Towards ligand design: Quantum Chemical Topology descriptors of heterocyclic compounds and pK\textsubscript{a} prediction from \textit{ab initio} bond lengths

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

2013

Mark Z. Griffiths
School of Chemistry
List of Contents

Preliminary Pages

List of Tables 8
List of Figures 9
List of Commonly used abbreviations 14
Abstract 15
Declaration 16
Copyright Statement 17
Acknowledgements 18

Chapter 1
Introduction and industrial context

1.1 Introduction 19
  1.1.1 Aims of the work 19
1.2 Importance to industry 20
  1.2.1 The Quantum Isostere Database's industrial application 20
  1.2.2 Target identification 21
  1.2.3 Hit identification 22
  1.2.4 Hit to lead 22
  1.2.5 The QID in hit to lead 24
Chapter 2  
Background theory

2.1 The QID  
2.1.1 Introduction of the QID  
2.1.2 SMILES generation  
2.1.3 Conformation generation  
2.1.4 Descriptor calculation  
2.1.5 QID fragment queries  
2.1.6 Score calculation  
2.1.7 Geometry scoring  

2.2 Quantum Chemical Topology  
2.2.1 Introduction to QCT  
2.2.2 Critical points  
2.2.3 Partitioning of the electron density  
2.2.4 Electrostatics  

2.3 Quantum mechanics  
2.3.1 Introduction to quantum mechanics  
2.3.2 The Schrödinger equation  
2.3.3 Level of theory  
2.3.4 Basis sets  

2.4 Bioisosterism and ligand design  
2.4.1 Introduction to bioisosterism  
2.4.2 Definition of bioisosterism  
2.4.3 Non-classical and classical bioisosterism  
2.4.4 2D molecular descriptors  
2.4.5 3D molecular descriptors  
2.4.6 Descriptor space  
2.4.7 Scoring functions  
2.4.8 Alternative methods to bioisosterism  
2.4.9 BROOD  

2.5 pKₐ prediction  
2.5.1 Introduction to pKₐ prediction  
2.5.2 The importance of pKₐ  
2.5.3 pKₐ prediction methods  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.1 Introduction of the QID</td>
<td>24</td>
</tr>
<tr>
<td>2.1.2 SMILES generation</td>
<td>26</td>
</tr>
<tr>
<td>2.1.3 Conformation generation</td>
<td>28</td>
</tr>
<tr>
<td>2.1.4 Descriptor calculation</td>
<td>28</td>
</tr>
<tr>
<td>2.1.5 QID fragment queries</td>
<td>28</td>
</tr>
<tr>
<td>2.1.6 Score calculation</td>
<td>30</td>
</tr>
<tr>
<td>2.1.7 Geometry scoring</td>
<td>30</td>
</tr>
<tr>
<td>2.2.1 Introduction to QCT</td>
<td>32</td>
</tr>
<tr>
<td>2.2.2 Critical points</td>
<td>32</td>
</tr>
<tr>
<td>2.2.3 Partitioning of the electron density</td>
<td>33</td>
</tr>
<tr>
<td>2.2.4 Electrostatics</td>
<td>36</td>
</tr>
<tr>
<td>2.3.1 Introduction to quantum mechanics</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2 The Schrödinger equation</td>
<td>38</td>
</tr>
<tr>
<td>2.3.3 Level of theory</td>
<td>39</td>
</tr>
<tr>
<td>2.3.4 Basis sets</td>
<td>41</td>
</tr>
<tr>
<td>2.4.1 Introduction to bioisosterism</td>
<td>42</td>
</tr>
<tr>
<td>2.4.2 Definition of bioisosterism</td>
<td>44</td>
</tr>
<tr>
<td>2.4.3 Non-classical and classical bioisosterism</td>
<td>45</td>
</tr>
<tr>
<td>2.4.4 2D molecular descriptors</td>
<td>46</td>
</tr>
<tr>
<td>2.4.5 3D molecular descriptors</td>
<td>48</td>
</tr>
<tr>
<td>2.4.6 Descriptor space</td>
<td>50</td>
</tr>
<tr>
<td>2.4.7 Scoring functions</td>
<td>51</td>
</tr>
<tr>
<td>2.4.8 Alternative methods to bioisosterism</td>
<td>53</td>
</tr>
<tr>
<td>2.4.9 BROOD</td>
<td>56</td>
</tr>
<tr>
<td>2.5.1 Introduction to pKₐ prediction</td>
<td>58</td>
</tr>
<tr>
<td>2.5.2 The importance of pKₐ</td>
<td>58</td>
</tr>
<tr>
<td>2.5.3 pKₐ prediction methods</td>
<td>59</td>
</tr>
</tbody>
</table>
Chapter 3
Bioisosterism of heterocyclic fragments

3.1 QCT of cyclic molecules 61
   3.1.1 QCT of ring atoms 61
   3.1.2 Ring atoms and substitutions 62
3.1 Relating Ring Critical Point Properties to ring atom properties 63
   3.1.1 Calculation of RCP and ring atom properties 63
   3.1.2 Non-substituted rings 64
   3.1.3 Substituted ring systems 65
   3.1.4 The effects of a ring’s atomic and RCP properties 66
   3.1.5 The role of ellipticity 67
   3.1.6 Local relationships between ring scaffolds 69
3.2 Ring atom properties 71
   3.2.1 Ring characteristics and ring properties 71
   3.2.2 Discovering all ring features 73
   3.2.3 Effect of substitution on ring properties 77
   3.2.4 Hammett constants 80
   3.2.5 Ring substitution effects 81
3.3 RCP space 84
   3.3.1 Ring characterisation through RCP space 84
   3.3.2 Calculation of RCP and ring atom properties 85
   3.3.3 Rings in RCP space 85
3.4 Practical use of ring properties: a case study 88
   3.4.1 Imidazoleglycerol-phosphate dehydratase 88
   3.4.2 Calculation of properties 89
   3.4.3 RCP properties as descriptors 90
   3.4.4 A combined RCP and ring atom property approach 92
3.5 Conclusions 94
# Chapter 4

## Evaluating the QID and addition of terminal groups

4.1 QID alignment
- 4.1.1 The current QID alignment method
- 4.1.2 Alignment of terminal or collinear fragments
- 4.1.3 Method for evaluating the alignment procedure
- 4.1.4 Evaluation of alignment procedure
- 4.1.5 Alignment of collinear fragments
- 4.1.6 Evaluation of alignment alternative
- 4.1.7 Terminal group alignment

4.2 Bioisosterism of terminal groups
- 4.2.1 Geometry from electrostatics
- 4.2.2 Calculating the multipole moments of geometry change
- 4.2.3 Finding unique electrostatics for a geometry
- 4.2.4 Determining a geometry from electrostatics alone
- 4.2.5 Geometry's effect on atom and molecular electrostatics

4.3 Testing the QID scoring function
- 4.3.1 QID search
- 4.3.2 Search results
- 4.3.3 QID vs BROOD

4.4 Conclusion
Chapter 5
Predicting pK\(_a\) values from an \textit{ab initio} bond length

5.1 Predicting pK\(_a\) values 120
   5.1.1 High Correlation Subsets 120
   5.1.2 Identifying problematic functional groups 121

5.2 Predicting the pK\(_a\) of enols 122
   5.2.1 Establishing a HCS 122
   5.2.2 Non-cyclic enols 124
   5.2.3 Bond length determination 124
   5.2.4 pK\(_a\) prediction from HCSs 128
   5.2.5 Calculation accuracy 130
   5.2.6 Calculation time 132

5.3 Predicting the pK\(_a\) of guanidines 134
   5.3.1 Establishing a HCS 134
   5.3.2 Bond length determination 136
   5.3.3 pK\(_a\) prediction of five biological molecules 137
   5.3.4 Tautomeric effect on pK\(_a\) prediction 139
   5.3.5 Relationships between HCSs 141
   5.3.6 Defining HCSs 144

5.4 Conclusion 148
Chapter 6
Conclusions and future work

6.1 Bioisosterism of heterocyclic fragments 149
6.2 Evaluating the QID and addition of terminal groups 150
6.3 Predicting the pKₐ values from an ab initio bond length 151

References 153

Appendix A - Architecture of the QID tool and its scripts 159
Appendix B - Rotation of atomic multipole moments 166
Appendix C - Data tables 168
List of Tables

**Table 2.1**: Grimm’s Hydride Displacement law.

**Table 3.1**: Non-substituted five-membered unsaturated rings in data set A.

**Table 3.2**: $r^2$ values for PLS models with different selected input and output variables. AP = Atom properties, RP = RCP properties. For clarity the RCP properties are highlighted. The third and fourth entries are the reverse of the first two entries.

**Table 3.3**: Data set of unsaturated ring scaffolds separated into the different substitution sites possible for each ring scaffold.

**Table 3.4**: Change in net charge upon substitution onto a non-substituted ring for eight functional groups split into the three possible atom types for the site of the substitution.

**Table 4.1**: Differences (in Å) between atoms for QID aligned fragments and ethane.

**Table 4.2**: Differences (in Å) between atoms for principal axis aligned fragments and benzene.

**Table 5.1**: Experimental and predicted $pK_a$ values of five tetracyclines.

**Table 5.2**: Absolute difference (right column) in $pK_a$ for five tetracyclines between a prediction error from the 10-point data set (left) and the 14-point data set (middle).

**Table 5.3**: Guanidine derivatives data set.
List of Figures

**Figure 2.1:** Flow chart of the stages of the data generation process for the QID.

**Figure 2.2:** amide linker joined to the rest of the molecule at connection points 1 and 2.

**Figure 2.3:** Ethyl, phenyl, and pyrrole capping groups used to emulate fragment environments.

**Figure 2.4:** User interface for QID tool. This includes the properties that can be selected for the query, the tolerance range for that property, and the importance value assigned to the property. The property selection, tolerance and importance are all changeable by the user.

**Figure 2.5:** Geometric parameters used to calculate geometry scores. Atoms 1 and 4 are the connection points $1^*$ and $2^*$ from Figure 2.2.

**Figure 2.6:** Gradient vector field in the molecular symmetry plane of furan, showing interatomic surfaces, atomic interaction lines and critical points.

**Figure 2.7:** Three-dimensional representation of topological atoms in imidazole with a 0.001 a.u iso-electron density boundary cut-off.

**Figure 3.1:** Normalised monopole moments for the non-hydrogen atoms within imidazole, represented by colours in the rainbow spectrum where violet = most positive and red = most negative.

**Figure 3.2:** Superposition of imidazole and imidazole-2-ol with the change in atomic basin highlighted for the carbon upon substitution of H (blue contour) with OH (red contour).

**Figure 3.3:** The general scaffold for the molecules the completed data set. $R_n$ represents the ring atom where $R = C, O, N, S$ or $P$ and $n$ is the position on the ring. $S_n$ is a substitution on the ring and can be $H$, $OH$, $OMe$, $NH_2$, $NHMe$, $NO_2^-$, $CN$, $F$, $COOH$ or nothing if the adjacent ring atom is already at full valence, e.g. $S$, $O$ or $=N—$. The two dashed bonds can each be either a single or double bond.

**Figure 3.4:** Correlation ($r^2=0.91$) between the observed (actual) RCP electron density and the one predicted by the first PLS model for non-substituted five-membered heterocyclic molecules (data set)
**Figure 3.5:** The six possible substitution patterns in furan. Note that the symmetry of furan would create unwanted duplicates if it were substituted on the two remaining sites.

**Figure 3.6:** PLS results for alcohol, ether, amine and non-substituted five-membered heterocyclic molecules.

**Figure 3.7:** Comparison of RCP properties with two different ring atomic properties (kinetic energy (K) and monopole) for data set B.

**Figure 3.8:** Comparison of four different ring scaffolds with all substitution data points for each scaffold using ring atom monopole moments and RCP properties. The monopole moment and RCP electron density plot is drawn in two forms, the data set divided into ring scaffolds (b) and substitution types (d).

**Figure 3.9:** Net charges for ring atoms of substituted five-membered heterocycles, where the heteroatom at the first position is varied for (a) single heteroatom ring scaffolds, (b) additional nitrogen at the second position, (c) additional nitrogen at the third position, (d) additional nitrogens at second and fifth positions.

**Figure 3.10:** Net charge of ring atoms for five-membered heterocyclics with respect to ring scaffolds and substitution types.

**Figure 3.11:** Ring atom net charges predicted from the sum of a mean net charge change of a given substitution and the net charge of a non-substituted ring.

**Figure 3.12:** Ring atom net charges predicted from the sum of a mean net charge change of a given substitution onto different atom types (sp<sup>3</sup> carbon, sp<sup>2</sup> carbon or N(-H) ) and the net charge of a non-substituted ring. Numbers along the x axis correspond to the rings in Table 3.3.

**Figure 3.13:** Effects of substitution on ring atom properties. Substitutions were chosen to represent the effects of a functional group’s proximity to the ring on the ring atom properties. The bold highlighted profiles are all substituents containing a cyano group.

**Figure 3.14:** (a) RCP space for unsaturated (one or two π-bonds) five-membered heteroatomic rings. (b) Unsaturated five-membered heteroatomic rings in ascending order in RCP space. The descending arrows gather rings that overlap in RCP space. Following the arrows down corresponds to an increase of electron density and Laplacian.
at the RCP relative to the other rings. For example, imidazole rings tend to have a higher electron density and Laplacian at their RCP than pyrazole rings.

**Figure 3.15:** Strongest known inhibitor of IGPD.

**Figure 3.16:** Nine different rings for IGPD inhibitor molecule where R1 denotes the rest of the molecule as shown in Figure 3.15.

**Figure 3.16:** QID distances from best IGPD inhibitor calculated with the electron density, its Laplacian and RCP ellipticity.

**Figure 3.17:** Euclidean distances from best IGPD inhibitor calculated with electron density, its Laplacian and RCP ellipticity (red squares) and distances calculated with QCT ring atom and RCP properties (green triangles) for nine IGPD inhibitors.

**Figure 4.1:** Standard QID orientation of an ester fragment from perspective (a) and (b) aligned by the two connection points.

**Figure 4.2:** Two separate orientations of a phenol fragment with collinear connection points.

**Figure 4.3:** 9 of the 30 fragments selected from the database to test the alignment method. Au and Ag represent connection points where the fragment is connected to the rest of the molecule.

**Figure 4.4:** Collinear molecules subjected to a principal axis alignment.

**Figure 4.5:** Atom number labelled propanoic acid.

**Figure 4.6:** Global axis system highlighted on a terminal phenol fragment.

**Figure 4.7:** Components of the (a) dipole (b) quadrupole moments plotted against the dihedral angle for propanoic acid.

**Figure 4.8:** Components of the (a) dipole (b) quadrupole moments plotted against the dihedral angle for propanamide.

**Figure 4.9:** Dipole moments for carbon (green) carbonyl oxygen (blue) and the alcohol oxygen / amine nitrogen (red) for propanoic acid (dark) and propanamide (light) expressed in a global frame.
**Figure 4.10:** Dipole moments for carbon in propanoic acid expressed in the ALF shown from perspectives (a) and (b).

**Figure 4.11:** Molecular dipole moments for propanoic acid.

**Figure 4.12:** QID scoring of amide replacement fragments derived from only electrostatic descriptors in a global frame compared to the fragment’s QID score with a full range of descriptors.

**Figure 4.13:** Relation of total number of hits for ethanol fragment in QID with respect to the complexity of the search criteria where the complexity of the search criteria increases as the search input tolerances, weightings and number of properties are increased.

**Figure 4.14:** Relation of propanol rank in QID with respect to increasing search criteria complexity for an ethanol fragment search where the complexity of the search criteria increases as the search input tolerances, weightings and number of properties are increased.

**Figure 5.1:** Common skeleton of enol in cyclic (top left) and non-cyclic (top right) configurations with bonds a to e. The stars are carbons in all molecules but one, in which they represent the location of oxygen atoms. Epichlortetracycline (bottom left) with the cyclic enol skeleton of interest highlighted in blue, and Sethoxydim (bottom right).

**Figure 5.2:** Relation between experimental pKₐ and *ab initio* bond lengths for cyclic enols for the five different bonds of the enol group (a to e in Figure 5.1).

**Figure 5.3:** The self-correcting process of HCS formation induced by conformational variation. Flipping the amide group from the high energy position (right, green) to the low position (left, red) forms the desired O-H...N intramolecular hydrogen bond, which establishes the full HCS.

**Figure 5.4:** Optimisation path of Isochlorotetracycline.

**Figure 5.5:** Original 15-data point HCS (red) distorted by applying noise to each data point, leading to “the worst possible scenario” HCS (black).

**Figure 5.6:** Predicted pKₐ values from the C=C bond length for each iteration of the geometry optimisation for three selected tetracyclines (blue). The experimental pKₐ value is marked in red.
Figure 5.7: The five tautomers of the guanidine functional group in the monosubstituted derivative. Panel 1 shows the labelling scheme for bond lengths a to e used throughout.

Figure 5.8: Correlation of ab initio bond lengths with experimental pKₐ values for guanidines in the first tautomeric form. Panel (d) represents both N-H bonds as they are almost indistinguishable. The bonds are labelled in one panel of Figure 5.7 for clarity. Panel (b) includes the alternative experimental pKₐ value for guanidine (13.6 taken from Foye's) given by the red triangle.

Figure 5.9: Five biological molecules used to test the guanidine HCS.

Figure 5.10: Correlation of the guanidine data set including five added biological molecules. Debrisoquine (red triangle) is highlighted as the only molecule with the guanidine functional group connected to the rest of the molecule through a tertiary amine. The four other biological molecules added to the set are highlighted in black (circles).

Figure 5.11: Correlation of ab initio bond lengths to experimental pKₐ values for the bond length of interest for each tautomers 1 to 5 in panels (a) to (e), respectively. The C=N bond appears in a different place depending on the tautomer, as shown in Figure 5.7.

Figure 5.12: Superimposed correlations of the C=N bond to experimental pKₐ for five guanidine tautomers

Figure 5.13: Separation into subsets of the HCS of the third tautomer (3) of guanidine derivatives.

Figure 5.14: Reduction of data points from subset C of guanidine tautomer 3 derivatives increasing strictness of subset definition.
List of Commonly use abbreviations

ADMET - Absorption Diffusion Metabolism Excretion Toxicity
ALF - Atomic Local Frame
BCP - Bond Critical Point
GSK - GlaxoSmithKlyne
IGPD - Imidazoglycerol-Phosphate Dehydratase
MAE - Mean Absolute Error
PLS - Partial Least Squares Regression
QCT - Quantum Chemical Topology
QID - Quantum Isostere Database
QSAR - Quantitative Structure Activity Relationship
RCP - Ring Critical Point
RMSE - Root Mean Squared Error
WDI - World Drug Index
Abstract

Bioisosterism is a field that is widely applied to biological molecules, including drugs and agrochemicals. Bioisosterism is the replacement of an active fragment in a molecule with another fragment similar in activity. The replacement is designed to alter the behavior of the molecule in its target environment.

In previous work a bioisostere database called the Quantum Isostere Database (QID) was built out of descriptors derived from the theory of Quantum Chemical Topology (QCT). The current work aims to expand the existing QID to include ring fragments. A series of rings were characterised by QCT properties taken from the ring. It was found that four features of a ring each independently have a systematic effect on the ring’s properties. In other words, each of the characteristics of a ring can be changed and have the same effect on the ring’s properties irrespective of the other ring features. The rings were also characterised using the three QCT properties taken from a point within the ring. The three properties established a space where rings were positioned based on their respective three properties. The positions of the rings showed that the space was able to discern between ring types, and that the features of a ring could be predicted if only its three properties were known.

To improve the QID the alignment method and scoring were tested. The alignment procedure is unable to correctly align collinear fragments. Therefore, a principal axis alignment procedure was successfully employed to align collinear fragments. For terminal fragments an alternative alignment procedure was proposed to account for the increased rotational freedom. A global axis system meant that the direction dependent properties for all fragments were expressed in this new axis system. This idea was extended further and it was found that the geometry of a molecule was imprinted in the electrostatics when they were expressed in the global axis system.

Finally, a pKₐ prediction method which correlates a single ab initio bond length was tested against two data sets (enols and guanidines). The method relies on subsets to form, where molecules within a subset share a chemical or structural commonality. These subsets were able to distinguish between the five tautomeric forms for the guanidines and different conformations for the enols. All predictions were within 1.0 pKₐ units of experimental values.
Declaration

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

Mark Z. Griffiths
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf, in any relevant thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgments

I would like to thank everyone who I have met throughout the course of this doctorate, whether our time together was scientific in nature or just a hootenanny, it has made the last few years both productive and enjoyable. I appreciate the time some of you took to help me with this endeavour. In particular I would like to thank my supervisors Paul Popelier and Nathan Kidley for the support and advice you have given. To my colleagues both at Syngenta and the University of Manchester for all the discussions we have had, fruitful or otherwise. A special thank you to Mike Devereux and Alex Harding who continued to give valuable feedback throughout. Thank you to my family and friends for your constant support.

And finally to the reader, for showing an interest in the subject I love and have been heavily immersed in for the last four years.

Köszönöm Mindenkinek

“A bosszú hidegen tálatva a legjobb” – Régi közmondás
Chapter 1
Introduction and industrial context

1.1 Introduction

1.1.1 Aims of the work

The aim of this work is to expand and improve the Quantum Isostere Database (QID), which is a tool for ligand design. This would result in a tool able to aid chemists in improving potential ligands in the pharmaceutical or agrochemical industries. Currently the QID is only able to improve molecules through suggesting linker fragment replacements. Linkers are fragments that connect any two regions of a molecule. This work aims to increase the range of application to rings and terminal groups. Rings were targeted because they are found within many ligand molecules and are frequently responsible for the molecule’s activity. Extending the QID to include terminal fragments would result in an increase in the range of fragments that the QID could suggest replacements for. Because the QID is currently structured towards linker fragments, work is done to establish a method of incorporating the rings and terminals into the QID. Once the QID is extended to include rings and terminals it will be possible to suggest replacements for any fragment of any molecule. In other words, a molecule can be fragmented into any combination of fragments and the QID will be able to suggest a replacement for any of the fragments.

The QID tool is further expanded by inclusion of a pKₐ prediction tool. Within the scope of ligand design the pKₐ of a molecule can be used as an indicator as to the appropriateness of the ligand to the target. Successfully predicting a pKₐ removes the need for experimental verification and therefore reduces time and cost through use of a computational pKₐ prediction tool. Therefore, a pKₐ predictor method is developed that can lie within the QID tool where the predictor and the database together form a ligand design tool with increased capabilities. In other words, the QID and pKₐ prediction combined form a single tool that can perform both functions.

Finally the current state of the QID is evaluated. The QID tool can be improved through analysis of its structure and method. This evaluation goes beyond verification of result hits and looks at how the QID returns hit fragments. This is done for areas such as the most suitable method of fragment alignment and method by which fragment hits
are returned and ordered. Having the best possible method for comparing the
fragments is vital because it is the order of the fragments that determines which ones
are the best replacements and should be considered for the ligand of interest. Naturally
improving the QID by eliminating the flaws makes it a stronger tool and better able to predict more suitable results.

1.2 Importance to industry
1.2.1 The Quantum Isostere Database’s industrial application

The QID tool is targeted at the lead optimisation stage of ligand design. Improving the efficiency of any stage of the ligand design process can be hugely beneficial because the costs are so high as well as the number of failed candidates. For a drug candidate to successfully go through trials and reach the market it is estimated to cost in around $1.8 billion\(^1\). A drug will take approximately twelve to thirteen years to get to market from its first inception\(^2\). For every 5000 to 10,000 compounds chosen at the beginning of development only one will make it through to production and be introduced to the market\(^3\). Therefore, streamlining a stage by either reducing the time required or improving the results of a stage by having a higher success rate of compounds is desirable. Although the QID is targeted for the hit to lead stage, all stages of the ligand discovery process are briefly discussed below for drug design to show where the QID fits into the context of ligand design. Note that for agrochemical ligand design the only difference from pharmaceuticals is for in vivo and biological testing. In other words, all stages up to and including hit to lead are the same for pharmaceuticals or agrochemicals. Biological testing is only applied after the ligand design process and confirms the success or failure of the designed ligands, i.e. biological testing occurs after the ligand has already been designed. If unsuccessful in the biological testing stage then ligands are returned to the lead optimisation stage and this stage is repeated. Therefore, biological testing is not discussed because it occurs after the ligand design process. The three stages of ligand design are discussed below in brief, as there is a huge literature available for each.
1.2.2 Target identification

The first stage in ligand design is to identify the target system. Target systems are identified by focusing on the function of a biological system (anything from a human body down to a single cell’s function) and identifying whether the system must be inhibited or activated. Various pathways control and regulate these functions and the pathways are controlled by protein actions. The objective is to either block the pathway by targeting certain proteins actions and inhibiting them, or to accelerate a pathway by activating the desired protein action. For example, prevention of cancer cell growth through inhibition of the proteins responsible for cell growth. Once the target protein is known the protein’s method of action is targeted by identifying the sites of the protein that are responsible for its action. There are numerous techniques that are able to identify the interactions a ligand is required to make with the protein for a successful action. A recent review highlights four key techniques and their recent advances in identifying ligand target sites. The four methods are: genetic assays, chemical proteomics, expression profiling and bioinformatics. Genetic assays find targets through genetic screening. Mutagens are introduced into an organism or cell. The resulting mutagens are screened for specific phenotypes such as alterations to appearance, behaviour or growth. The responsible genes are then identified from these phenotypes. Genetic assays have a short experimental time and small cost of materials but this method only works if essential gene targets are affected by the mutagens. Chemical proteomics identify targets by target specific small molecules. These chemical probes are each designed to affect specific families of enzymes. Therefore, the enzyme family can be identified through these probes. While chemical proteomics can provide information with respect to drug binding specificity, it is difficult to implement in a high-throughput setting. Expression profiling finds targets using the premise that if the gene responsible for a target’s activity is removed then the target acts as if it is inhibited. Expression profiling is able to identify the mechanism of the drug but require extensive data analysis. Bioinformatic approaches to target identification rely on computational models for predicting ligand target interactions and determining the target’s mechanisms. Most bioinformatic methods attempt this through knowledge of known ligand protein interactions and known protein structures, sequences and actions. Models are derived from the known ligand and protein data. Bioinformatic approaches are computational and therefore have a reduced cost relative to other methods.
However, the models require experimental data to function and the quality of the experimental data available can determine the strength of the model.

1.2.3 Hit identification

Once the target and its mode of action are identified a lead profile generated, which is based on the data of the target. The lead profile can be a collection of properties (physical, chemical or biological) that are deemed important for successful ligands. A set of hit compounds are generated by high throughput methods, such as databases of small molecules and their respective properties. From this set only the compounds that are confirmed as potential lead compounds progress in the ligand design process. A recent review identified three categories of techniques for hit discovery and confirmation\textsuperscript{16}. These three techniques are biochemical, affinity, and \textit{in silico}. Typically, biochemical methods involve \textit{in vitro} High Throughput Screening of up to a million compounds\textsuperscript{17}. A cut off in the allowed activity filters out any molecules that are too inactive. While biochemical methods can test a large number of molecules, they tend to return relatively large numbers of false positives. Affinity methods rely on NMR\textsuperscript{18}, Mass spectrometry\textsuperscript{19} and X-ray crystallography\textsuperscript{20} to detect molecule affinities to binding sites based on the ligand-target complex. Affinity methods suffer less from false positives than biological methods but have a larger number of false negatives. \textit{In silico} methods primarily utilise protein docking programs to predict if a hit is suitable\textsuperscript{21,22}. Protein docking methods rely on scores to determine the affinity between ligand and target. These scores are generated from any number of criteria including, hydrogen bond formation, steric clashes and interaction energies. \textit{In silico} methods reduce the number of compounds for screening before experimentation. However, experimentation is still required to confirm the molecules within the reduced set as potential lead candidates.

1.2.4 Hit to lead

The set of hits that are identified are turned to lead molecules through the hit to lead process before the leads advance to biological testing (e.g. pre-clinical and clinical trials). Methods of aiding the hit into lead process are discussed in more detail in chapter 2 with respect to bioisosterism and just a general overview is given here. When a hit candidate is found the molecule is refined by any of a large number of methods\textsuperscript{23}, which are techniques designed to improve and fine-tune the hit. Refinements to the hit are alterations made to the hit such as replacement of atoms or functional groups of the
hit with alternatives, which are guided by hit to lead methods. A tool that is able to successfully refine molecules and turn hits into leads greatly reduces the time and cost of the hit to lead process. By suggesting suitable hits, or alterations to hits, a method improves the chances of a given hit becoming a lead. This also means that hits that are not deemed suitable by the tool can be filtered out without requiring any form of experimental testing. The difficulties in hit to lead generation occur from attempting to predict and account for of the lead’s behaviour within the target biological environment. Exposing the lead to the target biological environment introduces a myriad of variables that are difficult to predict and account for. These variables include interactions with other molecules in the biological system, variations in the constitution of the solution and variance in the environments, for example, local pH differences within the human body. Therefore, hit to lead methods require much more specific, and focused, approaches that can encapsulate the mentioned difficulties and successfully turn a hit into a lead.

This stage of the ligand design process takes 1 to 3 years and costs approximately five million dollars\(^3\). Therefore a program that can aid in this stage is highly desirable and profitable, greatly increasing the speed of the whole process. This stage focuses more heavily on fine tuning the molecule to achieve the necessary activity in a biological environment. An accurate method of modelling the properties that are key in binding and activity of ligands with targets will no doubt lead to better predictions or modifications for lead compounds. Improved predictions result in better ligand molecules which have favourable properties such as a reduced toxicity or increased potency. Therefore, hit to lead design tools can increase the chances of a lead progressing through biological testing and into market by improving the hit. However, hit to lead design tools are not limited to increasing the chances of hits becoming leads and then progressing to market. Molecules that already exist on the market can be altered by replacing fragments of the molecule to create a new, but similar, molecule. This new molecule will ideally have an activity equal to or greater than the existing molecule. This feature is referred to as patent busting, and allows for a company to navigate around intellectual property and market their own molecules to compete with the existing molecules where they would otherwise not be able to compete due to patent restrictions. However, for successful patent busting the new molecule must differ enough from the patent protected molecule to circumvent the patent, which is achieved by novel fragment replacements. In summary, to successfully turn a hit into a lead the
hit must transform from one of millions of small molecules suggested by hit identification in to a ligand that has the desired effect on its target without any significant unwanted effects.

1.2.5 The QID in hit to lead

The QID tool is targeted at the hit to lead stage of the ligand design process. This largest hurdle with hit to lead design is the difficulty of modifying a molecule's structure to produce a lead with the correct activity within the target biological system. It is the complexity of the biological environment and its interactions with the lead molecule that cause this difficulty. The most difficult factor to account for is the activity of the lead with the target. There are a large number different features of a ligand that leads to its successful binding and action on a target, such as the lead’s 3D and electronic structures. Both the 3D and electronic structures can be complex, especially the electronic structure of a molecule, and both determine the ligand’s binding to the target and the ligand’s action upon it. Therefore, alterations to a hit in the hit to lead stage can have very specific requirements to balance all the competing features of the biological system while controlling the properties of the ligand. Such alterations can be made easier with more knowledge of the 3D and electronic structures. The QID relies on the principal that a more complete description of the ligand’s properties allows for a better understanding of the ligand and hence suggest better replacements. Such a rigorous treatment not only improves the chances of finding suitable replacements but allows for novel replacements to be found. By using this detailed description of a molecule’s properties replacements can be found that would not otherwise be found by simplistic methods.
Chapter 2

Background theory

2.1 The QID

2.2.1 Introduction of the QID

As discussed in chapter 1, ligand design is a complex problem. The QID is introduced to tackle this problem. In this chapter the function and structure of the QID are described. This is followed by a description of how the properties of the QID are calculated within the framework of Quantum Chemical Topology (QCT) and the quantum chemistry that QCT is built upon. This is followed by a discussion into bioisostersim and the core concepts that dictate bioisosterism to show the context in which the QCT properties are being used. Other bioisosteric replacement methods are discussed as well as alternative methods to bioisosterism. pK_a is then introduced to provide the basic concepts required for understanding pK_a prediction methods. Finally, the value of QID over the other bioisostere methods is discussed.

The Quantum Isostere Database (QID) is a webtool for ligand design, in particular the alteration and improvement of lead candidates based on changes to fragments within the molecule. This is achieved by suggesting alternative fragments with similar properties to the target fragment. The QID uses a database of fragments stored with their properties, which are derived from quantum chemistry. The properties are calculated from QCT. These properties are then compared with the properties of the query fragment and a suitable replacement recommended via a scoring function.

Figure 2.1 summarises all the processes involved with generating the data for the fragments up to their entry into the database. Each stage of Figure 2.1 will be discussed in detail below. Once the data (or descriptors) have been generated for a fragment the fragment and its associated descriptors are stored in the database. The QID builds on the belief that a high level of descriptor would be more suitable in characterising a fragment because it can capture more information about the fragment’s properties. Increased detail in the characterisation of a fragment allows for more detailed and refined comparisons to be made between fragments.
2.1.2 SMILES generation

The first stage of data generation is the fragmentation of the World drug index (WDI). The fragmentation refers to the process of extracting fragments of molecules from drugs within the WDI. These fragments pose as a logical starting point for populating a ligand design database because the WDI consists of drug molecules. To obtain the fragments molecules were cut using a set of cutting rules for molecule fragmentation. The rules ensured that cyclic structures were left intact, and π bonds were not cleaved. The molecules of the WDI were interpreted through a character string representation known as SMILES. SMILES is Simple Molecular Input Lines System, and allows for molecules to be encoded as character strings consisting of ASCII characters to represent the atoms and bonds of a molecule. For example, the SMILES string for acetic acid is CC(=O)O. Through the SMILES strings the molecules were safely cut and the resulting fragments stored as a set of SMILES strings. Any duplicate fragments were discarded.

The second stage of data generation caps the fragments with one of three functional groups (ethyl, benzyl or pyridine). The capping process is shown through SMILES in equation 2.1 and through Figure 2.2 where 1* and 2* are the first and second cleaved bonds and are referred to as the connection points. Because the fragments will lie within a molecule the connection points determine the fragments orientation within the molecule. Figure 2.3 shows the three possible capping groups that emulate the local environment the fragment will be calculated for. They simulate the corresponding environments for each capping group and were determined in a previous study24. For example, if the amide fragment in Figure 2.2 is replacing a fragment in a molecule where
it is connected to an aromatic group at 1* and alkyl group at 2* then the capping groups are benzene and ethyl, respectively. The pyridine group was selected to simulate a bond to a heteroatom. An amide fragment capped by two ethyl groups is treated as an entirely different fragment to the amide given in the example above and therefore is treated as such throughout. This greatly reduces the number the calculations required generating the descriptors for fragments due to the reduction of available environments to just three.

\[
[1^*]C(=O)N[2^*] \rightarrow [C2H5]C(=O)N[C2H5] \quad (2.1)
\]

**Figure 2.2:** amide linker joined to the rest of the molecule at connection points 1 and 2.

The same principle of capping groups is applied to fragments when queries are made. If the connection point of the target fragment is joined to the rest of the molecule by an alkyl environment then an ethyl capping group is attached to the fragment to emulate this environment. If the fragment is connected to an aromatic atom type or nitrogen a phenyl ring or pyrrole cap the fragment, respectively. Other environments are not included because fragment replacements at these sites would be less desirable and many are uncommon in ligands.

**Figure 2.3:** Ethyl, phenyl, and pyrrole capping groups used to emulate fragment environments.
2.1.3 Conformation generation

For the next two stages of data generation a series of 3D geometries are generated, undesirable geometries are filtered out, and a final set of geometries chosen. After the fragments are capped and stored as SMILES strings the SMILES string for the capped fragment is converted to 3D using the CORINA package\textsuperscript{25}. A conformational search with the MACROMODEL\textsuperscript{26} package produces a set of conformations for each fragment. B3LYP/6-311+G(2d,p) level \textit{ab initio} frequency calculations are run for each conformation to find any conformations that are in transition states. If a transition state is found then the bond responsible, identified by its negative vibrational frequency, is rotated into a new position. The frequency calculation is repeated to determine if the new geometry is no longer a transition state. Duplicate conformations are removed from the resulting set. Large sets are reduced down to the 20 most diverse conformations to ensure that the full range of available conformations can be best represented by a reduced set. This is done to reduce the number of geometries stored for each fragment as storing large sets of similar conformations for a fragment in the database can needlessly take up space and computation time.

2.1.4 Descriptor calculation

The fifth and sixth stages of data generation involve calculation of all properties (or descriptors) for the fragments. In the fifth stage each of the geometries are optimised at the B3LYP/6-311+G(2d,p) level of theory in GAUSSIAN03\textsuperscript{27}. For the sixth stage the QCT properties of the fragments are calculated with the QCT package PROAIMV\textsuperscript{28}. For QCT bond properties a modified version of the program MORPHY01\textsuperscript{29} is used. Once the full set of descriptors for the fragment has been generated they are entered into the QID tool using the MYSQL language. The QID tool currently has 300 1-atom linkers and 450 2-atom linkers entered, where each linker has anything from 1-20 geometries. A 1-atom linker has 1 atom between the 2 capping groups, a 2-atom linker has 2 atoms, and either linker type can have a side chain of up to 15 non-hydrogen atoms.

2.1.5 QID Fragment queries

To search for a replacement of a fragment of interest from within the QID, the properties of the query fragment are compared to the properties of the fragments within the database through a scoring function. The scoring function used in the QID
tool is a Euclidean distance measurement. To understand the scoring function the process of submitting a search to the QID tool is first explained. The user can select a series of properties on which to base the search, shown in Figure 2.4. Each property has an associated importance selected by the user, which helps to distinguish the properties that have a higher priority when scoring fragments. The selected properties are given a tolerance range that can be adjusted. The range is used to return a set of hit fragments from the database where each fragment lies within the tolerances set for all selected properties. For example, a search with charge tolerance as +/- 0.4 a.u. will only return fragments that have charge within 0.4 a.u. of the query fragment. Note that for properties where negative values are not possible, such as volume or number of atoms, the lower tolerance is always set to 0 for the calculation. Fragments that fall within all tolerance boundaries are returned as hits.

**Search for a Fragment with Overall:**

<table>
<thead>
<tr>
<th>Property</th>
<th>Allowed Tolerance</th>
<th>Units</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformation</td>
<td>±0.20</td>
<td>degrees</td>
<td>low</td>
</tr>
<tr>
<td>Charge</td>
<td>±0.6</td>
<td>eV</td>
<td>low</td>
</tr>
<tr>
<td>Dipole</td>
<td>±0.2</td>
<td>au</td>
<td>low</td>
</tr>
<tr>
<td>Quadrupole</td>
<td>±0.5</td>
<td>eV</td>
<td>low</td>
</tr>
<tr>
<td>Volume</td>
<td>±200</td>
<td>au</td>
<td>low</td>
</tr>
<tr>
<td>No. Atoms</td>
<td>±6</td>
<td>atoms</td>
<td>low</td>
</tr>
<tr>
<td>No. Heavy</td>
<td>±4</td>
<td>atoms</td>
<td>low</td>
</tr>
<tr>
<td>Dimensions</td>
<td>±1.5</td>
<td>Angstrom</td>
<td>low</td>
</tr>
<tr>
<td>SCF</td>
<td>±0.1</td>
<td>Hartree</td>
<td>low</td>
</tr>
<tr>
<td>Polarity(Q)</td>
<td>±0.5</td>
<td>au</td>
<td>low</td>
</tr>
<tr>
<td>Polarity(V)</td>
<td>±20</td>
<td>kJ/mol</td>
<td>low</td>
</tr>
<tr>
<td>Polarity(Weighted)</td>
<td>±20</td>
<td>kJ/mol</td>
<td>low</td>
</tr>
</tbody>
</table>

**Figure 2.4:** User interface for QID tool. This includes the properties that can be selected for the query, the tolerance range for that property, and the importance value assigned to the property. The property selection, tolerance and importance are all changeable by the user.
2.1.6 Score calculation

For each fragment hit a distance score from the query fragment is calculated. Each selected search property is calculated as a separate score and is normalized. These scores are then multiplied by the arbitrarily chosen numbers 1, 4 or 8 for low, medium and high importance, respectively. The final score is a combination of the individual scores for each property. The entire process is summarised by

\[ S = \sqrt{\sum_i l_i |X_{i,\text{ref}} - X_{i,\text{hit}}|^2} \]  

(2.2)

where \( X_{\text{ref}} \) and \( X_{\text{hit}} \) are the property \( i \) for the reference and hit fragments respectively, and \( l \) is the user allocated importance. The final score \( S \) for the fragment, represents how closely it matches the query fragment. Fragments are ordered by increasing score so when the user examines the set it is clear which fragments are suggested as the better replacements.

2.1.7 Geometry scoring

Geometry properties are an exception to the above scoring function, which is used for all other properties and the geometry score is calculated separately. However, once the score for the geometric properties is calculated the score is treated the same as other property scores and incorporated into the calculation of \( S \) as normal. The calculation of the score for geometry properties requires a separate calculation because the geometry must described by a combination of geometric values. There also exist special cases that can disrupt the accuracy of the score. This occurs because of the periodic nature of rotations. For example, the two angles 175° and -175° only differ by 10° in a 3D space. However the scoring function is unable to determine this and would consider them 350° apart. This is resolved by calculating the difference between the two angles with

\[ G_i = w(2\Pi - |X_{i,\text{ref}} - X_{i,\text{hit}}|) \]  

(2.3)

where \( i \) is the geometric variable, \( X_{\text{ref}} \) is the value of the geometric variable \( i \) for the reference fragment, \( X_{\text{hit}} \) is the value of the geometric variable \( i \) for the hit fragment, \( w \) is a weighting factor used to represent the different contributions of the geometric variables to the geometry score.
Each of the geometric values has a predetermined weighting depending on its importance towards describing the geometry of the fragment. For example, distances $r$ and $r'$ are deemed less important to describing the geometry then the angles and therefore $r$ and $r'$ have less weighting. Figure 2.5 shows the different geometric variables used for calculation of the geometry score.

Figure 2.5: Geometric parameters used to calculate geometry scores. Atoms 1 and 4 are the connection points $1^*$ and $2^*$ from Figure 2.2.

For each geometry variable the score is calculated with

$$G_i = w_i |X_{i, ref} - X_{i, hit}|$$  \hspace{1cm} (2.4)$$

and summed to give the score ($S$) for the geometry

$$S = \sqrt{\sum_i G_i^2}$$  \hspace{1cm} (2.5)$$

$w=2$ for $\theta$ and $\phi'$, $w=1$ for $r$ and $r'$. A final adjustment is made to $G_{\phi'}$ to account for the change in $\theta'$. When $\theta' \to 0$ the importance of $\phi'$ towards characterising the geometry decreases, therefore $G_{\phi'}$ is multiplied by $\sin(\theta')$ to represent this change in importance of $\phi'$. The score ($S$) is normalised to give the final score for the geometry

$$S_n = \frac{S}{(\|r_u\| + \|r_l\|)}$$  \hspace{1cm} (2.6)$$

where $t_u$ and $t_l$ are the upper and tolerance limits respectively. The QID database and data generation process discussed up to this point were established by Devereux. The aim of the following work is to expand and improve the existing QID tool.
2.2 Quantum Chemical Topology

2.2.1 Introduction to QCT

QCT refers to a body of work that uses the idea of a gradient vector field to extract chemical insight from modern wave functions or high-resolution experimental electron densities. The name QCT was first proposed in 2003\textsuperscript{30} in an attempt to bundle results that had been obtained under separate headers, the oldest and most important being the Quantum theory of Atoms in Molecules (QTAIM). This theory, proposed by Bader and co-workers, is in fact a generalisation of quantum mechanics to subspaces. Such a subspace can be identified with a topological atom (or QCT atom), which naturally arises by the electron density partitioning itself via its gradient vector field. How this happens is discussed below. The application of QTAIM and QCT has grown into many fields, including QSAR\textsuperscript{31}, molecular dynamics\textsuperscript{32}, free radicals\textsuperscript{33}, electron delocalisation\textsuperscript{34} and hydrogen bonding\textsuperscript{35,36}.

QCT partitions a molecule into (topological) atoms from an electron density, which in this work is calculated \textit{ab initio}. Figure 2.6 shows the electron density in the molecular plane of furan, which is a planar five-membered heterocyclic aromatic containing a single oxygen atom. Plotting all points in space with the same electron density value gives isodensity envelopes shown as contour lines in Figure 2.6. We can now trace a network of trajectories that are everywhere locally orthogonal to the set of constant electron density contour lines. This network is called the gradient vector field, as shown in Figure 2.6. This vector field consists of so-called gradient paths, which are trajectories of steepest ascent in the electron density. In other words, if one follows a gradient path then one will reach, after each step, the locally high electron density in the quickest possible way. Note that a gradient path has a direction, i.e. it has a starting point and an end point.

2.2.2 Critical points

The majority of gradient paths start at infinity and terminate at a point where the gradient of the density vanishes, known as a critical point. These critical points are marked as squares, circles or a triangle in Figure 2.6, depending on the type of critical point. The difference between types can be explained as follows. Taking the second derivative of the electron density at a critical point gives a three-by-three matrix known as the Hessian. The Hessian is symmetric and after diagonalisation yields three eigenvalues denoted \( \lambda_1, \lambda_2 \) and \( \lambda_3 \), which are the local curvatures in the electron density.
When summed, they give the value of the Laplacian of the electron density. The type of critical point can be categorised by the signs (positive or negative) of the three eigenvalues. Critical points are written in the form \((\omega, Z)\) where \(\omega\) is the number of non-zero eigenvalues, and \(Z\) is the sum of the signs of the eigenvalues where a negative value is represented by -1 and positive by +1. A critical point at the nucleus (circle in Figure 2.6) has three eigenvalues that are all negative and therefore summarised as \((3, -3)\). A bond critical point (BCP), shown as squares, exists at a saddle point in the electron density between two nuclei, and is marked by \((3, -1)\) because it has two negative eigenvalues and one positive eigenvalue. A ring critical point (RCP) (triangle) or \((3, +1)\) critical point appears in the centre of the furan ring in Figure 2.6 and has two positive eigenvalues and one negative one. In general, there also exist cage critical points \((3, +3)\) but they will not be discussed further because they do not feature in this work.

At a RCP, \(\lambda_2\) and \(\lambda_3\) are both positive while \(\lambda_1\) is negative, and by convention the eigenvalues follow the pattern \(\lambda_3 > \lambda_2 > \lambda_1\). The shape of the electron density at a given RCP can be calculated by an RCP ellipticity, which we define here. Note that this ellipticity is analogous to (but not to be confused with) the well-known ellipticity defined at a BCP. The BCP ellipticity is calculated as \((\lambda_1/\lambda_2) - 1\), whereas the ellipticity at the RCP is calculated as \((\lambda_2/\lambda_3) - 1\). The latter attains the minimum value of 0, when it is evaluated at the centre of a cylindrically symmetric electron density. The electron density of a benzene ring approaches such an electron density of this symmetry, as proven by the very low value of the RCP ellipticity of approximately \(2 \times 10^{-4}\).

### 2.2.3 Partitioning of the electron density

A molecule is partitioned into its constituent atoms through the gradient vector field, which is a collection of an infinite number of gradient paths. Figure 2.6 shows the gradient vector field of furan in its molecular symmetry plane. The vast majority consists of gradient paths that originate at infinity and terminate at a critical point. Figure 2.6 contains numerous special gradient paths that are not part of his majority. An example of a special gradient path is one that connects two critical points, such as the \((3,-3)\) at the nucleus (circle) and a BCP (square). Another example is a gradient path that connects the RCP (triangle) to a BCP (square). Two gradient paths that each begin at the same BCP but terminate at two different nuclei form a so-called Atomic Interaction Line (AIL). There are many examples in Figure 2.6, and they collectively form the molecular graph of furan, which recovers the expected Lewis diagram for this molecule. While
most typical gradient paths terminate at a nucleus there are gradient paths that originate at infinity but terminate at a BCP. Collectively these gradient paths form the so-called Interatomic Surface (IAS). An intersection of such an IAS with the plotting plane is shown in Figure 2.6, for example, as one of the bold lines that bisects the carbon oxygen AIL at the BCP. An IAS sharply divides the electron density between the two topological atoms that it bounds. In other words, the electron density at one of the IAS strictly belongs to one atom and the electron density at the other side to the other atom. Topological atoms do not overlap or penetrate each other. Moreover, they leave no gaps between them. From Figure 2.6 it is clear that an atom is a subspace that consists of gradient paths that are all attracted to its nucleus.

Figure 2.6: Gradient vector field in the molecular symmetry plane of furan, showing interatomic surfaces, atomic interaction lines and critical points.
Figure 2.7 displays the gradient vector field in three dimensions and shows that each atom has a distinct topological volume (known as the atomic basin) within the molecule. Note that, topologically, a QCT atom extends to infinity when it is not bound by other atoms and therefore a cut-off value of 0.001 a.u. iso-electron density is used to cap the atoms in Figure 2.7.

![Figure 2.7: Three-dimensional representation of topological atoms in imidazole with a 0.001 a.u iso-electron density boundary cut-off.](image)

The atomic basin \((\Omega)\) can be considered as a “volume of influence” of an atom within a given molecule. Volume integration of a property density over an atomic basin gives the atom’s property within the molecule. For example, if this property density is the electron density itself then the volume integration results in the atom’s electronic population. If corrected for the nuclear charge, this population leads to the QCT charge of an atom. This is discussed in more detail below. Clearly, whichever the property density being integrated, the integral yields the atom’s contribution towards the molecular property of interest. For example, if the property density is simply unity, as in eq 2.7,

\[
v(\Omega) = \int_{\Omega} d\tau
\]  

(2.7)
then the atom’s volume is obtained, where $d\tau$ is an infinitesimal volume element. The volume of the molecule, or a fragment thereof, is then simply obtained by adding the relevant atomic volumes. This perfect additivity is very convenient in any QCT analysis of properties.

A number of other atomic properties feature in this work. One type of atomic kinetic energy is defined by

$$K(\Omega) = -\frac{1}{4} N \int_{\Omega} d\tau' [\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*]$$  \hspace{0.5cm} (2.8)$$

where $N \int_{\Omega} d\tau'$ is a shorthand notation referring to an integration (over all space) over all electrons except one, which is integrated over the topological atom’s volume. Note that this compact notation also sums over the spins of the electrons. Topological atoms have the desirable property that they have a well-defined kinetic energy, which cannot be said of arbitrary subspaces within a molecule.

Because a topological atom has its own virial theorem\textsuperscript{37}, which expresses a balance between its kinetic and potential energy, the total energy of an atom can be expressed using its kinetic energy only, as shown in

$$E(\Omega) = -\int_{\Omega} d\tau K(\mathbf{r})$$  \hspace{0.5cm} (2.9)$$

where $\mathbf{r}$ is a position vector of the kinetic energy density $K(\mathbf{r})$ is a function. Again, summation of an atomic property over all atoms in a molecule produces the molecular property. Therefore, a property for a molecular fragment, such as a functional group or the set of ring atoms, is simply obtained by summing the constituent atoms. Any atomic property can be calculated by integrating the property function over the basin\textsuperscript{28}, this is achieved through different QCT packages\textsuperscript{38, 39}.

\subsection*{2.2.4 Electrostatics}

The molecular electrostatic potential is an important property, which has been widely utilised in studies on intermolecular interactions\textsuperscript{40-43} and hydrogen bonding\textsuperscript{44-47}, for example. The molecular electrostatic potential can be physically understood by imagining a unit charge that moves around from one position to another, probing the molecular charge density. We assume that the charge density is not polarised by this probing charge. For each position $\mathbf{r}$ of the charge, a new interaction energy will be
obtained. This energy, divided by the unit charge, is the molecular electrostatic potential. It can be written as

\[ U_{ele}(\mathbf{r}) = \sum_{i=1}^{N} \frac{Z_i}{|\mathbf{r} - \mathbf{R}_i|} - \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \] (2.10)

where \( \rho(\mathbf{r}') \) is the electron density and the first term takes care of the nuclear potential, for a molecule of \( N \) atoms, with nuclear positions \( \mathbf{R}_i \).

Evaluating the second term in equation 2.10 at every single point is time consuming. A Taylor series expansion can be used to reduce the cost of evaluating the electrostatic potential at every point. This expansion essentially factorises the expression \( |\mathbf{r} - \mathbf{r}'| \) such that a volume integral in \( \mathbf{r}' \) can be pre-computed independently of the \( \mathbf{r} \) coordinate. The expansion creates terms called 'multipole moments', which are capable of representing the electrostatics exactly provided the multipolar series expansion converges. More precisely, one can place multipole moments on the nuclear positions and have them generate the molecular potential via the multipolar expansion. The only disadvantage of this multipolar expansion is possible convergence problems, but an advantage of topological atoms is that they occupy a finite volume, and hence formal convergence is possible\(^{48}\). The higher order terms have a smaller effect on the result so it is possible to accurately represent the electrostatics using only the first few multipole moments of the expansion\(^{49}\). Each multipole moment can be categorised by its rank \( 'l' \), where \( l=0 \) for a monopole moment, \( l=1 \) for a dipole moment, \( l=2 \) for a quadrupole moment, and so on.

The first term of the expansion is the monopole moment given by

\[ Q_{00}(\Omega) = -\int_{\Omega} d\tau \rho(\mathbf{r}) \] (2.11)

The monopole represents the electronic population of the atom, which when corrected for the nuclear charge is the atomic charge. The second term of the expansion involves the three dipole moments, one of which is given by

\[ Q_{11c}(\Omega) = -\int_{\Omega} d\tau x \rho(\mathbf{r}) \] (2.12)

where \( x \) is referring to a global frame. The third term of the expansion involves only five rather than six quadrupole moments because we use the spherical tensor formalism.
rather than the Cartesian one. The five components of the quadrupole moment correspond to the well-known $d$ orbitals from quantum mechanics and read $3z^2 - r^2$, $x^2 - y^2$, $xy$, $xz$ and $yz$. The quadrupole moment represents how the electron density deviates from a spherical distribution. The next term involves seven (i.e. $2\ell + 1 = 7$) octupole moments but they do not feature in this work.

2.3 Quantum mechanics

2.3.1 Introduction to Quantum mechanics

A quantum mechanical treatment of a system, in essence, holds all the information that can be known about the state of the system. Electronic structure and properties may be determined for a system using quantum mechanics. While the exact properties can only be determined for a one electron system, such as hydrogen or the Helium ion, approximate solutions can be obtained to a high level of accuracy. However, quantum mechanical methods are computationally demanding and higher levels of accuracy require increased computing resources. The cost in computational power required increases exponentially with the size of the system to the extent where large molecules such as proteins are not considered feasible with a purely quantum mechanical treatment and instead mixed quantum mechanical and molecular dynamics models are used. As the power of computer increases so does the applications available for quantum mechanics, where larger systems can now be explored and more accurate and demanding methods utilised.

2.3.2 The Schrödinger Equation

To obtain the properties of a system through an ab initio approach an approximate solution to the Schrödinger equation is necessary. The time independent Schrödinger Equation treats molecules as isolated entities within a vacuum, for a solvated molecule a solvation model is required or explicit waters added to the system for a more accurate representation. The time dependent Schrödinger Equation, while making less approximations and therefore is more accurate, is much more computationally demanding. Of the two forms of the Schrödinger Equation, it is the time independent that is most commonly calculated in quantum chemistry.
The time-independent Schrödinger Equation is expressed in the simplified form

\[
H(r, R)\Psi(r, R) = E\Psi(r, R) \tag{2.13}
\]

where \(H\) is the Hamiltonian operator, \(R\) and \(r\) the nuclear and electronic coordinates, respectively, \(\Psi\) represents the wave function, and \(E\) the allowed energies of the system. The Hamiltonian takes the form \(^{51}\)

\[
\hat{H} = -\frac{1}{2} \sum_{i=1}^{N} \nabla_i^2 - \sum_{A=1}^{M} \frac{1}{2M_A} \nabla_A^2 - \sum_{i=1}^{N} \sum_{A=1}^{M} \frac{Z_A}{r_{Ai}} + \sum_{i=1}^{N} \sum_{j<i}^{N} \frac{1}{r_{ij}} + \sum_{A=1}^{M} \sum_{B>A}^{M} Z_A Z_B / R_{AB} \tag{2.14}
\]

Where \(M\) and \(N\) are the number of nuclei and electrons respectively, \(M_A\) is the ratio of the mass of nucleus \(A\) to the mass of an electron, \(Z_A\) is nucleus \(A\)'s atomic number and \(\nabla^2\) is the laplacian given by

\[
\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \tag{2.15}
\]

for the three dimensional Schrödinger Equation. The first and second terms for the Hamiltonian are the electronic and nuclear kinetic energies respectively, the third is the nuclei-electron coulomb attraction followed by the electron-electron coulomb repulsion and finally the nuclear repulsion term. Under the Born-Oppenheimer approximation (fixed nuclei) the nuclear kinetic energy and repulsion terms are constant, resulting in the electronic Hamiltonian which is used in \textit{ab initio} calculations.

The equation is an eigenvalue equation where \(\Psi\) and \(E\) are the eigenvector and eigenvalues respectively. The wave function “\(\Psi\)” can be determined upon solution of the equation, and while the physical meaning of the wavefunction remains unknown, the probability of finding an electron in a known space can be calculated by \(|\Psi|^2\) and a probability density derived for the entire system.

### 2.3.3 Level of theory

The Hartree-Fock 'Self-Consistent Field' (SCF) method is the most basic \textit{ab initio} method that is commonly used. Hartree-Fock solves the Schrödinger Equation for a single electron by expressing the electron-electron interaction term of the Hamiltonian as a single electron experiencing an average field from all other electrons. The electron repulsion term is reduced from an electron experiencing \(N\) interactions to \(N\) electrons...
experiencing a single interaction each. This approximation is referred to as the Hartree-Fock approximation and requires the wavefunctions of the rest of the electrons to be known for the interaction between the single electron and the field to be calculated. To circumvent this problem an initial wavefunction is guessed for the electrons in the field. Once the single electron’s wavefunction is calculated it is incorporated into the field to calculate the wavefunction of the next electron and the sequence is continued for all electrons with the updated single electron wavefunctions. The process is repeated once all electron wavefunctions have been calculated because the first electron now experiences a different, more accurate, field to its first computation. After a number of iterations the changes in the electric field and the wavefunction become negligible and the field has achieved self-consistency.

The Linear Combination of Atomic Orbitals (LCAO) approximation is also introduced. Atomic orbitals from their parent atom are combined to form molecular orbitals for the atoms concerned. The wavefunction of the molecular orbital can be expressed as the combination of orbitals \(a\) and \(b\) for their respective atoms

\[
\psi = c_a a + c_b b \quad (2.16)
\]

where \(c_a\) and \(c_b\) are coefficients.

The above approximations lead to an approximate solution to the Schrödinger Equation. However, Hartree-Fock methods do not consider Coulomb correlation energy that occurs through the interaction of electrons of opposing spin. Neglecting Coulomb correlation energy results in electrons moving too close to each other in space, and ignoring this phenomenon results in an energy that is always higher than the exact energy. The correlation energy of a system is the difference between these two energies. Correlation of two electrons in the same spin state is included in an exchange term and rises from the Pauli Exclusion Principle. The energy terms of a Hartree-Fock method can be partitioned as

\[
E^{\text{HF}} = E^{\text{nuclear}} + E^{\text{core}} + E^{\text{coulomb}} + E^{\text{exchange}} \quad (2.17)
\]

Higher level theories such as Density Functional and Møller-Plesset Perturbation theories, DFT and MP, respectively, correct for the absence of Coulomb correlation between two electrons of different spins in Hartree-Fock methods. Møller-Plesset Perturbation theory takes the Hartree-Fock methods and adds a corrective
contribution for the Coulomb correlation. Density functional theories take a different approach to Hartree-Fock and Møller-Plesset. Instead of taking a system’s many-electron wavefunction, hybrid Density Functional methods such as B3LYP\textsuperscript{54} use an electron density cloud. Density Functional theory methods are generally faster than Møller-Plesset because the degrees of freedom for \( N \) electrons are reduced from \( 3N \) to 3 (\( x, y, \) and \( z \)). The exchange-correlation energy term is introduced to account for Coulomb correlation and is based on the electron density. The partitioning of the energy for Density Functional Theory methods is given by

\[
E^{\text{DFT}} = E^{\text{nuclear}} + E^{\text{core}} + E^{\text{coulomb}} + E^{\text{exchange-correlation}} \tag{2.18}
\]

2.3.4 Basis sets

Atomic orbitals are represented by basis set functional, a molecular orbital (\( \phi_i \)) can be constructed from the atomic orbitals using

\[
\phi_i = \sum_{\mu=1}^{N} c_{i\mu} \chi_\mu \tag{2.19}
\]

where \( N \) is the set of functions, \( c_{i\mu} \) a coefficient and \( \chi_\mu \) is an arbitrary basis function. The anti-symmetric product of the single molecular orbitals is the overall molecular wavefunction. The functions that are most commonly used are Gaussian functions given by

\[
g(\alpha, r) = c x^n y^m z^l e^{-\alpha r^2} \tag{2.20}
\]

where \( c \) is a constant, \( \alpha \) a positive exponent that fixes the radial extent, and \( n, l, m \) are integers that determine directional dependence. An ‘s-type’ orbital is represented when \( n + m + l = 0 \) because there is no directional dependence of the Gaussian function. A ‘p-type’ orbital is represented by \( n + m + l = 1 \) because the Gaussian function lies along the axis. Gaussian functions are combined to give Gaussian type orbitals. As the number of atomic orbitals in the basis set increases, the solution of the energy becomes closer to exact with the cost of a dramatically increased computational demand.

Improved computational times can be achieved without sacrificing too much accuracy with Split Valence basis sets. Valence shell electrons are more important in the bonding of a compound and are modelled with the more accurate double or triple zeta
models. The number of functions for an atomic orbital is increased to two or three for double and triple zeta basis sets, respectively. Variation within the orbitals of the valence electrons due to interactions and bonding is allowed for by the added functions. Computational time is saved by assigning fewer functions to the core orbitals as it is assumed that there is little variance within them; however accuracy is improved by assigning extra functions to the valence orbitals.

Polarisation functions can be added to a basis set to improve its accuracy. Adding polarisation functions to a basis set such as 6-31G creates the basis set 6-31G(d,p) where d-functions are added to non-hydrogen atoms and p-functions to hydrogen atoms. For weakly bond electrons diffuse functions are important, the function is spread over a large radius and is present in the 6-31+G basis set.

2.4 Bioisosterism and ligand design
2.4.1 Introduction to bioisosterism

The QCT properties which are derived from quantum chemistry are applied by the QID with an aim towards ligand design through the concept of bioisosterism. This concept is introduced in this section and its key points discussed. There are a vast number of methods aimed at bioisosterism through various different means, and a small selection is discussed to introduce the reader to key work that is firmly rooted in the literature. As there are many methods of attempting bioisosterism, so too are there a large number of alternative methods that attempt to achieve successful ligand design through other means. A small selection of these methods is discussed with their merits and drawbacks with a final discussion towards the use of the QID with respect to all other methods.

In order for the pharmaceutical and agrochemical industries to successfully design ligands, a successful ligand must undergo an extensive process that iteratively improves potential candidates and filters undesirable candidates until a ligand with the desired properties is found. Modification of the ligand can take the form of a functional group conversion, CO₂H to CO₂Me for example, or alteration of the stereochemistry by converting a carbon-carbon double bond from cis to trans. The aim of these modifications is to create ligands that express a predefined or desired set of properties while also minimising any detrimental biological effects, such as toxicity. The process of
creating and honing the ligand has largely relied on chemical intuition mixed with a trial and error based approach. This process is usually found to be longwinded. Creating subsequent sets of laboratory results where each set builds on the previous one is a time consuming iterative cycle. Examples of such cycles of high throughput combinatorial chemistry are scattered throughout the literature. An example of ligand design through chemical intuition is Aberg et al. who studied possible variations to a ligand used for inhibition of Pilus Biogenesis in E.coli. The active COOH group was replaced by 12 different functional groups all showing varying levels of bioactivity of the inhibitor molecule. This is one of countless examples showing a “trial and error” method employed by synthetic chemists. This method can lead to laborious research and the possibility is small for discovery of truly novel replacements that would go against chemical intuition.

Introduction of computer-based methods has spawned logical and efficient approaches to suggesting suitable ligand replacements. Databases such as BIOSTER, which contains thousands of bio-analogous molecules, have been created to house large volumes of molecules and molecular fragments with a large range of associated experimental data obtained from the literature. BIOSTER is the benchmark to which other databases are compared. Existence of these databases still poses the problem of how to choose suitable ligand replacements from a large source of material. Quality of results is also dependent on the search criteria. It is now seen that many research groups are designing their own software using a wide variety of methods to seek and suggest suitable replacements for the ligands. The volume of different methods rises from the numerous possibilities of different molecular descriptors, search algorithms and scoring functions used to return a set of suggested ligand replacements. An example of the large differences between approaches are GOLD, which uses Gaussian shape overlays to describe 3D shapes and volumes of molecules, and Graham who decided to use a molecule’s “hardness” or “softness” as descriptors. The programs can be used to produce a list of potential candidates as possible replacements. Depending on the quality of the programs comprehensive lists of replacements can be called forward and the programs may even be capable of producing less obvious and fresh alternatives.
2.4.2 Definition of Bioisosterism

The concept of bioisosterism has been around since the earliest attempts at drug design. Bioisosterism is the modification of a ligand to produce a resultant ligand of similar activity and is widely used in Medicinal chemistry\textsuperscript{70-73}. Unfortunately, bioisosterism is loosely defined throughout the literature where most definitions focus on the phrase "similar activity", which is vague in itself. Originally, the first isosteres were defined as compounds or groups of atoms with the same number of electrons\textsuperscript{74} this was later clarified by Grimm on proposing the hydride displacement law\textsuperscript{75,76}. Table 2.1 summarises the hydride displacement law which states that adding hydrogen to an atom gives a compound of similar properties to the element in the previous period. Each column within the table represents a set of molecules with similar physical properties.

<table>
<thead>
<tr>
<th>C</th>
<th>N</th>
<th>O</th>
<th>F</th>
<th>Ne</th>
<th>Na</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH</td>
<td>NH</td>
<td>OH</td>
<td>FH</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CH\textsubscript{2}</td>
<td>NH\textsubscript{2}</td>
<td>OH\textsubscript{2}</td>
<td>FH\textsubscript{2}</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>CH\textsubscript{3}</td>
<td>NH\textsubscript{3}</td>
<td>OH\textsubscript{3}</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CH\textsubscript{4}</td>
<td>NH\textsubscript{4}</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Table 2.1: Grimm's Hydride Displacement law.

Erlenmeyer took the idea further and suggested that only the valence shell electron count needs to be similar for a pair of molecules to show isosterism\textsuperscript{77}. The concept of isosterism applied to a biological context coins the term bioisosterism. At present, bioisosteres are said to be molecules similar in bioactivity. Bioactivity refers to the molecule’s ADMET properties (absorption, distribution, metabolism, excretion and toxicity) pharmacophore properties, steric and electronic features of a molecule that ensure binding to the target molecule gives the optimum biological activity of the receptor. The IUPAC definition of a bioisostere is “a compound resulting from the exchange of an atom or group of atoms with another, broadly similar, atom or group of atoms”\textsuperscript{78}.
2.4.3 Non-classical and classical bioisosterism

There are two main classifications of bioisosteres, classical and non-classical. Classical bioisosteres constitute a direct replacement of an atom or functional group (of no more than 2 heavy atoms size), or a minimal alteration to the biological molecule, such as the hydride displacement law (Table 2.1). A more recent example of classical bioisosterism is shown by Sheridan who employed an algorithm to ensure that only bioisosteres with direct atom or functional group replacements were allowed\textsuperscript{79}.

Non-classical bioisosteres show a more complex design in the arrangement of the replacement atoms\textsuperscript{80}. Sometimes little similarity is shown between the original molecule and the resultant bioisostere. Non-classical replacements include ring substitutions, alteration of the molecule's carbon backbone, and introduction of more exotic functional groups. Non-classical bioisosteres are more desirable than classical because they are more difficult to conceive and most classical bioisosteres are now well known. There are countless other examples of classical and non-classical bioisosteric replacements\textsuperscript{24,81,82}.

In the quest to find non-classical bioisosteres a method is required to find suitable replacements without finding false positives because novel replacements do not always follow chemical intuition. It is in the method of how the molecules are described that allows for successful results of novel non-classical bioisosteres. The quality of a bioisostere search is dependent on how the molecules are described to the machine and how they are compared and ranked. Broad methods such as pharmacophore keys are able to describe a molecule by creating a description that omits the finer details of the molecule and relies on key features of the molecule, which are selected and defined by the user. This broad approach makes it possible to miss the finer details of a molecule that may play important roles in binding and its activity, such as the effects of conjugation. Conversely, overcomplicating the descriptors may lead to misclassification where two molecules are deemed to be different because of the level of detail required for a match. Therefore, descriptors must be able to capture the "essence" of a molecule making sure that all features important to a molecule's activity are accounted for, but without choking the molecular freedom, i.e. being too demanding on match criteria because of the level of detail. With most bioisostere databases there is a tendency to rely on simpler descriptors as opposed to the more recently used high level descriptors such as \textit{ab initio} calculated properties\textsuperscript{24,82,83}. 

45
2.4.4 2D molecular descriptors

To encode the chemistry of a molecule for a bioisostere search the molecules are described through molecular descriptors. The descriptors are chosen to best represent the features deemed important for determining a suitable replacement. The list of possible descriptors is endless, from counting the number of carbons within a molecule to a Molecular Electrostatic Potential (MEP) generated from a quantum chemical electron density. A small sample of molecular descriptors is discussed below.

A popular descriptor used throughout the pharmaceutical industry to describe hydrophobicity is LogP, where P is the partition coefficient (ratio of the concentrations of compound between 1-octanol and water\(^ {34} \)).

\[
\text{LogP} = \text{Log} \left( \frac{[\text{solute}]_{\text{oct}}}{[\text{solute}]_{\text{water}}} \right) \quad (2.21)
\]

Experimentally determining LogP for a compound can take over 30 minutes for the compound to equilibrate between the water and octanol. When predicting LogP theoretically, the contributions of each atom towards the overall LogP are calculated by fitting to a training set of experimentally determined LogP values\(^ {85-87} \). To increase the efficiency of the calculations the molecule itself is partitioned into fragments and partition functions for each fragment calculated and summed together. A method using this idea was performed by Leo\(^ {88} \). Leo began by defining an isolated carbon as a carbon not bound to a heteroatom and not having any double or triple bonds. Molecules were then fragmented into isolated carbons, non-isolated carbons and heteroatoms and the partition functions of each fragment calculated.

The simplest descriptors are based on molecular graphs. A molecular graph is a 2D representation of a molecule where the vertices are atoms (labelled with an atom type) and the edges are bonds (labelled with a bond type). Molecular graphs are very quick to construct but contain no 3D structural information. However, molecular graphs act as a simple framework for other descriptors. Typically, descriptors are assigned to the vertices and edges of the molecular graph and comparisons made between the descriptor values and their positions on the graph. Because a large portion of bioisosteric methods rely on molecular graphs, two examples are used to highlight the different uses of assigning descriptors to molecular graphs. The first example is the derivation for a molecule’s topological state descriptors developed by Kier and Hall. The fragment, molecule, and atom types are discriminated from each other using the \( \delta \) values: \( \delta_i \) and \( \delta_i^v \) denote the simple and valence \( \delta \), respectively\(^ {89,90} \). \( \delta_i \) is calculated by
subtracting the number of hydrogens bonded to species $i$ from the number of $\sigma$ electrons on species $i$. It discriminates atoms of the same element but in different local environments. $\delta^i$ is the subtraction of the number of hydrogens bonded to species $i$ from the number of valence electrons of $i$. Therefore it can discriminate between different elements. Each atom in the fragment or molecule is assigned an intrinsic state as defined by

$$I_i = \frac{\delta^i_v + 1}{\delta^i}$$  \hspace{1cm} (2.22)

The intrinsic state is independent of the surrounding atoms and any effects imprinted on it from the local environment. To account for the influence that the surrounding environment generates over the fragment, the perturbation is calculated between atoms $i$ and $j$ separated by path length $r_{ij}$ using

$$\Delta I = \sum_j \frac{I_i - I_j}{r_{ij}}$$  \hspace{1cm} (2.23)

When summed together, the intrinsic state and perturbation terms give the electrotopological state. To obtain the descriptor for the whole molecule the mean square of electrotopological state is taken for all atoms. It is also noted that the intrinsic state ($I$) on its own can be used as a descriptor by assigning bits to each different value of $I$ to create a bit-string.

The second example of descriptors assigned to molecular graphs is R-group descriptors by Holliday et al.\textsuperscript{65}. This approach is similar to the above in that a set of properties (descriptors) are assigned to the atoms of the molecular graph. This method also uses the summed properties as well as the properties of the atoms individually. The difference is the application of a distance to the individual atom properties. A weight is applied to an atom’s given property based on its position in the molecular graph. While both these example approaches use molecular graphs, they are limited to a 2D description of a molecule.
2.4.5 3D molecular descriptors

Naturally, calculation of descriptors for large biological molecules such as drugs, proteins and enzymes can be computationally expensive. Pharmacophore descriptors are specially designed to handle the larger classes of molecules. A pharmacophore is an atom or point in a molecule that is instrumental in forming the correct binding interactions with receptors. Common pharmacophores are hydrogen bond donors and acceptors, positively charged centres, aromatic centres, and hydrophobic centres. It can be a laborious task to calculate properties each time for a descriptor, especially when using 3D structures or properties dependent on the 3D conformation of the molecule. Pharmacophore descriptors are able to store 3D structures and search for particular conformational arrangements. Calculating structural descriptors for large biological molecules can be difficult because of freedom in the conformation of the molecule due to bond rotation. The same molecules can have a large range of high and low energy conformations. Therefore conformational freedom or flexibility needs to be considered. To solve the problem of conformational freedom a 3-point pharmacophore model is used\textsuperscript{91,92}. Three pharmacophoric centres are taken and are assigned a type: donor, acceptor or ring. The distances between centres are measured and ranges assigned to the distances, to allow for conformational freedoms. The smaller the range the tighter the fit of the bioisostere needs to be. A larger range is more forgiving and allows for more conformational variance. The distances and ranges are known as screens. Multiple screens are then grouped together and referred to as a bin. Bins are set to bits and the final product is a bit-string that can be used to find matches between biological molecules. It is possible to increase the number of points to four\textsuperscript{93}, however calculations using 4-point pharmacophoric descriptors are computationally more demanding. An example of a pharmacophore based bioisosterism is the method proposed by Wagener et al., which used fingerprints. A fingerprint for a molecule is constructed from the structural data or pharmacophore sites, which included pharmacophores such as hydrophobic sites and hydrogen bond sites. Bioisosteres were recommended based on the comparison between fingerprints. These comparisons were made with the Euclidean distances between the pharmacophore sites. Each pharmacophore was weighted with respect to its importance towards activity. By using these fingerprints constructed from pharmacophores Wagemar was able to make fast comparisons and cover a large portion of chemical space because not all the structural information was being encoded into the fingerprint.
Improvement in CPU speed means chemists are able to improve their descriptors and many new descriptors have evolved or been created to take on more physical forms, holding much more physical and or chemical information relative to 2D molecular graphs. Many new methods are emerging that rely on these higher-level molecular descriptors. The following are a few of these new methods used for defining molecules, and subsequently the difference or similarity between target molecules and their replacements.

Jennings and Tennant proposed a method that combined both shape and electrostatic properties. They chose shape and electrostatics because these are currently accepted to be the main driving forces behind binding affinities. Jennings and Tennant chose to use both shape and electrostatic descriptors simultaneously to create their molecular descriptors. They were chosen because they are both found to be instrumental in protein binding and the relative influence of each on the binding affinity is still unknown so both were incorporated.

Comparing the shape similarity of two molecules was calculated by overlapping the two molecules and calculating the similarity using

$$Tanimoto_{A,B} = \frac{O_{A,B}}{I_A + I_B - O_{A,B}}$$

(2.24)

where $O_{A,B}$ is the overlap between molecules $A$ and $B$, and $I_X$ is the self-overlap of molecule $X$. A Tanimoto score of 1 suggests an identical match. As the Tanimoto tends towards zero, the molecules $A$ and $B$ become more dissimilar. The electrostatics of the molecules was compared in a similar fashion, but the similarity of the molecules was calculated with the overlap of the electrostatic fields. Database molecules were compared to the target molecule based on their shape and electrostatic descriptions, using molecular overlap values to generate a similarity.

Another method based on chemically derived properties is implemented in the program Volsurf, which relies on a descriptor known as the molecule's “molecular field”. Volsurf generates a 3D molecular field by introducing a water or hydrophobic probe to the system and measuring the interaction. Using the field and the interactions a multitude of descriptors was generated that represented the size, shape, and hydrophobic or hydrophilic nature of local environments of the molecule. Some of the descriptors generated included: molecular volume (generated by the water excluded
volume), molecular surface area, the area of the water accessible surface, molecular volume to surface area ratio, molecular globularity (where a score of 1.0 is given to a system that is perfectly spherical), hydrophilic regions (areas that are water accessible and attract the water), and hydrophobic regions (generated in a similar fashion to hydrophilic but using a methyl or phenyl probe). More detailed and chemically relevant descriptors enable a more complete description of a molecule and allow for much more flexibility in comparing, with increased CPU cost. Molecular fields are also implemented in various other programs, such as Cresset’s ligand-based virtual screening program, CoMFA, GRID and ALMOND, where all packages generate molecular field based descriptors.

2.4.6 Descriptor space

Naturally, with the limitless number of descriptors available methods employ large numbers of descriptors simultaneously to describe a molecule. However, as the number of dimensions of a descriptor space is increased with the addition of extra descriptors creates a space of such high dimensions to describe a molecule, which can create problems. A study by Rupp et al. highlighted troubles of navigating the higher dimensional descriptor spaces. They found phenomena with high dimensional descriptor space, which were all attributed to the non-linear growth of volume as dimensionality is increased. To visualise this take a sample uniformly distributed in the descriptor space of $d$ dimensions. Split each dimension into 2 compartments so that an object lies in each compartment. For all dimensions this gives $2^d$ compartments. As the number of dimensions increases an exponential number of objects are needed to cover the descriptor space. Applied to a non-uniform sample this will lead to large redundant space. For example a data set of $10^7$ compounds described by Volsurf descriptors ($d = 70$) the fraction of space covered is $70 / 10^7 = 0$. The max number of dimensions the data set can cover is $\log_2(10^7) = 23$. This also makes distance comparisons difficult. If the number of points that lie within distance $r$ of a target for a $d$-dimensional Euclidean sphere with radius $r > 0$ were measured, as $d$ tends to $\infty$ the volume goes to zero. In practical terms as the number of dimensions increases the space sampled is exponentially reduced to the point where the target is the only point that lies within the sphere. Therefore, it is desirable to reduce the size of descriptor space. The ideal number of descriptors can fully describe the properties required for bioisosterism without venturing into higher dimensions where these phenomena are more noticeable.
Reducing the number of descriptors required to describe a molecule alleviates the problems mentioned above. Highly detailed descriptors will undoubtedly improve the quality of a bioisostere database as fewer descriptors will be required. They allow for a more complete description of a fragment or molecule making it easier to compare the similarities and differences between molecules. However, the creation of a set of high-level descriptors will by no means guarantee a high quality bioisostere method. When conducting a bioisostere search the package must be able to return a set of potential replacements generated by comparison of the descriptors. The extent of deviation or similarity from the search molecule is more commonly referred to as "scoring" and is the final stage of a database's inner working.

2.4.7 Scoring functions

Scoring functions compare the descriptors between molecules within a bioisostere database (or library). The purpose of creating a library of bioisosteres is to provide combinatorial and pharmaceutical synthetic chemists with a tool that is able to suggest a variety of alternative molecules for use in their studies. Commercial libraries such as WDI\textsuperscript{101} and BIOSTER contain up to tens of thousands of bioisosteres. There are also many libraries that are developed in-house that also contain anything from hundreds up to tens of thousands of bioisosteric replacements. Therefore, the results must be organised through scoring functions. As with the descriptors, the level of detail in the scoring function can create vastly different results for different methods. For example, very high levels of descriptor and a high level (or strict) scoring function means that very few possible replacements are suggested and their variation may be limited. If there is little room for deviation then no novel ideas would be generated. This would most likely result in simple classical replacements that chemists could do very easily by hand. Conversely, loose similarity scoring would suggest a huge amount of possibilities and many would not be bioisosteric in nature. The purpose of the library is to suggest suitable replacements. Therefore, if the scoring function is not suitable then no matter how well-crafted the descriptors are the library will be unable to suggest the best replacements. Scoring also enables a library to distinguish between varying degrees of similarity with the original subject of the search.

Scoring methods can range from simple comparison of bit strings, to more in-depth statistical models, and most recently the application of multi-objective searches\textsuperscript{102}. Comparing bit strings is the reduction of a molecule into a mathematical
expression or bit string. Various functional groups or pharmacophore keys are assigned bits that are turned on and off by their presence in the molecule. Once the bit string for a molecule is constructed other strings are compared and a similarity is generated for the two strings. This method loses all chemical information held within the molecule. However, it is simplistic in its nature and a very cheap method of scoring relative to other methods.

Similarity and distance coefficients are most commonly used for bioisosterism. Similarity coefficients measure the level of similarity between two molecules, and distance coefficients measure how they differ. Many similarity and distance coefficients are found to be equivalent as they both use the same method but one measures the degree of similarity the other the difference. The six most commonly used coefficients are the Tanimoto, Dice and Cosine coefficients and the Hamming, Euclid and Soergel distances. These coefficients either generate a score between 0 and 1, where for similarity coefficients 1 is maximum similarity i.e. identical and for distances 1 is maximum dissimilarity, or the scores can be normalised to lie within that range. There is no established choice for which coefficient to use, or whether distance or similarity is better. The choice of coefficient tends to be different for each case.

\[ D_{A,B} = \left[ \sum_{j=1}^{n} |x_{jA} - x_{jB}|^t \right]^{1/t} \quad (2.25) \]

\( x_{jA} \) is the value of observable \( j \) for object A, \( t=1 \) for the Hamming distance and \( t=2 \) for the Euclidean distance.

Soergel Distance

\[ D_{A,B} = \left[ \sum_{j=1}^{n} |x_{jA} - x_{jB}| \right] / \left[ \sum_{j=1}^{n} \max(x_{jA}, x_{jB}) \right] \quad (2.26) \]

Dice coefficient

\[ S_{A,B} = \left[ 2 \sum_{j=1}^{n} x_{jA} x_{jB} \right] / \left[ \sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 \right] \quad (2.27) \]
Cosine coefficient

\[ S_{A,B} = \left[ \sum_{j=1}^{n} x_{A,j} x_{B,j} \right] \left/ \left[ \sum_{j=1}^{n} (x_{A,j})^2 \sum_{j=1}^{n} (x_{B,j})^2 \right] \right]^{1/2} \] (2.28)

The QID was already established with the Euclidean distance as its method of scoring. Therefore, it remains the scoring function used in the QID, as most bioisosterism methods rely on either the Euclidean distance or Tanimoto coefficient for scoring.

2.4.8 Alternative methods to bioisosterism

Many new methods have emerged that utilise new techniques and algorithms to generate new ligands from existing ones. Not all of these methods rely on the same processes as bioisostere methods. These alternative methods do not necessarily rely on pre-existing libraries full of fragments with descriptors and use scoring functions as is the case with the majority of bioisostere methods. These methods are more like tools used to simply exchange or replace fragments on existing molecules to provide new suggestions. Some do not even score the generated bioisosteres and instead provide a list of unranked suggestions. These methods tend to be quicker but are very primitive in their nature. A chemist can easily swap two fragments of a molecule with pen and paper. These methods can be seen as just an automated tool and consequently are unable to suggest novel and new bioisosteres with the same level of sophistication as modern bioisosteric replacement methods. The largest drawback is that the effectiveness and ability to suggest multiple bioisosteres is proportional to the amount of data already collected for specific systems. These methods tend to focus on binding of particular enzymes and active sites and therefore have limited transferability between systems.

BREED is a package that does this using a ligand-fragment replacement method. A pre-docked ligand is superimposed onto another docked ligand\(^{103}\). A matching bond is found using a set of criteria, a single bond and its orientation, which can be adjusted to alter the strictness of the match. Fragments on either side of the matching bond are then exchanged to create a new set of ligands. The usefulness of this method lies in the nature of the new ligand generation. Because the new ligands are generated from previous docked ligands and fragments are exchanged at a common point in space the newly generated ligands should show similar docking properties as its predecessors. This
method of ligand-fragment swapping is heavily limited by its requirement of input data in the form of docked ligands. For BREED to be an effective method, large amounts of data are required of ligands that have already been tested, therefore there is less of a chance of it detecting a novel replacement.

GANDI\textsuperscript{\textsuperscript{104}} has gone on to incorporate multi-objective optimisation with a genetic algorithm to produce an altogether more sophisticated approach relative to the fragment swapping of BREED. The multi-objective optimisation was not used for the scoring as previously mentioned in its applications, but used in the library to predict the quality of bioisosteric replacements\textsuperscript{105}. The method involves a protein bound to an enzyme. Replacements were generated from a list of fragments and linkers and their suitability measured using force field energy and the 2D similarity of the inhibitor or the 3D overlaps of the known binding model were optimised. This was done for cyclin-dependent kinases with a fragment library of 14,000 fragments. The method relies on a genetic algorithm to evolve the docked molecule into new replacements. Each suggestion is evaluated in its suitability relative to the original molecule by multi-objective optimisation of the force field energy with either the 2D similarity of the inhibitor or the 3D overlap of the known binding model.

The algorithm is a parallel process that performs simultaneous evolution of multiple docking sites. Evolution of the models occurs through crossing over and mutation. Crossing over involves an exchange of fragments between two docking sites and mutations are the replacement of an existing fragment with a fragment present in another docking site. If the offspring docking sites contain unfavourable interactions or clash with previous models then they are removed from the population. Successful offspring continue to the next stage where a search is conducted to find efficient linking of the fragments into the model. This uses a pre-generated look-up table of bond lengths and angles for the linkers and fragment pairs. Linkers are incorporated into the model and a score is generated for the final molecule. From the scoring of the molecules, fitness is generated for further iterations.

Scoring is calculated with the weighted sum of force field binding energy, with 2D or 3D similarities to the target. Weighting is altered to favour binding energy or the structural binding mode of the molecule. This enables the user to prioritise between binding site or ligand based designs as both are optimised simultaneously. This method makes use of the fact that 2D and 3D shape are important along with the binding
energy. Similar shapes of molecules may not always bind in a similar fashion. The binding may depend on slight variations, such as presence of one vital functional group. If binding energies are found to be similar, it may be the presence of a key functional group that is responsible for the molecule’s activity.

Although GANDI is much more advanced than the previous example of BREED, both methods inherit the same flaw of requiring large amounts of experimental data before they are run. However, GANDI does help to show the successes of newer more sophisticated methods, such as genetic algorithms and multi-objective optimisation and how they can be utilised to create more powerful methods.

BREED and GANDI are fragments based methods, an upcoming alternative to bioisosterism. Fragments based designs use a library of fragments that are combined using a large variety of algorithm. Molecular Dynamics simulations dock the new ligands and scores are generated, usually based on comparison to known NMR or crystallographic data\textsuperscript{106-112}. There are many reviews on the subject that go into greater depth\textsuperscript{113-119}, here the methods are just being highlighted as alternative methods to achieving similar goals to bioisosterism.

However, this brings up the question that even with an evolution of methods and employing newer more powerful techniques, does this level of increased sophistication in the method mean much if the underlying theory is still not robust enough? Both BREED and GANDI rely on previously known data for the system and the accuracy of transferability between systems is not known. The transformation of one to the other is that of a very simple algorithm to more complex genetic algorithm with elements of multi-objective optimisation. However, both methods are limited in their able to generate entirely novel fragments that a synthetic chemist would not naturally think of. The suggested replacements are highly specific to the system they were generated for and have limited transferability. The replacements are only good for the enzyme system they were tested on, a new system means that a new set of starting data is required.
2.4.9 BROOD

In this work the QID tool is tested against the ligand design program BROOD\textsuperscript{120}. Therefore, the method of BROOD is discussed in detail. BROOD is a tool designed to do an isosteric swap, and is not concerned if the replacement is bioisosteric. This is because not all the descriptors in BROOD are chosen for biological importance. The database of fragments held in BROOD consists of 3 million fragments up to fifteen heavy atoms large. The database was built by fragmenting molecules from other databases using the CHOMP program, which cuts molecules into fragments using a specific set of rules to preserve chemical structure and feasibility, like not cutting through cyclic systems. Undesirable functional groups and duplicate entries are filtered out. Geometries for the final set of fragments to be stored in the database are calculated by Omega. Fragments are held in the database at pH 7.4 state.

BROOD determines isosteres by overlaying shape and properties of target and hit fragment, then calculating the Tanimoto. There are three different measures used in BROOD for isosteric replacements. Shape which is measured by standard shape overlay methods. Chemistry is determined by the positions of donors and acceptors, hydrophobic groups, ring groups, with the position and type of group determine, they are overlaid onto other molecules and compared. Electrostatics which are overlays in a method similar to the chemistry properties however instead of using group types, are the group types are replaced by a partial charge calculated from the MMFF\textsuperscript{121} force field. Charges are held within the database for all fragments.

For a molecule of interest, the desired fragment to replace is cut from the molecule by cleaving bonds where the fragment connects to the rest of the molecule. Dummy atoms are added to the connection points where the bonds were severed, up to 3 connection points can exist. Only the vector of the connection point bonds are retained, other information about connection to rest of the molecule is discarded. The user identifies which features they want such as location of donors, acceptors or rings. A protein can be overlaid onto the molecule but it is used at the end of the calculation only to rank ligand replacements. For easy alignment BROOD builds the fragments onto the connection point so they start in the same orientation.

There are three search types within BROOD, shape and chemistry overlap, shape and electrostatics overlap and finally a geometry comparison. The geometry comparison evaluates differences of geometry between molecules through positions of the
connection point bonds. A geometry comparison is used for both the first two search types however it is possible to perform a geometry only search type.

The hits are clustered based on the fragments having similar atomic graphs of their heavy atoms. Typically the results have approximately 50 to 100 clusters to browse through; therefore it is possible to eliminate clusters of similar molecules that are of no interest to the chemist. However, because hits are only clustered by heavy atom graphs, clusters can vary a lot, and higher ranked clusters may have lower ranked molecules within them. The clusters are ranked by the “head” of the cluster. The head of the cluster is the highest scoring fragment in the cluster.

Features are aspects the user would like to see in the hit fragments. The fragments are ranked higher if the feature exists within them. Constraints only return fragments that contain the specified constraint for example, if a ring is desired within the system. BROOD is designed to do searches for fragments that are not active as well, so BROOD searches are only isosteric and non bioisosteric in nature. The user can apply filters to reduce the number of hits, such as only filtering out any fragment that has a molecular weight above a user defined threshold. Other filters include logP, the number of hydrogen bond donors and/or acceptors, synthetic feasibility of the fragments, and the fragments Abbott bioavailability. The Synthetic feasibility score calculation is based on the number of specified connection types within a molecule and rings within the molecule, a set of rules determine the synthetic feasibility of a fragment such as rings with fewer connection points are more synthetically feasible than rings with more connection points, and the compositions of the rings and any substituents are more examples of factors affecting the synthetic feasibility determined by BROOD. The Abbott bioavailability belief model matches the known active fragment to the hit fragment and calculates a Tanimoto score for the overlay of the properties. Abbott bioavailability assumes that the greater the match, the greater the chance that the hit fragment will be bioactive. The Abbott bioavailability gives a probability of the hit fragment being active. BROOD has two forms of search, quick search and full search. A quick search only matches the target with a small random sample of fragments from the database of fragments. A quick search using quinine as the target structure returned no hits and took under 3 minutes. A full search for the quinine replacement can take over 15 to 20 minutes to return over one thousand hits. BROOD is used as a standard to compare the QID’s search capabilities against.
2.5 pKₐ Prediction

2.5.1 Introduction to pKₐ prediction

The QID tool is not currently able to perform pKₐ predictions. Knowledge of the pKₐ can be vital in ligand design. pKₐ prediction tends to lie in a different literature to bioisosterism. Therefore, the pKₐ is now introduced along with its importance to ligand design and the methods of pKₐ prediction discussed.

2.5.2 The importance of pKₐ

The extent of the dissociation of hydronium ions from an acid can be measured by the acid dissociation equilibrium constant (Kₐ) with the formula

\[
K_a = \frac{[H_3O^+][A^-]}{[HA]} \quad (2.29)
\]

where

\[
K_a = \frac{[H_3O^+][A^-]}{[HA]} \quad (2.30)
\]

The huge variance in molecules and by association can differ by many orders of a magnitude and is therefore more conveniently represented as the pKₐ, where

\[
pK_a = -\log_{10} K_a \quad (2.31)
\]

If the extent of dissociation is large, then Kₐ is larger, the pKₐ smaller, and the molecule is considered a strong acid. Knowing the extent to which a compound donates or accepts a proton, or what its preferred protonated state is for a given system of known pH is vital in both chemistry and biological contexts. Pharmacokinetic properties of a molecule are hugely reliant on the protonation state of a molecule. A drug molecule is usually required to pass through a membrane to reach its target site. Ionic species have great difficulty permeating through the membrane where neutral molecules can be more readily absorbed. Once the target is reached, the protonation state of the molecule will in most cases determine the nature of the binding to the active site of the protein. Similarly for agrochemicals, transport and binding of the molecules through phloem mobility behaves with a similar action to drug molecules with the same consequences of protonation state. The binding of the agrochemical to the target site also has a similar dependence as with pharmaceuticals. However agrochemicals have
the added process of soil binding. The ionic state of a molecule contributes towards the 
soil binding of the molecule, if the molecule binds too strongly then it leaches into the soil 
and cannot be taken up by the plant. If the binding to the soil is too weak then the 
agrochemical “runs off” and is lost before the plant can uptake it, or just is not taken up 
by the plant. 85% of current herbicides are weak acids and are only effective within a 2 
$pK_a$ unit range. If the $pK_a$ of a molecule lies outside the range for the target then the 
molecule will either not be taken up by the plant, transported to the correct site, or 
initiate the desired action.

In many cases where modelling biological processes is desirable, such as 
transport of a molecule through a living system, accurate knowledge of the $pK_a$ is 
required. In the case of plants, molecular transport through the system is 
achieved by the phloem. The Trapp model predicts the phloem mobility of a compound 
using the compound’s LogP and $pK_a$ values. This model predicts the extent to which 
the compound is able to travel to the target site. If the model predicts a molecule to be 
unsuitable then it cannot progress any further as a candidate in target development. 
Pre-emptively identifying unsuitable compounds before they reach the experiment 
stages is hugely beneficial, with cost and time saving implications. A further reduction in 
time and costs is achieved through $pK_a$ predictors, which remove the need for 
experimentally determining $pK_a$ values. However, current $pK_a$ predictors still have 
difficulty accurately predicting some novel functional groups found in biological 
molecules. Traditionally $pK_a$ values have been experimentally determined, however $pK_a$ 
predictors are being increasingly relied upon to produce accurate but quicker $pK_a$ values, 
but only for simpler molecules as they are not trusted enough to predict to within the 
desired accuracy range of 0.5 $pK_a$ units.

### 2.5.3 $pK_a$ prediction

The methods for $pK_a$ prediction have become more sophisticated as 
computational power increases. It is now possible to predict $pK_a$ values for larger 
complex molecules such as proteins. However, the complexity of an extensive ligand or 
a protein requires complicated procedures and can introduce large errors. When 
predicting for small molecules, many approaches involve a pharmacophoric method. 
Pharmacophores, which are predetermined features, are identified for a compound and 
are involved in $pK_a$ prediction. Because pharmacophoric features are usually derived 
from data with known $pK_a$ values, these methods tend to rely on databases of
compounds with known pKₐ values. Therefore, it is usually the size and quality of the data set that determines the accuracy of the predictor rather than the method itself.

Recently, quantum chemical methods have emerged for pKₐ prediction. The rate constant of proton dissociation and therefore the pKₐ of a molecule is predicted through \textit{ab initio} calculated thermodynamic properties. Calculating proton dissociation in solution is difficult and requires methods such as thermodynamic cycles and solvation models. These limitations in calculating the free energy cause errors in the pKₐ prediction because the exact free energy is related to the pKₐ. With so many different approaches to predicting pKₐ values, many descriptors and methods exist. This has led to a number of pKₐ prediction packages, all competing to predict pKₐ values the most accurately.
Chapter 3

Bioisosterism of heterocyclic fragments

3.1 Quantum Chemical Topology of cyclic molecules

3.1.1 Quantum Chemical Topology of ring atoms

This chapter establishes methods for characterising rings through QCT properties. If rings can be characterised through QCT properties then comparisons between the rings can be made for bioisosteric replacement. Both the RCP properties and the properties of the ring atoms are explored as potential descriptors.

The QCT partitioning of a molecule lends itself to localise chemical information in rings and their constituent atoms because individual atoms within the ring are naturally defined. Previous QCT work on rings calculated properties such as the delocalisation of the electron density onto the ring\textsuperscript{158,159} or the extent of aromaticity calculated for the ring\textsuperscript{34,160-163} but the QCT atoms themselves are rarely considered. The properties of a given QCT atom are influenced by other atoms in the molecule. Atomic properties respond to changes in environment and therefore the position of an atom within a ring and its neighbours determine the atom's properties.

Figure 3.1 visualises the changes observed in QCT atoms relative to their environment (or position in the ring). It is convenient to normalise the monopole moments of all five ring atoms, by setting the smallest and largest value to 0 and 1, respectively. The normalised monopole moments are represented by a colour on a scale that progresses through the rainbow from violet (most positive) to red (most negative).

The carbon in between the two nitrogens is found to be much more positive relative to the other carbons within the ring, while the -NH nitrogen is more negative than the =N- nitrogen. Interestingly, a difference in charges also appears between the two carbons adjacent to each other, although the difference is much more subtle. The carbon adjacent to the -NH nitrogen is marginally more negative than the carbon adjacent to the =N- nitrogen. This is because the more negative –NH nitrogen has a more pronounced effect on its adjacent atoms than the effect of the =N- nitrogen on its own adjacent atoms. These observations show that an atom's element and its
environment both affect the atom’s properties and these observations are reflected within the QCT atoms.

Figure 3.1: Normalised monopole moments for the non-hydrogen atoms within imidazole, represented by colours in the rainbow spectrum where violet = most positive and red = most negative.

3.1.2 Ring atoms and substitutions

QCT atoms of a ring can also reflect any changes in substituents onto the ring atoms. A substitution onto a ring atom will change this atom’s basin and therefore affect this atom’s properties as well. As an example, Figure 3.2 shows how the atomic basin of a carbon atom within an imidazole ring changes when the hydrogen bonded to the carbon is substituted for an alcohol group. The highlighted basin (green) is crushed upon substitution of the hydrogen with the alcohol, the basin changing from the blue outline to the red one. The oxygen of the alcohol has a much larger electronegativity than the hydrogen of the non-substituted imidazole, and therefore a larger portion of the electron density between the carbon and the oxygen is attributed to the oxygen atom. The two nitrogen basins are also affected by the substitution as the carbon basin is pushed downwards, which causes a noticeable spread in the basin at either side.
3.1 Relating Ring Critical Point properties to ring atom properties

3.1.1 Calculation of ring critical point and ring atom properties

A series of five-membered heterocyclic ring structures were geometry optimised at the B3LYP/6-311+G(2d,p) level of theory. The QCT properties were calculated from the resulting wave functions for the five atoms in a ring only. Note that the atomic ring properties listed below are always referring to the sum of these properties over all five atoms in the ring. Data set A consists of 20 molecules, which are shown in Table 3.1 as a series of five-membered ring structures without substituents. The structures in the set contain rings consisting of C, O, N, S and P atoms only, and all are unsaturated heterocyclics. The general scaffold of the rings is shown in Figure 3.3. The QCT atomic properties calculated were the volume, the total atomic energy $E$, the monopole moment and the magnitudes of the atomic dipole and quadrupole moments. The RCP properties calculated were the electron density $\rho$, its Laplacian $\nabla^2 \rho$, the curvatures $\lambda_1$, $\lambda_2$, $\lambda_3$, and the RCP ellipticity $\varepsilon$. 

Figure 3.2: Superposition of imidazole and imidazole-2-ol with the change in atomic basin highlighted for the carbon upon substitution of H (blue contour) with OH (red contour).
Table 3.1: Non-substituted five-membered unsaturated rings in data set A.

<table>
<thead>
<tr>
<th>Furan</th>
<th>Isothiazoline</th>
<th>Phosphole</th>
<th>Thiazole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Furazan</td>
<td>Isoxazole</td>
<td>Pyrazole</td>
<td>Thiazoline</td>
</tr>
<tr>
<td>Imidazole</td>
<td>Isoxazoline</td>
<td>Pyrazoline</td>
<td>Thiophene</td>
</tr>
<tr>
<td>Imidazoline</td>
<td>Oxazole</td>
<td>Pyrrole</td>
<td>1,2,3-triazole</td>
</tr>
<tr>
<td>Isothiazole</td>
<td>Oxazoline</td>
<td>Tetrazole</td>
<td>1,2,4-triazole</td>
</tr>
</tbody>
</table>

Figure 3.3. The general scaffold for the molecules the completed data set. Rn represents the ring atom where R = C, O, N, S or P and n is the position on the ring. Sn is a substitution on the ring and can be H, OH, OMe, NH$_2$, NHMe, NO$_2^-$, CN, F, COOH or nothing if the adjacent ring atom is already at full valence, e.g. S, O or =N—. The two dashed bonds can each be either a single or double bond.

3.1.2 Non-substituted rings

Finding a relationship between the ring atom properties and the RCP properties is achieved through Partial Least Squares regression (PLS). PLS reduces the complicated multivariable problem to a clearer picture and the quality of the relationship is quantified through $r^2$ values for the predicted output variables correlated with the observed output variables. A series of PLS models were created, which differed in the selection of the input and output variables. For the first PLS model all five atomic properties were selected as inputs and all six RCP properties as the output variables.

Figure 3.4 focuses on the prediction of only one property, made by this first PLS model, which is the RCP electron density. This figure plots the actual observed RCP electron density against the predicted one, with an $r^2$ value of 0.91. This shows that a clear relationship exists between the ring atom properties and the RCP properties. Therefore, it is possible to predict the RCP properties if the ring atom properties are known. The separate contributions of the electrostatic and energy variables were determined through two different models, one model containing only electrostatics input variables and the other only energy input variables. For both models the output
variables selected were all six RCP properties. The “electrostatic only model” (i.e. monopole, dipole and quadrupole moment) has an $r^2$ of 0.89. The “energy only model” has an $r^2$ of only 0.71. Therefore, it is the electrostatic variables that provide the greatest contribution towards predicting the RCP properties from the ring atom properties.

![Graph](image)

**Figure 3.4:** Correlation ($r^2=0.91$) between the observed (actual) RCP electron density and the one predicted by the first PLS model for non-substituted five-membered heterocyclic molecules (data set A).

### 3.1.3 Substituted ring systems

Data set B is an expansion of data set A, which now includes mono-substituted rings. For each of the 20 molecules, the hydrogen atoms were sequentially replaced by a simple functional group. The substituents were –OH, –NH$_2$ and –OCH$_3$, creating a set of 178 mono-substituted ring molecules. As an example, Figure 3.5 shows all six possible substituted rings (3 substitutions x 2 positions) for furan. The symmetry of furan makes substitution at the two remaining substitution sites unnecessary. The alcohol and amine groups were chosen because they are the simplest substituents that are found readily in many biologically relevant molecules. The methoxy group was chosen to elucidate both the effect of replacing the hydrogen of the alcohol with a methyl, and the perception of this replacement by the ring atoms and RCP properties. Therefore, it is possible to determine the physical distance in 3D at which the atoms of the substituent cease to significantly affect the ring atom and RCP properties. This distance allows for truncation of the substituents and therefore reducing the number of substituents required for modelling. For example, if the effect of butyl and longer alkyl substituents on the ring properties is comparable to the effects of a propyl then the propyl is sufficient to represent all larger alkyl substituents. Note that the number of possible substituted sites
varies from two to four, depending on the symmetry of the ring structure itself and the number of available substitution sites (determined by the atoms in the ring).

![Image of furan substitution patterns](Image)

**Figure 3.5**: The six possible substitution patterns in furan. Note that the symmetry of furan would create unwanted duplicates if it were substituted on the two remaining sites.

The PLS analysis was repeated for a new dataset (B) containing both non-substituted and substituted rings. As with the PLS models for data set A, variables included in the models were altered to assess the relationship of RCP properties to ring properties. The effect of reversing the ring atom and RCP properties as input and output variables of the model was also explored.

### 3.1.4 The effects of a ring’s atomic and RCP properties

Figure 3.6, which is the counterpart of Figure 3.4, shows the PLS model for all five atom properties as inputs and all six RCP properties as outputs for the data set B.

![PLS model for five-membered heterocyclic molecules](Image)

**Figure 3.6**: PLS results for alcohol, ether, amine and non-substituted five-membered heterocyclic molecules.
The corresponding $r^2$ value is 0.80 whereas before, for dataset A it was 0.83. Table 3.2 shows the $r^2$ values for a variety of PLS models (constructed in sets A and B), which differ in their selected input and output variables. The $r^2$ values of the data set that includes substituted rings (B) are always lower correlation than those of the data set without substituents (A), but only by 0.03 on average. The models show that prediction of the ring atom properties from the RCP properties is much more difficult than the reverse, as the largest $r^2$ of these models is only 0.59. However, the $r^2$ value never falls below 0.80 for any model (top two entries) that predicts RCP properties from the ring atom properties. This suggests the existence of some relationship between the atomic properties of the ring atoms and the RCP properties but the information contained in the RCP properties is not enough to predict the ring’s atomic properties. Further discussions regarding the models will always refer to the $r^2$ values of data set B only because correlations for larger data sets are more reflective of the model of interest.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input variables</th>
<th>Output variables</th>
<th>$r^2$ of Data set A</th>
<th>$r^2$ of Data set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AP</td>
<td>RP</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>AP</td>
<td>RP except ellipticity</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>RP</td>
<td>AP</td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td>4</td>
<td>RP except ellipticity</td>
<td>AP</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>RP except ellipticity</td>
<td>Ellipticity</td>
<td>0.93</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 3.2: $r^2$ values for PLS models with different selected input and output variables. AP = Atom properties, RP = RCP properties. For clarity the RCP properties are highlighted. The third and fourth entries are the reverse of the first two entries.

3.1.5 The role of ellipticity

Of the first four models that relate the atomic and RCP properties, the model with the largest $r^2$ is 0.90 (second entry) which has atomic properties as inputs and RCP properties as outputs, with ellipticity excluded. This suggests that the ellipticity at the RCP is the most difficult RCP property to predict. However, predicting the ring’s atomic properties from the RCP properties both with and without the ellipticity produces similar correlations of 0.57 and 0.56, respectively. But then again, an $r^2$ value of 0.57 is
rather poor and prevents one to be sure about causality in the first place. The high $r^2$ value of 0.9 for the model with RCP properties (except ellipticity) as inputs and the ellipticity as an output (fifth entry in Table 3.2) is anticipated because the ellipticity is a non-linear function of the curvatures $\lambda_2$ and $\lambda_3$ as explained in Section 2.2.2, that is, $\varepsilon_{\text{RCP}} = (\lambda_3 / \lambda_2) - 1$. Although PLS is a linear correlator it manages to capture most of this non-linearity. In summary, the decrease in the $r^2$ value when ellipticity is included in the output variables suggests that it must be more sensitive, or receptive, to changes in the ring atoms.

To determine the relationship of the ellipticity and ring atom properties a series of plots were constructed comparing a single RCP property to a single atomic property. Figure 3.7 compares each three (out of a total of six) RCP properties to two (out of a total of five) atomic properties, leading to $3 \times 2 = 6$ panels. Panels (a), (b) and (c) show the kinetic energy $K(\Omega)$ (which is more convenient than $E(\Omega)$, see eq. (2.9)) versus the electron density, its Laplacian and the ellipticity. Panels (d), (e) and (f) show the monopole against the same three RCP properties. Figures 3.7a, b, d, and e show that the electron density and its Laplacian have similar relationships to the ring atom properties. Indeed, the corresponding panels (3.7a and 3.7b; 3.7d and 3.7e) have the same general shapes. The general shape consists of three clusters, one small (top of panel), one medium (middle of panel) and one large (bottom of panel). The middle cluster has a subcluster directly beneath it and to the left of its centre. The small cluster (top) has the lowest RCP property (electron density or its Laplacian) followed by the medium cluster and finally the largest cluster. Interestingly, the ellipticity produces graphs (3.7c and 3.7f) of similar shape but over a larger range and mirrored, which is explained as follows. The majority of data points in the middle cluster lies to the left of the top cluster while the middle cluster’s subcluster now lies to the right (rather than the left). The bottom cluster lies predominantly to the bottom left of the graphs. The increased spread in the ellipticity stretches the graphs along the horizontal axis and distorts the shape relative to the top four panels (3.7a, b, d, and e). This increased spread marks a heightened sensitivity of the ellipticity towards the ring atom properties. In other words, if the ellipticity is predicted it will demand a more accurate model than the electron density or its Laplacian will demand. This observation may explain why the $r^2$ value for a PLS model with the ellipticity included (model 1 in Table 3.2) is lower than a PLS model without it (model 2).
3.1.6 Local relationships between ring scaffolds

In order to discover if relationships between ring scaffolds can be extracted from the RCP properties and ring atom properties, the raw data were examined “locally”, through four types of rings. Indeed, analysis of the whole of data set B can mask any local information, which may not be apparent when examining the data set in its entirety. The four types of rings are 1,2,3-triazole, 1,2,4-triazole, isoxazole and imidazole. The two triazoles were chosen because they differ by the position of their heteroatoms within in the ring. Imidazole was chosen because it only has two nitrogens.

Figure 3.7: Comparison of RCP properties with two different ring atomic properties (kinetic energy (K) and monopole) for data set B.
in the ring compared with the three in triazole, and oxazole because it has two different heteroatoms in the ring (N and O).

Figure 3.8a to 3.8d shows the overall monopole moment of the whole ring (i.e. number of electrons) for the four ring types as a function of three possible RCP properties. Figure 3.8a shows that each ring type is discernible from the three others as there is no overlap between the four clusters. In other words, each ring type has a distinct signature of monopole moment and RCP electron density. For example, all 1,2,3-triazole molecules lie within a range of 33.0 to 33.6 a.u. for their monopole moment and 0.059 to 0.063 a.u. for their electron density, while none of the other three rings have any data points within this range. This unique range of properties for each ring also occurs for the Laplacian in Figure 3.8b. However, there is an overlap between the 1,2,3-triazoles and 1,2,4-triazoles, which is not unexpected due to their close structural relationship. Again, the ellipticity recreates the general shape of the graphs for the other RCP properties but the shape is again stretched out (see Figure 3.7). This stretching leads to a greater overlap between the ring clusters, due to the ellipticity’s increased sensitivity to the variation of ring atoms, thus confirming the ellipticity’s difficulty in discerning between ring types and also explaining why PLS models struggle to predict the ellipticity. Figure 3.8 determines the effect of the substituents on a ring’s properties. Figure 8d plots the same data as Figure 3.8a but re-coloured (or relabelled) such that data points are grouped by substituent type rather than by ring type. The patterns of the different substituents within each ring type are remarkably similar throughout all four ring types. This repeated pattern begins with a non-substituted ring at the top, followed by a set of amines just below it, and followed again by the alcohol and methoxy substituents, further down. Interestingly, each alcohol data point is paired with a methoxy data point. These pairs are both substitutions onto the same available substitution site on the ring. Therefore, the RCP and ring atoms see the alcohols and ethers (i.e. methoxy) as very similar, suggesting that it is the atom of the substituent directly bonded to the ring that affects the properties more than the entire substituent.
Figure 3.8: Comparison of four different ring scaffolds with all substitution data points for each scaffold using ring atom monopole moments and RCP properties. The monopole moment and RCP electron density plot is drawn in two forms, the data set divided into ring scaffolds (b) and substitution types (d).

3.2 Ring atom properties

3.2.1 Ring characteristics and ring properties

In this section the ring's composition (marked by four indicators: number and type of heteroatoms, internal bonding structure, and substitutions onto the ring) is related to the atomic ring properties. Only the contributions of the five ring atoms towards the atomic ring properties are considered. In other words, all properties were calculated using only the five atoms in the ring (excluding hydrogens and the atoms of the substituents). Confining the calculations to these five atoms enables one to explore the extent of distortion to a ring atom's basin upon change of its environment, as explained in Section 3.1.2.
Figure 3.9 shows the net charge for the ring atoms of a new selection of 16 heterocyclic rings each with eight different substituents, which were substituted onto the rings in the same manner as outlined in beginning of Section 3.1.3. The eight substituents were OH, OMe, NO$_2$, NH$_2$, NHMe, CN, F and COOH. The rings were selected as four groups of four, totalling 16. Within each group of four, the rings only differed in the element of one of the heteroatoms. For example, pyrrole (N), furan (O), thiophene (S) and phosphole (P) are the first group of four (panel a). The second group of four differs from the first group by an additional nitrogen at the second position of the ring (panel b). For the third group the additional hydrogen is now at the third position of the ring (panel c). The fourth group has two nitrogens in addition to the heteroatom at the first position of the ring. For this group, the two nitrogens are at the second and fifth position in the ring.

Each profile line in any panel of Figure 3.9 represents the average net charge of rings for the substituent of interest. For example, the average net charge for the nitro-substituted pyrrole rings is -0.2 a.u. (panel a). The average is over the three possible substitution sites. The ninth profile line in each panel refers to the non-substituted rings.

For each substituent the net charge follows the same profile as the heteroatom at the first position of the ring is changed, for all four panels. This shows that a change of the elements within the ring has the same systematic effect on the net charge, such that replacing a nitrogen with an oxygen increases the net charge by approximately 0.4 a.u. irrespective of the ring. Moreover, this systematic nature is also found for substituents because each substituent again has the same consistent effect on the net charge, regardless of the ring characteristics. This regularity expresses itself through the nine parallel profiles found in each panel, one profile for each substituent. For example, substituting a F with NO$_2$ decreases the net charge by approximately 0.2 a.u. for all rings.

In summary, there are two "characteristics" of the rings that are independently responsible for the ring's net charge. The first is changing the element of the ring, i.e. moving horizontally between rings, and the second is changing the substituent, i.e. moving vertically between substituents. This introduces the concept of ring characteristic orthogonality, where different characteristics of a ring affect the ring's properties orthogonally. In other words, each characteristic is independently responsible for the ring's properties. The concept of ring characteristic orthogonality is extended to four characteristics as the data set is expanded to include a larger variety of rings.
Figure 3.9: Net charges for ring atoms of substituted five-membered heterocycles, where the heteroatom at the first position is varied for (a) single heteroatom ring scaffolds, (b) additional nitrogen at the second position, (c) additional nitrogen at the third position, (d) additional nitrogens at second and fifth positions.

3.2.2 Discovering all rings features

Figure 3.10 shows the net charges, as before summed over the five ring atoms, for 572 ring molecules separated into nine substitutions (8 substitutions + H [non-substituted]). The 572 consists of 20 different ring scaffolds, totalling 69 rings separated by possible substitution site. For example, furan, from the set of 20 ring scaffolds, becomes furan-2 and furan-3. This totals 69 ring structures shown in Table 3.3. Each of the 69 rings is substituted with 8 substitution types, and the 20 non-substituted rings gives (69 x 8) + 20 = 572 unique ring structures. The rings were ordered along the x axis.
of Figure 3.10 by a hierarchy of characteristics to maintain the orthogonality of the ring characteristics. The first separation in the hierarchy is into the elements of the rings, in a manner analogous to Figure 3.9. The first rings contain nitrogen as their only heteroatom, followed by rings with an oxygen atom, then rings with sulfur and finally rings with phosphorous. The second separation of the hierarchy divides the rings of each element into single then double π-bond rings. The third hierarchy separation orders the rings within each π-bond group by increasing number of heteroatoms in the ring, from lowest to highest.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pyrazoline-1</td>
<td>24</td>
<td>1,2,4-triazole-3</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>pyrazoline-3</td>
<td>25</td>
<td>1,2,4-triazole-5</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>pyrazoline-4</td>
<td>26</td>
<td>tetrazole-1</td>
<td>49</td>
</tr>
<tr>
<td>4</td>
<td>pyrazoline-5</td>
<td>27</td>
<td>tetrazole-5</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>imidazoline-1</td>
<td>28</td>
<td>oxazoline-2</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>imidazoline-2</td>
<td>29</td>
<td>oxazoline-4</td>
<td>52</td>
</tr>
<tr>
<td>7</td>
<td>imidazoline-4</td>
<td>30</td>
<td>oxazoline-5</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>imidazoline-5</td>
<td>31</td>
<td>isoazoline-3</td>
<td>54</td>
</tr>
<tr>
<td>9</td>
<td>pyrrole-1</td>
<td>32</td>
<td>isoazoline-4</td>
<td>55</td>
</tr>
<tr>
<td>10</td>
<td>pyrrole-2</td>
<td>33</td>
<td>isoazoline-5</td>
<td>56</td>
</tr>
<tr>
<td>11</td>
<td>pyrrole-3</td>
<td>34</td>
<td>furan-2</td>
<td>57</td>
</tr>
<tr>
<td>12</td>
<td>pyrazole-1</td>
<td>35</td>
<td>furan-3</td>
<td>58</td>
</tr>
<tr>
<td>13</td>
<td>pyrazole-3</td>
<td>36</td>
<td>oxazole-2</td>
<td>59</td>
</tr>
<tr>
<td>14</td>
<td>pyrazole-4</td>
<td>37</td>
<td>oxazole-4</td>
<td>60</td>
</tr>
<tr>
<td>15</td>
<td>pyrazole-5</td>
<td>38</td>
<td>oxazole-5</td>
<td>61</td>
</tr>
<tr>
<td>16</td>
<td>imidazoline-1</td>
<td>39</td>
<td>isoazoline-3</td>
<td>62</td>
</tr>
<tr>
<td>17</td>
<td>imidazoline-2</td>
<td>40</td>
<td>isoazoline-4</td>
<td>63</td>
</tr>
<tr>
<td>18</td>
<td>imidazoline-4</td>
<td>41</td>
<td>isoazoline-5</td>
<td>64</td>
</tr>
<tr>
<td>19</td>
<td>imidazoline-5</td>
<td>42</td>
<td>furazan-3</td>
<td>65</td>
</tr>
<tr>
<td>20</td>
<td>1,2,3-triazole-1</td>
<td>43</td>
<td>thiazoline-2</td>
<td>66</td>
</tr>
<tr>
<td>21</td>
<td>1,2,3-triazole-4</td>
<td>44</td>
<td>thiazoline-4</td>
<td>67</td>
</tr>
<tr>
<td>22</td>
<td>1,2,3-triazole-5</td>
<td>45</td>
<td>thiazoline-5</td>
<td>68</td>
</tr>
<tr>
<td>23</td>
<td>1,2,4-triazole-1</td>
<td>46</td>
<td>isothiazoline-3</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 3.3: Data set of unsaturated ring scaffolds separated into the different substitution sites possible for each ring scaffold.
Each of the characteristics of the hierarchy is examined in turn to determine their effect on the net charge and if they remain within the scope of ring characteristic orthogonality. The first separation in the hierarchy is for the heteroatom elements. This separation is already known to be orthogonal with respect to substituents based on Figure 3.9. Where Figure 3.9 created a profile from four rings at a time, Figure 3.10 recreates the same profile but for all 69 rings simultaneously. In Figure 3.10 four bands appear where each band corresponds to the element of the ring. For example, the range of net charges (of all ring atoms) for the nitrogen band is -0.6 to 0.2 a.u., that of the oxygen and sulfur band is -0.3 to 0.5 a.u., and that of the phosphorous band is 0.2 to 0.8 a.u.. These ranges are the same as those found in Figure 3.9. Therefore, if each band is considered as a single point on the profile, then the bands match up to the four-point profile of Figure 3.9. The orthogonality of heteroatom element successfully extends from the 16 scaffolds of Figure 3.9 to all 69 rings in Figure 3.10. In other words, the first data set of four ring type (Figure 3.9) is already representative for the full data set (Figure 3.10).
The second separation in the hierarchy divides each element’s band into rings with a single π-bond and a double π-bond. The single π-bond rings occupy the first half of a band while the double π-bond rings occupy the second half. There are no discernible differences between the first half (single π-bond) and the second half (double π-bond) of this hierarchy separation. However, the first halves show erratic spikes within the net charge. These spikes are caused by substitutions onto nitrogens or sp³ carbons. Note that the single π-bond rings also have sp² carbons and therefore some net charges are similar to the net charges of the double π bond rings, because substitutions occur onto the same type of carbon. This shows that the internal bonding structure of the ring indirectly affects the net charge profile as a single π-bond ring introduces sp³ carbons that affect the net charge differently to sp² carbons. Even though the removal of one π-bond from the ring does not directly affect the ring’s net charge, the effect of substituting onto an sp³ carbon relative to an sp² carbon is still orthogonal to the other characteristics. The orthogonality remains because the same spikes are found throughout all the element bands and are relatively parallel.

The third stage of the hierarchy is the ordering of rings within the single or double π-bond separations by increasing number of heteroatoms within the ring. For example, in the nitrogen double π-bond separation the rings are ordered from a single nitrogen in the ring (pyrrole) to four nitrogens in the ring (tetrazole). The effect of the number of heteroatoms within the ring is best observed in double π-bond separations. This is because the erratic spikes found in the single π-bond separation (discussed above) are greater than any changes in net charge due to the number of heteroatoms, and therefore mask the effect. Therefore, if the oxygen double π-bond separation is examined, a small increase in net charge is found as the number of heteroatoms in the ring increases. Again, the effect of the number of heteroatoms within a ring on the net charge is orthogonal to changes due to other characteristics. However, the change in net charge due to the number of heteroatoms in a ring is occasionally masked by changes due to the π-bond structure of a ring.

It is also worth noting that for any ring pattern with an N-H, substitutions at this position tend to give very different ring net charges relative to substitutions at any other position in the ring. However, the change in net charge for substitutions onto this position is consistent. Unexpected net charges occur when an NO₂ is substituted onto sp³ carbons, producing net charges similar to or lower than NH₂ and NHMe where the
NO$_2$’s net charge is expected to be larger than both. This explains why the pink and blue profiles in Figure 3.10 occasionally cross each other.

The three characteristics (hierarchy separations) discussed above and the parallel profiles of the substituents total four orthogonal characteristics that determine a ring’s net charge. In other words, each characteristic independently affects the ring’s net charge. For example, replacing the F with an OH in a mono-substituted pyrrole, where either F or OH can be on any position of the ring, decreases the net charge by 0.1 a.u.. Furthermore, if this new ring requires a net charge increase of 0.3 a.u., this is achieved through replacing the nitrogen with oxygen and forming furan-ol. This orthogonality of the four characteristics allows one to traverse the net charge landscape by manipulation of any of the ring’s four characteristics.

3.2.3 Effect of substitution on ring properties

Now it is possible to exploit the parallel nature of the profiles of net charge variation throughout the set of ring scaffolds in Table 3.3. More precisely, one can move from one profile to another and observe a change in net charge specific to a change in substituent. This observed change in substitution and net charge can be applied for prediction purposes. For example, the change in net charge going from a non-substituted to an F substituted ring is approximately 0.6 a.u. for all rings. Predictions can be made by taking the non-substituted ring and adding the mean change in net charge for each substitution. Therefore, if the net charge of a ring scaffold (Rn in Figure 3.3) is known then the new net charge of the substituted ring can be predicted. This procedure can be repeated for many cases as is done in Figure 3.11. Figure 3.11 is the result of applying the mean change in net charge upon substitution to non-substituted rings for each of the substituents. It should be noted that because the lines are parallel, the use of mean values to predict net charges can be generalized between any two substituents. Comparing Figure 3.11 to Figure 3.10, the three large bands are maintained and the parallel lines for the substitutions also remain. Hence, the general shape of Figure 3.11 matches that of Figure 3.10. However, many of the smaller peaks of Figure 3.10 are not present in Figure 3.11 because the site of the substitution is not considered in the calculation of the mean change.
Figure 3.11: Ring atom net charges predicted from the sum of a mean net charge change of a given substitution and the net charge of a non-substituted ring.

In order to recover the peaks that are absent in Figure 3.11 one must account for the site of the substitution in the calculation of the mean change in net charge. This calculation was refined by calculating a separate mean for substitutions onto nitrogens (N-H), sp\(^2\) or sp\(^3\) carbons. Table 3.4 shows the mean net charge change for each substitution with respect to the site of the substitution, as well as the corresponding standard deviations, and the latters’ proportion (as %) of the respective means. For each of the substituted rings the appropriate mean change in net charge (determined in Table 3.4) was applied to the non-substituted ring’s net charge. In all cases except five, the standard deviation of the change in net charge is below 20% of the mean change in net charge. Therefore, the changes in net charge due to substitution are similar within each substituent and site. COOH substitutions have a very large standard deviation relative to mean change. However, this may be due to the small mean change (i.e. -0.01) between COOH and non-substituted rings. The high percentage is caused by a typical standard deviation being divided by an untypically small mean. Indeed, the COOH substitution (0.04, 0.06 and 0.02, for sp\(^2\) C, sp\(^3\) C and N, respectively) is similar to the standard deviations of all other substituents (ranging from 0.02 to 0.07, 0.03 to 0.07, and 0.01 to
0.03, for sp\(^2\) C, sp\(^3\) C, and N, respectively) but the mean change is much smaller. Figure 3.12 shows the replotting of Figure 3.11 where the appropriate mean change for each substitution with respect to the site of the substitution is applied. Comparing the new predictive model (Figure 3.12) to the original calculated values (Figure 3.10) shows that the majority of peaks are now present in the graph where they were absent when the substitution site was ignored (Figure 3.11).

The site of the substitution has been shown to significantly affect the ring’s net charge. The changes show the same systematic orthogonal change as shown by the previous four ring features. However, the site of substitution is similar to the \(\pi\) bond feature, where it was the substitution onto sp\(^2\) and sp\(^3\) atoms that was found to affect the ring’s net charge and not the addition or removal of the \(\pi\) bonds themselves. Therefore, the number of orthogonal ring features remains at four but the \(\pi\) bond feature and the site of substitution are considered as the same feature.

<table>
<thead>
<tr>
<th></th>
<th>NO(_2)</th>
<th>CN</th>
<th>F</th>
<th>COOH</th>
<th>OH</th>
<th>OMe</th>
<th>NH(_2)</th>
<th>NHMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp(^2) C</td>
<td>Mean</td>
<td>0.29</td>
<td>0.14</td>
<td>0.55</td>
<td>-0.01</td>
<td>0.48</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>0.07</td>
<td>0.02</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>S. D. %</td>
<td>25</td>
<td>18</td>
<td>8</td>
<td>510</td>
<td>7</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>sp(^3) C</td>
<td>Mean</td>
<td>0.42</td>
<td>0.20</td>
<td>0.66</td>
<td>0.08</td>
<td>0.56</td>
<td>0.56</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>0.07</td>
<td>0.04</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>S. D. %</td>
<td>16</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>11</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>Mean</td>
<td>0.38</td>
<td>0.03</td>
<td>0.64</td>
<td>0.02</td>
<td>0.47</td>
<td>0.47</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>S. D. %</td>
<td>6</td>
<td>74</td>
<td>5</td>
<td>83</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

* Values in table only shown to 2 d.p.

**Table 3.4.** Change in net charge upon substitution onto a non-substituted ring for eight functional groups split into the three possible atom types for the site of the substitution.
Figure 3.12. Ring atom net charges predicted from the sum of a mean net charge change of a given substitution onto different atom types (sp\textsuperscript{3} carbon, sp\textsuperscript{2} carbon or N(-H) ) and the net charge of a non-substituted ring. Numbers along the x axis correspond to the rings in Table 3.3.

3.2.4 Hammett constants

The order of the substituents (Figure 3.10: highest profile line, F, to lowest profile line, H) expresses to what extent they alter the electronic charge of the ring. An interesting question is whether this order matches the order of functional groups of the corresponding Hammett constants. The Hammett constant is a measure of a substituent’s electron donating or withdrawing properties. Starting with the most electron withdrawing they are ordered as follows: NO\textsubscript{2} > CN > F > H > OMe > NH\textsubscript{2}. On the other hand, the order of the substituents by decreasing ring net charge is F > NO\textsubscript{2} > OMe > NH\textsubscript{2} > CN > H. Clearly, the net charge variation of the substituents does not match the order of their Hammett constants. A more informed way of answering the question above realizes that Hammett constants may correlate much better to the charge due to the rings’ π electron density only\textsuperscript{164}. In order to test this, the π molecular orbitals were determined for each molecule by visual inspection of the molecular orbitals using the program GAUSSVIEW\textsuperscript{165}. The program MORPHY was modified to recalculate the net charge of all ring atoms, but now from electrons occupying π orbitals only. This exclusion of σ molecular orbitals was tested for furan and imidazole. It turned out that the ring net charges for the substituents in both furan and imidazole now matched the order of the
Hammett constants, when only the $\pi$ electrons were considered. This confirmed that the substituents were having the correct effect on the ring electron density, which was not confirmed by initial calculations.

### 3.2.5 Ring substitution effects

The effect of substitutions on the net charge has already been explored with regards to the parallel profile lines. However, the substituents themselves have so far been ignored. In other words, the significance of the positions of the profile lines with respect to each other demands attention. Examining the profiles with respect to each other can determine if a pattern between the substituents exists or which features of a substituent affect the change it imposes on the net charge. All substituents generally form parallel profiles and therefore each substituent causes a systematic effect on the net charge specific to the substituent. However, strictly speaking, there are three other cases where the substituent breaks from this parallel arrangement. Fluorine remains parallel throughout but the changes in the net charge tend to be less pronounced, i.e. the F profile is flatter than that of other substituents. Conversely, rings substituted with a CN group tend to vary more than for other substituents, but this is only observed in the nitrogen band (first third of Figure 3.10). The oxygen, sulfur and phosphorous ring patterns with a CN substituent behave in a parallel fashion similar to other substituents. The third case is NO$_2$, where its profile occasionally drops below the NH$_2$ and NHMe profiles but this case has been discussed in the previous section.

Interestingly, the plots for OH and OMe almost follow each other exactly. The net charge of the ring does not seem to see the effect of a replacement of H by Me on the oxygen of the substituent. This case also occurs with NH$_2$ and NHMe. These observations call for further calculations attempting to generalize this phenomenon. The substitutions were chosen to explore three aspects of a substituent’s effect on the ring’s net charge:

1. **Replacement of a hydrogen atom with a methyl group** – This replacement determines the effect of extending the substitution fragment, and has already been witnessed for the substitution of OH to OMe, and NH$_2$ to NHMe.

2. **Inserting a methylene group between the ring and original substituent** – This addition determines if the methylene group acts as a "buffer" for the interaction of
the substituent with ring. In other words, the extent to which the effect of the original substituent permeates through the additional CH$_2$.

3. *Replacement of a hydrogen atom by CN* – This replacement determines the effect of a second functional group on the ring's net charge as this effect has to permeate through the original substituent, similarly to point 2. However, this replacement also determines if a second functional group will have a significant effect on the original substituents, and alter how the ring’s net charge responds to the original substituent.

The ten possible substituents are: COOH, COOMe (applying point 1. to COOH), C≡N, OH, OMe (applying point 1. to OH), OC≡N (applying point 3. to OH), NHC≡N (applying point 3. to NH$_2$), NHCH$_2$C≡N (applying points 2. and 3. to NH$_2$), CH$_2$OH (applying point 2. to OH) and CH$_2$CH$_2$OH (applying point 2. twice to OH).

![Figure 3.13](image)

**Figure 3.13**: Effects of substitution on ring atom properties. Substitutions were chosen to represent the effects of a functional group's proximity to the ring on the ring atom properties. The bold highlighted profiles are all substituents containing a cyano group.

Figure 3.13 shows the net charges of all the substituents onto all available substitution sites of imidazole. The substitutions are divided into three sets, the top set (between 0 and -0.15 a.u.), the middle set (between -0.17 and -0.35 a.u), and the lowest set (between -0.4 and -0.65 a.u.). The top set consists of OH, OMe and OC≡N, which are
all the substituents connected to the ring through an oxygen. The middle set consists of NHC≡N and NHCH₂C≡N, which are all the substituents connected to the ring through a nitrogen. The bottom set consists of COOH, COOMe, C≡N, CH₂OH and CH₂CH₂OH, which are all the substituents connected to the ring through a carbon. This separation into sets based on connecting element rather than substituent becomes more remarkable when examining all four substituents that contain the cyano group: C≡N, OC≡N, NHC≡N, and NHCH₂C≡N. Naturally, one would expect the four CN substituents to form a set, however, they are separated into the three sets by their connecting atom. Therefore, it is the element of the atom connecting the substituent to the ring that dominates the effect of the substituent on the ring’s net charge rather than the substituent itself.

The addition of the methyl onto a substituent (point 1.) has very little effect on the ring's net charge. The alterations of COOH to COOMe, or OH to OMe cause little to no difference in the ring’s net charge. Indeed, OH and OMe are almost identical. The addition of C≡N to a substitution (point 3.) marginally increases the ring’s net charge as all C≡N substituents appear at the top of their set. However, this effect is still relatively small with respect to the effect of the connecting atom to the ring, with regards to the ring’s net charge. The effect of inserting a CH₂ between the ring and substituent (point 2.) is relatively small. The profiles lines of CH₂OH and CH₂CH₂OH are almost identical. However, the OH substituent does differ greatly from its CH₂OH and CH₂CH₂OH counterparts. This large difference can be attributed to the connecting atom changing from oxygen to carbon. The OH belongs to the top set (oxygen connection to ring) but when the CH₂ is inserted between OH and the ring the connecting atom changes to carbon. Therefore CH₂OH and CH₂CH₂OH belong to the bottom set (carbon connection to ring). The large difference between the OH profile line and the CH₂OH and CH₂CH₂OH profile lines cannot be attributed to the effect of the CH₂ because it shows a small effect between CH₂OH and CH₂CH₂OH. While effects from the three separate alterations (points 1., 2. and 3.) are displayed in Figure 3.13, it is the element of the atom connecting the substituent to the ring that dominates the substituent's effect on the ring’s net charge.
3.3 Ring Critical Point space

3.3.1 Ring characterisation through Ring Critical Point space

To add to the ring atom properties as descriptors, the RCP properties are examined to determine if they can successfully characterise rings as well. This is achieved through an RCP space. The concept of RCP space is inspired on that of BCP space, which has been successfully used before\textsuperscript{166-168}. If a BCP is accepted as a signature of a chemical bond then it makes sense to evaluate properties at a BCP. The most straightforward property is the electron density itself, which can be related to bond order\textsuperscript{169}, provided the bonded atoms are the same. The same electron density value can appear at more than one BCP if one compares many different bonds \textit{AB}, where \textit{A} and \textit{B} can be any element. However, if a second property is evaluated at a given BCP then it becomes unlikely that two different BCPs share the same two property values. Typically, this second property is the Laplacian (of the electron density), which measures the local concentration of electron density and classifies bonds as closed-shell or shared interactions\textsuperscript{170}. A third is the BCP ellipticity, which is not to be confused with the RCP ellipticity or $\varepsilon_{RCP}$. The former is defined as $\left(\frac{\lambda_2}{\lambda_1}\right) - 1$ and acts as a simple shape descriptor of electronic structure, for instance, expressing the $\pi$-character of a CC bond. If we imagine these three properties discussed above to span a 3D space then each BCP appears as a point in this space. This was the original idea of BCP space, which can of course be extended to an arbitrary number of dimensions. A molecule is then a cluster of points in BCP space and this compact “quantum fingerprint” absorbs subtle changes due to substituent and conformational changes. Similarity between molecules can then be calculated in BCP space.

RCP space is also spanned by the electron density, the Laplacian and its own type of ellipticity $\varepsilon_{RCP}$. The RCP of a ring is plotted in RCP space using these three properties. RCP space assumes that the nature of the ring can be described by the electron density within the ring, that is, evaluated at the RCP in a similar fashion to BCP space, which assumes that a bond can be characterised by a single point in this space. The position of a ring in RCP space should show the nature of the ring and the nature of other rings with respect to it, based on their relative positions.
3.3.2 Calculation of RCP and ring atom properties

To test the "RCP space hypothesis" discussed above, a large number of five-membered rings were geometry-optimised and their QCT properties then calculated. The same set of rings is used from section 3.2.2. In addition to these rings, 192 saturated rings are calculated consisting of ten saturated ring scaffolds. This totals 764 ring structures when the 572 unsaturated rings are included. Each molecule was positioned in RCP space according to their electron density, Laplacian and ellipticity at the RCP.

3.3.3 Rings in RCP space

Figure 3.14a shows the RCP space for the unsaturated rings. Only the unsaturated rings are shown to maintain clarity. Only showing the unsaturated rings is sufficient because trends in RCP space positions found within the two π-bond ring (scaffolds) are the same for the one π-bond and no π-bond rings. Similarly, trends between the two π-bond and one π-bond ring sets are repeated between the one π-bond and no π-bond ring sets. In Figure 3.14a each colour corresponds to structures with a common ring scaffold where the number of points for each scaffold is determined by all of the possible mono-substituted structures. For example, each of the red points represents a mono-substituted 1,2,3-triazole ring where substitution type and position are varied. By colouring the data points in such a way it becomes clear that all points for a defined ring scaffold occupy the same region of RCP space. Therefore, RCP space can distinguish between ring types. The way the rings appear in RCP space highlights trends in the relative positions of the ring scaffolds. These trends are found by examining each separate ring scaffold (separate colours in Figure 3.14a). Figure 3.14b shows these trends when examined from the lowest values in both the electron density and Laplacian, to the highest electron density and Laplacian, i.e. moving in a sequence from left to right in Figure 3.14a. Rings that overlap in RCP space are shown as branches in Figure 3.14b, where following the arrows down corresponds to an increase of electron density and Laplacian at the RCP relative to the other rings. For example, pyrazole rings are higher up in RCP space than pyrrole rings, and imidazole rings are higher up in space than pyrazole but with an overlap between the pyrazole and imidazole ring sets. Figure 3.14b shows that there are three variables in the ring scaffold that affect its position in RCP space: (i) the number of π-bonds, (ii) which element a given heteroatom is, and (iii) the number of heteroatoms.
First the trends within two π-bond rings are examined (which are the same for the one and no π-bond rings), followed by the trends between two and one π-bond rings. The lowest two π-bond rings correspond to rings with a single phosphorous or sulfur atom within the ring. Moving on to the next set of rings in the aforementioned sequence, rings are found containing a single nitrogen or oxygen atom within the ring. These rings are followed by rings containing a second heteroatom within the ring. These are followed by rings with a third heteroatom and finally a fourth heteroatom within the ring. Therefore, as the element of the heteroatom changes from phosphorous or sulfur to nitrogen or oxygen, the ring’s position in RCP space from left to right. Similarly, an increase in the number of heteroatoms within the ring moves the ring’s position from left to right in RCP space. These trends are repeated for the rings with single π-bond. For example, the leftmost ring (in Figure 3.14b) contains one sulfur. When another heteroatom is added to the ring (column two of Figure 3.14b) it moves to the right in RCP space (Figure 3.14a). If the sulfur in the two-heteroatom ring is replaced by a nitrogen or oxygen (column four of Figure 3.14b) the ring moves further to the right in RCP space. However, there is an overlap between the three sets (zero, one or two π-bonds), where rings most left in RCP space of a π-bond set overlap with the rings highest in the RCP space for the previous π-bond set. An example of this is that of the overlap between rings containing a single sulfur with two π-bonds (column three of Figure 3.14b) and rings with two heteroatoms and a single π-bond (column four of Figure 3.14b). This overlap suggests that the increased number of heteroatoms in the rings with a single π-bond creates an electron density at the RCP comparable to that of rings with two π-bonds but containing fewer heteroatoms donating electron density to the ring. Because the rings in RCP space form this progression, the ring scaffold of an unknown ring can be determined through its position in RCP space relative to other rings. Rings can be categorised using only the electron density, Laplacian and ellipticity at the RCP.
Figure 3.14: (a) RCP space for unsaturated (one or two π-bonds) five-membered heteroatomic rings. (b) Unsaturated five-membered heteroatomic rings in ascending order in RCP space. The descending arrows gather rings that overlap in RCP space. Following the arrows down corresponds to an increase of electron density and Laplacian at the RCP relative to the other rings. For example, imidazole rings tend to have a higher electron density and Laplacian at their RCP than pyrazole rings.
In an attempt to quantify differences in the ring types in RCP space, the effects of ring alterations and distances in RCP space was examined. Consistent changes in RCP properties can predict similar changes for unknown ring scaffolds. It is possible to move through RCP space by making calculated alterations to a ring to alter its properties to match the desired properties. For example, a nitrogen atom was inserted into a furan ring replacing the carbon at the second or third position forming isoxazole and oxazole, respectively. The difference in electron density at the RCP between isoxazole and furan, and oxazole and furan, were calculated for all nine ring substituents attached to the ring. The mean change in electron density from furan to isoxazole was -0.015 a.u. The largest change was -0.020 a.u. and the smallest -0.012 a.u., or interval of only 0.008 a.u. On the other hand, the mean change in electron density in going from furan to oxazole was -0.058 a.u. Here, the change in electron density ranged from -0.052 to -0.064 a.u., again an interval of 0.008 a.u. The substituted derivatives of isoxazole and of oxazole display consistent changes in electron density at the RCP when a nitrogen is substituted into the ring. The difference in the position of the added nitrogen is also visible in the RCP position as both isoxazole and oxazole have no overlap in their ranges and the difference between the mean changes for both is much larger than the ranges for either. RCP properties have been found to distinguish between ring scaffolds, and RCP space can be navigated where changes in RCP properties are relevant to a ring’s position in RCP space.

3.4 Practical use of ring properties: a case study

3.4.1 Imidazoleglycerol-phosphate dehydratase

The previous sections in this chapter have shown how the QCT atom and RCP properties can reflect or capture a ring’s characteristics. An important application of this characterisation is its use in ligand design. As all rings can be successfully characterised by the QCT properties, comparison of QCT properties between rings leads to powerful ligand design capabilities. The potential for ligand design through RCP and ring atom properties is explored with Imidazoleglycerol-phosphate dehydratase (IGPD), an enzyme found only in plants. Figure 3.15 shows the structure of the strongest inhibitor of IGPD, which has the lowest inhibitor constant (K). The inhibitor constant is the concentration of the inhibitor required to displace 50% of the agonist from the binding site and is a unit less constant. Therefore, if the K value of an inhibitor is small, then a small
concentration is required to inhibit the target (by displacing the agonist), in which case the inhibitor is considered as stronger. IGPD was chosen because it is a well-studied system where it is known that the ring of the inhibitor is directly involved with binding to the protein\textsuperscript{171,172}. Alterations to the ring dramatically change the $K_i$ value of the molecule. Through crystal structures it is known that the inhibitors that have the strongest binding have hydrogen bond acceptors at the second and fourth positions of the ring (Figure 3.15).

![Figure 3.15: Strongest known inhibitor of IGPD.](image)

3.4.2 Calculation of properties

The nine inhibitors in Figure 3.16 were geometry optimised at the B3LYP/6-311+G(2d,p) level. The optimised structures were verified by eye to ensure that in each molecule the ring retained its orientation relative to the rest of the molecule (Figure 3.15). QCT properties were calculated with the program MORPHY01\textsuperscript{173,174}. It should be noted that the multipole moments such as the dipole moment (and higher rank moments) are direction dependent. Therefore, it is important to express the moments of all the rings in a global axis system, common to all molecules, so that the moments are comparable between molecules. The moments are also normalised so a Euclidean distance can be calculated with all the moments.

![Figure 3.16: Nine different rings for IGPD inhibitor molecule where R1 denotes the rest of the molecule as shown in Figure 3.15.](image)
The global axis system is fully defined by three atoms. First, the origin is defined as the carbon atom adjacent to the ring. Next, the x axis is defined by the bond between the carbon atom at the origin and the ring atom (N1) it is bonded to. Finally, the xy plane is defined by these two atoms and a third atom, which is the oxygen of the alcohol group bonded to the origin carbon. From the xy plane the z axis protrudes perpendicularly forming a right-handed axis system. This global axis system is chosen because the ring plane now approximately lies in the xy plane of the global axis system, by virtue of the geometry optimisation characteristics mentioned just above.

3.4.3 RCP properties as descriptors

The Euclidean distance from the best inhibitor (1) was calculated for rings 2 to 9 using only three RCP properties, which are the electron density, its Laplacian and the RCP ellipticity. Figure 3.16 plots the calculated distances against the $K_i$ values for all nine molecules. All values are also given in Table 3.5.

![Figure 3.16](image)

**Figure 3.16**: QID distances from best IGPD inhibitor calculated with the electron density, its Laplacian and RCP ellipticity.
Table 3.4: Euclidean distances from the strongest IGPD inhibitor and their corresponding $K_i$ values.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>RCP distance</th>
<th>RCP &amp; atom distance</th>
<th>$K_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.08</td>
</tr>
<tr>
<td>2</td>
<td>4.10</td>
<td>8.91</td>
<td>3.95</td>
</tr>
<tr>
<td>3</td>
<td>1.46</td>
<td>2.73</td>
<td>2.78</td>
</tr>
<tr>
<td>4</td>
<td>2.07</td>
<td>4.89</td>
<td>4.93</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>7.26</td>
<td>3.5</td>
</tr>
<tr>
<td>6</td>
<td>0.32</td>
<td>5.16</td>
<td>4.42</td>
</tr>
<tr>
<td>7</td>
<td>2.01</td>
<td>4.46</td>
<td>3.46</td>
</tr>
<tr>
<td>8</td>
<td>0.11</td>
<td>2.65</td>
<td>2.48</td>
</tr>
<tr>
<td>9</td>
<td>1.50</td>
<td>3.22</td>
<td>3.21</td>
</tr>
</tbody>
</table>

The best inhibitor (ring 1) has a $K_i$ of 1.08 and, naturally, a Euclidean distance of 0. According to the Euclidean distance the ring closest to ring 1 is ring 5, the $K_i$ of which is 3.5 but the distance only 0.02. The distance of ring 5 is too small for its $K_i$ value because it does not have hydrogen bond acceptors at positions 2 and 4 of the ring. However, the distance is small because it is the same ring as ring 1, except for a different orientation with respect to the rest of the molecule. The difference in orientation is not seen by the RCP properties because they are expressed in a local axis system that “travels” with the ring in 3D space, as discussed in Section 2.2.2. Therefore, the RCP properties of rings 1 and 5 are almost identical, but not quite because the distance is not exactly 0. Therefore, the influence of the rest of the molecule on ring 5 is different to that on ring 1, due to the different orientations of the rings with respect to the rest of the molecule. Ring 6 is a similar case to ring 5, where the ring structure is similar to that of ring 1 but its orientation differs. Hence, ring 6 has a distance (from ring 1) of only 0.32 because of the similar ring structure but the $K_i$ value of ring 6 differs substantially from that of ring 1, by 3.34. The relatively small distance of 0.32 places ring 6 much closer to the best inhibitor (ring 1) than it should, according to its $K_i$ value, because the ring lacks a hydrogen bond acceptor at its ring position 4. However, the distance is small because it is an isomer of ring 1 (ring 1 is a 1,2,4-triazole while ring 6 is a 1,2,3-triazole) and therefore the RCP properties are similar. Excluding the cases (rings 5 and 6) where it is
the ring’s orientation that affects the $K_i$ value (by not having the hydrogen bond acceptors at the correct position) there is a general increase in distance as the $K_i$ increases.

### 3.4.4 A combined Ring Critical Point and ring atom property approach

QCT ring atom properties were incorporated into the distance calculation to determine if the ring atom properties can improve the relationship between $K_i$ and distance. The following ring atom properties were calculated: the ring monopole, dipole, quadrupole moments and their respective magnitudes, the kinetic energy $K$, and the volume (0.001 a.u. cut-off). The ring properties were calculated by summing the individual atom properties for only the atoms within the ring. The moments for the atoms were taken with respect to the defined global axis system.

![Figure 3.17: Euclidean distances from best IGPD inhibitor calculated with electron density, its Laplacian and RCP ellipticity (red squares) and distances calculated with QCT ring atom and RCP properties (green triangles) for nine IGPD inhibitors.](image)

Figure 3.17 shows the distances for the ring atom and RCP properties combined (green triangles) compared to those of the RCP properties only (red square). The incorporation of the ring atom properties into the distance calculation increases the distance of ring 5 from 0.02 to 7.26. Now the new distance for ring 5 is not too small for its $K_i$ value. Although rings 1 and 5 are very similar and their ring atom properties are as well, expressing the multipole moments in the global axis system successfully
differentiates between the two rings. Ring 5 can be seen as a “flipped” ring 1. Therefore, the moments flipped with the ring and their direction component is different. If the moments were expressed in a local axis system instead of the global axis system, this flip would not be observed and the moments would be almost identical for both rings. Although the ring pattern is identical the flip is seen in the change of the moments with respect to the global axis system.

A similar case occurs with ring 6, where the distance increases from 0.32 to 5.16 upon incorporation of ring atom properties into the distance calculation. It is remarkable that the summed ring atom properties are able to create more suitable distance scores because the summed properties still have no information as to the ideal positions of the hydrogen bond acceptors. Individual atom properties are not expressed and it is through the directional dependence of the moments that it is possible to distinguish between rings 1 and 5.

Whether using only RCP properties or RCP and ring atom properties, ring 2 has the largest distance. Ring 2 differs most to ring 1 in ring pattern, with the exception of ring 8 but the amine substituent affects the properties of the ring, and so has the largest distance and the distance may be over-penalising. Rings 2, 4, and 8 are the only ring patterns with fewer than 3 nitrogens in the ring, and these rings are the three highest scoring distances using only RCP properties. Ring 4 has a better score than ring 2 because the two nitrogens are adjacent and therefore the ring pattern matches more closely that of ring 1. Ring 8 has the influence of the amine on the ring, so the RCP properties decide that ring 2 is has the largest distance of any ring. When the ring atom properties are included in the distance calculation (rather than just the RCP properties) then ring 2 retains it extreme position, and its distance increases to 8.91.

In summary, distances calculated with only the RCP properties correlate quite reasonably with Ki values where rings with similar RCP properties exhibit similar activities. The few outliers that occur are expected and can be explained. When ring atom properties are added to the distance calculation, then the outlier better fit the correlation.
3.5 Conclusion

A series of PLS models established a correlation between ring atom properties and RCP properties (however the reverse did not work, i.e. RCP properties to ring atom properties). For the correlation the ellipticity proved to be most difficult to predict and ruined the correlations. This is because the ellipticity is the most sensitive to the changes in the ring atoms and therefore more difficult to predict accurately.

Four features were found to determine a ring’s net charge. These features were the element of the heteroatom within the ring, the number of heteroatoms within the ring, the type of substituent, and the site of the substituent. The effect of each of these features causes a signature systematic change to the ring’s net charge. Each feature causes a change to the ring’s net charge that is independent to any changes caused by other ring features. The systematic change to ring atom properties is tested by predicting changes in ring net charge due to substitutions while using a mean change in net charge for substitution. The initial predictions were reasonable but improved by separating the mean change in net charge per substitutions into types, i.e. the atom it was substituted on to. It was found that the atom connecting a substituent was more important in determining the substituent’s effect on the ring’s net charge than the remaining atoms in the substituent.

An RCP space was established with three properties of the RCP, the electron density, its Laplacian and the RCP ellipticity. Rings were positioned in this space based on these three properties. Using only these three properties the RCP space was successful in distinguishing between ring types. The ring types were ordered in RCP space by their characteristics, which determined their position in the space. Therefore, is possible to predict the position of a ring within the RCP space without requiring the calculation of its properties.

Finally, the RCP space and ring atom properties were tested against the IGPD inhibitor molecule. While the RCP space produced a mediocre correlation, it struggled with certain cases where the rings orientation altered its activity. When ring atom properties were included then these outliers fell back into the correlation and the general trend was improved.
Chapter 4

Evaluating the QID and addition of terminal groups

4.1 QID alignment

4.1.1 The current QID alignment method

This chapter aims to improve the existing QID tool and expand its application to terminal fragments. To improve the QID tool the existing scoring function, alignment procedure and the tool’s search capabilities are evaluated to determine possible areas for improvement. This is followed by a brief study in the fragments that the QID suggests for a simple linker, which are compared to the fragments suggested by the program BROOD. The current method of data generation for linkers is not suitable for terminal fragments and therefore alternatives are explored.

Fragments must be aligned for meaningful comparison of properties such as volume and dipole moments so the fragments must be held in a standard orientation. Ideally, alignment of two fragments would arrange the fragments in space to provide give the maximum similarity, or overlap, between them. There are many variations of alignment methods and a standard form of alignment is the selection of key features on a molecule and overlay them which can be done with any set of features chosen by desirability\textsuperscript{175-178}. One example is pharmacophoric alignment\textsuperscript{179-181} that identifies key pharmacophoric groups and minimises the difference between all the pharmacophoric groups using a root mean square difference. A more recently devised approach uses hard spheres at each atom and rotates the fragments such that the difference in the volumes is minimised\textsuperscript{182}. A new development to this method replaces the hard spheres with volumes constructed by Gaussian shapes on each atom\textsuperscript{183,184}. Again, the volumes are rotated in space until the difference in volume minimised.

The current method for alignment in the QID tool is much less complex in its nature then the hard spheres and Gaussian shapes. Because the database consists of fragments systematically taken from the WDI, therefore, the two connection points are common to all fragments. The connection points of the fragments can be employed to easily position a fragment into a standard orientation. The position of the first connection point is common to all fragments and it is the position and direction of the
second connection point is compared. This method enables quick and easy alignment relative to the more computationally intensive Gaussian shapes.

To place a fragment into the QID standard orientation the first atom in the linker (the first connection point) is translated to the origin. The fragment is rotated to align the second atom in the linker along the z axis. The bond joining the connection point and its adjacent atom now lies along the z axis. The second rotation of the fragment is performed, this time around the z axis, until the second connection point is in the zy plane. The direction of the bond between the second connection point and its adjacent atom allows for comparison of fragment orientations. The position of the second connection point allows for comparison of fragment size and the relative position of one end of the fragment to the other. The standard orientation of a fragment is shown in Figures 4.1a and b.

![Figure 4.1: Standard QID orientation of an ester fragment from perspective (a) and (b) aligned by the two connection points.](image)

4.1.2 Alignment of terminal or collinear fragments

The current alignment method is not suitable for use on terminal fragments. Terminal groups present a series of problems which cannot be addressed by the current alignment method. Absence of the second connection point makes this method unusable for the terminals. The second connection point was selected to align the linkers because a position change of the second connection point accurately represents how a fragment would change the geometry of a molecule. The molecule continues
form the second connection point so a twist in the geometry of the linker would greatly affect the geometry of the whole molecule. Taking the two connection points as references for alignment and geometry makes sense for linkers as it represents how the linker lies in within a molecule. This is not the case with terminals. Taking the last atom of a terminal group as connection point 2 and repeating the alignment would not give meaningful comparisons. Terminal fragments are only bound to the rest of the molecule by one point and have much more freedom. Tethering the second point in the alignment process reduces this possible freedom as the coordinates for that point will be similar for fragments of similar size. An example of this would be the alignment of cis and trans isomers. Alignment of a fragment with the standard QID alignment for a terminal atom that is ether cis or trans would position this atom in the same place for both isomers. A more suitable alignment would position these terminal atoms at their respective positions (cis or trans), as would be the case if the double bonds were aligned. A stronger alignment would start closer to the first connection point as the substituents that lie cis or trans should be in different position to reflect their geometry.

Another trouble with terminal groups is that they also introduce a large freedom of rotation around the connection point bond. The rotation must be accounted for in the database. A terminal fragment’s rotation can dramatically change from being perpendicular to the rest of the molecule to parallel to the rest of the molecule. An example the rotational freedom is ethane changing from staggered to eclipsed geometries. It is important to account for this rotation because the properties of the terminal group can change upon rotation as the interactions it forms with the rest of the molecule change.

The current method for alignment is also unsuitable for a special case of linker, when both connection points of the linker fragment are collinear. It is possible to rotate the fragment into any orientation around the z axis because the second connection point already lies in the yz plane. Rotation around the z axis will not influence the direction and position of the second connection point and so all orientations are observed as the same. Figure 4.2 shows a phenol ring rotated into two separate orientations by the current QID alignment method where there is no control, or means of determining, which orientation the fragment will be stored as. This also means that the QID would consider the geometries of the two phenols exactly the same, which is not the case. The alignment method does work for collinear fragments if the fragment
has infinite fold symmetric around the z-axis because all possible orientations are indistinguishable, alkynes are such a case.

![Figure 4.2: two separate orientations of a phenol fragment with collinear connection points.](image)

**Figure 4.2:** two separate orientations of a phenol fragment with collinear connection points.

### 4.1.3 Method for evaluating the alignment procedure

First the current alignment method needed to be tested to ensure the extent of fragment alignment by QID was sufficient for comparisons between the descriptors to be meaningful. Second, an alternative alignment method is required for collinear fragments and also terminal fragments when they are added in the future. With regards to the QID tool’s structure, when a collinear fragment or terminal group is encountered the alternate method will be called to align the fragment instead.

The current alignment method was validated using a standalone code that replicated the alignment procedure to help identify flaws with the current method. Figure 4.3 shows a sample of 9 molecules from the test set of 30 fragments, all selected from the database. The Cartesian coordinates were recorded for the atoms before alignment by the QID method. The 30 fragments were aligned using the QID alignment procedure. Coordinates of the aligned fragments were compared to a QID aligned ethane fragment. The coordinates of the two atoms at each connection point were chosen to best ascertain the alignments of the fragments. Ethane was chosen because it is the simplest fragment and has no side chains. Comparing the atoms in the connection points will show how fragments have deviated during the alignments process.
Figure 4.3: 9 of the 30 fragments selected from the database to test the alignment method. Au and Ag represent connection points where the fragment is connected to the rest of the molecule.

4.1.4 Evaluation of alignment procedure

Table 4.1 shows the differences (in Å) for the two atoms of each linker where atoms 1 and 4 are the connection points. Atom 1 is the connection point Au, then moving along the fragment backbone are atoms 2 and 3. Using fragment 15 as an example, Au is atom 1, C is atom 2 and S is atom 3 and Ag is atom 4.
<table>
<thead>
<tr>
<th>Fragment ID</th>
<th>Atom 2</th>
<th>Atom 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.08</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>0.07</td>
<td>1.48</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.07</td>
<td>0.30</td>
</tr>
<tr>
<td>6</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.79</td>
</tr>
<tr>
<td>9</td>
<td>0.07</td>
<td>0.29</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td>11</td>
<td>0.07</td>
<td>0.35</td>
</tr>
<tr>
<td>12</td>
<td>0.01</td>
<td>0.74</td>
</tr>
<tr>
<td>13</td>
<td>0.00</td>
<td>0.69</td>
</tr>
<tr>
<td>14</td>
<td>0.01</td>
<td>0.69</td>
</tr>
<tr>
<td>15</td>
<td>0.00</td>
<td>1.04</td>
</tr>
<tr>
<td>16</td>
<td>0.02</td>
<td>0.32</td>
</tr>
<tr>
<td>17</td>
<td>0.00</td>
<td>0.41</td>
</tr>
<tr>
<td>18</td>
<td>0.00</td>
<td>0.77</td>
</tr>
<tr>
<td>19</td>
<td>0.01</td>
<td>0.74</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>21</td>
<td>0.02</td>
<td>0.83</td>
</tr>
<tr>
<td>22</td>
<td>0.02</td>
<td>0.32</td>
</tr>
<tr>
<td>23</td>
<td>0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>24</td>
<td>0.00</td>
<td>0.75</td>
</tr>
<tr>
<td>25</td>
<td>0.04</td>
<td>1.09</td>
</tr>
<tr>
<td>26</td>
<td>0.00</td>
<td>0.68</td>
</tr>
<tr>
<td>27</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>28</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>29</td>
<td>0.01</td>
<td>0.66</td>
</tr>
<tr>
<td>30</td>
<td>0.03</td>
<td>0.66</td>
</tr>
</tbody>
</table>

*Table 4.1:* Differences (in Å) between atoms for QID aligned fragments and ethane.
For all fragments atom 2 does not differ by greater than 0.08 Å, this shows that the atom 1 to atom 2 bond is being correctly aligned along the axis. Atom 2 is not exact for every fragment because bond lengths vary depending on the atom type and adjacent functional groups. Atom 3 shows greater deviations than atom 2. Most points vary by approximately 0.5 Å. This greater deviation compared to atom 2 can be attributed to a similar effect seen for atom 2. Because atom 3 is on the opposite side of the fragment there are now three bonds separating atom 3 from the origin. With more bonds separating atom 3 from atom 1 the differences in bond lengths accumulate and the variation in bond angles will also affect the final position of atom 3. Fragment 2 differs the greatest out of all atom 3 points. The difference of 1.48 Å appears because the fragment is ethyne, and is due to its difference in geometry to ethane. Fragments 15 and 25 also have large deviations for atom 3 both differ by 1 Å. They are both sulphur containing fragments, and therefore local geometries of the fragments are different and this is reflected in the results. The alignment between fragments is sufficient to retain this method for the QID, and no new special cases were found.

4.1.5 Alignment of collinear fragments

To align fragments outside the bounds of the current method, such as collinear and terminal groups, principal axis alignment was tested. Principal axis alignments were performed on a set of fragments absent from the database. Fragments were chosen that shared the collinear characteristic that the current method is unable to handle. Once the fragment was aligned by principal axis alignment the molecule was translated so atom 1 was centred on the origin. Atom 2 was aligned along the x-axis in the same fashion as the current alignment method. Therefore, all fragments shared the same origin. Correct alignment of these systems is vital as most collinear molecules will be cyclic fragments which are prolific throughout both drugs and agrochemicals. It is essential that the plain of a ring so that their properties can be meaningfully compared. This also ensures that and any substituent's on the rings are also in alignment correctly.

The test set chosen had to incorporate fragments that the current method could not align; the set is shown in Figure 4.4. All fragments were aligned with the principal axis method. The Cartesian coordinates of selected atoms were compared to a benzene molecule that was also aligned by a principal axis system. The atom numbering remains the same as for the test set against ethane (above) with the addition of atom 5, which is
within the ring. Atom 5 was selected to show the difference in alignment of the ring planes. Atom 5 is adjacent to the atom bonded to the connection point.

![Molecules](image)

**Figure 4.4:** Collinear molecules subjected to a principal axis alignment

### 4.1.6 Evaluation of alignment alternative

Table 4.2 shows the results of the alignments. Note that atom 4 is also included into the analysis whereas previously, with the ethane test set, atom 4 was excluded. This is because in the previous set was part of the alignment procedure and was directly rotated into the same position. Fragments A, B and C are all strongly aligned to the benzene ring with fragment D only slightly worse. This is found because fragment D is an eight atom ring, therefore, atoms 3 and 4 are further away from the origin. Atom 5 differs because the bond angle between atoms 1, 2 and 5 is smaller and therefore atom 5 is positioned further away. However, the difference is small enough to assume that the ring planes are well aligned.
<table>
<thead>
<tr>
<th>Fragment ID</th>
<th>Atom 3</th>
<th>Atom 4</th>
<th>Atom 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>B</td>
<td>0.22</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>C</td>
<td>0.22</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>D</td>
<td>0.49</td>
<td>0.49</td>
<td>0.56</td>
</tr>
</tbody>
</table>

**Table 4.2:** Differences (in Å) between atoms for principal axis aligned fragments and benzene.

The current method is adequate for alignment of basic fragments and preliminary tests of principal axis alignment show it to be capable of aligning collinear fragments. More test cases are needed to ensure the alignment works for more elaborate collinear systems and terminal groups. Testing across methods is also required to show the collinear fragments are in agreeable alignments to standard fragments.

### 4.1.7 Terminal group alignment

A separate alignment method is proposed for terminal fragments because they lack the second connection point of the linkers. To align the terminal fragments for meaningful comparisons a global axis system is proposed. With a global axis system the fragments are not rotated but properties dependent on coordinates and direction are expressed in a global axis system. Therefore, meaningful comparisons can be made for the terminal groups if they are expressed in the same axis system. The global axis system is an adaption of the Atomic Local Frame (ALF). The ALF assigns a local axis system onto an atom of interest. For the ALF, the origin is assigned to the atom of interest (atom 2 in Figure 4.5). Atom 3 is determined by Cahn-Ingold-Prelog (CIP) rules and the x axis projects through atoms 2 and 3. Atom 1 is the next CIP and the xy plane is formed with atoms 1, 2, and 3. The z axis is projected perpendicular to the xy plane and forms a right handed axis system. The ALF introduces an axis system that travels with an atom. Therefore, directional properties of atoms can be compared when expressed in the ALF.

The global axis system proposed for terminal fragments is analogous to an ALF. The connection point is assigned as the origin (atom 2). The x axis extends through the first atom of the terminal group (atom 3). The second carbon in the ethyl capping group (atom 1) is used to assign an xy plane and the z axis is projected from this plane. Figure 4.6 highlights the global axis system for a terminal phenol fragment. Previously the
capping group was not included into the alignment. Inclusion of the terminal group into the alignment method adds information regarding the terminal group's orientation relative to the rest of the molecule. Changes in geometry of terminal fragments in this axis system can be observed by changes relative to the origin. However, due to the rotational freedom of terminal groups, an increased number of geometries are required for each fragment to capture its possible rotations. Linkers in the database are stored in a number of different geometries up to a maximum of 20. Data for the terminal groups must account for the rotation, so a number of geometries will be held for a set of rotations around the connection bond. Every 15° rotation will hold 5 different geometries. The geometry set will allow for comparison of terminals in different rotations relative to the rest of the molecule. It is even possible to compare the same terminal group in different rotations to itself where properties may change as it rotates. This method allows for meaningful comparisons of properties between terminal fragments.

**Figure 4.5**: Atom number labelled propanoic acid.

**Figure 4.6**: Global axis system highlighted on a terminal phenol fragment.
4.2 Bioisosterism of terminal groups

4.2.1 Geometry from electrostatics

A purely quantum chemical descriptor space is possible if the geometry descriptors are removed from the descriptor space. However, geometry is important in ligand design and therefore must be accounted for. The nuclear coordinates are required to solve the Schrödinger equation to obtain the electron density. If the electron density is known then it may be possible to reverse the procedure and obtain the geometry from the electron density. It should be noted that the shape of the molecule cannot be obtained from the electrostatics alone. However, the shape can be accounted for if the atomic volumes are added to the electrostatic descriptors. Successfully obtaining the geometry from the electrostatics would suggest that the geometry is imprinted into the electrostatics. Such a case would lead to bioisosterism with only electrostatic descriptors, a pure quantum bioisostere database. Restructuring the database for use with only electrostatic properties would cut down the time for data generation and entry into the database. Only the electrostatic properties would be extracted from the calculation outputs and volume of data entered into the database reduced. Performing a search of the database for hits would also be quicker as the number of descriptors searched is smaller. A smaller descriptor space also benefits the search outputs as mentioned in section 2.4.6. This effect of geometry on electrostatics has been seen by Mills and Popelier who saw that the rotation of an ethanol dihedral angle had a marked effect on the moments of the atoms in the ethanol. This indicated that the geometry could be held in the electrostatics because the electrostatics of ethanol changed with the geometry. The reverse of this observation may also be possible, where knowledge of the electrostatics alone could derive the geometry of the molecule. Alternatively, if the geometry is fully imprinted into the electrostatics then the geometric parameters could be removed entirely from the descriptor space because the geometry is accounted for in the electrostatics. However, applying such a method that relies purely on electrostatics needs to satisfy two criteria. First the changes in moments must be noticeable against change in substituent. Second the electrostatics must correspond to a unique geometry. In other words, if there were multiple geometries that share the same electrostatics then they would be indistinguishable.
4.2.2 Calculating the multipole moments of a geometry change.

The two fragments chosen were a capped carboxylic acid (Figure 4.5) and amide group because they vary only by substitution of OH with NH$_2$. The substituent change is enough to significantly change the moments but make rational comparison between geometries easier. The fragments were capped with an ethyl as per QID convention. Each molecule was optimised at the B3LYP/6-311G+(2d,p) level. The dihedral angle of atoms 1, 2, 3, and 4 from Figure 4.5 was rotated in 10 degree increments and each geometry was stored for both molecules. Wavefunctions were calculated for each rotation of both molecules. MORPHY01 calculations were done for each of the wavefunctions. The electrostatics for the molecules were expressed in three ways; in terms of the global axis system, an ALF, and summed for the entire molecule moments.

4.2.3 Finding unique electrostatics for a geometry

Multipole moments for each dihedral angle were plotted against the corresponding dihedral angle. Figure 4.7a shows the three dipole moment components for the carbon in the COOH fragment. Taking only the Q[11c] moment, the value of -0.05 a.u. is shared by two dihedral angles. Both 120° and 320° have a Q[11c] of -0.05 a.u.. However, if the second and third dipole moments are incorporate then single valued results are obtained. Taking the same value of -0.05 a.u. for Q[11c] but adding the condition that Q[11s] must be 0.1, we now find that only the dihedral angle at 320° can have these two moments. With only two components a unique geometry has been found. This applies to any dihedral angle on the graph. Therefore, if the eight multiple moments (Figures 4.7a and b) and the monopole is used then it is reasonable to assume that a given geometry has a unique set of multipole moments.
Figure 4.7: Components of the (a) dipole (b) quadrupole moments plotted against the dihedral angle for propanoic acid.
Figure 4.7a and 4.8 show how the carbon atom moments change when the OH is substituted for an NH$_2$. Tracing multiple components still arrives at a unique geometry. However, the shape of the moments has become distorted. The moments show a similar change but have become much more exaggerated with the NH$_2$ present. The range of the multipole moments has doubled. Figures 4.7a, 4.7b, 4.8a and 4.8b show that the moments track the rotation of the terminal group. As the dihedral angle is altered the moments change with it, and can give a singular geometry that can be defined using a combination of only two of the nine calculated moments.

![Graph showing the change in dipole moment with dihedral angle](image-url)
4.2.4 Determining a geometry from electrostatics alone

To test the moments further the geometry parameter was removed from the plot. Each of the three dipole moments was plotted against each other in Figure 4.9, where the electrostatics are the only input. It should be noted that there are now no geometry variables in Figure 4.9. Each loop corresponds to the moments of a single atom. The green loops represent the carbon, the blue loops the carbonyl oxygen, and red loops the OH an NH$\textsubscript{2}$. The lighter colour variations are for atoms taken from the CONH$_2$ fragment and the darker loops for the COOH fragment. The moments for the atoms were expressed in the global frame. The moments travel through the space with respect to the axis system. Therefore, as the geometry is changed the moments change with it. This is further clarified when examining single points on each loop. The southernmost point on each loop corresponds to the same dihedral angle, as does the northernmost point, and all points in between. Tracking only the moments through space it is possible to determine the dihedral angle. Therefore, the geometry of the fragment can be predicted from only the electrostatic moments.
Non-uniform loops represent an uncharacteristic change in the moments where the loops break from their uniform nature. As a functional group is rotated the atoms experience different interactions with the rest of the molecule. These interactions manifest themselves as the bumps in the loops. The bumps can show that the rotation of the molecule can lead to noticeable changes to the moments themselves as a uniform loop would simply track the moment’s vector as it is rotated with the fragment. These bumps provide a unique bioisosteric opportunity. If the two closest points between the loops are considered to possess the most similar in properties then naturally a straight mapping of the identical corresponding dihedral angles would be expected. However, this may not be the case where the bumps exist. It may turn out that rotating one fragment beyond another produces moments that are a closer match to the other fragment, where the stronger match between the two fragments is not necessarily the same geometry for both.

Figure 4.9 shows that there is little movement in the loops with respect to the Q[11c] axis, relative to the movement of the loops in the other two axes. Q[11c] corresponds to the dipole moment along the x axis. As the fragment is rotated around the x axis the y and z components are greatly affected, however, the x component remains relatively similar and only changes due to intramolecular interactions.
Figure 4.9: Dipole moments for carbon (green) carbonyl oxygen (blue) and the alcohol oxygen / amine nitrogen (red) for propanoic acid (dark) and propanamide (light) expressed in a global frame.

4.2.5 Geometry’s effect on atom and molecular electrostatics

Tracing the moments with respect to the global axis frame show how the moments travel with the fragment as it is rotated. However, moments in the global axis frame cannot easily show changes to the moments of the atoms themselves. Expressing the moments in the ALF shows the moments of the atom irrespective of its position in space. Any change in the ALF moments would suggest that the intramolecular interactions are strong enough to have an imprint on the moments of said atom. Changes in the ALF moments could also help determine the geometry of the fragment.
Figure 4.10 shows the three dipole moments, taken in the ALF, for the carbon in the COOH fragment. There is a clear change in the moments as the geometry is changed. As the fragment interacts with the rest of the molecule the interactions influence the moments of the atoms. Notably these changes in the moments are on a small scale but still noticeable. Because the moments in the ALF travel with the atom the changes in the geometry can be detected through the moments in the ALF of single atoms. This leads to the conclusion that the electrostatics do hold information about the geometry of the molecule.

Figure 4.10: Dipole moments for carbon in propanoic acid expressed in the ALF shown from perspectives (a) and (b).
Figure 4.11 shows the summed moments for the entire molecule. The molecular moments show how the moments change for the molecule as it is rotated. The changes in the molecular moments show that the moments themselves do change with the geometry, and make evident that the fragment’s orientation to the rest of the molecule can influence its moments.

![Figure 4.11: Molecular dipole moments for propanoic acid.](image)

Use of moments without geometric parameters was tested within context of the QID. The first seven fragments for an amide fragment search were taken as the test set. The moments for all seven fragments were rotated into the global axis frame. Euclidean distances were calculated using the moments i.e. the geometry component was removed. Figure 4.12 shows the distance of the moment scores relative to the fragments rank in the original search, which includes the geometry parameters.
Figure 4.12: QID scoring of amide replacement fragments derived from only electrostatic descriptors in a global frame compared to the fragment’s QID score with a full range of descriptors.

With only the moments and excluding the geometric terms, the QID distance scores show a general increase with the rank of the geometry included search. Fragments are ranked based on score, therefore when only the moments are used two of the seven fragments are not ranked in the same order as a standard search but five of the fragments are. This does not conclusively show that the electrostatics alone can be successfully employed as the only form of descriptor. However, the close matching ranks suggest a model based solely with electrostatics is not unreasonable.
4.3 Testing the QID scoring function

4.3.1 QID search

To test the current scoring method, a series of bioisosteric searches were performed for an ethanol fragment. The ethanol fragment was chosen as it is a linker within the strongest inhibitor of IGPD (Figure 3.15). It is known that the alcohol group is vital for the activity of the molecule as it directly interacts with the binding site. For each search the propanol fragment was used as a marker as a bioisosteric replacement for the ethanol fragment. However, the largest linkers in the QID are 2 atoms long. Therefore, the propanol fragment is actually a 1-methyl-2-hydroxy-ethane fragment within the target molecule. Information regarding the ranking of the fragment and the search was used to evaluate the reliability of the scoring function.

The searches were varied by number of properties selected, the selection of properties, and the importance for each property. The values recorded were rank and total number of hits. Rank is the position of the propanol fragment in the output set after ordering by score. The rank shows the relative bioisosteric match of the hit fragment to the target fragment with respect to other fragments. The total number of hits is the number of fragments that fall within the tolerances of the search. The searches ranged from selecting conformation as the only search criteria to incorporation of all criteria in the search each with varying importance. The more criteria that were included in the search more complex it became. The searches were designed to highlight the dependency of fragment scores for the hit fragments to the search criteria.

4.3.2 Search Results

Figure 4.13 shows how the number of hits is dependent on the complexity of the search. As more search terms are introduced the number of hits decreases. This clearly shows that as the complexity of the search is increased such as introduction of more properties or tolerances increased in strictness, the resulting set reduces in size as more and more fragments are filtered out. The plateau reached for the last set of searches where the number of fragments hits for a search is at ten fragments. This suggests that the set of ten fragments are all strongly similar to the ethanol fragment as increasing the search criteria still holds the same set number of fragments and so must lie well within the tolerances of the searches.
Figure 4.13: Relation of total number of hits for ethanol fragment in QID with respect to the complexity of the search criteria where the complexity of the search criteria increases as the search input tolerances, weightings and number of properties are increased.

Figure 4.14 shows that in general the propanol fragment is ranked in the top twenty hits. This suggests that the scoring function is able to distinguish it as a strong match. The rank varies much greater in the broad searches, but as the search becomes stricter, the fluctuations in rank are reduced.

Figure 4.14: Relation of propanol rank in QID with respect to increasing search criteria complexity for an ethanol fragment search where the complexity of the search criteria increases as the search input tolerances, weightings and number of properties are increased.
For three of the searches, the first twenty fragments were examined with respect to the IGPD inhibitor where the alcohol group is required for a strong active molecule. The three searches were:

1. Conformation, molecular electrostatics and polarity
2. Charge and conformation
3. Charge, conformation and molecular electrostatics

The presence of the alcohol group in the correct position of the fragment was noted for each fragment within the first twenty results of the three search sets. Search 1 returned 17 fragments with the alcohol in the correct position, search 2 returned 14 correct fragments and search 3 returned 15. The remainder of the fragments within each set had either an amine or carbonyl oxygen where the alcohol should have been. The majority of the fragments for all three searches were found with the correct alcohol positioning, and any of them could be potential bioisosteric replacements.

4.3.3 QID vs BROOD

The fragments in the result set of the QID charge and conformation search (search 2 above) were compared to a set of fragments from a query in BROOD for the same fragment. The BROOD query was a shape and electrostatics search (see section 2.4.9 for information on BROOD searches). Fragments from the first cluster of the BROOD results are compared to the QID results. If the two methods return the same fragments then there is agreement that said fragment is a suitable bioisosteric replacement. However, if one method returns fragments that are absent from the other this suggests either a false positive or success at finding a replacement fragment the other method could not. As the rank of the BROOD cluster increases the results become less similar to the query. Analysing the quality of results for QID against all clusters would give an inaccurate representation as the BROOD database is vastly larger in size compared to the QID. The best matches for the query fragment should be found in the first cluster and is deemed sufficient for this test.

The total number of fragments in the first cluster of the BROOD search was 273, QID returned 189 fragments. Therefore, the resulting sets are similar in size for a reasonable comparison. Of the 189 QID fragments 25 were matched in the BROOD results and 164 unmatched. 19 of the brood results were found in the QID but were not returned by the QID in its result set. The large number of fragments found by the QID
but not BROOD may be within lower ranked clusters of the BROOD results; however, the QID suggests them as good matches for the query fragment where BROOD does not. It is difficult to definitively say which database is performing better as BROOD has a database of fragments orders of magnitude larger than QID. However if only the strongest set of BROOD results are compared to the QID set, the first cluster of the BROOD set was still larger than the QID result set but QID found 164 fragments not found by BROOD, whereas BROOD found 19 fragments that were not found by the QID. This conclusion is based on the assumption that all fragments of the QID are present in BROOD. It must be noted that only the first cluster of BROOD was used. Therefore, it is not clear as to which database is performing best but is a good indication that the QID is able to find fragments that other bioisostere databases find. A reanalysis is required with a larger dataset of molecules, which will be possible when the QID has more fragments, or more accurate knowledge of all fragments within BROOD to determine which fragments found by the QID were not in BROOD. However, the most difficult part of validating bioisosteres is with regard to the suggested replacements themselves. Unfortunately, it is difficult to ascertain which replacements would be successful bioisosteres as this would require the replacement fragments inserted into the target molecule and taken to biological testing, which would be a resource heavy process for a set of suggested replacements.

4.4 Conclusion

Evaluation of the current QID method for alignment highlighted the case of collinear molecules that cannot be successfully modelled within the QID. The current method of alignment’s shortcoming was explained and the alternative of alignment through a principal axis system rotation shown to successfully align the collinear fragments. A new alignment method was proposed for terminal functional groups because they lack the second connection point common to all the linkers. This method relies on the direction dependent properties to be expressed in terms of a global axis system. This idea of a global axis system was extended to remove all geometric variables. Through the electrostatics alone, which were expressed in the global axis system, it was found that geometry of the molecule was imprinted into the electrostatics. In other words, changes in geometry corresponded to changes in electrostatics where each unique geometry had a unique set of multipole moments.
Finally the scoring function of the QID was tested. First a simple fragment search showed that the QID was able to consistently find a suitable bioisosteric match for the query fragment when search criteria were altered. Secondly the QID was tested against the program BROOD. This brief test showed that the QID returned fragments that were comparable to BROOD, bearing in mind that the BROOD results were reduced in size to compensate for its much larger database size relative to the QID.
Chapter 5

Predicting pKₐ values from an ab initio bond length

5.1 Predicting pKₐ values

5.1.1 High Correlation Subsets

This chapter introduces a method of pKₐ prediction designed to work alongside the QID tool for ligand design purposes. A pKₐ predictor is valuable in ligand design because it can be used as an indicator to determine if a molecule will become a suitable ligand. However, the pKₐ is an experimentally determined value and therefore must be predicted from other descriptors. Here a single descriptor (a single ab initio bond length) is successful in correlating with the experimental pKₐ values of the molecule within the data set. This method relies on so-called high correlation subsets (HCSs). The formation of HCSs is still not fully understood and therefore two sets of molecules (enols and guanidines) were chosen to investigate the features that govern HCSs. This chapter is split into two, first HCS for the enol functional group is constructed and analysed, then the same for the guanidines.

The roots of this work is a method called Quantum Topological Molecular Similarity (QTMS)\textsuperscript{31,167}. The descriptors of QTMS are quantum mechanical functions, such as the electron density, which is evaluated at special points in three-dimensional space called bond critical points\textsuperscript{166}. Over the last decade this method has delivered successful quantitative structure-activity/property relationships over a wide application radius (e.g. environmental\textsuperscript{186}, cytotoxicity\textsuperscript{187}, cancer\textsuperscript{188}, ester hydrolysis\textsuperscript{189} and pKₐ prediction\textsuperscript{190}). However, taking advantage of the existence of local linear relationships between bond critical point properties and bond lengths\textsuperscript{169} one can simplify the descriptor space to just bond length. The current work makes full use of this insight.

Harding and Popelier found that a correlation between a given ab initio calculated bond length, where the chosen bond is found in all molecules for a defined set, and the experimental pKₐ of these molecules was very strong. For example, the C-O bond length of a set of phenols correlated to their pKa values gave an $r^2$ value of 0.91. The typical $r^2$ correlation was around 0.95 for different sets of molecules. A set of phenols was optimised at the HF/6-31G(d) level of theory. The correlation of the entire
set was poor, however, smaller subsets were found to have strong correlations. The phenol set was separated into ortho, meta and para subsets each with RMSE errors below 0.5. The bond length selected for the correlations was determined by correlating each bond length with the experimental pKₐ value and a single bond was found to give very high correlations. The method was applied to anilines¹⁹¹, benzoic acids¹⁹¹, thiophenols, pyridines, barbituric and carboxylic acids with similar results. In all cases the high correlation subsets have a commonality, and chemical space becomes partitioned into these sets. Once a high correlation subset is determined the equation of the line of best fit is used to predict the pKₐ value of a compound where the pKₐ value is unknown using the corresponding ab initio calculated bond length as the input for the equation. The largest error in prediction was below 0.5 log units. Because the method derives an equation for a single functional group at a time, it is possible to obtain pKₐ values for multiple sites of a molecule while most pKₐ predictors tend to output global pKₐ values. Harding and Popelier did not apply the method to any biological molecules; the data sets were constructed from small organic molecules.

The fact that the HCS method uses only a single descriptor, and an ab initio one, is an attractive feature. Models obtained from reliable descriptors are more predictive in nature, especially if a single descriptor can achieve a satisfactory prediction. Such a descriptor is then also robust in predicting pKₐ values for molecules that lie outside the training set. A powerful and "physically informed" descriptor is able to point out the conformation from which a pKₐ value can be predicted accurately. This method avoids the construction and use of models that rely on large data sets, which can be branded as lookup tables. Such methods typically need a sizeable number of descriptors. A method that relies on a descriptor rather than on data set size, demands that this descriptor is robust. This work is continued with the aim of successfully predicting the pKₐ values for biological molecules from relevant high correlation subsets and to determine the causes of the high correlation subsets and their meaning.

5.1.2 Identifying problematic functional groups

A study by Balogh et al. highlighted three functional groups that were particularly badly predicted using commercial pKₐ packages¹⁹². Two of these three were enols and weak acid nitrogens, in particular guanidines. The high correlation subset method was tested against the other packages by predicting the pKₐs for the enols taken from the Balogh study. Sethoxydim and tralkoxydim were both poorly predicted by the
packages, with experimental pK<sub>a</sub>s of 4.58 and 4.98 for the enol and alcohol respectively, the predictions ranged from 4.40 – 9.23 for sethoxydim and from 3.90 – 9.54 for tralkoxydim. VCC was the best package, predicting sethoxydim as 4.40 and tralkoxydim as 3.90 out by more than 1 pK<sub>a</sub> unit. ACD predicted pK<sub>a</sub>s of 3.83 for sethoxydim and 3.66 for tralkoxydim around 1 pK<sub>a</sub> unit below the experimental for both. Epik predicted pK<sub>a</sub>s of 6.09 for sethoxydim and 5.95 for tralkoxydim, around 1 pK<sub>a</sub> unit above the experimental for both. Marvin predicted pK<sub>a</sub>s of 7.59 and 7.53, around 3 pK<sub>a</sub> units above the experimental for both compounds. Finally, Pallas predicted pK<sub>a</sub>s of 9.23 and 9.54, approximately 5 pK<sub>a</sub> units above the experimental values. For the enols VCC was the most consistent and closest to the experimental pK<sub>a</sub>s, however, 1 pK<sub>a</sub> unit error is still relatively large and the huge variance in predictions between packages suggest that there is still room for improvement for pK<sub>a</sub> prediction of enols. Sethoxydim and tralkoxydim were the only two enols in the dataset, but with the enol functional groups identified as a difficult to predict the pK<sub>a</sub> of, a method that can predict the pK<sub>a</sub>s for any enol is desirable.

5.2 Predicting the pK<sub>a</sub> of enols

5.2.1 Establishing a HCS

The two enols of Balogh’s study (the pesticides tralkoxydim and sethoxydim) are β-diketones, a structural motif that inspired the data set of this work. β-diketones are also found in tetracyclines, which are important in medicinal chemistry. Therefore, all enols within this article are enols of a β-diketone structure, which corresponds to the most stable tautomer. Figure 5.1 shows the non-cyclic and cyclic molecular skeleton relevant to the current study, the latter being illustrated with two representative large molecules studied here, the tetracycline epichlortetracycline (bottom left Figure 5.1) and the pesticide sethoxydim (bottom right Figure 5.1). The data set of enols used to find the HCS for the enolic functional group consisted of 22 molecules. Eight molecules were taken from the Lange data set<sup>193</sup>, consisting of three molecules with a common cyclic enol skeleton (Figure 5.1 top left) and five with a non-cyclic enol skeleton (Figure 5.1 top right). The remaining 14 molecules were from the Prankerdl data set<sup>194</sup>, which included 7 biological molecules with a common tetracycline skeleton shown in Figure 5.1 (bottom left).
For each molecule of the set the bond lengths marked \( a \) to \( e \) in Figure 5.1 were obtained from an \textit{ab initio} calculation. The program GAUSSIAN03\textsuperscript{27} carried out full geometry optimisation of all molecules at HF/6-31G(d) level, which was established in refs.\textsuperscript{191,195} as sufficient (and remarkably) accurate compared to higher levels of theory. Starting geometries for the tetracyclines were prepared by the program GAUSSVIEW, which placed the amide group (see Figure 5.1, bottom left) in a position roughly perpendicular to the cyclic enol ring (marked in blue). This perpendicular position did not introduce a bias towards either the O-H...O or the O-H...N intramolecular hydrogen bond occurring in the final optimised geometry. The \textit{ab initio} bond length of each of the five bonds investigated (\( a \) to \( e \)) was taken in turn as a single descriptor to regress against experimental pK\textsubscript{a}. Typically, one of these five bonds emerges with the highest possible correlation coefficient, thus serving as an ideal bond to use in setting up a HCS.

\textbf{Figure 5.1:} Common skeleton of enol in cyclic (top left) and non-cyclic (top right) configurations with bonds \( a \) to \( e \). The stars are carbons in all molecules but one, in which they represent the location of oxygen atoms. Epichlorotetracycline (bottom left) with the cyclic enol skeleton of interest highlighted in blue, and sethoxydim (bottom right).
5.2.2 Non-cyclic enols

A preliminary analysis showed that cyclic and non-cyclic enols cannot form a single HCS. Indeed, several investigated bond lengths of the non-cyclic enols differed greatly from those of the cyclic enols. In non-cyclic compounds the enol form is very stable since it forms a very strong O-H...O intramolecular hydrogen bond (see Figure 5.1 top right). This is probably the reason why they do not belong to the same HCS as that of the cyclic derivatives. The biological molecules that we wish to predict the pK\textsubscript{a} of are cyclic enols. Therefore, the non-cyclic enols were removed from the data set and the possibility to predict the pK\textsubscript{a} of non-cyclic enols was not further investigated.

5.2.3 Bond length determination

Figure 5.2 explores each of the five bond lengths (each panel a to e corresponds to a bond length) as the single descriptor for the HCS. Comparing the five panels visually, with the help of the two enlarged insets for O-H and C=C, clearly demonstrates that the C=C bond emerges as the ideal bond for the cyclic enol HCS. Bond c is the ideal bond because out of the five possible bonds it shows the largest number of data points with the best local linear relationship (red points at the left in Figure 5.2c).
Figure 5.2: Relation between experimental \( pK_a \) and \textit{ab initio} bond lengths for cyclic enols for the five different bonds of the enol group (\( a \) to \( e \) in Figure 5.1).

The overall correlation for the Figure 5.2c is poor. However, the overall set divides itself into three HCSs highlighted as red, blue and green, which we now discuss in turn. The first HCS (red) contains the majority of points, which all lie between bond lengths 1.33 Å and 1.35 Å. For this HCS an outlier is identified at a bond length of approximately 1.34 Å and \( pK_a \) of 5.1 units. This data point was examined and compared to the rest of the points within the HCS. The common cyclic enol skeleton of the outlier contained two ether oxygens at the two starred positions of Figure 5.1, whereas the remainder of the molecules within the HCS were all pure carbon rings. The outlier was removed from the HCS as it did not belong to the HCS being investigated. There is a probability that the outlier belongs to a separate HCS but not enough data points are available to verify this.
The second HCS of Figure 5.2c (blue) contains two molecules of bond length approximately 1.36 Å. These two molecules are the pesticides tralkoxydim and sethoxydim. They both contain cyclic enols but are unique to the rest of the set as the intramolecular hydrogen bond forms between the enolic hydrogen and an imine nitrogen rather than the amide nitrogen (the latter appearing in Figure 5.1). This imine group appears only in these two molecules of the full data set (containing 17 molecules). Lack of further experimental data prevents an expansion of this HCS. However, the remaining green HCS data points demonstrate an important asset of the current method.

The third HCS consists of the five green data points between bond lengths 1.36 Å and 1.38 Å. These molecules all turn out to be tetracyclines (see bottom left panel of Figure 5.1). It was mentioned before that there are seven tetracyclines in the data set, which triggers the question as to why this third (green) HCS does not contain all seven. In fact, it works out that the two missing tetracyclines are in the first (red) HCS. This observation is curious because one would expect a well-defined congeneric set such as tetracyclines (containing the blue enolic ring in Figure 5.1) to form one single HCS. How can this problem be solved? The key to the solution is the way the intramolecular hydrogen bond between the enolic hydroxyl group and the amide group forms. We show that a simple conformational correction can make all tetracyclines gather under the same HCS, which turns out to be the first (red) HCS. The five “green” tetracyclines form an O-H...O intramolecular hydrogen bond between the enolic OH group and the oxygen of the amide group. However, the “red” tetracyclines form the intramolecular hydrogen bond between the enolic OH group and the nitrogen of this amide group, i.e. O-H...N. The latter hydrogen bonded conformation is the one shown in Figure 5.1. This is the lower energy conformation compared to the former conformation. The amide groups in the five green tetracyclines were “flipped over” so that all seven tetracyclines formed the intramolecular hydrogen bond in the same conformational manner. The amide group was rotated to encourage an intramolecular hydrogen bond between the amidic nitrogen and the enolic hydrogen by positioning these two atoms in close proximity. Each molecule of the third (green) HCS was then re-optimised starting from its new geometry with the O-H...N intramolecular hydrogen bond. It was observed that the intramolecular hydrogen was preserved during the geometry optimisation. The new fully optimised geometries of the five green tetracyclines were indeed of lower energy than the optimised geometries with the O-H...O hydrogen bond.
Figure 5.3 shows the success of the conformational correction described above. The five green tetracyclines now belong to the first (red) HCS after reoptimisation into the lower energy conformation. The arrows in Figure 5.3 show how each (green) tetracycline moves to the left, joining a single overarching HCS, which consist of red molecules. The lower energy conformations now adopt bond lengths between 1.34 Å and 1.35 Å, which is approximately 0.02 Å smaller than the bond lengths of the higher energy conformations (green). Note that the relative constellation of the five green data points is not invariant under translation to the left. This means that the way the C=C bond length is affected by substituent variation is slightly different depending on the amide conformation.

Figure 5.3: The self-correcting process of HCS formation induced by conformational variation. Flipping the amide group from the high energy position (right, green) to the low position (left, red) forms the desired O-H...N intramolecular hydrogen bond, which establishes the full HCS.
5.2.4 pK$_a$ prediction from HCSs

Once an HCS is established, the predictive capability of this HCS can be tested for molecules external to it. In this work we carry out this test using the red HCS of Figure 5.3, which contains 15 data points. We rebrand the set of the five formerly green tetracyclines in Figure 5.3 as an external test set, and treat them as if their pK$_a$ is unknown. This removal ensures an unbiased prediction. An equation was derived for the new HCS built from the remaining 10 molecules. Table 5.1 compares the pK$_a$ values predicted by the HCS method for the five tetracyclines of the external test set against experiment. It is pleasing to see that the absolute discrepancies between experiment and prediction are all smaller or just at the desirable threshold of 0.5 pK$_a$ units. The root-mean-square error of the five data points is 0.32 pK$_a$ units and the corresponding mean absolute error (MAE) is 0.296. The worst predicted tetracycline (oxytetracycline) had an error of 0.51 pK$_a$ units.

The MAE for the HCS was compared to the MAE of the VCC pK$_a$ prediction package. VCC was chosen because it is considered one of the strongest performers in a recent study by Balogh et al. In their study the MAE of VCC for the full GOLD data set, which includes enols, was 0.295. A comparison between the results generated the VCC and the HCS methods can be made even though both methods were evaluated against different test sets. Balogh et al. identified the enols of the GOLD set as one of a handful of functional groups that were the most difficult to predict accurately. The MAE is less sensitive to large discrepancies for a given number of compounds if they occur in large set compared to them occurring in a small set. This is so because the MAE is a mean value, and hence the result of a division by the total number of compounds. Therefore, if a set of enols is subsumed in a larger set of molecules for which the pK$_a$ is readily predicted in an accurate way, the MAE of the full set will give the false impression that the enols were also predicted accurately. This “masking” of the error in prediction for the enols does not occur for the HCS method as the test set consists exclusively of enols. In the light of this discussion, the HCS MAE of 0.296 compares favourably with the VCC MAE of 0.295.
<table>
<thead>
<tr>
<th></th>
<th>Experimental $pK_a$</th>
<th>Predicted $pK_a$</th>
<th>Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anhydrochlortetracycline</td>
<td>3.28</td>
<td>3.58</td>
<td>0.30</td>
</tr>
<tr>
<td>Epichlorotetracycline</td>
<td>3.65</td>
<td>3.37</td>
<td>0.28</td>
</tr>
<tr>
<td>Isochlorotetracycline</td>
<td>3.10</td>
<td>3.28</td>
<td>0.18</td>
</tr>
<tr>
<td>Oxytetracycline</td>
<td>3.27</td>
<td>3.78</td>
<td>0.51</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>3.30</td>
<td>3.51</td>
<td>0.21</td>
</tr>
</tbody>
</table>

**Table 5.1**: Experimental and predicted $pK_a$ values of five tetracyclines.

The size of the data set from which the HCS is derived is admittedly small. This concern triggers further investigation into the dependence of the HCS on the number of data points. We therefore expanded the data set from 10 to 14 data points, by including four of the five tetracyclines from the formerly external test set (green in Figure 5.3). The $pK_a$ is then predicted (from the 14 data points) for the one tetracycline that was not included into this expanded set. This procedure is repeated for each of the five tetracyclines (from the formerly external test set).

<table>
<thead>
<tr>
<th></th>
<th>Absolute error from original enol set prediction</th>
<th>Absolute error of tetracycline when excluded from set</th>
<th>Absolute difference in error between original and extended HCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anhydrochlortetracycline</td>
<td>0.30</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>Epichlorotetracycline</td>
<td>0.28</td>
<td>0.42</td>
<td>0.14</td>
</tr>
<tr>
<td>Isochlorotetracycline</td>
<td>0.18</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Oxytetracycline</td>
<td>0.51</td>
<td>0.50</td>
<td>0.01</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>0.21</td>
<td>0.15</td>
<td>0.06</td>
</tr>
</tbody>
</table>

**Table 5.2**: Absolute difference (right column) in $pK_a$ for five tetracyclines between a prediction error from the 10-point data set (left) and the 14-point data set (middle).

Table 5.2 shows that four of the five tetracyclines benefit from improved a $pK_a$ prediction when using the equation obtained from the 14-point data set. When there was an improvement it was not large and nor was the deterioration in the case of epichlorotetracycline. Therefore once a high correlation subset has been found the equation remains accurate and robust, irrespective of data set size. This means that the quality of the predictions is dependent on the quality of the descriptor itself, rather than the data set size used to derive the equation.
5.2.5 Calculation accuracy

The bond length differences between molecules within a given HCS are very small. We need to verify if such small changes in bond length can be trusted when used in a correlation, that is, if the changes are significant. The bond length differences between molecules are as minute as a few thousandths of an Ångström. If a change in bond length is so small that it cannot be distinguished from the "noise" of the ab initio calculation then the correlations cannot be trusted. In that case, the correlations are affected by noise and any underlying chemical reason for the correlation is masked by this noise. Noise arises as an artefact of the iterative geometry optimisation procedure, which terminates when the set of four convergence criteria are met (maximum force, root-mean-square force, maximum displacement and root-mean-square displacement). When convergence is reached, a bond length is uncertain beyond the number of its significant figures enforced by the convergence criteria. This uncertainty can be determined as follows. A standard geometry optimisation takes twenty to thirty iterations to reach convergence. Figure 5.4 shows a typical optimisation where it can be seen that rather large fluctuations in bond length diminish substantially in about the last half of the optimisation interval. Inspection of this last half shows a slowly decreasing zone between iteration 10 and 17, followed by a visually stagnant zone of 7 iterations. Estimating the noise from the last 10 iterations can be therefore considered as conservative. Indeed, it would be misleading to underestimate this noise in the analysis just below. The noise for each molecule was calculated as the difference between the largest and smallest bond length out of all the geometries taken from the last 10 iterations, and typically varies around the 0.0006 Å mark.

![Figure 5.4: Optimisation path of Isochlorotetracycline.](image)
Figure 5.5 shows the 15-point HCS (red set in Figure 5.3) again but now together with a "worst possible scenario" HCS (black data point and concomitant line) where each data point has suffered a maximum possible noise-induced shift (of 0.0006 Å) in a direction that aims at destroying the correlation. Practically, this means that a red data point that lies left of the red HCS line, will be shifted further to the left. Similarly, a red data point at the right hand side of the red HCS line will be shifted further to the right. The shifted data points are shown in black in Figure 5.5. Note that this "worst possible scenario" is actually unlikely because it applies the maximum noise to each data point in a consistently averse manner. In going from the original red line to the new "worst possible scenario" black line, the $r^2$ correlation decreases from 0.87 to a still respectable 0.77, a difference of 0.1. Although this is a sizeable reduction in correlation, this unlikely "worst possible scenario" does not destroy the HCS. However, in order to gauge the deterioration of prediction quality caused by the "worst possible scenario", it is more informative to predict the $pK_a$ directly for a given molecule from a new equation derived from the "worst possible scenario" HCS. The overall discrepancy between the $pK_a$ predictions and experiment for the original 15-point HCS is gauged by a root-mean-square error of 0.24 $pK_a$ units. The root-mean-square error between experiment and "worst possible scenario" $pK_a$ predictions amounts to 0.32 $pK_a$ units. Hence, the worse possible effect of noise leads to an increase of root-mean-square error of only 0.08 $pK_a$ units. In summary, the decrease in the predictive ability of the HCS due to noise is negligible, and the small differences in the bond lengths can be considered as meaningful and genuine results that are unaffected by computational noise. This finding is important because it confirms the validity of using minute bond length differences. Correctly picking up such minute differences is necessary in the HCS method because conformational variation (e.g. amide group rotation) induces only subtle changes in bond length.
5.2.6 Calculation time

With an eye on industrial applicability of the HCS method we explore here if the CPU computation required to geometry-optimise molecules of interest can be reduced. The \textit{ab initio} program \textsc{Gaussian03} applied its default threshold values for each of its four convergence criteria to decide when a molecule is geometry optimised. For example, Figure 5.4 shows that the number of iterations needed for isochlorotetracycline is 24 and takes approximately 14 CPU hours using only a single core. The small effect, of the second half of the geometry optimisation interval, on a bond length and the resulting correlations suggest that CPU time can be reduced by relaxing the convergence criteria. This is achieved by examining how HCS-predicted pK$_a$ values change as a function of the number of iterations in the geometry optimisation.

Three representative molecules were selected in order to answer the question how many iterations are sufficient. Oxytetracycline, epichlorotetracycline and isochlorotetracycline were chosen because they represent the worse, median and best predicted molecule, respectively, in terms of deviation from experiment (see Table 5.2).

Figures 5.6 a to c show, for each representative molecule, the pK$_a$ calculated from the bond length obtained from a geometry optimisation frozen at each iteration (using the equation of the “original” 15-point HCS obtained with the full number of iterations, e.g. 35 in the case of epichlorotetracycline).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure5.5.png}
\caption{Original 15-data point HCS (red) distorted by applying noise to each data point, leading to “the worst possible scenario” HCS (black).}
\end{figure}
Figure 5.6: Predicted pK<sub>a</sub> values from the C=C bond length for each iteration of the geometry optimisation for three selected tetracyclines (blue). The experimental pK<sub>a</sub> value is marked in red.
For all three molecules, the calculated bond length suddenly reaches a plateau after a “threshold iteration number”, which is typically about half the full number of iterations. This threshold number can be determined by noting that $\text{pK}_a$ values predicted from bond lengths obtained at any of the iteration numbers larger than this threshold are always within 0.5 $\text{pK}_a$ units. For epichlorotetracycline, oxytetracycline, isochlorotetracycline this threshold iteration number is 13, 11 and 10, respectively. In summary, the CPU time can be halved if one allows the predicted $\text{pK}_a$ value to be off by 0.5 log units. Cruder $\text{pK}_a$ predictions (allowing for error bars of 1 to 2 log units) do not introduce substantial further savings in CPU time. This is a consequence of the cliff-like nature of the profiles in Figure 5.6, in the vicinity of the plateau. However, there are some serendipitous predictions at a very low number of iterations (around iteration number 5) but these cannot be used in practice due to the lack of a clear local pattern. So overall, users interested in crude $\text{pK}_a$ predictions will save only half the CPU time used to set up an HCS equation, but will benefit from their prediction being accurate by 0.5 log units.

5.3 Predicting the $\text{pK}_a$ of guanidines

5.3.1 Establishing a HCS

The second functional group found by Balogh that has a $\text{pK}_a$ value that is consistently poorly predicted is weak $\text{–NH}$ acids, in particular guanidines. The possible tautomers of guanidine introduce an added difficulty for $\text{pK}_a$ prediction. An incorrect tautomeric structure may produce poor predictions. This problem is also displayed in the enols, but to a lesser extent. Any changes in tautomeric form of the guanidines will be apparent using high correlation subsets. If the bond length of interest is involved in the tautomerisation then changing between tautomeric forms will change the bond type from a single bond to a double bond, or the reverse. A change in bond order will result in a large difference in bond length relative to the other bond lengths of the molecule. This is evident when comparing the average bond lengths of C-N and C=C which are 1.47Å and 1.27Å, respectively. Therefore, two guanidine groups of different tautomers would have relatively large differences in bond lengths for the given bond. These two guanidines would not be comparable in the same subset.
The molecules for the guanidines were taken from the Williams data set. The guanidine data set consisted only of small organic molecules containing a guanidine functional group. This data set was chosen to derive the guanidine HCS’s equation for the prediction of the pKₘ values of five biological molecules. By excluding the biological molecules from the data set the derived equation remains unbiased towards the biological molecules. Therefore, the equation can be seen as a predictive model whose strength lies in the method rather than the training set. The data set consisted of only seven small molecules containing guanidine shown in table 5.3. The molecules were optimised at HF/6-31G(d) level. All molecules were optimised in the same tautomer. Figure 5.7 shows the five possible tautomeric forms of guanidine where all guanidine groups were of the first tautomeric form. The six highlighted bond lengths in Figure 5.7 were correlated with the experimental pKₘ for each molecule. The bond producing the strongest subset by visual inspection was assigned as the ideal bond length to use for further study. The N-H bond of the nitrogen adjacent to R1 was not measured.

<table>
<thead>
<tr>
<th>pKₘ</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AcetylGuanidine</td>
<td>8.33</td>
</tr>
<tr>
<td>Benzamidine</td>
<td>11.6</td>
</tr>
<tr>
<td>DiGuanidine</td>
<td>13.25</td>
</tr>
<tr>
<td>Ethylenediguanidine</td>
<td>11.76</td>
</tr>
<tr>
<td>Guanidine</td>
<td>13.71</td>
</tr>
<tr>
<td>MethylGuanidine</td>
<td>13.4</td>
</tr>
<tr>
<td>Phenyl diguanidine</td>
<td>10.71</td>
</tr>
</tbody>
</table>

Table 5.3: Guanidine derivatives data set.

Figure 5.7: The five tautomers of the guanidine functional group in the monosubstituted derivative. Panel 1 shows the labelling scheme for bond lengths a to e used throughout.
5.3.2 Bond length determination

Figure 5.8 shows that the C=N bond produces the strongest HCS. Figure 5.8e does not include the bond length for benzamidine because the corresponding bond is a C-C bond, instead of a C=N bond. The C-C bond length is 1.49 Å and lies well outside the subset. The $r^2$ values for the bond lengths a to e are 0.01, 0.97, 0.04, 0.07 and 0.50, respectively. These correlations shows that the C=N bond length should be used for deriving the equation for the guanidine.
Figure 5.8: Correlation of \textit{ab initio} bond lengths with experimental pK$_a$ values for guanidines in the first tautomeric form. Panel (d) represents both N-H bonds as they are almost indistinguishable. The bonds are labelled in one panel of Figure 5.7 for clarity. Panel (b) includes the alternative experimental pK$_a$ value for guanidine (13.6 taken from Foye's) given by the red triangle.

![Figure 5.8](image)

(e)

Famitodine

Metformin

Amiloride

Debrisoquine

Arginine

Figure 5.9: Five biological molecules used to test the guanidine HCS.

5.3.3 pK$_a$ prediction of five biological molecules

The equation for the guanidine subset was tested with the five drug molecules shown in Figure 5.9. The molecules were taken from the Williams and GOLD data sets. The five drug molecules were famitodine, metformin, amiloride, debrisoquine and
arginine. Of the five drug molecules debrisoquine was unique in that the nitrogen of the guanidine group that connects debrisoquine to the rest of the molecule is a tertiary amine. For the remaining four drug molecules the guanidine group is connected to the rest of the molecule through a secondary amine. This difference in the connecting amine was reflected in the results when using the equation derived from Figure 5.8b. The experimental pKₐ of debrisoquine is 12.5 but predicted as 17.3. The position of debrisoquine relative to the rest of the guanidine subset is shown in Figure 5.10, the C=N bond length was 1.266Å. Because debrisoquine lies so far outside of the subset it is probable that it belongs to a separate HCS of other guanidine containing molecules. However, this can only be verified if enough data points can be collected which are relevant molecules to create the desired HCS.

**Figure 5.10:** Correlation of the guanidine data set including five added biological molecules. Debrisoquine (red triangle) is highlighted as the only molecule with the guanidine functional group connected to the rest of the molecule through a tertiary amine. The four other biological molecules added to the set are highlighted in black (circles).

The experimental pKₐ of metformin is 12.4, and its pKₐ predicted as 11.8 with an error of only 0.6 pKₐ unit. The most accurately predicted biological molecule from the set of five was the amino acid arginine. The experimental pKₐ of arginine is 12.5. The calculated C=N bond length of arginine was 1.261 Å and predicted a pKₐ of 12.4, an error
of only 0.1 pKₐ units. The pKₐ predictions for amiloride and famitodine are discussed below with respect to tautomeric effects.

5.3.4 Tautomeric effect on pKₐ prediction

Functional groups that are subject to tautomeric effects are known to be problematic with respect to pKₐ prediction. The sensitivity of ab initio bond lengths to changes in tautomeric form was tested with famitodine. Famitodine was optimised in the first and third tautomeric forms of Figure 5.7. The famitodine in the third tautomeric form has a second C-NH₂ where the C=N bond of the first tautomer would be. Already this gives an indication that the two different tautomeric forms will produce greatly different pKₐ values when predicted from their bond lengths. The range of bond lengths for the equation is from 1.256 Å to 1.262 Å; however the C-N bond length of the third tautomer is 1.345 Å. The bond length lies well outside the range of bond lengths expected for the guanidines. The experimental pKₐ of famitodine is 6.98 and its predicted pKₐ from the third tautomer is 86.3 pKₐ units. The famitodine is in a different tautomeric form to the other molecules of the HCS. Famitodine in the first tautomeric form matches the tautomeric form of the guanidines in the set. The bond length was calculated as 1.255 Å and the predicted pKₐ is 7.49. The predicted pKₐ for the third tautomer had an error of 79.3 pKₐ units, whereas the predicted pKₐ for first tautomer had an error of just 0.5 pKₐ units. The same test was applied to metformin. An identical set of tautomers, both first and third tautomers were optimised and the corresponding bond length measured. The tautomer in the third form had a predicted pKₐ of 107.8. The bond length from the correct (first) tautomeric form for metformin predicted a pKₐ of 11.8 and an error of only 0.6 pKₐ units. Such a dramatic improvement in the prediction of the pKₐ for two different tautomers shows that the HCS method is capable of differentiating between tautomeric forms.

HCSs were derived for each of the five different tautomeric forms. Every molecule from the set was optimised in each tautomeric form. The data sets for each tautomer included the drug molecules because a larger data set produces more trustworthy correlations. Equations were derived for each tautomer where a tautomer is considered as a separate subset to the other tautomers. The correlations of the bond length chosen for each tautomer are in Figure 5.11a to e.
Figure 5.11: Correlation of *ab initio* bond lengths to experimental pK$_a$ values for the bond length of interest for each tautomers 1 to 5 in panels (a) to (e), respectively. The C=N bond appears in a different place depending on the tautomer, as shown in Figure 5.7.

For tautomers 1 to 4 there is always one bond length that has a noticeably stronger correlation than the other bonds of the given tautomer. The correlations for bonds a to e (corresponding to bonds a to e in Figure 5.11) for tautomer 1 were 0.70,
0.95, 0.03, 0.56, 0.49 and 0.15 respectively. For tautomers 1, 2, 3 and 4 the C=N bond was the ideal bond. Tautomers 1 and 2 share the same ideal bond, whereas the bond length for tautomer 4 is the same C=N bond type but lies in a different position with respect to the rest of the molecule. Tautomer 3 does not use the same bond relative to the rest of the guanidine functional group. The C=N bond of tautomer 3 is not a terminal bond, it is the C=N bond at the centre of the guanidine group. The remaining tautomer (5) does not have an obvious bond of choice. Both bonds b and e have correlations of 0.81 so neither is the obvious choice as the ideal bond for the HCS. However, bond b is the C=N bond that corresponds to the bonds chosen for the other tautomers. The bond is of the same type as the other tautomers and same position as tautomer 4, which is the tautomer it has strongest resemblance to. Therefore, it is not unreasonable to assume that it is in fact this bond which should be used to derive the equation for tautomer 5.

The type of the bond, and not the position, determines the best correlation to pKₐ, this “choice” of bond implies that the best correlation bond is not entirely random and has physical meaning. The experimental guanidine pKₐ values are determined for the protonated form of the guanidine. The protonation occurs on the nitrogen of the C=N bond and therefore it is clear that this bond is the most important in determining the pKₐ. While this choice of bond seems obvious at first it is important to remember that the ab initio calculation has no knowledge of the protonated state of the guanidine. The calculation is performed on a single neutral molecule in the gas phase. The fact that this quantum mechanical treatment has found this property of the molecules is remarkable and leads to the belief that there must be chemical meaning driving the HCSs. The pKₐ values for guanidines correspond to the dissociation of the guanidine group in the protonated form, and are calculated as 14 – pKₐ. For all tautomers the protonation site is the Nitrogen of the C=N bond and supports the choice of bond length selection for correlation by the high correlation subset method.

5.3.5 Relationships between HCSs

Within the scope of HCSs the subsets have been rigorously examined but as isolated data sets without a focus towards the significance of a subsets position in chemical space relative to other subsets. If each subset is determined by the chemical nature of the molecules within it then relationships must exist between different subsets. Such an analysis is easiest performed on the tautomer subsets for two reasons:
(1) There are both distinct and subtle differences between the tautomers. For example, tautomer 3 differs to tautomers 1, 2, 4 and 5 by the type of tautomer i.e. location of the C=N relative to the rest of the guanidine. Tautomers 1, 2, 4 and 5 can be subdivided into tautomers 1 & 2 and 4 & 5 based on the position of the C=N bond relative to the rest of the molecule.

(2) All correlations for the tautomers are for the same bond type (C=N), the bond type for correlations in other subsets do not always match. For example, the carboxylic acids and alcohols differ in their ideal bond.

The correlations for the five tautomers are superimposed in figure 5.12. Of the five tautomers the correlation for the third tautomer is completely different to other tautomers as expected. The remaining four tautomers are closer together and relatively parallel. Tautomer 3 also has a much larger range of bond lengths and its correlation has a negative gradient. It is clearly identified as the most different tautomer of the five by its HCS. The remaining four tautomers lie much closer together and would suggest a stronger relationship between them. This is the case when the structures are examined. The set of four can be divided into two sets where each set has two tautomers, tautomers 1 and 2 in one set and tautomers 4 and 5 in the other. As previously discussed tautomer 1 holds a much stronger relation to tautomer 2 than to 4 or 5 because tautomers 1 and 2 have the C=N bond in the same position relative to the rest of the molecule. The same holds when comparing the reverse, i.e. comparing tautomers 4 and 5 to 1 and 2.
Figure 5.12: Superimposed correlations of the C=N bond to experimental pKₐ for five guanidine tautomers

The choice of tautomer for the prediction of the pKₐ was assessed by comparing the pKₐ predictions of amiloride using the equation of each of the five tautomers. Balogh et al. evaluated five commercial pKₐ prediction packages, and found that amiloride was one of a handful of molecules poorly predicted. The experimental pKₐ of amiloride is 8.65 and for the five commercial packages the pKₐ predictions were; ACD 2.29, Marvin 3.29, Pallas 4.04, Epik 6.75 and VCC 9.00. The largest error in pKₐ prediction was 6.36 pKₐ units for ACD, whereas the error for the best prediction was 0.35 pKₐ units. Four of the five commercial packages had errors of greater than 1.5 pKₐ units. With HCSs the predicted pKₐ values for amiloride in tautomeric forms 1 to 5 were 7.78, 5.34, 8.53, 9.07 and 9.78, respectively with absolute errors of 0.87, 1.03, 0.12, 0.42 and 1.13, respectively. Errors between 0.12 and 1.13 pKₐ units highlight the importance of the tautomeric form of the molecule on its pKₐ prediction. The commercial packages evaluated in Balogh’s study use the user input structure exactly, preserving its tautomeric form for the calculation. It was noted by Balogh that the packages may have predicted the molecules from the data set so poorly because an incorrect structure for the molecule was executed as a query. Out of all five tautomers the third tautomer is the most considered stable tautomer of amiloride by HCSs because the prediction for this tautomer was the closest to the experimental value. The study by Smith et al.
confirms the most stable tautomer is tautomer 3 through NMR, agreeing with the HCS conclusion\textsuperscript{196}.

\subsection*{5.3.6 Defining HCSs}

It is possible to continue to divide the HCSs as they consist of subsets of their own. It has been proven that each HCS exists separate to other HCSs due to chemical or structural subtleties, as a subset clearly defines the molecule types that exist within the set. Within each subset there exist “sub” subsets of their own that can be considered in a similar fashion to the higher level sets which they belong to. Indeed, higher level subsets have already been discussed in the case of the five guanidine tautomers. Tautomers 1 & 2 and 4 & 5 form combined subsets respectively. As the level of the subset increases the correlation deteriorates as the description of the molecules that lie within the set broadens, but this correlation and prediction can still be meaningful. Taking the C=N bond length from tautomers 1, 2, 4 and 5 of amiloride the pK\textsubscript{a} values were predicted from each bond length using an equation derived from these four tautomers combined. The MAE of the four tautomers was 0.87. As the subsets are revealed that lie within larger sets the equations become more specific to molecular classifications, but the subsets formed within each layer continue to hold physical reasons as to its existence as a subset. Figure 5.13\textit{b} breaks the subset for tautomer 3 into its own subsets where the full set is shown in figure 5.13\textit{a}. Five molecules were added to the set to determine the strength of the subsets, guanabenz, guanethidine, guanoxan and two amiloride derivates were chosen due to their relation to the subsets that were found.
Figure 5.13: Separation into subsets of the HCS of the third tautomer (3) of guanidine derivatives.

Subset A defines guanidines in an alkyl environment. Subset B contains guanidines that are connected to the rest of the molecule through an amide bond, with
the exception of famitodine that lies furthest from the other points. Famitodine still belongs to the subset because in place of the C=O of the amide bond is a C=S and so can be considered part of the set. Subset C is comprised of diguanidine derivatives with the exception of guanabenz. Guanabenz does not have a diguanidine group however the guanidine group of guanabenz is connected to the rest of the molecule through a single nitrogen. Guanabenz does not strictly belong in subset C but it bears closest relation to the molecules within the subset relative to subsets A and B, which is reflected by its position relative to the other molecules of subset C. All subsets improve as the definition of the subset is increased in strictness. Figures 5.14a to d displays this increase in subset strength as the definition of the set is increased in strictness. Figure 5.14a shows subset C with famitodine. Famitodine is the most different to the other molecules within the set and as it is removed the set improves, as seen in figure 5.14b. The biggest outlier of the set is now guanabenz, and the rationale of this observation has already been discussed. Guanbenz’s removal from the set improves the definition of the subset to exclusively diguanidine derivatives in Figure 5.14c. The outlier in Figure 5.14c is diguanidine, the only molecule form the set that is not a diguanidine group attached to the rest of a molecule. Diganudine has a hydrogen where the rest of the molecule would be for other molecules in this set. The final set is in Figure 5.14d, where all molecules are diguanidines connected to alkyl groups. This is the strictest definition of a subset that can be made from subset C with the data available and has an almost perfect correlation.
Figure 5.14: Reduction of data points from subset C of guanidine tautomer 3 derivatives increasing strictness of subset definition.

Dividing a data set into subsets HCSs produces meaningful correlations that can predict pKₐ's accurately. As the subsets are further divided chemical meaning is not lost. Indeed, the reverse occurs, the relation between the molecules within the subsets is strengthened. Definition of the molecules within a subset is vital to accurate predictions. Broad subset definitions produce less accurate pKₐ predictions and predicting for a molecule that does not belong to a subset gives even worse predictions. When the subsets for tautomer 3 were divided into subsets A, B, and C, the five new drug molecules lay reasonably within a subset, and their positions relative to the subsets was justified. Inclusion of the five new drug molecules into the entire set for tautomer 3 and a new correlation gives an r² of 0.69 compared to the original r² of 0.88. The two outliers are famitodine and guanabenz. With a general description of the subset that of guanidine derivatives, guanabenz and famitodine are indeed the two molecules furthest from this generalised description for the set. The guanidine groups for both molecules are connected to the rest of the molecule through unique functional groups with respect to other molecules in the set. The point of interest is that when the definition of the subsets is changed to, for example, diguanidines then guanabenz can exist within the set without ruining the correlation. The same can be said for famitodine and subset B. The cases of famitodine and guanabenz and their effect on the correlations of two different subsets show the importance of the subset’s definition. The commonality shared by the molecules within the subsets is vital for accurate predictions.
5.4 Conclusion

The HCS method was applied to two functional groups (enols and guanidines) that both have difficult pK\textsubscript{a} values to predict. For the enol molecules it was found that the geometry of the enol was vital to accurate pK\textsubscript{a} prediction. This reliance on geometry was due to the enolic hydrogen forming hydrogen bonds with different atoms when in the two different geometries. It was the lowest energy geometry that was required for the most accurate predictions. The problem of bond length sensitivity was addressed applying an error of calculation of to the bond lengths of all molecules within the HCS. It was found that the variations in the bond lengths between molecules was significantly greater than the error within the calculation and therefore considered as meaningful and reliable results.

For the guanidine data set five separate tautomers each formed their own HCS. All five of the tautomers had the strongest correlation using the C=N bond of the tautomer. This C=N bond corresponds to the site of protonation. The pK\textsubscript{a} values of five drug-like molecules were all predicted accurately with the HCS of the first tautomer. However, it was found that the pK\textsubscript{a} of amiloride was more accurately predicted by its third tautomer HCS, which is confirmed by experiment to be its most stable tautomer. The HCSs themselves were found to consist of smaller subsets, where the commonality between the molecules of these subsets is stricter in its definition than for the original HCS.
Chapter 6

Conclusions and Future work

6.1 Bioisosterism of heterocyclic fragments

The aims of this work were to improve and evaluate the QID as a ligand design tool. The QID tool is only able to suggest replacement linker fragments and the work done aimed to establish methods to move the QID towards completion, where any fragment of any molecule can be replaced by the QID. The inclusion of a pKₐ prediction method into the QID tool improves its value as a ligand design tool. An existing method of pKₐ prediction was expanded in its application range and tested against biological molecules.

To extend the QID beyond linkers a method for characterising ring fragments was developed. Initially a correlation between ring atom properties and RCP properties was established. However, the reverse of this correlation (prediction of ring atom properties from RCP properties) was not possible. It was discovered that the RCP ellipticity was extremely sensitive to changes in a ring and therefore all correlations suffered as a result. The ring atom and RCP properties were separated and a method for ring characterisation developed for each individually. For the ring atom properties the combined net charge of the ring atoms was determined by four ring features. The four features were: the element of the heteroatoms within the ring, the number of heteroatoms within the ring, the type of substituent on the ring and the position of the substituent. Remarkably these four features affected the ring's net charge independently. Each feature had a systematic effect on the ring's net charge, which was always the same irrespective of the other features. Therefore, in terms of bioisosterism, ring features can be changed to alter the ring's net charge to a desired value based on the alterations made. For the RCP properties a 3D space was establish with the electron density, its Laplacian and the RCP ellipticity. By placing rings in this RCP space a ring's position within the space was determined by its scaffold. For example, rings with more heteroatoms were found higher up in the space. Therefore, the RCP space was able to characterise rings based on their ring scaffold and the position of any ring can be predicted just through comparison of its ring scaffold with other rings in RCP space. The importance of RCP space in bioisosterim is the positions of the rings relative to each other. Rings that were structurally similar to each other were placed closer in RCP space.
Finally, the RCP and ring atom properties were tested for an IGPD inhibitor. The RCP properties alone were not able to fully account for the relative activities of the inhibitors because they could not account for the positions of hydrogen bond acceptors on the ring. When ring atom properties were introduced the combination of RCP and ring atom properties were able to resolve the problem with hydrogen bond acceptors in the correct position of the ring. By successfully characterising ring fragments through QCT properties it is possible to make comparisons between the rings. If comparisons can be made then bioisosteric replacements can be suggested.

For the ring fragments to be incorporated into the QID the two key areas require attention. First, the number of substituent types requires expanding. The initial findings require conformation for a larger set of substituent types and the subtle effects of the atoms beyond the first atom of the substituent need to be characterised. This must be done to establish a method of determining how to model any substituent within the QID framework. In other words, if a substituent is presented to the QID, where the effects of the substituent are not yet known, then it should be possible to predict the effects of the substituent from its structure. Secondly, the findings of ring characteristics and RCP space need to be confirmed for rings of different sizes. Currently, only five-membered rings have been modelled. However, the RCP and ring atom properties are calculated for the ring atoms only, the different number of atoms within a ring is not expected to produce different conclusions, but this requires verification. The approach is also expected to work for fused rings as well because the fused rings only introduce additional atoms and RCPs which can be modelled with QCT as the number of atoms should not affect the model.

6.2 Evaluating the QID and addition of terminal groups

For the QID to be able to replace any fragment of any molecule the terminal fragments require entry into the database. However, the current structure of the QID is not suitable for terminal fragments. A new method of comparing the terminal groups was proposed that relies on the assumption that the geometry of a fragment is imprinted in its electrostatics. For rotations in two small fragments each rotation had a unique set of electrostatic properties. Therefore, it is possible to determine the geometry of the fragment from its electrostatics. However, this was taken further and the electrostatics examined without any geometry variables. Using only the electrostatics the geometry of the molecule could be traced through space with its
rotation in the form of loops made by the electrostatic properties. Deviations from a uniform loop suggested that the electrostatics were affected by the molecule enough so that the same rotation in another molecule was not in fact the closest match based on the electrostatics. Exclusion of geometry properties from the QID was tested by comparing the ranks of fragments from a standard QID search to the same fragment ranked only with electrostatics. The ranks correlated well between the two methods.

For terminals to be included in the QID further work is required to determine the extent to which the geometry is imprinted in the electrostatics. Further testing is required on a larger number of systems and testing against geometry parameters other than the dihedral angle as well, such as bond angle. Therefore, this would show that every aspect of the geometry is held in the electrostatics. If this is confirmed then the QID can be populated with terminal fragments and in combination with the work on rings the QID would be able to suggest replacements for any fragment of any molecule.

To improve the QID the alignment and scoring methods were addressed. The alignment method in the QID is unable to align collinear fragments. Because the alignment method rotates a molecule into a standard orientation using the atoms on either end of the molecule, collinear fragments were not aligned in the same orientation every time. This was fixed by introducing a principal axis alignment to replace the QID alignment of collinear fragments. The principal axis alignment was successful in aligning the collinear fragments and therefore suitable for use in the QID. The scoring method of the QID was tested against the program BROOD. Both packages queried the amide linker and the resulting fragments from both methods were examined. QID was able to return fragments that were also returned by BROOD. While the QID returned many fragments not returned by BROOD, it is unknown whether these fragments are stored in BROOD or not. Therefore, QID was deemed satisfactory when compared to BROOD, but unfortunately it is not possible to conclusively ascertain the QIDs ability against BROOD until the size of the data in the database is increased.

6.3 Predicting the \( pK_a \) values from an \textit{ab initio} bond length

The \( pK_a \) prediction method of high HCSs was explored to add a \( pK_a \) predictor to the QID. The HCS method introduced by Harding and Popelier correlates a single \textit{ab initio} bond length of a molecule (in the gas phase) to its experimental \( pK_a \) (in aqueous solution). Here the HCS method was applied to enolic functional groups to test the method against a set of biological molecules that contain this functional group, including
drugs and agrochemicals. The enols were selected because this functional group is known to be difficult to accurately predict a $pK_a$ value for. Correct conformational treatment of intramolecular hydrogen bonds improves $pK_a$ prediction by of allowing a HCS to merge with another, forming a single overarching HCS with more data points.

The correlations of the HCS remain similar after additional points were added to the HCS showing that once a HCS is derived it remains robust and successful at predicting $pK_a$ values. HCS correlations are based on very small variations in bond length, which were however found to be significant against the “computational noise” of standard geometry optimisation. A geometry optimisation truncated to about half the default number of iterations deteriorates the $pK_a$ prediction by only 0.5 log units.

The HCS method was also applied to guanidines. The C=N bond emerged as the ideal bond for use in the correlations. This bond corresponds to the site of protonation for the guanidine group. Through HCSs the effects of tautomerisation can be accounted for when predicting $pK_a$ values, as each tautomer separated itself from the others, into distinct subsets. Relations between the structures of the different tautomers are apparent through HCSs. In the case of amiloride, the HCS of the tautomeric form that produced the most accurate prediction, and therefore the ideal HCS, was also found to be the most stable tautomeric form by experiment. The HCSs themselves are comprised of smaller subsets, where the molecules within each subset are structurally related. In other words, the statistical subdivision of the subsets goes hand in hand with a chemical justification. HCSs have been shown to hold chemical information and predict $pK_a$ values accurately, using only a single bond length.

The HCS method requires equations to be derived for a larger number of functional groups. Currently only a small portion of functional groups have HCS equations. If the HCS method is to be incorporated into the tool then automation of the HCS and equation generation is invaluable. To automate the process of finding HCSs the most difficult step is identification of HCSs. This can be achieved through various statistical means were data points of a set are sampled and correlations calculated for any combination of data points. The strongest correlations will emerge as the HCSs but the method must be robust enough to find HCSs without creating HCSs from data points that do not belong in a HCS together. Constructing such an algorithm would greatly improve the speed at which HCS equations could be derived and therefore increase its applicability range so the $pK_a$ values of a diverse set of drugs can be predicted.
References

(20) Core Software Package 2010.


(43) Lamarche, O. Ph.D., Cardiff University, 2003.


(51) Atkins, P. W.; Friedman, R. S. Molecular Quantum Mechanics; Oxford University Press, 2005.


(54) Moller, C.; Plessset, M. S. Phys. Rev. 1934, 46, 618 LP


(76) Grimm, H. G. Naturwissen. 1929, 17, 557.
(78) IUPAC IUPAC Gold Book 2012.
(88) Leo, A. Chem. Rev. 1993, 93, 1281.
(100) Rupp, M.; Schneider, P.; Schneider, G. J. Comp. Chem. 2009, 30, 1754.
(101) WDI World Drug Index, Derwent Information 1996.
(132) Nielson, E. J.; McCammon, A. Protein Sc. 2003, 12, 313.
ADMET Predictor, http://www.simulations-plus.com/, Lancaster, California, USA.
Dennington, R.; Keith, T.; Millam, J. GaussView Version 5 2009, Shawnee Mission KS.
Appendix A

Architecture of the QID tool and its scripts

A.1.0 General overview

This appendix serves as a general overview to the structure and code of the QID and serves as a manual for future QID developers. A flow chart of a search process is given in figure 54.

The Database consists of three tables, fragment data, atom data, and bond data. Every fragment conformation is stored as a separate entry in each table, with a smiles string and conformation number used to identify the fragment. Bond properties of the fragment are stored in the bond data table, atom properties in the atom table and fragment properties in the fragment table. Customised queries allow for searches based on any combination of properties from either table, multiple queries from each table are allowed such as searching the atom table for two separate atoms. SQL is used to store and retrieve data from the database. Queries to the database are executed through SQL commands embedded in the Perl scripts.

When a user performs a search, the query returns a set of fragments that lie within the tolerances given for each property. For example, tolerances of property A are set at -1 to 1, and property B -10 to 10. (NB property A or B could be from any of the 3 tables, atom, bond or fragment, depending on the search) if the search fragment has a property A of 3 and B of 70. The query returns every fragment that has property A between 2 and 4, AND property B between 60 and 80. The set of fragments are ranked by their final score. The final score is determined by combining the weighted distances of each query property. Weights are assigned based on user selections.

Perl code is used to interact with the Database using Perl DBI module. Perl CGI module reads user inputs from the website such as smiles strings, search criteria, etc. within the Perl code, a print statement prints out an html page written in javascript, which creates the visual front end of the web tool pages. For example, for the Perl code below, (page_3.pl) the first half of the script in Perl initialises variables, such as user inputs, the second half is a print statement, printing out a html/javascript formatted page.
The blue portion of code is Perl, the red is html and javascript and bold statements are the Perl variables passed to the html/javascript. The print statement prints the html code, with the Perl variables, which is interpreted by the web browser and displayed as a web page.

A.1.2 The scripts of the QID

Home.html

The home page, with a brief outline of what the tool can be used for and (in very broad terms) how it is done. The only action here is to continue to the next page.

Page_2.html

The first step in a QID bioisostere search is to define the fragment that the user wishes to replace. This is a pure html code. The user enters a smiles string of the fragment and selects the capping group to represent the query fragments context in the molecule, this string is passed to Lookup_smiles.pl.

Lookupsmiles.pl
Searches the database and finds the matching smiles string of the query fragment with capping groups and displays the possible conformers of the fragment stored in the QID.

The smiles string of the fragment, including capping groups, is then converted into canonical SMILES using the Molconvert program. The search of the database is conducted through SQL command that matches the SMILES. All conformations of the query fragment are returned and displayed to the user.

The user selects which conformer to use as the query fragment, and the .xyz, and array of linker properties are passed to basic_query.pl

**Basic_query.pl**

The search properties and their importance are selected by the user. The search criteria offered are composite criteria made up of combinations of the raw *ab initio* based properties stored in the database. The criteria are sent to query.pl which searches the database and scores the hits. Search criteria are grouped by atom, bond and fragment. Information sent to query.pl takes the form of a list of properties, with their associated tolerances and importance.

Fragments that lie within the tolerance limits selected are returned. In basic_query.pl the tolerances are set and cannot be changed. Tolerances are +/- the value for the given property. For example the default tolerance for charge is +/- 0.6 au for the charge and only fragments that have a charge within 0.4 au of the query are returned. Hit fragments must lie within all tolerances that are specified. When the user selects a search criteria, a set of relevant tolerances are initialised. Continuing the example with charge, selecting the charge criteria turns on the charge, dipole and quadrupole tolerances and only fragments that with charge, dipole and quadruple properties within the tolerance for each respectively are returned as hits. In create_query.pl the upper and lower limits for the tolerance can be manually altered but requires the user to understand the meaning of the *ab initio* based properties.

Other than the basic search criteria, specific bond or atom properties may also be added to the list of query properties. The user defines the search type, bond or atom, and the query properties. The search type and query properties are passed to query.pl and incorporated into the database search. The search criteria listed are a combination of multiple properties from within the database.
When search criteria are selected the composite properties are passed to query.pl, so the format of the values mimics that of create_query.pl.

Create_query.pl

Is similar to basic_query.pl but differs in two ways.

1. A complete list of properties is given for search criteria, instead of the search criteria available in basic_query.pl which consist of combinations of properties, shown in table 9. This also holds for atom and bond properties. When the search type is selected, a complete list of properties is displayed. This gives much more flexibility over basic_query.pl.

2. The tolerances for each property are adjustable. Upper and lower limits can be independently altered.

Query.pl

Obtains a set of fragments that lie within ALL the tolerances of the search properties and each property of a linker is scored using the Euclidian distance and the importance. Individual property scores are combined to give an overall score for the linker. The linkers are then displayed in order of the increasing distance score.

The relevant properties from the query linker are obtained (only those properties that were selected for the search) and written into an SQL query as a list where x.property is a property from table x, where x = a for atom, b for bond or f for fragment.

An SQL query file is written using the search criteria (with their appropriate tolerances). A loop is used to sequentially add search criteria to the SQL query statement. Once the SQL statement is written it is submitted to the database and a set of fragments are
returned, that lie within the tolerances of all the search criteria stated in the SQL. Scores for each fragment are calculated as outlined in chapter 2.

A.1.3 SQL commands

The following SQL command selects the atom properties for a specified atom from within the atom data table using the fragment smiles string held in the array input.

```
SELECT a.SMILES, a.CONF_NO, a.ATOM_TYPE,
       a.ATOM_NO, a.X, a.Y, a.Z, a.CHARGE, a.DIP_X,
       a.DIP_Y, a.DIP_Z, a.DIPOLE, a.QUADRUPOLE,
       a.VOL, a.KE, a.I, a.V_ATOM, a.V_NE0_COR, a.V_EET_COR, a.FAA,
       a.V_MAX_003, a.V_MIN_003, a.MEAN_V_003, a.V_DEV_003,
       a.V_MEDIAN_003, a.N_PTS_003, a.V_MAX_005, a.V_MIN_005,
       a.MEAN_V_005, a.V_DEV_005, a.V_MEDIAN_005, a.N_PTS_005, a.L_INDEX,
       a.IE_MEAN, a.PKHB, a.PKHA
FROM $atom_table a
WHERE
       a.SMILES = '$input[$m][0]'
 AND a.CONF_NO = '$input[$m][1]'
 AND a.ATOM_NO = $input[$m][2]\n'';
```

This SQL command is taken straight from the perl code. Each variable corresponds to a value from the SQL tables. The prefix "a" refers the search to the atom property table. The variables passed to the SQL command are determined by the Perl scripts and the user inputs, such as the query fragment or atom within the fragment. In this SQL search the user only inputs the fragment and atom within the fragment of interest. The smiles of the fragment is held as $input[$m][0], the conformation of the fragment as $input[$m][1] and the atom number as $input[$m][2]. The SQL query returns the data from the SQL atom table for the atom within the selected conformation of the fragment with the correct smiles. The other elements from the fields given by the select statement are filled with the appropriate values for the atom of choice and passed further to the Perl script as an array.
When the data has been generated for each conformation of each fragment the data_entry.pl script extracts all the relevant data, including rotating the fragments into correct alignment, summing the moments etc, and prints the SQL command used to enter the correct data into the database. The following extract of code shows one such SQL command printed by the data_entr.pl script:

```sql
insert into ET_ATOM_DATA(SMILES, CONF_NO, ATOM_TYPE, ATOM_NO, FRAG_NO, X, Y, Z, DIP_X, DIP_Y, DIP_Z, CHARGE, DIPOLE, QUADRUPOLE, ERROR_L, VOL, KE, I, V_ATOM, V_NEO_COR, V_EET_COR, FAA, V_MAX_003, V_MIN_003, MEAN_V_003, V_DEV_003, V_MEDIAN_003, N_PTS_003, V_MAX_005, V_MIN_005, MEAN_V_005, V_DEV_005, V_MEDIAN_005, N_PTS_005, L_INDEX, Q20, Q21C, Q21S, Q22C, Q22S, IE_MIN, IE_MAX, IE_MEAN, EPNUC, PKHA, PKHB )
values ('[Ag]N=N[Au]', '1a', 'N', 1, 114, 4.58597190170634e-15, 9.48283053599964e-15, 2.78699468869353, -4.29616256499295e-05, 0.713260047990805, -0.0338028610016004, -3.39305296371079E-01, 7.14060593589535E-01, 1.19142019482727, 1.9792364733767E-03, 1.0120627998867E+02, 5.4717595974161E+01, 6.89662738614924E-01, -1.09619512827910E+02, -1.31064259648068E+02, 7.72740140849753E+01, 0.660979777013158, 10.571051, -113.261707, -49.634275, 33.667680, -46.790217, 836, -2.442067, -30.253374, -15.505130, 8.366203, -15.404832, 366, 5.95804, 0.540922921090977, 0.000307328281435266, 1.01621864030436, -0.306894050618981, -0.000310519895565504, 0.4908, 0.6860, 0.5695, -18.367781, -999, 0.27848623317)
```

This portion of SQL code inserts the data for atom N 1 from the 1a conformation of the fragment N=N where Ag and Au are the connection points for the linker into the atom data table. Each value corresponds to the SQL field it is assigned to. This SQL command is passed to the database and the data for the atom entered into the atom data table. If the data for this atom is to be later retrieved, the previous select statement would be written and $input[$m][0] would be [Ag]N=N[Au] for the smiles string of the fragment, $input[$m][1] would be the conformation number 1a, and $input[$m][2] would be 1 for the atom number. The above values would then be returned in their corresponding fields and stored as an array for the rest of the script.
Figure 54: General architecture of the QID webtool, where button icons in the figure show the exact text displayed for the button in the QID that performs the function.
Appendix B

Rotation of atomic multipole moments

The multipole moments are calculated, and expressed, with respect to an axis system defined for a molecule by MORPHY01. For comparison of the moments between molecules, the moments must be rotated into a global axis system defined for all molecules. The multipole moments calculated by MORPHY01 are expressed as real spherical harmonics. A method for rotating real spherical harmonics is outlined by Su et al. For the uses in this work, the moments were converted to their complex form, rotated using a transformation similar to a Wigner matrix, then converted back into their real values.

The complex spherical harmonics \( Y_l^m \) rotate as

\[
Y_l^m (\theta', \varphi') = \sum_{m'=-l}^{l} Y_l^{m'} (\theta, \varphi) D_l^{(1)} (\alpha, \beta, \gamma)
\]  

(A1)

where \( D_l^{(1)} (\alpha, \beta, \gamma) \) is the \((2l+1) \times (2l+1)\) matrix representation of the rotation about Euler angles \( \alpha, \beta, \gamma \), with each element of the rotation matrix defined as

\[
D_l^{(1)} (\alpha, \beta, \gamma) = \exp(-im'\alpha)d_l^{(1)} (\beta) \exp(-im\gamma)
\]  

(A2)

The elements of \( d_l^{(1)} (\beta) \) are calculated from

\[
d_l^{(1)} (\beta) = \left[ (l + m')!(l - m')!/(l + m)!/(l - m)! \right]^{1/2} \\
\times (-1)^{m-m'} \sum_k (-1)^k \binom{l + m}{k} \\
\times \left( \frac{l - m}{l - m' - k} \right) \cos(\beta/2)^{2k-m+m'} \\
\times \sin(\beta/2)^{2l-k-m'}
\]  

(A3)

where \( k \) is defined by

\[
\max(0, m - m') \leq k \leq \min(l - m', l + m)
\]  

(A4)

and

\[
\binom{a}{b} = a!/(a - b)!b!
\]  

(A5)
Converting the real forms of the MORPHY01 moments into their complex form is described by Su and can be done using

\[ y_{l0} (\theta, \varphi) = Y_{l0}^0 (\theta, \varphi) \]  \hspace{1cm} (A6)

\[ y_{lm}^+ (\theta, \varphi) = \left[ (-1)^m Y_{l}^m (\theta, \varphi) + Y_{l}^{-m} (\theta, \varphi) \right] / \sqrt{2} \]  \hspace{1cm} (A7)

\[ y_{lm}^- (\theta, \varphi) = \left[ (-1)^m Y_{l}^m (\theta, \varphi) - Y_{l}^{-m} (\theta, \varphi) \right] / \sqrt{2}i \]  \hspace{1cm} (A8)

Equation 42 is rearranged for the reverse of the transformation, the complex form to the real values, giving

\[ Y_{l}^m (\theta, \varphi) = \left[ \sqrt{2} y_{lm}^- (\theta, \varphi)i + y_{lm}^+ (\theta, \varphi) \right] / (-1)^m \]  \hspace{1cm} (A9)

The complex term \( Y_{l}^m (\theta, \varphi) \) in Equation 41 is then replaced by equation 43 to obtain \( Y_{l}^{-m} \) purely in terms of the real spherical harmonics

\[ Y_{l}^{-m} (\theta, \varphi) = \left( y_{lm}^+ (\theta, \varphi) - y_{lm}^- (\theta, \varphi)i \right) / \sqrt{2} \]  \hspace{1cm} (A10)

The corresponding expression for \( Y_{l}^{-m} \) in terms of the real spherical harmonics is obtained through similar re-arrangement. The resulting expressions to convert all real spherical harmonics to their complex equivalents are shown as

\[ Y_{l0}^0 (\theta, \varphi) = y_{l0} (\theta, \varphi) \]  \hspace{1cm} (A11)

\[ Y_{l}^m (\theta, \varphi) = \frac{1}{\sqrt{2}} (-1)^m (y_{lm}^+ (\theta, \varphi) + y_{lm}^- (\theta, \varphi)i) \]  \hspace{1cm} (A12)

\[ Y_{l}^{-m} (\theta, \varphi) = \frac{1}{\sqrt{2}} (y_{lm}^+ (\theta, \varphi) - y_{lm}^- (\theta, \varphi)i) \]  \hspace{1cm} (A13)

Multipole moments can then be rotated about Euler angles \( \alpha, \beta, \gamma \) into the global axis frame so meaningful comparisons can be made.
## Appendix C

## Data tables

A .3.0 pK$_a$ iteration calculation tables for three tetracycline molecules

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Tetracycline 1 Bond length</th>
<th>Predicted pK$_a$</th>
<th>Error less than 0.5 pK$_a$ units</th>
<th>1 pK$_a$ unit</th>
<th>2 pK$_a$ units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.36927</td>
<td>-0.1824012</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>1.38228</td>
<td>-2.0501168</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>1.36598</td>
<td>0.2899112</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>1.35327</td>
<td>2.1145588</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>1.35327</td>
<td>2.1145588</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>1.35034</td>
<td>2.5351896</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>1.36384</td>
<td>0.5971296</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>1.37123</td>
<td>-0.4637788</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>1.36694</td>
<td>0.1520936</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>1.36388</td>
<td>0.5913872</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>1.35561</td>
<td>1.7786284</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>1.347</td>
<td>3.01468</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>1.348</td>
<td>2.87112</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>1.34738</td>
<td>2.9601272</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>1.34723</td>
<td>2.9816612</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>1.34688</td>
<td>3.0319072</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>1.34653</td>
<td>3.0821532</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>18</td>
<td>1.34623</td>
<td>3.1252212</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td>1.34482</td>
<td>3.3276408</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>20</td>
<td>1.34416</td>
<td>3.4223904</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>21</td>
<td>1.34468</td>
<td>3.3477392</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>22</td>
<td>1.34398</td>
<td>3.4482312</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>23</td>
<td>1.34372</td>
<td>3.4855568</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>24</td>
<td>1.34343</td>
<td>3.5271892</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>25</td>
<td>1.34307</td>
<td>3.5788708</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>26</td>
<td>1.34325</td>
<td>3.55303</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>27</td>
<td>1.34354</td>
<td>3.5113976</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>28</td>
<td>1.34365</td>
<td>3.495606</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>29</td>
<td>1.34369</td>
<td>3.4898636</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>30</td>
<td>1.34368</td>
<td>3.4912992</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>31</td>
<td>1.34373</td>
<td>3.4841212</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>32</td>
<td>1.34371</td>
<td>3.4869924</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>33</td>
<td>1.34371</td>
<td>3.4869924</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>34</td>
<td>1.34369</td>
<td>3.4898636</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Iteration</td>
<td>Bond length</td>
<td>Predicted pK&lt;sub&gt;a&lt;/sub&gt;</td>
<td>0.5 pK&lt;sub&gt;a&lt;/sub&gt; units</td>
<td>1 pK&lt;sub&gt;a&lt;/sub&gt; unit</td>
<td>2 pK&lt;sub&gt;a&lt;/sub&gt; units</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>---------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td>1.36977</td>
<td>-0.2541812</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>1.38321</td>
<td>-2.1836276</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>1.36723</td>
<td>0.1104612</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>1.35133</td>
<td>2.3930652</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>1.3459</td>
<td>3.172596</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>1.36273</td>
<td>0.7564812</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>1.36551</td>
<td>0.3573844</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>1.36642</td>
<td>0.2267448</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>1.36357</td>
<td>0.6358908</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>1.35319</td>
<td>2.1260436</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>1.34284</td>
<td>3.6118896</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>1.34264</td>
<td>3.6406016</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>1.34477</td>
<td>3.3348188</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>1.34546</td>
<td>3.2357624</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>1.3449</td>
<td>3.316156</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>1.34491</td>
<td>3.3147204</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>1.34516</td>
<td>3.2788304</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>18</td>
<td>1.34531</td>
<td>3.2572964</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td>1.34535</td>
<td>3.251554</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>20</td>
<td>1.34523</td>
<td>3.2687812</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>21</td>
<td>1.34522</td>
<td>3.2702168</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>22</td>
<td>1.34523</td>
<td>3.2687812</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Iteration</td>
<td>Bond length</td>
<td>Predicted pK_a</td>
<td>0.5 pK_a units</td>
<td>1 pK_a unit</td>
<td>2 pK_a units</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>1.36862</td>
<td>-0.0890872</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>1.36667</td>
<td>0.1908548</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>1.36656</td>
<td>0.2066464</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>1.36162</td>
<td>0.9158328</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>1.3615</td>
<td>0.93306</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>1.36362</td>
<td>0.6287128</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>1.36326</td>
<td>0.6803944</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>1.36039</td>
<td>1.0924116</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>1.35548</td>
<td>1.7972912</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>1.34841</td>
<td>2.8122604</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>1.34868</td>
<td>2.7734992</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>1.34758</td>
<td>2.9314152</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>1.34723</td>
<td>2.9816612</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>1.34734</td>
<td>2.9658696</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>1.34651</td>
<td>3.0850244</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>1.34582</td>
<td>3.1840808</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>1.34599</td>
<td>3.1596756</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>18</td>
<td>1.34582</td>
<td>3.1840808</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td>1.34585</td>
<td>3.179774</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>20</td>
<td>1.34587</td>
<td>3.1769028</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>21</td>
<td>1.34587</td>
<td>3.1769028</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>22</td>
<td>1.34587</td>
<td>3.1769028</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>23</td>
<td>1.34586</td>
<td>3.1783384</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>24</td>
<td>1.34587</td>
<td>3.1769028</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>