more quickly, and provided useful information about the points.

Another area for application of the D3 visualisation was in conjunction with the classification work that was carried out in the QSAR analysis; both with the Gale group compounds (Chapter 2) and the combined transporter dataset (Section 3.9). The visualisations could be expanded to incorporate the ClassyFire classification groups and molecular similarities across the whole group of anion transporters. The use of force directed graphs and structural similarity could be beneficial to overcome the problems encountered with subset assignment.

D3 was also utilised in the Smart Lab Chapter of this research (Chapter 7). In that research the system handled dynamic data inputs from sensors to produce visualisation of the lab environment. The d3.js library was a powerful tool for charting the sensor data over time.

4.3.6 Expansion of functionality

To create a more cohesive visualisation a number of modifications could be made to the code, pulling together and expanding on functionality developed across the different sections of code.

- Allow selection of input file - once the file is selected the responsive axes code would identify available parameters for the plots. This would require development of filters to exclude non numeric columns and absent datapoints.
- User customisation of compound properties displayed - allow the user to select the properties from a list, these would be displayed in the right panel.

- Comparison of two points - incorporate a second panel to allow comparison of two molecules, showing their structures, properties and similarity.
- Save the state of a plot - set parameters within the page, so that specific views of a plot can be accessed directly from a link, this would allow better sharing and allow a user to return to a plot they created.
- Export plots - allow export of the plot in .svg or other static image format that can be used for documentation or publications. This has been implemented but requires expansion to capture the compound structures correctly in the .png format.

The visualisations developed here have the potential to be powerful tools for analysis, but they are contained in an isolated system and require the use of a webserver to access the interactive plots. A number of improvements could be made to make the system more accessible.

- Packaging files - Encapsulate all the required files/information for visualisation in a self-contained file or directory as this would allow the interactive visualisation to be circulated or included in supplementary information supporting a publication.
- Integration - Investigate integration with data analysis software. Additional libraries that build upon d3.js to work with Python and R have recently gained popularity [185, 186]. Adopting these could enable interactive data visualisation to be created from the output of data analysis and allow incorporation into notebooks.

Part II

Connected Science

Chapter 5

Introduction to Connected Science

The way which we interact with the world around us is changing rapidly. The use of computers has become ubiquitous in everyday life; developments in recent years mean it is possible to access almost any information at the touch of a few keys, adjust the controls in a house from anywhere in the world, see real-time traffic information on roads and public transportation. These are just a few ways in which advances in technology have changed our everyday lives, but how have these changes affected the way in which we work?

An overview of some of the key technology developments of the past few decades is shown in Figure 5.1. While this is by no means an exhaustive list of the technology developments it shows that much of the technology that is now completely ingrained in everyday life (smartphones, social media, broadband etc) were developed relatively recently and have been swiftly adopted. [187] Less than a decade since the smartphone revolution 70% of UK adults have smartphones. [188]

Technology is continuously developing at a pace that is often difficult to keep up with. In the digital age trends come and go just as quickly, but every now and then there comes a technology that revolutionises the landscape. The development of the computer is a key example of this, as is the development of the internet. However, potentially the biggest example of a revolutionary technology in recent years is smartphone and mobile internet technology.

Not only is technology being developed more swiftly but many devices are becoming more complex. Today's smartphones are not simply a phone with some smart capabilities. Instead they incorporate a wide variety of communication methods, sensors and hardware, allowing interaction with a huge number of devices and services from our pockets [189]. Numerous sensors are contained within a standard smart phone which

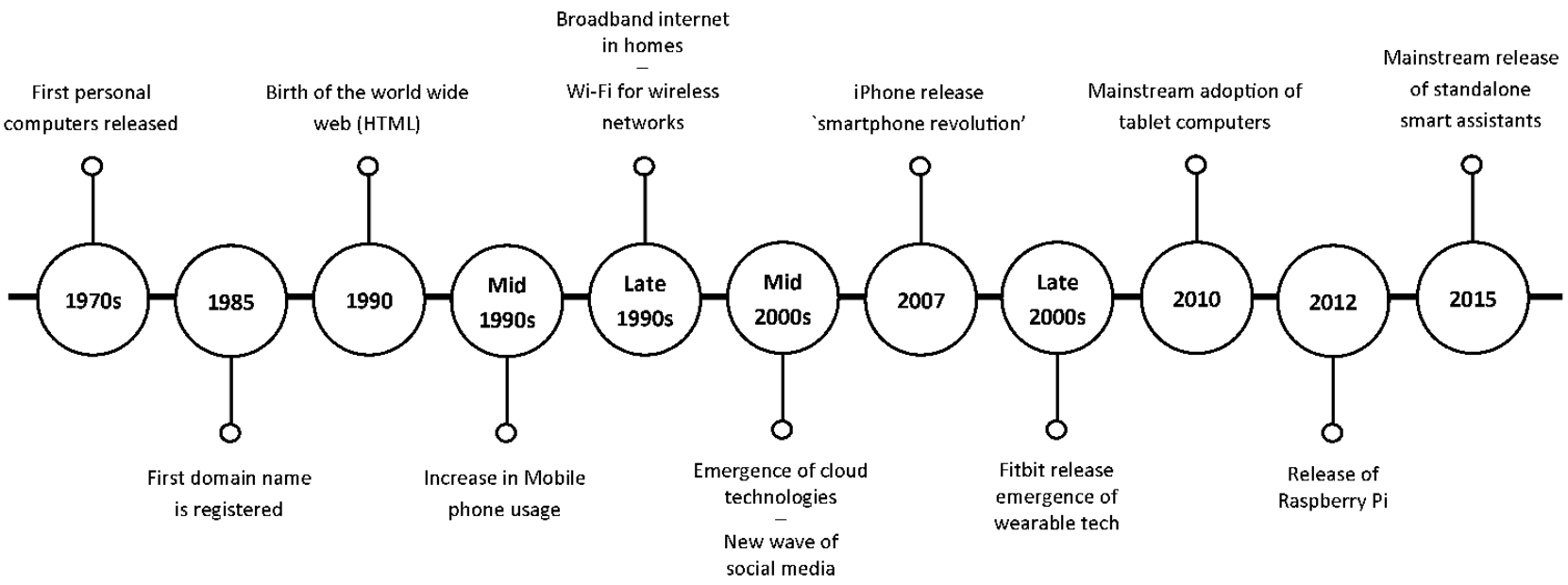


Figure 5.1: Overview of some key technology developments of the past decades [191–194]

could be used to take measurements of the environment around them. Phones these days contain accelerometers which can measure speed and direction of movement, thermometers, light sensors, proximity sensors and many more. The computing power that we can now carry around in a phone is huge, far surpassing the power of the computers that powered the Apollo space missions [190]. Not only have we achieved our school math teachers' fears of always carrying around a calculator in our pockets but we have surpassed this many times over.

The advances in computing have not only affected smartphones but many different aspects of computer hardware and peripherals. Processors, memory, cameras, and sensors have all increased in capabilities and decreased in size and price. This widens the range of potential applications for these technologies, as the technology is no longer the limiting factor. Sensors and processors have been embedded in everything from washing machines [195, 196] to toothbrushes [197]. There is little that someone somewhere isn't trying to make 'smart' by connecting it to the internet. Embedding technology in objects to connect them to the internet is referred to as the 'Internet of Things' (IoT) [198].

While many of the applications of 'smart' technology and IoT devices may seem like gimmicks, sensible application of connected technology has the potential to bring a multitude of benefits to any environment it is used in. In particular the lab environment could be enhanced by the addition of connected technology.

While the use of the computer in research is now widespread and researchers are often quick to adopt new algorithms and processing methods, the adoption of connected technologies within the lab environment is not as widespread as in everyday life. In the traditional science lab, many separate systems often exist which are isolated and specialised for a single type of measurement, sharing data between systems is not easy and frequently measurements must be made manually often even written into notebooks.

Implementation of connected technology within a lab could have a whole host of benefits, including but not limited to: allowing access and control remotely, creating more efficient interaction with the lab, increased data collection and retrieval, easier data sharing and collation of data in a central system. All of these aspects are ways in which connected technology can create digital synergies with the lab environment and add value for the scientist.

The investigation into technology in the lab had multiple areas of interest including; the presence of technology in the lab, interaction with the technology and the retrieval and presentation of data. The ideal method in which to examine technology in the lab would be to develop two labs, one created with integrated technology in all possible aspects, and the other one a more traditional lab with a combination of systems. This would allow evaluation between the two systems of usability, functionality and possible added value. In the busy research environment of a university; however, it would be very

complex to carry out an evaluation such as that. For this reason the development that has been undertaken has worked mainly within those lab environments already present in the university.

The research carried out looked at creating a more connected lab experience through two different pathways. In the first, remote experiments were created for the purposes of undergraduate practical teaching, allowing students to carry out experiments via the internet when they are not physically in the lab. In the second, interaction with the physical lab environment was investigated. This was achieved through the creation of a framework for the connected lab (Talk2Lab), connecting sensors and equipment up to the internet and enabling researchers to interact with their lab environment through voice.

Chapter 6

Remote experiments

6.1 Background

In a traditional teaching environment at a school or university, students carry out practical experiments in a laboratory, involving the potential use of expensive equipment, costly chemicals and hazardous compounds. These practicals also usually require a large amount of specialised space (e.g. laboratories with fume hoods), supervision from skilled technicians and lecturers and a considerable amount of time to carry out the experiments.

All of these factors have prompted research into the areas of remote experiments for a number of years, although remote experiments have come a long way from their inception in the 1990s they have not yet become mainstream. Online teaching resources such as MOOCs (Massive Online Open Courses) have experienced a surge in participation [199,200], and interactive remote teaching environments for computer programming and engineering [201, 202] have increased, but experiments in other science disciplines are comparatively slow in their development. [203]

Many of the remote teaching environments developed focus on provision of pre-recorded material or computer simulations. A number of systems have been developed to provide a more integrated experience; however, these are very limited in number. [204] This is likely to be caused by many factors; however, a major issue is the complexity of remotely controlling scientific chemistry experiments in a method that allows them to be run multiple times, and the ingrained association that the sciences have with carrying out practical experiments in a lab.

Science experiments are traditionally carried out in a lab, but in many cases the knowledge acquired is not completely dependent on physically carrying out the experiment.

A lot of skills can be learnt when partaking in physical experiments within the lab environment but in many cases experiments utilise skills that have been learnt in previous experiments and often the new knowledge is related to the analysis and application of theory rather than the experimental procedure. This new knowledge can be learnt equally well through a well designed remote experiment.

The principle behind designing the remote experiments was to create a variety of educational resources. In Southampton these have been proposed for use in the undergraduate teaching labs as 'practical experiments'; they can reduce the strain on the teaching labs schedule and capacity by giving remote access to the experiments. The experiments can be accessed at any time which allows flexibility in the student's schedule and a higher capacity for the total number of students that can use the experiments.

Although the positive aspects focus on the time and capacity constraints, remote experiments are unlikely to replace all of the practical experiments. This is because a large learning outcome from the practical course as a whole involves the learning of laboratory skills such as lab safety, handling and measuring chemicals, and using specialised equipment.

Within the traditional strands of chemistry, remote experiments are better suited to certain areas, such as physical chemistry, as the experiments in this area often involve taking measurements and observations of a system rather than organic chemistry where students carry out reactions often with multistep processes involving many reagents.

To examine whether remote experiments were a viable option for UoS (University of Southampton) undergraduate teaching a series of remote experiments were created for the physical chemistry teaching course. The first experiment that was created focused on investigating the Beer-Lambert Law (BLL), a key concept in UV-visible spectroscopy. This built on work previously carried out by the Frey Group [205] in creating an online interface for remote experiments. A second experiment was also developed for use alongside the BLL experiment in the teaching labs, this second experiment focused on the Ideal Gas Law by examining the effect of changing temperature on gas volume.

6.2 Beer-Lambert Law Experiment

The Beer-Lambert Law defines the relationship between the concentration of a solution (c/mol dm^{-3}), the path length of the cell (l/cm) and the absorbance of light (A). It is given by equation 6.1, where ε /M^{-1} cm^{-1} is the molar extinction coefficient of the solution.

$$A = \varepsilon cl \tag{6.1}$$

The normal method of determining the value of ε is to carry out sequential dilution on a solution and measure the absorbance of a constant sized cell of the resulting solution in a colorimeter; this gives a constant path length, l and a varying concentration. The value of ε can then be determined from a plot of absorbance vs. concentration. Figure 6.1 shows the resultant plot for determination of the extinction coefficient of Rhodamine 6G in ethanol from a traditional experiment, where the gradient of the line is the molar extinction coefficient.

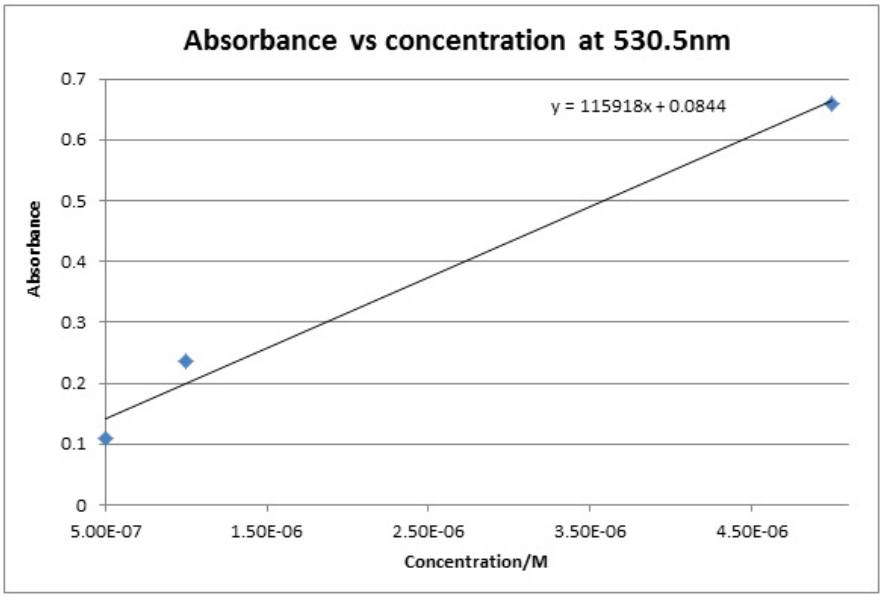


Figure 6.1: Example plot from Traditional BLL experiment - Determination of extinction coefficient of Rhodamine 6G in ethanol

The design of the remote experiment approached the problem from a different direction. The aim of having a remotely controlled experiment meant that physically changing the concentration of a solution during an experiment would not be feasible. Instead the concentration of the solution was kept constant and the variable that was changed was the distance, or path length, through the use of a longer optical cell.

This experiment was designed to use a solution of Rhodamine 6G, which is a fluorescent laser dye. Rhodamine 6G in water absorbs light at approximately 530nm and fluoresces

with a wavelength of 550nm [206], this absorption maxima corresponds well to the wavelength of a Green Laser at 532nm which can be used to cause fluorescence.

The concept behind this experiment assumes that the amount of fluorescence emitted by the solution is proportional to the intensity of the light present at that point in the solution, and that the wavelength of the fluoresced light is shifted far enough that it is not significantly re-absorbed in a different area causing excitation and 're-fluorescence' of the Rhodamine which could lead to a distortion of the intensity data.

The experimental set-up allowed the determination of the intensity of light present at each distance point through the cell. This was achieved through analysis of an image taken of the cell when it was fluorescing. Using equation 6.2 this can be converted to show the absorbance at any given distance (x) using the relationship between absorbance and intensity.

$$A_x = -\log_{10} \frac{I_x}{I_0} \tag{6.2}$$

where A_{x} is the absorbance at distance x, I_{x} is the light intensity at point x in the sample and I_0 is the initial light intensity of the beam before it is absorbed by the sample. The absorbance and distance (path length) measurements can then be used in conjunction with the Beer-Lambert Law.

The data extracted from the image could be output to the student in a variety of different forms. These forms ranged from the raw intensity data to the calculated value for ε. The type of data output could be adjusted to correspond to the depth of subsequent analysis required by the level of the course.

This experiment provided an educational resource which demonstrates the theory of the Beer-Lambert Law without the necessity of a time consuming 'wet' practical and the only requirements of the experiment were a computer with internet access and a spreadsheet program. The experiment could also be used to teach additional data analysis skills by requiring the students to calibrate the results.

An experimental set-up for a BLL experiment was created previously [205, 207], and had been accessible through `http://soton-altc.ourexperiment.org/`. This set-up was investigated to increase the scope of the original experiment and develop it for use in teaching labs. However, it was discovered that the experiment was non-operational and could not produce new results, despite appearing operational through the webpage.

The set-up of the previous experiment can be seen in Figure 6.2. This set-up contained a small laser which was controlled by an Arduino, a type of microcontroller board which is frequently used in small electronics projects. Due to the nature of an Arduino it was not possible to retrieve the coding which had previously been programmed onto it. A lack of accessible documentation about the Arduino code and the rest of the experimental set-up meant that the decision was made to begin the experiment again 'from scratch'.

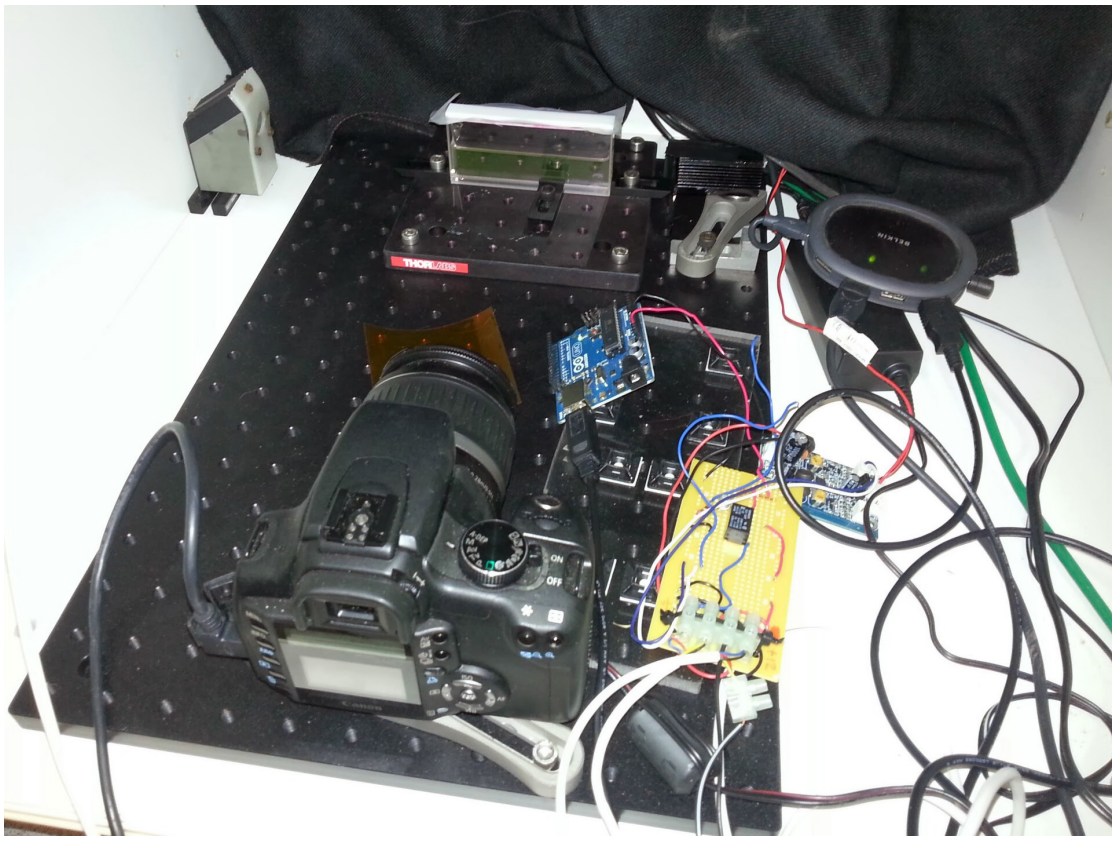


Figure 6.2: Old set-up of BLL experiment

Beginning the experiment again meant that improvements could be made on a number of aspects from the previous experiment which were not optimal. In addition to this extra features could be added which were not previously incorporated.

Areas for improvement included:

- Creating the experiment within a moveable holder. This will entirely enclose the system (with interlock), allowing the experiment to be moved if necessary. Previously the experiment was installed directly in a cupboard.
- Use a smaller camera, previously a Canon DSLR was used which is a bulky camera, using a webcam may give a suitable resolution and sensitivity whilst taking up significantly less space.
- Not storing the last captured image in the 'memory' of the experiment. Previously it would simply publish the same image repeatedly if a new image was not captured.
- Control user access to the website so only one user can be in control of the laser/ camera at a time.

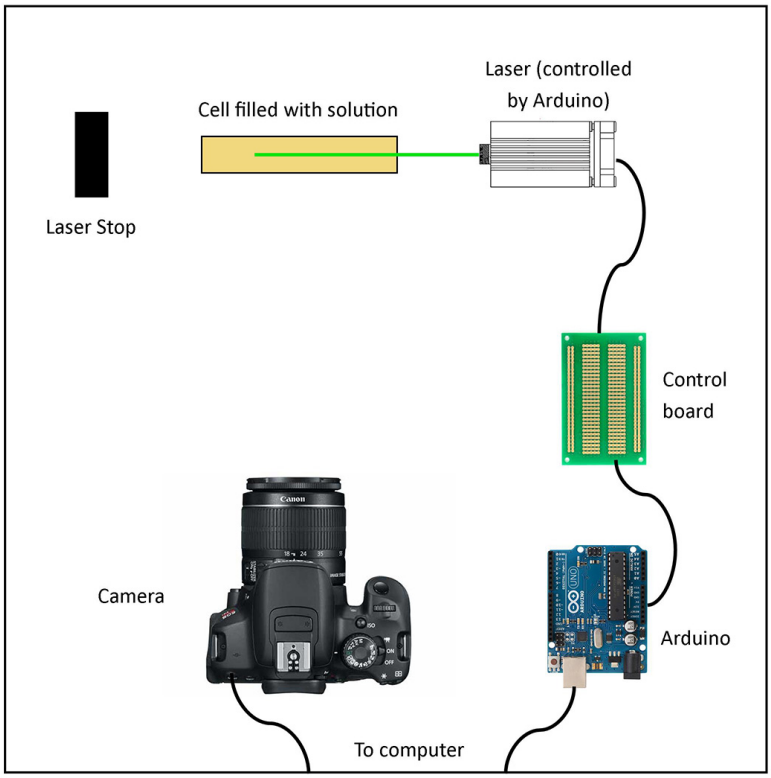


Figure 6.3: Simplified Scheme of the BLL experiment set-up

- Better documentation relating to the set-up and Arduino coding which is crucial for anyone wishing to develop the experiment later or carry out maintenance/ troubleshooting.
- Install lights and allow independent control of both the laser and lights. This can show the students the effect of not running the experiment in dark conditions in addition to allowing them to view the set-up.

6.2.1 Experimental Laser Set-up

Following discovery that the previous BLL set-up could not be fixed a new set-up was developed following a similar format to the old experiment. Figure 6.3 shows a simplified scheme of the set-up shown in Figure 6.2. The key components taken from this set-up were a laser that could be controlled, a long optical cell which contained the solution of interest and a camera for capturing images.

The same sized cell and same style laser were used in the development of the new set-up along with the use of an Arduino for the the controls. To enable portability and reduce the storage requirements of the experiment the whole set-up was condensed, with the connections being streamlined and the DSLR camera being replaced by a smaller camera.

The experimental set-up was developed in a self-enclosed housing which could be moved if required. This housing contained a laser board with mounted lights, camera, laser, optical cell and laser stop. Alongside the main board a separate section housed the electronic controls, circuitry controlling the power to the system and an Arduino board taking signals from the server. An interlock was integrated into the housing to interrupt power if the housing was opened.

Light strips were added to the housing to allow the user to view the experimental set-up in addition to observing it under fluorescence. Controls were developed allowing the user to operate these light strips in addition to the control of the laser module and camera which had been implemented in the previous set-up.

For the excitation of the Rhodamine 6G solution a 532nm Green Laser (30-40mw - Odicforce Lasers) was selected as this corresponds well to the absorption maxima of Rhodamine 6G. After investigation with multiple concentrations of Rhodamine 6G solution a concentration of $5 * 10^{-6} moldm^{-3}$ was selected as the optimum concentration for the experiment. This concentration allowed the laser to penetrate sufficiently far through the solution for analysis but ensured that it was fully absorbed before reaching the end of the cell.

The DSLR was removed as the camera in the set-up and replaced with a less bulky webcam. A Logitech C920 HD web camera and network-enabled D-Link DCS-942L camera were both examined as replacements for the DSLR. The Logitech camera was selected as the snapshot camera as it produced a higher resolution image. The D-Link camera only gave a 640x480 pixel image, which was too low for image analysis. The network-enabled camera was instead installed as a live-feed camera, giving a continuous live-feed view for the experiment alongside the snapshot functionality.

Figure 6.4 shows the equipment inside the housing of the new experimental set-up. This set-up was much clearer and more compact compared to the previous experiment whilst containing all necessary functionality. The lights in the housing cannot be seen in these images.

6.2.2 Experimental Control

To create an operational BLL remote experiment a number of sections of development were required. Control was required for the hardware contained in the experimental housing, experimental scripts were required to run the process of the experiment and carry out the analysis, and an interface was required to allow the student to access the experiment.

The circuit for controlling the hardware in the housing was redesigned and built with separate controls for the lights and the laser using relay switches (See Figure 6.5); this

Figure 6.4: Front and Top views of the new BLL experimental set-up

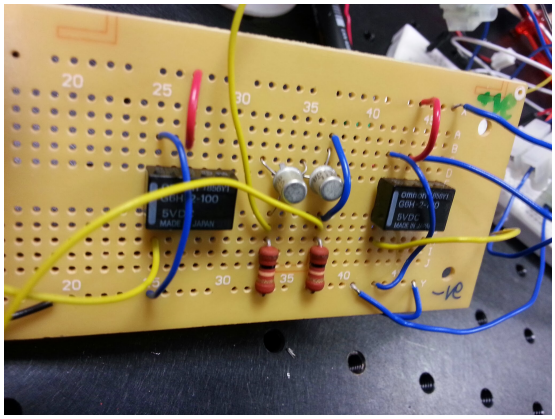


Figure 6.5: Circuit design for Laser and Light control

design allowed the user to control the light of the experimental environment separately from the laser. 4 different combinations of settings existed; 'laser on lights on', 'laser off lights on', 'laser on lights off' and 'laser off lights off'. The default setting in this set-up was 'laser off lights off' which the system would revert to if the power was cut or the interlock broken. The interlock was a safety feature which cut power to the laser and the lights, but not to the cameras. This would be activated if the housing in which the experiment was stored was opened. This was included to eliminate the risk of injury due to accidental exposure to the laser beam.

6.2.2.1 Arduino Control

It is not possible to retrieve code from an existing Arduino controller as Arduino code is compiled when it is transferred to the Arduino, and no documentation was available for the previous set-up, therefore the code for the Arduino was developed again. The code on the Arduino was used in conjuction with the circuit design to control the laser and the lights, based upon input from an external source. In this case the external source

was the computer connected to the Arduino. There are a variety of different methods by which the Arduino can be controlled; in the development of this experiment the two control methods investigated were serial input from a computer and direct control through use of MATLAB [16] scripts.

The Arduino Uno board was used to control the signals to the laser and lights. In the set-up the Arduino 5V digital pins were defined as follows:

Pin 8: laser

Pin 12: lights

These pins were controlled by setting them to HIGH (on) or LOW (off), based on input from the computer. The serial input method was selected for development as using MATLAB required the MATLAB script to rewrite the code on the Arduino each time a command was triggered. This added more complexity to the system and more chance that the Arduino would malfunction.

Figure 6.6 shows the code that was created for the Arduino. When a pin was set to HIGH this triggered the relay switch in the circuit and the relevant section would be powered up, following a switch to LOW the relay would reset and the section would be powered off.

The coding which controlled the experimental process also had to be re-written, as there was no documentation for the previous installation. The overall process was controlled by the host server with the laser section being further controlled through the use of the Arduino, and the acquisition of results being carried out through scripts in MATLAB.

6.2.2.2 Image acquisition

The previous version of the experiment was created using a Linux machine with the use of image capture program gphoto2 (a free image capture script built for the UNIX framework). When the whole system was rewritten it was converted to a Windows server as the programs and interface were more familiar. This required a new image processing program.

The programs 'Robot-Eyez' [208] and MATLAB both exhibited the capability to capture an image from the webcam (Logitech C920 HD) at 2304x1296 pixels. The previous image resolution obtained via the DSLR camera was 3456x2304 pixels; however, this slightly reduced resolution was still sufficient for analysis of the image.

A network-enabled camera (DCS-924L by D-Link) was also examined as an alternative image source, due to its simpler image acquisition process. However, the highest resolution of the network-enabled camera at 640x480 pixels was not high enough to allow

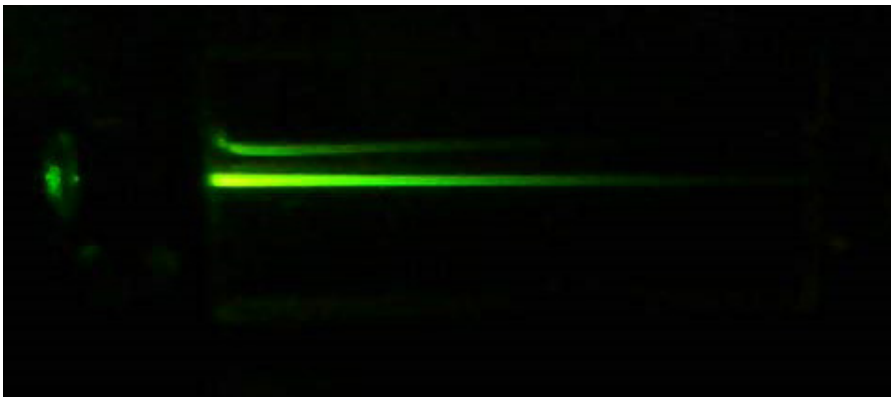
```
/*
 It waits for a byte in the serial port, and
 powers up pin 8 if it is a and powers down if the value is b.
 powers up pin 12 if it is m and powers down if the value in n.
*/
int inByte = 0;         // incoming serial byte
int laser = 8;
int lights = 12;

void set-up()
{
  pinMode(lights, OUTPUT);
  pinMode(laser, OUTPUT);
  // start serial port at 9600 bps:
  Serial.begin(9600);
  while (!Serial) {
    ;
        }
}
void loop() {
  // if we get a valid byte, read analog ins:
  if (Serial.available() > 0) {
    // get incoming byte:
    inByte = Serial.read();
  }
    if (inByte == 'a'){
         digitalWrite(laser, HIGH);
  }
        if (inByte == 'b'){
                digitalWrite(laser, LOW);
  }
        if(inByte == 'm'){
                digitalWrite(lights, HIGH);
  }
        if(inByte == 'n'){
                digitalWrite(lights, LOW);
  }
}
```

Figure 6.6: Arduino Code for controlling lights and laser - written in Arduino IDE

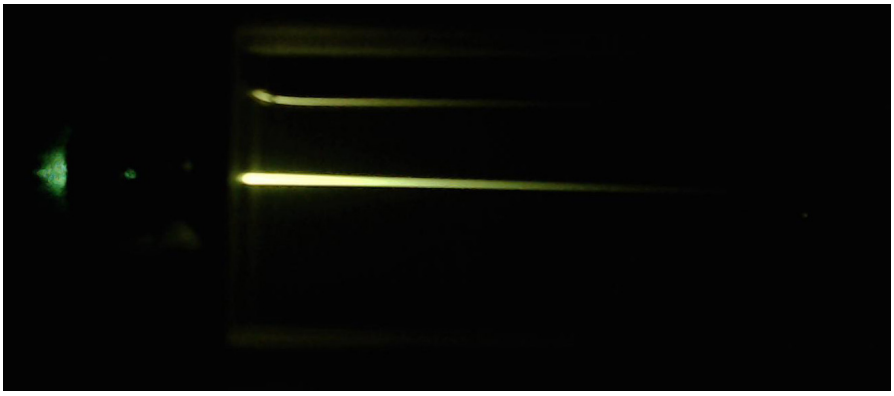
analysis. Figure 6.7 shows a comparison of the two camera outputs, the data obtained from the D-Link camera contained much more noise than the Logitech camera.

Due to this the Logitech webcam was selected as the snapshot camera to obtain the image for analysis, giving the resolution of 2304x1296 pixels. Robot-Eyez was selected as the method for image capture, as the process was quicker and less memory intensive than using MATLAB. Robot-Eyez was operated through the command-line, passing parameters to the Robot-Eyez application that would define the filetype and resolution for the image capture.

The analysis that was done on the fluorescence image to obtain absorbance values was executed in MATLAB. A script was written for MATLAB which took a fluorescence image as its input and carried out analysis to convert the data to absorbance values. Image analysis was performed on a specific section of the image, which was identified in development. This section was the horizontal location of the optical cell with the vertical selection encompassing the position of the laser beam. The image toolbox



(a) Rhodamine Fluorescence with D-Link DCS-942L camera



(b) Rhodamine Fluorescence with Logitech C920 camera

Figure 6.7: Comparison of image resolution through two different cameras

was used to obtain light intensities, these intensities could be converted to absorbances using Equation 6.2 where the initial intensity is the intensity at the start of the cell. The distances were converted from pixels to centimetres using conversions identified from the image composition. As output the script produced three files; the raw image of the snapshot that was used as the input, a plot of distance vs. absorbance and a spreadsheet with distance and absorbance values that could be used by a student for further analysis.

6.2.3 Experimental Interface

Beyond the coding that controlled the hardware in the experimental housing and carried out the analysis a system was required that provided the students with access to the system, carried out the experimental run and returned results to the user.

The backend server controlling all the processes of the experiment was created using the Node.js framework [19]. Node.js is an open source server framework built on the Chrome JavaScript engine. It is lightweight and efficient, making it suitable for this application. Node.js was used to create a server on the host computer which could be accessed

through a web interface by permitted users. The experiment was controlled through a combination of Node.js functionality and Serial port and command line processes to control the lights, laser and snapshot functionality. The scripts for controlling this process including error handling, page view designs and data processing were all designed specifically for this experiment.

The Node.js server was designed to respond to the user's queries and display specific pages based upon the variables provided by the user e.g. laser on/off, lights on/off, etc. enabling the user to interact with the experiment and obtain processed results from the server. Figure 6.8 shows the user's path through the experiment. Appendix C contains screenshots of the experimental interface in use and the code for the server can be found in the ESI. [30]

Once the user started the acquisition of the data, the experiment controlled the laser, obtained an image and passed this image to the MATLAB script for processing to obtain results in the form of distance and absorbance measurements. This process had checks built in to see if the experiment was in use, and error-handling if stages of the processing could not be carried out. A troubleshooting guide was also included on the server to help students out if the experiment did not perform as expected.

The students could access the experiment at any point when connected to the SOTON.ac.uk university network, either through a wired network or via the VPN connection. This allowed the students to carry out the experiment from their own computer or a computer terminal in university at a time which was convenient for them, provided that the experiment was not in use by another student.

6.2.3.1 Design Features

User control

The interface was designed to only allow one user to control the laser set-up at a time. It required the input of a username to begin the session. This did not check the input name against a list of students carrying out the practical but it did require an input to proceed which was stored against the results of the experimental run. Once the user had started their session they had 5 minutes to complete the experiment. If the experiment was not completed within that time then their session timed out and they would have to begin the session again. In the future the system could be linked into the university login systems, requiring students to log-in at which point it can check if they are eligible to run the experiment and if necessary a scheduling system could be implemented to allow students to book their slot on the experiment.

If another user was active on the system then anyone trying to begin a session would be informed that it was in use and to try again shortly. Once a user had completed their

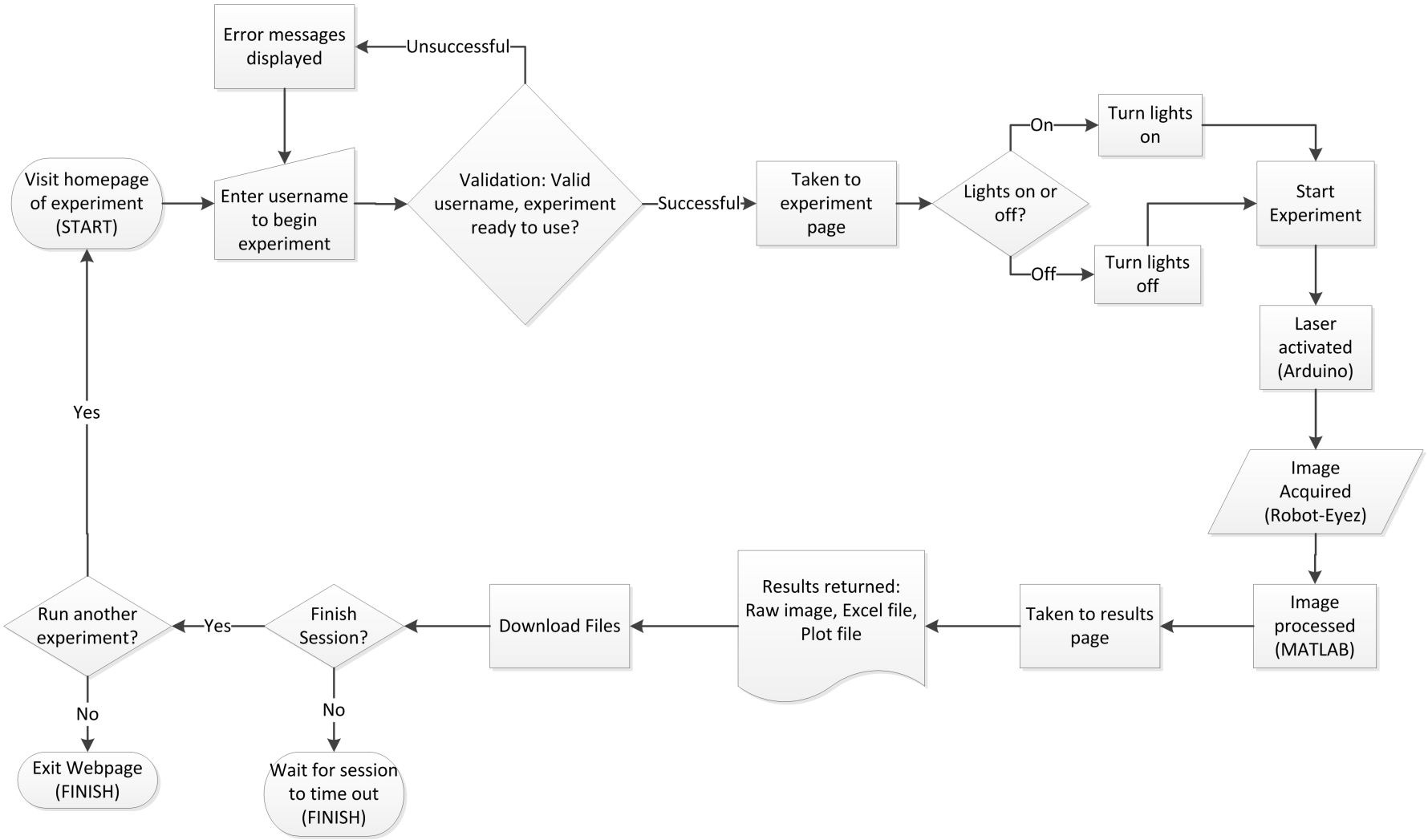
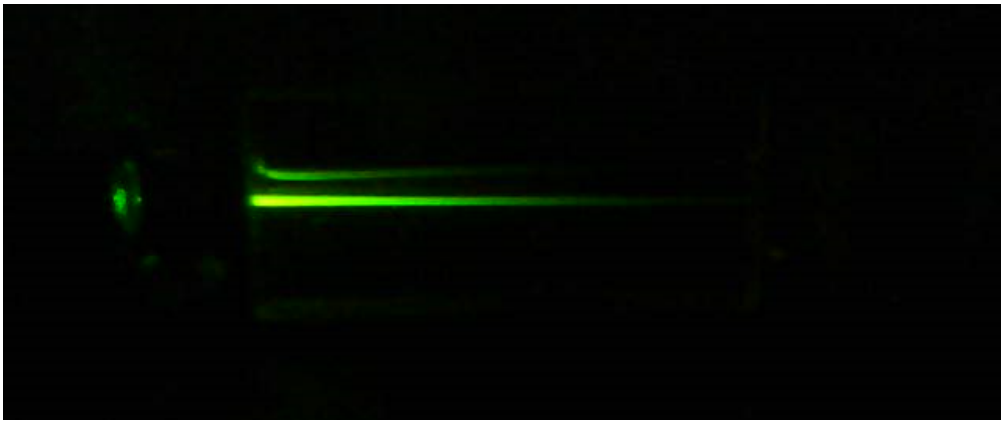


Figure 6.8: Flow through the BLL Experiment

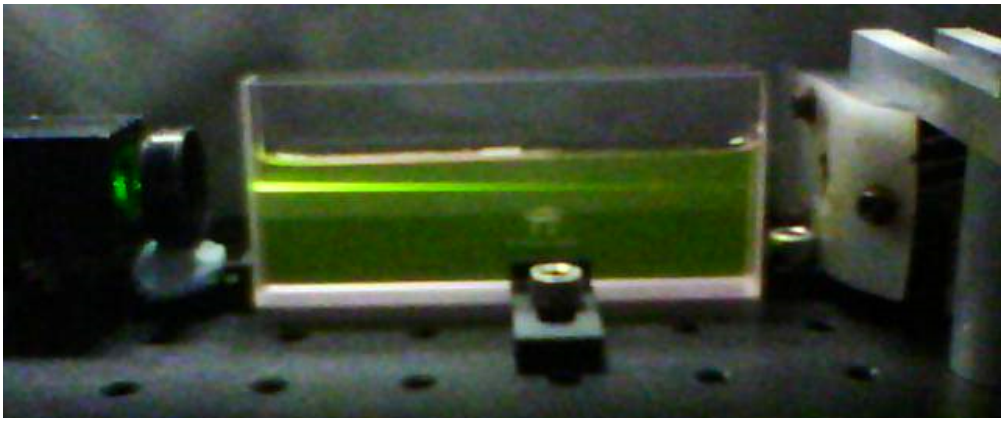
experiment and obtained their results then they could manually end their session to free up the experiment. If they did not do this then their session would automatically time out once the 5 minute time limit was reached.

Light conditions

The server allowed the experiment to be carried out under both light and dark conditions to allow the student to compare the two outcomes and the effect that the light has on the results. Figure 6.9 shows fluorescence images obtained under both dark and light conditions. In the dark conditions the fluorescence emission could be clearly identified; however, under light conditions the fluorescence was less defined due to the background light level, leading to noise in the analysed data.



(a) Rhodamine Fluorescence under dark conditions



(b) Rhodamine Fluorescence under light conditions

Figure 6.9: Comparison of Fluorescence images taken under dark and light conditions

Level of analysis

MATLAB scripts were created which could automate the entire analysis, taking the input of an image and outputting the calculated value for ε. However, by doing this, there would be little learning value for the students, as they do not process the data. Reducing the amount of processing that MATLAB carried out on the data increases the amount of work required from the students. The script implemented in the practicals gave the output as absorbance values.

By making small alterations to the MATLAB code it would be possible to tailor this set up for a range of teaching needs. For future expansion of the experiment it would be possible to incorporate multiple levels of analysis that would be carried out dependent on the students username. A script could check the username; if it was from list A run the analysis to the first level, if on list B run to the second level, etc. This would allow potential application to earlier stages of education, or as an outreach activity.

6.2.4 Implementation

This experiment was implemented in the teaching labs for the 2014/15 teaching period in first year undergraduate lab. It was implemented alongside the Gas Law experiment and a third remote experiment not designed in this work.

The response to the experiments was positive and the BLL experiment was successfully run over 400 times. Feedback was obtained from the students during the course of the implementation and this is discussed in the 'Discussions & Future Work' section. However, review and changes to the curriculum of the first year teaching course meant that the experiments were replaced by computational modelling practicals covering new course content.

These experiments could be implemented in other areas of teaching e.g. schools or used as an outreach activity for science education with minimal modification.

6.3 Gas Law Experiment

The second experiment that was created for the teaching labs was designed to investigate the properties of gases. When discussing the properties of gases there are equations of state which the substances obey; these relate the quantity, volume, temperature and pressure of a given gas. In the case of low pressure gases this relationship is described by the following equation (6.3) which is the perfect gas equation of state.

$$P\,V = n\,R\,T \tag{6.3}$$

Where P is the pressure of the gas (in Pa), V is the volume of the gas (in m^3), n is the quantity of gas (in mol), R is the Gas Constant (8.314 J K^{-1} mol^{-1}) and T is the absolute temperature of the gas (in K). [209]

The perfect gas equation combines observations from Boyle's law, Charles's law and Avogadro's principle. A hypothetical substance whose properties can be modelled by the perfect gas law at all pressures is termed a perfect gas. In reality gases do not actually obey the perfect gas law uniformly at all pressures due to molecular attractions and repulsions. These are referred to as real gases and, as the pressure approaches zero, they behave more like perfect gases. Normal atmospheric pressure is in practice low enough for real gases to behave almost perfectly and therefore can be treated as perfect gases for the purposes of many calculations.

The perfect gas equation of state is also widely referred to as the ideal gas law and underpins many of the calculations relating to gases, including thermodynamic processes and meteorology. It is a key concept in physical chemistry and as such it is important for chemistry students to understand the equation and be able to apply it.

The gas law experiment must allow the students to carry out measurements of a gas system with the ability to measure and record 3 of the 4 variables in the gas law equation over a period of time, these variables are; pressure, volume, temperature and quantity (no. of moles). A common experimental set-up for investigating gas pressures is the use of a manometer, see Figure 6.10.

Usually in a manometer both ends of the tube are open, one to a gas of interest and the other to a gas of known pressure, usually atmospheric pressure. Alternatively a closed-end manometer can be used where one end of the tube contains enclosed gas. A barometer, used to determine atmospheric pressure, is an example of a closed-end manometer. The difference in height between the two arms allows calculation of the pressure through this equation:

$$P = P_0 + \rho\,g\,h \tag{6.4}$$

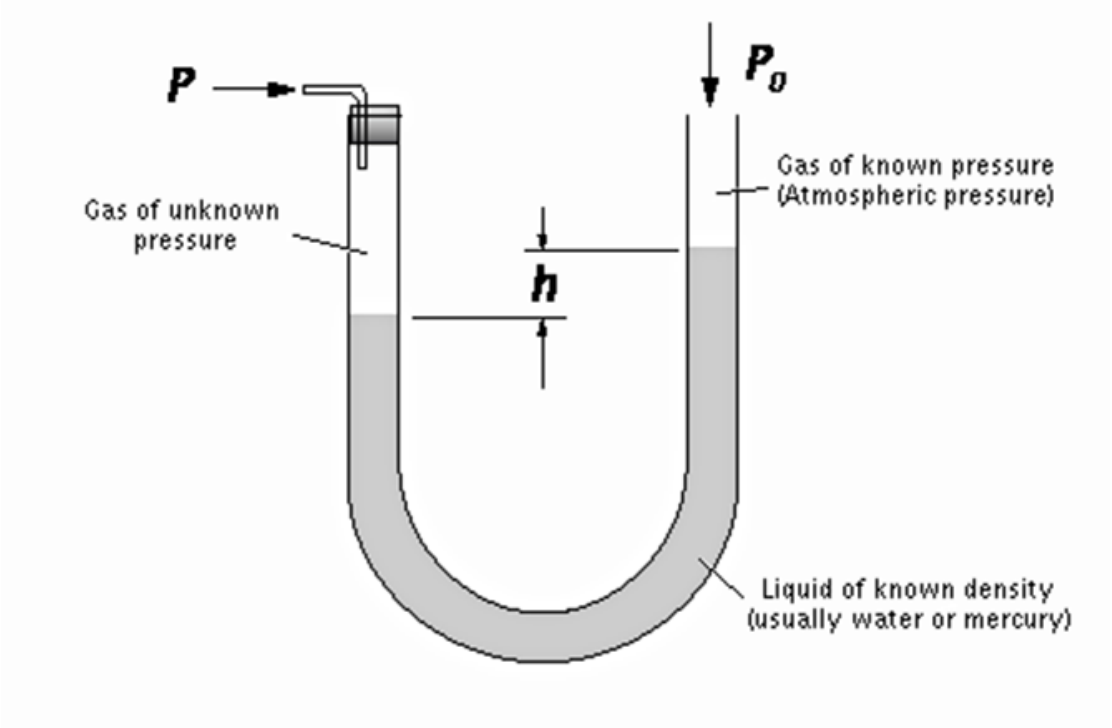


Figure 6.10: Example of a Manometer, used to determine gas pressure

Where P_0 is the known pressure (in Pa), ρ is the density of the liquid (in kg m^{-3}), g is the acceleration of gravity (9.81 m s^{-2}) and h is the height difference between the arms (in m).

6.3.1 Experimental Set-up

This second experiment was designed to practically demonstrate the ideal gas law, whilst teaching aspects of measurement acquisition and analysis. The experimental set-up is an implementation of a closed-end manometer, using an inverted container of gas within a temperature controlled water bath, where the top of the water bath is open to the lab, see figure 6.11 for a simplified depiction. The pressure of the gas within the vessel can be calculated through the difference in heights of the two water levels.

This experiment was in theory a simpler experimental set-up than the BLL experiment as it required less dynamic controls. It was designed to require a single experimental set-up, from which multiple students can carry out their experiment. The experiment design required the control of only one variable and the measurement of other variables which were affected.

In order to demonstrate the principles of the gas law one of the parameters had to be varied throughout the experiment and the values of the dependent variables measured. Temperature was the variable chosen to be controlled throughout the experiment, this

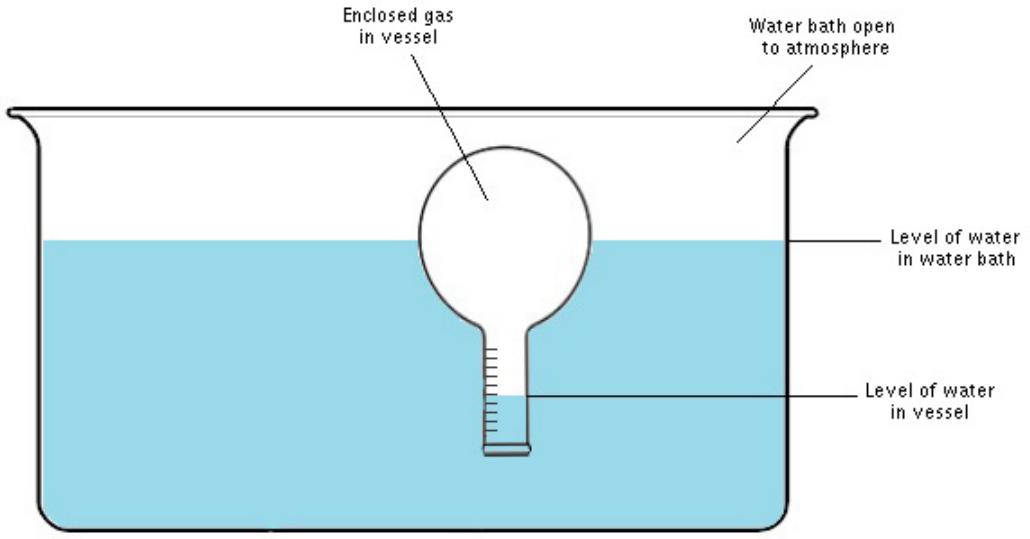


Figure 6.11: Simplified depiction of the experimental set-up of the manometer used in the gas law experiment

can be done via a heater plate under the water bath and measured through the use of a temperature probe. The atmospheric pressure in the lab can be measured with a pressure gauge, the unknown pressure of the gas in the vessel can be calculated through the use of the closed-end manometer and the volume can be measured through gradations on the manometer set-up. The quantity of this gas was unknown; however, this set-up was a closed system so the quantity of gas was a constant value and could be calculated from the experimental measurements.

To create a useful remote experiment for the undergraduate (UG) experience it must:

- Be able to be carried out from a remote location
- Not require physical input in the lab
- Either automate or have controls for the temperature adjustment
- Allow measurements to be taken remotely

There are a number of ways in which this experimental set-up could have been developed; however, much of the design is limited by the capabilities of the equipment available. There are instruments which can be controlled via a computer to produce readings on command, such as temperature and pressure or change parameters, such as the temperature of a heater plate. This sort of equipment can be significantly more expensive than basic options and was not already present in the labs, consequently some of the equipment employed in this set-up was more basic than desired.

The main control required in this experiment was for the heater plate which regulates the temperature of the water bath and therefore the temperature of the gas. It would be desirable for the student to be able to set the temperature of the heater plate and control its use in their experiment; however, the existing heater plates did not allow electronic adjustment, so any alteration of the heater would require physical input and eliminates the remote aspect.

It would be possible; however, for the student to turn on the heater remotely at a set temperature and allow the water in the bath to heat up. Then at a specific temperature or after a certain time period they would turn off the heater and start to take measurements as the temperature of the gas returned to its original temperature.

The other measurements that were required included the atmospheric pressure in the lab, the relative height of the levels in the water bath and the vessel, the volume of gas in the vessel and the temperature of the gas.

6.3.2 Development of Remote Experiment

Aspects to consider in the implementation of this remote experiment involved:

- Multiple students trying to use the equipment at one time (this was a significantly longer experiment than the BLL experiment).
- Implementation of live feed cameras in the secured network of Southampton University.
- Remote control of equipment, needs to be controlled by computer not by physical push buttons.
- Recording of measurements

When carrying out a practical in the lab a student would be able to manually control the various aspects of the experiment and position themselves where required to obtain the measurements, whether these were physical readings or from digital displays. However, when creating a remote experiment the user is limited by the initial set-up and constraints of the system.

A user requires a view of the experimental system when carrying out an experiment. This is most easily achieved through the use of a live feed camera. In the development of this experiment there were concerns that it would not be possible to mount a live feed camera unless it was connected to the computer controlling the web server for the experiment. However, during the development of the BLL remote experiment, network enabled cameras (D-Link DCS-942L) were successfully tested. These could be installed

within the Southampton University network and viewed and embedded into other pages as necessary.

For the user to take measurements it was either necessary for the equipment to provide a digital output that can be picked up by a computer, or for the values to be displayed in a way that can be viewed by the user. Digital probes for temperature and digital pressure gauges did exist but were not in use in the teaching labs at the time. It would be possible to develop a remote experiment which displayed digital readouts for pressure and temperatures on a web page; however, the existing (non-networked) equipment was selected to ensure costs were kept to a minimum.

The measurements of height of water and volume of gas cannot easily be obtained electronically, the straightforward method of measuring these involves reading off a value from a measuring cylinder or other volume measuring mark. In a remote experiment this would still be possible as the equipment can be oriented to display the relevant measuring marks in view of cameras.

6.3.2.1 Control of equipment

Due to the limitations of existing equipment most of the equipment could not be controlled electronically; however, the heater plate must be controllable. At a minimum it must be capable of being switched on/ off as required.

With the recent expansion in home automation systems it could be expected that a wide range of computer controlled power systems would be available. In particular this experiment required a power strip in which each socket could be individually controlled by a computer; however, there were limited products which provided this functionality. Many products would monitor one socket and control other sockets based on whether or not the main socket was powered up, e.g. for turning off a monitor and printer when a computer is turned off. There were also numerous sockets which could be controlled by IR, but these involve physically pressing a button. Neither of these types of power strip were suitable for this experiment.

The one product found that matched the specification was the Energenie Power Management System[1] this product can be controlled by a program on the computer but it also has a command line interface which would be a useful method of control in a remote experiment.

The initial concept for control of this experiment was an online interface, through which a user could control the heater plate, allowing the system to reach a required temperature before they turned off the heater and begin acquisition of their measurements. This

[1] `https://energenie4u.co.uk/index.php/catalogue/product/ENER011`

would have required direct control from the online interface to the heater plate, a method of ensuring only one student was in control of the experiment at one time and a number of safeguards on the system to ensure it would not cause any hazards.

To safeguard the system it would be desirable for the heater plate to have a threshold temperature where it would turn off if the system reached this temperature. As the temperature probe does not output its temperature, this would not be possible in this system. Another method could be that that heater plate would turn off automatically after a set period of time if it had not already been turned off. This would reduce excess power usage, excess water evaporation and risk of fire.

If the heater was user-controlled then, to ensure that another user could not interfere with the running of an experiment, it would be necessary to lock the experiment down to a single user at a time. In the BLL experiment this was achieved by the web server only permitting one session at a time, with timeouts if a user did not complete their experiment within the timeframe. If the experiment was in use it will tell the user to try again shortly. The time required for the gas law experiment was significantly longer than the BLL experiment. The experimental set-up required at least 20 minutes to get up to temperature and following that measurements were taken over a period of approximately 1 hour. As similar system for this experiment would have risked the students incurring lengthy waiting periods.

Issues with waiting times could be addressed by allocating each user a specific time in which to carry out the experiment. However, this would rely on the users sticking to their allocated slots and removes some of the benefit of the remote experiments, which was to allow the experiment to be carried out when the user chooses. An alternative would be a system in which the user could book a convenient slot on the experiment.

It was decided that many of the technical challenges that would be encountered in developing the safeguards and systems to control the user interface in this experiment could be overcome by slight alterations to the experimental procedure. For the first version of this experiment the complexity would be reduced as much as possible. Then, dependent on feedback and evaluation of the experiment, it could be expanded to include more controllable equipment. Instead of the user controlling the heater, it would be automated to turn on and heat the water for a specific length of time, after that time period the heater would be turned off, allowing the user to take measurements on the experiment. This change to the set-up eliminated the need for many of the safeguards by automating the control of the heater plate. This also resolved the issue of single-user sessions as multiple users could all take their own measurements on the equipment at the same time.

The power management system used had functionality which allowed control via a program, command line or scheduled power control for each socket individually. The implementation of the scheduling function in this simplified set-up eliminated the need for a computer to continually control the experiment as the schedule was saved on the power strip. It also eliminated the need for an online interface for the user to control the experiment, as the control of the equipment was handled automatically.

In the modified set-up the heater turned on at set times specified in the experimental procedure (at 2 hour intervals), after the heater had been on for a set amount of time (30 minutes) it was automatically turned off, and the student made their measurements over the period of 1 hour. The necessary measurements were; pressure in the lab (pressure gauge), volume of gas in vessel (from measuring cylinder), height of water bath and water in vessel (from height ruler), temperature of water bath (from thermometer display). All of these measurements could be ascertained through the display of a network enabled camera (D-Link DCS-942L). The pressure gauge, digital thermometer, ruler and measuring cylinder are all visible on the image, shown in Figure 6.12. [210]

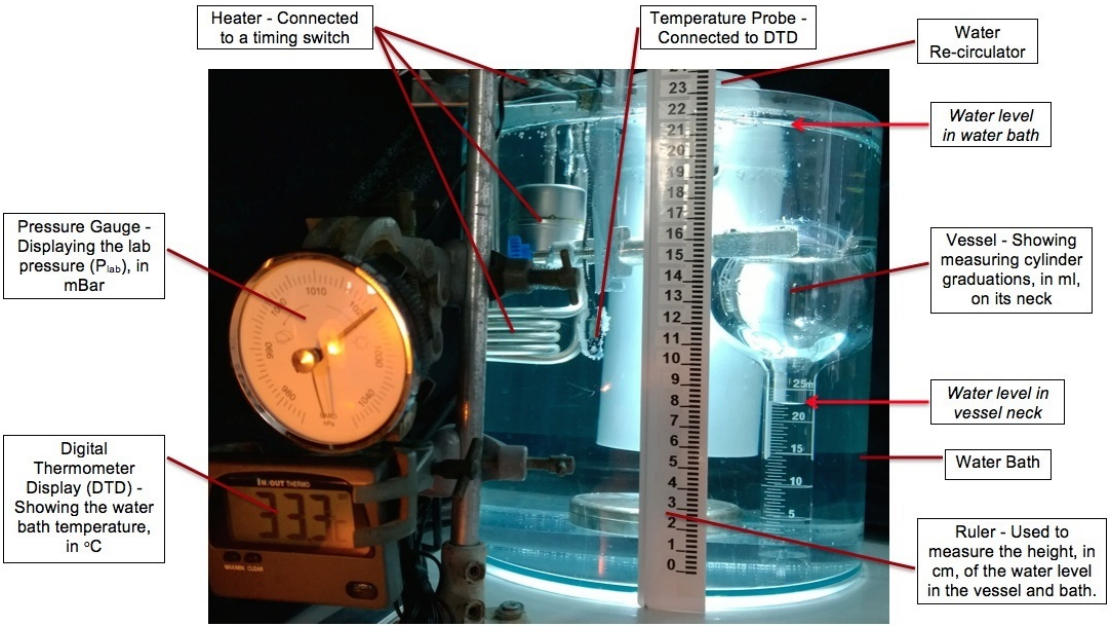


Figure 6.12: Annotated view of the Gas Law Experimental set-up

The measurements, taken at a number of time points, allowed the students to calculate the pressure of the gas inside the vessel for each time point (equation 6.4) and therefore create a plot of PV vs T. The application of the gas law (equation 6.3) meant that the gradient of this plot was equal to nR, which in this system was constant as the amount of gas in the vessel did not change. As such the students could use the gas law to calculate the amount (moles) of gas in the system and answer some follow up questions relating to the gas.

The students could view the experiment at any point when connected to the SOTON.ac.uk university network, either through a wired network or the VPN connection. This allowed the students to carry out the measurements for the experiment from their own computer or a computer terminal in university at a time which was convenient for them (dependent on the times at which an experiment was being run). This also allowed multiple students to access the experiment at the same time as the camera allows multiple simultaneous connections.

The experimental set-up created did not; however, give any estimate or log of how many students have used the experiment as it did not have individual runs for each use. Page view statistics for the website linking to the camera feed could be used to give an indication of the numbers, but it did not require the students to input their username or a similar identifier as was the case in the Beer-Lambert Law experiment.

6.4 Discussions & Future Work

These remote experiments were developed with the aim of creating practicals for the undergraduate teaching labs which combined aspects of automation and computer control with chemistry apparatus and scientific theory. The two experiments developed here were combined with a third experiment to fill an experimental slot in the lab rotation. The third experiment used pre-recorded experiments to demonstrate viscosities of different liquids.

The development and implementation of these experiments has largely been concluded as the experiments have been successfully established and utilised in the undergraduate practicals. The BLL experiment was established mostly using the functionality of Node.js, MATLAB and an Arduino. In this experiment the students could successfully access the set-up remotely, run the experiment and retrieve their results. In one semester the experiment had been run over 400 times.

The gas law experiment was reduced to a simpler set-up and so it relied upon the in built controls of the power management system and the network enabled camera. The combination of these experiments allowed an increase in capacity of the teaching labs because a portion of the student cohort did not require space in the lab for each session.

Initial implementations for these experiments were quite 'low-tech' solutions as they were required to keep the costs to a minimum. The amount of equipment involved was kept to a minimum to reduce the complexity of interactions with the equipment. In these 'proof of concept' experiments there were no parameters for the students to adjust. However, going forwards in the development of remote experiments this would be a key area to work on.

6.4.1 Technical Considerations

The design and development of these experiments produced a number of differing technical issues, with various methods being used to resolve these dependent on the nature of the problem. For the gas law experiment the redesign and simplification of the experimental set-up helped to reduce potential equipment cost and removed complex technical aspects of the remote experiment. For the BLL experiment most of the problems encountered related to the coding of the equipment and server. The initial problems were solved by restarting and writing the code from scratch. As could be expected, a number of bugs were produced within the development of the system. These issues were resolved largely by the use of situational testing and troubleshooting of the code.

There were some issues encountered with the power supply to the experimental set-ups. In the gas law experiment the power management system would retain its schedule in the event of a power cut. However, it would be necessary to ensure that the system time was correct following any extended loss of power. This should not arise frequently but it could have a large impact on the experiment if it were to occur. The BLL server would also have issues with power cuts if the machine was not automatically switched back on. This issue was mitigated by a script to ensure the server would start when the machine is turned on, this means that no user input is required and the server machine can simply be switched on if the power is lost for any reason.

During the experimental runs some additional issues were encountered with the BLL experiment. The solution in the cell evaporated more quickly than was expected; however, this was easily overcome by periodically monitoring the experimental set-up, to ensure the solution does not fall below a threshold level. This could be carried out via the live-feed camera. The BLL set-up also experienced fluctuations in the connection to the Arduino controlling the experiment. This may have been caused by the power supply that fed the Arduino as the connection was stabilised by connecting the Arduino through a USB hub rather than directly to the computer.

6.4.2 Feedback on Experiments

During the assessment of the practical, students were required to complete a questionnaire about the use of the remote experiments. This allowed the collection of comments about the students' enjoyment of carrying out the remote experiments, ways in which the experiments could be improved and the educational benefit gained from them.

Examination of this feedback allows evaluation of the experiments and ways in which they could be developed further to increase the benefit to the students. As feedback was only collected at the end the practicals it was not possible to incorporate the feedback

into the experiments as they were being run. However, it would be possible to use the comments to improve the experiments for future use.

Although the comments on the experiments varied significantly a number of areas were commented on frequently which highlighted some aspects of the experiments which were important to the users:

- The user needs to have sufficient interaction with the system so they feel engaged and that they are influencing the experiment rather than just watching another user's experiment.
- Ensure the equipment is set up correctly for ease of use. As the experiments are remote the students cannot adjust any equipment.
- Having a reasonable amount of analysis on the results from the experiments was perceived as useful, analysis consisting of simple unit conversions or input into a formula was seen as too easy and pointless.
- The most educational benefit appeared to be for experiments which introduced or worked with new concepts rather than concepts which had been encountered previously in practicals. This does; however, depend on the previous experience of the student as they will have had different practical experiences prior to university.
- Introduction of new equipment or different uses of equipment makes the experiment more interesting as the students feels they are learning new skills.
- Including some aspects of problem solving or application of theory was enjoyed, as was demonstration of extra factors that may influence the experiment (such as the presence/absence of lights in the BLL experiment).
- Applicability to concepts being covered in lectures was useful and gives the most educational benefit.
- Even though an experiment was not the most enjoyable it often gave more perceived educational benefit.

Feedback and comments were used to identify areas which may not have performed as well as desired and could be improved for future implementations along with possible ways in which they could be improved.

The interactivity within the BLL experiment could be slightly increased without significantly increasing the experiment time. The stages of data acquisition could be separated out so that the user is required to carry out the stages of the experiment manually. This would require them to turn on the laser and acquire the output image separately. As it would not affect the safe running of the experiment it would be possible to allow the

student to make mistakes in the experiment, for example if they acquired the image without a laser beam present.

On the BLL webpage, or the webpage for the remote experiments practical, the theory behind the experiment and the data processing could be expanded to aid the students' understanding. If it is appropriate it could also include some theory about lasers as additional reading for those students who are interested.

A simple adjustment could be made to the amount of prior analysis that is carried out on the BLL data to create different levels of complexity. In the initial experimental set-up the result output was configured to produce distance and absorbance data. This was converted from intensity data on the server through the use of a MATLAB script. To give the student more analysis to carry out the script could be adjusted to produce intensity data instead, which the student must convert. Although this is not significantly more analysis it adds an additional step to the process.

In the gas law experiment it would be more complex to create a system in which the user could control the experiment. However, it would be possible, with a longer timescale, to modify the experiments to allow control. The gas law experiment could be improved through the simple addition of snapshot functionality to the experiment page. Allowing users to snapshot the apparatus at each time point would mean that users do not have to worry about the time they take to make measurements. As such it may reduce the amount of error that they encounter on their results. Users could also use the snapshot to zoom in on equipment if readings are difficult to decipher.

The position of the single camera within the gas law experiment makes it difficult for all measurements to be taken equally as some measurements are subject to parallax error due to the angle at which they are viewed. This could be overcome by the addition of extra cameras, each positioned at the correct height and orientation to minimise parallax error on the respective gradation measurements.

6.4.3 Expansion of Experiments

Based upon the development of the system and feedback from the students there are a number of areas which could be investigated in the future to improve and expand the functionality of these experiments. With a longer timescale for development more complexity could be introduced to the experiments, creating more interactive and dynamic experiments for the students.

Areas of expansion for the BLL law experiment include:

- Separation of the experimental controls to allow the student to control each stage of the acquisition individually. This would increase the interactivity of the experiment with minimal additional complexity in the coding. Additional safeguards would have to be incorporated to ensure that the equipment could not be damaged.
- Introduction of multiple beams containing varying light source colours. Although the Rhodamine 6G was excited with a green laser (532nm) it could also be examined with a red or blue laser to investigate how the fluorescence/extinction coefficient was affected. If the laser sources are small enough multiple sources could stack vertically on one side of the cell. Alternatively the light sources could possibly be mounted opposite one another, but laser safety must be considered to ensure the beams cannot be scattered if the solution level drops too low.
- Addition of further solutions either containing different concentrations of Rhodamine 6G or different laser dyes such as Fluorescein or Coumarin. These would require a separate laser set-up for each solution if they were to be used in a single experimental run as automating the change of solution in the set-up would be very complex. Separate set-ups could be isolated and therefore run simultaneously by different users. As an alternative the solution could be changed by a technician at pre-determined intervals. This would reduce the automation of the experiment but increase the scope of experimentation.

Areas for expansion for the gas law experiment include:

- Increase the user interaction with the system by allowing users to control the setting of the heater plate. If this is controlled via command line then the user could set the plate for a certain time e.g. 'Heater plate is on for 30 minutes'. Messages could be displayed on the webpage so other users could choose to also carry out measurements at the same time, or wait until it has finished. Example messages include: 'Heater plate on for 30 minutes, 15 minutes remaining', 'Heater plate turned off 5 minutes ago'. Safeguards could be implemented to ensure a new user cannot start a new experiment by controlling the heater plate whilst another user is still taking measurements.
- Expand the remit of the experimental set-up to incorporate more complex chemical concepts, such as heat capacities or rate of heat transfer. Similar style experiments could be developed with controllable heat sources and temperature readings to investigate a variety of different physical chemistry concepts.

Although these experiments are not currently in use due to curriculum changes their implementation was a success. The application of these experiments could be adapted with minimal effort for future use in outreach at the university or other levels of education before university.

Chapter 7

Smart Lab Interaction

7.1 Introduction

Take a moment to think. If you redesigned the scientific laboratory with any of the possible technology of the 21st century, what might it look like?

In most current research laboratories notes are hand written, instruments are isolated, measurements are not digital and there is no interaction with the system. [211] But with all the technology at your fingertips you could revolutionise the way in which a lab functions. Integrating technology at all levels of the lab could create a seamless lab environment.

When working in the lab, it would recognise who you are and automatically adjust the lab to your preferred settings, displaying the information you need at your workstation. Instruments would all be connected so measurements can be fed into your system, you could interact with your instruments through a variety of methods; voice, touch, computer interface. You could view and track past measurements and record information into a digital notebook entry with simple commands, eliminating the need for paper notebooks. Voice interaction with the lab would allow you to talk to the instruments and even the lab as a whole to request readings, enquire about the state of systems and start experiments running.

The lab environment would be continuously monitored and if any sensors detected out of specification readings then warnings would be triggered, generating audible alerts and messages transmitted to the users in charge of those areas of equipment. Users running unattended experiments would be able to check in remotely to see if their experiments are running smoothly and adjust any parameters as necessary.

This could be the reality of the lab of the future, but completely redesigning a laboratory from scratch to give the full connected lab experience would be a huge undertaking. However, integrating some aspects of these technologies into a laboratory is an an achievable undertaking and the first step towards creating a fully connected lab.

In the process of carrying out scientific research, scientists are always looking at better ways to interact with their equipment, process their data and add value to their work. The objective of this project was to investigate the interaction of the scientist with their research environment, aiming to increase the connectivity of the lab, bringing data together from multiple sources to allow them to interact with it.

A connected lab environment has many potential uses, not just for those users working inside the lab but also for collaboration and engagement with a wider audience. If a lab can be accessed or controlled remotely the resources could be shared between researchers from different institutions or it could be used as an educational tool to teach or inspire students and young minds.

There are laboratories that have implemented connected systems; however, these are mostly in the form of industrial laboratories which utilise LIMS (Laboratory Information Management System) and ELN (Electronic Lab Notebook) systems to link up their equipment and manage their data, helping them to adhere to the stringent regulations and requirements. Examples of work in research environments include work looking at linking experiment plans with a tablet interface to facilitate quicker data entry during experiments [212]; and more recent work in developing a prototype system incorporating an ELN which linked to electronic devices used in experiments [213].

The recent expansion in development of IoT (Internet of Things) devices has been widespread in areas outside of that lab, but adoption of newer technologies including ELNs and other screen based systems in a lab has been slow. [161, 168] Laboratory systems that allow user voice interaction have recently been reported by Helix [214], but these are only able to recall data from a fixed database rather than allowing the user to interact with data generated in the lab. This gives a user access to chemical information and procedures that have been populated, but does not give any real-time or lab specific data.

This work builds on research carried out previously in the Frey group in which the idea of a fully automated lab was proposed where instruments can 'talk to each other'. [215] This was expanded with work trialling IBM's middleware software to link up to sensors in a lab. [216, 217] This new work looked at utilising recently developed technologies to create new interaction methods within the lab environment, allowing the scientist to talk to their lab, linking them up to equipment and real-time sensor data. In addition to data retrieval other use cases were discussed including safety alerts and data display.

The interactive lab system should be developed in a way which makes it easy for the system to be expanded or implemented in other locations and with this in mind the technologies used should be commodity technologies wherever possible. Such technologies are becoming readily available and can be implemented with minimal customisation rather than requiring bespoke systems made specifically for the individual case.

7.2 Lab Environment

The lab environment which was selected for use in the development of the connected lab system (Talk2Lab) was a laser lab run by Professor Brocklesby and Professor Frey situated in the physics building at the University of Southampton.

This laboratory houses a class 4 laser set-up which is used in a variety of different experiments using high harmonic generation of coherent soft x-ray radiation for imaging and spectroscopy. The high power of the laser present in this system is a potential hazard to all people working in the lab and results in a highly safety-conscious environment. The laser must be controlled with great care as the beam or indirect reflections of the beam can cause permanent eye damage and ignition of combustible materials. Monitoring the system requires a number of ancillary machines to maintain safe operation.

Within the laser lab there are 5 separate areas;

- Main laser room (this contains the laser system with 4 separate experimental beam lines, the vast majority of the sensors and equipment, and a mezzanine write-up area)
- Entry room (2 parts)
- Chiller room
- Gas cupboard (housing the ventilation system)

Figure 7.1 shows a schematic of the rooms of the laser lab and the equipment they contain, Figure 7.2 is a view of the main laser table, which is depicted in the centre of the schematic (with the appearance of a letter 'E').

As can be seen from the figures, the schematic presented is a highly simplified depiction of the lab layout and does not contain all of the equipment present in the lab, but it shows the main equipment and sensors of interest that were present and relevant to the Talk2Lab development. A specific group of equipment that was excluded from the schematic were the optics used to control the beam. These would be difficult to track as they are numerous and frequently change when an experimental set-up is altered.

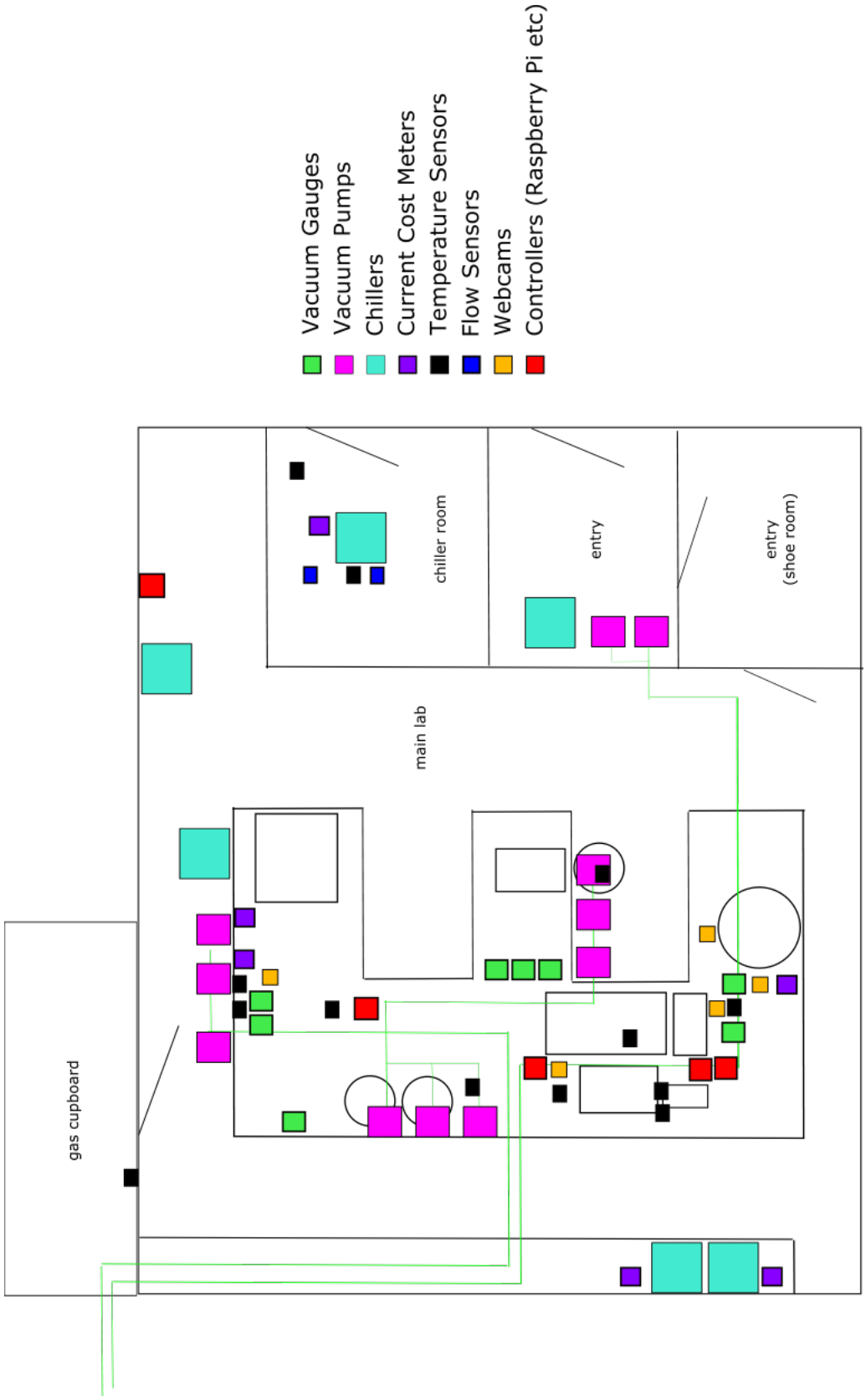


Figure 7.1: Schematic of the laser lab

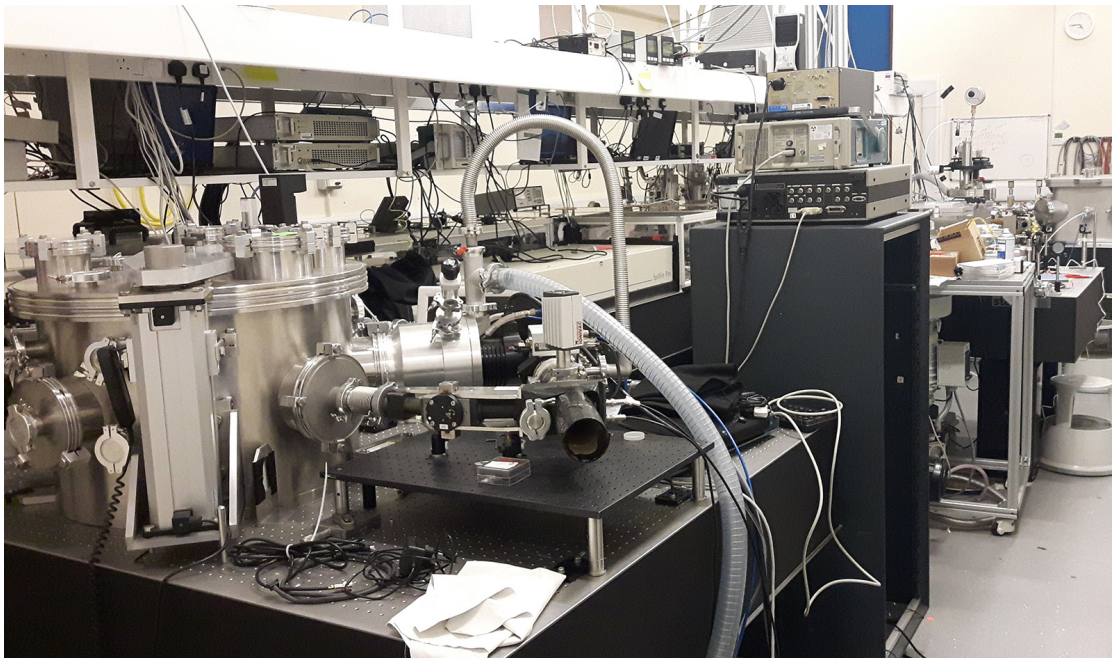


Figure 7.2: Main laser table in Physics laser lab

Within the lab environment the equipment present can be broadly separated into three categories: safety critical, experiment specific and environmental monitoring. The safety critical systems are concerned with maintaining a safe working environment and minimising risk, an example would be the laser interlock which shuts off the beam if the door is opened unexpectedly. The experiment specific equipment varies significantly depending on the work being carried out, but examples include imaging stages and vacuum chambers. The environmental monitoring equipment covers the equipment which monitors the lab, an example is the temperature sensors present around the lab.

This proof of concept focused on integrating the equipment in the environmental monitoring category. These were chosen as they were the simplest items to modify and they would not impact on the safety of the working environment if they were to malfunction. In future implementations of a connected lab the remit would be expanded to integrate all equipment into the system, but those areas that concerned the safety of the lab would need to be subject to much more rigorous testing to ensure reliability.

In an established lab such as this one there were a number of pieces of equipment that were controlled by ageing systems. Often the equipment is too specialist and expensive to upgrade regularly to incorporate newer networking capabilities. These are challenges that have to be overcome when trying to connect up legacy equipment to created a connected lab. For example, in the Southampton network it is prohibited to connect many older computers up to the network as the outdated operating systems can introduce security risks. Frequently the older systems are not equipped with adequate

connectivity or use an obsolete protocol, these require an intermediate system to convert analog to digital interface of translate the protocol and allow them to be networked.

7.2.1 Sensor Systems

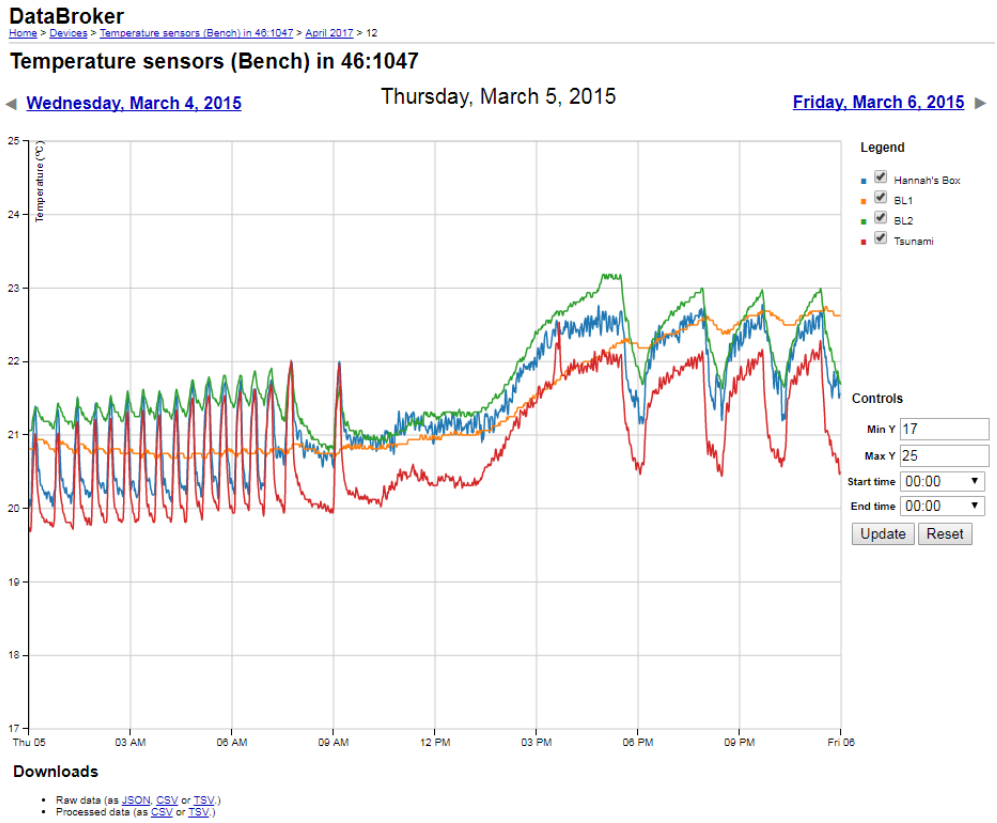
As a precursor to the connected lab project a number of sensors were installed in the physics lab to monitor environmental variables in a number of locations around the lab and provide accessible data about the lab environment to the users. The initial sensor installation consisted of a number of temperature sensors installed around the lab and power consumption monitors (current cost) attached to items of equipment.

The sensors in the lab obtained readings automatically at set time intervals. Once measured the readings for each individual sensor were published to a database[1] which stored all of the sensor data over time and could be accessed by the researchers through a website. The website was created using the d3.js framework which was also used in the 'Data Handling and Visualisation' section of this thesis for the visualisation of chemical data. When a user viewed the web-page, interactive graphs were generated for the sensor data based upon the time period selected. Figure 7.3 shows a daily plot and monthly overview for the temperature sensor readings.

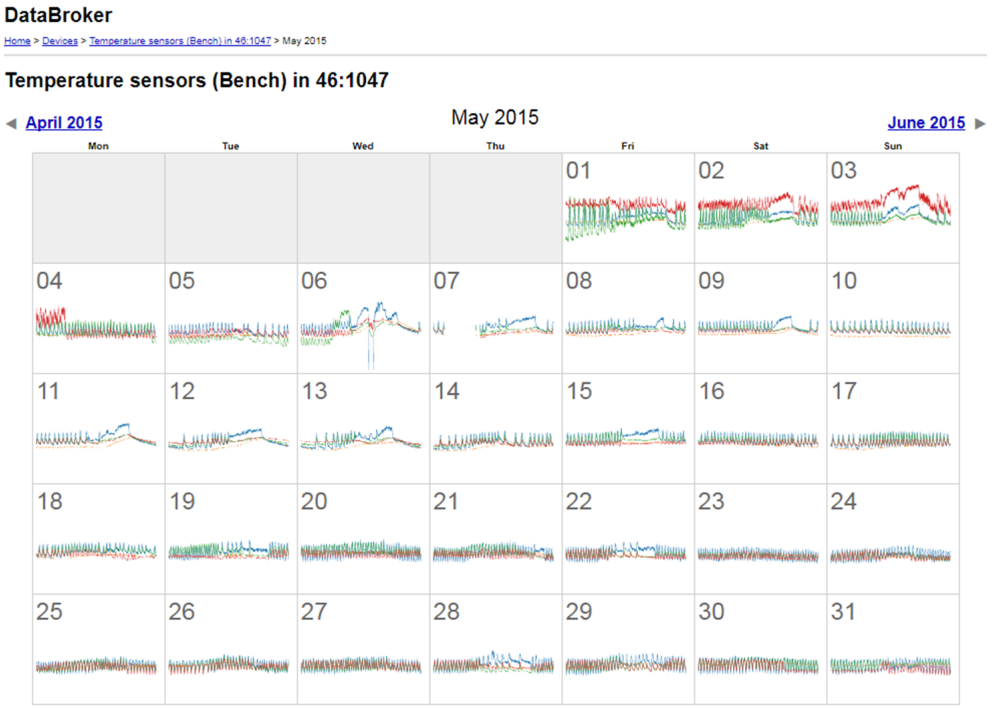
For the development of the prototype connected lab system a number of additional sensors were installed in the lab and incorporated into the sensor database. This widened the scope of the environmental monitoring with additional temperature readings, some in water and some air, water flow readings, additional power consumption and network enabled cameras. Examples of a number of sensors present in the lab can be seen in Figure 7.4. Although the network cameras were not strictly sensors and cannot provide readings in the same manner as other sensors they can be positioned to allow users to remotely read values from equipment displays that may not otherwise be accessible.

These sensors were central to the implementation of the first phase of the connected lab as they provided the underlying data for the system, allowing the scientist to interact with the lab and retrieve real-time data about the lab. A number of the technologies that were used in the implementation of the sensor network are covered in the following section.

[1] created by a colleague from Electronics and Computer Science



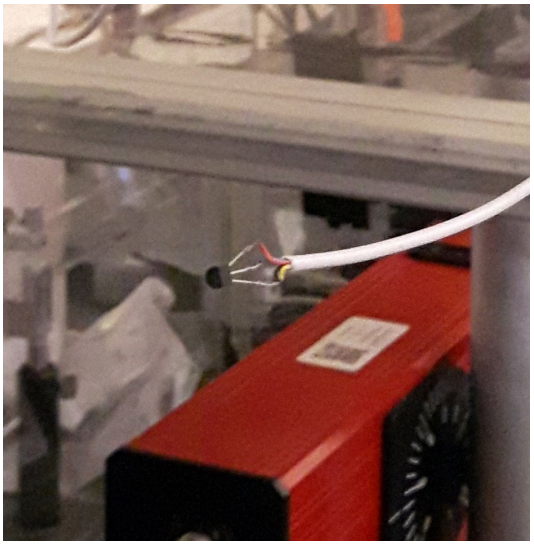
(a) Daily view of temperature sensor readings



(b) Monthly view of temperature sensor readings

Figure 7.3: Example of temperature sensor data viewed through web interface

(a) Vacuum Pressure Gauge



(b) Temperature Sensor

(c) Current Cost Meter

(d) Network Camera

Figure 7.4: Examples of sensors installed in the lab

7.3 Technologies

In the development of the connected lab system a number of different hardware and software technologies were utilised. Outlined in this section are a number of the technologies that were incorporated into the system either with the sensors for data acquisition or to facilitate the interaction and processing of the data. The technologies used in this prototype were mainstream and readily available as the aim was to create a framework which was simple to modify rather than creating a completely custom development.

- Raspberry Pi
- MQTT
- Node-RED
- Smart Assistants
- Slack

Many of the technologies investigated in this project have been facilitated by the expansion of the IoT technology. The rise of systems where objects are embedded with microchips and sensors and where 'things talk to things' has spurred the rapid development of these technologies.

7.3.1 Raspberry Pi

The Raspberry Pi[2] is one of many single board computers (SBC) that exist. First released in 2012, it has gone through an number of models with the most recent release being the Raspberry Pi 3 [218]. It was introduced as a device to advance computer science education through programming, but with its small size of just larger than a credit card and low cost it has found applications in a vast array of small projects that require more than just the basic microcontroller. [219] Beyond educational objectives [220] examples of projects range from twitter controlled hand gestures [221] and lighting [222] to RFID entry systems [223] and hydroponic automation systems [224].

Despite its small size the Raspberry Pi packs in a quad core processor, 1GB of RAM and Micro SD slot for storage and an operating system (OS). For connectivity it contains; 40 GPIO (general purpose input/output) pins, 4 USB ports, HDMI video output, 3.5mm audio output, Ethernet, WiFi and Bluetooth networking, allowing interaction with a wide variety of peripheral devices. Figure 7.5 shows the Raspberry Pi 3.

The primary OS for the Raspberry Pi is Raspbian Jessie. This is a system based upon the Debian OS that has been optimized for use on the Raspberry Pi CPUs. It is possible though to install a wide variety of third party OS images on the Raspberry Pi such as Ubuntu Mate or Windows 10 IoT Core.

Many alternatives to the Raspberry Pi exist in the SBC market although the Raspberry Pi is easily the most recognisable name in this area being the most popular SBC with developers [225] and selling over 10 million units by 2016 [226]. Another single board computer used is the Beaglebone [227]. The beaglebone slightly outranks the Raspberry Pi in terms of performance but its real strength lies in the number of GPIO pins it contains allowing a huge range of hardware interaction. However, for those projects

[2]Raspberry Pi is a trademark of the Raspberry Pi Foundation

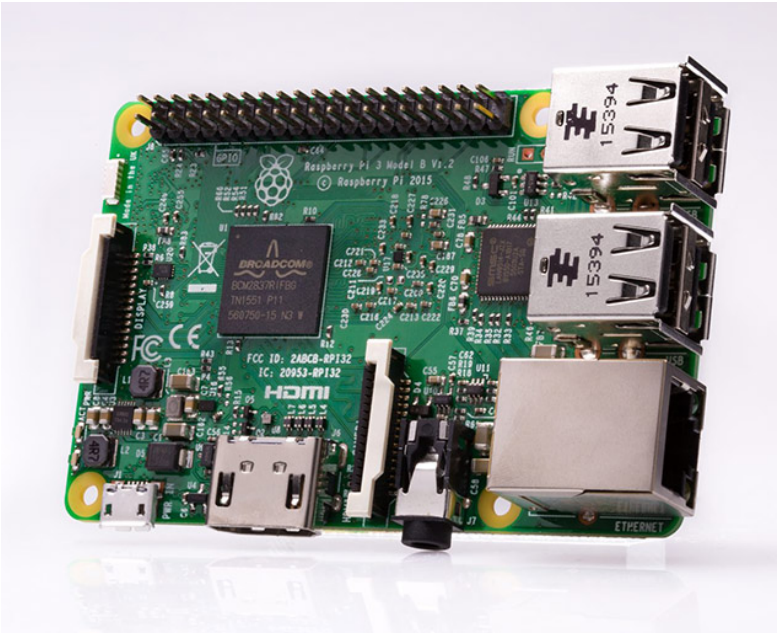


Figure 7.5: Raspberry Pi 3 single board computer

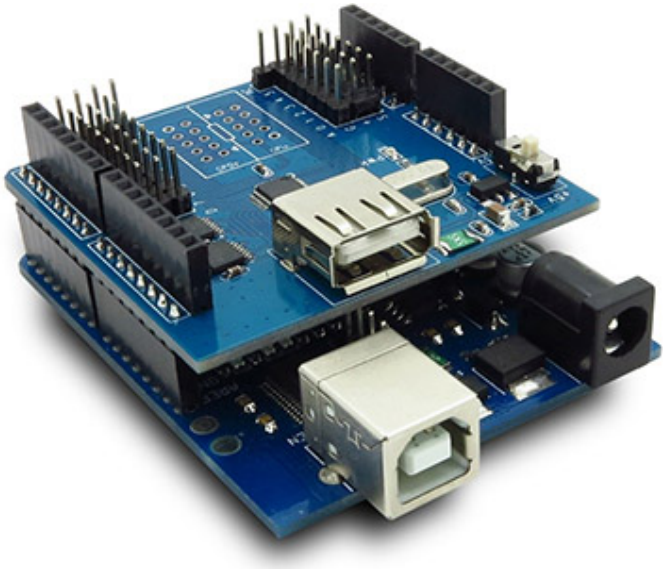


Figure 7.6: An example of an Arduino with added shield, giving USB functionality

with modest hardware interaction the strength of the Raspberry Pi lies in the extensive documentation and community surrounding the devices.

Another widely used device useful in the construction of electronic projects are the Arduino range of controllers. [228, 229] Unlike the Raspberry Pi and Beaglebone the Arduino is only a microcontroller which cannot run an operating system, but instead must be programmed in machine code. The Arduino integrated development environment (IDE) supports C and C++ languages. Although the Arduino does not run an OS it can be

very useful in many electronics projects and for prototyping systems. The Arduino was used in the development of the BLL remote experiment discussed earlier in this part of the thesis. The functionality of the Arduino can easily be expanded through the use of shields which are boards that can be plugged on top of the Arduino. Figure 7.6 shows an Arduino with a USB shield, but many other shields exist to incorporate a wide range of additional functions.

7.3.2 MQTT, Pub/Sub

Message Queue Telemetry Transport (MQTT) is a lightweight messaging protocol [230], designed to facilitate message transfer on networks that may be unreliable or low-bandwidth, such as within sensor systems. It is also widely used in the implementation of IoT devices, where machine-to-machine communication is taking place. MQTT was originally designed by IBM but since 2014 it has been a standard protocol of the Organization for the Advancement of Structured Information Standards (OASIS) [231].

Pub/Sub The underlying principal of the MQTT protocol is the publication of messages, and subscription to topics, which is known as 'Pub/Sub'. At the core of the messaging system is an MQTT broker, to which devices (clients) connect to send and receive messages. There are a wide number of MQTT brokers available as the MQTT protocol is open. These different brokers possess varying features potentially extending their capabilities beyond MQTT; however, most of them support the core MQTT features. The implementation of MQTT carried out in this project used Mosquitto [232], a broker implementing the most recent MQTT protocol version. The broker for this system was installed within the SOTON firewall.

The MQTT system allows many devices to all listen (subscribe) to messages that are provided (published) by another device. Figure 7.7 shows a potential use of an MQTT system in a house. In this example there are two devices, one publishing temperature readings and one publishing light information to different topics. These published messages are handled by the broker and can be subscribed to by other devices. A multitude of different devices can subscribe to receive the different messages and, as shown in this example, devices can connect to one or more topics.

MQTT messages are published by clients to a topic. In the example case it would have a topic containing temperature measurements, which could be structured in a hierarchy such as:

HOUSE1/temp/livingroom

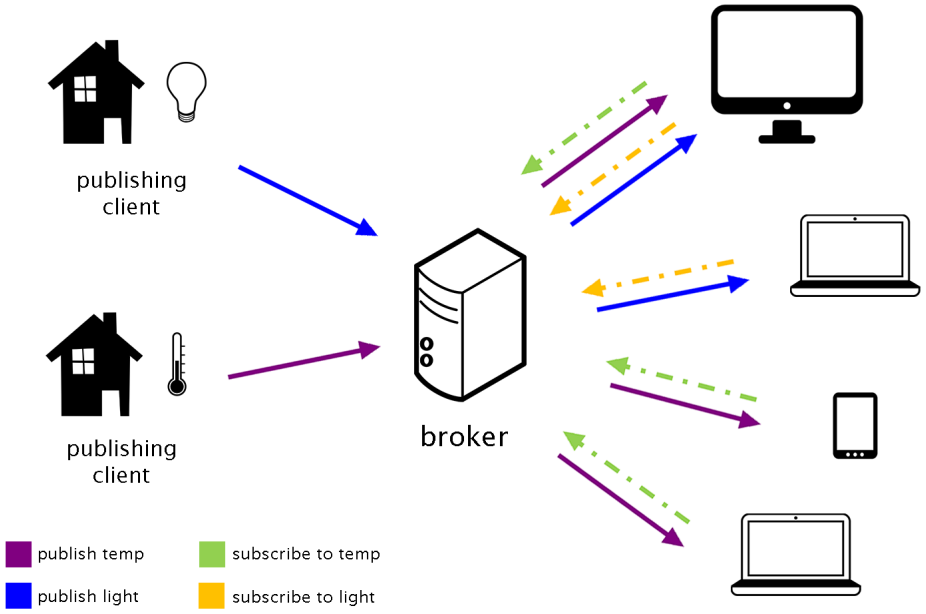


Figure 7.7: Example structure of messaging system using MQTT Pub/Sub

Clients create subscriptions to a topic or topics to allow them to receive all messages published on those topics. Subscriptions can be made to a specific topic or a range of topics through the use of wildcards, + can be used as a wildcard for a single level, or # as a wildcard for all remaining levels. Examples of subscriptions include:

- HOUSE1/+/livingroom - all messages from livingroom
- HOUSE1/# - all messages from HOUSE1
- # - all messages from the broker

Topics may also contain retained messages. Unlike the normal messages, messages that are flagged as 'retained' by the publisher are stored by the broker even after they have been sent to all subscribers. If a new client were to subscribe to that topic they would receive the retained message immediately. This can be used as a 'last known good' message, which is useful on systems that update topics infrequently. Retained messages can also be used to provide the status of a client, for example whether the device 'temperature1' is online or offline.

There are two other features in MQTT which are useful for ensuring the reliability of the system. These are Quality of Service (QoS) level flags and Last Will and Testament (LWT).

Quality of service: QoS refers to the level the broker/client will guarantee that a message is received. There are 3 levels to QoS in MQTT:

- 0 : At most once - delivers the message once, no confirmation
- 1 : At least once - delivers the message at least once, requires confirmation
- 2 : Exactly once - delivers the message exactly once, using a four step handshake

The publisher of a message will set the QoS when publishing a message to the broker, but this QoS level can get downgraded if the client subscribing to the topic has subscribed at a lower level as the broker sends messages to the subscribing clients using the QoS level in their subscription.

The level of QoS implemented in a system will depend on the use case; QoS 0 is mostly used when there is a stable connection, or it doesn't matter if occasional messages might get lost, QoS 1 when you need every message and can handle duplicates or when you can't handle the overheads needed for QoS 2, and QoS 2 when your system must receive every message exactly once, and would be damaged by receiving duplicates. Due to the four step handshake to acknowledge receipt QoS 2 requires larger overheads and takes longer to complete than QoS 1 or QoS 0.

Last Will and Testament: LWT is a feature used to notify clients when another client has disconnected abruptly (ungraceful disconnect). This is useful functionality in sensor networks as there are many cases where a client could lose connection, get damaged or run out of battery. When a client initially connects to a broker it can set a last will message which will be held by the broker until the client disconnects ungracefully, at this point the message will be sent to all subscribed clients on the topic.

LWT can be used in conjunction with retained messages to provide accurate states for clients. Clients can set a retained 'Online' status message upon connection and upon graceful disconnect set a retained 'Offline' status message, if they ungracefully disconnect then LWT can set a 'Disconnected' retained status message.

The overall structure of an MQTT PUBLISH message from a client can be seen in Figure 7.8 containing information such as the topic, QoS level and payload content of the message. CONNECT, SUBSCRIBE or other type of command messages have varying structures containing different fields in the variable header and structured payloads instead of the data contained in a PUBLISH message's payload. A key concept of MQTT is that it is agnostic with regard to the content of the payload, the payload of the message can contain data in any form as long as it does not exceed the maximum message size.

Part of Message	Content	Fields	Example
Fixed Header	Contains the control header and Packet length	Message type DUP flag QoS level RETAIN flag Remaining length (variable header + payload)	0011 (PUBLISH) 0 (FALSE) 1 0 (FALSE)
Variable header	Contains additional control information	Topic name MessageID (if QoS 1 or 2)	"HOUSE1/temperature/kitchen" 425
Payload	Contains data for publishing	Data	"temperature:22.4"

Figure 7.8: Outline structure of an MQTT publish message

7.3.3 Node-RED

Node-RED [18] is a flow based programming tool originally created by IBM for connecting IoT devices [233]. It utilises a browser based programming editor to connect chunks of code, called nodes, together to create flows connecting hardware devices and services, in particular IoT devices.

Node-RED is built upon the Javascript runtime Node.js, which is a lightweight and efficient runtime using an open source package system, npm, to deliver expandable features through its library. It can easily be installed on a local device, a microcomputer, such as the Raspberry Pi or Beaglebone, or installed in the Cloud. The latest versions of the Raspberry Pi and Beaglebone already come with Node-RED pre-installed on the default image so they are simple to get running.

Within Node-RED, flows are created in the visual editor by dragging and dropping nodes into the flow workspace and connecting them up to create the pathways. Each node takes in some data, carries out a defined function on the data and passes it on to the next node. The various nodes can be implemented and customised with little additional coding. The beauty of the system stems from the ease with which devices can be linked up without having to focus as much on the underlying code. Flows can also be shared and imported into other systems which help create standardised systems which only require minimal modification when implemented in a new location.

An example of a simple flow can be seen in Figure 7.9; this flow is a simple weather check and email. The first node in the flow is an inject node used in this example to trigger the flow, but a multitude of other triggers also exist allowing integration with MQTT, HTTP and other communication methods. The second node checks the weather of a specific city, in this case Southampton GB, passing the response message to a function node which creates an email. Finally the last node in the flow sends an email containing the weather message. Also present in this small example is a debug node which is useful

when working with the system to see where errors are being generated, in this case it displayed *msg.payload*.

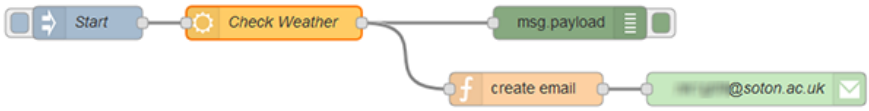


Figure 7.9: An example of a simple flow within Node-RED

The Node-RED library contains a wide variety of built-in functions and this palette of nodes can be extended by installing community created nodes or, if a node with the desired functionality does not exist, the user can create a custom node for a specific task. These flows can be used to connect up a vast array of different devices, in particular the functionality of interest for the connected lab is the ability to receive and respond to http requests and subscribe to MQTT channels.

7.3.4 Smart Assistants

Smart assistants are a type of virtual assistant which can perform tasks at the request of a user. The offerings in the smart assistant market have recently boomed with devices from Amazon (Alexa, Echo), Samsung (Bixby), Google (Google Assistant) and others such as Siri and Cortana. These products capitalise on widespread use of smartphones and increasing connectivity of devices through the Internet of Things. Smart assistants can either be present as an app on a device such as a tablet or phone or more recently as a stand alone device.

A key concept of these emerging assistants is the use of voice based user interfaces where questions and commands are spoken to the device, with actions or spoken responses being returned to the user. Voice commands given to the assistant are usually processed in the 'cloud' to harness the power of machine-learning algorithms and allow continual development.

Although a large number of smart assistant devices have been released recently the development in this project focused on the use of Amazon's smart speaker devices. These were selected as they were standalone rather than smartphone based, were well established in their functionality and they had good resources available to allow the development of custom interactions.

Echo & Echo dot: The Amazon Echo is a voice interactive smart speaker designed to function as a standalone smart assistant in the home. The Echo aims to bring connectivity to a range of services and products such as smart home systems and online

Figure 7.10: Echo Dot by Amazon

purchasing, through conversational interaction with the Alexa voice services (AVS). The functionality of the Echo devices can be expanded through the addition of skills, which are continually being developed by their providers. Although a number of the skills could be termed 'junk skills' and do not provide significant benefit to the user, skills have been developed which allow interaction with smart home thermostats, such as Nest Learning Thermostat and Hive Active Heating.

The Echo is activated through the use of a 'wake word' and then a question or command, such as: *"Alexa, play radio 4"*, *"Alexa, what is the weather today?"*, or *"Alexa, ask EDF Energy to check my account balance"*. The processing of each request is carried out on AVS which generates a response to the query. This response is then spoken over the speakers, or if it is an action such as playing music the music will play over the speakers.

Figure 7.10 shows the Echo's smaller sister device, the Echo Dot, which was released in March 2016 [234]. The Echo Dot is very similar to the Echo in functionality as the processing for both speakers is carried out externally to the device. The Echo Dot can be connected up to external speakers as the integrated speaker is not very powerful; however, for most non musical purposes the integrated speaker is sufficient.

The Echo and Echo Dot both have an array of seven microphones in the device to give superior noise cancellation far-field voice recognition to enable voice recognition in noisy environments. This also allows multiple devices with the same wake work to be linked together, Alexa uses echo spatial perception to automatically determine which device is closest to the request and respond through that device.

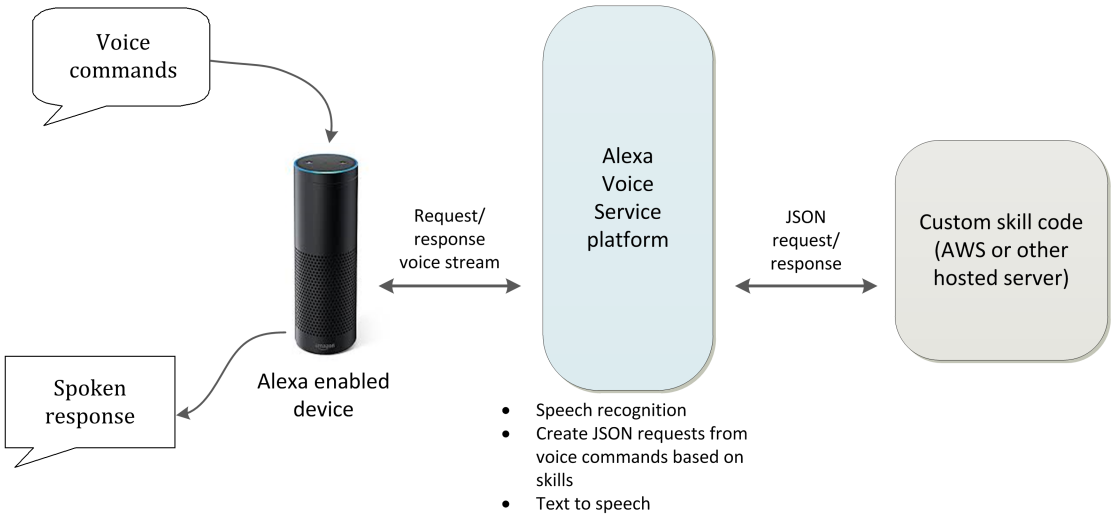


Figure 7.11: Alexa voice command processing

Alexa Processing: The Alexa Voice Service is the processing engine behind the Echo devices. It combines automatic speech recognition (ASR) and natural language processing (NLP) with a library of 'skills' to enable conversational intelligent voice control.

AVS is not solely available on the Echo devices but has recently been made available in the UK [235] for installation in any connected device with a speaker and microphone.

Alexa devices are always on, listening for a wake word (Alexa, Echo, Amazon or computer depending on settings) and record the speech directly following the wake word. This speech recording is streamed to AVS where it is analysed through speech processing and skill details to identify and structure the request in JSON format. Requests are then passed to the relevant skill to generate the response. Once the response is generated it is passed back to AVS and converted to the spoken response, Figure 7.11 shows the processing of requests.

Skills on Alexa are like an app on a mobile phone or a program on a computer. Skills give different functionality to the device such as ordering a taxi, setting an alarm or interacting with smart home lighting. Each skill can be added to an Amazon account, from which it will work with all Alexa devices set-up to that account. Skills can be added from a library of skills published by companies and developers, with the developers area (Alexa skills kit) being open to all, allowing anyone to develop their own skills. As the skills are hosted in the cloud rather than on each device the skills can easily be updated without requiring users to download new updates.

7.3.5 Slack

Slack [29] is a communication platform designed for work teams, to bring together multiple areas of communication. Slack contains a wide variety of functionalities that can be tailored to the group or users needs. At the basic level teams can set up open or private channels in which to send messages, share files or make calls to other team members. Users can send messages to channels, direct messages to specific people or tag people in messages.

The Slack interface can be seen in Figure 7.12 accessed through the web interface. Slack can also be accessed through apps on Windows, Mac, Linux, iPhone and Android, giving access to the platform wherever the user is.

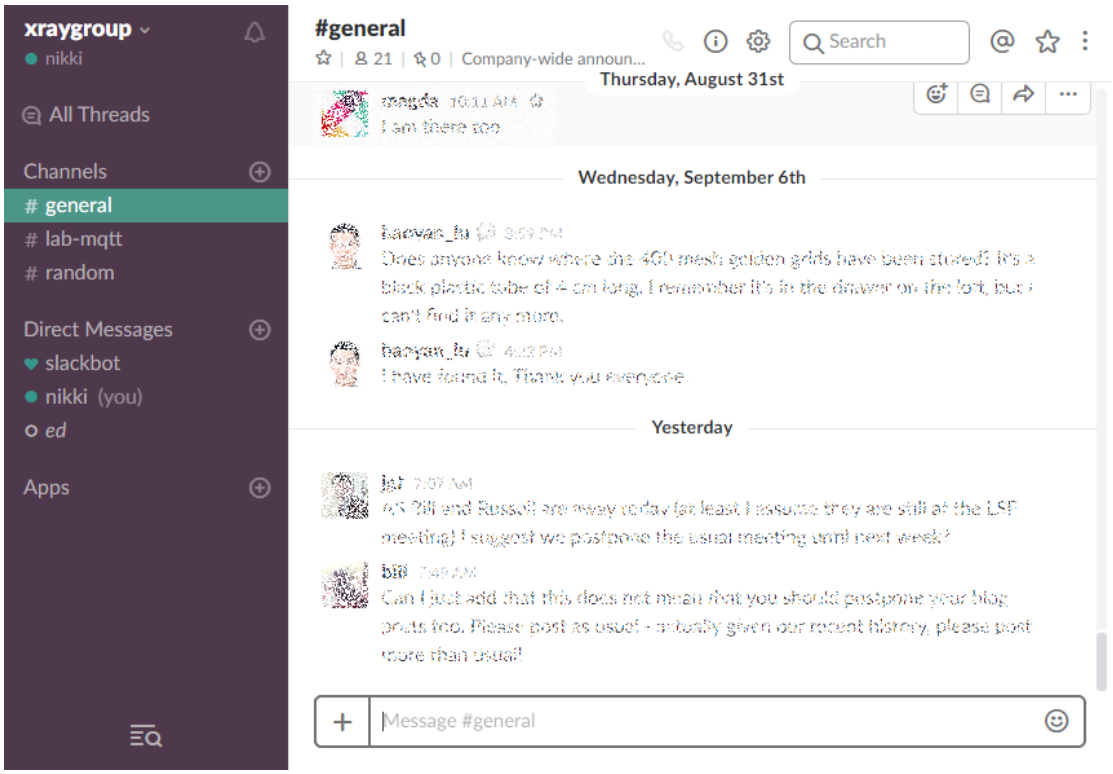


Figure 7.12: Slack communication platform - accessed via web interface

There is also extended functionality allowing integration with external applications such as Twitter, cloud storage platforms and most importantly Node-RED. Previous messages sent on the channels remain visible and can be searched at a later date if necessary.

The Slack platform was already being used by the group in the Physics lab as their communication platform which made it the obvious choice for integration with the connected lab development.

7.4 Talk2Lab Development

For the creation of a connected lab environment the Talk2Lab project was created. This project aimed to create a prototype Smart Lab system using commodity technologies to give the scientists access to new data and allow interaction with the lab environment in new ways.

The first implementation of this connected lab project focused on creating connections to lab sensor data collected in the lab, enhancing the data being collected and the methods by which the scientists could retrieve it. There were no plans initially to incorporate functionality to control equipment in the lab, although use cases for ways in which control could be added were discussed as future ideas. The addition of remote control for equipment would have had significant safety implications due to the nature of the equipment. That functionality would require extensive testing to ensure that the system functions correctly under all circumstances and would not impact of the safe working environment in the lab.

Throughout the project a number of workshop sessions were run at the University, bringing together expertise from different research areas within the University and external companies for collaborative brainstorming about the connected lab. Alongside these workshops work was also carried out with the scientists in the lab discussing the needs for the systems and use cases in which it could operate. To develop the prototype of the connected lab the sensor systems were updated and expanded to allow access to more data in the lab, new interaction methods were developed to allow users to access the data through voice and frameworks were created for expanding the capabilities of the system in the future.

For the creation of a system like this there were a huge variety of technologies available, ranging from open-source to purchased software to self coded/built items. The technologies outlined in previous section were the technologies selected during this development. They may not be the best solutions in the long run as the requirements for the system develop; however, they were selected for the prototype for their low cost, familiarity and ease of implementation. In particular it may be desirable to revisit the choice of voice based user interface to determine if other systems are available which do not rely on Amazon.

7.4.1 Use Cases

Throughout the development of the Talk2Lab system multiple sessions were held with scientists who work in the laser lab. The aim of these sessions was to identify ways in which the system could enhance the working environment within the lab, using the

experiences of the subject matter experts. In some sessions this focused on cases which would be straightforward to integrate into an Alexa/NodeRED interface, while other sessions encouraged more blue sky thinking to discuss what could be possible if there were no limitations.

The discussion of these use cases covered what functionality would be desired and how it could be achieved, alongside some of the potential issues and pitfalls that may be experienced in implementing them.

These use cases were developed by focusing on those aspects which would be of use to the scientists in the lab, so although they were not all possible to implement straight away, they will form the focus of the development as it continues.

Single source questions: Using the sensor data as inspiration these were questions which use information from a single data source to provide a response. The simplest of this style of questions were those which require a single data point to be returned. For example requesting the temperature or pressure of a specific sensor, *"What is the temperature of tsunami?"* This would be the simplest dynamic question that could be asked, requiring the system to retrieve only the current value for a specific sensor.

More complex questions could also be incorporated for a single source, such as; *"How has the temperature of tsunami changed in the last hour?"* or *"What is the average temperature for tsunami?"*. These questions require the system to have more context and process more data points to generate the response. As the complexity of the question increases it requires additional data and coding to process the request.

A method for getting the system access to more data for analysis would be for it to always store more data points, which could be very memory intensive depending on the scale of the stored data. Alternatively it would need to query a database to retrieve the relevant data each time there was a request, which could be more time-consuming but would allow for easier addition of extra questions.

Additionally the system would require more context for some questions, when taking the average it is necessary to know what time period this average is over. This additional context could be determined from a multi-step interaction where the system asks for additional parameters that it requires. Currently voice interaction software is unable to easily handle all the context that is involved in multi-step interactions; however, these products are continually developing and their ability to handle more complex situations will increase in the future. An alternative method is to use default parameters, if a user specified the time period in their request the system would use that, otherwise it would set to the default parameter. For example if the user asked *"What is the weekly average for tsunami?"* the parameter would be weekly, otherwise it would return the average for a set time frame such as the last 24 hours.

Qualitative and Multi-source questions: These cover more specific questions that would be asked by the scientists related to the behaviour of the equipment - for example: *"Is the laser on?"*, *"Is the laser working well?"*, *"How is beam line 2 doing?"* Although these might seem like simple questions, they are significantly more complex than questions such as *"What is the temperature of tsunami?"*

Questions which enquire about qualitative measures, such as how well a device is working, require additional behaviours to be defined and programmed for each piece or group of equipment. These behaviours could either be thresholds which determine an acceptable reading, or banding which could show good, satisfactory and bad levels for the equipment.

In addition to defining the acceptable levels it is not always possible to determine the state of a piece of equipment from a single sensor reading. Some equipment may require multiple sensor readings to be combined in processing, to give an overall state in which several criteria must be met to attain the acceptable behaviour.

The laser system in the lab is complex and composed of many different parts. When querying about the state of the laser there are many different parameters which must be considered. For example, the laser 'on' state cannot simply be inferred from sensors like the power consumption, as the power is always switched on to the laser, unless the entire system has been shut down. Similarly there is no single piece of equipment that determines the overall behaviour of the laser or its output.

To create questions about the state of the equipment it is necessary to identify the relevant sensors that are required to determine the behaviour, then define the combination of these sensor readings that gives 'good' or 'bad' behaviour. Once this structure is defined it could be implemented into the voice interaction system.

Dashboards: Digital Dashboards are frequently used in business to display a key selection of values or to visualise a set of data to give overall views of performance. In the lab environment it would be beneficial to have access to multiple readings from the sensor system in one place. This would involve a different method of interaction to a voice interaction system as the smart speakers can only respond directly to the user with audio.

An efficient interaction method would be a touchscreen display which presents a digital dashboard. The display could give the user an at-a-glance overview of the state in the lab, pulling in and visualising data from different areas of the sensor systems. The d3.js system that has been implemented in the existing visualisations could be implemented in the dashboard, creating dynamic displays using the lab data.

The dashboard could show the overall state of the lab but also allow access to more in depth data when requested. For example it could show the current temperatures for all temperature sensors, but when a single sensor is selected it could display the temperature for that sensor over the last 24 hours. Separate users could also create custom views on the display which show the sensor readings which they are most interested in.

Access to the dashboards could also be provided via a web interface allowing users and supervisors to monitor the lab status from a remote location.

Viewing camera feeds: Within the lab there were a number of network enabled cameras which were positioned to view the output from equipment or included in the experimental setups. In particular cameras were positioned to capture the output of equipment that is crucial to alignment of the laser beam. The adjustment of the laser beam previously required two people to be present as the output from the monitoring equipment was not visible from the position where the equipment is adjusted.

Using a network enabled camera allowed the alignment to be more easily carried out by a single person, but it required the use of a laptop in the vicinity of the laser. Using the Talk2Lab system the network camera could be visualised through the use of a command, *"Show laser alignment camera"*. The Alexa system would control the display of the video feed through integration with a Raspberry Pi. The camera feed could be displayed on a large flatscreen in a position that is easily visible from where the laser beam is adjusted, allowing the user to adjust the laser beam more easily and get instant visualisation of the monitoring output without use of a laptop.

Additional commands could also be included to stop the feed and show the feeds from the other network cameras. If the screen was not visible from all necessary points additional screens could be installed which could also display the network camera feeds if commanded. Following on from the visualisation of the camera feeds the system could be used to capture snapshots from the feeds and save the snapshots to a specified location. Commands like these would require additional context; however, this could be included in the command such as *"Save laser alignment camera feed for Greg"*, where the save locations are associated with individual users.

Control of the Laser: Control of the laser was functionality that was strongly desired by the lab users, in particular the ability to remotely turn the laser on. This was because the laser requires a lengthy start up period before it can be used for experiments. If it was started up remotely this could maximise the experimental time when the user is in the lab. This use of the connected lab was the least likely to be implemented as it concerned the control of the laser itself. Control of the system rather than monitoring introduces a

raft of additional complexities and safety concerns, particularly when involving expensive and potentially dangerous equipment.

While it would be an exciting and beneficial use of the connected lab the start up of the laser will be a future project. In addition to the safety concerns, the IoT sphere is still rapidly developing and an area of concern is the security of the internet connected devices. [236] The system would need to be very robust to minimise as far as possible the possibility of unauthorised access to the control of the laser.

Another area of laser control is the shutdown of the laser system. If the steps of the laser shutdown procedure are not followed correctly this can cause issues with the laser system. Automating the shutdown procedure could ensure that the steps are consistently carried out in the correct order. Automation of the laser shutdown would be less risky than the start-up of the system; however, it would still require extensive testing and safety checks to be in place.

Image analysis: The users in the lab are frequently required to be able to identify the behaviour of the laser from the visual output of equipment, without additional analysis being required. One output which is used to identify laser behaviour is the Frequency-resolved Optical Gating (FROG) trace. This output is a spectogram, the shape of which can characterise the behaviour of the laser pulses.

This FROG trace does not have an output which can be directly linked to the connected lab system as due to the age and OS of the machine it is not permitted to connect this machine to the internet. Instead the visual output is used to characterise the laser pulses and if further analysis is carried out it must be manually copied across. If the visual output could be analysed automatically then the behaviour of the laser could be linked in to the connected lab system.

Image analysis could be applied to the FROG trace via a network camera which would capture the screen output from the non-networked PC. The analysis could be used to determine the shape of the output in the image and trained using sample images to identify what pulse behaviour this corresponds to.

Another situation in which image analysis could be used is in one of the beam line experiments where the output is an image from which roundness is measured. Carrying out image analysis on an experimental result could be the first step towards automating the experimental run.

Alerts & Warnings: In the connected lab system the sensors and measurements will be continually monitored and recorded. This data collection can be used help the lab users by activating warnings and alerts when unexpected conditions are detected in

the lab. Alerts would notify users of issues which could be a hazard to equipment in the lab or the safety of users as well as informing them about issues which may affect experimental results. If a user can rectify the issue during an experiment it may prevent them from having to repeat the experiment.

The unexpected behaviour could come in many forms, such as sensor measurements being outside of an acceptable specification. For certain sensors such as temperature this would be a simple range of acceptable values, but sensors like vacuum pressure may have a number of different ranges including whilst under vacuum and not under vacuum, which would require the system to identify which range is applicable.

Other specifications may require a combination of multiple sensor readings to determine if it is out of specification. One situation that could cause damage to the equipment is if the laser interlock, which blocks the beam for safety purposes, is accidentally tripped and not re-set shortly after. This would require knowledge that the beam is powered up and reaching the final chamber of the laser, but no output beam is detected.

When out of specification behaviour is detected a warning should be triggered. The recipient and method of alert should be dependent on the type and severity of the unexpected condition. A slightly out of specification temperature would not be very severe, a significantly out of specification temperature or multiple out of specification temperatures would be more severe and a situation like the interlock being tripped would be urgent. The alerts could be sent through a number of channels such as Slack, email or text message, where possible urgent alerts could also be delivered audibly in the lab.

7.4.2 Talk2Lab workshops

A series of workshop sessions were run for the Talk2Lab project with attendees from within the University and external companies. These workshops covered a variety of topics incorporating both the theoretical side of the development in addition to practical development in 'hackathon' sessions.

The workshops began with discussions on the possibilities of what a connected lab could do and how it could be achieved. Topics included hardware and software solutions that could be useful in the creation of the system and areas in which the lab could be connected, along with the safety and security implications of these.

Suggestions raised at these workshops were wide-ranging with some that would be easily achievable and other ideas that were more suited to future development, a number of these discussions cross over with the use cases that were explored with the lab users discussed in the previous section. Items covered in the discussions included:

- Turning the laser on/off

- Alerts for equipment if there was an issue, and what the issue was
- Long term monitoring and tracking of equipment status
- Microphones monitoring equipment vibration for abnormal behaviour
- Recording data - automated experimental runs, dictation to a lab notebook, snapshots from a camera
- Monitoring laser performance - is the laser working well
- Communicating with the system via other means, Slack, text message etc

Following on from these discussions the development focused on the creation of an interface for the existing temperature sensors that had already been installed in the lab. The aim of the initial hackathon sessions was to create a simple interface, allowing a voice command to retrieve live environmental data from the lab. For the first hackathon session we were joined by a number of collaborators including Andy-Stanford Clark[3] from IBM, whose knowledge of Node-RED helped to create the initial flows in Node-RED to retrieve data from the lab.

7.4.2.1 Initial Alexa Integration

The initial Node-RED flow was designed to create the interface between the lab data and the Amazon Alexa system, pulling in lab data from a number of sensors and allowing requests for information to be made through the Echo Dot. Figure 7.13 shows the initial Node-RED flow that was created. This flow allowed the user to retrieve the most up-to-date reading for each sensor.

The flow had 3 sections which monitored sensor data and stored the values in the system; these linked in to the lab sensor data by subscribing to the MQTT feeds published by the temperature, flow rate and current cost sensors in the lab. The temperature feeds contained readings for all sensors in every message which required no additional processing. However, in the current cost feed each message only contained the reading for a single unit which required pre-processing to ensure the data was correctly stored for all units.

The flow also contained a section monitoring for requests coming from the Echo via AVS. When a request was received this triggered the flow which processed the request and generated a response for the user.

In order for the Echo device to interact with our system via AVS it was necessary to create a custom Alexa skill in which specifications were defined for the questions (intents)

[3]Developer of the MQTT messaging protocol

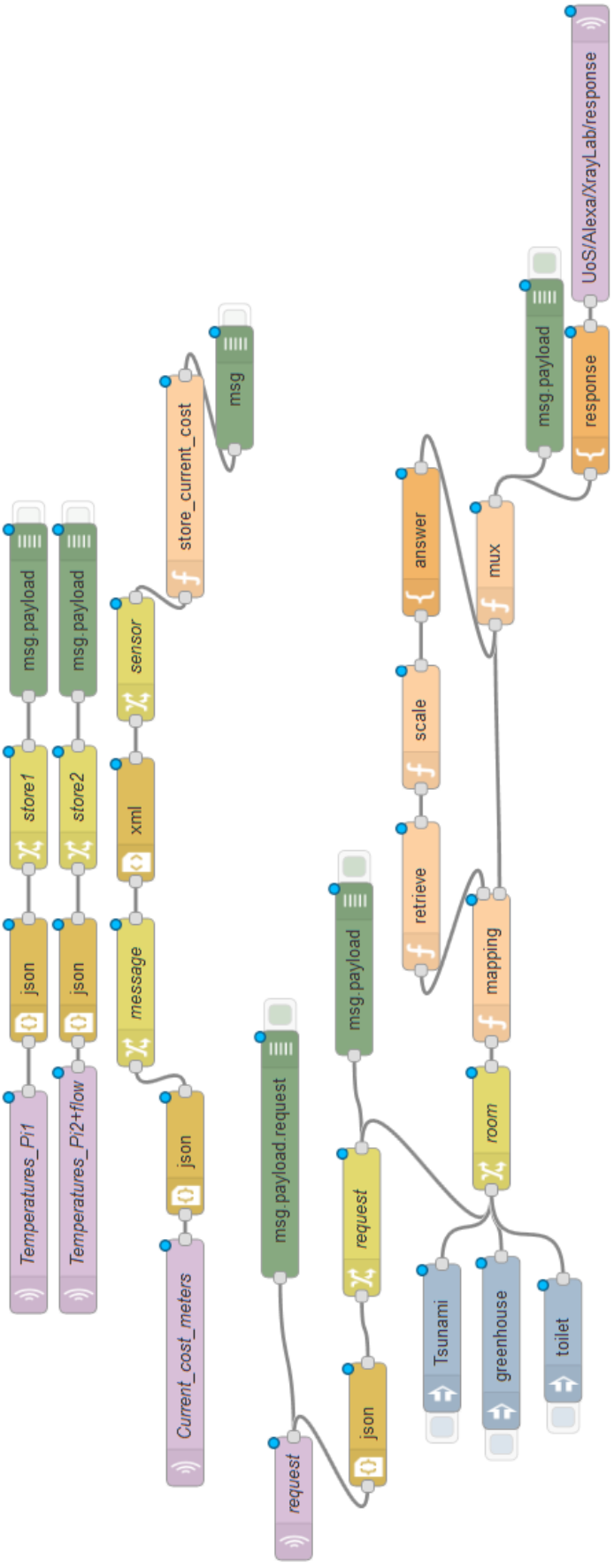


Figure 7.13: Initial Node Red flow from first Hackathon session

that would be posed to the Echo device relating to the lab, and the method by which Alexa should handle any requests to the skill.

Initially a single skill was created through the Alexa skills kit, with a single intent defined as "GetTemperature". The intent schema that was defined for the skill is shown in Figure 7.14. This skill was intended to take requests for the temperature of a number of 'rooms' which are the various sensors located in the lab.

A limitation of the Alexa system was that the slots had to be populated via the skills website and could not be pulled in from a list or database. For a static system this would not be an issue. However, in a system that was dynamic or required expansion this could cause a number of issues. This may be addressed in a future expansion of the Alexa system or if it causes significant issues the choice of Alexa as the initial processing engine could be reconsidered.

```
{
  "intents": [
    {
      "slots": [
        {
          "name": "room",
          "type": "LIST_OF_ROOMS"
        }
      ],
      "intent": "GetTemperature"
    }
  ]
}
```

Figure 7.14: Intent Schema in Alexa Developer for X Ray skill

Prior to the interaction request entering the Node-RED flow it has already undergone the first stage of processing carried out on the Amazon server. This performed the speech to text processing and natural language processing to attempt to identify the skill, intent and any slots based upon the structure of the skill detected by the Alexa device. Once this first stage of processing had been completed the request was passed to the endpoint specified by the skill, in this case the request was passed to the Talk2Lab Node-RED server.

The incoming request from Alexa was formatted in JSON with the structure outlined in Figure 7.15, in this example it contained a request for the temperature of tsunami. Alexa has identified the intent of 'GetTemperature' and value of 'tsunami' for the slot 'room'.

In processing the request in Node-RED the room slot value (in the example above 'tsunami') was extracted from the request and mapped to the sensor ID that was present in the incoming MQTT feeds. The relevant temperature value was then retrieved from

```
{
  "session": {
    session information
  },
  "request": {
    "type": "IntentRequest",
    "requestId": "EdwRequestId.ecf64ee1-e0eb-4cd1-887b-547c5dc8ee9b",
    "intent": {
      "name": "GetTemperature",
      "slots": {
        "room": {
          "name": "room",
          "value": "tsunami"
        }
      }
    },
    "locale": "en-GB",
    "timestamp": "2017-09-15T17:30:36Z"
  },
  "context": {
    context information
    }
  },
  "version": "1.0"
}
```

Figure 7.15: JSON request structure from Alexa

the stored sensor data, scaled and the response message created for return to the Alexa device. Upon receipt of the response message the Alexa device converts this to speech.

This initial flow successfully gave the lab user access to the current temperature readings for all of the temperature sensors attached to Raspberry Pi 1. The way in which the sensor MQTT feeds were structured meant that the sensor feeds came in based upon which Pi they were connected to in the lab. This was convenient for their installation but the MQTT feeds could be better structured based upon sensor type or having all sensors in the same group, which could give access to all of the sensors in one section of the flow. The structure of the MQTT feeds would also need to take into account any pre-processing that is carried out during the storage of the data.

The power consumption values (current cost) were also pulled into the Node-RED flow from the MQTT feeds; however, these feeds were not useful in their raw form as the sampling rate can cause the reported power consumption values to fluctuate quite considerably. Values like these would require storage of data over a period of time which could be processed on the fly, for example providing an average of the 10 most recent readings. Processing such as this would also be required for some of the ideas raised in the workshops where pieces of equipment are tracked over time or compared to their 'normal' state.

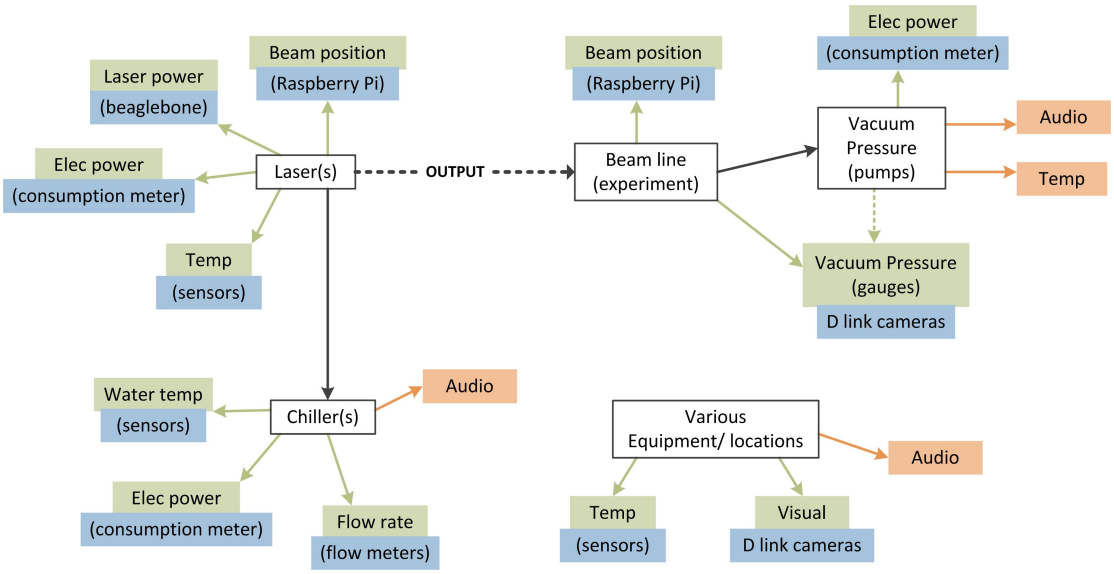


Figure 7.16: Equipment and Sensors map showing connections for physics lab Green items are 'measurements' taken and blue the equipment taking the measurement

7.4.3 Cataloguing Equipment

The processing of MQTT feeds in Node-RED and incorporation of additional sensors for the system highlighted the necessity for the equipment in the lab to have a defined structure, rather than all equipment being named and referred to differently which is often how a lab environment naturally evolves due to the cycle of users through the lab.

In order to consider how to define the structure it was necessary to catalogue all of the equipment present in the lab. A database was created where each item in the lab was recorded. Over 100 items were added to the database ranging from beam lines that contain experiments to environmental sensors.[4]

The items in the database were assigned a UUID (Universally unique identifier) that can be used to ensure that each item can be referred to specifically without any risk of duplication. The database also included the names that the lab users used for the equipment, in some cases multiple names or nicknames may refer to the same item of equipment where the name has evolved over time.

In addition to the item's name the database was also populated with further information about the item including location, device type, manufacturer and serial numbers where the information was available. Although only a small portion of this information would be used in the initial stage of the development, future development of a robust system will benefit from a thorough inventory carried out with the help of lab users.

[4]The full equipment database can be found in the ESI. [30]

To aid the organisation of information and retrieval of data, records are often organised into a structure or processed through a classification system. This enables more efficient processing of the data in search, retrieval and analysis.

When considering the different methods for organisation of the equipment data it is necessary to understand how the equipment and readings are related. Figure 7.16 shows a simplified connection map created to link the equipment present in the lab. This map links the various types of equipment and the measurements that were associated with them. The black boxes show key items of equipment, the green boxes show 'measurements' taken and the blue boxes are the devices taking the measurement, which were also pieces of equipment. The orange items are measurements that were not initially taken but were planned for future expansion.

There are two major different methods for structuring the equipment, one method is to use a flat-file structure where all of the equipment is at the same level and the other is to implement a hierarchical structure, which relies on the creation of a tree structure through which the equipment can be classified. For this system a flat structure was selected, this was the simplest option as it does not require a taxonomy to be created or implemented for the equipment.

If the hierarchical method was selected a classification tree would have been required. Due to the complex nature of the lab connections there were many different ways in which the equipment be classified which included; equipment type, location, function and how it is connected. However, none of these classifications simply and sufficiently captured the structure of the equipment. The connections were challenging as items often had multiple functions, were connected with more than one piece of equipment or location, and were not static in their positions.

The flat structure contained all the equipment at the same level and information about the items can be stored as the item's metadata. Although this method does not allow selection of groups of devices through the tree it allows far more flexibility in the relationships between items. It can be also be used in conjunction with taxonomies which define relationships between categories and their subcategories for item selection.

7.4.4 HyperCat

The method selected for the catalogue implementation was HyperCat [237], this is a JSON-based open specification for cataloguing IoT assets. A HyperCat catalogue contains a collection of uniform resource identifiers (URI) and metadata from the item, contained in RDF-like (Resource Descriptor Framework) triple statements. Hypercat is designed to expose information about IoT assets and allow interrogation of the asset information.

A Hypercat catalogue is a collection of resources, which may themselves be another collection or a simple item object. Each item in the catalogue must contain a `href` which is the identifier for the object and `item-metadata`. Information in `item-metadata` is given by a JSON array of `rel val` pairs. The only required `item-metadata` is a resource description, other metadata are defined by the catalogue creator.

The catalogue of lab items was used as the basis for the Hypercat creation. Implementation of a catalogue which can be queried would allow the automatic population of data in the processing system. In the initial Alexa integration slots had to be manually populated via the Alexa skills kit, but potential future developments may allow Alexa to pull the slots in from a database. Alternatively a different NLP engine could be implemented to link up to the catalogue.

Pulling data in from the catalogue would also extend to metadata for warnings and alerts, such as limits for specifications and other parameters necessary for triggering alerts. The catalogue also allows functionality to be built around groups of sensors, for example showing all of the temperature values on a screen or asking if the temperature in the lab is okay, where the system would be required to examine all of the temperature values and respond. This facilitates the creation of digital dashboards and multi-source questions from the use cases.

Looking toward the expansion and creation of new systems, the use of a catalogue of equipment would be beneficial as new sensors and items can simply be added into the catalogue and this would incorporate them into the interactive system. The addition process would require certain metadata fields to be populated when new items are added to the system, ensuring that items and their data can be properly linked up.

7.4.5 Further Integration

Following on from the initial integration of the sensor system with the Alexa voice interaction additional functionality was explored, including the incorporation of extra intents to the Alexa requests, additional interaction pathways and further sensor measurements.

Laser Power: Further development allowed the addition of laser power values to the system. The laser power monitor which operated on a Beaglebone was modified to publish its values via an MQTT feed in addition to its existing web interface. A second intent (GetLaserPower) was added to the Alexa skill to handle questions about the laser power. As the laser power did not have multiple sources, such as multiple power readings, the skill did not require any slots to be added.

The Node-RED system was modified to split the flow dependent on the intent that was identified by Alexa, Figure 7.17 shows the modified section of the flow concerned with

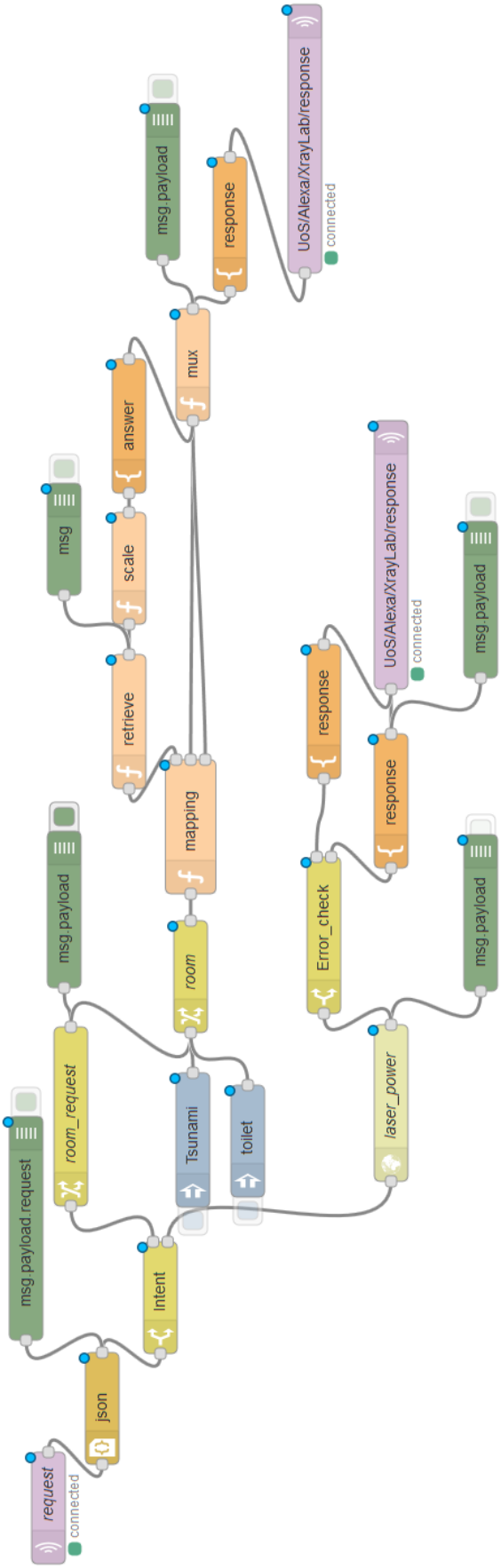


Figure 7.17: Node Red flows for Alexa requests with 2 intents

the Alexa requests. The first step in the flow carried out identification of the intent present in the request. Requests for 'GetTemperature' were directed through the upper pathway and intents of 'GetLaserPower' were directed through the lower pathway. Each pathway carried out retrieval of the relevant data and returned a response to Alexa. If the response could not be generated then an error message would be returned.

Following input from the lab users this skill was modified to calculate a power ratio from the obtained readings and report this alongside the laser power as this is a useful diagnostic for the laser.

Alerts (Slack output): As discussed in the use cases, having a system which continually monitors equipment in the lab gives potential to alert the users when an unexpected situation is encountered. To test the generation of alerts a simple alert system was created to trigger a message when a sensor reading was outside of a specification.

The code for testing the specification and generation of alerts was placed into the flows which monitors the sensor feeds so it did not rely on a user making a request to the system, but would be continuously monitoring the readings. If a reading received via MQTT was out of specification a message was generated and sent out to the users using integration between Node-RED and the Slack messaging platform used by the physics group. The Slack output uses a Slack bot to send a message in a specified channel.

An example alert generated was: *"Alert! The temperature of tsunami is out of spec at 28.7 degrees!"*. The alert informs the user of the piece of equipment that was out of specification and what its reading was. This provides the user with enough information to tell how urgent the alert is and the location to look to fix it.

In this implementation the specification for the sensors had to be hard-coded into the Node-RED flow; however, the incorporation of a catalogue for the items in the lab would allow specifications for each sensor to be recorded alongside the device information. The Node-RED flow could then interface with the catalogue, retrieving specifications when required. Additional specifications could be defined allowing sensors multiple levels of specification, such as inner and outer thresholds. It would also be possible to create an alert for specific users in the Slack channel by tagging users that are recorded in the catalogue as being responsible for that piece of equipment.

The alert system should be further developed so that alerts are not repeated continuously every time a new reading is received if they are still out of specification. The temperature readings are received once a minute, which would cause frustration if the alerts were not urgent but sent every minute until fixed. Alerts should be repeated only if the reading goes back into specification and then out again, or if the reading is still out of specification after a specified time period. Different time periods could be set dependent on how important it is that the reading stays inside the specification. A temperature

value being a little high would be significantly less urgent than an alert which says that the interlock has been tripped on the laser whilst it is on.

Creating a more complex alert system would require additional variables to be stored along side the temperature data in the Node-RED system. These variables would be used to flag the time that the reading went out of specification or the time the last alert was sent. This would be used in conjunction with the specification to determine if an alert is required each time the reading comes in. Once the reading goes back in specification the flag would be cleared. The alerts could also be escalated to other methods such as email, or SMS if a situation is not resolved within a given time frame.

The integration with Slack is useful for alerts; however, it could also be used as the output for other functions such as monitoring or status reports. It could also be integrated as part of a text interaction as an alternative to the Echo Dot voice interaction.

Slack input: Slack input was also examined to explore the possibility of creating a text interaction method, where interaction with the system was carried out entirely through Slack. The request would be input in Slack and following processing response would be returned to Slack.

Slack input was made possible through use of a Slack bot to monitor a specific Slack channel for posts, any messages on the channel are used as inputs to the Node-RED system. Capturing the Slack input was a straightforward process; however, the integration of Slack requests with the flows that handle request processing was more complicated.

Unlike the Alexa requests from the Echo Dot, the Slack messages entered into the system in their raw text form without any language processing. This required the identification and extraction of the request parameters before the request can be processed. If the request is posed as a natural question, such as *"What is the temperature of tsunami?"*, then the text requires natural language understanding (NLU) to be applied to identify the intent and parameters of the request.

NLP/NLU engines do exist that can be integrated with the Node-RED system; however, the use of most cloud engines (Amazon Alexa/Lex, IBM Watson) required a subscription to their service. This could be a viable option for a future development; however, this would be dependent on the scope of the project and the cost incurred. Open source engines would be an alternative to paid cloud engines, although these often require more development and customisation.

An alternative method to using natural language commands would be to require a specific structure for requests. If the request could be structured in the format 'Intent,parameter1, parameter2 (if required)' e.g. 'Temperature,tsunami' then the complexity of the processing would be reduced. This method of interaction was created and

tested, successfully allowing the retrieval of simple requests which mirrored the Alexa system; however, the strict syntax required was not very straightforward for making requests. It would require significant expansion of error handling to cope with variations and potential syntax errors in the user input to create a usable interface through Slack.

Following expansion of the system to include the sensor data from all sensors and numerous intents covering the wide range of queries, the Slack interaction method would need to be integrated with the main Alexa processing flows to avoid unnecessary replication. Implementation of an NLU engine would allow the requests to be structured in a similar format to the Echo requests, parameters extracted from a structured request could be converted to a similar JSON format or input at a later node in the flow. The system would also need to track the interaction method (Slack, Echo Dot, etc) for each request so the response can be returned to the correct end point.

The interaction with Slack is outlined in the 'System Interaction' section.

Vacuum pressures: Through a number of existing sensors temperature values were fully integrated into the Node-RED system; however, during an experiment these are not always the most useful readings to access. For users running experiments it was more important to be able to access other readings, for example the values for the vacuum pressures.

The beam lines are placed under vacuum when carrying out experiments. The vacuum is controlled by vacuum pumps and monitored by multiple vacuum gauges in the lab. The gauges can take input from up to three vacuum sensors and display the readings on the front of the unit.

These gauges were not easily compatible with the sensor system as they were not equipped with modern connectivity ports, some gauges had RS232 connections and others had analog pins. It was possible after some experimentation to connect the analog output of the vacuum gauge to a Raspberry Pi to receive data from the gauge.

Once the data were received by the Raspberry Pi they were sent via MQTT and could be incorporated into the Node-RED system. However, the presence of multiple different models of vacuum gauge may require the creation of multiple different procedures to handle the inputs from the different gauges.

The age of the vacuum sensors meant they were not optimised for an IoT system and so cannot identify themselves when plugged in or provide an ID along with their readings. This removed the ability to request a reading from a specific vacuum sensor, instead the requests must be directed to a vacuum gauge which would return the pressure readings for all sensors attached to it.

7.5 System Interaction

In the creation of this connected lab the aim was to produce a system which enhanced the users interaction with the lab. The methods through which the user interfaces with the lab should not interfere with the users work but operate alongside to improve the lab experience.

The primary interaction within this prototype system was voice interaction through the Amazon Alexa devices. Full functionality of the system was initially limited to requests made via the Echo dot. However, text interaction via the Slack interface was implemented which mirrored the functionality in the Echo voice interface. The interaction through Slack was not as user-friendly as the Echo dot but it can be used in different situations.

7.5.1 Voice interaction - Echo Dot

In the lab environment where the level of background noise was quite high the Echo dot performed well at close range, and understood the majority of commands at distances up to 1.5m. This was in areas of high noise, in areas of lower noise such as the mezzanine write-up area and entry rooms the distance was increased. It was possible to operate multiple Echo devices together to cover a larger area, as they have Echo spatial perception to determine location, allowing the closest device to handle the request and response.

It was not possible to interact with an Echo dot through a Bluetooth microphone, which would have overcome most problems with background noise. There is an Alexa voice remote which contains a microphone, this could be utilised if multiple Echo devices could not cover the whole lab adequately.

Interaction through Alexa is triggered by a spoken request from the user. This takes the form of request qualifiers followed by a question in natural language. Figure 7.18 shows an example of the structure of an Alexa request. The request must begin with the wake-word for the device and contain the name of the skill followed by a question or command containing words that are linked to the relevant intent and any parameters that are required for the intent. Devices are limited in their possible wake words, the options are ‘Echo’, ‘Alexa’, ‘Computer’ or ‘Amazon’.

Intents were available for the request of temperatures and laser power with sample utterances defined for each skill. These are phrases which will be identified as that intent. Examples for temperature were: *‘What is the temperature of...?’*, *‘How hot is..?’* When the request matches (or is similar) to the sample utterances then it is passed to

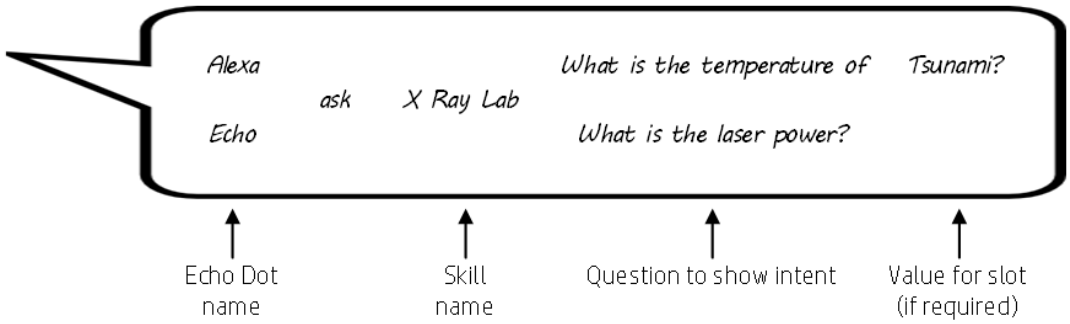


Figure 7.18: Structure of example requests made to Alexa Device

the specified intent. However, the 'similarity' of the matches can be somewhat hit and miss and it is better to directly match an utterance where possible.

Although the use of NLP allows some flexibility in the structure of the question in the request, the user must follow the overall structure of *'wake word, skill, question'*, as well as learning the different intents that are available. Like many new systems that are implemented this may require some training and experimentation to ensure the users become comfortable with the interaction. If necessary additional sample utterances could be added to the system if users encounter difficulties.

To aid the training of users more information about the system was incorporated into the skills in the form of 'help files'. These were designed to inform the user of the options available to them through Alexa. This could be triggered through a launch request to the skill which informed users of the questions they could ask. Alternatively, use of the help intent would give users more information about a specific intent including an example structure of the request and the sensors that were available.

With the responses voiced by the Echo dot it was important to try and convey the information in the most concise form whilst still maintaining a natural speech pattern. The shortest possible response for a temperature request would be: *'Temperature tsunami 20.1'* but this is abrupt and lacks any form of speech pattern. Long responses on the other hand became quite monotonous unless the response was properly defined using speech synthesis markup language (SSML)

7.5.2 Text interaction - Slack

Although voice interaction is useful in many situations it is not always the most suitable interface and other interfaces, such as text or visual displays with click/touch may be better suited. Text interaction would be useful in situations where audible interaction is not desired, such as in a quiet shared office, in a noisy area which disrupts the quality of the system, or where a user has hearing difficulties.

The interaction created with Slack was mostly used for output from the Node-RED system, handling the generation of alerts for sensor measurements. These were generated automatically when a sensor measurement was received which was outside the specification. Figure 7.19 shows an example of an alert generated in Slack when a temperature sensor was outside of specification.

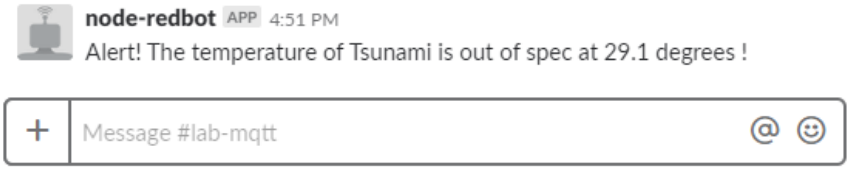


Figure 7.19: Example of a temperature alert in Slack

Request and response interaction through Slack was also developed; however, the syntax required by the input meant that it not mirror a natural way of interacting. Requests could be made for both temperatures and laser powers. Figure 7.20 shows 3 separate requests for temperature made via Slack. When the request was correctly formatted the system generated the correct response. The first two interactions show requests for the temperature of *'tsunami'* and *'BL1'* respectively, with the third request having the sensor misspelt as *'tsuanami'*.

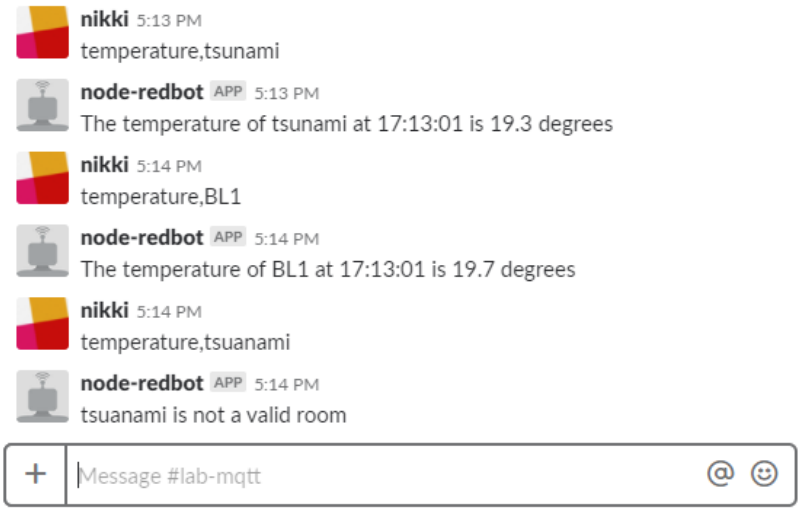


Figure 7.20: Interaction in Slack for temperature requests

It is not unreasonable that spelling errors would cause the system to generate an error; however, errors were also generated if the wrong delimiter or spacing was present in the request. The error handling could be expanded for this text interaction to allow slight variation in the format of the request. If full sentence requests were used in Slack then a NLP engine would be required; however, short text requests may be more beneficial than full sentences for simple requests as typing full sentences takes much more time.

The benefit of using a text interaction was that it did not take time for responses to read out, as such longer help messages could be generated, including full examples of the syntax required and the options available to the user. These were incorporated so they triggered when no intent was recognised in the request.

7.5.3 Future interaction

With the swift development of 'smart home assistants' new products are continuously being released in this area. Developments are also being made incrementally to the Alexa system with new features being released regularly. Voice interaction could potentially be improved if each user was assigned their own wireless device or headset. This would enable each user to be identified when linking in with the lab system and the system could be tailored to their needs. In addition this would also combat many of the issues associated with background noise as the user would never be too far away from the microphone.

The implementation of digital dashboards to display summary information from the sensors would bring an additional interaction method to the system as these are visual displays which would show visualisations obtained from the sensor data. Depending on the type of display the user would interact with the dashboard directly through touch or via mouse interactions. The dashboards would pull the data from the Node-RED system and use d3.js to create dynamic visualisations that can be easily manipulated by the user. Prototypes of these are currently in the early stages of development.

An extension of the voice interaction method would be triggering actions through the use of commands. This would be implemented initially for the display of network enabled cameras on screens in the lab. Beyond this it could be extended to enable capture of snapshots from the cameras and the recording of data from sensors/equipment for a lab notebook. Moving from retrieval of information to triggering events and recording information is a large step towards the implementation of a completely connected lab.

7.6 Security & System Design

7.6.1 System Security

System security is an area of the design that should not be overlooked. In the current system design none of the data available were safety critical or sensitive data. Although a data breach is never desirable, if this data were to be accessed it would not be a significant issue. However, if the system were expanded to contain controls for the lab

then a security survey should be undertaken. Additional security measures would then be implemented to ensure that no unauthorised access could occur.

Node-RED system
The sensor data and Node-RED system were both installed within the SOTON domain network. This meant that the data was behind the university firewall and could only be accessed by devices directly on the network or connected to the VPN. The server itself was password protected so changes could only be made by authorised users.

Echo Dot & Alexa
The Echo Dot is one area of the system that could be vulnerable to hacking; however, currently the only report of hacking specific to the Echo devices requires physical access to the device, and only impacts devices produced in 2015/16 as newer devices have had the physical vulnerability removed. [238]

Currently the main other areas of threats include malicious use by other people and triggering commands through playback of voice recordings. [239] The current design of this system means that it is not very susceptible to malicious use by other people. Although it may be possible for an unauthorised user to physically access the system they would be unable trigger any malicious actions. The Alexa interface is not able to control any equipment nor can it carry out any unauthorised purchases, which is a frequent misuse of Alexa devices, as the account settings do not allow this.

To ensure the security of the Alexa system and minimise risk of hacking there are also a number of additional security measures:

- Encrypted communication between the Echo dot and Amazon servers.
- Skill publication requires HTTPS security certificates and validation of signatures and timestamps of Alexa requests.
- Security patches and updates are regularly pushed out to connected Echo devices.

7.6.2 Privacy

Privacy within this system had two areas of focus, these were personal privacy for the users and data privacy for the data contained within the system.

Personal Privacy: The Amazon Alexa devices are designed to always be listening for their wake-word which triggers activation. Although the manufacturers maintain that the devices only stream the audio following the wake-word, the always listening aspect is a concern to many users as they feel it is an invasion of their privacy. The apprehension of users may be reduced over time as voice interfaces become more widespread in the

home and workplace. If the always listening is a concern to users this could be overcome by activating the mute buttons on top of the device when the user is not interacting with Alexa. Another solution would be to develop push to talk devices which would only record when a button is pressed. But both of these solutions reduce the benefits of having handsfree interaction.

In addition to the always listening behaviour of the devices, the history of voice commands is stored with the Amazon account to allow the system to learn and improve the accuracy of responses. This data cannot be accessed directly from the Echo device, but can be accessed through the Alexa app or via the web interface which require knowledge of the login credentials. If this was highlighted as a concern to the users then policies could be introduced which require frequent deletion of this data, although this may reduce the accuracy of the request identification.

Data Privacy: In the current Echo dot integration requests and responses are passed through servers which are hosted in the cloud by IBM and Amazon. In an environment which contains sensitive data in the responses this would not be a suitable set-up.

Whilst processing can be carried out on an internal server, responses which are spoken via an Alexa device must pass through the Alexa servers. If that response contained data which must remain private then the use of an Alexa system would not be possible. Alternative solutions could be investigated using open source NLU engines that could be self-hosted. However, it is unlikely that these would have the same power as the cloud based services.

For the purposes of this installation the data does not need to be kept private. In fact most of the data was already made publicly accessible though the web interface where data values can be seen by anyone.

7.6.3 Data Validity

The system was designed to carry out the processing for the request out of sight of the user and simply return the response to the query. However, the user needs to be confident that when the system returns a response it is returning the correct data. This can be achieved through a combination of quality controls.

- In the initial system design and set-up, care must be taken to ensure the sensors are configured correctly and assigned the correct metadata. This makes certain that when a sensor value is queried it is being retrieved from the correct sensor.
- When sensor data is published often a timestamp is included with the measurement. This timestamp was stored with the sensor reading and utilised in the

generated response. In this prototype the timestamp was included in the response to provide the user with the time of the most recent temperature reading. This is shown in the example in Figure 7.20.

An alternative method could use the timestamp as a validity test where the system checks if the reading was provided within a specific timeframe, e.g. temperatures must have been taken in the last 5 minutes. If the reading was not recent enough then it indicates a problem with the sensor or connection and the system would return an error message such as *'The measurement for **sensor** is not up to date, please check the connection'*. The specification for the timeframe could adjusted dependent on the frequency of measurements for each type of sensor.

Where sensors do not provide timestamps the system could add a timestamp at the point when the reading is received into the system. However, care must be taken to ensure that this reading was an up-to-date reading. Distinction could be signified by the use of *'The measurement of **reading** was **value**, received at **time**'*.

Timestamps could also be used in conjunction with the Last Will and Testament (LWT) functionality of MQTT to notify the user if a sensor has been disconnected or has its status set to offline.

7.6.4 Error Handling

When interacting with a system request cannot always be expected to follow a perfect scenario. System design requires error handling to ensure that any issues are caught in processing and useful error messages are returned to the user. Throughout the processing of a request in the Talk2Lab system there were a number of points where errors could be generated.

Within the Alexa system the errors were handled through Amazon's system. If Alexa did not understand the request at the first stage of processing or if the skill was identified but no intent had been understood then it returned a rather generic error message of *"I'm not sure"*. If Alexa identified both the skill and intent then the request was passed to the endpoint (Node-RED server) for processing. Other error messages could be generated by a lack of internet connection, or a lack of response from the requested skill.

In Node-RED there were also multiple points where error handling was incorporated and errors could be returned to the user. The error messages were designed to be helpful to the user rather than just returning 'error'. As some error handling was carried out on the Alexa system the error handling in Node-RED related to each skill seperately. In the GetTemperature intent a slot was required to identify the sensor, if no slot had been identified by Alexa then the following message was returned by Node-RED; *"Sorry I did*

not detect a sensor name". If a slot value was identified; however, the 'room' identified was not a valid sensor then Node-RED returned a message of *"**roomvalue** is not a valid room"*. With the Laser Power intent no slots were required. The Node-RED system checked if the laser power monitor was active, if it received an error then the following error message was returned, *"The Beaglebone is unavailable, please reboot."*

Similar error handling was also employed in the Slack text interaction; however, it also required error handling for the identification of the intent. As the responses were in text form a longer error message could be generated which informed the user of the correct way to format their request and the options available for the 'slots'. Each intent also had the same error handling as in the Alexa system. As more intents are added to the system extra error handling methods would need to be incorporated into the system.

7.6.5 Limitations of Current System

Although the proof of concept system has been developed successfully with the combination of Alexa and Node-RED this system has a number of limitations. Some of these limitations may be overcome by future updates to Alexa or tweaks to the systems but others may have to be accepted as a compromise for the other powerful functionality of the Alexa system.

7.6.5.1 Alexa Limitations

When carrying out interactions with the Alexa system each command must be made separately and prefixed by the skill name. The system cannot handle a query which contains multiple requests or multiple values, e.g. *'Ask X Ray Lab what is the temperature of tsunami and what is the Laser power?'* This is a restriction that users must learn to live with.

During the development of the Talk2Lab system the Alexa system was restricted to single stage questions which meant that it could not handle context and multi-stage questions. If a user asked *'What is the weather in Southampton?'* and following the response asked *'What is the time there?'* it would not understand what 'there' meant. Functionality which handles multistage dialog has since been added to the AVS processing, so in future development multi-step questions could be developed where the system retains knowledge of context to create additional questions and responses. This also allows for extra error handling where the system has not fully understood the command or to get confirmation of an action before it is carried out.

Initial set-up and expansion of the system required manual configuration for the intents on the Alexa skill developer site. For a system that is not very large or does not change

frequently this would largely be a one off time requirement. If the system required frequent changes then it would be beneficial to investigate alternative processing solutions which can be populated in bulk.

Amazon have recently developed new tools for managing the skills kit through an API rather than via the web [240], which would simplify the population of a skill for slots with a large number of possible values. This could also open up the possibility of populating the skill from the database of equipment created during this project.

7.6.5.2 Other Limitations

The system requires continual internet access as all of the processing is carried out in the cloud. This should not be a significant issue for this installation as the network connection within the university is reliable and should not frequently be interrupted. If a connected lab system was implemented in a network where the internet connection was intermittent then an alternative processing system may be required which is locally hosted.

Although Node-RED was very useful for implementing the initial prototype and outlining the flow of requests through the system, the introduction of additional intents and error handling creates a significant number of nodes. The increasing number of nodes becomes difficult to manipulate in the visual display of Node-RED. Further development may be more manageable in a node.js server instead.

The current implementation handled all requests as though they originated from a single user. For current functionality this was sufficient as there were no user specific commands or variables. However, development of commands that allow users to store data for their lab notebooks or commands that have different set-ups dependent on user may benefit from the introduction of voice recognition that identifies a user. This functionality currently exists, although it is not yet fully reliable.

7.7 Discussion & Future Work

Development in the area of IoT devices and Home-automation is very swift; the technology utilised in this system was released in the UK less than two years ago, but since then a whole range of products and services have been developed surrounding it. Although voice interaction is spreading through the sphere of home-automation little development had been done in lab environments.

This connected lab project has successfully developed a working prototype system centred around Amazon Alexa and the Echo Dot. This allowed users to interact with

their lab environment via voice and retrieve real-time information from sensors and equipment, a design that had not been previously reported. This work lays the foundations for expansion into other lab environments with more opportunities for sensor installation and monitoring. The system was implemented using a combination of Node-RED and MQTT to access and process sensor data. Sensors were installed in the lab to give access to temperature, power consumption, vacuum pressures, laser power and visual feeds.

The Amazon Alexa system was integrated through the Echo Dot to give voice interaction with the sensor system, where users could request temperature and laser power readings. Text interaction was also created through Slack both for request and response as well as alerts when readings were out of specification. The interaction through slack was very rigid in its structure and not as user-friendly as Alexa. The interaction pathways through Slack could be expanded to give more flexibility in the input structure. Additionally they should link into the Node-RED flows which process the Alexa requests as this would allow expansion of both interaction methods without duplication.

A full survey of the lab was undertaken to create a catalogue of the items present in the lab, these included equipment, sensors and locations. Metadata for each item was assembled in a database for use in integration with new intents and functionality in the connected lab system.

Running a series of Talk2Lab workshops and hackathon sessions provided a great deal of expertise from a wide variety of backgrounds, providing a number of suggestions for functionality in the connected lab. Sessions with the lab users enabled the creation of a number of 'use cases', identifying scenarios in which the implementation of the connected lab technology would be of most benefit. These use cases will influence the direction of future development of this system.

7.7.1 Areas for Development

As this project focused on creating a proof-of-concept connected lab system a number of the ideas raised throughout the development have not been fully implemented in the system. This gives a range of areas in which expansion and development could be carried out to expand the scope of the Talk2Lab system.

- Fully integrate the additional sensor measurements into the Alexa skill and create the corresponding intent pathways in Node-RED
- Full implementation of database containing all of the lab equipment (HyperCat catalogue)
- Incorporate additional interaction methods: dashboards for displaying data, camera feed display and capture

- Apply NLU to Slack input allowing natural questions to be asked through Slack. This could be run through Alexa or an alternative NLU engine.
- Expand the error handling methods to cover more and create a more robust system.
- Link the Slack alert systems in with the equipment database to pull specification limits through.

7.7.1.1 Further Expansion

Beyond the areas mentioned for development, this system could be expanded to include a wider range of integration. Here are a number of areas which could be investigated, many of these ideas come from the use cases and workshops that were run throughout the project.

- Headsets for interaction with Alexa
- Creation of more complex questions which combine sensor measurements to provide more insight in the lab.
- Dashboards showing live-feed information about the lab systems, such as temperatures, power consumption, and laser information
- Integration with lab inventory software e.g. Alexa where is sodium chloride?
- Pushing data outside of the University environment - either for dissemination of data or potential interaction with the public - public access science.
- Image recognition for analysis of legacy equipment
- Integrate with Lab Notebook system to allow information to be recorded
- Link up with static systems to look-up information e.g. safety data and protocols

Bibliography

[1] P. J. Goodford. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.*, 28(7):849–857, 1985.

[2] ACD/I-Lab. Version 2.0, Advanced Chemistry Development, Inc., Toronto, ON, Canada, available at: https://ilab.acdlabs.com/iLab2/, [accessed: 02/03/2017].

[3] Virtual Computational Chemistry Laboratory. ALOGPS2.1. available at: http://www.vcclab.org/lab/alogps/, [accessed: 11/12/2017].

[4] CambridgeSoft. Chemdraw Ultra. Version 12.0, 2012.

[5] ChemAxon. Chemicalize.org. available at: www.chemicalize.org, [accessed: 02/11/2017].

[6] Royal Society of Chemistry. Chemspider. available at: www.chemspider.com, [accessed: 02/08/2015].

[7] Daylight Chemical Information Systems, Inc. available from: http://daylight.com, [accessed:12/05/2014].

[8] Talete.srl. DRAGON 6.0. available from: http://www.talete.mi.it/products /dragon_description.htm, [accessed: 01/02/2018].

[9] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, and V. V. Prokopenko. Virtual Computational Chemistry Laboratory Design and Description. *J. Comput. Aided. Mol. Des.*, 19(6):453–463, 2005.

[10] ChemAxon. Marvin Sketch 6.1.3. available from: https://chemaxon.com /products/marvin, [accessed: 21/02/2015].

[11] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An Open chemical toolbox. *J. Cheminform.*, 3(1):33, 2011.

[12] Wavefunction Inc. Spartan '08. depreciated software, www.wavefun.com, [accessed: 16/03/2015].

[13] Cresset. Torchlite. Version 10, http://www.cresset-group.com/products/torch/, [accessed: 24/07/2014].

[14] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. Wong, F. Paesani, J. Vanicek, R. Wolf, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G.Cui, D. Roe, D. Mathews, M. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman. AMBER 12. http://www.ambermd.org, [accessed: 13/02/2015], 2012.

[15] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, X. Caricato, M.; Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, J. A., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09. Gaussian, Inc., Wallingford CT, 2009.

[16] MATLAB R2013a. The MathWorks, Inc., Natick, Massachusetts, 2013.

[17] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–9, dec 2011.

[18] JS Foundation. Node-RED. available from: https://nodered.org, [accessed: 03/03/2018].

[19] Node.js Foundation. Node.js. 2015, v0.12.0, available from: https://nodejs.org/, [accessed: 14/05/2015].

[20] JMP Version 11.0.0. SAS Institute Inc., Cary, NC, USA, 2013.

[21] JMP Version 9.0.0. SAS Institute Inc., Cary, NC, USA, 2011.

[22] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria., 2008.

[23] M. Borkum. Blog3 Demo Site. available at: http://blog3-demo.mylabnotebook.ac.uk/, [accessed: 07/02/2018].

[24] University of Southampton. Labtrove. available at: www.labtrove.org, [accessed: 03/02/2015].

[25] Microsoft Corporation. Microsoft OneNote 2010. https://www.onenote.com, [accessed: 06/01/2018].

[26] RStudio Team. RNotebooks. available at: https://rmarkdown.rstudio.com/r_notebooks.html, [accessed: 04/03/2017].

[27] Wishart Lab. ClassyFire: A Comprehensive, Computable Chemical Taxonomy. available at: http://classyfire.wishartlab.com/, [accessed: 15/04/2016].

[28] Chemical Abstracts Service. SciFinder. American Chemical Society, available at: https://scifinder.cas.org/, [accessed: 07/08/2014].

[29] Slack Technologies. Slack. available from: https://slack.com, [accessed: 18/11/2017].

[30] N. J. Knight. Electronic Supplementary Information for thesis: The Connected Lab: Digital Synergies from Data to Models. available at: http://dx.doi.org/10.5258/SOTON/D0563, 2018.

[31] A. Leach and V. Gillet. *An Introduction to Chemoinformatics*. Springer, Dordrecht, The Netherlands, revised edition, 2007.

[32] M. Abhilash. Quantitative structure activity relationship (QSAR). *Int. J. Pharma Bio Sci.*, 1(1), 2010.

[33] G. M. Maggiora. On outliers and activity cliffs - Why QSAR often disappoints. *J. Chem. Inf. Model.*, 46(4):1535, 2006.

[34] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194:178–180, 1962.

[35] C. Hansch. A Quantitative Approach to Biochemical Structure-Activity Relationships. *Chem. Res.*, 2(8):232–239, 1969.

[36] R. Collander. The Partition of Organic Compounds Between Higher Alcohols and Water. *Acta Chem. Scand.*, 5:774 – 780, 1951.

[37] C. A. M. Hogben, D. J. Tocco, B. B. Brodie, and S. Schanker. On the mechanism of intestinal absorption of drugs. *J Pharmacol Exp Ther*, 125:275–282, 1959.

[38] L. F. Fieser, M. G. Ettlinger, and G. Fawaz. Naphthoquinone Antimalarials. XV. Distribution between Organic Solvents and Aqueous Buffers. *J. Am. Chem. Soc.*, 70(10):3228–3232, 1948.

[39] R. Collander. The permeability of plant protoplasts to non-electrolytes. *Trans. Faraday Soc.*, 33:985–990, 1937.

[40] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.*, jan 2014.

[41] J. Gasteiger. Chemoinformatics: Achievements and challenges, a personal view. *Molecules*, 21(2):151, 2016.

[42] M. Leist, B. A. Lidbury, C. Yang, P. J. Hayden, J. M. Kelm, S. Ringeissen, A. Detroyer, J. R. Meunier, J. F. Rathman, G. R. Jackson, G. Stolper, and N. Hasiwa. Novel technologies and an overall strategy to allow hazard assessment and risk prediction of chemicals, cosmetics, and drugs with animal-free methods. *ALTEX*, 29(4):373–388, 2012.

[43] J. Xu and A. Hagler. Chemoinformatics and drug discovery. *Molecules*, 7(8):566–600, 2002.

[44] C. M. Dobson. Chemical space and biology. *Nature*, 432(7019):824–8, dec 2004.

[45] R. Perkins, H. Fang, W. Tong, and W. J. Welsh. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.*, 22(8):1666–1679, 2003.

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

[47] K. Roy, S. Kar, and R. N. Das. Statistical Methods in QSAR/QSPR. In *A Prim. QSAR/QSPR Model.* Springer, 2015.

[48] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese, and R. K. Agrawal. Validation of QSAR Models - Strategies and Importance. *Int. J. Drug Des. Discovery*, 2(3):511–519, 2011.

[49] A. Leach and V. Gillet. *Introduction to Chemoinformatics.* Springer, 2007.

[50] G. James, D. Witen, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[51] D. C. Montgomery, E. A. Peck, and G. G. Vining. Multiple Linear Regression. In *Introd. to Linear Regres. Anal.* John Wiley & Sons, Inc., 5th edition, 2012.

[52] M. Dehmer, K. Varmuza, and D. Bonchev, editors. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR.* Wiley-VCH, Weinheim, Germany, 2012.

[53] Minitab Blog. How do I interpret R squared and assess goodness of fit. available at: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit, [accessed: 19/04/2015].

[54] S. Maitra and J. Yan. Principle Component Analysis and Partial Least Squares : Two Dimension Reduction Techniques for Regression. 2008 Discussion Paper Program, available from: https://www.casact.org/pubs/dpp/dpp08/, [accessed: 19/07/2015].

[55] B. H. Mevik and R. Wehrens. The pls Package: Principle Component and Partial Least Squares Regression in R. *J. Stat. Softw.*, 18(2):1–24, 2007.

[56] Statistica Help. Principal Component Analysis (PCA) and Partial Least Squares (PLS) Technical Notes. available from: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=mspc/PCAandPLSTechnicalDetails, [accessed: 02/03/2015].

[57] D. C. Montgomery, E. A. Peck, and G. G. Vining. Introduction to nonlinear regression. In *Introd. to Linear Regres. Anal.* John Wiley & Sons, Inc., 5th edition, 2012.

[58] V. Consonni, D. Ballabio, and R. Todeschini. Comments on the definition of the Q2 parameter for QSAR validation. *J. Chem. Inf. Model.*, 49:1669–1678, 2009.

[59] A. Golbraikh and A. Tropsha. Beware of q2! *J. Mol. Graph. Model.*, 20:269–276, 2002.

[60] P. Gramatica. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.*, 26(5):694–701, 2007.

[61] R. Wehrens, H. Putter, and L. M. Buydens. The bootstrap: A tutorial. *Chemom. Intell. Lab. Syst.*, 54(1):35–52, 2000.

[62] P. Gramatica. External evaluation of QSAR models, in addition to cross-validation: Verification of predictive capability on totally new chemicals. *Mol. Inform.*, 33:311–314, 2014.

[63] G. Schuurmann, R. U. Ebert, J. Chen, B. Wang, and R. Kuhne. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient's - Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.*, 48:2140 – 2145, 2008.

[64] V. Consonni, D. Ballabio, and R. Todeschini. Evaluation of model predictive ability by external validation techniques. *J. Chemom.*, 24(3-4):194–201, 2010.

[65] Guidance document on the validation of (quantitative) structure-activity relationship [QSAR] models. Environment Health and Safety Publications Series on Testing and Assessment. Organisation for Economic Co-operation and Development, 2007, available from: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2, [accessed: 24/05/2016].

[66] D. M. Hawkins, S. C. Basak, and D. Mills. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, 43(2):579–586, 2003.

[67] D. M. Hawkins. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.*, 44(1):1–12, 2004.

[68] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2000.

[69] M. Randic. Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.*, 31:311–320, 1991.

[70] A. R. Leach. The Use of Molecular modelling and Chemoinformatics to Discover and Design New Molecules. In *Mol. Model. Princ. Appl.* Pearson Education Limited, 2nd edition, 2001.

[71] R. Todeschini and V. Consonni. *Molecular Descriptors for Chemoinformatics*. John Wiley & Sons, Inc., 2 edition, 2009.

[72] W. A. Warr. Representation of chemical structures. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(4):557–579, 2011.

[73] H. A. Favre and W. H. Powell. *Nomenclature of Organic Chemistry. IUPAC Recommendations and Preferred Name 2013 (Blue Book)*. The Royal Society of Chemistry, Cambridge, UK, 2013.

[74] N. M. O'Boyle. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.*, 4(9):1, 2012.

[75] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[76] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.*, 7(1):1–34, 2015.

[77] I. V. Tetko, V. Y. Tanchuk, and A. E. P. Villa. Prediction of n -Octanol / Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.*, 41:1407–1421, 2001.

[78] F. M. Ashcroft. *Ion Channels and Disease*. Academic Press, San Diego, 1999.

[79] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman and Company, New York, 6th edition, 2013.

[80] D. J. Voet, J. G. Voet, and C. W. Pratt. *Principles of Biochemistry*. John Wiley & Sons, Inc., 3rd edition, 2008.

[81] S. K. Ko, S. K. Kim, A. Share, V. M. Lynch, J. Park, W. Namkung, W. Van Rossom, N. Busschaert, P. A. Gale, J. L. Sessler, and I. Shin. Synthetic ion transporters can induce apoptosis by facilitating chloride anion transport into cells. *Nat. Chem.*, 6(10):885–892, 2014.

[82] A. Kondratskyi, K. Kondratska, R. Skryma, and N. Prevarskaya. Ion channels in the regulation of apoptosis. *Biochim. Biophys. Acta - Biomembr.*, 1848(10):2532–2546, 2015.

[83] B. Nilius and G. Droogmans. Amazing chloride channels: An overview. *Acta Physiol. Scand.*, 177(2):119–147, 2003.

[84] F. M. Ashcroft. From molecule to malady. *Nature*, 440(7083):440–447, 2006.

[85] N. Busschaert and P. A. Gale. Small-molecule lipid-bilayer anion transporters for biological applications. *Angew. Chem. Int. Ed. Engl.*, 52(5):1374–1382, 2013.

[86] A. Fürstner. Chemistry and biology of roseophilin and the prodigiosin alkaloids: A survey of the last 2500 years. *Angew. Chemie - Int. Ed.*, 42(31):3582–3603, 2003.

[87] W. A. Harrell, Jr., M. L. Bergmeyer, P. Y. Zavalij, and J. T. Davis. Ceramide-mediated transport of chloride and bicarbonate across phospholipid membranes. *Chem. Commun.*, 46(22):3950–3952, 2010.

[88] S. Bahmanjah, N. Zhang, and J. T. Davis. Monoacylglycerols as transmembrane Cl- anion transporters. *Chem. Commun.*, 48(37):4432–4434, 2012.

[89] S. Matile, A. Vargas Jentzsch, J. Montenegro, and A. Fin. Recent synthetic transport systems. *Chem. Soc. Rev.*, 40(5):2453–2474, 2011.

[90] P. A. Gale. From anion receptors to transporters. *Acc. Chem. Res.*, 44(3):216–26, 2011.

[91] N. Busschaert, S. J. Bradberry, M. Wenzel, C. J. E. Haynes, J. R. Hiscock, I. L. Kirby, L. E. Karagiannidis, S. J. Moore, N. J. Wells, J. Herniman, G. J. Langley, P. N. Horton, M. E. Light, I. Marques, P. J. Costa, V. Félix, J. G. Frey, and P. A. Gale. Towards predictable transmembrane transport: QSAR analysis of anion binding and transport. *Chem. Sci.*, 4(8):3036, 2013.

[92] N. Busschaert, M. Wenzel, M. E. Light, P. Iglesias-Hernández, R. Pérez-Tomás, and P. A. Gale. Structure-activity relationships in tripodal transmembrane anion transporters: the effect of fluorination. *J. Am. Chem. Soc.*, 133(35):14136–48, 2011.

[93] C. J. E. Haynes, N. Busschaert, I. L. Kirby, J. Herniman, M. E. Light, N. J. Wells, I. Marques, V. Félix, and P. A. Gale. Acylthioureas as anion transporters: the effect of intramolecular hydrogen bonding. *Org. Biomol. Chem.*, 12(1):62–72, 2014.

[94] C. J. E. Haynes, S. J. Moore, J. R. Hiscock, I. Marques, P. J. Costa, V. Félix, and P. A. Gale. Tunable transmembrane chloride transport by bis-indolylureas. *Chem. Sci.*, 3(5):1436, 2012.

[95] S. J. Moore, M. Wenzel, M. E. Light, R. Morley, S. J. Bradberry, P. Gómez-Iglesias, V. Soto-Cerrato, R. Pérez-Tomás, and P. A. Gale. Towards drug-like indole-based transmembrane anion transporters. *Chem. Sci.*, 3(8):2501, 2012.

[96] N. J. Andrews, C. J. E. Haynes, M. E. Light, S. J. Moore, C. C. Tong, J. T. Davis, W. A. Harrell Jr., and P. A. Gale. Structurally simple lipid bilayer transport agents for chloride and bicarbonate. *Chem. Sci.*, 2(2):256, 2011.

[97] N. Busschaert, I. L. Kirby, S. Young, S. J. Coles, P. N. Horton, M. E. Light, and P. A. Gale. Squaramides as potent transmembrane anion transporters. *Angew. Chem. Int. Ed. Engl.*, 51(18):4426–30, 2012.

[98] P. B. Cranwell, J. R. Hiscock, C. J. E. Haynes, M. E. Light, N. J. Wells, and P. A. Gale. Anion recognition and transport properties of sulfamide-, phosphoric triamide- and thiophosphoric triamide-based receptors. *Chem. Commun. (Camb).*, 49(9):874–6, 2013.

[99] L. E. Karagiannidis, J. R. Hiscock, and P. A. Gale. The influence of stereochemistry on anion binding and transport. *Supramol. Chem.*, 25(9-11):626–630, 2013.

[100] S. J. Moore, C. J. E. Haynes, J. González, J. L. Sutton, S. J. Brooks, M. E. Light, J. Herniman, G. J. Langley, V. Soto-Cerrato, R. Pérez-Tomás, I. Marques, P. J. Costa, V. Félix, and P. A. Gale. Chloride, carboxylate and carbonate transport by ortho-phenylenediamine-based bisureas. *Chem. Sci.*, 4(1):103, 2013.

[101] C. J. E. Haynes and P. A. Gale. Transmembrane anion transport by synthetic systems. *Chem. Commun.*, 47:8203–8209, 2011.

[102] H. J. Clarke, W. Van Rossom, P. N. Horton, M. E. Light, and P. A. Gale. Anion transport and binding properties of N N -(phenylmethylene)dibenzamide based receptors. *Supramol. Chem.*, 28(1-2):10–17, 2016.

[103] E. W. Pelikan. Glossary of Terms and Symbols used in Pharmacology. 2004, available from: http://www.bumc.bu.edu/busm-pm/academics/resources/glossary/, [accessed: 06/05/2015].

[104] M. J. Hynes. EQNMR: a computer program for the calculation of stability constants from nuclear magnetic resonance chemical shift data. *J. Chem. Soc., Dalt. Trans.*, (2):311–312, 1993.

[105] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.*, 50(7):1189–1204, 2010.

[106] I. Filippov and M. Nicklaus. Optical Structure Recognition Software To Recover Chemical Information: OSRAAn Open Source Solution. *J. Chem. Inf. Model.*, 49(3):740–743, 2009.

[107] J. Park, G. R. Rosania, K. A. Shedden, M. Nguyen, N. Lyu, and K. Saitou. Automated extraction of chemical structure information from digital raster images. *Chem. Cent. J.*, 3(1):1–16, 2009.

[108] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini. Dragon Software: an easy approach to molecular descriptor calculations. *Commun. Math. Comput. Chem.*, 56:237–248, 2006.

[109] M. Waldman, R. Fraczkiewicz, and R. D. Clark. Tales from the war on error: The art and science of curating QSAR data. *J. Comput. Aided. Mol. Des.*, 29(9):897–910, 2015.

[110] J. Klump, R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, M. Lautenschlager, U. Schindler, I. Sens, and J. Wächter. Data publication in the open access initiative. *Data Sci. J.*, 5(June):79–83, 2006.

[111] B. Lawrence, C. Jones, B. Matthews, S. Pepler, and S. Callaghan. Citation and Peer Review of Data : Moving Towards Formal Data Publication. *Int. J. Digit. Curation*, 6(2):4–37, 2011.

[112] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC Technical Report). *Pure Appl. Chem.*, 78(11):2051–2066, 2006.

[113] M. E. Wieser, N. Holden, T. B. Coplen, J. K. Böhlke, M. Berglund, W. A. Brand, P. De Bièvre, M. Gröning, R. D. Loss, J. Meija, T. Hirata, T. Prohaska, R. Schoenberg, G. O'Connor, T. Walczyk, S. Yoneda, and X. Zhu. Atomic weights of the elements 2011 (IUPAC Technical Report). *Pure Appl. Chem.*, 85(5):1047–1078, apr 2013.

[114] W. P. Walters. Modeling, informatics, and the quest for reproducibility. *J. Chem. Inf. Model.*, 53(7):1529–30, jul 2013.

[115] F. A. Bulat, A. Toro-Labbe, T. Brinck, J. S. Murray, and P. Politzer. Quantitative analysis of molecular surfaces: Areas, volumes, electrostatic potentials and average local ionization energies. *J. Mol. Model.*, 16(11):1679–1691, 2010.

[116] SAS Institute Inc. *JMP 10 Modeling and Multivariate Methods*. SAS Institute Inc., Cary, NC, USA, 2012.

[117] M. D. Troutt. Regression, 10k Rule of Thumb for. In *Encycl. Stat. Sci.* John Wiley & Sons, Inc., 2004.

[118] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, 15:361–387, 1996.

[119] P. L. Flom and D. L. Cassell. Stopping stepwise : Why stepwise and similar selection methods are bad , and what you should use. In *NESUG 2007 Stat. Data Anal.*, 2007.

[120] B. Sribney. What are some of the problems with stepwise regression? available at: https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/, [accessed: 10/10/2016].

[121] SureChEMBL. Tanimoto Coefficient and Fingerprint Generation. available from: https://www.surechembl.org/knowledgebase/84207-tanimoto-coefficient-and-fingerprint-generation, [accessed: 04/04/2015].

[122] Y. D. Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, and D. S. Wishart. ClassyFire : automated chemical classification with a comprehensive , computable taxonomy. *J. Cheminform.*, 8(61):1–20, 2016.

[123] C. Morley. OpenBabel 2.3.1. available from: http://openbabel.org, [accessed: 28/04/2017].

[124] T. A. Halgren. Merck Molecular Force Field. *J. Comput. Chem.*, 17(5-6):490–519, 1996.

[125] R Core Team. prcomp - stats package. version 3.6.0, available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html, [accessed: 01/02/2017].

[126] B. Carté and D. J. Faulkner. Defensive Metabolites from Three Nembrothid Nudibranchs. *Am. Chem. Soc.*, 48(2):2314–2318, 1983.

[127] M. Carbone, C. Irace, F. Costagliola, F. Castelluccio, G. Villani, G. Calado, V. Padula, G. Cimino, J. Lucas Cervera, R. Santamaria, and M. Gavagnin. A new cytotoxic tambjamine alkaloid from the Azorean nudibranch Tambja ceutae. *Bioorganic Med. Chem. Lett.*, 20(8):2668–2670, 2010.

[128] A. J. Blackman and C. P. Li. New Tambjamine Alkaloids From the Marine Bryozoan Bugula dentata. *Aust. J. Chem.*, 47(8):1625–1629, 1994.

[129] N. R. Williamson, P. C. Fineran, T. Gristwood, S. R. Chawrai, F. J. Leeper, and G. P. C. Salmond. Anticancer and immunosuppressive properties of bacterial prodiginines. *Future Microbiol.*, 2(6):605–618, nov 2007.

[130] B. Díaz De Greñu, P. I. Hernández, M. Espona, D. Quiñonero, M. E. Light, T. Torroba, R. Pérez-Tomás, and R. Quesada. Synthetic prodiginine obatoclax (GX15-070) and related analogues: Anion binding, transmembrane transport, and cytotoxicity properties. *Chem. - A Eur. J.*, 17:14074–14083, 2011.

[131] K. Papireddy, M. Smilkstein, J. X. Kelly, D. Shweta, S. M. Salem, M. Alhamadsheh, S. W. Haynes, G. L. Challis, and K. A. Reynolds. Antimalarial activity of natural and synthetic prodiginines. *J. Med. Chem.*, 54:5296–5306, 2011.

[132] B. C. Cavalcanti, H. V. N. Júnior, M. H. R. Seleghim, R. G. S. Berlinck, G. M. A. Cunha, M. O. Moraes, and C. Pessoa. Cytotoxic and genotoxic effects of tambjamine D, an alkaloid isolated from the nudibranch Tambja eliora, on Chinese hamster lung fibroblasts. *Chem. Biol. Interact.*, 174(3):155–62, 2008.

[133] K. Dairi, S. Tripathy, G. Attardo, and J. F. Lavallée. Two-step synthesis of the bipyrrole precursor of prodigiosins. *Tetrahedron Lett.*, 47(15):2605–2606, 2006.

[134] D. M. Pinkerton, M. G. Banwell, and A. C. Willis. Total syntheses of tambjamines C, E, F, G, H, i and J, BE-18591, and a related alkaloid from the marine bacterium pseudoalteromonas tunicata. *Org. Lett.*, 9(24):5127–5130, 2007.

[135] D. M. Pinkerton, M. G. Banwell, M. J. Garson, N. Kumar, M. O. De Moraes, B. C. Cavalcanti, F. W. A. Barros, and C. Pessoa. Antimicrobial and cytotoxic activities of synthetically derived tambjamines C and E-J, BE-18591, and a related alkaloid from the marine bacterium Pseudoalteromonas tunicata. *Chem. Biodivers.*, 7(5):1311–1324, 2010.

[136] L. N. Aldrich, S. L. Stoops, B. C. Crews, L. J. Marnett, and C. W. Lindsley. Total synthesis and biological evaluation of tambjamine K and a library of unnatural analogs. *Bioorganic Med. Chem. Lett.*, 20(17):5207–5211, 2010.

[137] P. I. Hernández, D. Moreno, A. A. Javier, T. Torroba, R. Pérez-Tomás, and R. Quesada. Tambjamine alkaloids and related synthetic analogs: efficient transmembrane anion transporters. *Chem. Commun.*, 48:1556–1558, 2012.

[138] V. Saggiomo, S. Otto, I. Marques, V. Félix, T. Torroba, and R. Quesada. The role of lipophilicity in transmembrane anion transport. *Chem. Commun.*, 48:5274–5276, 2012.

[139] E. Hernando, V. Soto-Cerrato, S. Cortés-Arroyo, R. Pérez-Tomás, and R. Quesada. Transmembrane anion transport and cytotoxicity of synthetic tambjamine analogs. *Org. Biomol. Chem.*, 12:1771–8, 2014.

[140] N. J. Knight, E. Hernando, C. J. E. Haynes, N. Busschaert, H. J. Clarke, K. Takimoto, M. García-Valverde, J. G. Frey, R. Quesada, and P. A. Gale. QSAR analysis of substituent effects on tambjamine anion transporters. *Chem. Sci.*, 7:1600–1608, 2016.

[141] A. Canty and B. Ripley. R Package boot version 1.3-17. available from: https://cran.r-project.org/web/packages/boot/index.html, [accessed: 06/01/2015].

[142] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997.

[143] D. Bates, M. Maechler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 61(1):1–48, 2015.

[144] N. J. Knight, E. Hernando, C. Haynes, N. Busschaert, H. J. Clarke, K. Takimoto, M. Garcia-Valverde, J. G. Frey, R. Quesada, and P. A. Gale. QSAR analysis of substituent effects on tambjamine anion transporters: Supplementary Data. http://dx.doi.org/10.5258/SOTON/384138, 2015.

[145] A. V. Hill. The Combinations of Haemoglobin with Oxygen and with Carbon Monoxide. *Biochem. J.*, 7(5):471–480, 1913.

[146] J. S. Lolkema and D. J. Slotboom. The Hill analysis and co-iondriven transporter kinetics. *J. Gen. Physiol.*, 145(6):565–574, 2015.

[147] S. Bosale and S. Matile. A Simple Method to Identify Supramolecules in Action: Hill Coefficients for Exergonic Self-Assembly. *Chirality*, 18:849–856, 2006.

[148] IUPAC. *Compendium of Chemical Terminology (the "gold" book)*. Blackwell Scientific Publications, Oxford, 2nd edition, 1997.

[149] M. Wu, M. Vogt, G. M. Maggiora, and J. Bajorath. Design of chemical space networks on the basis of Tversky similarity. *J. Comput. Aided. Mol. Des.*, 30(1):1–12, 2016.

[150] X. Wu. Private communication. email exchange, 2016.

[151] W. Van Rossom, D. J. Asby, A. Tavassoli, and P. A. Gale. Perenosins: a new class of anion transporter with anti-cancer activity. *Org. Biomol. Chem.*, pages 2645–2650, 2016.

[152] M. Olivari, R. Montis, L. E. Karagiannidis, P. N. Horton, L. K. Mapp, S. J. Coles, M. E. Light, P. A. Gale, and C. Caltagirone. Anion complexation, transport and structural studies of a series of bis-methylurea compounds. *Dalt. Trans.*, 44:2138–2149, 2015.

[153] R. B. P. Elmes, N. Busschaert, D. D. Czech, P. A. Gale, and K. A. Jolliffe. pH switchable anion transport by an oxothiosquaramide. *Chem. Commun.*, 51(50):10107–10, 2015.

[154] M. J. Spooner and P. A. Gale. Anion transport across varying lipid membranes the effect of lipophilicity. *Chem. Commun.*, 51:4883–4886, 2015.

[155] S. N. Berry, N. Busschaert, C. L. Frankling, D. Salter, and P. A. Gale. Aromatic isophthalamides aggregate in lipid bilayers : evidence for a cooperative transport mechanism . *Org. Biomol. Chem.*, pages 3136–3143, 2015.

[156] L. E. Karagiannidis, C. J. E. Haynes, K. J. Holder, I. L. Kirby, S. J. Moore, N. J. Wells, and P. A. Gale. Highly effective yet simple transmembrane anion transporters based upon ortho-phenylenediamine bis-ureas. *Chem. Commun. (Camb).*, 50(81):12050–3, 2014.

[157] H. Valkenier, C. J. E. Haynes, J. Herniman, P. A. Gale, and A. P. Davis. Lipophilic balance a new design principle for transmembrane anion carriers. *Chem. Sci.*, 5(3):1128, 2014.

[158] N. Busschaert, R. B. P. Elmes, D. D. Czech, X. Wu, I. L. Kirby, E. M. Peck, K. D. Hendzel, S. K. Shaw, B. Chan, B. D. Smith, K. A. Jolliffe, and P. A. Gale. Thiosquaramides: pH switchable anion transporters. *Chem. Sci.*, 5(9):3617–26, 2014.

[159] J. G. Frey. Curation of Laboratory Experimental Data as Part of the Overall Data Lifecycle. *Int. J. Digit. Curation*, 3(1):44–62, 2008.

[160] C. Sansom. Exploiting the data mine. aug 2015, available from: https://www.chemistryworld.com/feature/exploiting-the-data-mine/8850.article, [accessed: 26/09/2016].

[161] M. P. Long and R. C. Schonfeld. Supporting the Changing Research Practices of Chemists. 2013, available at: http://www.sr.ithaka.org/sites/default/files/reports/Supporting-the-Changing-Research-Practices-of-Chemists-FINAL.pdf, [accessed: 04/06/2017].

[162] T. Hey and A. Trefethen. The data deluge: an e-science perspective. In *Grid Comput. Mak. Glob. Infrastruct. a reality.*, number January, pages 809–824. Wiley, Chichester, 2003.

[163] A. Eisenberg. Keeping a Laboratory Notebook. *J. Chem. Educ.*, 59(12):1045–1046, 1982.

[164] E. Wilson. Laboratory Notebooks: Sacred Works. *Plant Mol. Biol. Report.*, 8(4):220–222, 1990.

[165] K. Taylor. The status of electronic laboratory notebooks for chemistry and biology. *Curr. Opin. drug Discov. Dev.*, 9(3):348–353, 2006.

[166] J. M. Wright. Make it better but don't change anything. *Autom. Exp.*, 1(1):3–5, 2009.

[167] S. Y. Nussbeck, P. Weil, J. Menzel, B. Marzec, K. Lorberg, and B. Schwappach. The laboratory notebook in the 21st century. *EMBO Rep.*, 15(6):631–634, 2014.

[168] S. Kanza, C. Willoughby, N. Gibbins, R. Whitby, J. G. Frey, J. Erjavec, K. Zupančič, M. Hren, and K. Kovač. Electronic lab notebooks : can they replace paper ? *J. Cheminform.*, 9(31):1–15, 2017.

[169] S. Carpenter. Must a Paper Trail be Paper. available at: http://www.sciencemag.org/careers/2012/09/must-paper-trail-be-paper, [accessed: 05/05/2016].

[170] N. Morris. To lab book, or not to lab book, that is the question? available from: http://www.nature.com/scitable/blog/bioscience-elearning/to_lab_book_or_not, [accessed: 06/08/2015].

[171] A. J. Milsted, J. R. Hale, J. G. Frey, and C. Neylon. LabTrove: a lightweight, web based, laboratory "blog" as a route towards a marked up record of work in a bioscience research laboratory. *PLoS One*, 8(7):e67460, jan 2013.

[172] K. A. Badiola, C. Bird, W. S. Brocklesby, J. Casson, R. T. Chapman, S. J. Coles, J. R. Cronshaw, A. Fisher, J. G. Frey, D. Gloria, M. C. Grossel, D. B. Hibbert, N. Knight, L. K. Mapp, L. Marazzi, B. Matthews, A. Milsted, R. S. Minns, K. T. Mueller, K. Murphy, T. Parkinson, R. Quinnell, J. S. Robinson, M. N. Robertson, M. Robins, E. Springate, G. Tizzard, M. H. Todd, A. E. Williamson, C. Willoughby, E. Yang, and P. M. Ylioja. Experiences with a researcher-centric ELN. *Chem. Sci.*, 6(3):1614–1629, 2015.

[173] M. N. Robertson, P. M. Ylioja, A. E. Williamson, M. Woelfle, M. Robins, K. A. Badiola, P. Willis, P. Olliaro, T. N. C. Wells, and M. H. Todd. Open source drug discovery A limited tutorial. *Parasitology*, 141(1):148–157, 2014.

[174] OSM - Open Source Malaria. available at: http://opensourcemalaria.org, [accessed: 20-04-2017].

[175] Optical-Research-Centre. ORC Xray group blogs. available at: http://xray.orc.soton.ac.uk/, [accessed: 03/03/2014].

[176] A. E. Day, S. J. Coles, C. L. Bird, J. G. Frey, R. J. Whitby, V. E. Tkachenko, and A. J. Williams. ChemTrove: Enabling a generic ELN to support chemistry through the use of transferable plug-ins and online data sources. *J. Chem. Inf. Model.*, 55(3):501–509, 2015.

[177] R Studio Team. RStudio: Integrated Development for R. available at: https://www.rstudio.com/, [accessed:19/10/2016].

[178] M. A. C. Gatto. Making Research Useful : Current Challenges and Good Practices in Data Visualisation. Technical Report May, Reuters Institute for the Study of Journalism, 2015.

[179] SAS Institute. Data Visualization: What it is and why it matters. available at: https://www.sas.com/en_sg/insights/big-data/data-visualization.html, [accesed: 05/04/2017].

[180] Daily Infographic. What Is an Infographic? The History and Evolution of Data Visualization. available at: http://www.dailyinfographic.com/blog/what-is-an-infographic-history-and-evolution, [accessed:10/12/2017].

[181] M. Krystian. 18 Surprising Data Visualizations in Your Everyday Life. available at: https://infogram.com/blog/18-surprising-data-visualizations-in-your-everyday-life/, [accessed: 10/12/2017].

[182] W3schools. HTML DOM Events. available at: https://www.w3schools.com/jsref/dom_obj_event.asp, [accessed: 14/10/2015].

[183] A. Bender and R. C. Glen. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, 2(22):3204, 2004.

[184] M. Kaufmann and D. Wagner, editors. *Drawing Graphs - Methods and Models*. Springer, 2001.

[185] Plot.ly. plotly.js JavaScript Graphing Library. available from: https://plot.ly/javascript/, [accessed: 02/02/2017].

[186] R. Vaidyanathan. rCharts. available from: http://ramnathv.github.io/rCharts/, [accessed: 04/01/2017].

[187] M. DeGusta. Are Smart Phones Spreading Faster then Any Technology in Human History? available from: https://www.technologyreview.com/s/427787/are-smart-phones-spreading-faster-than-any-technology-in-human-history/, [accessed: 13/01/2017].

[188] Ofcom. Landline and mobile statistics. Fast Facts, 2016, available from: https://www.ofcom.org.uk/about-ofcom/latest/media/facts, [accessed: 17/03/2017].

[189] C. Phillips. How Smartphones Revolutionized Society in Less than a Decade. available from: http://www.govtech.com/products/How-Smartphones-Revolutionized-Society-in-Less-than-a-Decade.html, [accessed:04/06/2017], 2014.

[190] J. E. Tomayko. Part I: Manned Spacecraft Computers. In *Comput. Spacefl. NASA Exp.* NASA, 1988.

[191] K. A. Zimmermann. History of Computers: A Brief Timeline. available from: https://www.livescience.com/20718-computer-history.html, [accessed: 18/11/2017], 2017.

[192] M. Foster. History and Growth of Mobile Phone Technology. available from: http://www.tccohio.com/blog/telephone-technology, [accessed: 02/04/2017], 2015.

[193] M. Strain. 1983 to today: a history of mobile apps. available from: https://www.theguardian.com/media-network/2015/feb/13/history-mobile-apps-future-interactive-timeline, [accessed: 09/01/2017], 2015.

[194] GCN Staff. 25 years: A technology timeline. available from: https://gcn.com/Articles/2007/12/06/25-years–A-technology-timeline.aspx, [accessed: 05/01/2017], 2007.

[195] J. Elliot. Clean Machines: Washers and Dryers Designed for the Smart Home. available from: https://www.mansionglobal.com/articles/76815-clean-machines-washers-and-dryers-designed-for-the-smart-home, [accessed: 14/11/2017], 2017.

[196] J. Kiruthika and D. Arulanantham. Making Washing Machines Smart through IoT. *Int. J. Mod. Trends Eng. Sci.*, 3(6):39–41, 2016.

[197] DentistryIQ. Kolibree introduces Ara, the first toothbrush with artificial intelligence. available from: https://www.dentistryiq.com/articles/2017/01/kolibree-introduces-ara-the-first-toothbrush-with-artificial-intelligence.html, [accessed: 04/07/2017], 2017.

[198] A. McEwen and H. Cassimally. The Internet of Things: An Overview. In *Des. Internet Things*. John Wiley & Sons, Inc., 2014.

[199] J. Sinclair, R. Boyatt, C. Rocks, and M. Joy. Massive Open Online Courses: a review of usage and evaluation. *Int. J. Learn. Technol.*, 10(1):71–93, 2015.

[200] A. Finder. A Surge in Growth for a New Kind of Online Course. available from: https://www.nytimes.com/2013/09/26/technology/personaltech/a-surge-in-growth-for-a-new-kind-of-online-course.html, [accessed: 07/05/2014], 2013.

[201] C. A. Jara, F. A. Candelas, S. T. Puente, and F. Torres. Hands-on experiences of undergraduate students in Automatics and Robotics using a virtual and remote laboratory. *Comput. Educ.*, 57(4):2451–2461, 2011.

[202] F. A. Candelas, S. T. Puente, F. Torres, P. Gil, F.G. Ortiz, and J. Pomares. A Virtual Laboratory for Teaching Robotics. *Int. J. Eng. Educ.*, 19(3):363–370, 2003.

[203] M. Cooper. Remote laboratories in teaching and learning issues impinging on widespread adoption in science and engineering education. *Int. J. Online Eng.*, 1(1), 2005.

[204] M. Cooper and J. M. M. Ferreira. Remote Laboratories Extending Access to Science and Engineering Curricular. *IEEE Trans. Learn. Technol.*, 2(4):342–353, 2009.

[205] S. Perry. Lab Book 6174. University of Southampton, 2011.

[206] F. M. Zehentbauer, C. Moretto, R. Stephen, T. Thevar, J. R. Gilchrist, D. Pokrajac, K. L. Richard, and J. Kiefer. Fluorescence spectroscopy of Rhodamine 6G : Concentration and solvent effects. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, 121:147–151, 2014.

[207] N. Knight, A. Foster, and R. Stoneman. Experimental Analysis of BLL experiment. available at: http://altc.ourexperiment.org/beer_lambert_log/338/Experimental_Analysis_of_Beer_Lambert_Law_Experiment.html, [accessed: 19/11/2013], 2012.

[208] T. Burke. RobotEyez. available at: https://github.com/tedburke/RobotEyez, [accessed: 02/08/2014], 2011.

[209] P. Atkins and J. de Paula. *Elements of Physical Chemistry*. Oxford University Press, Oxford, 5th edition, 2009.

[210] C. Flowers. Remote Experiments Practical Script for Undergraduate Labs. unpublished, 2015.

[211] S. Iyengar. What do the IoT and R&D Labs Have in Common? Not Much. Yet. available from: https://medium.com/@sridhariyengar/what-do-the-iot-and-r-d-labs-have-in-common-not-much-yet-84cdb7bf094c, [accessed: 12/12/2017], 2017.

[212] G. Hughes, H. Mills, D. De Roure, J. G. Frey, L. Moreau, M. C. Schraefel, G. Smith, and E. Zaluska. The semantic smart laboratory: a system for supporting the chemical eScientist. *Org. Biomol. Chem.*, 2(22):3284–3293, 2004.

[213] D. S. Lütjohann, N. Jung, and S. Bräse. Open source life science automation: Design of experiments and data acquisition via "dial-a-device". *Chemom. Intell. Lab. Syst.*, 144:100–107, 2015.

[214] H. Bethany. Meet your new lab assistant. *Chem. Eng. News*, 95(19):26–27, 2017.

[215] J. G. Frey. Dark Lab or Smart Lab : The Challenges for 21st Century Laboratory Software. *Org. Process Res. Dev.*, 8:1024–1035, 2004.

[216] S. Wilson and J. Frey. The SmartLab: Experimental and environmental control and monitoring of the chemistry Laboratory. *2009 Int. Symp. Collab. Technol. Syst. CTS 2009*, pages 85–90, 2009.

[217] J. M. Robinson, J. G. Frey, A. J. Standford-Clark, A. D. Reynolds, and B. V. Bedi. Sensor Networks and Grid Middleware for Laboratory Monitoring. In *First Int. Conf. e-Science Grid Comput.*, Melbourne, Vic., 2005.

[218] J. Martin. Raspberry Pi 3 release date, price and specifications. *Tech Advis.*, feb 2016.

[219] N. Normal. 47 Raspberry Pi Projects to Inspire Your Next Build. https://makezine.com/2013/04/14/47-raspberry-pi-projects-to-inspire-your-next-build/, [accessed: 01/05/2015], 2013.

[220] X. Zhong and Y. Liang. Raspberry Pi : An Effective Vehicle in Teaching the Internet of Things in Computer Science and Engineering. *Electronics*, 5(3):56, 2016.

[221] Whiskeytangohotel.com. Hand of PI (Twitter controlled Robot Hand). http://www.whiskeytangohotel.com/2013/04/hand-of-pi-tweeter-controlled-robot-hand.html, [accessed 02/03/2017], 2013.

[222] Falldeaf.com. The Make Contest. http://falldeaf.com/2013/04/the-make-contest/, [accessed 02/03/2017], 2013.

[223] L. Upton. Raspberry PI Blog: Hackspace Security System. https://www.raspberrypi.org/blog/hackspace-security-system/, [accessed 02/03/2017], 2013.

[224] T. Reed. Internet control of Hydroponics using Raspberry Pi and Arduino. http://hapihq.com/blog/2013/04/10/internet-control-of-hydroponics-using-raspberry-pi-and-arduino/, [accessed 02/03/2017], 2013.

[225] E. Brown. 2017 hacker board survey results. http://linuxgizmos.com/2017-hacker-board-survey-raspberry-pi-still-rules-but-x86-sbcs-make-gains/, [accessed: 02/08/2017], 2017.

[226] E. Upton. Ten Millionth Raspberry Pi, and a New Kit. https://www.raspberrypi.org/blog/ten-millionth-raspberry-pi-new-kit/, [accessed: 01/11/2016], 2016.

[227] J. Kridner. BeagleBone: open-hardware expandable computer. available at: http://beagleboard.org/Support/bone101/, [accessed: 12/07/2017].

[228] D. Kushner. The Making of Arduino. https://spectrum.ieee.org/geek-life/hands-on/the-making-of-arduino, [accessed: 12/03/2015], 2011.

[229] What is an Arduino? https://learn.sparkfun.com/tutorials/what-is-an-arduino, [accessed: 17/04/2015].

[230] MQTT.org. MQ Telemetry Transport. www.mqtt.org, [accessed: 01/11/2017].

[231] OASIS. MQTT Version 3.1.1 becomes an OASIS Standard. https://www.oasis-open.org/news/announcements/mqtt-version-3-1-1-becomes-an-oasis-standard, [accessed: 15/02/2017], 2014.

[232] R. A. Light. Mosquitto : server and client implementation of the MQTT protocol. *J. Open Source Softw.*, 2(May):10–11, 2017.

[233] N. Heath. How IBM's Node-RED is hacking together the internet of things. available at: https://www.techrepublic.com/article/node-red/, [accessed: 30/06/2016], 2014.

[234] Amazon. Amazon Echo Dot arrives in the UK as Alexa adds new UK features. press release, 20 Oct 2016, available at: http://phx.corporate-ir.net/phoenix.zhtml?c=251199&p=irol-newsArticle&ID=2213230, [accessed: 09/01/2017].

[235] T. Karczewski. Alexa Voice Service Now Available for the UK and Germany. Alexa Blogs, 07 Feb 2017, available at: https://developer.amazon.com/blogs/post/f6d79f37-1de1-4369-9660-c4347a91e76b/avs-now-available-for-the-uk-and-germany, [accessed: 13/04/2017].

[236] M. A. Razzaq, S. H. Gill, M. A. Qureshi, and S. Ullah. Security Issues in the Internet of Things (IoT): A Comprehensive Study. *Int. J. Adv. Comput. Sci. Appl.*, 8(6):383, 2017.

[237] Hypercat Alliance. HyperCat Standard. available at: http://www.hypercat.io/standard.html, [accessed: 06/03/2017], 2016.

[238] M. Barnes. Alexa, are you listening. available from:: https://labs.mwrinfosecurity.com/blog/alexa-are-you-listening, [accessed: 10/12/2017], 2017.

[239] C. Wueest. A guide to the security of voice-activated smart speakers. white paper, available from: https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-security-voice-activated-smart-speakers-en.pdf, [accessed: 19/01/2018], 2017.

[240] P. Cutsinger. Announcing Alexa Skill Management API, Alexa Skills Kit Command-line Interface, and Events in the Alexa Skills Kit. available from: https://developer.amazon.com/blogs/alexa/post/e7a57ec4-f9e0-4efa-9052-06d320245f9b/announcing-alexa-skill-management-api-alexa-skills-kit-command-line-interface-and-events-in-alexa-skills-kit, [accessed: 10/12/2017], 2017.

Appendix A

Anion Transporter Plots

Graphs plotting various models created in Chapter 2 - Modelling Anion Transport in Vesicles

A.1 Full dataset

All Possible Models

Ordered up to best 10 models up to 3 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC
VE2_H2	1	0.1526	0.9027	228.085	235.116
DLS_04	1	0.1413	0.9087	229.214	236.246
VE2_A	1	0.1380	0.9104	229.535	236.567
VE2_X	1	0.1372	0.9108	229.615	236.647
EE_L	1	0.1354	0.9118	229.796	236.827
SM6_L	1	0.1347	0.9122	229.862	236.893
MPC10	1	0.1329	0.9131	230.040	237.072
SM5_L	1	0.1328	0.9131	230.046	237.078
SM4_L	1	0.1289	0.9152	230.428	237.459
SM6_H2	1	0.1261	0.9167	230.707	237.739
SpDiam_Dz.e.,ECC	2	0.2975	0.8269	214.355	223.626
SpDiam_Dz.i.,ECC	2	0.2882	0.8324	215.472	224.743
SpPosA_D,WiA_Dz.e.	2	0.2701	0.8429	217.604	226.874
SpMaxA_D,WiA_Dz.e.	2	0.2701	0.8429	217.604	226.874
SpMAD_D,WiA_Dz.e.	2	0.2698	0.8430	217.637	226.908
SpPos_D,AVS_Dz.e.	2	0.2655	0.8455	218.140	227.410
SpMax_D,AVS_Dz.e.	2	0.2655	0.8455	218.140	227.410
SpAD_D,AVS_Dz.e.	2	0.2655	0.8455	218.140	227.411
SpPosA_D,WiA_Dz.i.	2	0.2546	0.8517	219.385	228.655
SpMaxA_D,WiA_Dz.i.	2	0.2546	0.8517	219.385	228.655
CSI,TI1_L,WiA_Dz.v.	3	0.3820	0.7803	205.718	217.172
SpPos_D,VE3_X,AVS_Dz.e.	3	0.3819	0.7804	205.731	217.185
SpMax_D,VE3_X,AVS_Dz.e.	3	0.3819	0.7804	205.731	217.185
SpAD_D,VE3_X,AVS_Dz.e.	3	0.3819	0.7804	205.731	217.185
CSI,TI1_L,SpMAD_Dz.v.	3	0.3770	0.7835	206.402	217.855
CSI,TI1_L,SpPosA_Dz.v.	3	0.3746	0.7850	206.733	218.186
CSI,TI1_L,SpMaxA_Dz.v.	3	0.3746	0.7850	206.733	218.186
AVS_Dz.i.,SpAbs_Dz.i.,DLS_04	3	0.3704	0.7876	207.293	218.747
AVS_Dz.i.,SpAD_Dz.i.,DLS_04	3	0.3663	0.7902	207.848	219.302
SpPos_D,VE3_X,AVS_Dz.i.	3	0.3654	0.7908	207.972	219.425

Figure A.1: Top 10 models for Log(1/EC_{50}) through fit-all - up to 3 parameters ranked by R^2 for the full Gale anion transporter dataset

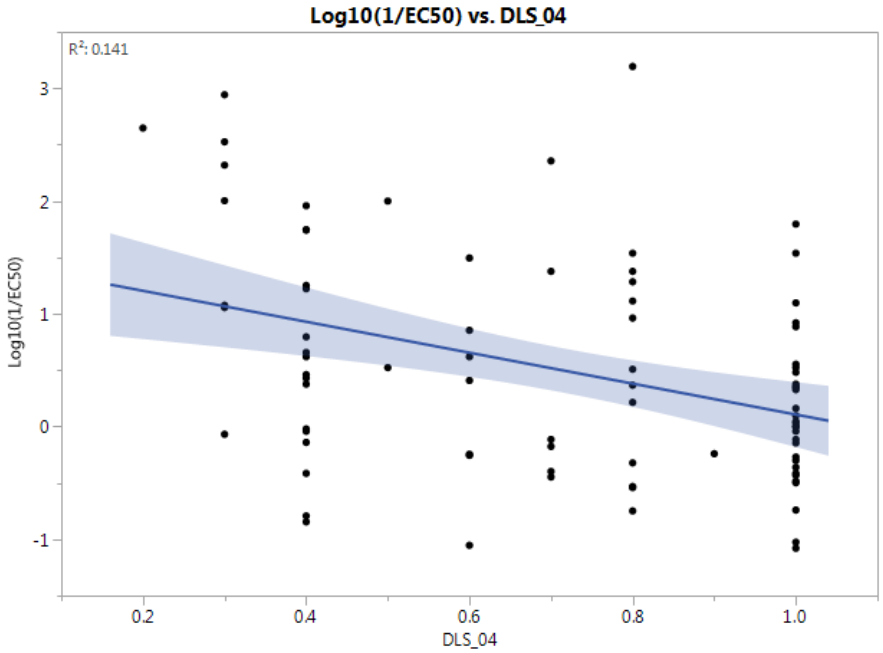


Figure A.2: DLS_04 vs. Log(1/EC_{50})
$R^2 = 0.1413$, $R^2_{adj} = 0.1309$

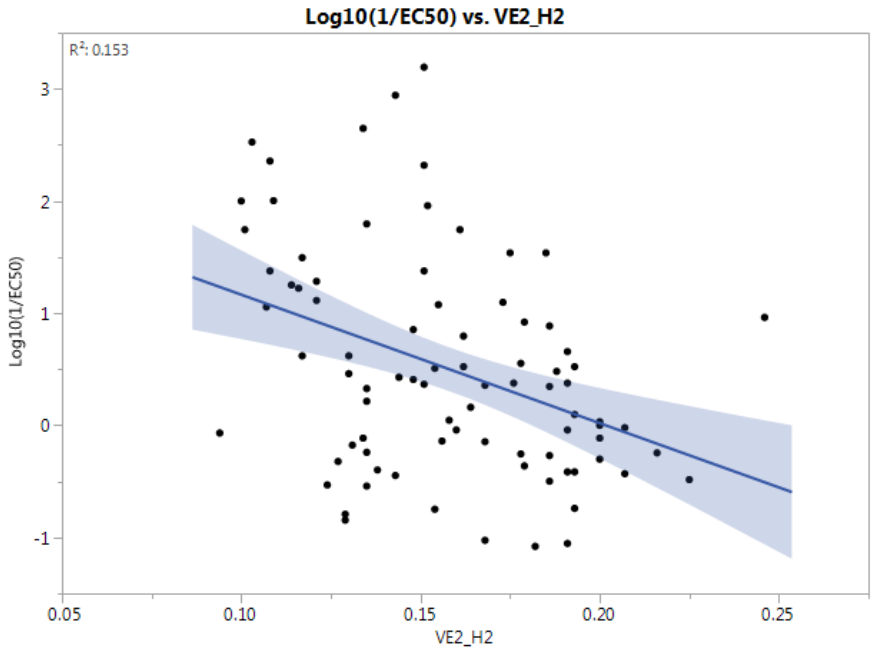


Figure A.3: VE2_H2 vs. Log(1/EC_{50})
$R^2 = 0.153$, $R^2_{adj} = 0.142$

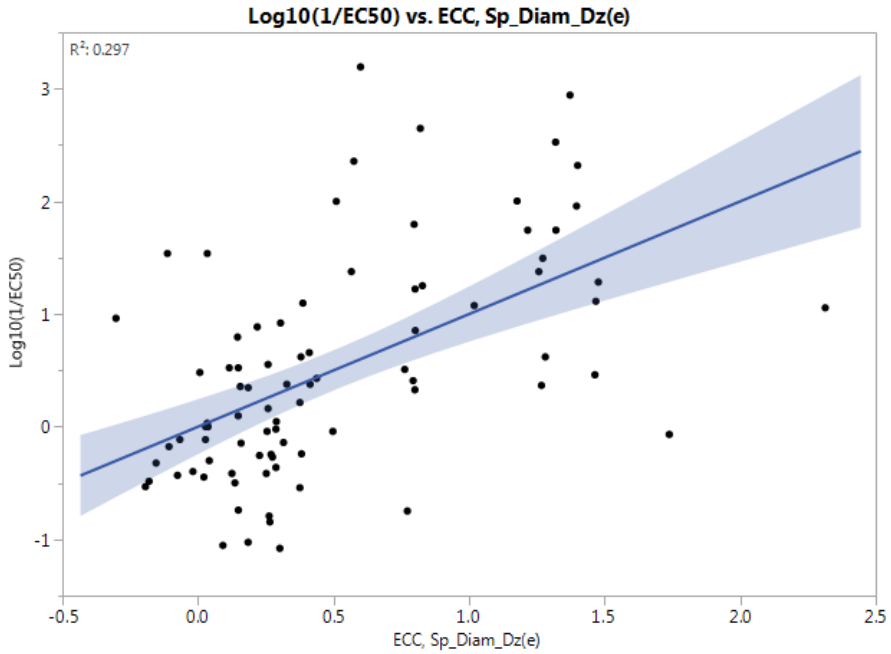


Figure A.4: Model of ECC & SpDiam_Dz(e) vs. Log($1/EC_{50}$)
$R^2 = 0.2975$, $R^2_{adj} = 0.2803$

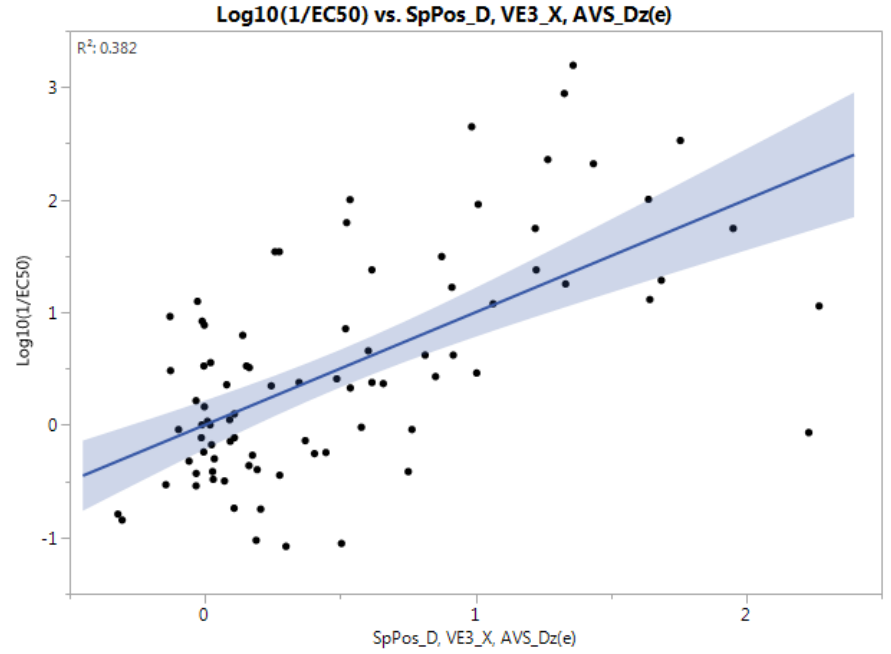


Figure A.5: Model of SpPos_D, VE3_X & AVS_Dz(e) vs. Log($1/EC_{50}$)
$R^2 = 0.3819$, $R^2_{adj} = 0.3590$

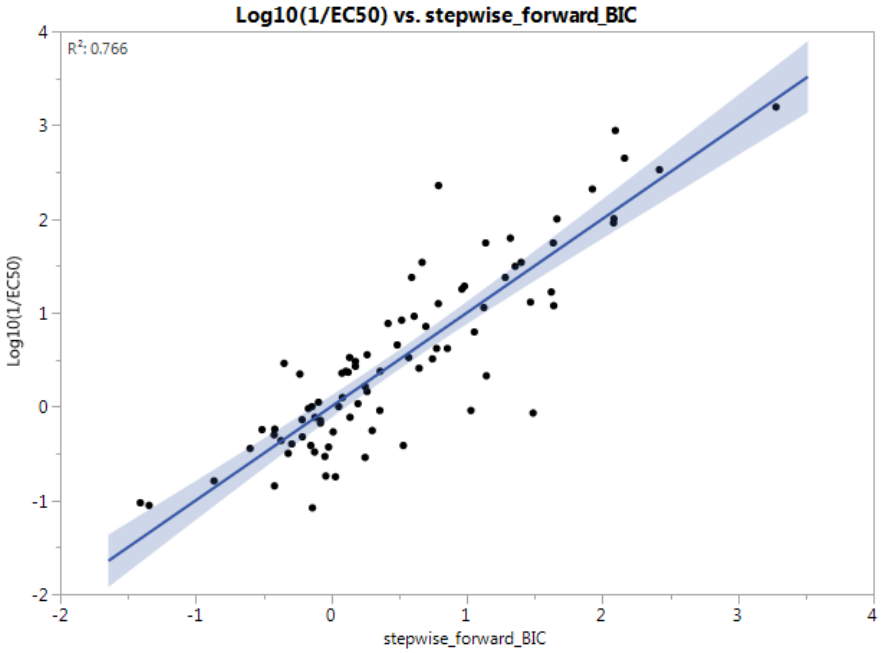


Figure A.6: Actual vs. Predicted Log($1/EC_{50}$) for Forward stepwise model using BIC
$R^2 = 0.821$, $R^2_{adj} = 0.753$

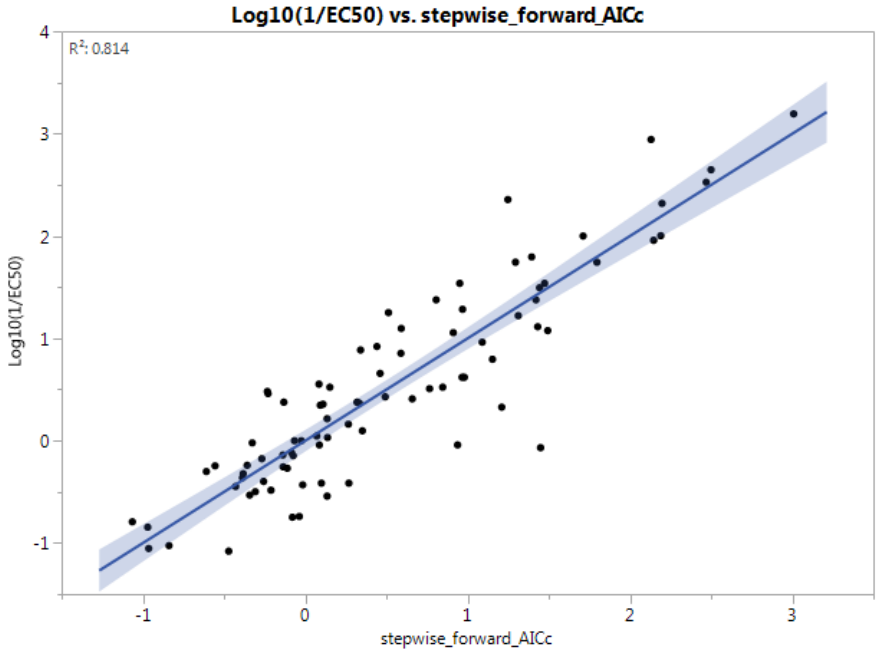


Figure A.7: Actual vs. Predicted Log($1/EC_{50}$) for Forward stepwise model using AICc
$R^2 = 0.8726$, $R^2_{adj} = 0.809$

	DLS_04	ICR	LOC	VE2_X	ChiA_Dz(i)	SpMax_B(p)	AVS_H2	VE2_H2	P_VSA_MR_2	RBF	EE_B(m)	SpMax_B(s)	SpMaxA_L	Psi_e_0	S0K	Hy	J_Dz(p)	ON0V
DLS_04	**1.000**	-0.017	0.764	0.511	0.424	-0.405	-0.747	0.450	-0.274	0.635	-0.293	-0.442	0.502	-0.418	-0.386	-0.332	0.489	-0.132
ICR	-0.017	**1.000**	-0.228	-0.510	-0.694	0.072	0.218	-0.566	0.191	0.440	0.269	0.191	-0.625	0.422	0.670	0.634	-0.572	0.720
LOC	0.764	-0.228	**1.000**	0.686	0.562	-0.419	-0.769	0.520	-0.278	0.536	-0.311	-0.383	0.669	-0.499	-0.526	-0.705	**0.814**	-0.373
VE2_X	0.511	-0.510	0.686	**1.000**	**0.895**	-0.671	**-0.885**	**0.895**	-0.565	0.195	-0.462	-0.514	**0.915**	-0.778	**-0.879**	-0.746	0.729	-0.780
ChiA_Dz(i)	0.424	-0.694	0.562	**0.895**	**1.000**	-0.499	-0.771	**0.885**	-0.576	0.040	-0.501	-0.532	**0.955**	-0.780	**-0.861**	-0.724	0.660	**-0.807**
SpMax_B(p)	-0.405	0.072	-0.419	-0.671	-0.499	**1.000**	0.712	-0.653	0.245	-0.314	0.455	0.332	-0.449	0.340	0.455	0.318	-0.429	0.287
AVS_H2	-0.747	0.218	-0.769	**-0.885**	-0.771	0.712	**1.000**	**-0.837**	0.582	-0.521	0.470	0.586	**-0.826**	0.744	0.741	0.552	-0.650	0.541
VE2_H2	0.450	-0.566	0.520	**0.895**	**0.885**	-0.653	**-0.837**	**1.000**	-0.582	0.061	-0.480	-0.534	**0.865**	-0.769	**-0.876**	-0.606	0.618	-0.770
P_VSA_MR_2	-0.274	0.191	-0.278	-0.565	-0.576	0.245	0.582	-0.582	**1.000**	-0.158	0.335	0.524	-0.632	**0.831**	0.640	0.287	-0.255	0.608
RBF	0.635	0.440	0.536	0.195	0.040	-0.314	-0.521	0.061	-0.158	**1.000**	-0.089	-0.329	0.103	-0.136	0.065	0.042	0.190	0.298
EE_B(m)	-0.293	0.269	-0.311	-0.462	-0.501	0.455	0.470	-0.480	0.335	-0.089	**1.000**	0.181	-0.507	0.491	0.490	0.362	-0.326	0.420
SpMax_B(s)	-0.442	0.191	-0.383	-0.514	-0.532	0.332	0.586	-0.534	0.524	-0.329	0.181	**1.000**	-0.503	0.509	0.472	0.273	-0.298	0.305
SpMaxA_L	0.502	-0.625	0.669	**0.915**	**0.955**	-0.449	**-0.826**	**0.865**	-0.632	0.103	-0.507	-0.503	**1.000**	**-0.854**	**-0.904**	-0.768	0.717	**-0.840**
Psi_e_0	-0.418	0.422	-0.499	-0.778	-0.780	0.340	0.744	-0.769	**0.831**	-0.136	0.491	0.509	**-0.854**	**1.000**	**0.879**	0.586	-0.515	**0.812**
S0K	-0.386	0.670	-0.526	**-0.879**	**-0.861**	0.455	0.741	**-0.876**	0.640	0.065	0.490	0.472	**-0.904**	**0.879**	**1.000**	0.758	-0.678	**0.931**
Hy	-0.332	0.634	-0.705	-0.746	-0.724	0.318	0.552	-0.606	0.287	0.042	0.362	0.273	-0.768	0.586	0.758	**1.000**	**-0.852**	0.752
J_Dz(p)	0.489	-0.572	**0.814**	0.729	0.660	-0.429	-0.650	0.618	-0.255	0.190	-0.326	-0.298	0.717	-0.515	-0.678	**-0.852**	**1.000**	-0.565
ON0V	-0.132	0.720	-0.373	-0.780	**-0.807**	0.287	0.541	-0.770	0.608	0.298	0.420	0.305	**-0.840**	**0.812**	**0.931**	0.752	-0.565	**1.000**

Figure A.8: Correlation matrix of variables in the stepwise AICc model coloured by correlation value

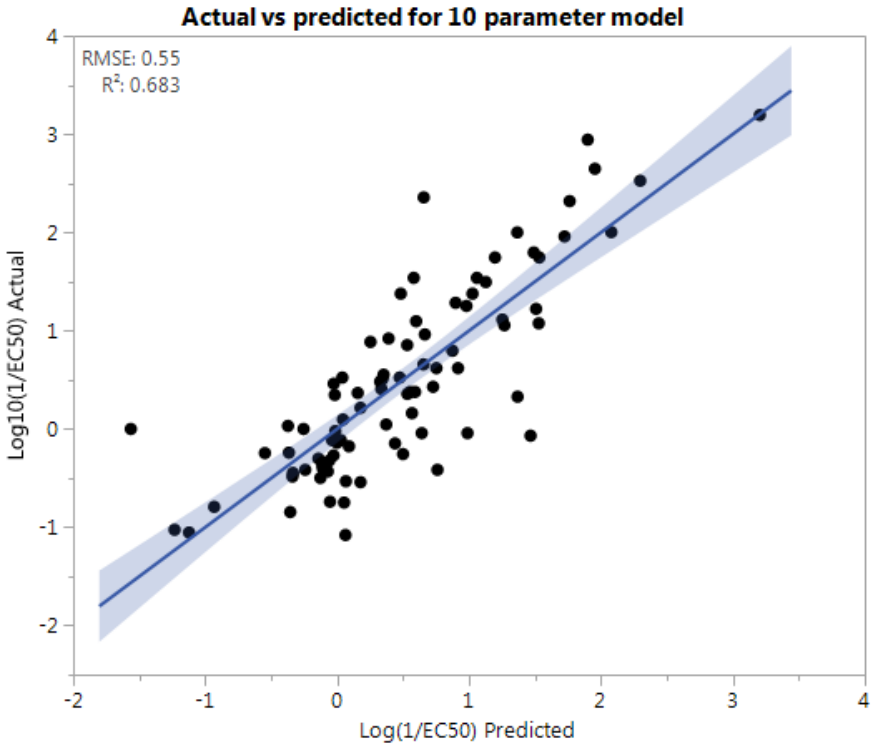


Figure A.9: Actual vs. Predicted Log(1/EC_{50}) for 10 parameter model $R^2 = 0.683$, $R^2_{adj} = 0.644$

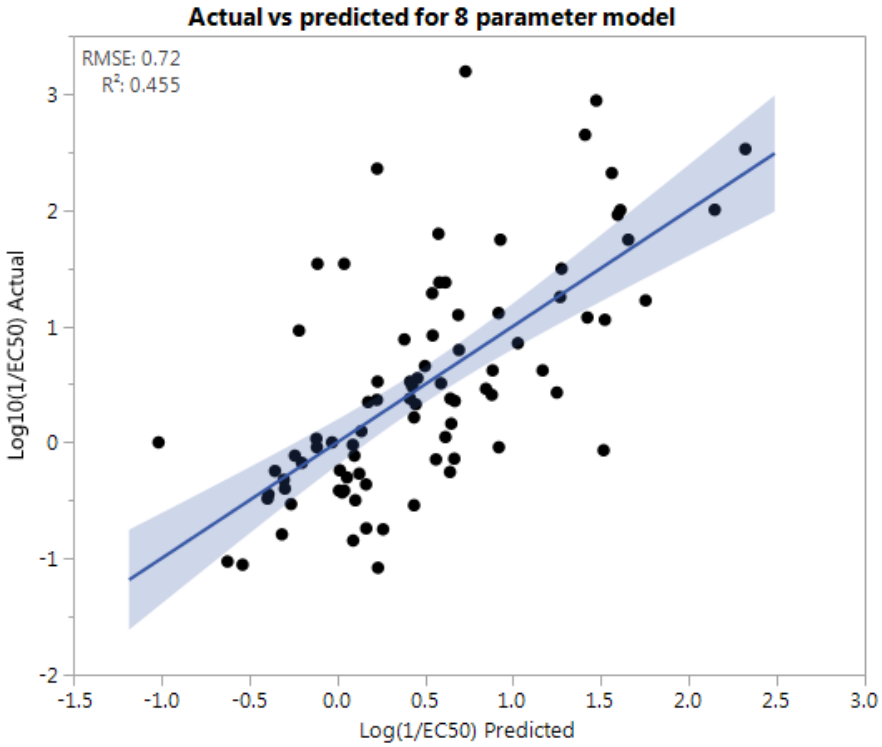


Figure A.10: Actual vs. Predicted Log(1/EC_{50}) for 8 parameter model $R^2 = 0.455$, $R^2_{adj} = 0.406$

	DLS_04	LOC	VE2_X	ChiA_Dz(i)	SpMax_B(p)	VE2_H2	P_VSA_MR_2	RBF	AVS_H2
DLS_04	1.0000	0.7642	0.5112	0.4241	-0.4053	0.4502	-0.2744	0.6355	-0.7470
LOC	0.7642	1.0000	0.6860	0.5624	-0.4188	0.5196	-0.2777	0.5364	-0.7690
VE2_X	0.5112	0.6860	1.0000	0.8950	-0.6706	0.8949	-0.5648	0.1948	-0.8853
ChiA_Dz(i)	0.4241	0.5624	0.8950	1.0000	-0.4989	0.8850	-0.5758	0.0401	-0.7711
SpMax_B(p)	-0.4053	-0.4188	-0.6706	-0.4989	1.0000	-0.6526	0.2450	-0.3141	0.7124
VE2_H2	0.4502	0.5196	0.8949	0.8850	-0.6526	1.0000	-0.5823	0.0607	-0.8373
P_VSA_MR_2	-0.2744	-0.2777	-0.5648	-0.5758	0.2450	-0.5823	1.0000	-0.1585	0.5820
RBF	0.6355	0.5364	0.1948	0.0401	-0.3141	0.0607	-0.1585	1.0000	-0.5210
AVS_H2	-0.7470	-0.7690	-0.8853	-0.7711	0.7124	-0.8373	0.5820	-0.5210	1.0000

Figure A.11: Correlation matrix of variables in the 10 parameter model coloured by correlation value

A.2 Expanded Subset Analysis

List of compounds in expanded subset (with EC_{50} values)

- 101039_c0sc00503g-6
- 101039_c2sc20041d-10
- 101039_c2sc20041d-11
- 101039_c2sc20041d-12
- 101039_c2sc20041d-7
- 101039_c2sc20041d-8
- 101039_c2sc20041d-9
- 101039_c2sc20551c-11
- 101039_c2sc20551c-2 ($\vee$)
- 101039_c2sc20551c-3
- 101039_c2sc20551c-4
- 101039_c3sc51023a-1
- 101039_c3sc51023a-10
- 101039_c3sc51023a-11
- 101039_c3sc51023a-12
- 101039_c3sc51023a-13
- 101039_c3sc51023a-14
- 101039_c3sc51023a-15
- 101039_c3sc51023a-16
- 101039_c3sc51023a-17
- 101039_c3sc51023a-18
- 101039_c3sc51023a-19
- 101039_c3sc51023a-2
- 101039_c3sc51023a-20
- 101039_c3sc51023a-21
- 101039_c3sc51023a-22
- 101039_c3sc51023a-3
- 101039_c3sc51023a-4
- 101039_c3sc51023a-5
- 101039_c3sc51023a-6
- 101039_c3sc51023a-7
- 101039_c3sc51023a-8
- 101039_c3sc51023a-9
- 10610278_2013_806809-1 (*)
- 10610278_2013_806809-2 (*)
- 10610278_2013_806809-4 (*)

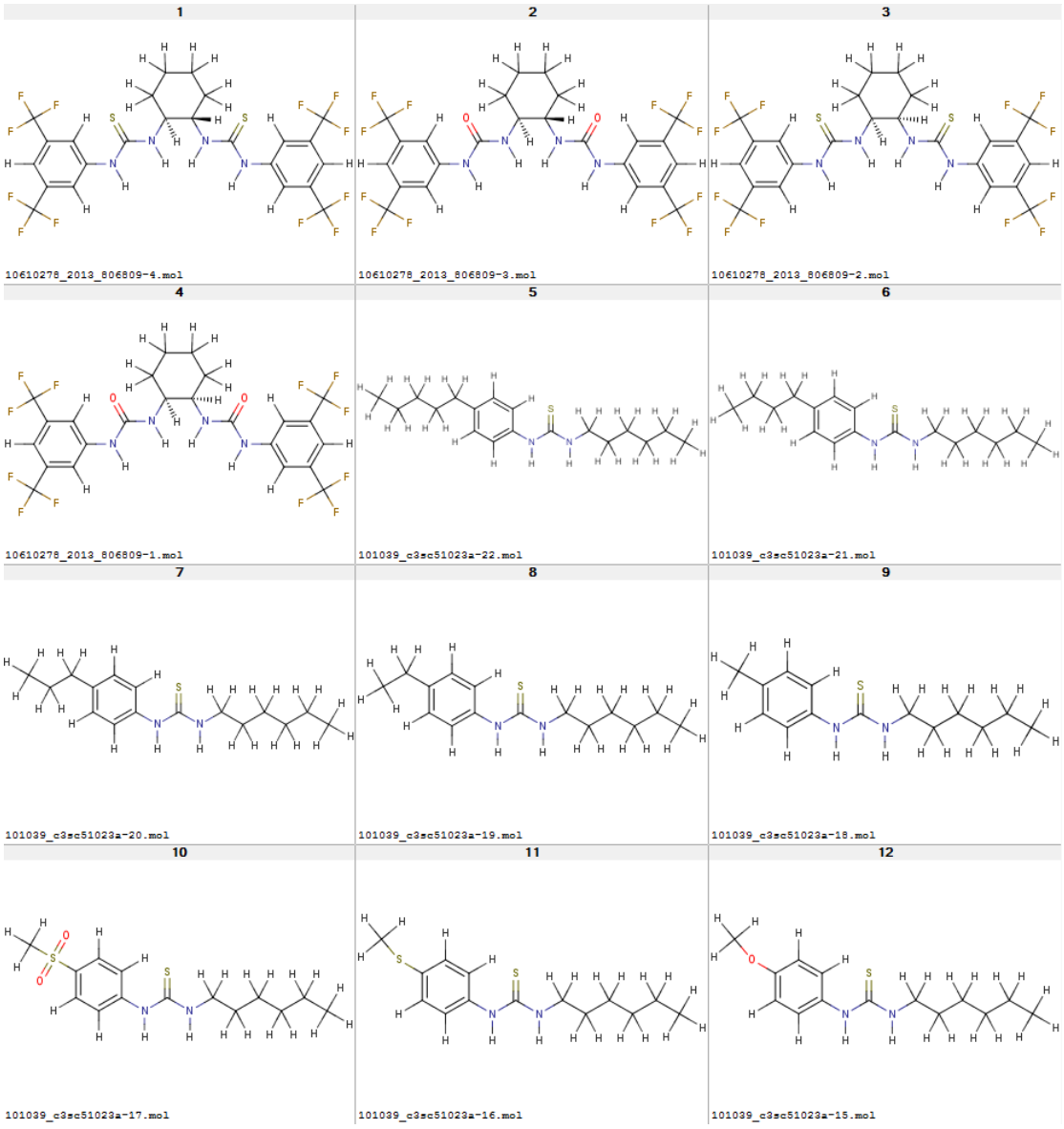


Figure A.12: Structures of Expanded Subset - with and without EC_{50} values - part 1

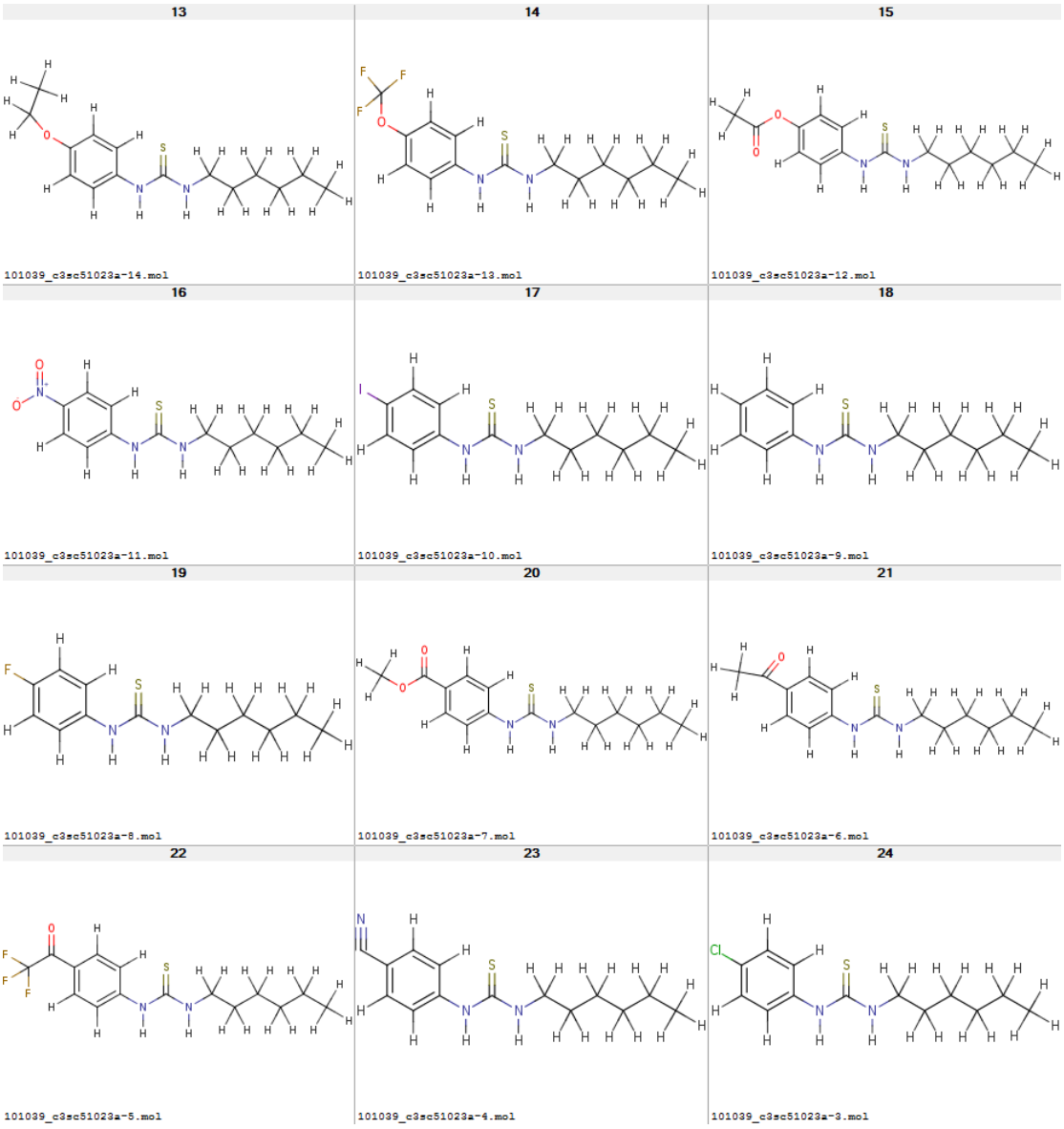


Figure A.13: Structures of Expanded Subset - with and without EC_{50} values - part 2

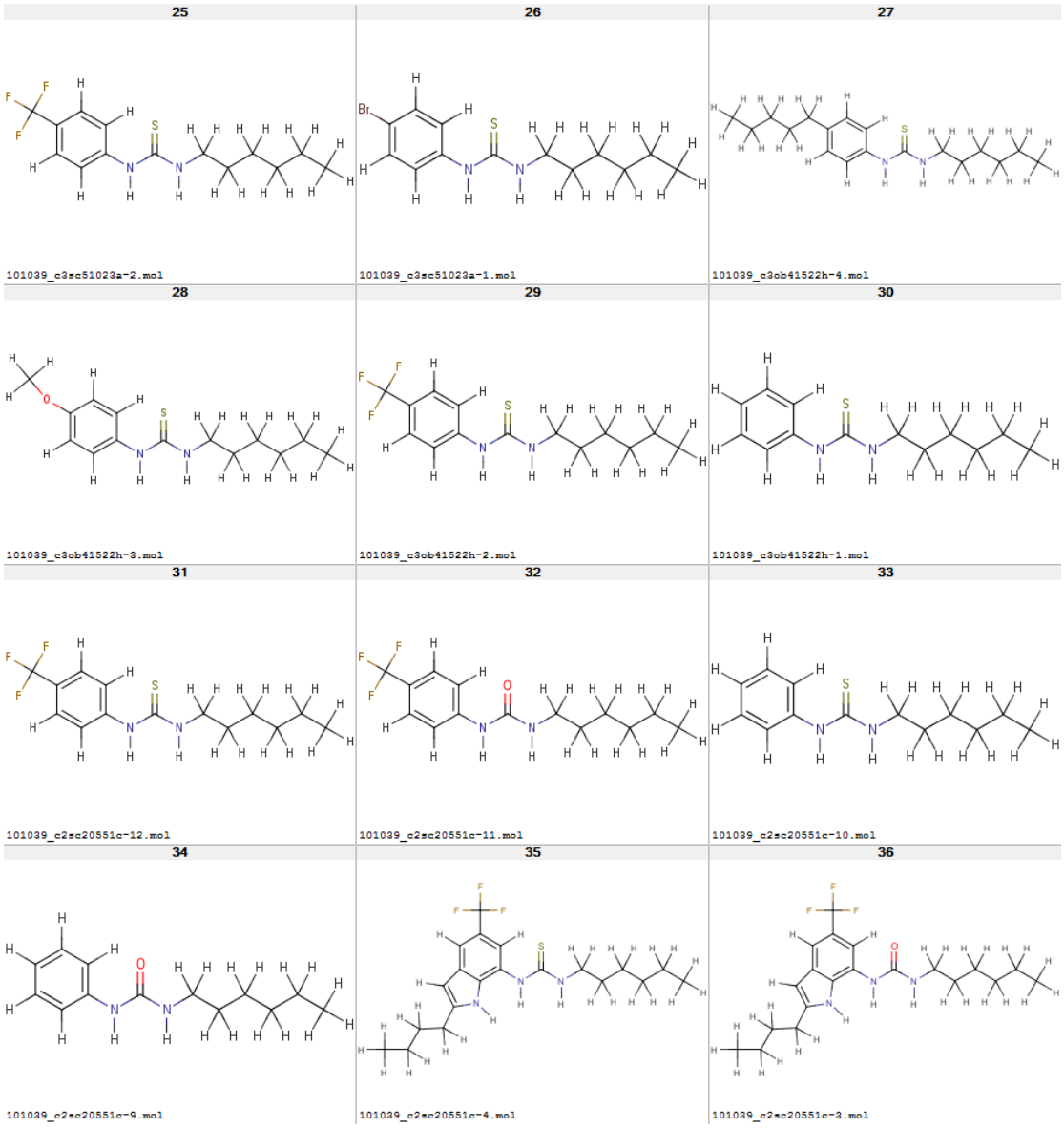


Figure A.14: Structures of Expanded Subset - with and without EC_{50} values - part 3

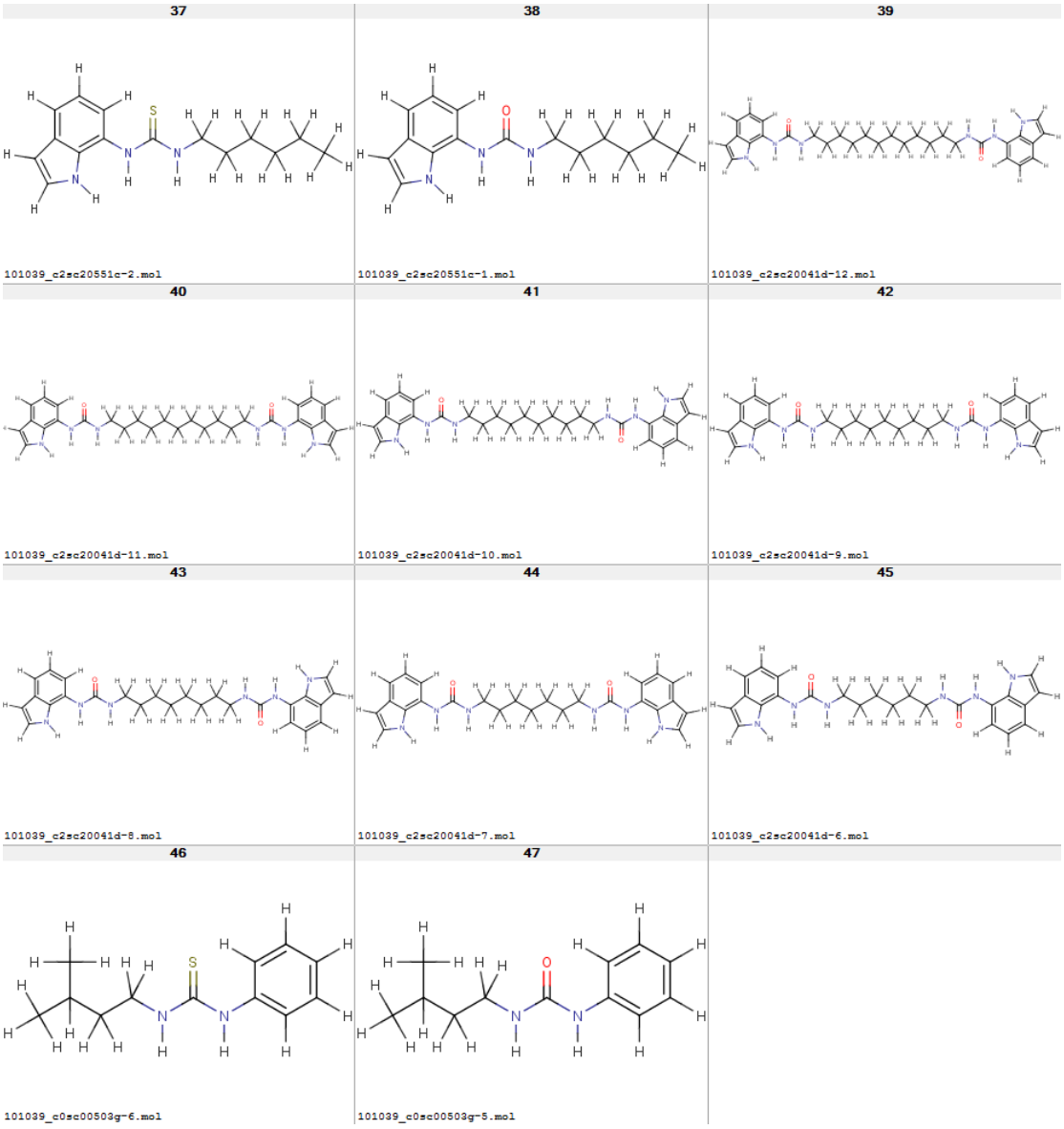


Figure A.15: Structures of Expanded Subset - with and without EC_{50} values - part 4

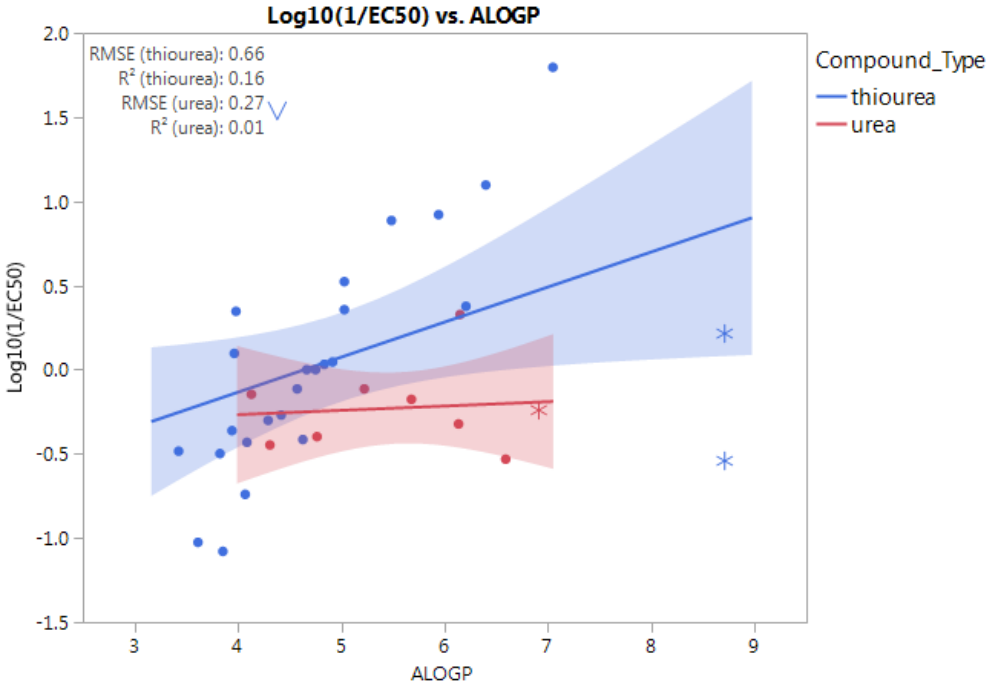


Figure A.16: Linear fit of ALOGP for expanded subset split by compound type

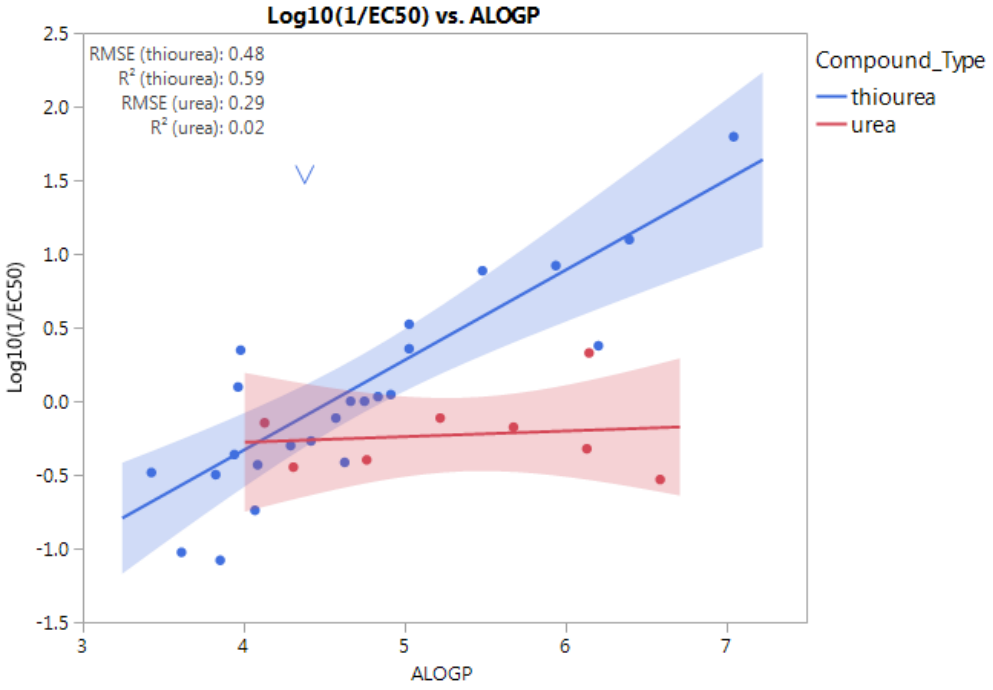


Figure A.17: Linear fit of ALOGP for expanded subset split by compound type, excluding bis functional group

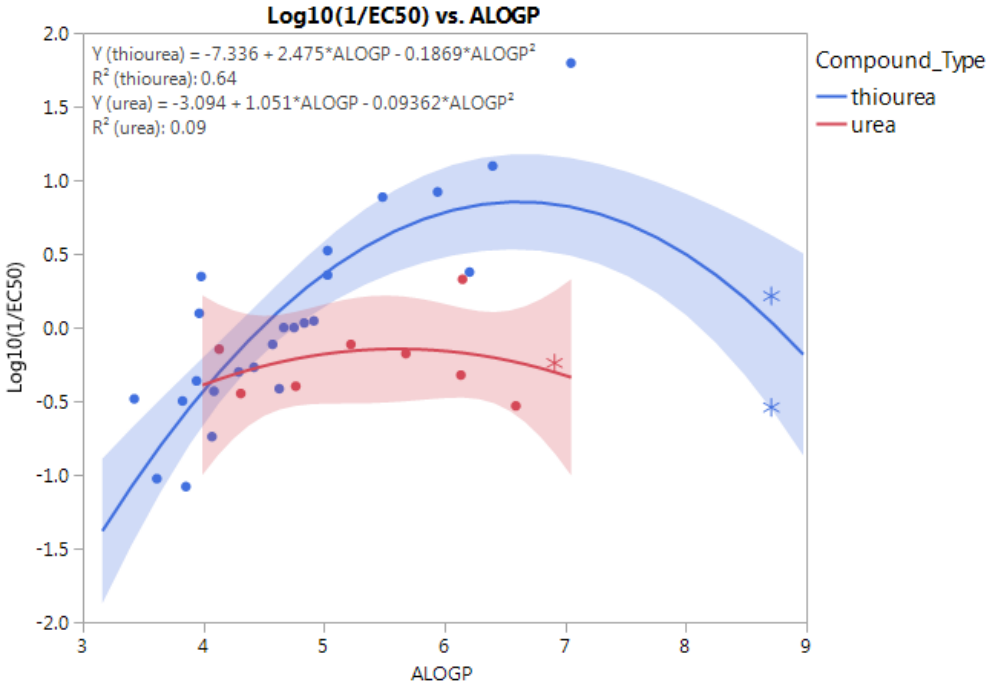


Figure A.18: Quadratic fit of ALOGP for expanded subset split by compound type, excluding outlier (∨)

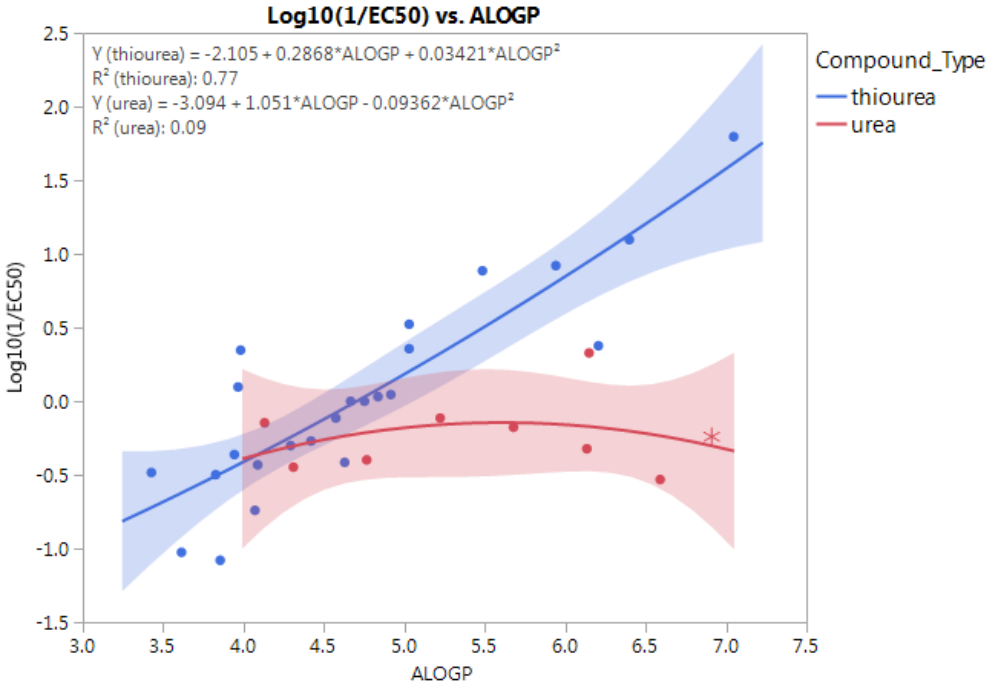


Figure A.19: Quadratic fit of ALOGP for expanded subset split by compound type, excluding bis functional group and outlier (∨)

Appendix B

Tambjamine data

On the following pages are some data from the tambjamine compounds, additional data, including the descriptors for the full dataset, can be found in the ESI of our paper. [140]

Table B.1 : A subset of the data from our tambjamine compounds, includes experimental data and ALOGPs descriptors.

Figure B.1 : Correlations of logP descriptors for the tambjamine dataset.

Table B.1: The tambjamine dataset with Experimental and ALOGPs values

No.	Compound group	NH sub.	Ring sub.	Top sub.	R-type	EC50 (%)	log(1/EC50)	n	RT	ALOGPs	ALOGPs-sq
1	Tambjamine	Me	OMe	NH-Ph	alkyl	0.00719	2.1433	1.19331	10.4	3.08	9.4864
2	Tambjamine	Et	OMe	NH-Ar	alkyl	0.00613	2.2125	1.23492	11	3.74	13.9876
3	Tambjamine	Pr	OMe	NH-Ar	alkyl	0.00699	2.1555	1.2497	11.6	4.17	17.3889
4	Tambjamine	Bu	OMe	NH-Ar	alkyl	0.00779	2.1085	1.32175	12.2	4.63	21.4369
5	Tambjamine	tBu	OMe	NH-Ar	alkyl	0.01036	1.9846	1.29244	11.9	4.72	22.2784
6	Tambjamine -test set	Pen	OMe	NH-Ar	alkyl	0.00951	2.0218	1.2484	12.7	5.02	25.2004
7	Tambjamine	Dec	OMe	NH-Ph	alkyl	0.28839	0.54	0.96494	14.5	7.11	50.5521
8	Tambjamine -test set	OH	OMe	NH-Ar	O-R’	0.06884	1.1622	1.41974	8.8	2.58	6.6564
9	Tambjamine	OMe	OMe	NH-Ar	O-R’	0.01967	1.7062	1.28872	9.8	2.86	8.1796
10	Tambjamine	OEt	OMe	NH-Ar	O-R’	0.01356	1.8677	1.2827	10.5	3.37	11.3569
11	Tambjamine	OCF3	OMe	NH-Ar	O-R’	0.02313	1.6358	1.31434	11	3.76	14.1376
12	Tambjamine	SMe	OMe	NH-Ar	S-R’	0.02603	1.5845	1.2931	10.4	3.2	10.24
13	Tambjamine	F	OMe	NH-Ar	Halogen	0.02077	1.6826	1.17885	9.6	2.92	8.5264
14	Tambjamine	Cl	OMe	NH-Ar	Halogen	0.0155	1.8097	1.26522	10.3	3.49	12.1801
15	Tambjamine	Br	OMe	NH-Ar	Halogen	0.02362	1.6267	1.37262	10.5	3.62	13.1044
16	Tambjamine	I	OMe	NH-Ar	Halogen	0.02212	1.6552	1.28762	10.8	3.76	14.1376
17	Tambjamine	CN	OMe	NH-Ar	Cyanide	0.01674	1.7762	1.47844	9	2.68	7.1824
18	Tambjamine	SO2Me	OMe	NH-Ar	S-R’	0.0494	1.3063	1.59143	8.2	2.11	4.4521
19	Tambjamine	2-picoline	OMe	NH-CH2-Py	N/A	0.19654	0.7065	0.85324		1.88	3.5344
20	Tambjamine	Me	OMe	NH	alkyl	0.34586	0.4611	1.29565	7	1.03	1.0609
21	Tambjamine -test set	Et	OMe	NH	alkyl	0.09208	1.0358	1.07582	7.7	1.55	2.4025

Continued on next page

Table B.1 – *Continued from previous page*

No.	Compound group	NH sub.	Ring sub.	Top sub.	R-type	EC50 (%)	log(1/EC50)	n	RT	ALOGPs	ALOGPs-sq
22	Tambjamine	Pr	OMe	NH	alkyl	0.02736	1.5629	1.03739	8.5	2.03	4.1209
23	Tambjamine	Bu	OMe	NH	alkyl	0.0116	1.9355	0.85699	9.3	2.46	6.0516
24	Tambjamine	Pen	OMe	NH	alkyl	0.00648	2.1884	1.17753	10.2	2.99	8.9401
25	Tambjamine	Hex	OMe	NH	alkyl	0.005	2.3010	1.18867	10.9	3.52	12.3904
26	Tambjamine	Hep	OMe	NH	alkyl	0.00451	2.3458	1.51257	11.5	4.02	16.1604
27	Tambjamine -test set	Oct	OMe	NH	alkyl	0.00312	2.5058	1.0697	12.1	4.79	22.9441
28	Tambjamine	Non	OMe	NH	alkyl	0.0038	2.4202	1.09531	12.6	5.1	26.01
29	Tambjamine	Dec	OMe	NH	alkyl	0.0053	2.2757	1.32819	13.1	5.36	28.7296
30	Tambjamine	Dodec	OMe	NH	alkyl	0.00731	2.1361	1.15405	13.8	6.14	37.6996
31	Tambjamine	Iso	OMe	NH	alkyl	0.01125	1.9488	1.20205	9.2	2.24	5.0176
32	Tambjamine	31	OMe	NH	alkyl	0.00668	2.1752	1.05064	10	2.84	8.0656
33	Tambjamine	C3H6OMe	OMe	NH	alkyl-O-R	0.09767	1.0102	0.96266		1.5	2.25
34	Tambjamine -test set	H	OBn	NH-Ar	alkyl	0.01565	1.8055	1.32058	11.5	4.4	19.36
35	Tambjamine	tBu	OBn	NH-Ar	alkyl	0.01162	1.9348	1.19969	13.1	5.94	35.2836
36	Tambjamine	Pen	OBn	NH-Ar	alkyl	0.01234	1.9087	0.85732		6.46	41.7316
37	Tambjamine	OMe	OBn	NH-Ar	O-R'	0.01334	1.8748	1.44971	11.6	4.38	19.1844
38	Tambjamine	Py	OBn	NH-Py	N/A	0.00878	2.0565	1.42571	11.3	3.3	10.89
39	Tambjamine -test set	Pr	OBn	NH	alkyl	0.01957	1.7084	1.13632	10.7	3.62	13.1044
40	Tambjamine	Pen	OBn	NH	alkyl	0.00968	2.0141	1.74418	11.9	4.49	20.1601
41	Tambjamine	Non	OBn	NH	alkyl	0.00517	2.2865	1.14582		6.07	36.8449
42	Tambjamine	Dec	OBn	NH	alkyl	0.02044	1.6895	0.92909	13.9	6.42	41.2164
43	Tambjamine	Dodec	OBn	NH	alkyl	0.07442	1.1283	0.26473		7.14	50.9796

27/07/2015 12:25
Data Table=Tambjamines_New_numbers_classified_cleaned

Multivariate

Correlations

	Retention time	TorchLiteSlogP	ALOGPs	miLOGP	AC LogP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3	Log P	LogP (AB/LogP v2.0)	LogP (ACD/Labs)
Retention time	1.0000	0.9419	0.9858	0.9784	0.9802	0.9585	0.9421	0.9782	0.9760	0.9730	0.9741	0.9616	0.9643
TorchLiteSlogP	0.9419	1.0000	0.9770	0.9693	0.9762	0.9856	0.9298	0.9799	0.9631	0.9655	0.9825	0.9288	0.9301
ALOGPs	0.9858	0.9770	1.0000	0.9921	0.9941	0.9819	0.9656	0.9916	0.9855	0.9897	0.9896	0.9534	0.9648
miLOGP	0.9784	0.9693	0.9921	1.0000	0.9985	0.9776	0.9784	0.9929	0.9935	0.9975	0.9917	0.9674	0.9824
AC LogP	0.9802	0.9762	0.9941	0.9985	1.0000	0.9823	0.9741	0.9964	0.9924	0.9959	0.9946	0.9639	0.9786
ALOGP	0.9585	0.9856	0.9819	0.9776	0.9823	1.0000	0.9358	0.9826	0.9729	0.9807	0.9907	0.9320	0.9379
MLOGP	0.9421	0.9298	0.9656	0.9784	0.9741	0.9358	1.0000	0.9644	0.9789	0.9750	0.9621	0.9557	0.9759
KOWWIN	0.9782	0.9799	0.9916	0.9929	0.9964	0.9826	0.9644	1.0000	0.9862	0.9895	0.9957	0.9513	0.9699
XLOGP2	0.9760	0.9631	0.9855	0.9935	0.9924	0.9729	0.9789	0.9862	1.0000	0.9921	0.9860	0.9622	0.9804
XLOGP3	0.9730	0.9655	0.9897	0.9975	0.9959	0.9807	0.9750	0.9895	0.9921	1.0000	0.9899	0.9652	0.9780
Log P	0.9741	0.9825	0.9896	0.9917	0.9946	0.9907	0.9621	0.9957	0.9860	0.9899	1.0000	0.9485	0.9654
LogP (AB/LogP v2.0)	0.9616	0.9288	0.9534	0.9674	0.9639	0.9320	0.9557	0.9513	0.9622	0.9652	0.9485	1.0000	0.9789
LogP (ACD/Labs)	0.9643	0.9301	0.9648	0.9824	0.9786	0.9379	0.9759	0.9699	0.9804	0.9780	0.9654	0.9789	1.0000

There are 4 missing values. The correlations are estimated by Pairwise method.

Scatterplot Matrix

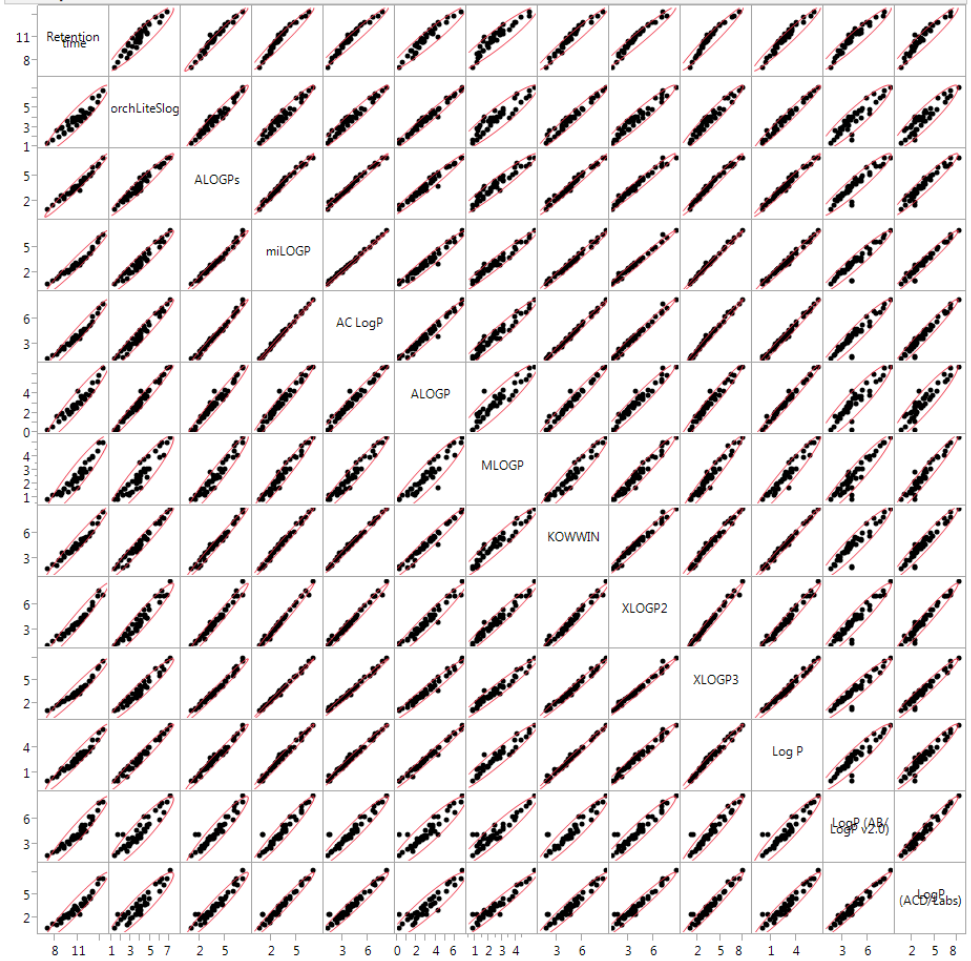


Figure B.1: Correlations of various logP descriptors

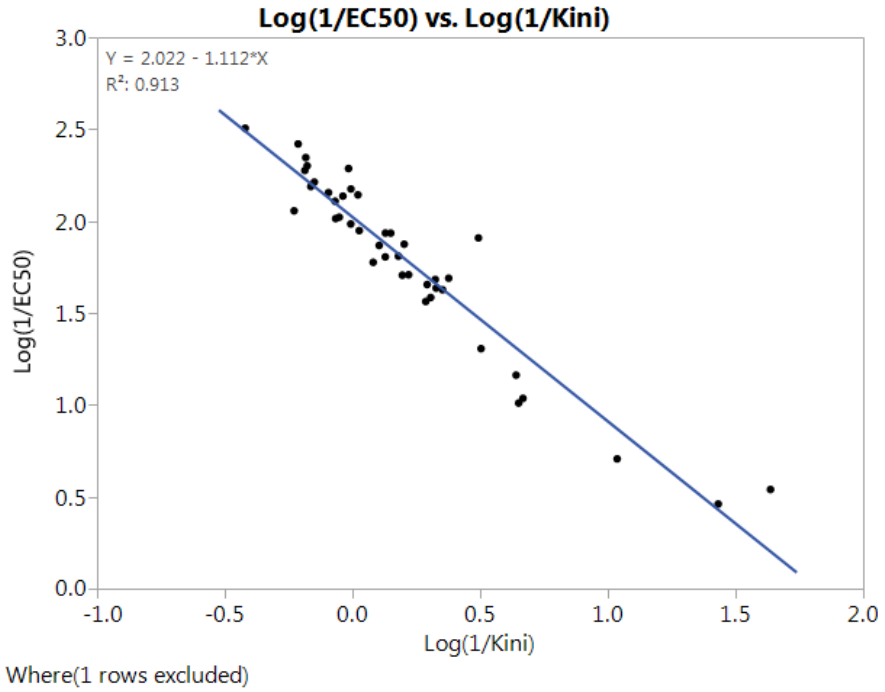


Figure B.2: Correlation of $\log(1/EC_{50})$ and $\log(1/k_{ini})$ excluding compound 43 - used in validation of EC_{50} value for compound 43

Appendix C

BLL Experiment

On the following pages are a number of images showing the Online interface for the BLL experiment in use.

Figures

Figure C.1 : Home Page of the Online Experiment

Figure C.2 : Online experiment running under light conditions
The experiment can be run under dark or light conditions to show the effect of background light.

Figure C.3 : Screenshot of the interface once the experiment has finished
Once the experiment has completed the interface displays a link to a results page.

Figure C.4 : Results page allowing download of files
The files available for download are: a raw image of the fluorescence, an excel file containing the raw data and a image plot of absorbance vs. distance

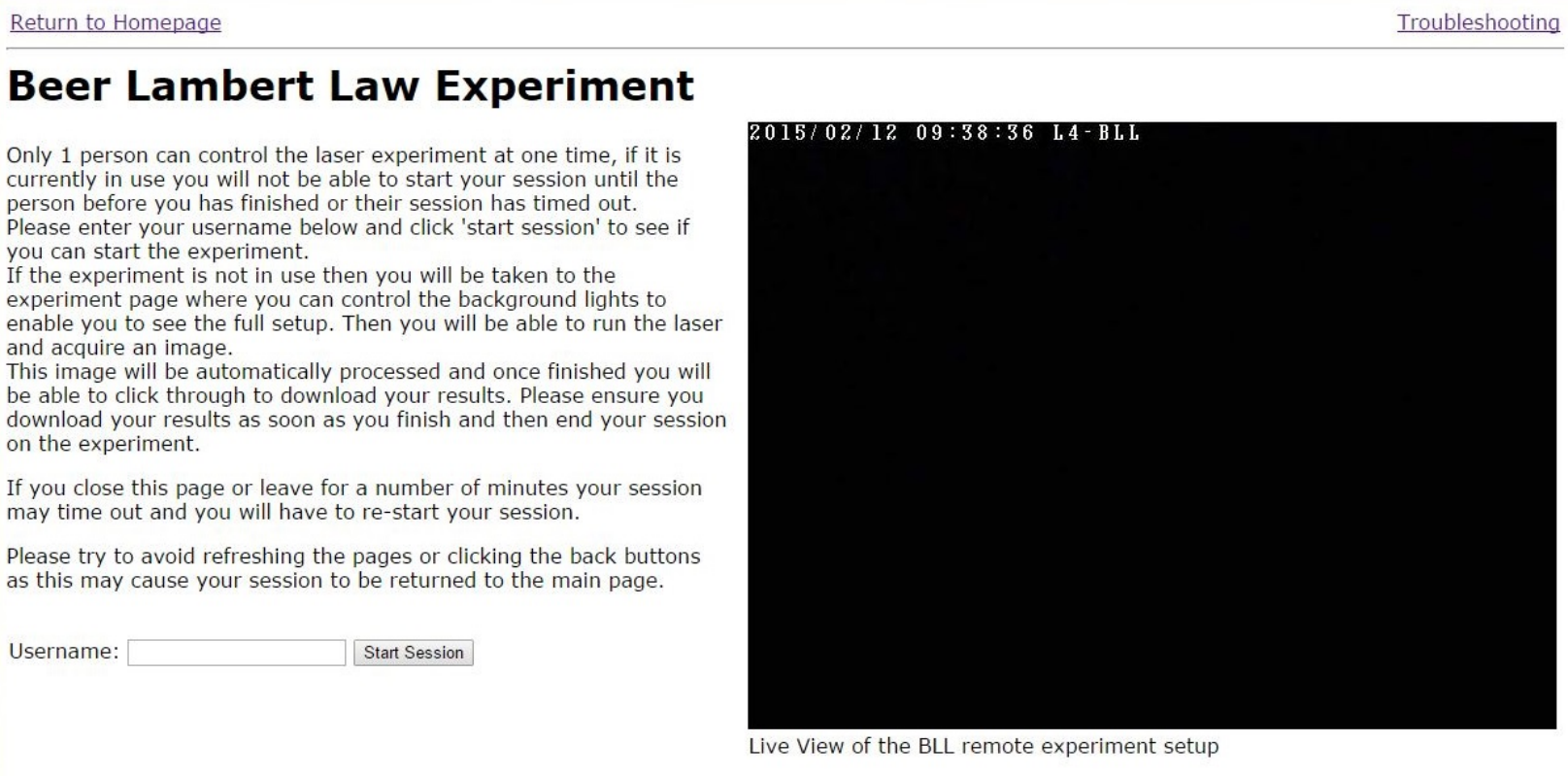


Figure C.1: Home Page of the Online Experiment

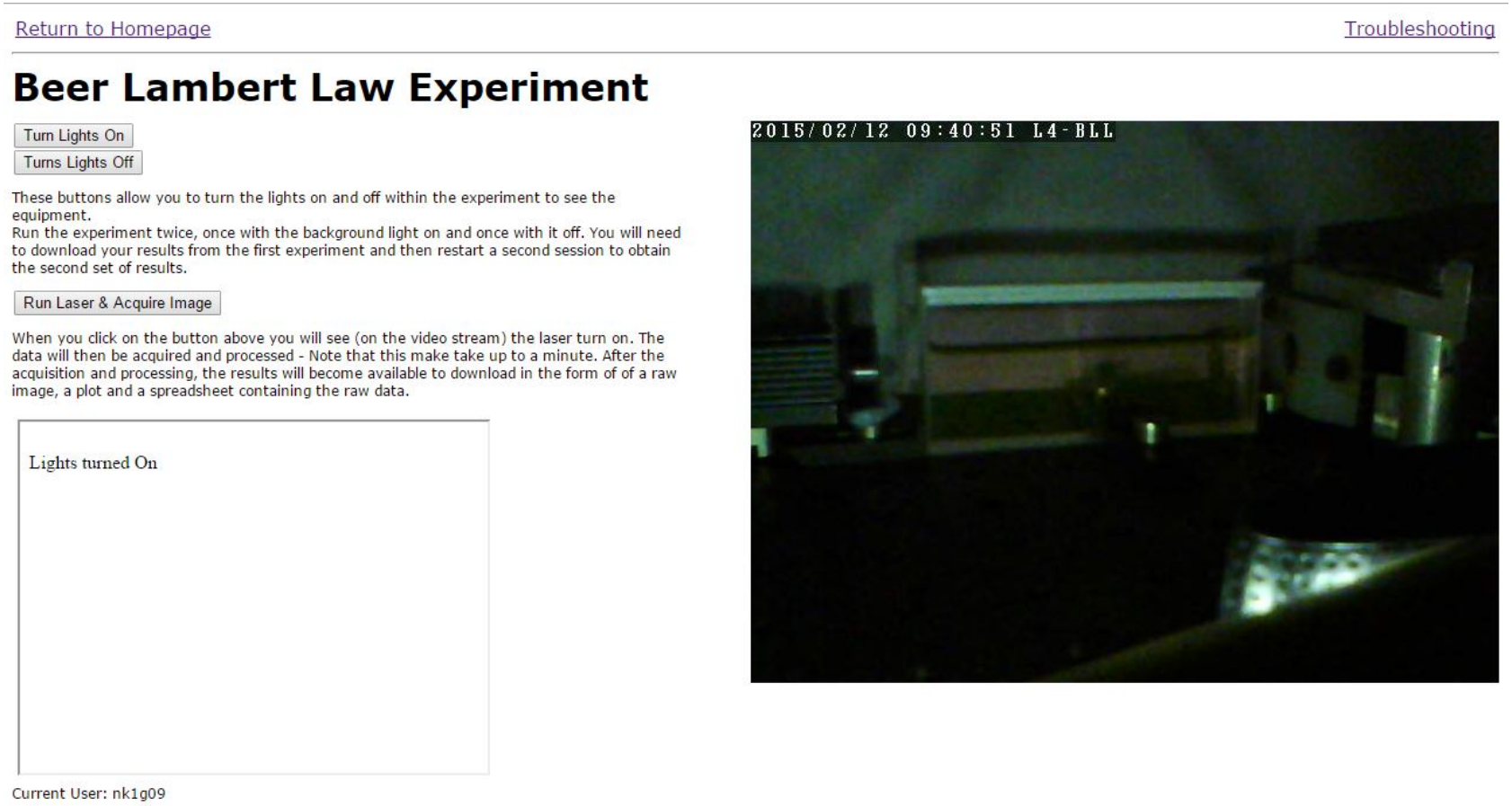


Figure C.2: Online experiment running under light conditions

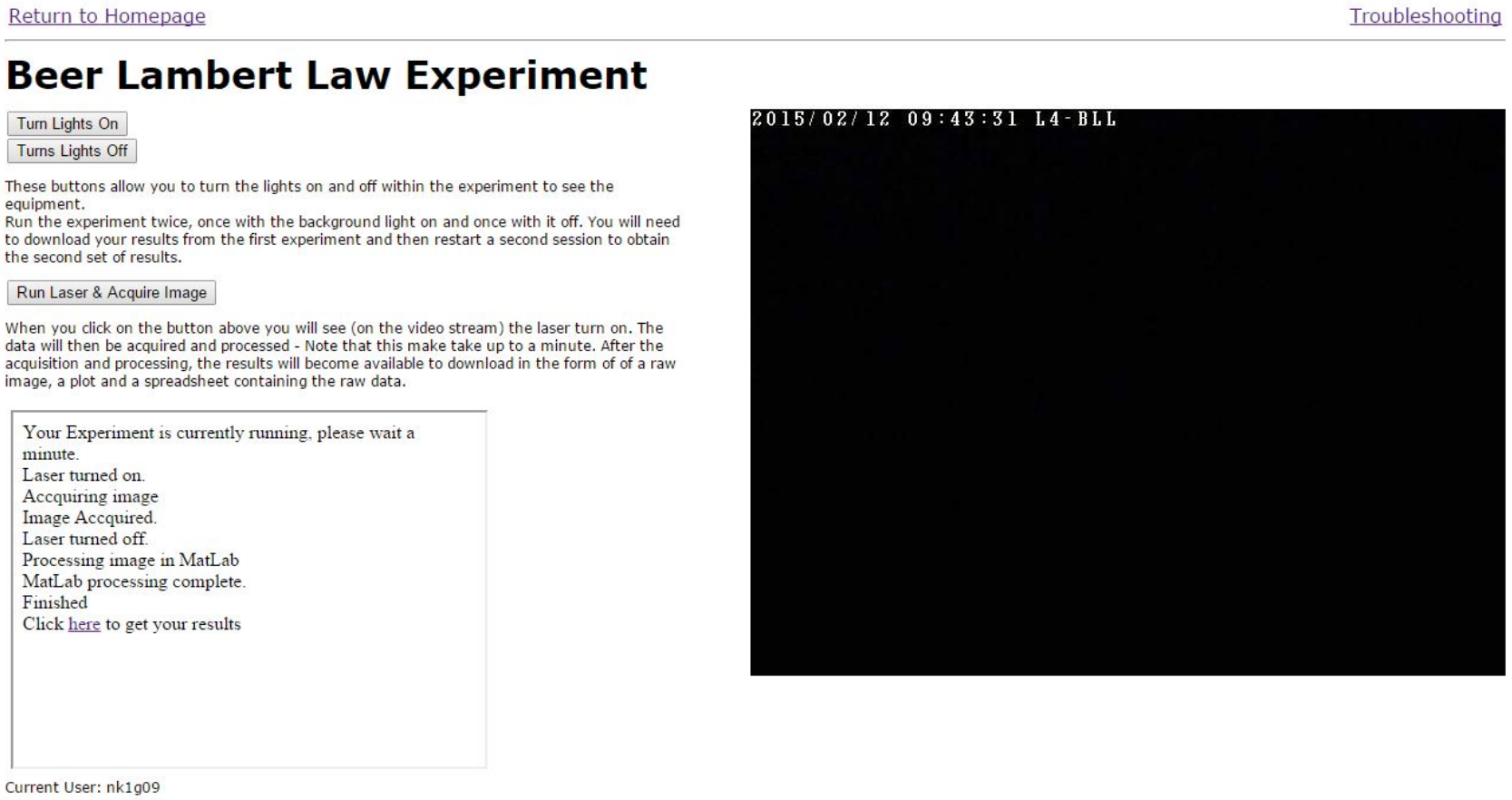


Figure C.3: Screenshot of the interface once the experiment has finished

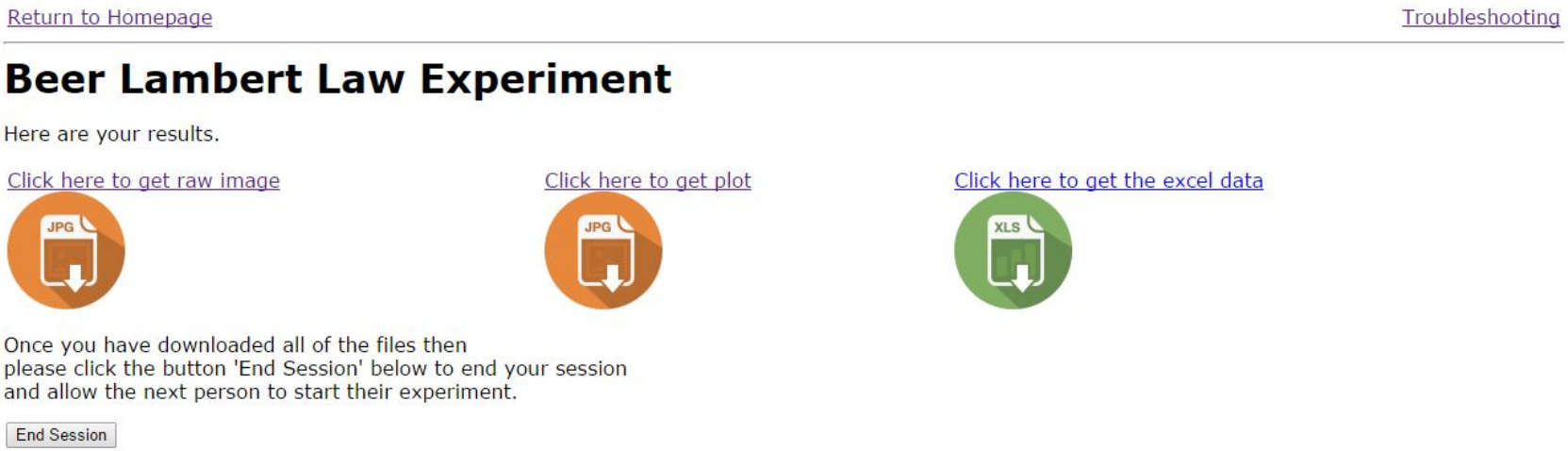


Figure C.4: Results page allowing download of files

Appendix D

ESI Contents

D.1 Chapter 2 - Modelling Anion Transport

Folder: C2_AnionTransport

Files:

Experimental_Chemicals_database.xlsx Initial database of compounds following extraction from papers

PAG_dragon_chemical_values.xlsx Experimental values and DRAGON descriptors for original anion transport compounds, excluding duplicates

Chemicals_mechanisms_cleaned.xlsx Additional data on biological action and mechanisms - only for compounds with measured EC_{50} values

logEC50_with_errors.xlsx Excel file containing experimental errors and $\log(1/EC_{50})$ vs ALOGP plot

PAG_2D_structures.zip Structures of anion transporters, individual .mol files

PAG_2D_structures.sdf structures in .sdf format - all compounds in one file

PAG_structures_image.pdf Structures of anion transporters in image format

c3sc51023a_expanded_compounds.sdf Structures of the subset expanded from paper 10.1039/c3sc1023a - Section 2.4.3

Folder: PCA

2D_Descriptors_PAG_Originalset.csv DRAGON descriptors generated from 2D structures

2D_Structures_PAG_Originalset.sdf 2D structures of anion compounds (original set)

3D_Descriptors_PAG_Originalset.csv DRAGON descriptors generated from 3D structures

3D_Structures_PAG_Originalset.sdf 3D structures of anion compounds (original set)

PCA_R.nb.html RNotebook containing PCA code and plots

Folder: Model_Fits Files showing fit statistics from various model fits

D.2 Chapter 3 - Tambjamine Anion Transporters

Folder: C3_TambAnionTransport

Files:

Compound_structures_2D.zip Tambjamine 2D Compound Structures, individual .mol files

logEC50_with_errors.xlsx Excel file containing experimental errors and $\log(1/EC_{50})$ vs ALOGPs plot

43compounds_DRAGON_descriptors.csv Raw descriptors generated in DRAGON, includes additional synthesised compounds

Tambjamines_dataset_cleaned.csv Dataset from Tambjamine paper [140] contains experimental values, classifiers and descriptors from multiple sources for the 43 tambjamines

4parameter_subset.txt Subset of descriptors used in fit of 4 parameter models.

model_coefficients_CI.xlsx Confidence intervals for selected 2, 3 & 4 parameter models.

model_fit_plots.png Actual vs. predicted plots for selected 2, 3 & 4 parameter models.

Folder: Initial_Models - files for Quesada's models

Test_set_cleaned_descriptors.xlsx 380 descriptors for compounds in test set

Training_set_cleaned_descriptors.xlsx 380 descriptors for compounds in training set

Descriptors_original38_modeldescriptors.csv Descriptors used in initial models, only the original 38 compounds

Descriptor_sources.xlsx Sources for the descriptors generated by Quesada's group

fit_all.png Results of fit all models for the initial modelling

Folder: RCode

linear_code_results.r Code and results for 3 & 4 parameter fits

lme4_code_results.r Code and results for mixed effect linear fits within alkyl R group

Tambjamines_New_numbers_classified_cleaned.csv .csv file used in R code - contains tambjamine descriptors and identifiers

Folder: CombinedQSAR

combinedQSAR_structures.sdf Structures from old_PAG, new_PAG and tambjamine datasets - Section 3.9

combinedQSAR_structures.zip Structures from old_PAG, new_PAG and tambjamine datasets, individual .mol files - Section 3.9

compounds_noduplicates_Alogps_EC50.xlsx Experimental values and AlogPs for all compounds, no duplicates

compounds_noduplicates_cleaned_descriptors.xls DRAGON descriptors for all compounds, no duplicates

classyfire_compounds_noduplicates.xlsx Classyfire classification groups for all compounds, with counts

D.3 Chapter 4 - Data Handling & Visualisation

Folder: C4_DataHandling_Vis

Files:

Example_blog_post_blog3.html Sample blog post in Blog^3 - note links require login

Example_blog_post_LabTrove.html Sample blog post in LabTrove - note links require login

solubility_initial_data.nb.html Example Output from RNotebook

solubility_initial_data.Rmd Example input file to generate RNotebook html output

Folder DataVis

.html and .js files Code for Data Visualisations - see README.txt for the details of each file

D.4 Chapter 6 - Remote Experiments

Folder: C6_RemoteExp

Files:

BeerLambertRawdata.m MATLAB script for image processing

BLL_server.js Node.js Code for running BLL server and experiment

troubleshooting.html Troubleshooting page displayed on the website for the students

Folder: views (pages displayed by BLL experiment)

end.ejs Displayed when ending an experiment

experiment.ejs View during experimental run

index.ejs Homepage for experiment

results.ejs Displaying results following an experiment

session_in_use.ejs Displayed if the session is already in use

Folder: styles (contains .css file for experiment)

Folder: sample_data (contains example data from experiment)

D.5 Chapter 7 - Smart Lab Interaction

Folder: C7_SmartLab

Files:

node_RED_nodes.json code for nodes, taken from node-RED system

Alexa_Interaction.json interaction model from Alexa developer console

Item_catalog.csv spreadsheet containing catlogue of items