SMARTS Approach to Chemical Data Mining
and
Physicochemical Property Prediction

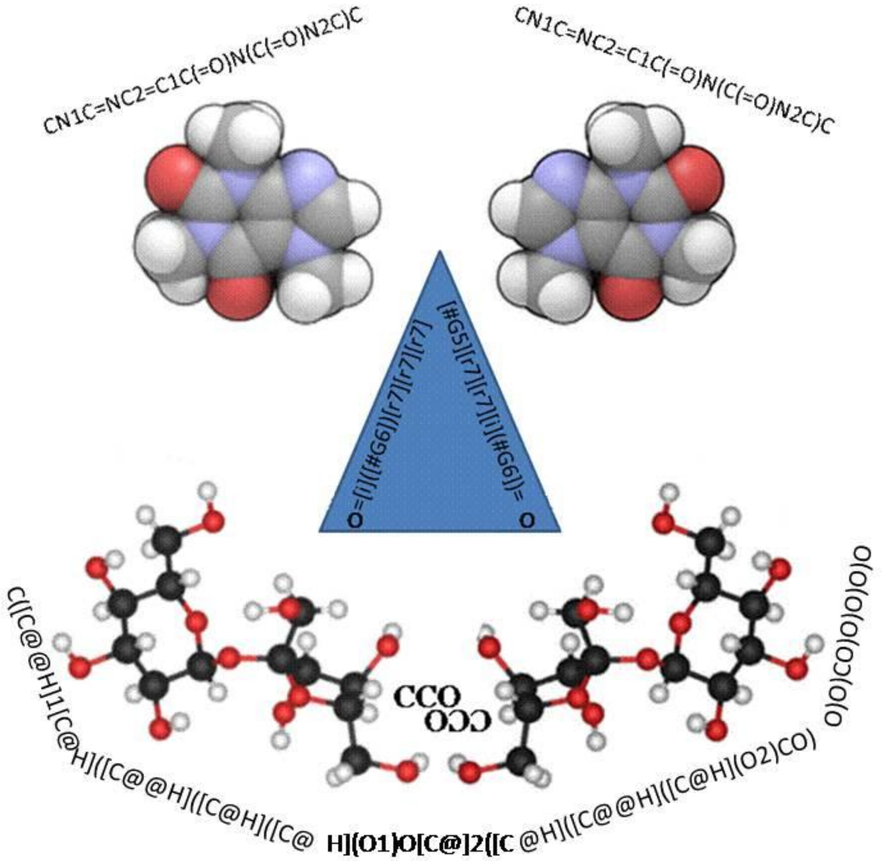
by

Adam C. Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in The University of Michigan
2009

Doctoral Committee:

Professor Gordon M. Crippen, Chair
Associate Professor Gustavo Rosania
Associate Professor Kerby A. Shedden
Assistant Professor Oleg V. Tsodikov


CN1C=NC2=C1C(=O)N(C(=O)N2C)C
CN1C=NC2=C1C(=O)N(C(=O)N2C)C
O=[i]([#G6])[r7][r7][r7]
[#G5][r7][r7][i]([#G6])=
O
CCO
OCC
C([C@@H]1[C@H]([C@@H]([C@H]([C@
H](O1)O[C@]2([C
@H]([C@@H]([C@H](O2)CO)
O)O)CO)O)O)O)O

Milla – now, it is your turn… All my love, Daddy!

Vanessa – focus returns home… My heart is yours forever!

Mom – I understand patience… You are the model!

Dad – the journey is now finished… Hunting we will go!

Acknowledgements

I would like to thank three MOE vectors of people. First, my advisors: [‘Dr. Gordon M. Crippen’, ‘Dr. Gus Rosania’, ‘Dr. Kerby Shedden’, and ‘Dr. Oleg Tsodikov’]. Dr. Crippen, you are the most wonderful mentor I could have imagined. Patient, understanding, and providing guidance at every turn, you have allowed me to travel my own path independently with reassurance and your steadfast support, cultured stories, grounded logic, and ever so slightly twisted dry sense of humor. [A^2 Friends, !A^2 Friends], I may not be in town, but I will always make time for you. [‘Vanessa’, ’Milla’, ‘Mom’, ‘Dad’], you are my foundation, my reason for being, and the true loves of my life. I revel in your ability to support my endeavors and am finally approaching the pivotal moment where the favors can be continually returned without reserve. I am truly a better person for knowing you all!

Preface

In *The Hitchhiker's Guide to the Galaxy*, the English humorist and science fiction novelist Douglas Adams wrote, "Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the chemist's, but that's just peanuts to space." Here, when Mr. Adams refers to "the chemist's", he is actually referring to the minute geometric distance between the reader and the pharmacy or drugstore (chemist is a common term for pharmacist in the UK and other countries like Australia and New Zealand), however the statement itself is accidentally ironic. When one thinks of chemists, the image comes to mind of a wild-eyed soot streaked sinister cartoonish figure with hair blown back dressed in a white jacket wearing eye-protection and mixing colorful liquids in assorted glassware precariously situated over an open flame. The chemists' reaction-based environment is limited to the chemicals at hand. However, the witty Doug had no concept of cheminformatics and the even more perplexingly large concept of chemical space(s) that the computational chemist deals with on a daily basis.

Chemical space refers to the space spanned by all the energetically stable stoichiometric combinations of atomic nuclei and electrons in molecules, including their isomers. It has been estimated that if one considers only molecules containing 30 or fewer C, N, O, and S atoms, that the number of potential molecules is 10^{60}. Furthermore, there are well over 20^{100} conceivable protein sequences up to 100 amino acids in length. In fact, summing the mass of the individual small molecules would likely exceed the theoretical mass of the universe. Therefore, it is physically impossible to synthesize this list within 10’s of orders of magnitude. The question then arises, “how does one realize the potentials for data mining the vastness of chemical space or at least begin to explore a small portion for a specific purpose, such as drug design?”

One answer involves the use of Structure Activity Relationships (SARs), allowing the medicinal chemist to investigate how small modifications to the scaffold of a known drug can affect its overall activity, selectivity, and specificity. SAR research is based upon the principle of “like produces like,” or molecules with similar structure are expected to behave similarly. While it is well known that even minute changes can render compounds inactive, the principle is commonly applied in drug development in order to modify medications so as to optimize efficacy, minimize side-effects, and establish new patents. Albuterol is an example such a medication. It is a racemic mixture, acting as a β_2-adrenergic receptor agonist, used to relieve bronchospasms in conditions such as asthma and chronic obstructive pulmonary disease. Marketed as levalbuterol, the r-enamtiomer is the active component; the l-enantiomer is completely inactive. While the first β_2-adrenergic receptor agonist was likely ephedrine extracted from Ma Huang or *Ephedra sinica*, having its roots in Chinese medicine, modern civilization accepted theophylline as the first β_2-adrenergic receptor agonist. Theophylline and caffeine, both members of the xanthine family, bear a strong structural resemblance. Understanding this, it is obvious how caffeine, most commonly administered in a hot cup of coffee, found its way into home remedy books for asthma attacks and shortness of breath. “Like produces like.” There is also a significant structural similarity between ephedrine, isoproterenol, metaproterenol, albuterol, and epinephrine (also the r-enantiomer), all of which can be used to treat the acute onset of asthma. “Like produces like.” Hence SARs are a favored approach in medicinal chemistry.

Unfortunately, SARs are limited for the very same reason that makes them useful: the scaffold or backbone of the chemical entity is constant. “Like produces like” can still be applied in a more abstract manner that involves scaffold-hopping based on the similarity between molecules in chemical property space. A chemical property space is defined by descriptors, typically measurable physicochemical properties: log*P*, molecular weight, polar surface area, number of hydrogen bond acceptors/donors, etc. Thereby, a computational chemist is able to map a library of known molecules into some *n*-dimensional hyperspace and localize regions of activity based on the presence of known drug molecules. “Like produces like” says that nearby molecules should also exhibit similar biological activity, pharmacokinetics, and pharmacodynamics. The drawbacks for this method are: (1) selecting the optimal set of orthogonal physicochemical properties, which may not be the same for each class of drug molecules, (2) the experimental data required may not be available for compounds on hand in the laboratory, much less for theoretical (*in silico*) compounds, and (3) the process of inferring physicochemical property from structure is unidirectional, that is, it is not currently possible to predict novel structures based on known physicochemical properties. It is not possible to reverse engineer chemical structures based on a set of physicochemical property parameters.

The goal of this work centers on developing a solution to the second issue.

Table of Contents

Dedication ... ii

Acknowledgements ... iii

Preface ... iv

List of Figures ... x

List of Tables ... xiii

Abstract ... xv

Chapter 1. Introduction

- 1.1 Time and Money ... 1
- 1.2 The ABC's of Chemical Data Mining: SMILES and SMARTS ... 2
- 1.3 What is Data Mining and How is It Used to Answer Questions? (An Example) ... 4
- 1.4 Overview ... 7
- 1.5 References ... 9

Chapter 2. Predicting pK_a

- 2.1 Introduction ... 10
- 2.2 Proteins ... 12
  - 2.2.1 Experimental Data ... 12
  - 2.2.2 Predictive Methods ... 13
    - 2.2.2.1 The Null Model ... 13
    - 2.2.2.2 Electrostatic Models ... 16
    - 2.2.2.3 Empirical Models ... 28
- 2.3 Small Molecules ... 31
  - 2.3.1 Experimental Data ... 32
    - 2.3.1.1 Data Curation ... 32
    - 2.3.1.2 Experimental Methods ... 34
  - 2.3.2 Predictive Methods ... 40
    - 2.3.2.1 Linear Free Energy Relationships ... 40

2.3.2.2 Quantitative Structure-Property Relationships 42
2.3.2.3 Quantum Mechanical and Continuum Electrostatic Methods 48
2.3.2.4 Artificial Neural Networks 52
2.3.2.5 Database Methods..53
2.3.2.6 Decision Trees 56
2.4 Protein-Ligand Complexes 57
2.5 Statistics and Benchmarking..58
2.6 Conclusions..63
2.7 References..68

Chapter 3. Chemical Data Mining of the NCI Human Tumor Cell Line Database
3.1 Introduction..85
3.2 Principles and Practices of Knowledge Discovery in Databases..........................88
3.3 Data Cleaning, Integration, Selection, and Transformation 93
3.3.1 Characterization of Chemical Compounds 93
3.3.2 Characterization of the Cell Line Screening Growth Inhibition Values.......96
3.3.3 Characterization of the Gene Expression Results..98
3.3.4 Predicting Missing Activity Values 99
3.4 Data Mining 101
3.4.1 Relating Dictionary-Based Structural Keys to Cellular Screening Activities...101
3.4.2 Predictive Models of Activity..104
3.4.3 Relating Freely Generated SMARTS Structures to Cellular Screening Activities...110
3.5 Conclusions and Future Directions..120
3.6 References..121

Chapter 4. Data Mining the NCI60 to Predict Generalized Cytotoxicity
4.1 Introduction..126
4.2 Methods..130
4.2.1 Data Gathering...130
4.2.2 Data Analysis...130
4.2.3 Imputing Missing Data 135
4.2.4 Data Mining with PCA 135
4.2.5 Predictive Binary Decision Trees 137
4.2.6 Prediction Using Least Squares Fit..139
4.3 Results and Discussion 140
4.4 Conclusions..154
4.5 References..157

Chapter 5. pK_a Prediction of Monoprotic Small Molecules the SMARTS Way
5.1 Introduction..159
5.2 Methods..162
5.2.1 Preliminary Studies..162
5.2.2 Training and Validating the Final Model...176
5.2.3 Cross Validation and pK_a Shuffling Studies. 178

5.3 Results and Discussion 179
5.4 Conclusions 191
5.5 References 194

Chapter 6. Future Directions: Removing the Human Element from Physicochemical Property Predictions
6.1 Introduction 198
6.2 Methods 203
6.2.1 SMARTS and SMILES 203
6.2.2 Data Selection 204
6.2.2 Training 205
6.2.3.1 Comprehensive Fragmentation to Obtain Chemical Descriptors 205
6.2.3.2 Screen Training Set Molecules to Eliminate Redundancies 207
6.2.3.3 Identify Predictive SMARTS String Descriptors 207
6.2.3.4 Refine and Expand Predictive SMARTS String Descriptors 209
6.2.3.5 Grow the Regression Tree 210
6.2.3.6 Fine Tune the Regression Tree by Generalizing SMARTS String Descriptors 211
6.2.3.7 Test for Overfitting Using Dependent Variable Randomizations 212
6.2.3.8 Validation: Test on External Data 212
6.3 Conclusions 212
6.4 References 215

Chapter 7. Conclusion 217

List of Figures

Figure 2.1. Simulation of the null model used to predict the pK_a of 18 residues (9 Asp, 7 Glu, and 2 Lys). 60

Figure 3.1. Comparative distribution of various properties for the compounds in the MRTD set (first column), tumor cell line set (second column), and two PubChem subsets (third and fourth columns). 95

Figure 3.2. Distribution of Tanimoto similarity values (*x*-axis) between compounds in the MRTD set and the most similar compound for each in the tumor cell line set 96

Figure 3.3. Fraction of the compounds with missing data for each of the 60 cell lines ...97

Figure 3.4. Fraction of compounds with growth inhibition values of 4.0 for each of the 60 cell lines 98

Figure 3.5. Fraction of compounds showing activity in each of the 60 cell lines............98

Figure 3.6. Distribution of the microarray gene expression data across all the 60 cell lines (left) and for five randomly selected cell lines (right) 99

Figure 3.7. Difference in active structural ratio and overall structural ratio..................103

Figure 3.8. Example compounds which are active in all cell lines (top row) or inactive in all cell lines (bottom row) 104

Figure 3.9. Accuracy of the prediction based on various structural features: (a) structural features are ranked across all cell lines and (b) structure features are ranked over cell line 60....................106

Figure 3.10. A box and whisker summary of the prediction accuracy for the 60 random forest models developed for the NCI DTP cell lines 109

Figure 3.11. Algorithm workflow 112

Figure 4.1. The number of available log GI_{50} values for the 43474 substances with measurements in units of molarity was color coded according to the class of cancer cell lines. The cell lines within each class have been alphabetized and numbered 131

Figure 4.2. The mean log GI_{50} of the NCI60 substances is shown with a granularity of 1000 bins 132

Figure 4.3a. Substance log GI_{50} correlation for 40131 data points between a non-small cell lung and a CNS cancer cell lines having $R^2 = 0.86$ 133

Figure 4.3b. Substance log GI_{50} correlation for 35680 data points between two non-small cell lung cancer cell lines having $R^2 = 0.73$............134

Figure 4.4. The x-axis represents the 59 cancer cell line assays and has been color coded and organized as described in Figure 4.1. The y-axis represents the components of the eigenvectors for each respective principal component 141

Figure 4.5. (1) The fit line through 543 substances determined by PC2 (▪), which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (2) The fit line for 239 substances determined by PC2 (⊙), which are responsible for decreased cytotoxicity of the leukemia cell lines compared to the non-leukemia cancer cell lines 145

Figure 4.6a. This histogram depicts the 543 substances determined by PC2 which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines 145

Figure 4.6b. This histogram depicts the 239 substances determined by PC2 which are most responsible for decreased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (—) and (········) represent the mean log GI_{50} of the leukemia cell lines and all cell lines, respectively............146

Figure 4.7. PC9 specifically identifies the Leukemia_SR cancer cell line related eigenvector component. PC13 identifies NSC_Lung_Hop_92 and CSN_SNB_75........147

Figure 4.8. Comparison of molecules from the 45 substances having specific toxicity for the Leukemia_SR cancer cell line............148

Figure 4.9. Histogram of the highest S_{tan} derived by comparing pairs of substances using the MACCS fingerprint on the set of 45 substances specifically targeting the Leukemia_SR cancer cell line 149

Figure 4.10. Prediction results using a decision tree............151

Figure 4.11. Prediction results using the least squares fit and forward stepwise regression 154

Figure 5.1. Each pK_a distribution is based on 25 bins for the monoprotic compounds having one of the three ionizable atom types considered here: oxygen (—●—), nitrogen (—◆—) and sulfur (—▲—)............164

Figure 5.2. The scatter plot on the left depicts experimental vs. predicted pK_a values for 202 monoprotic alcohols from the training set using the fingerprint defined by SMARTS string subset A, having an $R^2 = 0.90$.....172

Figure 5.3. Workflow used to achieve the SMARTS pK_a decision tree.....176

Figure 5.4. Scatterplots of the experimental vs. predicted pK_a values for the 185 compounds test set with SMARTS pK_a, SPARC, MARVIN and the ACD/I-Labs Web service ACD/ pK_a 8.03 and ADME Boxes.....187

Figure 5.5. Scatterplot of the experimental vs. predicted pK_a values for 112 compounds outside the training and first test sets.....188

Figure 5.6. Close aromatic and aliphatic analogs with pK_a values (experimental: predicted) using the SMARTS pK_a method.....190

Figure 6.1. a. The XOR decision made at a single node. **b.** XOR processed by two separate decisions complexes the decision tree by increasing the number decisions made and doubling the number of decision sub-pathways from the parent node.....202

Figure 6.2. a. Ethanol and halogenated derivatives and their SMILES string representations. **b.** An example of a decision tree that can express an XOR relation with a single SMARTS string.....204

Figure 6.3. How to extract aromatic SMARTS strings in MOE.....207

List of Tables

Table 2.1. Proposed null models 14

Table 2.2. Software for pK_a calculations using FDPB methods 17

Table 2.3. Selection of protein pK_a predictors outperforming the null model 20

Table 2.4. pK_a data sources 34

Table 2.5. Beilstein database (DE.MET): dissociation exponent method category....................35

Table 2.6. Survey of pK_a prediction methods 45

Table 2.7. The predictive abilities of ten pK_a prediction utilities....................61

Table 3.1. Accuracy of the prediction using only fingerprint information 100

Table 3.2. Accuracy of prediction using fingerprint and cell line information....................100

Table 3.3. SMARTS and interpretation for the bits associated with global activity and inactivity 103

Table 3.4. Accuracy of the prediction based on various structure features 106

Table 3.5. SMARTS bond types 115

Table 3.6. Most common single atom SMARTS strings in the NCI/DTP data set....................115

Table 3.7. Description of results 116

Table 3.8. Some of the most significant SMARTS strings 116

Table 3.9. Scoring selective MACCS SMARTS strings....................118

Table 4.1. Comparative summary of recent cytotoxicity study results 150

Table 4.2. Decision tree results for cutoff log GI_{50} = –5.0....................151

Table 5.1. Example MOE SMARTS strings 167

Table 5.2. Hand selected SMARTS string fingerprint that uniquely identifies the alcohol training set........169

Table 5.3. SMARTS string refinement process........175

Table 5.4. Overfitting test: accepted model vs. 100 models with randomized pK_a data. 180

Table 5.5. Literature survey of pK_a calculators with benchmarking using SPARC, MARVIN, ACD/I-Labs v8.03 and ADME Boxes........183

Table 5.6. Evaluation of test set prediction errors: SMARTS pK_a vs. available on-line calculators........186

Table 5.7. pK_a prediction for 4-(benzyloxy)benzoic acid using the final model........189

Table 5.8. Analysis of molecules falling within IoPs........191

Table 6.1. Relationship between potential fragment count and database size........208

Abstract

The calculation of physicochemical and biological properties is essential in order to facilitate modern drug discovery. Chemical spaces dimensionalized by these descriptors have been used to scaffold-hop in order to discover new lead and drug-like molecules. Broadening the boundaries of structure based drug design, these molecules are expected to share the same physiological target and have similar efficacy, as do known drug molecules sharing the same region in chemical property space.

In the past few decades physicochemical and ADMET (absorption, distribution, metabolism, elimination, and toxicity) property predictors have been the subject of increased focus in academia and the pharmaceutical industry. Due to the ever increasing attention given to data mining and property predictions, we first discuss the sources of experimental pK_a values and current methodologies used for pK_a prediction in proteins and small molecules. Of particular concern is an analysis of the scope, statistical validity, overall accuracy, and predictive power of these methods. The expressed concerns are not limited to predicting pK_a, but apply to all empirical predictive methodologies.

In a bottom-up approach, we explored the influence of freely generated SMARTS string representations of molecular fragments on chelation and cytotoxicity. Later investigations, involving the derivation of predictive models, use stepwise regression to determine the optimal pool of SMARTS strings having the greatest influence over the property of interest. By applying a unique scoring system to sets of highly generalized SMARTS strings, we have constructed well balanced regression trees with predictive accuracy exceeding that of many published and commercially available models for cytotoxicity, pK_a, and aqueous solubility. The methodology is robust, extremely adaptable, and can handle any molecular dataset with experimental data. This story details our struggles of data gathering, curation, and the development of a machine learning methodology able to derive and validate highly accurate regression trees capable of extremely fast property predictions.

Regression trees created by our method are well suited to calculate descriptors for large *in silico* molecular libraries, facilitating data mining of chemical spaces in search of new lead molecules in drug discovery.

Chapter 1

Introduction

1.1 Time & Money

Two factors familiar to every industry both big and small are time and money. “Time is money,” and “money cannot buy back time.” This is the huge impediment looming over Big Pharma like Pfizer. Lipitor, the $12.4 billion per year cash cow is going off patent in 2011 and there is nothing ready to take its place.[1] According to statements on Yahoo Finance, Pfizer grossed approximately $48 billion during the last fiscal year.[2] Considering the 9 – 15 year gestational period to take a new chemical entity (NCE) to an FDA approved drug and that the price tag of research and development averages $800 million for NCEs and $200 million for non-NCEs,[3] it stands to reason that Pfizer stands to lose over 25% of its revenue for the indefinite future.

The two main reasons new drugs do not make it to market is that they fail to be efficacious or pass safety standards late in testing due to poor pharmacokinetics and pharmacodynamics: absorption, distribution, metabolism, excretion, and toxicity (ADMET). Traditionally, the most expensive and time consuming stages of the drug development pipeline are phase I–III trials where the unfavorable ADMET properties are identified. By 2004 standards, approximately one-third of the new drug failures, representing $8 billion in losses, were due the inability to accurately predict toxicity.[4]

Therefore, it is essential to explore every means possible to curtail the losses and accelerate the drug development process.

The fastest and substantially less expensive stages of drug development rely on computational research. Once configured, *in silico* screening, docking, and property prediction can be made into batchable processes, requiring minimal to no monitoring and running 24 hours a day, 7 days a week. As high-level object oriented programming and operating systems have become more and more resource dependent, one way to increase speed is to improve the quality and processing power of the hardware by purchasing more memory, faster processors, and relying on parallel processing techniques, such as distributed and grid computing, which supports massively parallel CPU capacity.[5] An example of grid computing is the Search for Extra-Terrestrial Intelligence (SETI) project.[6] The other way to increase computational speed is through programming efficiency, reducing redundancies, and relying on computer language(s) best suited to the task at hand. Even though the computational phases of the drug development pipeline tend to be the fastest, great emphasis is still placed on optimizing information throughput, improving physical resources, and more efficient programming, following the "time is money" principle.

1.2 The Abc's of Chemical Data Mining: SMILES and SMARTS

One of the early *in silico* optimizations was the derivation of a chemical notation capable of consolidating chemical nomenclatures, encoding two-dimensional molecular representations in simple line notation, along with the incorporation of a sister language that increased the efficiency of *in silico* high throughput screening (HTS). The notation and language are SMILES (Simplified Molecular Input Line Entry System)[7] and SMARTS (A Language for Describing Molecular Patterns).[8]

A SMILES string is a string of ascii characters, representing a molecular connectivity table without spatial coordinates. For compactness, hydrogens are implied except when dealing with stereocenters. Furthermore, while many SMILES strings can be used to represent the same molecule, a unique canonical form exists. Uniqueness is important because it allows for faster queries. For example, if one had a large molecular database, such as PubChem with 37,321,069 unique compounds,[9] and wanted to support exact structure searching without having to rely on an engine which parsed molecular connectivity tables for every molecule, indexing PubChem using canonical SMILES for each molecular entry would solve the problem. The solution is possible as the canonical SMILES representation of a molecule results in a unique ordering of atoms and notations. As SMILES strings are generic one-dimensional textual objects, the problem has been reduced to a simple text matching issue instead of a chemistry-specific multidimensional problem. The canonical SMILES for Lipitor is CC(C)C1=C(C(=C(N1CCC(CC(CC(=O)[O-])O)O)C2=CC=C(C=C2)F)C3=CC=CC=C3)C(=O)NC4=CC=CC=C4.CC(C)C1=C(C(=C(N1CCC(CC(CC(=O)[O-])O)O)C2=CC=C(C=C2)F)C3=CC=CC=C3)C(=O)NC4=CC=CC=C4.[Ca+2], where the periods represent breaks between non-covalently linked molecules. While the SMILES language does not halt the evolution of the International Union of Pure and Applied Chemistry (IUPAC) nomenclature, it does provide an important means for more efficient storage of chemical data, as SMILES strings are a robust notation for compressing two-dimensional molecular structural data, while maintaining structural chemistry.

In silico HTS data mining is supported through SMARTS strings, which are an extension of SMILES. If a string is a SMILES string, then it is also a SMARTS string. SMILES and SMARTS follow conditional logic, as the reverse of the previous statement is not true. Rather, if a string is a SMARTS string, it is *not* necessarily a SMILES string. SMARTS strings are typically used for substructure searching to identify molecules based on pattern matching, either as a singular string or as a group of SMARTS strings (molecular fingerprint). The set of SMARTS strings represented by the vector [‘c1ccccc1’, ‘F’, ‘[Ca+2]’, ‘[Sn]’] would return [1,1,1,0] when used to query Lipitor, as Lipitor contains a benzene ring, a fluorine, a calcium ion with charge +2, but no tin. The capabilities of SMARTS for substructure searches are limited only by the individual and the particular SMARTS implementation used. While Daylight[8] is the originator of SMILES and SMARTS, we have chosen to use the MOE representation, due to our lab’s close relationship with its creators, the Chemical Computing Group.[10] The use of SMARTS strings is the unifying theme for this work. Here we have extended the use of SMARTS beyond that of simple molecular screening, and have shown that the language can be used to reverse engineer measurable chemical properties from structural data.

1.3 What is Data Mining and How is It Used to Answer Questions? (An Example)

Data mining is the modern term which represents the process of searching for “a needle in a haystack.” Fortunately, it is not as physically laborious as digging through a haystack or nearly as painful as accidentally falling on one of the needle’s pointed ends. Here, we provide an example of how freely generated SMARTS strings and successive screens of a molecular database can be used to answer questions relating to medicinal chemistry. First, the question: do chelating substances induce toxicity by trapping and eliminating or transporting complexed metal ions to specific compartments within the body or all of the above?

In an attempt to answer this question, we considered data from the National Cancer Institute's Developmental Therapeutics Program's human tumor cell line data, the NCI60.[11] In every data mining effort the type, quantity, and quality of the available data must be taken into consideration. While high quantity and high quality data can lead to more robust, accurate data models, the type of data needs to be carefully considered. The NCI60 consists of cancer cell growth inhibition data (GI_{50}: 50% after 48 hours) for over 43000 substances assayed against 60 cancer cell lines. Structural and growth inhibition data was obtained from PubChem's Substance and BioAssay databases.[12] The NCI60 is considered one of the highest quality, most well curated, publically available data sets.

First, we investigated congeneric series of chelating compounds, where the only variants are (1) the presence of metal ions and (2) the metal ion type. A chelant is an organic structure that can form an ionic bond with one or more metal ions, such as EDTA or ethylenediaminetetraacetic acid, whereas a chelate is the metal-bound form. EDTA has many uses, but as it relates to medicinal chemistry, it is used in chelation therapy to treat heavy metal poisoning.[13] Structural screening efforts identified several chelant / chelate pairs and congeneric series of chelates in the NCI60, in which the organic chelant was complexed with a variety of metal ions: Cu, Zn, Pd, Mn, Sn, Au, Ag, etc. Unfortunately, based on the small amount of data, we were only able to draw limited conclusions regarding the toxicities of chelates and had no way of determining if the toxicity was due to the accumulation metal ions within some organ or subcellular compartment. We were able to show that chelants share similar toxicity profiles to their metal-complexed chelate counterparts and tend to have minimal fluctuations in mean log GI_{50}, regardless of the metal ion type. Notable exceptions were Pt and Sn which both exhibited greater toxicity in all cases. Fe, Mn, and Pd showed varied levels of toxicity based on the chelant. If we exclude the data for Pt and Sn, metal-complexed substances appear to be no more toxic than the organic substances from this data set.

A second and more robust attempt was to data mine the PubChem Compound database for a set of molecular motifs that were commonly found in chelating compounds. First, all data from the PubChem Compound database were downloaded as SDF files[14] and imported into a MOE database. Second, compounds having toxicity data (the NCI60) were excluded and set aside as a test set. The remainder of the compounds represented the training set. Third, SMARTS strings were generated starting from a single atom and grown to chains up to seven atoms in length. At each stage of growth the SMARTS strings were screened against the molecules in the training set and scored, as detailed in chapter 3. In this case, the score for each respective SMARTS string equaled the number of identified molecules containing metals divided by the number of identified organic molecules. SMARTS strings with scores at least an order of magnitude above the ratio of metal containing to non-metal containing compounds of the training set (score = 0.01) were saved for testing. Over 2000 SMARTS strings, representing organic substructures commonly found in metal containing substances, were identified.

The test set was divided into two groups: the first group (likely to contain chelates) was identified by at least one of the 2000 SMARTS strings and also included any unidentified substances containing metal ions, while the second group (likely to contain non-chelates) consisted of the organic substances that were not identified.

Comparing histograms for the two groups and ignoring substances containing Pt and Sn, no significant differences in the distribution, mean, median, or standard deviation of the mean log GI_{50} could be identified. Therefore, we obtained more evidence in support of our earlier observation that the organic portion of the chelate (the chelant) is more likely to be responsible for cytotoxicity than metal ions accumulating inside the cell.

While chemical data mining may not always provide an answer to the question at hand, the process can unearth other potentially significant facts leading to new discovery.

This material was presented at a poster session, *Cheminformatics: Chelation vs. Toxicity*, held at the 233rd American Chemical Society National Meeting in Chicago, IL on March 25-29, 2007.

1.4 Overview

Chapter 2 is a perspective on pK_a prediction for both small molecules and proteins. Experimental data, curation, methodologies, and benchmarking are discussed.

In chapter 3 we discuss a formal knowledge discovery approach to characterizing and data mining the NCI60. The results of some of the initial experiments from a cheminformatics perspective are reported, including methodology which has identified a large set of SMARTS strings likely to be found in cytotoxic molecules and a complementary set found in molecules that show no signs of cytotoxicity. These sets of SMARTS strings can be used as filters for future data mining experiments.

We continue exploring the NCI60 in chapter 4 by deriving predictive methodologies for cytotoxicity. First a completely *in silico* methodology using decision trees and the MACCS keys is discussed. Finally, using stepwise regression and least squares fit, we were able to identify 9 cancer cell lines which could be assayed to provide extremely accurate predictions for generalized cytotoxicity. Having log GI_{50} values at the maximum threshold, a considerable portion of the substances considered for NCI60 testing showed no sign of activity for all cancer cell lines. The proposed regression equation would save time and materials by using an initial screen, which would consider only 9 cancer cell line assays, instead of the entire NCI60.

Chapter 5 has received considerable attention from both the pharmaceutical industry and academia. Here, the methodology from the previous chapter is expanded using highly generalized and manually derived SMARTS strings to predict pK_a for small molecules. The publication has received interest from Roche, Novartis, the Chemical Computing Group, the Unilever Centre for Molecular Sciences Informatics, the Institute for Parallel Processing of the Bulgarian Academy of Sciences, as well as other international academics. This work has resulted in several requests for a review article on the subject of pK_a prediction, Chapter 2.

In Chapter 6 we discuss a series of machine learning techniques applicable to the derivation of regression trees for physicochemical property modeling. Here, a more objective way to create a set of SMARTS strings for these models is considered. Given a reliable and chemically diverse training set with well curated experimental data, the methodology is capable of deriving regression trees suitable for predicting any dependent variable. The prediction of aqueous solubility for small molecules is used as an example.

Chapter 7 summarizes what we have discovered, and we attempt to predict the future for regression trees as a machine learning technique in physicochemical and biological property predictions.

1.5 References

(1) Leckey, A. Lipitor's patent status a challenge for Pfizer. *Richmond Times-Dispatch*, Apr. 6, 2009. http://www.timesdispatch.com/rtd/business/local_other/article/ LECK06_20090403-231732/248829/ (Accessed on Jun 25 2009).

(2) Yahoo Finance: http://finance.yahoo.com/q/is?s=PFE (accessed Jun 28, 2008).

(3) Light, D. W. Misleading Congress about Drug Development. *J. Health Politics, Policy and Law* **2007**, *32*, 895–913.

(4) Acton, G. Toxicogenomics and Predictive Toxicology: Market and Business Outlook, 2004 http://www.the-infoshop.com/study/cd25153_toxicogenomics.html (accessed Feb 29, 2008).

(5) Bersitis, V. Fundamentals of Grid Computing, 2002. IBM Redbooks. http://www.redbooks.ibm.com/redpapers/pdfs/redp3613.pdf (accessed Jun 28, 2009).

(6) SETI@home http://setiathome.berkeley.edu/index.php (accessed Jun 28, 2009).

(7) Daylight Chemical Information Systems Inc. http://www.daylight.com/smiles/ (accessed Jun 25, 2009).

(8) Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Jun 25, 2009).

(9) NCBI: PubChem Project: National Center for Biotechnology Information, Bethesda, MD 2008. http://pubchem.ncbi.nlm.nih.gov/ (accessed Jun 25, 2009).

(10) *MOE: Molecular Operating Environment*, version 2007.0902: Chemical Computing Group; Montreal, Quebec, Canada 2007.

(11) DTP: Developmental Therapeutics Program NCI/NIH. http://dtp.nci.nih.gov/ (accessed Jan 8, 2008).

(12) NCBI: PubChem Project: National Center for Biotechnology Information, Bethesda, MD 2008. http://pubchem.ncbi.nlm.nih.gov/ (accessed Feb 1, 2008).

(13) Hawkins, E. B.; Ehrlich, S. D. http://www.umm.edu/altmed/articles/ethylenediaminetetraacetic-acid-000302.htm (Accessed on Jun 28 2009).

(14) Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 244–255.

Chapter 2

Predicting pK_a

"A man would do nothing if he waited until he could do it so well that no one would find fault with what he has done." John Henry Cardinal Newman

2.1 Introduction

One of the most important physicochemical properties of small molecules and macromolecules are the dissociation constants for any weakly acidic or basic groups, generally expressed as the pK_a of each group. This is a major factor in the pharmacokinetics of drugs and in the interactions of proteins with other molecules. For both the protein and small molecule cases, we survey the sources of experimental pK_a values and then focus on current methods for predicting them. Of particular concern is an analysis of the scope, statistical validity, and predictive power of methods, as well as their accuracy.

When a weak acid dissociates according to the schematic equation $HA \rightleftharpoons A^- + H^+$, the equilibrium constant is $K_a=[A^-][H^+]/[HA]$, which is conveniently rearranged into the Henderson-Hasselbach equation

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\frac{[A^-]}{[HA]} \qquad (1)$$

where pK_a = –$\log_{10}K_a$, analogous to the definition of pH. Of course, one can describe the protonation of a weak base in the same terms. When the weak acid/base is titrated against a strong base/acid, the titration curve is a plot of pH as a function of equivalents of added titrant, and the curve shows a characteristic inflection when pH = pK_a. Experimental determination of pK_a is straightforward as long as there is only one titratable group involved. When the molecule in question has n protonatable sites, there are 2^n microspecies (particular combinations of protonations at the n sites) to be considered and $2^n - 1$ independent micro-pK_as (equilibrium constants between two microspecies), so that at any given pH there is an equilibrium mixture of some of these microspecies at non-negligible concentrations. Particularly if some of the micro-pK_as have similar values, the titration curve may show only a few inflections corresponding to macro-pK_as between the macrospecies that are the predominant mixtures of microspecies. Ullman points out that a maximum of $n^2 – n + 1$ parameters can be extracted from a titration curve of all ionizable sites, and since $2^n - 1 > n^2 – n + 1$ for $n > 3$, it is not possible to obtain micro-pK_as for polyprotic acids with more than 3 independent ionizable sites without additional information or special assumptions.[1] Expressed in these terms, there is a concern in predicting the pK_as for all microspecies.

This review focuses on pK_a predictions for molecules in an aqueous environment with the typical pK_a range of interest lying between that of the hydronium ion (–1.74) and the hydroxide anion (15.74), as the pH of blood is generally regulated to the range of 7.35–7.45 and the pH range encountered in the normal human gastrointestinal tract is 1–8.[2,3] Most of the available experimental data were obtained at 25° C in aqueous solutions having ionic strength less than 0.1 M. Obviously 37° C would better match human *in vivo* conditions. Many of the predictive methods discussed herein could be re-parameterized for other environments, considering both temperature and solvent, as long as sufficient data were available for training and validating the model.

2.2 Proteins

Calculating the macro-pK_as for globular proteins may seem easy because there are only a few different kinds of protonation sites involved, particularly Asp, Glu, and His sidechains. The problem is non-trivial because particularly the somewhat buried acidic and basic groups have long been known to have pK_as that are sometimes substantially shifted from what is observed in oligopeptides. Worse yet, there are so many protonation sites on most soluble proteins that experimentally determining the macrospecies is challenging. Prediction methods are therefore developed on the basis of several sites each of rather few well-studied proteins, which is something of a concern for validating the methods. The standard prediction task takes as input not only the amino acid sequence, i.e. the covalent structure of the molecule, but also the experimentally determined three-dimensional structure typically from X-ray crystallography, which can be a problem when conformational flexibility is important.

2.2.1 Experimental Data

The favored method for determining the pK_as of individual ionization sites on proteins is by NMR.[4] It can also be used for small molecules in both aqueous and mixed solvents. It is applicable to proteins both in their folded and in their denatured states,[5] and thus is useful in determining the correct order of deionization. The chemical shift of an assigned proton near the ionization site (in terms of covalent bonds or even through space) varies with pH, so the chemical shift vs. pH curve is fitted to a simple model involving three adjustable parameters: the chemical shifts in the protonated and deprotonated states and the pK_a.[6] The main concern is that the chemical shift of a proton can also be influenced by other environmental factors, such as nearby solvent or other parts of the protein, or by multiple conformations interconverting rapidly on the NMR time scale. It is also not possible to accurately determine the pK_a of residues that are not fully titrated at low pH (for example Glu73, Asp93, and Asp101 of barnase), as no true baseline representing the protonated state can be established. Furthermore, experimentally determining the pK_as for coupled ionizable residues is difficult, as fitting ideal titration curves to NMR chemical shift data of these residues leads to poorly resolved pK_as (for example Asp54, Asp75 and Asp86 of barnase).[7] Sometimes these questions can be resolved by difference titrations where the sequence of the protein is changed to eliminate sources of confusion.[8]

Currently, the Protein pK_a Database (PPD) exists as a free data source, providing over 1400 experimental data points taken from literature for the ionizable amino acid sidechains in proteins, as well as N and C termini. The vast majority of the available measurements are for Asp, Glu, His, and Lys. Very little data exists for Arg, as titrations at high pH tend to denature proteins.[9]

2.2.2 Predictive Methods

2.2.2.1 The Null Model

The simplest method is the trivial null model, where the experimental pK_a of the amino acid sidechain in some oligopeptide is taken as the predicted value for all amino acid residues of the same type. Several variations on this theme are shown in Table 2.1. Many earlier works unintentionally demonstrate, or at least mention, that the performance of the null model can be difficult to beat. This is especially true when the proteins being considered have a preponderance of ionizable residues exhibiting small pK_a shifts. In these instances, the null model can be expected to have a root mean square error (RMSE) less than or equal to 1.0.[10,11,12] In protein pK_a prediction, outperforming the null model is essential, as the residues showing the most significant pK_a shifts are often the most interesting, for example buried residues, residues participating in salt bridges, or residues found in enzyme active sites.[13] The overall success of the null model is mainly due to the available data, which are dominated by surface residues and other residues that do not participate in strong intramolecular interactions.[11] Nonetheless, null model values are an important starting point for most prediction methods.

Table 2.1. Proposed null models

Group	1943[a]	1967[b]	1973-4[c]	1978[d]	1993[e]	2006[f]	2007[g]	2009[h]
α-Carboxyl	3.0–3.2	3.8	3.3	—	3.5–4.3	3.67	—	—
Asp	3.0–4.7	4.0	3.91	3.9	3.9–4.0	3.67	3.47	3.49
Glu	4.4	4.4	4.145	4.2	4.3–4.5	4.25	4.16	4.39
His	5.6–7.0	6.3	(6.8)	6.9	6.0–7.0	6.54	6.30	6.6
α-Amino	7.6–8.4	7.5	8.1	—	6.8–8.0	8.00	—	—
Cys	9.1–10.8	9.5	—	—	9.0–9.5	8.55	4.67	—
Tyr	9.8–10.4	9.6	(10.0)	10.2	10.0–10.3	9.84	9.90	—
Lys	9.4–10.6	10.4	10.47	11.0	10.4–11.1	10.40	10.04	9.78
Arg	11.6–12.6	12.0	—	—	12.0	—	—	—
RMSE[i]	1.44	1.54	1.48	1.56	1.46	1.46	1.43	1.42

[a-f]Data taken from reference 18. [a,b]Determined using various model compounds at 25° C.[14,15] [c]Determined with Gly-Gly-X-Gly-Gly pentapeptides by ^{13}C NMR with unblocked termini,[16,17] while values in parenthesis were taken as reported in reference 18. [d]Determined in Gly-Gly-X-Ala tetrapeptides with unblocked termini by ^{13}C NMR at 35°C.[19] [e]Data taken from Creighton.[20] [f]Determined in Ala-Ala-X-Ala-Ala pentapeptides with blocked termini using potentiometry.[18] [g]Mean values taken from 475 different sites from 73 proteins.[21] [h]Values for each residue type were obtained by minimizing the RMSE in a benchmark set of 80 residues. [i]In all cases where a range is indicated, the midpoint was taken and used to compute the RMSE for a benchmark set of 80 residues.

In an attempt to determine which set of null values represents a good starting point, we considered a benchmark set of 80 interesting residues from 30 different proteins used to compare some known macromolecular pK_a prediction utilities.[11] For each of four residue types (Asp, Glu, His, and Lys), the data set consists of 20 pK_a measurements, 10 having pK_a shift less than 1.0 and the other 10 having pK_a shift greater than or equal to 1.0. The dataset is diverse, as it consists of a more balanced collection of residues having varied solvent exposure with over half exhibiting large pK_a shifts, and it includes residues found in active sites, as well as structurally important regions.[22] There are arguments both for[12,21] and against[23,24] the correlation between solvent exposure and pK_a shift. However, in this case focus is placed on the diversity of the local environment. RMSE was calculated for each of the null models applied to the benchmark data set, as shown in Table 2.1. Of course our 2009 null model performs the best, since it was trained by a least squares fit to the very benchmark dataset. It is interesting to note that the most commonly used or traditional set of intrinsic values comes from the 1967 set, compiled by Nozaki and Tanford.[15] The traditional set is outperformed by almost every other proposed set of null values. The best and least controversial datasets which can be used as intrinsic values are those in the columns labeled 1943, 1993, and 2006.[14,18,20] All proposed pK_a values from these columns are based on experimental measurements and come close to matching the results from the optimized values as well as the values acquired by taking the mean value for each residue type over a large set of curated experimental values from 73 proteins.[21]

When developing any predictive model, care should be given to the separation of training and test data. That is, any protein residues used to parameterize a model should not be used for validation purposes. In this regard, some have taken a cross-validation approach to model optimization, commonly used in cases where there is a lack of available data. In the cross-validation approach, the null values can be set to the mean pK_a values, or some other statistical variant thereof, for each residue type in the training set. In one example, the pK_a values for the ionizable residues in each individual protein of a set of 27 proteins were predicted using the data from the other 26 proteins. The resulting RMSE of 0.853 was a significant improvement over the traditional null model's RMSE of 1.069. The study also showed that the results of the optimized null model surpassed those of a Poisson–Boltzmann equation based approach for a dataset of 122 ionizable residues of five different types that had been previously evaluated by the same group.[12] Clearly, optimizing the intrinsic null values is simple and may prove beneficial in more elaborate models.

2.2.2.2 Electrostatic Models

In order to account for deviations from the null model due to the spatial arrangement of the rest of the protein and the general solvent, one must estimate the free energy difference between the protonated and deprotonated states of the ionizable group in question. Most protein pK_a prediction methods are based on solving the linearized Poisson-Boltzmann equation using atomic partial charges from a molecular modeling force field. Typically, the three terms considered in calculations using finite difference Poisson-Boltzmann (FDPB) methods are the Born solvation energy, the energy due to Coulomb interactions of the ionizable group in question with fixed partial charges of the protein atoms, and pairwise Coulomb interactions of it with other titratable sites of the protein. For those interested in the underlying parameterization of an FDPB model, we direct the reader to a work by Fitch and García-Moreno[25] where the basic protocols are described for implementing the University of Houston Brownian Dynamics (UHBD) software for pK_a calculations developed by McCammon and colleagues.[26] This work serves as an excellent review of FDPB methodology, provides a list of downloadable FDPB software shown in Table 2.2, and discusses model parameterization including the selection of the most appropriate dielectric constant for electrostatic calculations.[25] Bashford provided a comprehensive review covering the major macroscopic electrostatic models and approximations that are used to calculate the relative energies of protonation states and the pH-titration properties of ionizable groups in proteins as well as their applications to small molecules.[27] The methods discussed are rooted in solving the Poisson-Boltzmann equation, which has been thoroughly discussed in literature.[28,29,30]

Table 2.2. Software for pK_a calculations using FDPB methods[25]

Software	URL
Delphi	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi
MEAD	www.scripps.edu/mb/bashford/
PEP	casegroup.rutgers.edu/
UHBD	mccammon.ucsd.edu/uhbd.html
ZAP TK	www.eyesopen.com/products/toolkits/modeling-toolkits.html
APBS	apbs.sourceforge.net/
H++	biophysics.cs.vt.edu/H++/
HYBRID	gilsonlab.umbi.umd.edu/software1a.html
KARLSBERG KARLSBERG+	agknapp.chemie.fu-berlin.de/agknapp/
MCCE	www.sci.ccny.cuny.edu/~mcce/
PCE webserver	bioserv.rpbs.jussieu.fr/Help/PCE.html
WHAT IF	swift.cmbi.kun.nl/whatif/
Wade lab scripts	projects.villa-bosch.de/mcm/software/pka

When using FDPB models to calculate pK_a, two other problems need to be considered: electrostatic and thermodynamic. The electrostatic problem involves understanding the dielectric properties of the different phases in a protein-water system, thereby making it difficult to accurately calculate pK_a for all ionizable sites of a protein without considering the environment of the individual ionizable sites. For instance, buried ionizable residues can experience substantial pK_a shifts when compared to the null value for the respective residue type. It has been shown that regions with low dielectric medium provide an unfavorable environment for inducing a net charge.[31,32,33] Therefore, it would be expected that the pK_a of an aspartic acid residue located in a region of moderately low polarity would be shifted higher due to the absence of strong ionic and hydrogen bonding interactions found in an aqueous environment. On the other hand, buried environments exist such that a carboxylate group is favored over the unionized form, such as those in close proximity to another polar or positively charged group which lowers the pK_a relative to the null value. This is common in protein functional regions such as active sites.[34,35,36] Finally, the dielectric constant is inversely proportional to the electrostatic energy and can thereby significantly affect the calculated energies in continuum electrostatic calculations as well as the pK_a shift of ionizable residues. Even though the definition and parameterization of the dielectric constant is model dependent,[63] it is considered the most important adjustable parameter when performing continuum electrostatic calculations.[25] The statistical thermodynamic problem involves making FDPB calculations to solve the state of ionization and electrostatic energy for each ionizable site. The issue is purely combinatorial and limits the application of FDPB models to smaller proteins. The number of calculations increases exponentially with the number of ionizable sites in a protein. As discussed earlier, a protein with N ionizable residues can have up to 2^N microstates. The practical computational limit is thought to be 30 ionizable groups.[25] Treating larger proteins undoubtedly would require approximations.

Still other problems exist, such as protein flexibility and partial charge parameterization. One of the most common and potentially problematic simplifications used when applying Poisson-Boltzmann calculations is the assumption that protein structures are rigid and identical to the crystal structure. It is known that ions can co-crystallize with a protein affecting the fine details of the protein's surface structure.[13] Furthermore, this assumption places limitations on the calculations of changes in free energy when the state of an ionizable group changes.[37] Ionizable sidechains of proteins in crystals grown at fixed pH have a fixed charge. In solution at fixed pH proteins are flexible to various degrees and constantly undergo conformational changes, as seen by NMR. During titration, the pH changes and alters the ionization state of the protein in solution, which may lead to different conformations. Static models based solely on the crystal structure coordinates are therefore not likely to be appropriate at all pHs.[38,39] On the other hand, considering conformational flexibility requires computationally expensive methods such as Monte Carlo sampling. Now, instead of considering 2^N ionizable states in the case of rigid structures, one could consider up to $(2M)^N L^K$ possible states where each of N ionizable residues has M potential conformations and K nonionizable groups have L conformations.[40] This is further complicated by the use of molecular modeling force fields, which may lack proper parameterization or provide less than adequate partial charge assignments. Alternatively, a Quantum Mechanical (QM) or mixed Quantum Mechanical Molecular Modeling (QM/MM) approach would rely on significantly fewer empirical parameters, but would be far more demanding of computational resources. In spite of the many challenges, Table 2.3 shows the continual improvement of FDPB and other methods for the prediction of protein pK_a.

Table 2.3. Selection of protein pK_a predictors outperforming the null model

Ref	Author	Year	Method[a]	# of Para	Training	RMSE	# of Proteins	Valid[b]	Residues[c]
10	Antosiewicz	1994	FDPB		60	0.89	7		DEHIKYcn
41	Antosiewicz	1996	FDPB		52	0.7	4		CDEHKYcn
42	Demchuk	1996	FDPB		48	0.5	3		CDEHKRYcn
13	Nielsen	2001	FDPB		124	0.91	9		CDEHKRYcn
43	Georgescu	2002	FDPB		166	0.83	12		DEHKYcn
44	Czodrowski	2006	FDPB/PEOE		132	0.88	9		DEHKY
45	Barth	2007	FDPB		31	0.38	10		DE
46	Dimitrov	1997	DH		70	0.79	6		DEHKYcn
47	Warwicker	1999	DH		53	0.72	6		DEHKRcn
48	Warwicker	2004	FDPB/DH		117	0.86	15		CDEHKRYcn
49	Sham	1997	PDLD/S-LRA		9	0.73	2[d]		DE
63	Schutz	2001	PDLD/S-LRA		11	0.31	4		DEHK
50	Sandberg	1999	PDS		40	0.83	3		DEHKRYcn
51	Mehler	1999	SCP		103 (8)	<0.5 (0.504)	7	ext	DEHKRYcn (DE)
52	Mongan	2004	GB		18	0.82	4[d]		DEHKY
53	Kuhn	2004	GB		69	1	9		DEHKYcn
54	Pokala	2004	GB		226	0.92	15		DEH
55	Spassov	2008	GB	1	21 (310)	0.45 (0.51)	1 (23)	ext	DEHKYcn
56	Khandogin	2006	GB/DH		135	0.95	10		DEHc
57	Wisz	2003	Emp	24	260	0.95	41		DEHKYcn
12	Krieger	2006	Emp	4	227	0.879	(27)	cv	DEHKYcn
21	He	2007	Emp	13	405	0.593 (0.775)	73	cv	DEHKcn
22	Li	2005	Emp	30	314 (77)	0.89 (0.56-0.97)	44 (4)	ext	CDEHKRYcn

Information in parentheses pertains to external data used for validation. [a] FDPB – finite difference Poisson-Boltzmann; DH – Debye-Hückel; PDLD/S-LRA – protein dipoles Langevin dipoles/solvation linear response approximation; PDS – position dependent screening; SCP – screened coulomb potential; GB – generalized Born; Emp – empirical. [b] Validation method used: cross-validation (cv) or external data set (ext). [c] Modeled amino acid residue types in single character notation plus c for C-terminus and n for N-terminus. [e] Lysozyme crystal structures having different conformations were considered.

In the 1990's, pK_a predictors tended to optimize their parameters in order to beat the null model.[41,42,58,59,60] It was quite common to ignore external validation and focus only on selecting the optimal dielectric constant(s) to optimize the RMSE using experimental pK_a data for the ionizable sites of a few proteins, such as hen egg white lysozyme (HEWL), ribonuclease A (Rnase A), and bovine pancreatic trypsin inhibitor (BPTI). The Antosiewicz and the Demchuk models improved their calculations by adjusting a single parameter, the dielectric constant. Antosiewicz used a fixed dielectric constant of 20 for best results.[10] The Demchuk model assumed different local dielectric constants (ranging from 15 for buried residues to 80 for highly solvated residues) in the neighborhood of each ionizable site in order to better fit the experimental data.[42]

When developing any model, being able to fit the experimental training data is essential. Even though the FDPB models are based on solid physics, the dielectric constant needs to be tweaked in order to improve performance, much as the variables in empirical linear regression models are adjusted to fit the training data. A model's predictions are not likely to be as good as its fit to the training data. One must keep in mind that an unvalidated model tells us very little about that model's ability to predict new data. While the fixed dielectric constant model was not presented with external data for validation, the creators of the multiple dielectric constant models included a small external test set which was used to compare the ability of both models to predict the pK_as of 12 buried histidine residues from triose phosphate isomerase (TIM).[42] The RMSE for the fixed dielectric was 0.40, and for the multiple dielectric was 0.42, compared to 1.26 for the null model. In Table 2.3 the multiple dielectric model shows better fits than the fixed dielectric model, but the multiple local dielectric constants were freely varied throughout for best results, so no fair comparison can be made. Second, as the external test set is not diverse, one cannot draw any general conclusions about either model's ability to predict new data.

When considering the protein dielectric constant, ε_p, for use in electrostatic calculations, it is worth noting that ε_p depends on the method and system used to define it.[26,61,62] Furthermore, the best value for ε_p in modeling electrostatic effects has been found to have little to do with the protein dielectric constant, but rather is a measure of the electrostatic interactions which have not been included explicitly in the model,[63] such as conformational variations due to flexibility, specific water binding,[64] or proton/hydrogen-bond network relaxation.[65]

FDPB methods have made strides by incorporating protein flexibility,[40,66] solvation models,[67] and by improving efficiency through Monte Carlo sampling of the many microstates of a protein.[13] 329 data points were calculated using Kieseritzky's PACs method,[68] built on the Karlsberg+ framework, which combines continuum electrostatics with multiple pH adapted conformations selected by Monte-Carlo sampling. While the overall RMSE failed to beat the null model, improvements in accuracy were found for the residues having a pK_a shift greater than or equal to 1.0. Also, the multi-conformational continuum electrostatics approach of Georgescu, Alexov, and Gunner has been included in recent surveys and found to perform well on residues experiencing strong pK_a shifts.[11,43,68,69] Some other FDPB methods make use of significantly larger training sets, by including data from other proteins, and have been implemented as web applications, such as WHAT IF.[13,70]

Alternative methods used to approximate FDPB calculations include Partial Dipole Langevin Dipole, Debye-Hückel, Generalized Born, charge screening based methods, and their combinations.

Warshel and colleagues have developed the Protein Dipole Langevin Dipole (PDLD) method, in which protein dipoles are modeled.[63,49] This method tends to relax the atom-centered partial charge assumption and treats protein relaxation in a microscopic framework using linear response approximation (LRA) allowing for structural reorganization during charge formation. The net effect of PDLD/S-LRA is reduced reliance on ε_p, such that less variance of the parameter is required in order to accurately explain pK_a. Improved accuracy is seen when using both ε_p and ε_{eff}, the effective dielectric constant of the solvent, as adjustable parameters when fitting a set of 11 protein sidechains experiencing significant pK_a shift, ranging from –4.9 to 5.3 log units.[63] According to Schutz and Warshel, ε_p and ε_{eff} are model dependent scaling factors. Having less variance between parameters is a big plus, as the optimal values for other models (Generalized Born, Tanford and Kirkwood and modified Tanford and Kirkwood)[71,72] ε_p in their survey ranged from –80 to 80. Unfortunately, there is no way to forsee what values should be used for new data, and the predictive power of these models was not evaluated.

The Debye-Hückel approximation for electrostatic pair interactions has been used individually and in combination with FDPB calculations to model pK_a. Dimitrov[46] showed how it could be used as a standalone alternative to FDPB methods with superior results for the same six proteins considered by Antosiewicz. Warwicker found that the Debye-Hückel method matched the overall pH-dependent stability, while the FDPB method provided more accurate results for active-site groups. Combining both methods provided a computational framework for distinguishing solvent-accessible groups from buried groups. In doing so, significant improvements were found for a larger set of residues than considered in previous works using only the Debye-Hückel method.[47,48]

In search for a simple, less time consuming way for modeling electrostatic interactions in proteins, Sandberg described a distance and position dependent screening technique for the electrostatic potential. The goal was to calculate pK_as by applying this technique in conjunction with a Monte Carlo algorithm to speed up protein molecular dynamic simulations. While the results were slightly less accurate than those using FDPB calculations with the dielectric constant of 20, execution time of the algorithm was two orders of magnitude faster than the traditional grid based Poisson-Boltzmann calculations, with one pK_a calculation every 10 seconds compared to 30 minutes, respectively.[50]

In order to deal with significant prediction errors for the pK_a of specific residues by various FDPB methods, Mehler introduced sigmoidally screened coulomb potentials and considered microenvironment hydrophobicity, based on the hypothesis that the key factor responsible for pK_a shift is the protein microenvironment around each ionizable residue.[51] By considering the log$P$ contributions of groups of atoms very near the ionizable sites, they improved their fit to the training data, which in this case consisted of 103 ionizable residues from 7 proteins. A total of 8 Asp and Glu residues from turkey ovomucoid inhibitor domain 3 and the aspartyl dyad of HIV protease were used to validate the method, having an RMSE of ~0.50. While the fit to the training data was an improvement, the test set was not diverse, and it is not possible to tell how this model will perform on residues without carboxyl groups. Instead of using the Rekker fragmental hydrophobic constants,[73,74] the authors suggested future results might be improved by using more recent atomistic hydrophobicity values. Perhaps Slog*P* would help.[75] One might also consider combinations of other atomistic descriptors, such as the ratio of hydrogen bond donors or acceptors to carbon atoms within an ellipsoid of some radius surrounding the ionizable site.

The generalized Born model (GB) is an approximation of the Poisson-Boltzmann equation, and it can efficiently describe the electrostatics of molecules in an aqueous environment by implicitly representing the solvent as continuum with the dielectric properties of water and thus reducing the computational demand associated with molecular dynamics (MD) simulations. The basic idea behind the GB model is to assign each atom an effective radius, such that the solvation free energy can be calculated using the Born formula. Therefore, it is important to accurately calculate the Born radii, when using any GB model.[76] Efficiency is achieved by describing the instantaneous solvent dielectric response, which eliminates the need for equilibration of water in explicit water simulations. Furthermore, as the GB model corresponds to solvation in an infinite volume of solvent, it avoids artifacts associated with replica interactions in periodic system.

Mongan describes a method used to evaluate four different crystal structures of HEWL applying implicit solvent GB electrostatics and performing MD at constant pH with periodic Monte Carlo sampling of protonation states.[52] In contrast to most electrostatically based methods, this model was shown to be independent of the starting crystal structure.

Similar to the FDPB methods, the dielectric constant is used as an adjustable parameter. Kuhn used a molecular mechanical MM/GBSA approach, which is among the most commonly used implicit solvent model combinations and typically used for calculating biomolecular binding free energies.[53] In MM/GBSA, the GB model is augmented by a term representing the hydrophobic solvent accessible surface area (SA). Kuhn et. al. question the overall accuracy of the GB method noting that in theory, complex electrostatic interactions involving several charges and electric dipoles in close proximity should be better handled by Poisson-Boltzman continuum solvation calculations.[53]

Prior to Pokala's publication of EGAD (Egad! A Genetic Algorithm for Protein Design) in 2004,[54] the published GB methodologies considered relatively few macromolecular pK_a data points for training and validation. EGAD is particularly attractive, as electrostatic calculations are reported to be performed six orders of magnitude faster than FDPB methods. This increase in speed can be attributed to an approximation the Born radii. By applying these approximations to the GB continuum dielectric model and extending the methodology to approximate solvent accessible surface area, EGAD method was used to provide calculations for 226 ionizable groups from 15 proteins with similar accuracy to other GB models of the same time. Here, it is noteworthy that a subset of five proteins was used to parameterize the model and identified the optimal $\varepsilon_p = 8$. On the other hand, the predictive power of the model on new data is questionable. First, the number of data points used to parameterize the model was not disclosed. More importantly, statistics were provided for all 226 together, instead of independently evaluating the training and test sets. Finally, due to overall lack of data only the Asp, Glu, and His sidechains were considered. Data for a small number of Lys measurements, compared to the other residues, were omitted as the inclusion can significantly increase the correlation coefficient, while minimally affecting the RMSD, as discussed in section 5.

Khandogin and Brooks presented a first principles model based on continuous constant pH molecular dynamics simulations, utilizing replica-exchange protocol for enhance conformational and protonation state sampling.[56] The method is based on a GB implicit solvent model, which is modified by an approximate DH screening function. The DH screening function is used to account for salt effects. One of the more interesting features of this method is the scaling of the dielectric constant based on the DH length, instead of simply adjusting ε_p to empirically find the best fit for the data. RMSE for proteins with ionizable residues exhibiting low pK_a shift was approximately 0.6, whereas it was approximately 1.0 for the proteins with more highly shifted residues.

Spassov uses GB approximations with an iterative mobile clustering (IMC) approach to calculate the equilibria of protons binding to titration sites in proteins.[55] IMC is used to halt the exponential growth of GB calculations when considering conformational flexibility. Here, binding and conformational states are fully enumerated for an ionizable site within a local cluster of ionizable sites. Ionizable sites outside the cluster are treated by mean field approximation. The procedure continues to iterate through the list of all ionizable sites, repeating the calculations and using the results from previous iterations outside of the present cluster for the mean field terms of the current iteration until some convergence criteria is met.[77] This method has been incorporated into the Accelrys Discovery Studio. Not only does it appear to be highly accurate with a low RMSE, approximately equal to 0.50, but it considers the largest external set of test data of all the GB models in this review. It is of note that this model trained its single adjustable parameter, ε_p, only on HEWL (2lzt) and validated the model on over 300 external data points from 23 proteins with RMSE of approximately 0.5. A survey[55] of 105 ionizable residues from 7 proteins showed improved accuracy for this model over the top methods in the other classes mentioned in this review.[22,43,48,51,56] A close inspection of the dataset used in the survey revealed that approximately 20% of the residues were shifted more than one pK_a unit. Evaluating the dataset with the null model proposed by Thurkill[8] resulted in an RMSE and mean absolute error of less than 0.80 and 0.70, respectively, and a maximum error of 2.43. While the IMC method posted the best results, four of the other five methods surveyed had lower RMSE than the null model. It is rather curious that the IMC model exhibited poorer performance on its own training set, residues from hen egg white lysozyme (2lzt), than five of six of the other proteins considered in the survey. Typically, it is expected that a model's performance decreases when used to evaluate external test data. However, this may be explained, as 2lzt has a larger proportion of highly shifted residues than most of the other proteins considered. Apparently, the IMC approach to handling conformational flexibility is responsible for the high accuracy reported by this model.

2.2.2.3 Empirical Models

Wisz used four model equations to determine 24 independent parameters, which could be used to simulate the electrostatic interactions in proteins. Monte Carlo methodology was used to achieve convergence for all 24 parameters based on titration curves derived at 1 pH unit intervals from 0 to 14 using the model equations.[57] The training set consisted of 260 ionizable groups of which over 20% of the residues had pK_a shift greater than or equal to 1 unit. In order to investigate the stability of the parameters, additional rounds of Monte Carlo simulations were run, but no true validation was performed on external data.

Krieger published a method in which the electrostatic potential is evaluated using Ewald summation. Ewald summation is fast and can be used to monitor pK_a shifts during MD simulations and effectively handles periodic crystal environments.[12] Naive electrostatic calculations in periodic systems may diverge. Here Ewald summation allows for simplification within the periodic environment by combining a rapidly converging short-range variable with a long-range term evaluated in reciprocal space. The particle-mesh Ewald algorithm, standard in many MD programs, was used to identify models based on three and four parameters. 227 ionizable sites from 27 proteins were considered and a leave-one-protein out cross-validation was performed. Both three and four parameter models outperformed the null method, which also had RMSE less than 1.0. It is worth noting that the null model was optimized using a similar cross-validation technique where the mean was taken of the respective pK_a values of the amino acids of 26 proteins and used to predict the pK_a of the remaining protein. The optimization technique improved null model predictions by over 0.2 RMSE units.

From an empirical standpoint, it is highly unlikely that a model trained on data, having relatively low variance from their respective null values could accurately predict the pK_a for ionizable sites having significantly shifted pK_a values. This is especially true if the pK_a shift is due to an environment which was not considered by the model. This is best explained by the statistical optimization performed by He et. al.[21] Here a significantly larger data set was considered by mining the PPD, such that 1122 pK_a values belonging to 667 ionizable sites could be utilized for training and validation purposes.

Structures were validated against the PDB.[78] After curation 475 unique sites from 73 proteins were accepted. According to He et. al. “In the data set, the pK_a values of 46 sites are unusual because of physical or chemical factors, such as salt bridges or disulfide bridges (Table II). To obtain reasonable parameters, these data were excluded from the fitting procedure and were predicted using parameters obtained from the remaining sites.” In total 13 parameters were considered using multiple linear regression, where each parameter represented one or more amino acid types. pK_a shift was induced by residue-residue interaction determined by the amino acids surrounding the C^α of the ionizable residue within a sphere of some radius (minimum RMSE at 11 Å). Based on the data set used to train the model, it is obvious why the model performed reasonably well on the 405 residues with low pK_a shift (RMSE = 0.775, using six-fold cross-validation) and quite poorly on the external data, composed of the 46 unusual ionizable sites, which also had highest pK_a shifts (RMSE = 4.258).

Seemingly, the most accepted empirical method for predicting protein pK_a in literature today is PROPKA.[22] The most thorough surveys, often entitled benchmarks, include PROPKA performance as the method to beat.[11,55,68,69] In each survey, PROPKA’s RMSE is less than 1.0, including those that consider highly shifted residues. PROPKA’s origins began with quantum mechanical/molecular mechanical studies by Jensen, where analyses of pK_a determinants led to a set of quantitative structure property relationships forming the basis of PROPKA.[79] The model was trained on 314 experimental values using 30 parameters, 20 of which are distance related. Validation was performed on four proteins not appearing in the test set where the RMSE of the predictions for each individual protein ranged from 0.56 to 0.97 with a maximum predicted error of 2.0 units.

The original release of PROPKA version 1.0 was noted to ignore pK_a shifts caused by ligands, ions and waters interacting with the protein. More recently, version 2.0 incorporates protein-ligand interactions affecting ionizable groups as well as providing predictions for the ionizable groups of ligands in the protein environment.[80] Calculations are usually complete for an entire protein in a matter of seconds. Desolvation, hydrogen-bonding, and charge-charge interactions are considered in calculating the shifts, as well as groups with a fixed charged, such as Zn^{+2}. Current limitations noted by the developers include the assumption that intra-ligand interactions are included in the pK_a model value, while both pK_a shifts due to inter-ligand interactions and the effects of sidechain motion as well as other conformations are not considered.

Desirable qualities for empirical models are large diverse training sets fitted to as few parameters as possible and high predictive accuracy on a diverse data set that was not used for training purposes.

2.3 Small Molecules

Predicting the pK_as for small molecules is a substantially different problem than for proteins, based on their respective isolated environments. There are far fewer microstates to consider, and most three-dimensional effects are commonly neglected, such as the local electrostatic field and degree of solvent exposure. However, the range of chemical structures is far greater, so care must be taken when assessing the diversity of training and test sets. It has been shown that understanding the site-specific charges and concentrations of the microspecies can allow for more realistic predictions regarding a molecule's pharmacokinetic behavior.[81] Unfortunately, most models are simplifications and provide only macrospecies predictions due to the limitations of the data available for training. However it is possible to make predictions regarding the microspecies by interpolating the approximated titration curves based on the macrospecies. At least three available applications, ChemAxon's MARVIN[82] and ACD/PhysChem Suite[83], and Simulations *Plus* ADMET Predictor[84] provide microspecies predictions for small molecules.

2.3.1 Experimental Data

2.3.1.1 Data Curation

While a great deal of experimental pK_a data can be found in the literature and 'in house', its reliability is sometimes questionable. A large portion of the literature containing data on small molecules is recorded in the Beilstein database and is accessible using the MDL Crossfire Commander.[85] Lange's Handbook of Chemistry also provides a good source of pK_a data.[86] Software tools, such as ACD and SPARC, provide access to experimental data with references. SPARC will only provide the reference data based on queries for an exact structural match.[87] ACD iLabs has the added benefit of providing literature references for molecules based on a structural similarity search, returning references for exact structural matches and a limited number of similar structures based on a user defined similarity score threshold.[88] Another potential source which can identify articles that may contain pK_a data based on a structure or text based query is SciFinder Scholar.[89] Unfortunately, such data obtained from sources other than the original literature references are not necessarily clean, complete, or standardized. Sometimes even published data is unreliable, and often conflicts are found between sources. For example, when mining Beilstein for the experimental pK_a measurements of phenol, approximately 30 records exist in a bimodal distribution. The distribution consists of a cluster of six values around the pK_a of –1.0 and over 20 values clustered around the pK_a of 10. Evidently, –1.0 values arose from entering the negative logarithm of the pK_a. Other common but less frequently identified errors include: typographical errors, predicted values (rather than experimental), incorrect transcription of temperature and/or solvent used, and the incorrect associations between the experimental pK_a and the ionizable sites on polyprotic molecules. In a previous effort to identify and resolve some of these problems, a method for curating pK_a data from multiple sources having redundancies was discussed.[90] Finally, so much pK_a data is associated with proprietary chemical structures that public training sets are not as chemically diverse as they could be. Hence predictive models may be less accurate for some molecules in proprietary datasets. Table 2.4 provides an updated list of free and commercial electronic sources of pK_a data from the review article published in 2006 on '*in silico*' prediction of ionization constants by Fraczkiewicz.[91] Another significant source of pK_a data is the six-volume "Critical Stability Constants" by Martell and Smith.[92] This data is available for purchase as the NIST Standard Reference Database 46 from the National Institute of Standards and Technology.[93] Buyer beware, as the database is a self-contained application which has very limited search capability and does not support data export to common machine readable formats such as comma space delimited (CSV) or standard data files (SDF).[94]

Table 2.4. pK_a data sources

Data source	Vendor	url
ACD/pK_a DB	Advanced Chemistry Development	www.acdlabs.com
ADME INDEX	Lighthouse Data Solutions	www.bio-rad.com
Beilstein Database	Elsevier B.V.	www.info.crossfiredatabases.com/
BioLoom Database	BioByte Corporation	www.biobyte.com
The Merck Index, 14th edition	Cambridgesoft Corp.	www.cambridgesoft.com
Lange's Handbook of Chemistry, 15th Edition	Knovel	www.knovel.com
CRC Handbook of Chemistry and Physics, 89th Edition	CRC	www.hbcpnetbase.com/
HSDB	National Institutes of Health	toxnet.nlm.nih.gov/
LOGKOW	Sangster Research Laboratories	logkow.cisti.nrc.ca/logkow/
MolSuite DB	ChemSW	www.chemsw.com
Pallas	CompuDrug	www.compudrug.com
pK database	University of Tartu, Estonia	mega.chem.ut.ee/tktool/teadus/pkdb/
PHYSPROP	Syracuse Research, Inc.	www.syrres.com
SPARC	University of Georgia	ibmlc2.chem.uga.edu/sparc/

In the absence of available data, ideally one should simply measure the pK_a by titration. However, this is not an option for large libraries of *in silico* small molecules that have yet to be synthesized. When dealing with the vastness of chemical space, often a good computational approximation is superior to experiment in order to overcome cost and time limitations.

2.3.1.2 Experimental Methods

Analytical chemistry has provided a plethora of experimental tools for making pK_a measurements, some of which are amenable to automation. For those interested in the history of titration and its development for the use of colorimetric and electrometric analysis used in the determination of pK_a, The *History of Analytic Chemistry* describes the achievements during the early 20th century.[95] Today, there are two main titration methods: volumetric and coulometric. The volumetric method entails adding titrant directly to the sample, whereas the coulometric method generates titrant electrochemically. Some of the associated indicator methods include: colorometric, potentiometric, conductometric, spectrophotometric, amperometric, thermometric, solubility, cryometric, and NMR (discussed above for protein pK_as). In order to show the most commonly used and preferred methods, queries of data obtained from the Beilstein database relating to the analytical methods were performed as shown in Table 2.5. The potentiometric, spectrophotometric and conductometric methods have been used predominantly to determine pK_a. Interestingly enough, approximately half of the pK_a measurements obtained from the Beilstein database are not associated with a method. Furthermore, over the past year the total number of measurements increased by approximately 20%. Only 10% of the new pK_a data are associated with a method. Regardless, over 98% of the measurements with stated indicator methods used potentiometry, spectrophotometry and conductometry. Seemingly, there is a wealth of available information, however after curating the data for monoprotic molecules and accounting for redundancies, the authors have found reliable pK_a data for fewer than 2000 molecules.[90]

Table 2.5. Beilstein database (DE.MET): dissociation exponent method category

Indicator Method	Count based on Beilstein[a] Version	
	2007.04	2008.03
(Blank)	56832	79602
Potentiometric	45639	46906
Spectrophotometric	18339	18872
Conductometric	2127	2163
NMR	687	774
Kinetics	359	367
Calorimetric	76	143
Solubility	31	32
Polarographic	11	15
Distribution	5	6
Total	124106	148880

Numbers reflect pK_a measurements in all conditions including same molecule measurements in various solvents and temperatures. [a] Beilstein was accessed using the Molecular Design Limited (MDL) CrossFire Commander.

Potentiometric titrations are possible in turbid, deeply colored, highly absorptive, or dilute solutions. Simplicity, speed of measurement, robustness, and ease of automation have made it the historically preferred method to measure pK_a. On the other hand, potentiometry does have pitfalls when it comes to pK_a measurement. Measurements on compounds sparingly soluble in water require mixed solvent extrapolation.[96] Non-high throughput applications require large amounts of reagent (not favorable for newly synthesized compounds or natural products) and time to prepare solvent from carbonate free solutions.[97] Furthermore, impurities of both reagent and analyte can affect the observed pK_a values.[96,98] When dealing with nonaqueous media, repeated measurements are often recommended, as the signals from the glass electrodes are less reliable than those from aqueous media, due to high liquid junction potentials.

Similar to the potentiometric method, conductometric pK_a measurements are possible in turbid, deeply colored, highly absorptive, and dilute solutions. They are relatively simple, quick to perform, and have been automated. On the other hand, it does not suffer from the same degree of reagent solubility limitations experienced in potentiometry, although conductometric methods are rather problematic in the presence of foreign electrolytes, which decrease the accuracy of the measurement. Purity is not the only issue, as conductometric methods are also sensitive to temperature. Raising the temperature one degree results in a 2–2.5% increase in the conductivity of most salts. Furthermore, conductometry is generally considered inferior in accuracy compared to potentiometric and spectrophotometric methods.[99]

Ultraviolet spectroscopy hinges on the principle that the uncharged and ionic species of a compound exhibit different spectra. Spectroscopy is known for its excellent precision in pK_a measurements. While it is comparable in accuracy to potentiometry, spectroscopy can be used to measure pK_a for compounds having poor aqueous solubility. Amenable to high throughput automation, UV spectroscopy has the added advantage of requiring ten times lower concentrations of reagent than similar high throughput potentiometric titrations do.[97,100]

In the early 1990's, capillary electrophoresis (CE) began to show usefulness as a universal analytical technique for determining pK_a over a wide pH range.[101,102] It relies on the principle that the solute exhibits an electrophoretic mobility continuum versus pH (uncharged has no mobility; charged has maximum mobility). It exhibits both higher sensitivity and selectivity than potentiometry and spectrophotometry do, producing accurate pK_a values for small molecules. The method is capable of handling compounds of diverse solubility and samples of low concentration since it relies on migration times and does not require measurement of the reagent or titrant concentrations, as potentiometry does. The solute is purified during migration, as impurities have inherently different migration times, so purity is not an issue with CE. Finally, CE is not limited by media, as measurements can be made in aqueous, aqueous-organic, or nonaqueous media.[96, 103]

pK_a can also be determined through reverse phase high performance liquid chromatography (HPLC) by measuring the capacity factor based on a compound's retention time in a column against a series of solvents having mobile phases at different pH values.[104,105] The advantages of using HPLC for pK_a screening include the minimization of solubility restrictions, eliminating the effect of impurities on measurements (thus allowing for screening combinatorial complexes directly), and the potential for high throughput automation when coupled with a mass spectrometer. However, the main potential disadvantage is the loss of accuracy when using organic solvents in a nonpolar column, due to the potential interaction between analytes and the stationary phase.

Several high throughput pK_a assays have been recently reviewed, such as capillary electrophoresis in conjunction with ultraviolet detection, capillary electrophoresis coupled with mass spectroscopy, pH gradient HPLC with mass spectroscopy, and a mixed buffer linear pH gradient system with measurements based on ultraviolet absorbance.[106] While it is not the goal of this paper to focus on the details of experimental methods, it is useful to note the development of newer technologies increasing the efficiencies of physicochemical property measurements for large chemical libraries of compounds lacking such data.

Familiarity with potential experimental problems can help in data curation. When performing a titration to determine pK_a, the following conditions are preferable to avoid unnecessary errors.[100] First, the analyte needs to be water soluble. For those molecules with poor water solubility, Yasuda–Shedlovsky plots can be used to extrapolate the theoretical aqueous pK_a from a gradient of semiaqueous (water:methanol) pK_a measurements.[107] In a similar approach, a series of measurements at various ionic strengths allows for extrapolation to zero ionic strength. An alternative approach, sacrificing accuracy for number of measurements, requires only one measurement and uses a linear equation with slope and intercept based on limited families of compounds, such as phenols and protonated amines.[108,109] Second, in order to acquire an accurate measurement, compounds need to be stable enough to establish an equilibrium between two states of ionization. Third, compounds need to be pure in simple titration experiments, although with high throughput techniques purity is much less of an issue. Fourth, in order to eliminate experimental errors and for the purpose of validation, it is usually preferred to use a series of homologous compounds. Fifth, each titration should be performed in a thermostatic environment, as dissociation is an endothermic process. It has been found that pK_a values for some common organic acids change by less than 1 pK_a unit between 5° C and 60° C, typically decreasing with the increase in temperature. Anomalies may result when approaching 0° C when the solvent is water.[110] Sixth, the presence of carbonate has been shown to affect data measurements leading to anomalies in the titration curve, which may require correction.[111] Carbonate acts as a base in aqueous solution and has pK_b of 6.36. Carbonic acid is diprotic and has pK_as of 3.60 and 10.25. The amount of carbonate in solution is proportional to the partial pressure of carbon dioxide in the atmosphere. Therefore, when performing titrations using a low concentration of acid, anomalies in measurements and the titration curve can occur when the pH is around 3.6, 6.36, and 10.25. Seventh, in the cases of most manual titrations, high amounts of test materials are needed. Finally, care must be taken when performing acid-base titrations with perfluorinated compounds, which exhibit an artifact of sorption. The unionized form of a perfluorinated compound tends to preferentially adhere to interfaces, both water surface and glassware. This non-uniform distribution affects the overall measurable concentrations of the unionized species in solution. Unfortunately, the extent to which sorption occurs with “ordinary” compounds is unknown, but it can be conjectured that a similar but lesser effect may be experienced for molecules having highly halogenated lipophilic tails. In cases where the concentration of the nonionic species is reduced due to sorption, one can expect the titration curve to be right-shifted, indicating higher than actual pK_a values for acids.[112,113] Furthermore, there are cases where the addition of co-solvents did not decrease the sorption of an organic acid in the aqueous phase.[114] Considering these points, even seemingly well curated data may be corrupted due to potential artifacts beyond simple human transcription errors, leading to unexpected biases in otherwise well constructed and validated predictive models.

2.3.2 Predictive Methods

2.3.2.1 Linear Free Energy Relationships

Seminal publications[115,116] and reviews[110,117] on the prediction of pK_a base their model on linear free energy relationships (LFER), applying the Hammett equation where pK_{a0} is the ionization constant for the parent molecule, pK_{aS} is that of the substituted molecule, ρ is the constant for a particular class of molecules, and σ_i is the effect of the i^{th} substituent on the ionization constant of the parent molecule.

$$pK_{aS} = pK_{a0} - \rho\sum\sigma_i \qquad (2)$$

The flaw in this method is that the parent molecules inherently carry the majority of the chemical information, and without training on a particular parent, predictions for such compounds are impossible. A good example of this problem is the MCASE study. Until the mid 1990s, there had been few attempts to model diverse sets of chemicals. Like eq. 2, the MCASE approach used a linear regression equation with terms for quantitative properties such as log$P$, water solubility, molecular weight, absolute electronegativity, hardness, Hückel molecular orbital charge densities, and HOMO and

LUMO coefficients, plus possibly 58 indicator variables for the presence of certain molecular substructures. The set of compounds was broken into 22 subsets based on the presence of molecular fragments called biophores that were particularly associated with acidity, the most important one being the carboxyl group. Different regression equations were determined for each subset. The total training and test set consisted of 2464 organic acids. When using 1848 molecules in the training set, the r^2 based on the predictions of the remaining 616 compounds was 0.91 with standard deviation 0.774. When training the model with the entire dataset and attempting to predict the pK_a for 214 drug molecules, which were most likely not well represented by the parent molecules in the training set, the r^2 was 0.70 with a standard of deviation of 1.44.[118]

LFER models are still used and have been implemented in popular commercial and freely available software packages, such as EPIK and SPARC.[119,120,121,122] To help overcome the aforementioned problem, the SPARC methodology combines LFER, perturbed molecular orbital theory and QSPR to deal with the effects of π-bonding and electron delocalization. SPARC scored an r^2 of 0.80 with RMSE of 1.05 on a set of 537 molecules from an internal Pfizer dataset[122] and is able to predict both macro and micro pK_as calculated from molecular structure alone. It is worth noting that while SPARC is not parameterized for all atom types, one can modify the molecule being tested by substituting the closest weight atom of the same group and achieve reasonable, often good, results. For example, we have found that most predictions where silicon atoms were replaced by carbons and seleniums by sulfur were within 1 pK_a unit from the experimental value. However, since there are few heavy atom compounds to test, the accuracy of such replacements remains questionable.

2.3.2.2 Quantitative Structure-Property Relations

One of the most common techniques used in pK_a prediction is quantitative structure activity/property relationships (QSAR/QSPR) deriving their fit equations from partial least squares (PLS) or multiple linear regression (MLR).[118,123,124,125,126,127,128,129] Other methods include artificial neural networks,[84,91,130,131] quantum mechanical continuum solvation models,[132,133,134] anti-connectivity models,[135,136] and tree based methods.[90,127,137,138] It has often been the case that a model was based on a relatively small set of experimental data for a specific ionizable group, such as carboxylic acids.[124,129,130,132,134,135,136] Others have tackled the problem of chemical diversity by devising and combining multiple models, each applied to a relatively small set of molecules when compared to the complete set of experimental data.[127,130] Here the overall combination of models is more robust at handling novel chemical structures, but the individual training sets may suffer from a lack of chemical diversity due to their small size. This may allow for a good fit on the training sets, but has the potential disadvantage of leaving little freedom for cross-validation. The following sections address some of the most significant methods that have been used in property prediction. Table 2.6 shows the performance of several commercial pK_a prediction models and other methods used in literature within the past two decades.

One of the reasons QSPRs are so popular is that linear regression always leads to a unique, easily computed model. Typically, highly correlated descriptors are undesirable, but with partial least squares, highly correlated descriptors are handled appropriately. Like the MCASE approach, when dealing with complex properties such as pK_a, it is common to break up the training set into groups based on ionizable site type and chemical group, because it is impossible to establish a single robust model on a chemically diverse set of molecules. In doing so, multiple models are generated which may or may not use the same set of descriptors. Descriptors can be qualitative or binary, representing a <has> *vs.* <has not> property, or they can be quantitative experimental or calculated features. Reducing the size of the training sets by separating the molecules into classes often leads to a better fit, but does not necessarily reduce the number of descriptors considered. Caution needs to be exercised in order to avoid overfitting, especially when nonlinear descriptors are considered. If we have a descriptor named x, the regression equation can also have terms for $\ln x$, $\log_2 x$, $\log_{10} x$, x^n (where n is any real number), and so on. Overfitting the model leads to a high correlation between the calculated and experimental values in the training set, but a poor or non-existent correlation in cross-validation or validation on a separate test set. Forward or reverse stepwise linear regression are ways to select the smallest subset of descriptors which still fit the property being modeled. In 1972 Topliss and Costello pointed out the risk of chance correlations in quantitative structure activity relationships and gave recommendations for the number of descriptors to be used in linear regression models given the number of observations to be fit.[139] Hence, the major step in deriving a robust QSPR model is finding the smallest set of molecular descriptors that best represent the structural variations in a set of chemically diverse molecules. In QSPR the most common methods to identify a good set of descriptors are stepwise multiple linear regression, partial least squares (PLS), and principal components analysis.

Comparative molecular field analysis (CoMFA) has been used to model pK_a values for small series of chemical homologs where partial least squares found a linear correlation using four parameters.[140-143] While the statistics appear very good, one must note two limiting factors: first the models were trained on very small, chemically similar training sets, and second the results depend on the chosen conformations and spatial alignments. More recently, a CoMSA (comparative molecular surface area analysis) study using similar 3D descriptors and PLS fitted the pK_a values for a series of benzoic acids.[128] The previous CoMFA methods appear to outperform the CoMSA method in predicting pK_a.

Table 2.6. Survey of pK_a prediction methods

			Training Set			Test Set			External Test Set		
Method	**Ref.**	**Class**	***N***	***r*^{2}**	***RMSE***	***n***	***q*^{2}**	***RMSE***	***n***	***r*^{2}**	***RMSE***
QSPR/PLS	137	all subclasses							25	0.95^a	0.78^a
		acids	625	0.98	0.405	10%	0.86	1.04			
		bases	412	0.99	0.298	10%	0.87	1.12			
QSPR/PLS – MoKa	127,141	33 subclasses	24617						39	0.80	0.90
		acidic nitrogen	421	0.97	0.41	20%	0.87	0.41			
		6 member N-heterocyclic bases	947	0.93	0.60	20%	0.85	0.86			
QSPR/PLS (CoMSA)	128		49			49	0.86		23	0.77	
QSPR/PLS (CoMFA)	140	benzoic acids	49	0.916	0.102 ^{b}						
	142	imidazoles	23	0.99	0.19^b			0.27^b	5	0.98^a	0.15^{ab}
		imidazolines	16	0.99	0.35^b			0.69^b			
	143	nucleic acid components	18	0.99	0.19^b		0.89				
QSPR/MLR	125		15	0.97	0.12				3	0.99^a	0.10^a
QSPR/MLR	126	carboxylic acids	1122	0.81	0.42^b	20%	0.81	0.43^b			
		alcohols	288	0.82	0.76^b	20%	0.81	0.78^b			
QSPR/MLR	129	aromatic acids	74						33	0.99	0.27
QSPR/LFER	124	monoprotic oxy acids	135	0.993	0.455				14		0.471
QSPR/LFER – MCASE	118		2464			616	0.91	0.774^b	214	0.70	1.52^b
QSPR/LFER – EPIK	119,120		4057		1.27^b				123		1.37^b
QSPR Anti-Connectivity	136		31			31	0.87	0.463			
ANN – ChemSilico	130	12 classes	>16000						665	0.83	
		primary amine	1100	0.95		20%	0.92				
		tertiary amines	870	0.92		811	0.80				
		monoprotic acids	1640	0.95		1640	0.88				
		aromatic nitrogen	1480	0.92		1367	0.80				
		alcohols	1302	0.88		1302	0.85				
ANN/PCA/GA	131	nitrogen	170(282)	0.99	0.30						
ADMET Predictor	84,91		9075	0.971	0.593				2253	0.961	0.644

Method	Ref.	Class	Training Set			Test Set			External Test Set		
			N	r^2	*RMSE*	*n*	q^2	*RMSE*	*n*	r^2	*RMSE*
Semiempirical/PLS (Novartis In-House)	123	all			0.48				350		0.81
		alcohols	202	0.87	0.58		0.80				
		amines	1403	0.89	0.49		0.84				
		anilines	311	0.90	0.49		0.78				
		carboxylic acids	681	0.90	0.34		0.86				
		imines	84	0.98	0.55		0.88				
		pyridines	397	0.95	0.58		0.86				
		pyrimidines	91	0.95	0.43		0.87				
Semiempirical MO	144	phenols				175	0.93	0.599 [b]			
		benzoic acids				99	0.85	0.357 [b]			
	145	amines & anilines				132	0.94	0.985 [b]			
		N containing heterocycles				150	0.69	1.168 [b]			
Semiempirical RM1+solv.	146	carbon containing aliphatic amines	26	0.948	0.68[b]						
Quantum (MEP)	147	phenols & carboxylic acids	228	0.896							
Quantum (MEP-V_{min})	148	anilines	36	0.945	0.301[b]						
(MEP-$V_{S,min}$)				0.932	0.336[b]						
($I_{S,min}$)				0.949	0.285[b]						
(Hammet σ)				0.940	0.310[b]						
Quantum (MEP-$V_{S,min}$)	149	phenols	19	0.938	0.300[b]						
(MEP-$V_{S,max}$)				0.932	0.314[b]						
($I_{S,min}$)				0.941	0.292[b]						
($I_{S,min}$ & $V_{S,max}$)				0.953	0.271[b]						
Quantum (MEP-$V_{S,min}$)	149	benzoic acids	17	0.942	0.120[b]						
(MEP-$V_{S,max}$)				0.970	0.085[b]						
($I_{S,min}$)				0.941	0.120[b]						
Quantum (philiity)	150		63	0.98	0.57[b]						
Quantum Solvation	132	carboxylic acids	16	0.69	0.72						
Quantum Solvation	151	phenols	20		0.38						
Quantum Solvation	152		11	0.88	2.2						
COSMO-RS	133	bases	43	0.98	0.56[b]				58		0.66
	134	acids	64	0.98	0.49[b]						
Jaguar	153								191	0.98	0.66
MD continuum	154	diprotic acids	12	0.96	2.02						

Method	Ref.	Class	Training Set			Test Set			External Test Set		
			N	*r*^{2}	*RMSE*	*n*	*q*^{2}	*RMSE*	*n*	*r*^{2}	*RMSE*
solvation											
QSPR/LFER/PMO – SPARC	87,121,122		2500	0.99	0.36[b]				4338	0.99	0.37[b]
		Pfizer dataset[c]							123	0.92	0.78[b]
		Pfizer internal dataset[d]							537	0.80	1.05[b]
									185[c]	0.84	1.15
MARVIN	82,155					208[c]	0.98	0.38[b]	185[c]	0.88	1.03
ACD/I-Lab v8.03	88		>31000						185[c]	0.90	0.93
ADME Boxes	156								185[c]	0.93	0.69
SMARTS p*K*$_a$	90		1693	0.95	0.65	10%	0.91	0.80	185	0.94	0.68
Consensus[e]	90								185	0.96	0.60

In all training sets *n* refers to the number of pK_a measurements; in the test sets *n* refers to the number of pK_a measurements or percentage of the training set. [a] External set statistics were calculated from data presented in the referenced material. [b] Standard deviation. [c] It is unknown whether these molecules were used in the training set. See reference 90. [d] These molecules were unlikely to be found in the SPARC training set. [e] The consensus model used predictions from SPARC, MARVIN, ACD/I-Lab 8.03 and SMARTS pK_a.

2.3.2.3 Quantum Mechanical and Continuum Electrostatic Methods

Continuum electrostatics models and quantum mechanical descriptors have also been a focal point in small molecule pK_a prediction. Similar to the work by Antosiewicz on protein pK_a,[10] a continuum electrostatics model for small molecule diamines using UHBD to make FDPB calculations resulted in an r^2=0.86 and RMSE=1.1 for the 12 ionizable sites in six aliphatic diamines with experimental pK_as ranging from 1.09 to 10.34. Significant errors occurred in the calculations for the primary pK_a of 1,2-diaminopropane and the secondary pK_a of succinic acid, which were both calculated approximately 3 units below the experimental values.[154]

A polarizable continuum model was used to evaluate 15 small simple monoprotic molecules with experimental pK_a ranging from –6 to 33 with r^2=0.96 and RMSE = 2.02. All in all, there were 11 compounds that deprotonated within the range 0 to 16 with r^2=0.88, RMSE = 2.2, and a maximum error of 4.7.[152] Electrostatic models have also been used to predict pK_a for small multiprotic tetrahedral and triangular oxyacids, such as arsenic (H_3AsO_4) and arsenious (H_3AsO_3) acid, with close to the same accuracy as for the simpler organic acids.[157]

QM descriptors offer a promising means to accurately calculate pK_a. The *ab initio* aspect allows for greater confidence when calculating pK_a for molecules than when using strictly empirically derived descriptors. That is, QM methods are not restricted to the chemical diversity of a training set of molecules. However, the calculations are time consuming and not feasible when considering large databases of theoretical molecules or the analysis of macromolecules. Some QM descriptors that have a strong correlation to pK_a include superdelocalizability,[123] polarizability,[123] group philicity,[150,158] molecular electrostatic potential (MEP),[147,148,149,161] and molecular surface local ionization energy ($I_{S,min}$). [148,149,159,160,161]

Group philicity refers to the electrophilic nature of the ionizable group, such as a carboxyl group, and is equivalent to the sum of the local electrophilicities of each group atom, which are determined by the electrophilicity of their respective bonded neighboring atoms and calculated using density functional theory. The philicity descriptor is a modification of a molecule's electrophilic index.[162] The reciprocal of the group philicity showed a strong correlation to the pK_a for 63 molecules including carboxylic acids, substituted phenols, anilines, phosphoric acids, and alcohols.[150]

Three classes of MEP calculations have shown a strong correlation to pK_a: spatial minima (V_{min}), surface minima ($V_{S,min}$), and surface maxima ($V_{S,max}$).[148,149] It was recently shown that the MEP minus a given reference value for each category of compounds (as in the FDPB calculations for protein pK_a prediction with the electrostatic methods) has a single unique linear relationship to the experimental pK_a data for thiols, sulfonic acids, alcohols, carbonyl acids, amines, and analines.[147]

The investigations of Brinck et al. established the correlation between a molecule's pK_a and an ionizable atom's minimum surface local ionization energy ($I_{S,min}$), as defined by self-consistent-field molecular orbital theory.[159-161] The locations of the $I_{S,min}$ is related to charge/transfer polarization and indicates the areas where the least amount of energy is needed to abstract an electron from the surface of the molecule. Early investigations found a single linear relationship between the $I_{S,min}$ and pK_a for four sets of carbon and oxygen acids, as well as three nitrogen acids. It is easy to see from the scatter plots correlating $I_{S,min}$ to pK_a that some calculations missed by approximately 10 pH units, and that the high correlation coefficient (r=0.97) was due largely to the broad range (–5 to 40) of experimental pK_as considered.[160] Later investigations confirmed that not all ionizable groups could be represented by a single linear equation due to key structural differences between the ionizable groups.[149]

While MEP descriptors and $I_{S,min}$ both show good pK_a correlations for different series of compounds taken separately, it is interesting to note that the derivation of $I_{S,min}$ and $V_{S,min}$ correspond to different atoms and generally do not correlate.[149,161] When performing simple linear regression on one descriptor, the $I_{S,min}$ has been shown to be slightly superior to the V_{min}, $V_{S,min}$, and $V_{S,max}$ as well as the natural charge, and relative proton-transfer enthalpy.[148,149] Furthermore, it was found that no significant improvement could be obtained by linear regression on combinations of $I_{S,min}$, $V_{S,min}$, and $V_{S,max}$ QM descriptors.[149] Although these QM descriptors appear quite promising, no strong correlations have been found between them and pK_a for aliphatic amines. More recently, good correlations between neutral amines (excluding ammonia and hydroxylamine) and their cations were found by using the SM5.4A solvent model and performing calculations at both the semiempirical RM1 and density functional theory (DFT) B3LYP/6-31G* levels.[146]

Ab initio quantum mechanical methods are always the slowest but have often been found to be the most accurate, such as Jaguar, which performs geometry optimization at the DFT B3LYP/6-31G* level.[153] Shields et al. used QM calculations to accurately predict pK_a for 20 phenols[163] and 6 carboxylic acids[164] using a CPCM[165] continuum solvation method in Gaussian 98.[166] CPCM utilizes COSMO,[167] a conductor-like screening model to calculate the polarization charges of a molecule, in a polarizable continuum model (PCM) framework.

Another popular QM package that has been shown to perform as well as Jaguar in aqueous environments, capable of accurate pK_a calculations is COSMO-RS,[168,169] where the RS stands for real solvents. It is a statistical thermodynamics post-processing of COSMO calculations that extends the applicability of quantum chemistry to the entire range of fluid thermodynamics including mixtures and variable temperatures. An assessment of several *ab initio* programs for the prediction of pK_a, not including Jaguar, tested neutral, cationic, and carbon acids with experimental pK_as ranging from 14 to 36. Ignoring three outliers, the overall r^2 was 0.89, while considering all data points lowers the r^2 to 0.72. [170] Still there are foreseeable complications. The COSMO-RS model was able to fit a set of 43 bases very well, but when aliphatic amines were considered, correction factors needed to be introduced for secondary and tertiary amines which uniformly deviated from the regression line. Furthermore, two compounds (hexamethylenetetramine and 1,2-diazabicyclo[2,2,2]octane), did not share this deviation. In these cases, ionizable nitrogens act as bridging atoms for a bicyclic ring system.[133] Strictly empirically based methods could not even hope to achieve this level of accuracy based on the relatively small sampling of chemical space considered for both acids and bases.

For pK_a predictions outside the physiological range, the *ab initio* QM methods tend to be more robust and often more accurate than the empirical and less complicated continuum electrostatic methods. This class of pK_a predictors also allows for a broader range of analysis than is afforded by empirical models trained on pK_a data obtained from titrations in H_2O alone. On the other hand, validation has been performed on rather small data sets, and it is not clear that even the QM methods will be able to maintain their statistics, based on a study of carbonaceous ionizable sites, involving COSMO-RS and other contemporary theoretical methods.[170]

2.3.2.4 Artificial Neural Networks

The theory behind artificial neural networks (ANN) has been described in the literature.[171,172,173] ANNs are a powerful tool for making non-linear approximations and are designed to emulate how the brain processes information through a network of neurons. As such, the neurons of an ANN act as interconnected units, processing information based on mathematical functions. Like the internet, each neuron acts like a minicomputer receiving requests and sending responses to other neurons. Multilayered ANNs with enough neurons have been said to have the capability to approximate almost any nonlinear mapping of input to output to any required accuracy.[174] They are well suited to handle large datasets and identify complex non-linear patterns that could easily be missed by a simple equation or set of equations modeling a system, such as those derived through linear regression.

A principal component, genetic algorithm, artificial neural network was used to calculate the pK_a of 282 various nitrogen containing molecules in water.[131] The training set consisted of 170 molecules, the cross-validation set 56 molecules, and the prediction set 56 molecules. The model uses 406 descriptors to identify 179 principal components that explained 99.9% of the total pK_a variance. Of these, 15 principal components were included in the final model. While it appears that the model generation made use of the training and validation sets, we were not able to determine if only the molecules in the training and validation sets were used to perform the principal components analysis. The RMSE of the 56 molecule prediction set for the artificial neural network was 0.0750, compared to 1.4863 when multiple linear regression was used to build the model. The extreme accuracy of the artificial neural network model, the excessive number of descriptors used to perform the principal components analysis, and the small number of molecules suggests that PCA was performed on the entire dataset to identify the relevant principal components, and that overfitting is an issue. As with many other empirical models, this one is unable to identify the site of ionization.

A well validated ANN model has been implemented by Simulations Plus Inc.[84,91] This model has diverse training and test sets consisting of 9075 and 2253 pK_a data points, respectively. It is also the only neural network model and one of the few prediction utilities that predicts micro-pK_as. Based on the large training and test sets and the respectable score from the Chem Silico dataset, this utility appears to be not only the most accurate, but also one of the most robust methods for pK_a prediction. On the other hand, this model, like most others, is only parameterized for C, N, O, S, P, F, Cl, Br, and I atoms. However, it will process molecules with other atom types, as well as salts (which are washed away), providing a warning to the user. Substances that are mixtures of compounds are not processed.

2.3.2.5 Database Methods

Tree methods and database lookup methods are becoming more popular tools for pK_a prediction. The simplest form of database lookup method is to assign to the test molecule the pK_a of the database molecule that is most similar, according to some similarity metric, such as the Tanimoto similarity coefficient, based on some fingerprint.

The fingerprint is typically made up of qualitative descriptors. The accuracy of the model depends on how comprehensive (large and chemically diverse) the molecular database of experimental pK_a data is and the design of the fingerprint. To represent the database well, ideally every molecule in it would have a unique fingerprint. In one study it was found that pK_a assignment based on a simple atom type method using SMARTS strings was able to assign pK_a to simple molecules with an r^2 of 0.80 and a standard deviation of 0.95.[175]

Other methods are far more rigorous, such as that of Kogej and Muresan.[138] Here a pK_a database was mined using fingerprints based on 64 atom types represented by SMARTS strings and bond distance from ionizable site. A fingerprint was taken at each level of removal from the ionizable atom (level 1). Atoms one covalent bond removed from the ionizable site were at level 2, atoms at two bonds removed were at level 3, etc. After the entire training database is fingerprinted, a fingerprint exists for every concentric level of removal from the ionizable site for each ionizable site in the molecule. This style of submolecular fingerprints has also been coined ‘circular fingerprints’, as it dissects local structural information in expanding concentric levels of bond removal from the ionizable site.[176] This method allows for identification and predictions for each ionizable site in a molecule based on the theory that the ionization of a particular group is dependent on these topological subenvironments. Furthermore, it allows for predictions of the microspecies, but it is likely that microspecies would be poorly represented, as the vast majority of published pK_as are for the macrospecies. pK_a assignment is made by identifying the highest level where exact matches to the fingerprints are found. In the event that there are multiple matches, the average pK_a value for the highest level fingerprint matches are taken. Typically accuracy was good for level 4 or greater matches, but level 6 or more is needed for substituted aromatic ring systems such as substituted phenol. Note that 48% of the compounds cannot be predicted with level greater than 4, which is a problem for the substituted aromatic compounds and amines, indicating the need to expand the database of 4700 compounds. One advantage of this approach is that when attempting to predict the pK_a for a compound found in the database, the exact value is returned as in a lookup. While the authors mention a 20-fold cross-validation (training the model using 95% of the data and testing on the remaining 5%) no statistics (r^2 or RMSE) for overall performance were provided. At level greater or equal to 5, 16% (4%) of the tests had a mean absolute deviation greater than 0.5 (1.0), indicating high accuracy for the vast majority of compounds and a need to increase the chemical space covered by the database.

Xing also used molecular tree structured fingerprints, but included the number of hits at each level of removal.[137] Like the previous approach, this method allows the individual treatment of specific chemical classes, as it generates tree-structured molecular descriptors for each class. A problem was experienced with an external dataset where four ionizable sites could not be classified because of the specificity trained into the chemical classes. Applying more general rules in their previous work[177] allows the missed ionizable sites to be combined with a similar class of molecules. While this appears to reduce the overall fit of the model on the training data, it shows that generalizations can lead to an improvement in the overall robustness, while refinements lead to improvements in accuracy. Here we would suggest enlarging the training set while retaining the more general classifications in a parallel scheme for background operations. This way one could gain all the improvements of the more specific descriptors without suffering loss of information. A training set consisting of 625 acids and 214 bases had a standard error of 0.41 for acids and 0.30 for bases. Similar tree based methods were also investigated in industry. In 2007, Jelfs et. al. described a method extending the molecular tree structured fingerprints by including 2D substructural fingerprints, which were used to flag the presence of other important structural features that affect pK_a.[123] As before, they found that the molecular trees (circular fingerprints) needed to consider at least five bonds of removal from the central atom for adequate results. This makes sense, for example in the case of the acidic OH group of phenol, where a para-substituent would be five bonds removed or at level 6 in the database lookup method of Kojeg. The software package MoKa implements this concept, where the descriptors are based on molecular interaction fields precomputed on a set of molecular fragments.[141]

2.3.2.6 Decision Trees

Decision tree methods have also been considered recently in pharmaceutical research, both for the prediction of biological activity and for physicochemical property predictions.[90,178,179] The benefits of decision tree methods include: (1) explaining nonlinear response, (2) ease of interpretation, as they provide a clear decision path for better understanding of the test compound, (3) the ability to ignore irrelevant descriptors, (4) the ability to handle large sets of both quantitative and qualitative descriptors, (5) the ability to handle large sets of structurally diverse compounds, and (6) speed. The main drawback with decision tree methods is how to deal with multiple data points for the same molecule, as in the case of polyprotic acids. Other drawbacks include instability and lack of accuracy when compared to other algorithms. While decision trees have commonly been used for classifications, it is possible to derive a regression tree to provide a quantitative rather than qualitative result.[180] Predictive decision trees can be derived through recursive partitioning, which often leads to an unbalanced tree, when smaller groups of compounds having similar property values are filtered out early on, rather than making clever choices that favor a more balanced tree. By defining a pool of both backbone and substituent molecular fragments in terms of highly generalized and specific SMARTS strings, Lee and Crippen were able to iteratively construct a more balanced decision tree where each decision gave weight to the evenness of each split and the reduction in pK_a variance at each child node for a large set of monoprotic molecules.[90] Performance of SMARTS pK_a was competitive with and even exceeded that of several well known applications as described in Table 2.6. It appears that accuracy and stability can be maintained by not heavily relying on highly specific descriptors or by making decisions leading to terminal nodes (where predictions are assigned) close to the root node. As with any other empirical model, regression trees can only be as good as the data used in training, particularly the accuracy and spread of the experimental pK_as, and the chemically diversity of the training set compounds.[90]

2.4 Protein-Ligand Complexes

In rational drug design, one of the ultimate goals is to understand protein-ligand interaction, therefore it is significant to note a recent change in direction for pK_a prediction, which attempts to account for binding effects. Experimental studies have demonstrated that ligand binding induces protonation state changes.[181,182,183,184] Dullweber et al. examined a series of congeneric ligands and identified significant changes in protonation states when binding to thrombin and trypsin.[184] Recently, Czodrowski et al. have developed a method for predicting protein pK_a that also accounts for pK_a shift due to a bound ligand.[44,185,186] Here, the pK_a shifts for the ionizable sites in proteins were calculated with MEAD[187], parameterized with partial charges from a modified version of Gasteiger's PEOE[188] method. As part of the method validation, pK_as were calculated for 132 ionizable groups of 9 proteins (RMSE = 0.88) and showed significant improvements over the null model and FDPB calculations using partial charges from PARSE[189] and CHARMM22[190].

PROPKA 2.0 also attempts to address the issue of pK_a shifts in relation to protein-ligand binding for the ionizable groups of both protein and ligand.[80] The underlying empirical rules of PROPKA 1.0[22] have been modified in PROPKA 2.0 to include the effects of the functional groups of the ligand. The model (or null) pK_a values for the ligand are taken from literature or from MARVIN, when no experimental data is available. In all 26 protein-ligand complexes were studied. Of these, PROPKA 2.0 was shown to identify changes in protonation states that agree with the majority of the experimental data, however no statistics on the pK_a predictions were provided. Clearly, more NMR pK_a data for protein-ligand complexes is required for retraining and validating both models, but it is encouraging to see that the combination of methodologies can provide some promising results.

2.5 Statistics and Benchmarking

Often experimental data for a particular ionizable site is provided as a range. The range is due to estimated experimental error or data curation, where multiple measurements were obtained from different sources possibly using different techniques.

One can only hope that care was taken in the curation, as outliers can significantly extend the range. Using an experimental range leads to a lower RMSE than when single values are used. Typically any predicted value falling within the range has an error of 0, and otherwise the error is the absolute difference between the value and the nearer limit of the range. Ranges are more commonly used when computing statistics for protein pK_a predictors, whereas the small molecule pK_a predictors generally compare to single experimental values in terms of Pearson's correlation coefficient squared, r^2. On the other hand, r^2 is often not considered when comparing experimental to calculated data for proteins, as pK_a data for proteins is dominated by carboxyl groups in aspartic and glutamic acid residues, and by the imidazole group of histidine residues, most of which tend to have pK_as in the range 2 to 6. If only a few lysine or tyrosine residues (pK_a 9 to 11) are added to the comparison, the r^2 increases substantially, while the RMSE will remain close to the same, providing that there is no significant difference in the range of errors for the different residue types. It was for this very reason that Pokala omitted the few available Lys data points from the evaluation, as it exaggerated the null model's apparent accuracy. Including the Lys data points increased the r^2 for the null model from 0.36 to 0.90.[54] When considering r^2, it is better to apply the statistic to the correlation between experimental and predicted pK_a shifts for each residue type separately. Figure 2.1 illustrates how using the correlation coefficient to evaluate data can be misleading. Similarly, it can be worthwhile to consider r^2 for different classes of small molecules to best evaluate the strengths and weaknesses of different pK_a predictors. When comparing methods, it is important that a discriminative benchmark is used.[63]

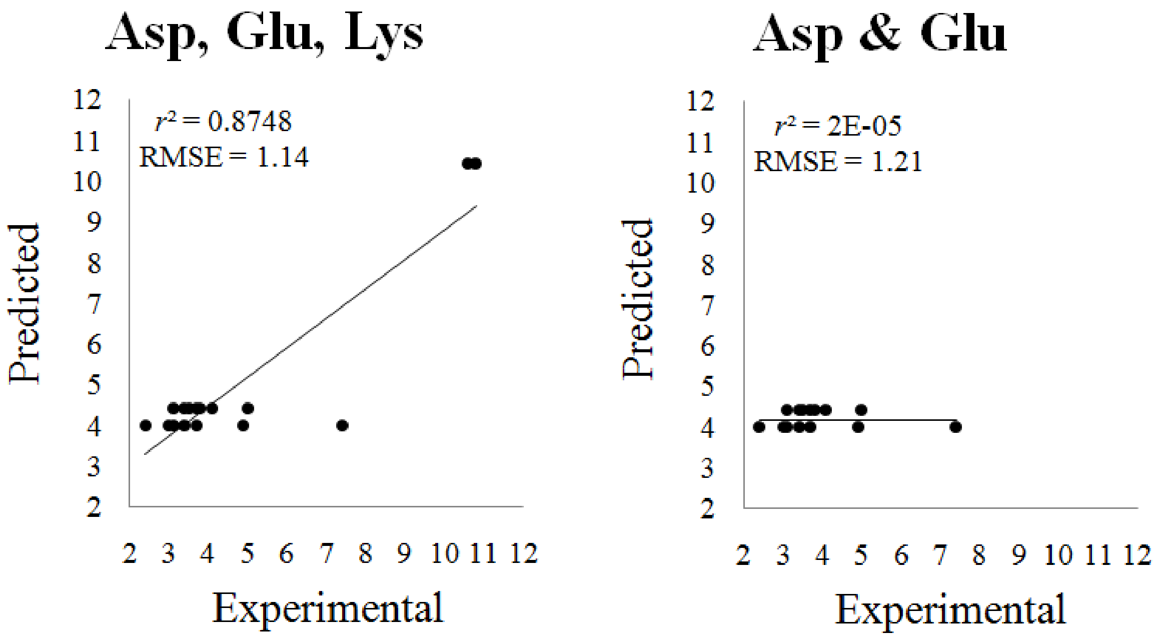


Figure 2.1. Simulation of the null model used to predict the pK_a of 18 residues (9 Asp, 7 Glu, and 2 Lys). One predicted value is assigned to each residue type, although the respective experimental values vary due to environmental effects. Adding in the two Lys data points raises the correlation coefficient (r^2) from nearly zero to 0.87 while improving the RMSE only slightly due to the close fit for the two extra points.

Empirical pK_a predictors are capturing ever increasing attention, given the vast amount of available data. We once again refer to Table 2.4 for a comprehensive list of large data sources. While the data available to the public may be vast, it is by no means comprehensive, as is shown by a survey of several commercial and public small molecule pK_a predictors made by Dearden et. al.[191] Chem Silico provided a 653 molecule test set for the survey and verified that it was not used in training their model, but it could not be verified what portion of the data was external to the other models. Table 2.7 is a reproduction of this survey in hopes that it, along with the broader survey in Table 2.6, can help the reader select the most suitable software. No statistics for training and or validation data were found for Pallas, Pipeline Pilot, and QikProp. Furthermore, the Chem Silico pK_a prediction utility is no longer described or offered on their website. It should also be noted that the only software packages which provided predictions for the entire data set were ADMET Predictor and MARVIN. One other hidden anomaly in the results is that VCCLAB uses pK_a prediction data from ADME Boxes. VCCLAB obtains log$P$, log$S$, and pK_a predictions from ADME Boxes as part of their suite of properties returned by their web utility ALOGPS.[192] For all of the molecules that were predicted in common between ADME Boxes and VCCLAB, the performance was equivalent. Apparently the performance differences are due to either differing SMILES interpretations or a transmission problem between the two web sites.

Table 2.7. The predictive abilities of ten pK_a prediction utilities[191]

Software	url	# Molecules Predicted	r^2	MAE[a]
ADME Boxes	ap-algorithms.com	627	0.959	0.32
VCCLAB	vcclab.org	610	0.931	0.40
ADMET Predictor	simulationsplus.com	653	0.899	0.67
Pipeline Pilot	accelrys.com	626	0.852	0.43
SPARC	ibmlc2.chem.uga.edu/sparc	644	0.846	0.78
MARVIN	chemaxon.com	653	0.778	0.90
QikProp	schrodinger.com	645	0.768	0.93
ACD/Labs	acdlabs.com	644	0.678	1.07
PALLAS	compudrug.com	646	0.656	1.17
CSpK_a	chemsilico.com	642	0.565	1.48
Not surveyed				
ASTER	www.epa.gov			
COSMOtherm	www.commologic.de			
Epik	www.schrodinger.com			
Jaguar	www.schrodinger.com			
MoKa	www.moldiscovery.com			

[a] MAE – mean absolute error.

Benchmarking models is a major issue in literature. To date, no true benchmarks for pK_a exist. Ideally, one would have a universal training set to train all models, and a universal disjoint and similarly diverse test set would be used to test their predictions. Even *ab initio* methods may be based on some small subset of the universal training set. A more practical way to compare two methods would be (1) to examine the fit of both on the intersection of their training sets and then (2) compare their predictions on a test set outside the union of their training sets.

With no true benchmarks for pK_a prediction utilities, the only way to identify a superior model is by trusting the statistics. All empirically based models should have r^2 close to 1.0 and RMSE as close to 0.0 as possible over a wide range of compounds. The statistics for both training and test data should be separate; unfortunately this is not the case with current surveys for both macromolecules[11,68,69] and small molecules,[193] including the one in this review. Consensus models further confuse the issue, as the training sets of all models are considered and the test set of molecules under consideration is more likely to be represented by one or more of those data sets. Statistics on the training data indicate the upper threshold for accuracy, while statistics on the test set (data outside the training set) suggest how the model will perform on data having similar chemical diversity. Therefore, it would be useful to have a diversity statistic based on some chemical property space defined by calculable orthogonal descriptors based on a large chemical database such as PubChem. The other obvious but nonetheless relevant aspect of the training and test sets are their respective sizes. Empirical pK_a models with small training sets and good statistics are either specific to an ionizable group, such as carboxylic acids, or their accuracy cannot be trusted without more rigorous validation. Robust evaluation of predictive models comes down to five factors: the statistics, the range and variance of the property values, and the size and chemical diversity of the test set. As shown in Figure 2.1, clustering of data in chemical property space can drastically affect the statistics, especially leave-some-out cross-validation. When choosing the method(s) keep the following questions in mind. *How chemically diverse is the dataset?*

Testing on substituted phenols alone only indicates how well the pK_a of phenols will be predicted. *What is the range and distribution of experimental values of the external test set?* It is impossible to trust predictions unless validation across the entire pK_a spectrum of the user's desired application is performed. For example, it is useless to only validate pK_a predictions in the range of 10 to 16, if the intended application is pharmaceutical research relating to the oral bioavailability of potential lead candidates. *What is the size of the test set?* Large and diverse test sets not only validate the accuracy of a predictive model, but also its robustness. *What can be learned from the outliers (poor predictions)?* If model A has superior performance on acids and model B has superior performance on bases, it is common sense to use both in their respective areas of strength. If it is not possible to discern a superior group of models, it has been shown that consensus models tend to improve accuracy,[69,90] apparently due to the increased diversity of the combined training sets of the models used in the consensus.

2.6 Conclusions

The main advantage of *in silico* pK_a prediction is that physical samples are not needed. Still, some new compounds surely need to be synthesized for experimental evaluation of physicochemical properties to better understand chemical space and expand the diversity of the molecules available to update existing models and develop new prediction methods. Even when one considers newer methods of high throughput pK_a, there are two limiting factors: the costs and time associated with obtaining or synthesizing the molecules of interest. Hence, there is a need for a quick, accurate, and robust model for pK_a prediction for large as well as small molecular libraries. There is also a need for better benchmarking and comparison of methods. For example, the common belief is that consensus models tend to improve the accuracy of predictions. Here, common sense should throw up red flags. At least with the methods and software discussed above, there is no comprehensive database indicating what molecules were used for training and testing. Therefore, the consensus could be based on a selection of methods where some or all of the training sets included the molecule being evaluated. Statistics obtained from a consensus model may not reflect its performance on new data.

So, which method is best? While *ab initio* quantum mechanical methods are broadly applicable, they are computationally expensive. With regard to small molecules, QM descriptors are easier to calculate, but they suffer some of the same limitations as Hammett based methods, as one needs to first group compounds and establish linear correlations between the descriptors and pK_a. Unlike the Hammett and Taft equation, a σ-like variable needs to be established for each descriptor type for each class of compounds. In order to accurately predict pK_a using QM methods, it is clear that solvation needs to be considered for some if not all compound classes. Any improved accuracy invariably is associated with large computational cost; hence these methods may be impractical when calculating pK_a for large *in silico* molecular databases. On the other hand, QM analysis can identify the sites and order of ionization, which many empirical methods cannot. Furthermore, QM may be used to provide added insight in the analysis of microspecies. It has been shown that QM continuum-solvation methods are still a viable tool for providing predictive regression equations for various chemical classes. While it may not represent the fastest solution to high throughput *in silico* pK_a analysis, it would behoove us to identify a single comprehensive level of DFT and solvation theory such that a single equation could deal with all compound classes.

Again, which method is best? It is impossible to know until true assessments have been made. Being able to fairly assess these models is paramount, but how? Data is limited, fallible, hidden, unorganized, and often found to be conflicting, yet it is the basis of each and every model discussed in this review. These factors are compounded when considering pK_a models for macromolecules. NMR titrations are by far the most difficult, and it is not obvious which shift (^{1}H, ^{13}C, ^{15}N) will provide the most interpretable titration curve. In some cases the titration curves were obtained by measuring the shifts of atoms in the vicinity of the ionizable site, but far removed considering covalent bond connectivity. The main limiting factor in order to improve pK_a prediction is the need for a large quantity of new well curated data for both small and macromolecules. Especially, more data are needed for buried ionizable residues and ionizable residues in active sites of proteins. Steps toward a comprehensive pK_a database have already begun with the PPD, but the data cannot be efficiently downloaded into a text delimited or other common computer readable format such as SD files. There is need for a similar database for experimental pK_a data for small molecules. The initial challenge is to collect and curate all of the freely available information, then offer data to the public in an organized and computer accessible format.

Proprietary data is still an issue, but it has already been seen that Big Pharma has been willing to participate in the assessment of predictive utilities on their own proprietary data sets.[194] These tests, while relevant to the pharmaceutical industry, may not be a fair assessment of the overall predictive quality of the models being tested. The proprietary data sets are likely to contain highly skewed sets of molecules based on the investigation of structure activity relationships. Therefore, it is likely that the proprietary datasets do not represent a well distributed sampling of chemical space, resulting in a less than adequate predictive performance of empirical and semi-empirical methodologies. A robust predictive model is expected to exhibit uniform performance across a defined segment of chemical space. It is possible and in fact desirable that performance will be maintained outside such definitions, but, it cannot be expected.

Accepting a new definition of chemical space, capable of differentiating all known classes of chemical compounds, could serve as a basis for identifying the strengths and weakness of existing physicochemical property prediction utilities. This is important, as different methodologies are likely to demonstrate higher accuracy when accessing molecules in various localized regions of chemical space. Such an analysis would allow researchers to select the optimal combination of predictive models for their specific purposes.

A final note: regardless of the model used, validation on an external test set is necessary. In this regard, we would draw the reader's attention to an article entitled *Beware q^2!*,[195] where leave one out (LOO) cross-validation was explored utilizing *k* nearest neighbors QSARs for three datasets. All-in-all 160 LOO models for each of the datasets were explored, very few of which were found to have desirable statistics for predicting new data, and it was impossible to identify the best LOO model without assessing it on new data. In order to validate that their models did not suffer from overfitting, chance correlations were explored by performing 160 randomizations on each dataset and respectively retraining the dependent variables for each randomization. It was verified that the q^2 for chance correlations were significantly lower than those of the trained models derived from the non-randomized dependent variables. Furthermore, there was little to no correlation between the q^2 of the cross-validated training sets and the respective r^2 on the external test data. The authors concluded that LOO cross-validation could not be used to identify a robust model, nor could it be used to identify the best model for making predictions without validation on an external test set consisting of new data. We again emphasize, the statistics for a model based on fitting training data represents the maximum predictive power for that model and in no way determines how the model will predict new data.

2.7 References

(1) Ullmann, G. M. Relations between protonation constants and titration curves in polyprotic acids: A critical view. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

(2) Prasad, R.; Mahajan, V.; Verma. S.; Gupta, N. Arterial blood gas: Basics and interpretation. *Pulmon* **2007**, *9*, 82–87.

(3) Hoener, B. A.; Benet, L. Z. In *Modern Pharmaceutics*, 3rd ed.; Banker, G. S., Rhodes, C. T., Eds.; Marcel Dekker Inc.: New York, 1996; pp 121-153.

(4) Nielsen, J. E. Analyzing protein NMR pH-titration curves. In *Annual Reports in Computational Chemistry*, 1st ed.; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Vol. 4, pp 89-106.

(5) Quijada, J.; López, G.; Versace, R.; Ramírez, L.; Tasayco, M. L. On the NMR analysis of pK_a values in the unfolded state of proteins by extrapolation to zero denaturant. *Biophys. Chem.* **2007**, *129*, 242–250.

(6) Bartik, K.; Redfield, C.; Dobson, C. M. Measurement of individual pK_a values of acidic residues of hen and turkey lysozymes by two-dimensional ^{1}H NMR. *Biophys. J.* **1994**, *66*, 1180–1184.

(7) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. pK_a Values of carboxyl groups in the native and denatured states of barnase: The pK_a values of the denatured state are on average 0.4 units lower than those of model compounds. *Biochem.* **1995**, *34*, 9424–9433.

(8) Thurkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. Hydrogen bonding markedly reduces the p*K* of buried carboxyl groups in proteins. *J. Mol. Biol.* **2006**, *362*, 594-604.

(9) Protein pK_a Database, http://www.jenner.ac.uk/PPD/ (accessed on June 8, 2009).

(10) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent properties of proteins, *J. Mol. Biol.* **1994**, *238*, 415–436.

(11) Stanton, C. L.; Houk, K. N. Benchmarking p*K* prediction methods for residues in proteins. *J. Chem. Theo. Comp.* **2008**, *4*, 951–966.

(12) Krieger, E.; Nielsen, J. E.; Spronk, C. A. E. M.; Vriend, G. Fast empirical pK_a prediction by Ewald summation. *J. Mol. Graph. Modell.* **2006**, *25*, 481–486.

(13) Nielsen, J. E.; Vriend, G. Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK_a calculations. *Proteins: Struct. Func. Genet.* **2001**, *43*, 403–411.

(14) Edsall, J. T. In *Proteins, amino acids and peptides as ions and dipolar ions.* 1st ed.; Cohn, E. J., Ed.; Reinhold Publishing Corp.: New York, 1943; Chapter 20, pp 444–505.

(15) Nozaki, Y.; Tanford, C. Examination of titration behavior. In *Methods in Enzymology*, 1st ed.; Hirs, C. H. W., Ed.; Academic Press: New York, 1967, Vol. 11, pp 715–734.

(16) Keim, P.; Vigna, R. A.; Morrow, J. S.; Marshall, R. C.; Gurd, F. R. N. Carbon 13 nuclear magnetic resonance of pentapeptides of glycine containing central residues of serine, threonine, aspartic and glutamic acids, asparagine, and glutamine. *J. Biol. Chem.* **1973**, *248*, 7811–7818.

(17) Keim, P.; Vigna, R. A.; Nigen, A. M.; Morrow, J. S.; Gurd, F. R. N. Carbon 13 nuclear magnetic resonance of pentapeptides of glycine containing central residues of methionine, proline, arginine, and lysine. *J. Biol. Chem.* **1974**, *249*, 4149–4156.

(18) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups in proteins. *Prot. Sci.* **2006**, *15*, 1214–1218.

(19) Richarz, R.; Wüthrich, K. Carbon-13 NMR chemical shifts of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers*, **1978**, *17*, 2133–2141.

(20) Creighton, T. E. In *Proteins: Structures and molecular properties,* 2nd ed.; W. H. Freeman and Company: New York 1993; pp 6.

(21) He, Y.; Xu, J.; Pan, X.-M. A statistical approach to the prediction of pK_a values in proteins. *Proteins: Struct. Func. Bioinf.* **2007**, *69*, 75–82.

(22) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins: Struct. Func. Bioinf.* **2005**, *61*, 704–721.

(23) Forsyth, W. R.; Antosiewicz J. M.; Robertson, A. D. Empirical relationships between protein structure and carboxyl pK_a values in proteins. *Proteins: Struct. Func. Genet.* **2002**, *48*, 388–403.

(24) Edgcomb, S. P.; Murphy, P. M. Variability in the pK_a of histidine side-chains correlates with burial within proteins. *Proteins: Struct. Func. Genet.* **2002**, *49*, 1–6.

(25) Fitch, C. A.; García-Moreno, E. B. Structure-based pK_a calculations using continuum electrostatics methods. *Curr. Prot. Bioinf.* **2006**, 8.11.1–8.11.22.

(26) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A; Antosiewicz. J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics Program. *Comp. Phys. Comm.* **1995**, *91*, 57–95.

(27) Bashford, D. Macroscopic electrostatic models for protonation states in proteins. *Front. Biosci.* **2004**, *9*, 1082–1099.

(28) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson–Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recognit.* **2002**, *15*, 377–392.

(29) Bashford, D.; Karplus, M. pK_as of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochem.* **1990**, *29*, 10219–10225.

(30) Yang, A.-S.; Gunner, M. R.; Samponga, R.; Sharp K.; Honig, B. On the calculation of pK_as in proteins. *Proteins: Struct. Func. Genet.* **1993**, *15*, 252–265.

(31) Hill, T. On intermolecular and intramolecular interactions between independent pairs of binding sites in proteins and other molecules. *J. Am. Chem. Soc.* **1956**, *78*, 3330–3336.

(32) Tanford, C.; Kirkwood, J. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(33) Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5250–5244.

(34) Warshel, A. What about protein polarity? *Nature*, **1987**, *330*, 15–16.

(35) Elcock, A. H. Prediction of functionally important residues based solely on computed energetics of protein structure. *J. Mol. Biol.* **2001**, *312*, 885–896.

(36) Ondrechen, M. J.; Clifton, J. G.; Ringe, D. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12473–12478.

(37) Honig, B.; Nicholls A. Classical electrostatics in biology and chemistry. *Science*, **1995**, *268*, 1144–1149.

(38) Barker, P. D.; Mauk, A. G. pH-Linked conformational regulation of a metalloprotein oxidation-reduction equilibrium: Electrochemical analysis of the alkaline form of cytochrome *c*. *J. Am. Chem. Soc.* **1992**, *114*, 3619–3624.

(39) Turano, P.; Ferrer J. C.; Cheesman, M. R.; Thomson, A. J.; Banci, L.; Bertini, I.; Mauk.; A. G. pH, electrolyte, and substrate-linked variation in active site structure of the Trp5l Ala variant of cytochrome *c* peroxidase. *Biochem.* **1995**, *34*, 13895–13905.

(40) Alexov, E. G.; Gunner, M. R. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **1997**, *74*, 2075–2093.

(41) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The determinants of p*K*s in proteins. *Biochem.* **1996**, *35*, 7819–7833.

(42) Demchuk, E.; Wade, R. C. Improving the continuum dielectric approach to calculating p*K*s for ionizable groups in proteins. *J. Phys. Chem.* **1996**, *100*, 17373–17387.

(43) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophys. J.* **2002**, *83*, 1731–1748.

(44) Czodrowski, P.; Dramburg, I.; Sotriffer, A.; Klebe, G. Deelopment, validation, and application of adapted PEOE charges to estimate pK_a values of functional groups in protein-ligand complexes. *Proteins: Struct. Func. Bioinf.* **2006**, *65*, 424–437.

(45) Barth, P; Alber, T; Harbury, P. B. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4898–4903.

(46) Dimitrov, R. A.; Crichton, R. R. Self-consistent field approach to protein structure and stability. I: pH dependence of electrostatic contribution. *Proteins: Struct. Func. Genet.* **1997**, *27*, 576–596.

(47) Warwicker, J. Simplified methods for pK_a and acid pH-dependent stability estimation in proteins: Removing dielectric and counterion boundaries. *Prot. Sci.* **1999**, *8*, 418–425.

(48) Warwicker, J. Improved pK_a calculations through flexibility based sampling of water-dominated interaction scheme. *Protein Sci.* **2004**, *13*, 2793–2805.

(49) Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent calculation of p*K*s of ionizable residues in proteins: Semi-microscopic and microscopic approaches. *J. Phys. Chem. B* **1997**, *101*, 4458–4472.

(50) Sandberg. L.; Edholm, O. A fast simple method to calculate protonation states in proteins. *Proteins: Struct. Func. Genet.* **1999**, *36*, 474–483.

(51) Mehler, E. L.; Guarnieri, F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys. J.* **1999**, *75*, 3–22.

(52) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(53) Kuhn, B.; Kollman, P. A.; Stahl, M. Prediction of pK_a shift in proteins using a combination of molecular mechanical and continuum solvent calculations. *J. Comput. Chem.* ***2004*, *25*,** 1865–1872.

(54) Pokala, N.; Handel, T. M. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **2004**, *13*, 925–936.

(55) Spassov, V. Z.; Yan, L. A fast accurate computational approach to protein ionization. *Prot. Sci.* **2008**, *17*, 1955–1970.

(56) Khandogin, J.; Brooks III, C. L. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochem.* **2006**, *45*, 9363–9373.

(57) Wisz, M. S.; Hellinga, H. W. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins: Struct. Func. Genet.* **2003**, *51*, 360–377.

(58) Karshikoff, A. A simple algorithm for the calculation of multiple site titration curves. *Prot. Eng.* **1995**, *8*, 243–248.

(59) Antosiewicz, J.; Briggs, J. M.; Elcock, A. H.; Gilson, M. K.; McCammon, J. A. Computing ionization states of proteins with a detailed charge model. *J. Comput. Chem.* **1996**, *17*, 1633–1644.

(60) Gibas, C. J.; Subramaniam, S. Explicit solvent models in protein pK_a calculations. *Biophys. J.* **1996**, *71*, 138–147.

(61) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta.* **2006**, *1764*, 1647–1676.

(62) Simonson, T.; Perahia, D. Microscopic dielectric properties of cytochrome c from molecular dynamics in aqueous solution. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.

(63) Schutz, C. N.; Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct. Func. Genet.* **2001**, *44*, 400–417.

(64) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; and Garcia-Moreno, E. B. Experimental pK_a values of buried residues: Analysis with continuum methods and role of water penetration. *Biophys. J.* **2002***, 82,* 3289–3304.

(65) Nielsen, J. E.; Andersen, K. V.; Honig, B.; Hooft, R. W. W.; Klebe, G.; Vriend, G.; Wade, R. C. Improving macromolecular electrostatics calculations. *Protein Eng.* **1999**, *12***,** 657–662.

(66) Alexov, E. G.; Gunner, M. R. Calculated protein and proton motions coupled to electron transfer: Electron transfer from Q_A- to Q_B in bacterial photosynthetic reaction centers. *Biochem.* **1999**, *38*, 8253–8270.

(67) Simonson, T.; Carlsson, J.; Case, D. A. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(68) Kieseritzky, G.; Knapp, E.-W. Optimizing pK_a computation in proteins with pH adapted conformations. *Proteins: Struct. Func. Bioinf.* **2008**, *71*, 1335–1348.

(69) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. Benchmarking pK_a prediction. *BMC Biochem.*[Online] **2006**, *7*, Article 18. http://www.biomedcentral.com/1471-2105/7/18 (accessed Jun 09, 2009).

(70) Vriend G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52–56. http://swift.cmbi.kun.nl/whatif/ (accessed May 27 2009)

(71) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of Poisson–Boltzman equation. *J. Phys. Chem.* **1997**, *101*, 1190–1197.

(72) Warshel, A.; Russell, S. T.; Churg, A. K. Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability. *Proc. Natl. Acad. Sci U.S.A.* **1984**, *81*, 4785–4789.

(73) Rekker, R. F. In *The hydrophobic fragmental constant, its derivation and application: A means of characterizing membrane systems.*; Elsevier Scientific Pub. Co.: New York, 1977.

(74) Rekker, R. F.; Kort, H. M. The hydrophobic fragmental constant: An extension to a 1000 data point set. *Eur. J. Med. Chem.* **1979**, *14*, 479–488.

(75) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.

(76) Onufriev, A; Case, D. A.; Bashford, D. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(77) Spassov, V. Z.; Bashford, D. Multiple-site ligand binding to flexible macromolecules: Separation of global and local conformational change and iterative mobile clustering approach. *J. Comp. Chem.* **1999**, *20*, 1091–1111.

(78) RCSB Protein Data Bank, http://www.rcsb.org/pdb/home/home.do (Accessed on Jul 15, 2009).

(79) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and rationalization of protein pK_a values using QM and QM/MM methods. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

(80) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein-ligand complexes. *Proteins: Struct. Func. Bioinf.* **2008**, *73*, 765–783.

(81) Marosi, A.; Kovács, Z.; Béni, S.; Kökösi, J.; Noszál, B. Triprotic acid-base microequilibria and pharmacokinetic sequelae of cetirizine. *Eur. J. Pharm. Sci.* **2009**, *37*, 321–328.

(82) *Calculator Plugins for structure property prediction, Marvin version 5.1.4*; ChemAxon: Budapest, Hungary, 2009. http://www.chemaxon.com/demosite/marvin/index.html (accessed May 7, 2008).

(83) *ACD/PhysChem Suite, version 12.0*; Advanced Chemistry Development Inc.: Toronto, Ontario, Canada, 2009.

(84) *ADMET Predictor*, version 3.0; Simulations *Plus*, Inc.: Lancaster, CA, 2009.

(85) *MDL CrossFire commander,* V*ersion 7.1*; Elsevier MDL: San Leandro, CA, 2009.

(86) Dean, J. A. In *Lange's Handbook of Chemistry,* 15th ed.; McGraw-Hill: New York, 1999; Chapter 8, pp 8.24–8.72. http://www.knovel.com (accessed Apr 2007).

(87) SPARC Performs Automated Reasoning in Chemistry v4.2. http://ibmlc2.chem.uga.edu/sparc/ (accessed Dec 16, 2008).

(88) Advanced Chemistry Development ACD/Labs Online (I-Lab). http://www.acdlabs.com/ilab/ (accessed May 7, 2008).

(89) *SciFinder Scholar*, version 2007; Chemical Abstract Services: Columbus, OH, 2007.

(90) Lee, A. C.; Yu, J.-Y.; Crippen, G. M. pK_a prediction of monoprotic small molecules the SMARTS way. *J. Chem. Inf. Model.* **2008**, *48*, 2042–2053.

(91) Fraczkiewicz, R. In silico prediction of ionization. In *Comprehensive Medicinal Chemistry II*, Testa, B.; van de Waterbeemd, H. Eds.; Elsevier: Oxford, UK, 2006, Vol. 5, Chapter 25, pp 603–626.

(92) Martell, A. E.; Smith, R. M. In *Critical Stability Constants*, Plenum Press: New York, NY, 1974, Vols. 1–6.

(93) NIST Standard Reference Database 46, version 8.0; National Institute of Standards and Technology: Gaithersburg, MD, 2009. http://www.nist.gov/srd/nist46.htm (Accessed 07/15/2009).

(94) Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comp. Sci.* **1992***, 32*, 244–255.

(95) Szabadváry, F. Electrometric Analysis. In *History of Analytic Chemistry*, 1st English ed.; Belcher, R.; Gordon, L.; Pergamon Press: Long Island City, New York, 1966; Vol. 26, pp 375–387.

(96) Barbosa, J.; Barrón, D.; Jiménez-Lozano, E.; Sanz-Nebot, V. Comparison between capillary electrophoresis, liquid chromatography, potentiometric and spectrophotometric techniques for evaluation of pK_a values of zwitterionic drugs in acetonitrile–water mixtures. *Anal. Chim. Act.* **2001**, *437*, 309–321.

(97) Kim, H.-s.; Chung, T. D.; Kim, H. Voltametric determination of the pK_a of various acids in polar aprotic solvents using 1,4-benzoquinone. *J. Electroanal. Chem.* **2001**, *498*, 209–215.

(98) Ishihama, Y.; Oda, Y.; Asakawa, N. Microscale determination of dissociation constants of multivalent pharmaceuticals by capillary electrophoresis, *J. Pharm. Sci*. **1994***, 83*, 1500–1507.

(99) Kolthoff, I. M. Conductometric titrations. *Ind. Engin. Chem.* **1930**, *2*, 225–230.

(100) Niazi, S. Dissociation, Partitioning, and Solubility. In *Handbook of Preformulation: Chemical, Biological, and Botanical Drugs*, 1st ed.; Informa Healthcare: New York, 2007; pp 112–115.

(101) Beckers, J. L.; Everaerts, F. M.; Ackermans, M. T. Determination of absolute mobilities, p*K* and separation numbers by capillary zone electrophoresis. Effective mobility as a parameter for screening. *J. Chromatogr. A* **1991**, *537*, 407–428.

(102) Cai, J.; Smith, T.; Rassi, Z. E. Determination of the ionization constants of weak electrolytes by capillary zone electrophoresis. *J. High Resolut. Chromatogr.* **1992**, *15*, 30–32.

(103) Cleveland Jr., J. A.; Benko, M. H.; Gluck, S. J.; Walbroehl, Y. M. Automated pK_a determination at low solute concentrations by capillary electrophoresis. *J. Chromatogr. A* **1993**, *652*, 301–308.

(104) Kaliszan, R.; Wiczling, P.; Markuszewski, M. J. pH gradient high-performance liquid chromatography: Theory and applications. *J. Chromatogr. A* **2004**, *1060*, 165–175.

(105) Wiczling, P; Markuzewski, M. J.; Kaliszan, R. Determination of pK_a by pH gradient reversed-phase HPLC. *Anal. Chem.* **2004**, *76*, 3069–3077.

(106) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. **2006**, *2*, 139–155.

(107) Avdeef, A; Comer, J. E. A; Thomson, S. J. pH–metric log *P*. 3. Glass electrode calibration in methanol–water applied to pK_a determination of water–insoluble substances. *Anal. Chem.* **1993**, *65*, 42–49.

(108) Rosés, M; Rived, F.; Bosch, E. Dissociation constants of phenols in methanol–water mixtures. *J. Chrom. A.* **2000**, *867*, 45–56.

(109) Ruiz, R.; Ràfols, C.; Rosés, M.; Bosch, E. A potentially simpler approach to measure pK_a of insoluble basic drugs containing amino groups. *J. Pharm. Sci.* **2003**, *92*, 1473–1481.

(110) Harris, J. C.; Hayes M. J. Acid dissociation constant. In *Handbook of Chemical Property Estimation Methods.* Lyman, W. J.; Reehl W. F.; Rosenblatt D. H. Eds.; McGraw-Hill, Inc.: New York, 1982, pp 6.1-6.28.

(111) Chen, J.-F.; Xia, Y.-X.; Choppin, G. R. Derivative analysis of potentiometric titration data to obtain protonation constants. *Anal. Chem.* **1996**, *68*, 3973–3978.

(112) Goss, K.-U.; Bronner, G.; Harner, T.; Hertel, M.; Schmidt, T. C. Partition behavior of fluorotelomer alcohols and olefins. *Environ. Sci. Technol.* **2006**, *40*, 3572–3577.

(113) Goss, K.-U. The pK_a values of PFOA and other highly fluorinated carboxylic acids. *Environ. Sci. Technol.* **2008**, *42*, 456–458.

(114) Lee, L. S.; Bellin, C. A.; Pinal, R.; Rao, P. S. C. Cosolvent effects on sorption of organic acids by soils from mixed-solvents. *Environ. Sci. Technol.* **1993**, *27*, 165–171.

(115) Clark, J.; Perrin, D. D. Prediction of the strengths of organic bases. *Q. Rev. Chem. Soc.* **1964**, *18*, 295–320.

(116) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. In *pK$_a$ Prediction for Organic Acids and Bases*; Chapman & Hall: New York, 1981.

(117) Livingstone D. J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003**, *3*, 1171–1192.

(118) Klopman, G.; Fercu, D. Application of the multiple computer automated structure evaluation methodology to a quantitative structure–activity relationship study of acidity. *J. Comput. Chem.* **1994**, *15*, 1041–1050.

(119) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput. Aid. Mol. Des.* **2007**, *21*, 681–691.

(120) EPIK, version 2.0109; Schrödinger LLC: New York, NY, 2009. http://www.schrodinger.com (Accessed 05/18/2009).

(121) Hilal, S. H.; Karickhoff, S. W. A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK_as. *Quant. Struct. Act. Relat.* **1995**, *14*, 348–355.

(122) Lee, P. H.; Ayyampalayam, S. N.; Carreira, L. A.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. In silico prediction of ionization constants of drugs. *Mol. Pharm.* **2007**, *4*, 498–512.

(123) Jelfs, S; Ertl, P.; Selzer, P. Estimation of pK_a for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.

(124) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.* **1993**, *14*, 1460–1467.

(125) Soriano, E.; Cerdán, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theo)* **2004**, *684*, 121–128.

(126) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.

(127) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pK_a prediction method using GRID molecular interaction fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(128) Gieleciak, R.; Polanski, J. Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: Application for modeling benzoic acid pK_a values. *J. Chem. Inf. Model.* **2007**, *47*, 547–556.

(129) Ghasemi, J.; Saaidpour, S.; Brown, S. D. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct. THEO.* **2007**, *805*, 27–32.

(130) CS_prpKa; ChemSilico: Tewksbury, MA, 2008. http://www.chemsilico.com/CS_prpKa/PKAexp.html (accessed Mar 11, 2008 – no longer available).

(131) Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water. *Monatsh. Chem.* **2009**, *140*, 15–27.

(132) Schüürmann, G.; Cossi, M.; Barone, V.; Tomasi, J. Prediction of the pK_a of carboxylic acids using the *ab initio* continuum-solvation model PCM-UAHF, *J. Phys. Chem. A* **1998**, *102*, 6706–6712.

(133) Eckert, F.; Klamt, A. Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19.

(134) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First principles calculations of aqueous pK_a values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK_a scale. *J. Phys. Chem. A* **2003**, *107*, 9830–9386.

(135) Pompe, M. Variable connectivity index as a tool for solving the 'anti-connectivity' problem. *Chem. Phys. Lett.* **2005**, *404*, 296–299.

(136) Pompe, M.; Randić, M. Variable connectivity model for determination of pK_a values for selected organic acids. *Acta. Chim. Slov.* **2007**, *54*, 605–610.

(137) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(138) Kogej, T.; Muresan, S. Database mining for pK_a prediction. *Curr. Drug Discov. Tech.* **2005**, *2*, 221–229.

(139) Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1166–1068.

(140) Kim, K. H.; Martin, Y. C. Substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substituted benzoic acids. *J. Org. Chem*. **1991**, *56*, 2723–2729.

(141) *MoKa*, version 1.1; Molecular Discovery Ltd: Middlesex, UK, 2009 http://www.moldiscovery.com (accessed June 09, 2009).

(142) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK_as) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem*. **1991**, *34*, 2056–2060.

(143) Gargallo, R.; Sotriffer, C. A.; Liedl, K. R.; Rode, B. M. Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: Proton affinities and pK_a prediction for nucleic acids components. *J. Comput.-Aided Mol. Des*. **1999**, *13*, 611–623.

(144) Tehan, B. G.; Lloyd, E. L.; Wong, M. G.; Pitt, W. R.; Montana, J. G. ; Manallack, D. T.; Gancia, E. Estimation of pK_a using semiempirical molecular orbital methods. Part 1: Application to phenols and carboxylic acids. *Quant. Struct. Act. Relat.* **2002**, *21*, 457–472.

(145) Tehan, B. G.; Lloyd, E. L.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pK_a using semiempirical molecular orbital methods. Part 2: Application to amines, anilines, and various nitrogen containing heterocyclic compounds. *Quant. Struct. Act. Relat.* **2002**, *21*, 473–485.

(146) Seybold, P. G. Analysis of the pK_as of aliphatic amines using quantum chemical descriptors. *Int. J. Quant. Chem.* **2008**, *108*, 2849–2855.

(147) Liu, S.; Pedersen, L. G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J. Phys. Chem. A* **2009**, *113*, 3648–3655.

(148) Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J. S.; Politzer, P. Comparison of quantum chemical parameters and Hammett constants in correlating pK_a values of substituted anilines. *J. Org. Chem.* **2001**, *66*, 6919–6925.

(149) Ma, Y.; Gross, K. C.; Hollingsworth, C. A.; Seybold, P.; Murray, J. S. Relationships between aqueous acidities and computed surface-electrostatic potentials and local ionization energies of substituted phenols and benzoic acids. *J. Mol. Model.* **2004**, *10*, 235–239.

(150) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a prediction using group philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

(151) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124,* 6421–6427.

(152) Pliego, J. R.; Riveros, J. M. Theoretical calculation of pK_a using cluster-continuum model. *J. Phys. Chem. A* **2002**, *106*, 7434–7439.

(153) Jaguar; version 4.2; User Guide; Schrödinger LLC: New York, NY, 1991-2000. http://yfaat.ch.huji.ac.il/jaguar-help/manTOC.html (Accessed May 18, 2009).

(154) Potter, M. J.; Gilson, M. K.; McCammon, J. A. Small molecule pK_a prediction with continuum electrostatics calculations. *J. Am. Chem. Soc.* **1994**, *116*, 10298–10299.

(155) Szegezdi, J.; Csizmadia, F. New method for pK_a estimation. Proceedings of the eCheminformatics 2003 - Virtual Conference and Poster Session, Zeiningen, Switzerland, 2003; Hardy, B., Ed.; Douglas Connect: Zeiningen, Switzerland, 2003.

(156) ADME/Tox WEB, version 3.5; Pharma Algorithms: Toronto, ON Canada, 2008 http://pharma-algorithms.com/webboxes/ (accessed July 9, 2008).

(157) Bickmore, B. R.; Rosso, K. M.; Tadanier, C. J.; Bylaska, E. J.; Doud, D. Bond-valence methods for pK_a prediction. II. Bond-valence, electrostatic, molecular geometry, and solvation effects. *Geochim. Cosmochima. Acta* **2006**, *70*, 4057–4071.

(158) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Subramanian, V.; Chattaraj, P. K. Intermolecular reactivity through the generalized philicity concept. *Chem. Phys. Lett.* **2004**, *394*, 225–230.

(159) Brinck, T.; Murray, J. S.; Politzer, P.; Carter, R. E. A relationship between experimentally determined pK_as and molecular surface ionization energies for some azines and azoles. *J. Org. Chem.* **1991**, *56*, 2934–2936.

(160) Brinck, T.; Murray, J. S.; Politzer, P. Relationships between the aqueous acidities of some carbon, oxygen, and nitrogen acids and the calculated surface local ionization energies of their conjugate bases. *J. Org. Chem.* **1991**, *56*, 5012–5015.

(161) Brinck, T.; Murray, J. S.; Politzer, P. Molecular surface electrostatic potentials and local ionization energies of group V – VII hydrides and their anions: Relationships for aqueous and gas-phase acidities. *Int. J. Quant. Chem.* **1993**, *48*, 73–88.

(162) Parr. R. G.; Szentpaly, L. V.; Liu, S. J. Electrophilicity index. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.

(163) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.

(164) Toth, A. M.; Liptak, M. D.; Phillips, D. L.; Shields, G. C. Accurate relative pK_a calculations for carboxylic acids using complete basis set and Gaussian-*n* models combined with continuum solvation methods. *J. Chem. Phys.* **2001**, *114*, 4595–4606.

(165) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(166) *Gaussian 98,* revision A.6; Gaussian, Inc.: Pittsburg, PA, 1998.

(167) Klamt, A.; Schüürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans. 2* **1993**, 799–805.

(168) Klamt, A. Conductor-like screening model for real solvents: A new approach to quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(169) Klamt, A.; Jonas, V.; Brürger, T.; Lohrenz, C. W. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A*. **1998**, *102*, 5074–5085.

(170) Ho, J.; Coote, M. pK_a calculation of some biologically important carbon acids - An assessment of contemporary theoretical procedures. *J. Chem. Theory Comput.* **2009**, *5*, 295–306.

(171) Zupan, J.; Gasteiger, J. In *Neural Networks for Chemists; An Introduction*; VCH Publishers: New York, NY, 1993.

(172) Zupan, J.; Gasteiger, J. In *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Verlag GmbH, Weinheim, FRG, 1999.

(173) Hagan, M. T.; Demuth, H. B.; Beale, M. In *Neural Network Design*; PWS: Boston, MA, 1996. (Chapters 1–4: http://hagan.ecen.ceat.okstate.edu/nnd.html accessed June 7, 2009.)

(174) Khayamian, T.; Kardanpour, Z.; Ghasemi; J. A new application of PC-ANN in spectrophotometric determination of acidity constants of PAR. *J. Braz. Chem. Soc.* **2005**, *16*, 1118–1123.

(175) Sayle, R. Physiological ionization and pK_a prediction. Metaphorics LLC. 2000 http://www.daylight.com/meetings/emug00/Sayle/pkapredict.html (accessed May 18, 2009).

(176) Glen, R. C.; Bender, A; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *iDrugs*, **2006**, *9*, 199–204.

(177) Xing, L.; Glen, R. C. Novel methods for the prediction of log$P$, pK_a, and log$D$. *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 796–805.

(178) Blower, P. E.; Cross, K. P. Decision tree methods in pharmaceutical research. *Curr. Top. Med. Chem.* **2006**, *6*, 31–39.

(179) Lee, A. C.; Shedden, K.; Rosania, G. R.; Crippen, G. M. Data mining the NCI60 to predict generalized cytotoxicity. *J. Chem. Inf. Model.* **2008**, *48*, 1379–1388.

(180) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. In *Classification and Regression Trees*; Wadsworth Inc.: Belmont, CA, 1984; Chapter 8, pp 216–265.

(181) Yamazaki, T. NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.

(182) Wang, Y. X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia D. A. Solution NMR evidence that the HIV-1

protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochem.* **1996**, *35*, 9945–9950.

(183) Singer, A. U.; Forman-Kay, J. D. pH titration studies of an SH2 domain-phosphopeptide complex: Unusual histidine and phosphate pK_a values. *Prot. Sci.* **1997**, *6*, 1910–1919.

(184) Dullweber, F.; Stubbs, M. T.; Musil, Đ.; Stürzebecher, J.; Klebe, G. Factorising ligand affinity: A combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.* **2001**, *313*, 593–614.

(185) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation upon ligand binding to trypsin and thrombin: Structural interpretation based on pK_a calculations and ITC experiments. *J. Mol. Biol.* **2007y**, *367*, 1347–1356.

(186) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Atypical protonation states in the active site of HIV-1 protease: A computational study. *J. Chem. Inf. Model.* **2007**, *47*, 1590 – 1598.

(187) Bashford, D. Scientific computing in object-oriented parallel environments. In *Lecture Notes in Computer Science*, 1st ed.; Ishikawa, Y.; Oldehoeft, R. R.; Reynders, J. V. W.; Tholburn, M. Eds.; Springer, New York, NY, 1997; Vol. 1343, pp 233–240.

(188) Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. Tetrahedron 1980;36:3219–3228.

(189) Sitkoff, D.; Ben-Tal, N.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(190) MacKerell Jr.; A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr.; R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Reiher III, W.E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wirkiewicz-Kuczera, D.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(191) Dearden, J. C.; Cronin, M. T. D.; Lappin, D. C. A comparison of commercially available software for the prediction of pK_a. *J. Pharm. Pharmacol.* **2007**, *59*, A7.

(192) Virtual Computational Chemistry Lab. http://www.vcclab.org/ (Accessed July 9, 2008).

(193) Meloun, M.; Bordovská, S. Benchmarking and validating algorithms that estimate pK_a values of drugs based on their molecular structures. *Anal. Bioanal. Chem.* **2007**, *389*, 1267–1281.

(194) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log *P* methods on more than 96000 compounds. *J. Pharm. Sci.* **2009**, *3*, 861–893.

(195) Golbraikh, A.; Tropsha, A. Beware q^2! *J. Mol. Graph. Modell.* **2002**, *20*, 269–276.

Chapter 3

Chemical Data Mining of the NCI Human Tumor Cell Line Database

3.1 Introduction

The NCI Developmental Therapeutics Program Human Tumor cell line data set is a publicly available database that contains cellular assay screening data for over 40000 compounds tested in 60 human tumor cell lines. The database also contains microarray assay gene expression data for the cell lines, and so it provides an excellent information resource particularly for testing data mining methods that bridge chemical, biological, and genomic information. Here we describe a formal knowledge discovery approach to characterizing and data mining this set and report the results of some of our initial experiments in mining the set from a cheminformatics perspective.

Since 1990, National Cancer Institute Developmental Therapeutics Program (DTP) has been screening compounds against a panel of 60 human tumor cell line assays. The results are available on the DTP Web site.[1] Approximately 10000 compounds are screened each year, and at the time of writing, results were available for 44653 compounds including growth inhibition (GI_{50}), lethal dose (LD_{50}), and total growth inhibition (TGI). The untreated cell lines have also been run through microarray assays, yielding gene expression information.

The tumor cell line data set is interesting in several ways relating to current research in finding biomarkers that cross different kinds of data and in using chemical, biological, and genomic information together. First, it provides a well curated set of tumor-related cellular assay screening results for a large number of compounds (the 60 cell lines include melanomas, leukemias, and cancers of the breast, prostate, lung, colon, ovary, kidney, and central nervous system[2]), which can be considered as a surrogate for high-throughput screening data. Second, the gene expression profiles of untreated cell lines allow some level of integration of genomic information with chemical and biological information. Third, the program is ongoing and so the tumor cell line data set is continually growing, but the cell lines themselves are stable (both in terms of number and comparability of results). Fourth, and most importantly, the data are made freely available through the DTP Web site and are thus available for research and publication.

A substantial amount of research on the tumor cell line data set has been carried out locally at the NCI laboratories including development of the COMPARE algorithm[3,4] which measures similarity between vectors of screening results of compounds using a Pearson correlation coefficient. A searching program based on COMPARE is available online.[5] Zaharevitz et al.[4] cite several examples of the successful application of these approaches in drug discovery projects. The original authors of COMPARE also introduced the use of the *mean graph*[3] that gives a visual bar graph representation of the difference between the screening result for a particular compound and the mean for all compounds, across the 60 cell lines. This representation has been widely used alongside COMPARE.

Other research has used neural networks[6] to classify compounds in the set. In their 2000 paper, Scherf et al.[7] examine correlations between compounds' high-throughput screening results (the activity pattern set) and mRNA expression levels. Recently, Rabow et al.[8] performed a clustering of the tumor cell line data set based on the activity profiles, using a self-organizing map (SOM). Other work at the NCI focused on ellipticine analogs and the potential relationship between the mechanism of action and the 60 cell line activity profiles. The compounds were grouped using hierarchical clustering, and a significant difference in activity profiles was found for groups with different mechanisms of action[9] which led to a follow-up QSAR study.[10]

Researchers at Leadscope Inc. have applied their Leadscope software[11] to relate the information in the tumor cell line data set to structural feature analysis of the DTP compounds, including analysis similar to that done by Scherf[12] and correlations of chemical structural features of cytotoxic agents with gene expression data.[13] Blower et al.[14] also applied a three-stage pipeline to the data set, including filtering for drug-likeness, structure alerts, promiscuity and diversity; structural feature based classification using a variant of Recursive Partitioning (requiring separation of actives and inactives) and organization based on hierarchical clustering; and SAR analysis through R-group assembly, macrostructure assembly, and predictive models. The researchers found a close match between classifications and clusters found by Leadscope and manual classifications previously identified at NIH.

Recently, Richter et al.[15] have evaluated an activity prediction model based on both structural information and genomic information, and at Bristol-Myers Squibb, a version of recursive partitioning derivative was applied.[16] Fang et al.[17] developed a set of Internet-based tools that permit correlations to be found between the activity profiles, gene expression profiles, and compounds using COMPARE as well as Spearman & Kendall correlation coefficients and a p-test to indicate significance of correlation results.

In this work, we have focused on characterizing the compounds present in the data set and applying a variety of methods to discover relationships between the compounds and the biological activity values. We have tried to take a more formal approach to data mining, such as has been applied in other domains where large volumes of information need to be searched for important associations. *Data Mining*, and more generally *Knowledge Discovery in Databases (KDD*), is an area of computer science that has attracted a significant amount of research, industry, and media attention in the past decade, as the amount and complexity of information in databases has increased. Many KDD techniques, such as cluster analysis and decision trees, are already well established in chemical and bioinformatics, while others, such as data cleaning and pattern verification and discovery, are less widely applied.

3.2 Principles and Practices of Knowledge Discovery in Databases

KDD is usually defined as the process of identifying *valid*, *novel, potentially useful*, and ultimately *understandable* patterns from large collections of data. At an abstract level, it is concerned with the development of methods and techniques for making sense of data. Since its debut in 1989, KDD has become the most rapidly growing field in the database community and was soon adopted in other business and scientific areas, such as marketing, fraud detection, and bioinformatics. In practice, this field covers techniques often applied in cheminformatics including cluster analysis, machine learning, and visualization techniques. Several KDD models have been proposed in the past decade. For the discussion in this paper, we adopt the 7-step KDD process presented in the most popular data mining textbook by Han and Kamber:[18] data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.

Knowledge Discovery goals are defined by the intended use of the system. Goals may be *verification goals*, in which the system is limited to verifying users' hypotheses, or *discovery goals,* in which the system is required to autonomously find entirely new patterns. Discovery goals may be *descriptive* (requiring characterization of general properties of the data in the database) and *predictive* (requiring predictions to be made using the data in the database).

Discovery goals are generally achieved through *data mining*. Data mining involves fitting models to, or determining patterns from, observed data. Model fitting may be stochastic or deterministic, although stochastic approaches are the most frequently used.

The first task of data mining is concept description. A concept is a labeling of a collection of data, such as labeling a set of "graduate students," "best-seller books," etc. The goal of concept description is to summarize the data of the class under study in general terms (*data characterization*) and to provide a description comparing two or more collections of data (*data discrimination*). Several methods have been proposed for efficient data summarization and discrimination. For example, a data cube[19] can be used for user controlled data summarization among concept hierarchies; analytical characterization can be used for unsupervised data generalization and characterization. After concept description, classification may be applied. The purpose of data classification is to find a set of models that describes and distinguishes data classes or concepts. Usually, finding such models is not the ultimate goal but rather the first step of using such models to predict the class of objects whose class is unknown or to predict future data trends. Decision trees are one of the most popular methods for data classification and predication.

In addition to classification, unsupervised clustering may be applied. The goal of cluster analysis is to examine data objects without consulting known class labels and is generally used as a way of organizing the database. In cluster analysis, objects are grouped based on *maximizing the intraclass similarity* and *minimizing the interclass similarity*. An excellent overview of clustering in cheminformatics is given by Downs and Barnard.[20] Popular clustering algorithms used in data mining include partitioning methods such as k-means,[21] k-mediods,[22] and CLARANS[23] algorithm; hierarchical methods such as agglomerative and divisive algorithms, BIRCH[24] algorithm, CURE[25] algorithm, and Chameleon[26] algorithm; density-based methods such as DBSCAN,[27] OPTICS,[28] and DENCLUE;[29] grid-based methods such as STING,[30] WaveCluster,[31] and CLIQUE;[32] and model-based methods such as classification trees and neural networks. It is interesting to note that there is only limited overlap between the methods popularly applied in cheminformatics and those applied in the data mining community as a whole.

Finally, association analysis may be applied. The goal of association analysis is the discovery of association rules showing attribute value conditions that occur together frequently in a given set of data. The Apriori[33] algorithm family has variants that are suitable for various data types and database models. Combining the association analysis and concept hierarchies, one may generalize the association rules with ISA relationship or various aggregations on different granularities.

Raw data are often not suitable for data mining, due to noise, missing or inconsistent data points, or lack of normalization across data sources. Preprocessing must therefore be applied. The purpose of *data cleaning* is to fill in incomplete data, smooth out noise, and correct inconsistencies. Data may be incomplete when attributes of interest are missing. Approaches for filling missing values include ignoring entries with missing values, filling missing values manually, using a global constant, using the attribute mean to fill in missing values, using the attribute mean for all samples belonging to the same class as the given entry, using the most probable value, and so on. *Noisy data* usually refers to data that contain errors or outlier values that deviate from the expected values. Approaches for noise elimination include the following: binning (smoothing a sorted data value by consulting its neighborhood), clustering (clustering data to detect and eliminate outliers), hybrid methods combining computer and human inspection, and regression (fitting the data to a function). *Inconsistent data* may be the result of errors that happen during data entry or due to the heterogeneous nature of data. The first usually needs to be handled manually. The inconsistency and data redundancy caused by heterogeneous data resources are usually handled in the data integration process.

Data integration and transformation are needed when data from heterogeneous resources are merged and transformed into forms appropriate for mining. In the data integration process, ontology is usually used for schema integration. Additional attention is needed to detect and resolve data value conflicts, such as attributes representing the same concept but using different units. Data transformation techniques include smoothing – removing the noise from data, aggregation, generalization – low level data are replaced by high level concepts, normalization – attribute data are scaled to fall within a small specific range and attribute construction – construct a new attribute to help mining.

Besides precision, performance is another important issue in data mining. The purpose of data selection is to obtain a data representation that is much smaller, yet closely maintains the integrity of the original data. Data reduction is the most common practice used in data selection. Many strategies have been proposed for data reduction: (1) Data cube aggregation,[19] where aggregation operations are applied to the data in the construction of a data cube. (2) Dimension reduction, where irrelevant, weakly relevant, and redundant attributes or dimensions are removed. The most popular dimension reduction algorithms are stepwise forward selection, stepwise backward elimination, hybrid (combination of forward selection and backward selection), and decision tree induction. (3) Data compression, where encoding is used to reduce the data size. Techniques include wavelet transformation, principal components analysis, etc. (4) Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations such as parametric models, by regression and log-linear models, histograms, clustering, or sampling. (5) Discretization and concept hierarchy generation where raw data values for attributes are replaced by ranges or higher conceptual levels.

A data mining system can generate thousands or even millions of clusters, classes, patterns, and rules. Not all of them are interesting to all users. The measurement of the "interestingness" of a pattern is subjective. Typically, a pattern is considered interesting if it is novel, valid with some degree of certainty, potentially useful, and easy to understand. It is unrealistic to expect a data mining system to generate all interesting patterns or only interesting patterns. This makes the measuring of pattern interestingness an essential component in KDD. A desirable feature of any data mining system is the development of a proper measurement model for a given field or user group and the use of it not only after all patterns are detected but also in the process of data mining as a guide for pruning uninteresting patterns and to speed up the mining process.

The data mining results, whether they are clusters or association rules, need to be presented to users (who usually are in the area of applications and are not database or data mining experts) before they can be deployed. Visualization and knowledge representation techniques are required to present the mining result to users, to improve the understandability. This is especially important for supervised mining tasks, where the user's involvement is required in the mining process.

3.3 Data Cleaning, Integration, Selection, and Transformation

At the time of writing the tumor cell line data set contained 257547 compounds in total. Among those compounds, 44653 compounds have cell line screening data (GI_{50}, LC_{50}, TGI data), and the total number of cell lines is 159, although only 60 of those cell lines have gene expression data. The gene expression data consist of 961 gene expression values for each cell line.[23] For the experiments reported here, we implemented a local version of the database containing the 44653 compounds, screening results and gene expression values using PostgreSQL along with the gNova CHORD extension to allow chemical searching and generation of fingerprint bits.[34] 166-Bit structural key fingerprints were produced with gNova, based on a SMARTS-based interpretation of the public MACCS key set available from MDL.[35]

3.3.1 Characterization of the Chemical Compounds

There are several well-established methods of characterizing compounds by chemical properties or structural features. We applied two methods to characterize the compounds: first, calculation and profiling of predicted property values compared to two other well-established data sets, and second, a 2D fingerprint based structural feature comparison with compounds in one of the data sets.

In our first experiment, we chose three compound data sets for comparison to the tumor cell line set. The first is the FDA's Maximum Recommended Therapeutic Dose (MRTD) set containing 1220 current prescription drugs available in SMILES format from the FDA Web site.[36] We chose this set as a representative of current marketed drugs. The second two sets were randomly selected 40000 compound subsets of PubChem, a freely available chemical database,[37] used as representatives of a diverse set of chemical structures. We calculated properties (Molecular Weight, XLogP, Polar Surface Area, and Numbers of Hydrogen Bond Donors and Acceptors) for all of the structures in the data sets using OpenEye FILTER[38] and then generated property distribution plots for each of the properties for each of the data sets. These profiles can be seen in Figure 3.1. The most striking result is that the profiles for the tumor cell line set are very similar to those for the MRTD set, indicating that the compounds in the tumor cell line set are very "drug-like". The noticeably different (but consistent) profiles for the two PubChem subsets indicate that the compounds in PubChem are more diverse.

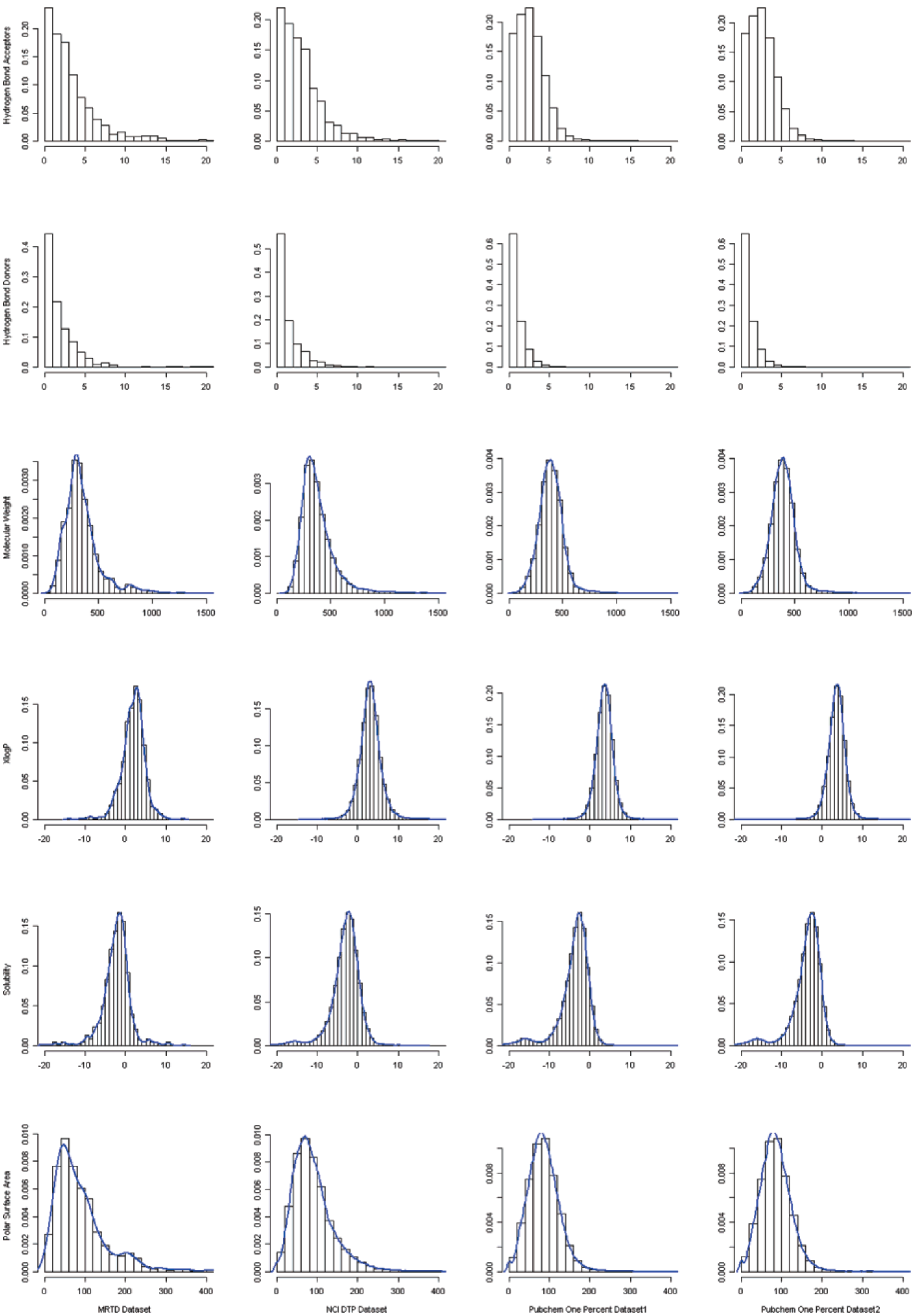


Figure 3.1. Comparative distribution of various properties for the compounds in the MRTD set (first column), tumor cell line set (second column), and two PubChem subsets (third and fourth columns).

In our second experiment we compared the similarity of the drug compounds in the MRTD with the most similar compounds in the tumor cell line set: the distribution of the Tanimoto similarity values of the 166-bit fingerprints is shown in Figure 3.2. Overall 29% of the compounds in the MRTD set have a counterpart in the tumor cell line set with similarity greater than 0.8.

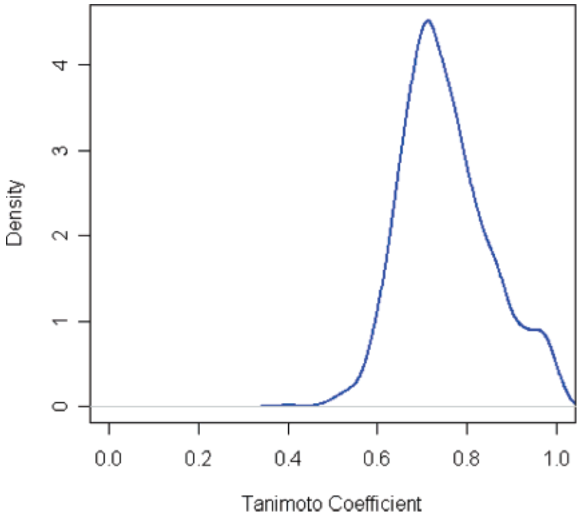


Figure 3.2 Distribution of Tanimoto similarity values (*x*-axis) between compounds in the MRTD set and the most similar compound for each in the tumor cell line set.

3.3.2 Characterization of the Cell Line Screening Growth Inhibition Values

We then went on to examine the distribution of the –log GI_{50} data points (henceforth referred to as growth inhibition values) across cell lines and compounds. First, it is important to note that there is missing data: overall 12.1% of the cell line screen data points are missing. Figure 3.3 shows the percentage of compounds with missing data for each cell line. Only 2696 compounds (6%) have the growth inhibition values for all the 60 cell lines.

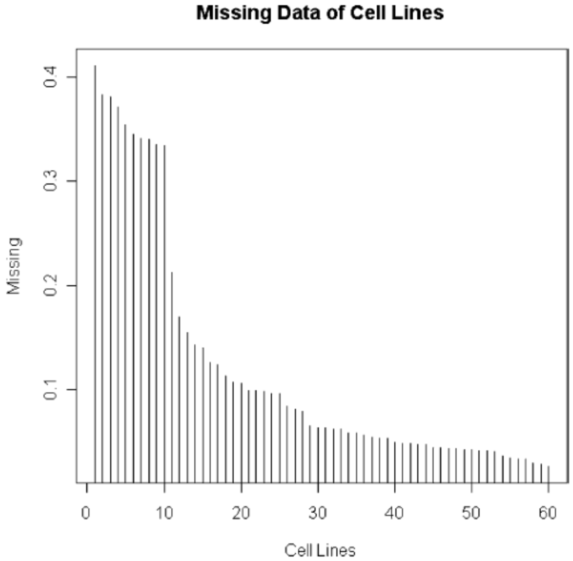


Figure 3.3 Fraction of the compounds with missing data for each of the 60 cell lines.

Growth inhibition values at or near 4.0 indicate inactivity of compounds (i.e., doses of less than 10^{-4} molar did not inhibit growth). Overall 44.9% of growth inhibition values are equal to 4.0 (see Figure 3.4 for the distribution across cell lines). When these compounds are removed from the set, a normal distribution can be seen with a peak of values less than 5.0, indicating inactive or extremely weakly active compounds. Based on this data distribution, we decided for our experiments to set the cutoff for determining whether a compound was active or inactive at 5.0: we consider the data which are less than 5 as inactive (set as 0) and the data which are greater or equal to 5 as active (set as 1). Overall, 19.6% compounds are considered active using this cutoff. The percentage of compounds considered "active" using this cutoff for each of the 60 cell lines is shown in Figure 3.5.

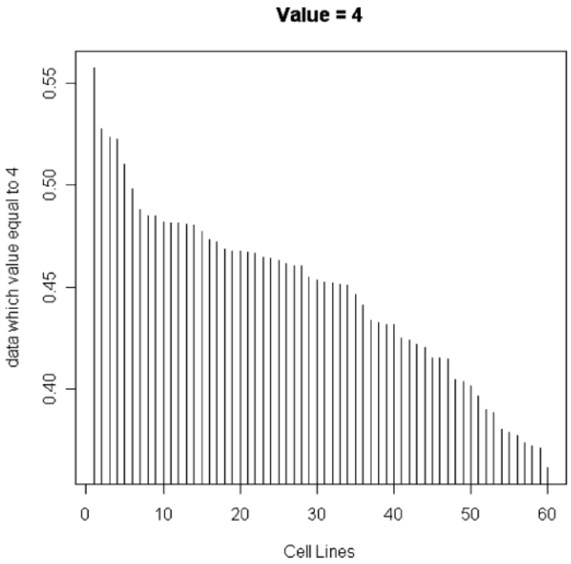


Figure 3.4 Fraction of compounds with growth inhibition values of 4.0 for each of the 60 cell lines.

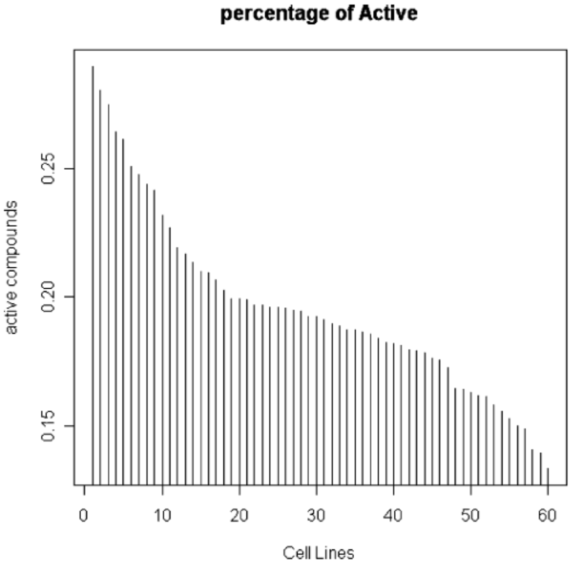


Figure 3.5 Fraction of compounds showing activity in each of the 60 cell lines.

3.3.3 Characterization of the Gene Expression Results

Although this paper does not directly address data mining of the gene expression results, we carried out some initial experiments to characterize the data, for completeness and as a basis for future data mining experiments. The distributions of the microarray gene expression data are shown in Figure 3.6. The values less than zero represent underexpression from the norm and the values above zero represent overexpression. As shown, the overall distribution and the distribution for individual cell lines are very similar. Based on these distributions, for our work we decided to consider values less than or equal to -1.0 and greater than or equal to 1.0 to indicate under or overexpression, respectively.

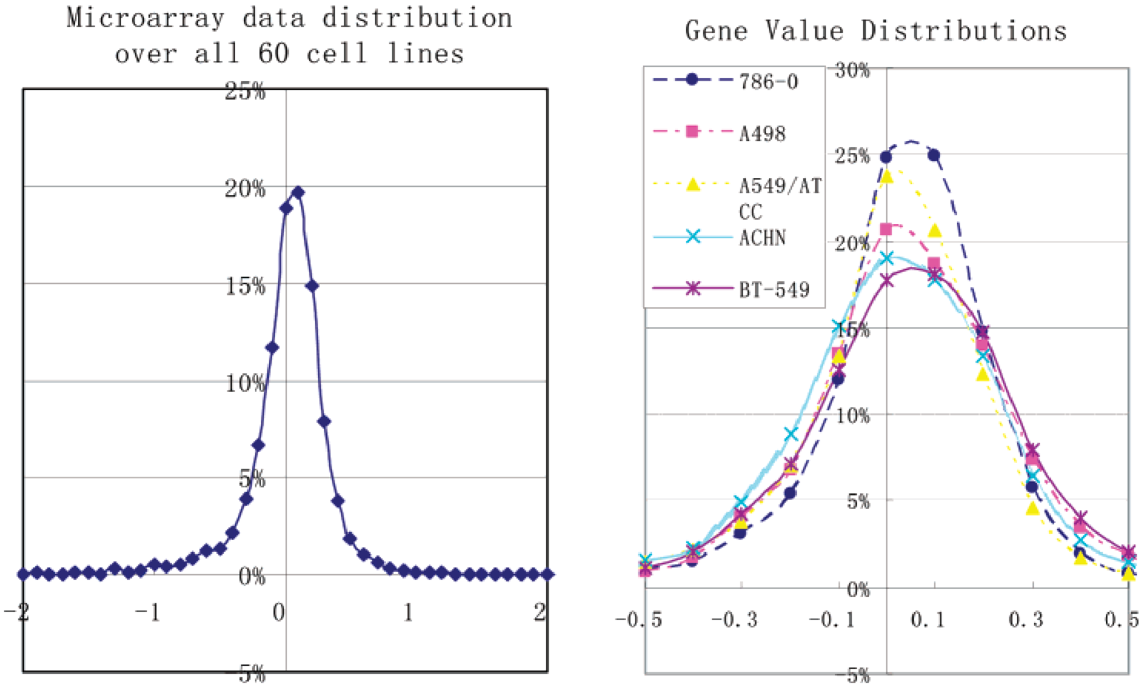


Figure 3.6. Distribution of the microarray gene expression data across all the 60 cell lines (left) and for five randomly selected cell lines (right).

3.3.4 Predicting Missing Activity Values

In order to test whether it might be possible to estimate the missing data points using computational prediction, we applied a machine learning tool, WEKA,[39] on the 2696 compounds which have values for all 60 cell lines. We did two prediction experiments using various methods: first using only 166 known attributes to predict one attribute (the 166 fingerprint is known and the cell line information is unknown); second a leave-one-out approach, using 255 known attributes to predict one attribute (the 166 fingerprint and 59 cell line growth inhibition values as known attributes, one cell line growth inhibition value as unknown). Tables 3.1 and 3.2 show the accuracy of the prediction using various methods (ADTree and REPTree, two decision tree methods; RIDOR, a rule-based method; AODE and BayesNet, two Bayesian methods; and VFI, a voting feature interval classifier). The columns show the true and false positive rates, precision, and activity class for each of the methods. Clearly the accuracy is poor when only fingerprint bits are used, but is much improved when other cell line data are included. We may therefore assume that activity in one cell line is related to activity in others. While we would have liked to use this method to predict missing values, we are not confident that the set is complete enough to warrant it: 90% of the compounds miss some cell line data and only 10% of compounds are missing only one cell line data.

Table 3.1 Accuracy of the prediction using only fingerprint information

methods	TP_Rate	FP_Rate	Precision	Class
ADTree	0.087	0.047	0.434	0
	0.953	0.913	0.713	1
REPTree	0.192	0.098	0.451	0
	0.902	0.808	0.727	1
Ridor	0.029	0.008	0.59	0
	0.992	0.971	0.709	1
AODE	0.389	0.227	0.418	0
	0.773	0.611	0.751	1
BayesNet	0.436	0.303	0.376	0
	0.697	0.564	0.747	1
VFI	0.545	0.413	0.356	0
	0.587	0.455	0.755	1

Table 3.2 Accuracy of prediction using fingerprint and cell line information

methods	TP_Rate	FP_Rate	Precision	Class
ADTree	0.813	0.092	0.787	0
	0.908	0.187	0.92	1
REPTree	0.781	0.087	0.789	0
	0.913	0.219	0.909	1
Ridor	0.785	0.091	0.784	0
	0.909	0.215	0.91	1
AODE	0.815	0.102	0.771	0
	0.898	0.185	0.921	1
BayesNet	0.82	0.109	0.759	0
	0.891	0.18	0.922	1
VFI	0.83	0.118	0.746	0
	0.882	0.17	0.925	1

3.4 Data Mining

Having obtained some broad characterizations of the compounds and cell line screening results in the set, we performed several experiments to find relationships between 2D chemical structure and activities across the 60 cell lines. Our intention in these experiments was to use both statistical and predictive modeling methods to look for associations and relationships between chemical structure features (as encoded by the 166-bit fingerprints) and the actual activities of the compounds in the 60 cell lines. Specifically, we applied a standard statistical ratio technique across all the cell lines, a random forest predictive modeling technique (as might be used in QSAR studies) to each cell line individually, and a novel rule-based SMARTS matching procedure that effectively generates "on-the-fly" structural descriptors related to activities.

3.4.1 Relating Dictionary-Based Structural Keys to Cellular Screening Activities

The activity classifications (active, inactive) and the structural key fingerprint bits described previously were used to determine which structural features were either more prevalent or scarce in active compounds compared with inactives. Two ratios, the active-structural ratio and overall-structural ratio, were created. The active structural ratio $R_{a,j}$ for a structural feature j is defined as

$$R_{a,j} = \frac{T_{a,j}}{C_a} \tag{1}$$

where $T_{a,j}$ is the total number of compounds with the feature j, and C_a is the set of active compounds. The overall-structure ratio R_j is defined as

$$R_j = \frac{T_j}{|C|} \tag{2}$$

where T_j is the total number of compounds with a structural feature j, and C is the complete set of compounds. We may then calculate the difference between these values, which provides a statistical value for how much more prevalent or absent a feature j is in the active compounds compared with the feature in all compounds:

$$\mathrm{diff}_j = R_j - R_{a,j} \tag{3}$$

Figure 3.7 plots the difference between the active ratio and the overall ratio for each of the 166 keys. A positive value indicates the greater percentage of this feature appearing in the active cells. Alternatively, a negative value indicates the lack of the feature in the active compounds compared with all compounds. Each feature was evaluated across all 60 cell lines, and thus each bar on the x-axis of the chart (each structural attribute) is based on 60 y-values. The effects of the substructure on the compounds' activities are very consistent as shown in the figure. Nearly all 60 cell lines follow the same track. Thus, we can use the average difference of the active ratio and the overall ratio to find the most important substructures in determining the "global" activity and inactivity. We may consider the features associated with global activity to be indicative of promiscuity (i.e., the tendency to bind to anything) and those associated with inactivity to be ones that tend to stop binding to tumor growth related proteins in a variety of situations. We found that the bits 105, 127, 145, 152, and 99 are the most important bits for activity and the bits 117, 110, 92, 77, and 95 are the most significant bits for inactivity. The Daylight SMARTS strings[40] and reasonable interpretations of those significant bits are shown in Table 3.3. In interpreting these results, it should be noted that approximately 5% of the structural keys differ only in the number of features present in the molecule, and some that almost never occur in biological molecules.

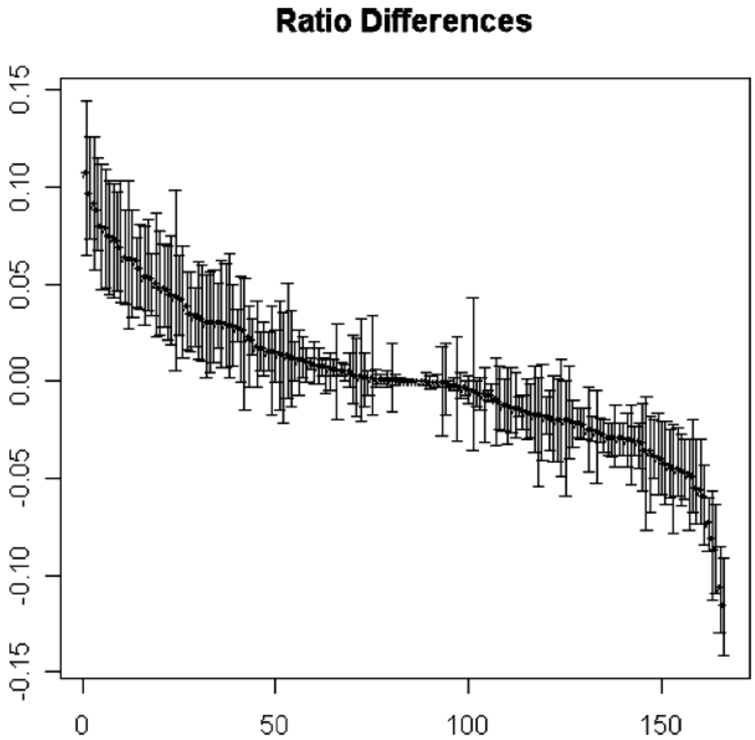


Figure 3.7 Difference in active structural ratio and overall structural ratio. Each of the 166 structural attributes is represented across the *x*-axis with the 60 cell lines displayed in various points. The central line shows the mean difference in active structural ratio and overall structural ratio.

Table 3.3 SMARTS and interpretation for the bits associated with global activity and inactivity

SMARTS	bit	Interpretation
		Significant Features for Activity
@(@*)@*	105	multiple ring system
@!@[#8].*@*!@[#8]	127	>1 aliphatic oxygen joined to a ring
~1~~*~*~*~*1.*~1~*~*~*~*~*1	145	>1 6-membered rings
[#8]~[#6](~[#6])~[#6]	152	tertiary carbon with 2 carbons and 1 oxygen attached
C=C	99	double-bonded carbons
[CH3].[CH3]	149	>1 methyl group
[CH3].[CH3].[CH3]	141	>2 methyl groups
[CH3]~*~*~[CH2]~*	116	methyl 3 bonds away from a chain carbon
[CH3]~*~[CH2]~*	115	methyl 2 bonds away from a chain carbon
[#7]~[#8]	71	NO
		Significant Features for Inactivity
[#7]~*~[#8]	117	nitrogen one bond away from an oxygen
[#7]~[#6]~[#8]	110	N−C−O
[#8]~[#6](~[#7])~[#6]	92	OC(N)C
[#7]~*~[#7]	77	two nitrogens separated by one bond
[#7]~*~*~[#8]	95	nitrogen two bonds away from an oxygen
[!#6]~*(~[!#6])~[!#6]	106	heteroatom bonded to atom with 2 branched heteroatoms
[#16]	88	sulfur
[#16]~*(~*)~*	81	sulfur off a branched system
[!#6]~[#7]	94	heteroatom bonded to a nitrogen
[#7]~[#6](~[#6])~[#7]	38	NC(C)N

We may deduce from this that compounds with multiple ring systems, particularly involving oxygens and methyl groups, tend to be associated with activity, and close non-amide formations of nitrogens and oxygens as well as sulfur-containing compounds tend to not be active. This is borne out by looking at compounds which are active or inactive in all cell lines: a few examples are given in Figure 3.8.

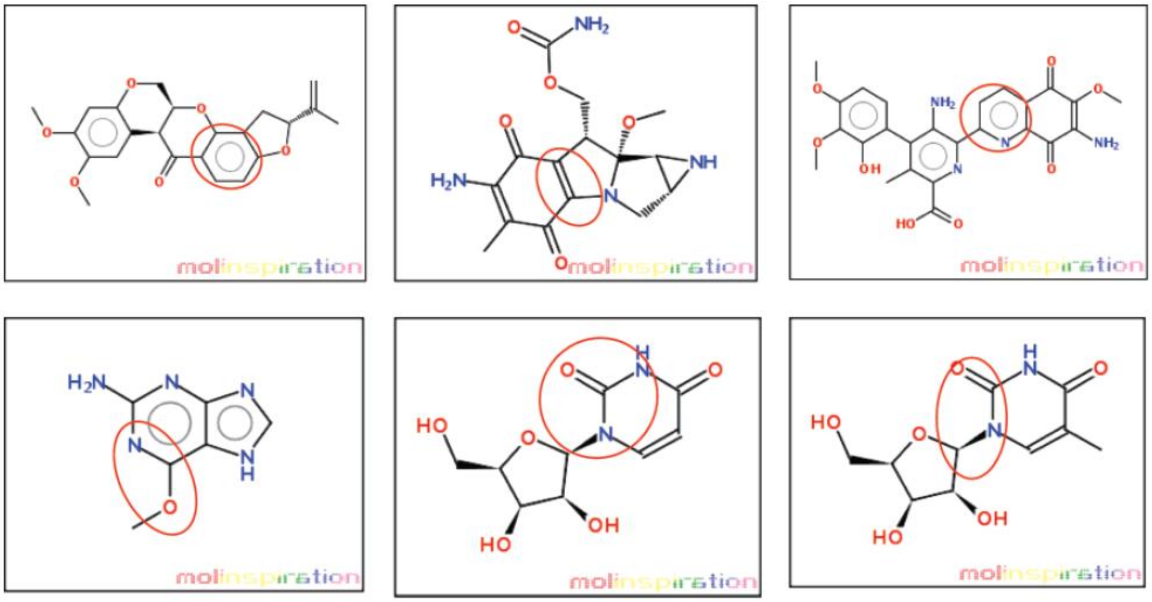


Figure 3.8 Example compounds which are active in all cell lines (top row) or inactive in all cell lines (bottom row). Depictions were generated by the Molinspiration package, www.molinspiration.com. Features identified in Figure 3.3 are highlighted.

3.4.2 Predictive Models of Activity

As shown in the last section, some structure features are highly correlated with activity or inactivity across the cell lines. We next performed experiments to see if it would be possible to build a predictive machine-learning model that can predict individual activity in each of the 60 cell lines. Our previous study with WEKA shows the AD-Tree and Ridor methods work best of the models available in that package. As an example, we initially applied those two methods on various feature subsets using cell line 60 (UO-31). The features are selected based on the rank of active and inactive features across all 60 cell lines and the rank of activity and inactivity features on cell line 60. For example, 20 features contain the top 10 active features and top 10 inactive features.

The results of these experiments are shown in tabular form (Table 3.4) and graphically (Figure 3.9). Clearly, not all 166 structural features are useful in determining the cell line activity. Our experiments show that the best prediction accuracy for AD-tree only uses 60 structural features and that the best prediction accuracy for Ridor only uses 80 structural features if the features are chosen based on the rank cross all cell lines. By limiting the number of features, we can increase the prediction accuracy for the inactive group from 43% to 62% for AD-tree and from 51% to 71% for Ridor. The best prediction accuracy for AD-tree only uses 40 structural features, and the best prediction accuracy for Ridor only uses 80 structural features if the features are chosen based on the rank over cell line 60. It also shows that the feature selection helps increase the prediction accuracy. Interestingly, the feature selection based on cell line 60 is slightly worse than the feature selection based on all 60 cell lines.

In addition to these methods, we also considered the random forest.[41] This technique has become popular in the data mining community, and there are a number of examples of its use in the chemical informatics literature.[42-44] The random forest is essentially an ensemble of decision trees and is thus an example of a bagging method.[45] The ensemble character of this method leads to some useful characteristics. Most important for our purposes is the fact that to develop a random forest model, one is not required to perform feature selection a priori. In addition, it can be shown that a random forest model does not overfit. That is, increasing the number of trees in the ensemble does not lead to overfitting, and the only real disadvantage is the increase in memory consumption.

Table 3.4 Accuracy of the prediction based on various structure features

	based on the rank cross all cell lines				based on the rank over cell line 60			
	AD-Tree		Ridor		AD-Tree		Ridor	
features	inactive	active	inactive	active	inactive	active	inactive	active
10	0.35	0.71	0.24	0.70	0.22	0.71	0.33	0.71
20	0.48	0.71	0.46	0.71	0.47	0.71	0.52	0.71
40	0.60	0.72	0.54	0.71	0.61	0.72	0.48	0.71
60	0.62	0.72	0.56	0.71	0.58	0.72	0.61	0.71
80	0.51	0.71	0.71	0.71	0.33	0.71	0.62	0.72
100	0.44	0.72	0.64	0.71	0.46	0.75	0.62	0.72
120	0.41	0.71	0.63	0.71	0.46	0.75	0.62	0.72
140	0.46	0.72	0.61	0.71	0.49	0.74	0.61	0.72
166	0.43	0.71	0.59	0.71	0.43	0.71	0.59	0.71

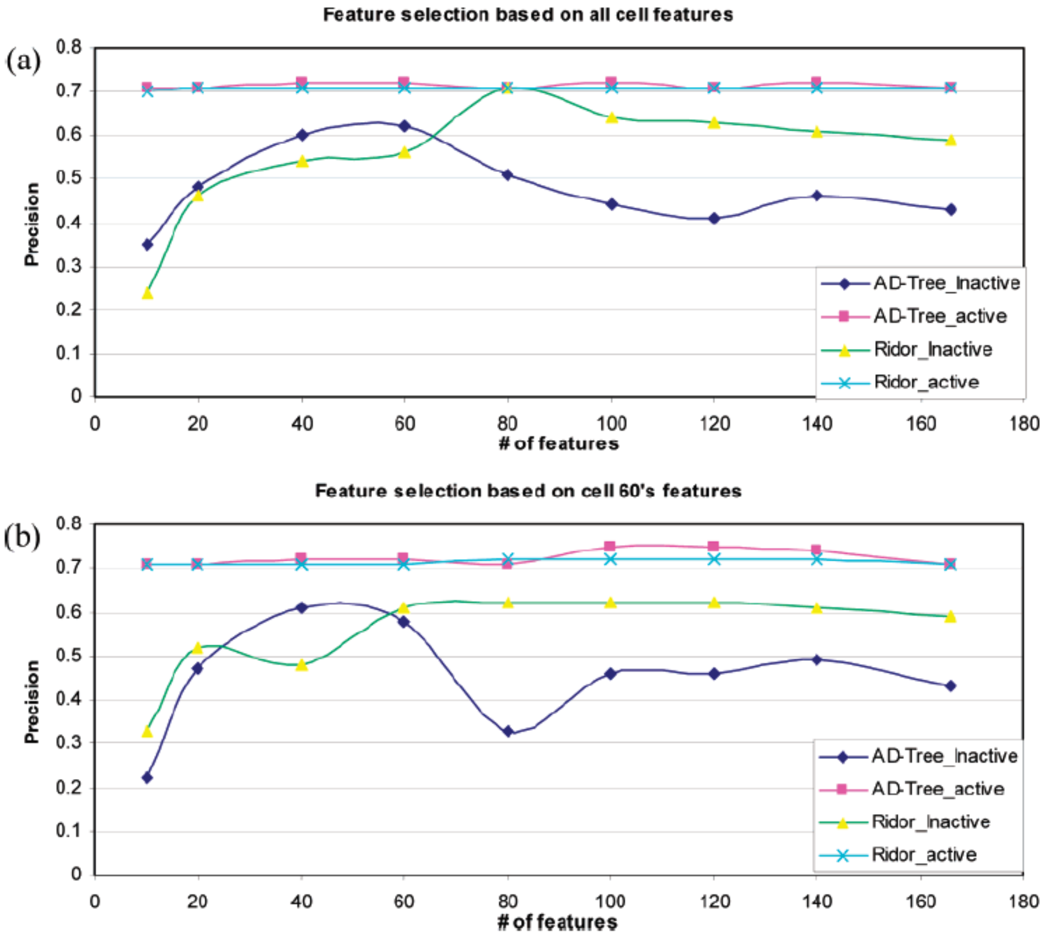


Figure 3.9 Accuracy of the prediction based on various structural features: (a) structural features are ranked across all cell lines and (b) structure features are ranked over cell line 60.

We developed 60 random forest models, one for each cell line, using the random forest package available in R.[46] We considered the 166-bit fingerprints previously described for the input features. For general usage the default settings for the method lead to good results. The main parameter of interest is the number of trees in the ensemble. As noted above, a higher number of trees do not lead to overfitting. However the default value of 500 trees led to excessive memory consumption when we built all 60 models. We investigated a number of values for this parameter and settled on 250 trees. The models were developed on a machine equipped with a 3.2 GHz dual core Xenon CPU and 2 GB RAM running Fedora Core 5. On average, the development of a single model took 16.5 min. Since we had a dual core CPU, we processed two cell lines at a time, thus leading to a total run time of 8 h to develop all 60 models. Note that the speed of this process could easily be increased by utilizing one of the many parallel execution packages available for R (such as snow) and a cluster of machines. Alternatives to the random forest could also be considered. Since we are mainly interested in pure predictive ability (as opposed to developing a model of the underlying distribution) one possible approach would be to consider a k-nearest neighbor classification. Though simplistic in nature, this method would be relatively fast, though for larger data sets this may not be such an advantage unless appropriate nearest neighbor detection algorithms were employed. The downside to this and other methods is that some sort of feature selection would need to be performed prior to the prediction step.

As has been noted above, the data sets for each cell line represent an unbalanced classification problem, with the actives being the minor class. As can be seen from Table 3.4, this leads to very poor predictive performance, since new observations will tend to be classified as inactive, by default. To alleviate this problem in our random forest models, we specified that for each tree in the ensemble the algorithm should consider all the actives as well as a set of randomly selected inactives in the ratio of 1.0:0.6. Thus each tree in the ensemble would not see the highly unbalanced data set but would in fact see a subset that was enriched by the actives. By including a smaller number of inactives, one can effectively force each individual tree to exhibit a high predictive accuracy for the minor (active) class. It is clear that this is simply the reverse of the current situation, where we have very good predictive accuracy for the major (inactives) class. As a result, we experimented with a variety of ratios until we obtained a ratio, where the predictive accuracy for the minor and major classes were approximately equal. We realize that this approach does lead to a model biased in favor of the actives. We believe that this is justified since our aim is to try and avoid false negatives. Thus by biasing toward the active class, we not only improve the true positive rate but also increase the false positive rate at the expense of the false negative rate. Finally, for each cell line we considered only those observations that had measured values of growth inhibition and split the data sets, such that 70% was placed in a training set and 30% in a test set.

The plots in Figure 3.10 summarize the predictive accuracy for the 60 models that were developed using the above approach. We consider the predictive accuracy in three ways: Box A represents the range of percentage correct prediction for the test set overall, across the 60 cell lines. For this case we utilized the g-mean measure of accuracy described by Kubat et al.[47] which takes into account the unbalanced nature of the test set. The worst model exhibited a 67% correct accuracy, while the best model exhibited close to 77% correct. Box B represents the percent correct prediction for the actives, across all

60 cell lines. It is clear that the variation in the accuracies for the 60 models is much smaller when the actives are considered in isolation. This is not surprising, since by construction the models are expected to fare better on the actives. Thus we see that the accuracies range from 74% to 79% correct. In contrast, Box C represents the percent correct prediction for the inactive class over the 60 cell lines. It is clear that the spread of accuracy is much more than for the actives, and once again this is a result of our model construction. As we noted above, our focus is on identifying actives, thus we accept a slightly poorer performance on the inactive class.

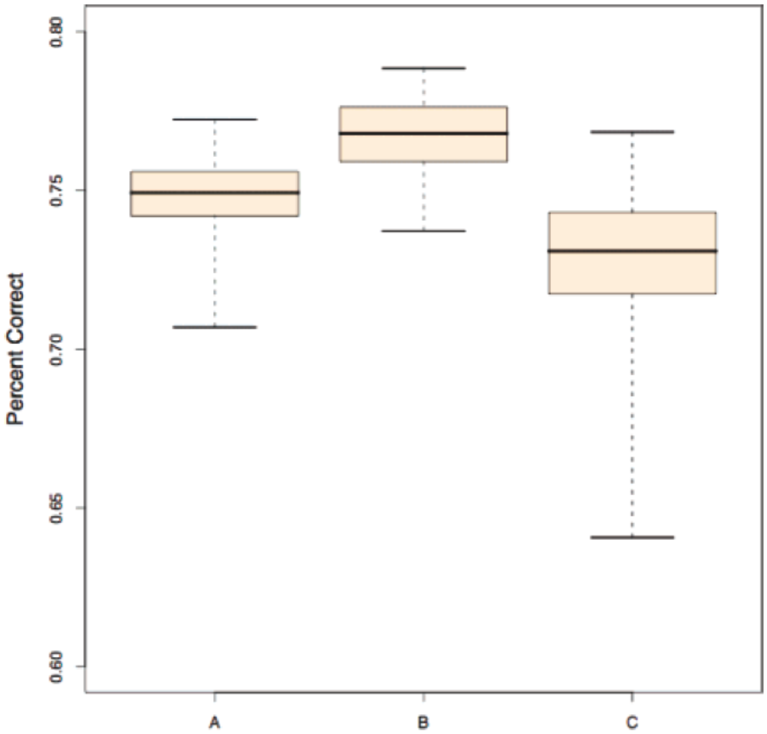


Figure 3.10 A box and whisker summary of the prediction accuracy for the 60 random forest models developed for the NCI DTP cell lines. Box A is the percent correct accuracy for the overall test set, box B is that for the actives, and Box C is that for the inactives. In each case, the whiskers extend to the extremes of the observed accuracy over the 60 cell lines.

The models have been deployed in our Web service infrastructure,[48] allowing access to predictions from any client that supports SOAP. As an example we have provided a Web page client that allows one to supply a set of SMILES and obtain the predicted activity class for all 60 cell lines. In addition, the probability associated with each classification is also provided. Thus, values greater than 0.5 indicate an increasingly higher probability of being predicted active and correspondingly for values lower than 0.5. The Web page can be accessed at http://www.chembiogrid.org/cheminfo/ncidtp/dtp.

3.4.3 Relating Freely Generated SMARTS Structures to Cellular Screening Activities

Our previous experiments used a constrained dictionary of 166 SMARTS fragments. We were also interested in applying a free-form approach that has been developed at the University of Michigan in which a larger number of SMARTS-based fragment keys are generated. A brute force method of lengthening and scoring SMARTS strings was applied in order to establish SMARTS strings up to seven atoms long that have a strong tendency to identify active and inactive compounds across the cell lines. For this experiment we used an updated version of the NCI/DTP 60 cancer cell line data set obtained through PubChem. A MOE database was created for the 42888 compounds that had both structural and growth inhibition data in order to perform iterative scoring based SMARTS structural similarity searches. This method tracks active and inactive hits for a set of SMARTS strings across the entire data set. SMARTS strings are then scored, evaluated, ranked, pruned, and extended for subsequent searches.

Scoring is determined by the ratio of active compounds identified by a SMARTS string divided by the number of inactive compounds identified by the same SMARTS string. With this method, scores will range from 0 to ∞. The ratio of active to inactive compounds in the NCI/DTP data set is 7274 to 35664. If we took a random sampling of the data set we would expect to find one active compound to every five inactive compounds selected. Therefore, the ratio of significance is 1:5 or 0.2. Here we will consider SMARTS strings that demonstrate a tenfold improvement in active or inactive hits as significant. That is, the score of significance for SMARTS strings identifying active compounds is greater than or equal to 2.0 and less than or equal to 0.02 for inactive compounds. Weight can further be given to SMARTS stings, which have a high number of total hits. For example, if SMARTS string A has a score of 5.0 with a total of six hits, five active and one inactive, it is not as significant as SMARTS string B with a score of 5.0 with 240 total hits, 200 active and 40 inactive. In this case SMARTS string A may likely be an artifact of the data set.

Adjusting the scores of significance with the ratio of significance allows one to deal with an unbalanced data set with an even greater skew than the NCI/DTP data set. If the active:inactive ratio of significance were much smaller, for example 1:100 or 0.01, the score of significance for an inactive substance would be taken to be greater than or equal to 0.1. Furthermore, with this strong bias in the data set toward inactives, we would expect that there would be fewer SMARTS strings associated with active substances and more associated with the inactives.

The specific algorithm applied for identifying and lengthening SMARTS strings incorporates three pruning rules at various stages to eliminate redundancies, to improve computational efficiency, and to eliminate artifacts. The workflow of our algorithm is depicted in Figure 3.11. This procedure was performed on a Dell Precision 380 workstation with 3 gHz CPU with 1 GB RAM. Runtimes for each iteration of the algorithm were based on the size of the SMARTS string set and ranged from to 2 min to

11 h, for sets on the order of 100 and 20000, respectively. The details of the steps performed are as follows:

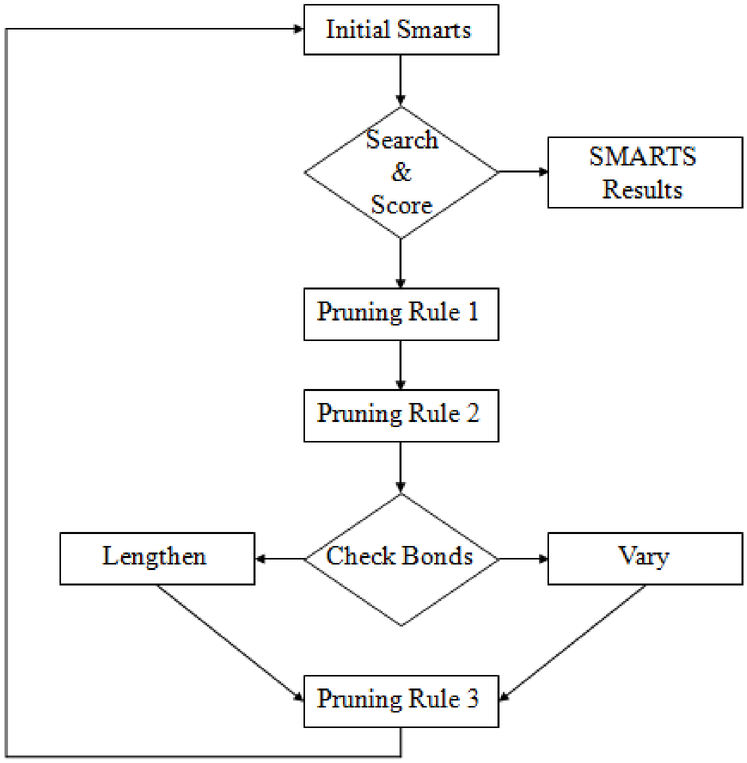


Figure 3.11 Algorithm workflow.

1. Select Initial SMARTS Strings.

   For the sake of generality, elements 2–105 of the periodic table were selected as single atom SMARTS strings. Hydrogen was not included in this SMARTS string set, as SMILES strings and the molecular connectivity tables provided typically suppress hydrogen atoms.

2. Search & Score

   A substructure search was performed against the NCI/DTP data set using the SMARTS string set. Scores were tabulated, and a bit string hit profile was maintained for each individual SMARTS string across all 42888 compounds. A bit string hit profile consists of a string of 42888 1's and 0's, where 1 means that the SMARTS string is found within the compound, and 0 means that the SMARTS string could not be found within the compound.

3. Record incremental SMARTS String Results.

   If SMARTS Strings contain seven atoms and no general bond types, then terminate the algorithm.

4. Apply Pruning Rule 1 to eliminate redundancies.

   Maintain only one SMARTS string child per unique bit string hit profile. The lengthening of SMARTS strings is a tree process leading to the exponential generation of child SMARTS strings. Bit string profiles are used in order to limit branching as they serve to identify all duplicate SMARTS strings as well as SMARTS strings that do not hit any compounds. Pruning will improve the efficiency of subsequent substructural searches.

5. Apply Pruning Rule 2 to improve computational efficiency.

   If the number of SMARTS string children exceeds 24000, then drop all parent SMARTS strings having scores in the range [0.2/*X* and 0.2**X*]. Starting with *X* = 1.5, increase *X* in increments of 0.1 until the number of SMARTS string children is less than or equal to 24000.

6. Check Bonds to select rules for generating child SMARTS strings.

   a. Vary Bond: If the parent SMARTS strings contain general bonds, then generate all possible SMARTS string children by varying the bond type. For SMARTS strings with fewer than five atoms all six specific bond types were used. For SMARTS strings with five atoms or more, the triple bond was disregarded. See Table 3.5 for a description of the bond types.

b. Lengthen: If the parent SMARTS strings do not contain any general bonds (~), then generate all possible SMARTS string children by joining a single atom to all the potential locations on the SMARTS strings with a general bond. For SMARTS strings with fewer than five atoms, the following atoms were appended to the parent SMARTS string: B, C, N, O, Si, P, S, F, Cl, Br, and I. These elements were selected, as they are among the most common in the PubChem compound data set. Table 3.6 shows the 14 most common single atom SMARTS strings found in the NCI/DTP data set based on the number of compounds identified. Na, Sn, and Pt were not included because our SMARTS strings only consider covalently bound atoms. For SMARTS strings with five or more atoms, C, O, N, P, and S were appended to the parent SMARTS strings. We limit the number of atoms based on the most common nonmetals in order to keep the number of children SMART strings in check. Using common elements allows generation of SMARTS string children that will hit compounds in the data set.

7. Apply Pruning Rule 3 to eliminate artifacts and improve computational efficiency.

For SMARTS strings with fewer than five atoms, drop all children SMARTS strings with less than 20 total hits. For SMARTS strings having scores with five atoms or more, drop all children SMARTS strings with fewer than 100 total hits.

8. Go to Step 2.

Table 3.5 Smarts bond types

~	general bond, any possible bond
-!@	single bond, not part of a ring
=!@	double bond, not part of a ring
#	triple bond
-@!:	single ring bond, not aromatic
=@!:	double ring bond, not aromatic
:	aromatic bond

Table 3.6 Most common single atom SMARTS strings in the NCI/DTP data set

SMARTS strings	element	42888 compounds	score
[#6]	C	42 845	0.2044
[#8]	O	38 674	0.1965
[#7]	N	34 992	0.1967
[#16]	S	11 969	0.1555
[#17]	Cl	8483	0.2772
[#9]	F	2557	0.2246
[#35]	Br	1832	0.2820
[#15]	P	1305	0.1929
[#53]	I	617	0.2390
[#14]	Si	349	0.2246
[#11]	Na	302	0.0942
[#50]	Sn	198	2.1936
[#78]	Pt	189	0.4427
[#5]	B	136	0.1525

Table 3.7 describes the overall results generated by our algorithm. It includes the data for SMARTS strings with modifications to all possible positions at which atoms may be added, subject to pruning as noted within the algorithm. Table 3.8 gives examples of the most selective SMARTS.

Table 3.7 Description of results[a]

no. of SMARTS atoms	SMARTS (possible)	SMARTS (used)	SMARTS (hits)	Active (only)	Active (mostly)	Inactive (only)	Inactive (mostly)	Score Range	Data Set Covered
1	105	104	67	1(0)	4(3)	13(3)	0(0)	0.0313−6.25	42888
2	6930	690	133	4(0)	11(7)	16(1)	1(1)	0.0127−23.0	42876
3	914760	2094	540	13(1)	26(21)	88(12)	3(3)	0.0132−23.0	42871
4	1.81E+08	10 248	2470	45(1)	73(49)	481(89)	12(12)	0.0127−31.0	42862
5	4.78E+10	22 584	18815	52(1)	48(19)	1232(318)	36(36)	0.00873−20.0	42752
6	5.98E+12	8150	8146	31(1)	66(55)	877(264)	83(83)	0.00532−12.5	31762
7	8.97E+14	17 155	6470	161(1)	304(204)	1814(359)	121(121)	0.00532−18.0	21253

[a] SMARTS (possible) is the total number of possible SMARTS strings. SMARTS (used) represents the set of SMARTS strings used in each iterative search after pruning. SMARTS (hits) is the number of SMARTS strings with unique bit string profiles. Active/Inactive (only) represent SMARTS strings identifying compounds that are only active and inactive, respectively. Active/Inactive (cutoff) represent SMARTS scoring >2.0 and <0.02, respectively. Integers within parentheses () indicate the number of significant SMARTS that have a minimum of 10 active or inactive hits. Score Range is minimum − maximum score. Data Set Covered is the number of compounds hit out of 42888. The significant drop in Data Set Covered for the last two rows resulted from Pruning Rule 2.

Table 3.8 Some of the most significant SMARTS strings

Order	SMARTS	TotalHits	Score
Active (only)	[#90]	1	∞
	[#8]-!@[#25]	3	
	[#6]-@!:[#6]-!@[#50]	16	
	[#6]-@!:[#6]-!@[#50]-!@[#6]	13	
	[#8]:[#6]-!@[#7]-!@[#7]=!@[#6]	10	
	[#6]:[#6]-@!:[#6](=@!:[#7])-@!:[#6]:[#6]	13	
	[#7]-!@[#6]:[#6]-@!:[#6](=!@[#7])-@!:[#6]:[#6]	12	
Active (mostly)	[#79]	29	6.25
	[#15]-!@[#79]	24	23.0
	[#6]-!@[#15]-!@[#79]	24	23.0
	[#7]-@!:[#6]-@!:[#16]-@!:[#29]	32	31.0
	[#6]-@!:[#6]:[#6]-!@[#6]=!@[#7]	21	20.0
	[#6]:[#6]-@!:[#6]-!@[#8]-!@[#6]-@!:[#8]	81	12.5
	[#6]-!@[#8]-!@[#6]-@!:[#6]:[#6]:[#6]-@!:[#6]	80	15.0
Active (mostly)	[#16]-@!:[#8]	80	0.0127
	[#7]:[#16]:[#6]	77	0.0132
	[#6]-@!:[#7]-!@[#7]-@!:[#6]	80	0.0127
	[#8]=!@[#6]-!@[#6]-!@[#6]=!@[#7]	231	0.00873
	[#8]=!@[#6]-!@[#6]-!@[#6]=!@[#7]-!@[#7]	190	0.00529
	[#8]=!@[#6]-!@[#6]-!@[#6](=!@[#7]-!@[#7])-!@[#6]	189	0.00532
Inactive (only)	[#12]	14	0
	[#7]-!@[#27]	46	
	[#7]=!@[#6]-!@[#5]	29	
	[#7]=!@[#6]-!@[#7]=!@[#7]	75	
	[#6]=!@[#6]-@!:[#7]-@!:[#6]-@!:[#7]	147	
	[#7]:[#6](:[#6]-!@[#6]-!@[#6])-!@[#8]	200	
	[#6]:[#7]:[#6](-!@[#6]-!@[#6]):[#6]-!@[#8]	178	

We then tested the SMARTS strings from the 166-bit fingerprints with the scoring system from this method. Based on the ratio of significance, the individual SMARTS strings for identifying the active and inactive compounds showed minimal increase and decrease in relative score. We identified all compounds that contained all active motifs and inactive motifs, respectively. When considering collections of low and high scoring motifs in a Boolean AND operation, a 2–4-fold respective increase in selectivity was identified. Furthermore, it was found that when combining more than five MACCS SMARTS strings the score minimally increased or decreased; however, the total number of hits significantly decreased. See Table 3.9 for details. We then tabulated the Boolean OR incorporating all active and inactive SMARTS strings from the MACCS example. Almost all compounds were selected, and the score of significance for both active and inactive sets was ~0.2.

Table 3.9 Scoring selective MACCS SMARTS strings

Type	MACCS SMARTS String	no of. Active Hits	no. of Inactive Hits	Score
Active	*@*(@*)@*	5102	19 855	0.2570
Active	*@*!@[#8].*@*!@[#8]	3192	10 256	0.3112
Active	*~1~*~*~*~*~*1.*~1~*~*~*~*~*1	5589	24 358	0.2295
Active	[#8]~[#6](~[#6])~[#6]	4836	19 565	0.2472
Active	C=C	3275	12 144	0.2697
Active	[CH3].[CH3]	3853	15 978	0.2411
Active	[CH3].[CH3].[CH3]	2470	9006	0.2743
Active	[CH3]~*~*~[CH2]~*	2099	7013	0.2993
Active	[CH3]~*~[CH2]~*	1921	5691	0.3376
Active	[#7]~[#8]	804	4296	0.1872
Active	Boolean AND (5 highest scoring Active)	440	1034	0.7407
Active	Boolean AND (All Active)	9	21	0.7500
Active	Boolean OR (All Active)	7246	35 104	0.2064
Inactive	[#7]~*~[#8]	2595	17 823	0.1456
Inactive	[#7]~[#6]~[#8]	2383	16 376	0.1455
Inactive	[#8]~[#6](~[#7])~[#6]	2123	14 314	0.1483
Inactive	[#7]~*~[#7]	2039	13 063	0.1561
Inactive	[#7]~*~*~[#8]	2407	14 617	0.1647
Inactive	[!#6]~*(~[!#6])~[!#6]	2121	13 468	0.1573
Inactive	[#16]	1611	10 358	0.1555
Inactive	[#16]~*(~*)~*	1458	9636	0.1513
Inactive	[!#6]~[#7]	2261	12 817	0.1764
Inactive	[#7]~[#6](~[#6])~[#7]	944	6456	0.1462
Inactive	Boolean AND (5 lowest scoring Active)	81	927	0.08738
Inactive	Boolean AND (All Active)	32	272	0.1176
Inactive	Boolean OR (All Active)	5275	28 573	0.1846

We took the Boolean OR for the four sets of SMARTS from this example. As our sets of SMARTS strings were tailored to the NCI60, we expected and confirmed that they outperform the MACCS fingerprints. As one would expect the Active(only) and Inactive(only) sets had scores of ∞ and 0, respectively. The Inactive(mostly) set hit a total of 165 active compounds and 9372 inactive compounds, yielding a score of 0.01761. The Active(mostly) set hit a total of 2999 active compounds and 9949 inactive compounds, respectively, yielding a score of 0.3014. It appears that the Inactive(mostly) set has been better tailored to identifying inactive compounds due to the low threshold score of 0.02 for each SMARTS string. From this, it can be inferred that there was very little overlap of inactive and active compounds identified. However in the case of the Active(mostly) set, there was obviously considerable overlap. Suppose SMARTS string A identifies two active compounds and one inactive compound, while SMARTS string B identifies the same two active compounds, it identifies a different inactive compound. If we were to use Boolean OR, tabulating a new score when both SMARTS A and B were used together, the new score would be equal to 1.0 as two active compounds are identified by both SMARTS and two inactive compounds are identified, one by SMARTS string A and the other by SMARTS string B. Therefore, due to the low threshold score required for the Active(mostly) SMARTS strings, we cannot group their properties with the Boolean OR and expect significant active hit enrichment, but rather they must be used discretely in order to maintain scores greater than or equal to 2.0. At this juncture, it would be wise to identify the Active(mostly) SMARTS strings with overlapping active and inactive compounds. Further pruning needs to be performed on the SMARTS strings sharing the same set of active compounds in order to obtain the most orthogonal set. This can be accomplished by maintaining only one SMARTS string identifying a specific set of active compounds and dropping all SMARTS strings identifying equal sized or larger sets of different inactive compounds.

Finally, the most significant SMARTS strings can be used to create molecular fingerprints to give a general prediction regarding the activity of compounds yet to be assayed. This method may be further complemented by addressing the activity profiles of compounds identified by multiple selective SMARTS strings. Also one might consider creating profiles for each of the individual 60 cancer cell line assays and weighting the SMARTS strings based on the growth inhibition value, rather than the binary interpretation used in this method with ‘1’ representing an active hit and ‘0’ an inactive hit in order to give a more quantitative growth inhibition predictions.

3.5 Conclusions and Future Directions

In this work, we have conducted broad characterizations of the compounds, biological activities, and gene expression values in the NIH DTP Tumor cell line data set. We have shown that compounds active or inactive across the 60 cell lines tend to have structural features in common. We have also demonstrated that a Random Forest model can be used to predict the activity profiles of unknown compounds across the cell lines reasonably well. Finally, we show that a novel SMARTS-based algorithm can be used to give finer resolution structure-activity correlations than a constrained dictionary-based fingerprint.

We are currently in the process of extending our data mining to include the gene expression information, in particular finding features that tend to be associated with activity or inactivity in subgroups of the cell lines which share particular gene expression profiles. We also wish to extend our random forest models to include information from other cell lines in our prediction of individual cell line activities.

This work was published in the Journal of Chemical and Information Modeling, reference Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A. C.; Guha, R.; Wu, Y.; Crippen, G. M.; Wild, D. J. Chemical Data Mining of the NCI Human Tumor Cell Line Database *J. Chem. Inf. Model.* **2007,** *47,* 2063-2076. This work was a collaborative effort between the Indiana University School of Informatics and Chemical Informatics and Cyberinfrastructure Collaboratory and the University of Michigan’s School of Pharmacy.

3.6 References

(1) Developmental Threapeutics Program Web site. http://dtp.nci.nih.gov (accessed July 23, 2007).

(2) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, Jr, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **1997**, *275*, 343–349.

(3) Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A. P.; Scudiero, D. A.; Rubinstein, L. V.; Plowman, J.; Boyd, M. R. Display and Analysis of Patterns of Differential Activity of Drugs against Human Tumor Cell Lines: Development of a Mean Graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088–1092.

(4) Zaharevitz, D. W.; Holbeck, S. L.; Bowerman, C.; Svetlik, P. A. COMPARE: A Web Accessible Tool for Investigating Mechanisms of Cell Growth Inhibition. *J. Mol. Graphics Modell.* **2002**, *20*, 297–303.

(5) DTP Data Search Page. http://dtp.nci.nih.gov/docs/dtp_search.html (accessed July 23, 2007).

(6) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**, *258*, 447–451.

(7) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **2000**, *24*, 236–244.

(8) Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining the National Cancer Institute's Tumor-Screening Database: Identification of Compounds with Similar Cellular Activities. *J. Med. Chem.* **2002**, *45*, 818–840.

(9) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: Cluster Analysis of Ellipticine Analogs with p53-Inverse and Central Nervous System-Selective Patterns of Activity. *Mol. Pharmacol.* **1998**, *53*, 241–251.

(10) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI Anticancer Drug Discovery Databases: Genetic Function Approximation for the QSAR Study of Anticancer Ellipticine Analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189–199.

(11) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. Leadscope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.

(12) Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. Pharmacogenomic Analysis: Correlating Molecular Substrucure Classes with Microarray Gene Expression Data. *Pharmacogenomics J.* **2002**, *2*, 259–271.

(13) Huang, Y.; Blower, P. E.; Yang, C.; Barbacioru, C.; Dai, Z.; Zhang, Y.; Xiao, J. J.; Chan, K. K.; Sade´e, W. Correlating Gene Expression with Chemical Scaffolds of Cytotoxic Agents: Ellipticines as Substrates and Inhibitors of MDR1. *Pharmacogenomics J.* **2005**, *5*, 112–125.

(14) Blower, P. E.; Cross, K. P.; Fligner, M. A.; Myatt, G. J.; Verducci, J. S.; Yang, C. Systematic Analysis of Large Screening Sets in Drug Discovery. *Curr. Drug Discovery Technol.* **2004**, (1), 37–47.

(15) Richter, L.; Rückert, U.; Kramer, S. In *Learning a Predictive Model for Growth Inhibition from the NCI DTP Human Tumor Cell Line Screening Data: Does Gene Expression Make a Difference?* Pac. Symp. Biocomput., 2006; 2006; pp 596–607.

(16) Cho, S. J.; Shen, C. F.; Hermsmeier, M. A. Binary Formal Inference-Based Recursive Modeling Using Multiple Atom and Physicochemical Property Class Pair and Torsion Descriptors as Decision Criteria. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 668–680.

(17) Fang, X.; Shao, L.; Zhang, H.; Wang, S. Web-Based Tools for Mining the NCI Databases for Anticancer Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 249–257.

(18) Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 1st ed.; Morgan Kaufmann: 2000.

(19) Gray, J.; Chaudhuri, S.; Bosworth, A.; Layman, A.; Reichart, D.; Venkatrao, M.; Pellow, F.; Pirahesh, H. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Min. Knowledge Discovery* **1997**, 29–53.

(20) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1–40.

(21) MacQueen, J. B. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*; 1967; pp 281–297.

(22) Kaufman, L.; Rousseeuw, P. J. *Findings Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: New York, 1990.

(23) Ng, R. T.; Han, J. In *Efficient and effecti*V*e clustering methods for spatial data mining;* 1994 International Conference Very Large Data Bases (VLDB'94), Santiago, Chile, 1994; Santiago, Chile, 1994; pp 144–155.

(24) Zhang, T.; Ramakrishnan, R.; Livny, M. In *BIRCH: An efficient data clustering method for* V*ery large databases;* 1996 ACM-SIGMOD International Conference Management of Data (SIGMOD '96), Montreal, Canada, 1996; Montreal, Canada, 1996; pp 103–114.

(25) Guha, S.; Rastogi, R.; Shim, K. In *Cure: An efficient clustering algorithm for large databases*; 1998 ACM-SIGMOD International Conference Management of Data, Seattle, WA, 1998; Seattle, WA, 1998; pp 73–84.

(26) Karypis, G.; Han, E.-H.; Kumar, V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *COMPUTER* **1999**, 68–75.

(27) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. In *A density-based algorithm for disco*V*ering clusters in large spatial databases;* 1996 International Conference of Knowledge Discovery and Data Mining (KDD'97), Portland, OR, 1996; Portland, OR, 1996; pp 226–231.

(28) Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. In *OPTICS: Ordering points to identify the clustering structure;* 1999 ACMSIGMOD International Conference Management of Data (SIGMOD'99), Philadelphia, PA, 1999; Philadelphia, PA, 1999; pp 49–60.

(29) Hoschka, P.; Klosgen, W. A support system for interpreting statistical data. In *Knowledge Discovery in Databases*; AAAI/MIT Press: Cambridge, MA, 1991; pp 325–346.

(30) Wang, W.; Yang, J.; Muntz, R. R. In *STING: A statistical information grid approach to spatial data mining;* 1997 International Conference of Very Large Data Bases (VLDB'97), Athens, Greece, 1997; Athens, Greece, 1997; pp 186–195.

(31) Sheikholeslami, G.; Chatterjee, S.; Zhang, A. In *WaveCluster: A multiresolution clustering approach for very large spatial databases;* 1998 International Conference of Very Large Data Bases, New York, 1998; New York, 1998; pp 428–439.

(32) Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. In *Automatic subspace clustering of high dimensional data for data mining applications;* 1998 ACM-SIGMOD International Conference Management of Data (SIGMOD'98), Seattle, WA, 1998; Seattle, WA, 1998; pp 94–105.

(33) Agrawal, R.; Imielinski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD* **1993**, 207–216.

(34) Available from gNova.com.

(35) Elsevier MDL. http://www.mdl.com (accessed July 23, 2007).

(36) FDA MRTD data set. http://www.fda.gov/CDER/Offices/OPS_IO/MRTD.htm (accessed July 23, 2007).

(37) Pubchem. http://pubchem.ncbi.nlm.nih.gov/ (accessed July 23, 2007).

(38) OpenEye. http://www.eyesopen.com (accessed July 23, 2007).

(39) Frank, I. H. W. a. E. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann: San Francisco, CA, 2005.

(40) Daylight SMARTS. http://www.daylight.com (accessed July 23, 2007).

(41) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and regression trees*; CRC Press: Boca Raton, FL, 1984.

(42) Guha, R.; Jurs, P. C. Development of a Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Model.* **2004**, *44*(6), 2179–2189.

(43) O'Brien, S. E.; deGroot, M. J. Greater than the Sum of its Parts: Combining Models for Useful ADMET Prediction. *J. Med. Chem.* **2005**, *48* (4), 1287–1291.

(44) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; R. P., S.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *42*, 1947–1958.

(45) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *26*, 123–140.

(46) Team, R. D. C. *A language and en*V*ironment for statistical computing;* Foundation for Statistical Computing: Vienna, Austria, 2006.

(47) Kubat, M.; Holte, R. C.; Matwin, S. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning* **1998**, *30* (2-3), 195–215.

(48) Dong, X.; Gilbert, K. E.; Guha, R.; Heiland, R.; Kim, J.; Pierce, M.; Fox, G. C.; Wild, D. J. Web Service Infrastructure for Cheminformatics. *J. Chem. Inf. Model.* **2007**, *47* (4), 1303–1307.

Chapter 4

Data Mining the NCI60 to Predict Generalized Cytotoxicity

4.1 Introduction

Elimination of cytotoxic compounds in the early and later stages of drug discovery can help reduce the costs of research and development. Through the application of principal components analysis (PCA), we were able to data mine and prove that ~89% of the total log GI_{50} variance is due to the non-specific cytotoxic nature of substances. Furthermore, PCA led to the identification of groups of structurally unrelated substances showing very specific toxicity profiles, such as a set of 45 substances toxic only to the Leukemia_SR cancer cell line. In an effort to predict non-specific cytotoxicity based on the mean log GI_{50}, we created a decision tree using MACCS keys that can correctly classify over 83% of the substances as cytotoxic/non-cytotoxic *in silico*, based on the cutoff of mean log $GI_{50} = -5.0$. Finally, we have established a linear model using least squares in which 9 of the 59 available NCI60 cancer cell lines can be used to predict the mean log GI_{50}. The model has R^2=0.99 and root mean square deviation between the observed and calculated mean log GI_{50} (RMSE) = 0.09. Our predictive models can be applied to flag generally cytotoxic molecules in virtual and real chemical libraries, thus saving time and effort.

With the advent of high throughput screening (HTS), mountains of biological screening data have been produced and continue to accumulate. In fact, as of 2002, ~14%

of research and development in the pharmaceutical industry is spent on biological screening.[1] In 2003 approximately one-third of the capital lost on all drug failures, $8 billion, was due to the inability to accurately predict toxicity during the early stages of drug development.[2] As it relates to the investigation of cytotoxicity and growth inhibition studies, numerous quantitative endpoints have been used including: protein analysis, enzyme release, exclusion or inclusion of dyes or radioactive markers, and metabolic alterations such as oxygen consumption, and ATP levels. As pharmacokinetics and toxicity (ADMET) are now a consideration in the early stages of drug development, many recent efforts have been made by both academia and industry to address the prediction of specific and general cytotoxicity. Methods commonly utilized to assign a cytotoxicity score or to classify substances as being cytotoxic/non-cytotoxic include neural networks, proteomic profiling, QSPR and QSAR.[3,4,5,6,7] With a plethora of data sources available, it is possible to merge information from multiple HTS libraries in order to obtain a highly diverse set of drug-like molecules which can be used to model both physiochemical and biological properties. Applying *in silico* screens to filter out molecules likely to fail ADMET, especially toxicity, is fiscally necessary considering that it can take over a decade and close to one billion dollars to release a new and federally approved drug.[8]

PubChem, the database component of the National Institutes of Health (NIH) Molecular Libraries Initiative (MLI), serves as a public repository of chemical and biological activity data generated by the Molecular Libraries Screening Center Network and other screening centers.[9,10,11] PubChem is a user/depositor system, which accepts annotated chemical structures and related biological activity data. PubChem is broken down into three main database components: Compound, Substance and BioAssay. PubChem Compound contains over 18 million unique chemical structures with their respective calculated physiochemical properties. PubChem Substance contains over 28 million records with structural data, descriptions of chemical samples from multiple sources, and links to 3D protein structures as well as PubMed citations. Finally, biological screening results for over 800 assays are stored in PubChem BioAssay, including the NCI60 human tumor cancer cell line HTS.[12]

When procuring data from multiple sources, quality and reliability are often issues. Although PubChem has taken some measures to ensure quality control (QC) in terms of referencing between compounds and substances, the quality of the structural data related to the bioassay data of PubChem was termed 'user beware' because the structural content submitted by the depositor is accepted without review.[13] It should be mentioned that PubChem has neither the resources nor the assigned responsibility to curate the data. Notably, if errors are identified, they may be reported to and corrected by the depositor. Accepting screening data from multiple sources, one might expect inconsistent endpoints and instrument variations leading to standardization issues and precision errors.

In order to address the issues of QC and reliability, we have chosen to work with the NCI60 human tumor cell line anti-cancer drug screen, as it is one of the most recognized datasets assembled by a single organization with all assays run at the same location. It provides a well curated publicly available dataset of toxicity profiles for 43899 substances assayed *in vitro* against nine distinct organ based classes of cancer: breast, colon, CNS, leukemia, lung, melanoma, ovarian, prostate and renal. Furthermore, cytotoxic concentrations of substances determined *in vitro* have been shown to correlate well to lethal doses in laboratory animals and humans for a range of selected drugs and chemicals.[14,15] With a rich history spanning over 20 years, 59 of the 60 cancer cell lines are still currently available. The NCI60 contains *in vitro* screening data for up to three IC_{50} endpoints: GI_{50}, TGI and LD_{50}, referring to the concentration of a substance in units of molarity or μg/mL, required for 50% growth inhibition, total growth inhibition, and 50% lethal dose, respectively. The GI_{50} is our measurement of choice, as the lowest concentrations of substances are used for the observed effect. In this paper we use only the log GI_{50} values where the concentration unit is molarity.

In a previous work we contributed our novel method for automatically generating selective SMARTS strings which are able to classify cytotoxic molecules based on mean log GI_{50} cutoff of –5.0.[16] Each substance is associated with its respective mean log GI_{50}, which is the calculated mean of the available log GI_{50} data from the NCI60 for each respective substance. While the SMARTS produced work quite well as a filter, they fail to classify a significant portion of the NCI60. Here, we are more concerned with a robust model that accurately predicts mean log GI_{50}.

In this study, we take a closer look at the NCI60, as it refers to the existing 59 cancer cell lines. First, we address the issues of dataset acquisition, analysis and completeness. Next we examine how principal component analysis (PCA) can be applied to large chemical datasets to extract hidden relations and attribute meaning to orthogonal toxicity profiles. We then apply binary decision trees for the *in silico* prediction of general cytotoxicity. Finally, we demonstrate how stepwise regression was used to develop a least squares fit (LSF) model allowing data from 9 of the NCI60 cell lines to be used as an accurate predictor of generalized cytotoxicity across the 59 cell lines.

4.2 Methods

All calculations were performed using the Molecular Operating Environment (MOE)[17] on a Dell Precision 380 workstation utilizing Red Hat Linux Enterprise version 4.0.

4.2.1 Data Gathering

The PubChem FTP site[9] was our preferred data source, as structural data for each molecule could easily be obtained through the association of the PubChem Substance and BioAssay databases. The entire PubChem database is available in the following file formats: abstract syntax notation (ASN) and extensible markup language (XML). PubChem Substance and Compound downloads are also available in standard data file (SDF) format,[18] while Bioassay downloads are also available in comma separated value (CSV) format. MOE's built in functions were used to import and merge the required structural data and log GI_{50} profiles into a flat table for computational analysis.

4.2.2 Data Analysis

Understanding the landscape of a dataset is necessary in order to avoid 'garbage in garbage out', especially in cases where the dataset is an incomplete matrix, as is the NCI60 dataset. Preliminary analysis of the toxicity profiles showed that only 88.2% of the assay data was complete, i.e. data were available for 88.2% of the 59×43889 possible experimental endpoints. Approximately half of the experimental data consists of upper or lower threshold concentrations signifying minimal, log GI_{50} = –4.0 or –5.0, or maximal activity, log GI_{50} = –8.0. The remaining portion of the experimental data shows some quantitative level of activity which is not threshold. Furthermore, only 4824 substances have been screened against the entire NCI60. The NCI60 dataset provides log GI_{50} data based on the measurements taken in one of three concentrations units: molarity (M) (43474 substances), μg/mL (369 substances), and volumetric (48 substances). There is no log GI_{50} data for 108 substances. In order to determine the units used for the volumetric formats, one must contact the contributor, according to the NCBI help desk. See Figure 4.1 for the distribution of available experimental log GI_{50} values in molarity across the different cell lines.

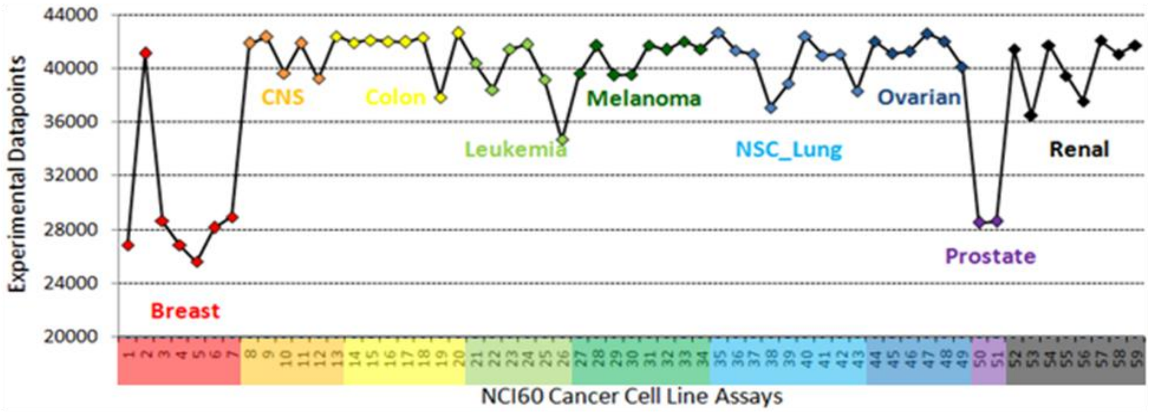


Figure 4.1. The number of available log GI_{50} values for the 43474 substances with measurements in units of molarity was color coded according to the class of cancer cell lines. The cell lines within each class have been alphabetized and numbered.

Figure 4.2 describes the mean log GI_{50} for 43474 compounds for the NCI60 using 1000 bins with the following statistics: mean = –4.518, standard deviation = 0.7447, mode = –4.0, minimum = –11.74 and maximum = 4.0. Interesting features to note are the significant skew of the data toward –4.0, a shoulder at –5.0, and the maximum mean log GI_{50} = 4. The maximum mean log GI_{50} = 4.0 is an anomaly, which we assume to be a data entry problem where the submitter intended to instead enter –4.0. In order to conform to most common standard upper threshold, the mean log GI_{50} of all substances greater than −4.0 was adjusted −4.0. Approximately 80% of all substances have mean log GI_{50} greater than −5.0 explaining the skew towards inactivity. Finally, 5782 and 413 substances, respectively, have mean log GI_{50} = –4 and –5 (indicating no activity) for all assays against which those substances were screened. These values correspond to the maximal allowable concentration of a substance used for assays over a specific timeframe. The National Cancer Institutes Developmental Therapeutics Program (NCI/DTP) home page designates a link to important changes to the NCI60 cell screen, specifying the addition of a one-dose 59 cell assay at concentration 10^{-5} M (corresponding to log GI_{50} = –5) in an attempt "to increase substance throughput and reduce data turnaround time to suppliers while maintaining efficient identification of active compounds." This is followed by the regular 5-dose assay used to determine the GI_{50}, TGI and LD_{50}. [12] For some period in the past, only the 5-dose assay was used with maximum recorded concentration of 10^{-4} M, corresponding to log GI_{50} = –4 (inactive).

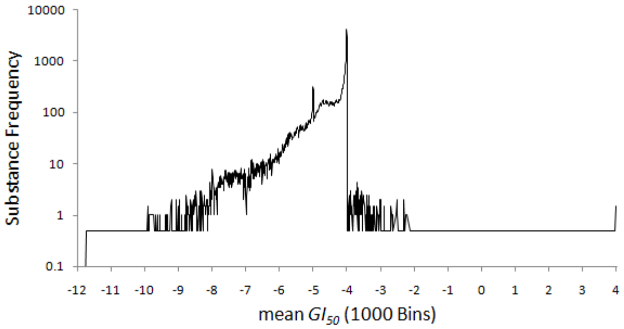


Figure 4.2. The mean log GI_{50} of the NCI60 substances is shown with a granularity of 1000 bins.

Correlation analysis between the activity data from different NCI60 screens of the same substances, i.e. on different cell lines, can also be applied to detect anomalies. In this case two of the threshold values are visualized. In Figures 4.3a and b the upper log GI_{50} threshold is apparent at –4, and a lower threshold is seen at –8. The lower threshold was used for only 178 substances. In this case the threshold values are the maximum and minimum cutoffs in the NCI60 dose response experiments. Note, another apparent threshold of log GI_{50} = –10 can also be detected for some experiments.

MOE's correlation matrix tool allowed us to select *n* database fields (in this case cancer cell lines containing log GI_{50} data) for which it calculates an $n \times n$ matrix of pairwise log GI_{50} correlations (R), where $n \leq 25$. By selecting a specific pairwise correlation from the matrix, we can visualize the scatter plot between any two of the selected cancer cell lines. When examining the scatterplots of pairs of NCI60 log GI_{50} data, very strong correlations were observed. The R^2 ranged from 0.66 to 0.88. It is interesting to note that some cell lines of similar tissue type were less correlated than those of differing organ types as shown in Figures 4.3a and b.

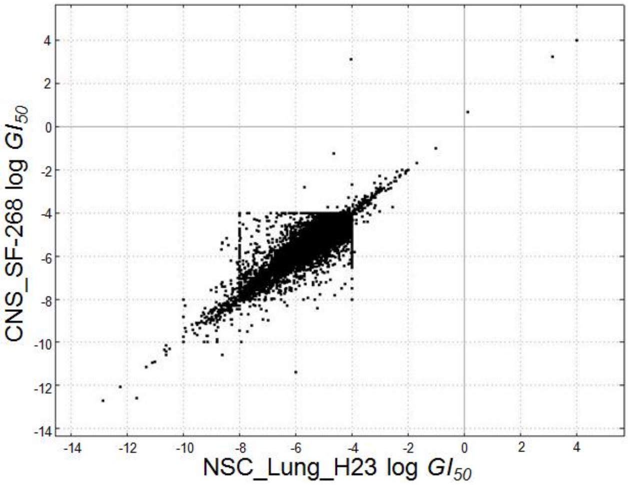


Figure 4.3a. Substance log GI_{50} correlation for 40131 data points between a non-small cell lung and a CNS cancer cell lines having R^2 = 0.86.

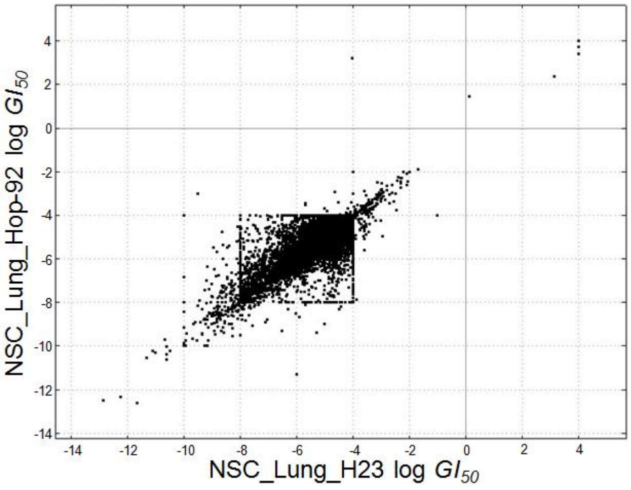


Figure 4.3b. Substance log GI_{50} correlation for 35680 data points between two non-small cell lung cancer cell lines having $R^2 = 0.73$.

The initial analysis is performed to ensure the quality and mining potential for the dataset. In this case, the vast majority of the log GI_{50} data points occur over the range [–8, –4], corresponding to four orders of magnitude in concentrations. Just over 10% of the substances have complete toxicity profiles, 43% of the data points are at threshold values, and 12% of the possible data points are missing. It is encouraging to note that there are four orders of magnitude difference between the high and low threshold values, representing the difference between 0.1 mM and low 10 nM concentrations. This range is good for the differentiation of profiles based on activity, as required concentrations for drug leads are typically in the low μM range. It is expected that the vast majority of molecules in an HTS dataset will be inactive against multiple or all targets of interest, however it is desirable that the distribution of activity data span several orders of magnitude for a large set of substances.

4.2.3 Imputing Missing Data

Missing values are a problem for all data analysts. If only a few substances had missing values, we could simply omit them from the dataset, but in this case many substances have missing log GI_{50} observations for multiple assays. Removing these substances drastically reduces the dataset from 43474 to 4824 substances, which would diminish the predictive power of the resulting model. We implemented a common strategy used in linear models in order to preserve the size of the predictor subset by imputing the missing data, null values, with the mean log GI_{50} of all the screening data for each respective substance.[19]

4.2.4 Data Mining with PCA

Principal component analysis was used for two purposes. First, PCA was used for validation purposes to ensure that we did not inadvertently skew our matrix of toxicity profiles by imputing 12% of the missing assay data with the mean log GI_{50} for each respective substance. PCA was performed on a matrix of random values from [–9.0, –4.0] to show that the first principal components of the datasets containing experimental and imputed data was not due to mean centering. Also, PCA of the non-imputed dataset was performed using the covariance matrix derived from eq 1, such that all existing pairs of log GI_{50} values between assays were considered. The covariance matrix is a square 59×59 matrix, where each row *i* and column *j* correspond to their respective NCI60 cancer cell line. Let the matrix *X* be defined by $x_{k,j}$. Where $x_{k,j}$ is the log GI_{50} for substance *k* and cell line *j*. Then, $\bar{x}_j$ is the mean log GI_{50} of the substances for cell line *j*. Now, let *k* refer to the substances having experimental log GI_{50} data for assays *i* and *j*, such that $\bar{x}_i$ and

$\bar{x}_j$ refer to the mean log GI_{50} values for the substances having experimental log GI_{50} data for both assays *i* and *j*. Note, $i = j$ is allowed in eq 1. Validation for using an imputed dataset can be established by showing a high correlation between the components of the eigenvectors responsible for the majority of log GI_{50} variance from the imputed and non-imputed datasets.

$$C_{i,j} = n^{-1}\sum_{k=1}^{n} x_{k,i}x_{k,j} - \bar{x}_i\bar{x}_j \tag{1}$$

PCA was then used to extract interesting features of the NCI60 for further investigations. MOE's PCA tool can output the importance of substance contribution to the principal components based on the respective log GI_{50} toxicity profile of *X*. This is done by first calculating the sample average vector $\mathbf{x} = \left[\bar{x}_1, \cdots, \bar{x}_{59}\right]$ and covariance matrix *C* based on the matrix of toxicity profiles for all *n* substances. *C* is diagonalized such that $C = Q^T DDQ$, where *Q*, the PCA transform, is orthogonal and *D* is diagonal-sorted from top left to bottom right. There are *p* non-zero diagonal values in *D*, the square roots of the eigenvalues of *S*, corresponding to the principal components. If we take the $p \times n$ matrix Z $= Q(X - \mathbf{x})$, such that *Z* has identity covariance and zero mean, there exists a *p*-vector of the form $z_i = Q(x_i - \mathbf{x})$, where the *p* components of each z_i can be taken as the relative weights for the respective principal components corresponding to each substance. Hence, the z_{ip} values are weights for each substance *(i)* and principal component (*p*). The substances most responsible for a particular principal component's variance have the largest magnitudes of z_{ip} values.

4.2.5 Predictive Binary Decision Tree

We applied two concepts in the construction of a binary decision tree. First, consider the 'ideal' fingerprint where the different descriptors' occurrence are statistically independent and each descriptor evenly divides a dataset of n molecules with some property value, such as mean log GI_{50}. The 'ideal' fingerprint has length n_d, where n_d is the minimum number of descriptors to uniquely identify all n molecules in the training set.

$$n_d = ceil(\log_2 n) \tag{2}$$

Using this concept we have applied the MACCS keys as our base fingerprint. The MACCS key which most evenly divides the substances at any given node in our tree is given weight when making our branching decision. As our goal was to not only reduce the number of substances at each child node, but also to reduce the range of mean log GI_{50}, a weighted accuracy factor was also included in our branching decision. Branching is allowed to continue as long as a MACCS key exists that can divide a node into children nodes, each containing no fewer than two substances. If branching cannot occur, the node is taken to be a leaf. Thus the trained binary tree has a MACCS key at each nonterminal node, and a prediction value at each leaf equal to the average of the mean log GI_{50}'s of the training set substances at each respective leaf node. Eqs 3-8 describe how decisions are made at each node.

$\boldsymbol{h}_k$ and $\hat{\boldsymbol{h}}_k$ are binary vectors with length equal to the number of substances m at a node with $\boldsymbol{h}_k$ representing the hit profile and $\hat{\boldsymbol{h}}_k$ the inverse hit profile with respect to MACCS key k. $\left|\mathbf{h}_k\right|$ and $\left|\hat{\mathbf{h}}_k\right|$ are the sum totals of the hits and misses for the respective profiles. See eqs 3-5.

$$\left|\mathbf{h}_k\right| = \mathbf{h}_k \bullet \mathbf{h}_k = \sum_{j=1}^{m} h_{kj} \tag{3}$$

$$\left|\widehat{\mathbf{h}}_k\right| = \widehat{\mathbf{h}}_k \bullet \widehat{\mathbf{h}}_k = \sum_{j=1}^{m} \widehat{h}_{kj} \tag{4}$$

$$m = \left|\mathbf{h}_k\right| + \left|\widehat{\mathbf{h}}_k\right| \tag{5}$$

Eq 6 defines the idealness score, $S_{1,k}$, for MACCS key k where $0.5 \le S_{1,k} \le 1.0$. A value of 0.5 indicates that the substances at a particular node are evenly divided into children nodes, whereas a score of 1.0 indicates that one child node contains all of the substances and the other contains none.

$$S_{1,k} = \frac{\max[\left|\mathbf{h}_k\right|, \left|\widehat{\mathbf{h}}_k\right|]}{m}, \text{ for } k = 1,\ldots,166 \tag{6}$$

Eq 7 defines the accuracy score $S_{2,k}$, for MACCS key k where $0 < S_{2,k} \le 1$. r_k and $\hat{r}_k$ are the log GI_{50} ranges at the two child nodes for MACCS key k. Values less than 1.0 represent child nodes with smaller log GI_{50} ranges than the parent node. Values closer to zero reflect child nodes with more narrow log GI_{50} ranges.

$$S_{2,k} = \frac{r_k\left|\mathbf{h}_k\right| + \hat{r}_k\left|\widehat{\mathbf{h}}_k\right|}{\max\limits_{k}[r_k, \hat{r}_k] \times m}, \text{ for } k = 1,\ldots,166 \tag{7}$$

Eqs 6 and 7 have been normalized such that the lower values of the scores are desirable. Eq 8 describes how the final decision is made at each node and considers the ability of a particular MACCS key to evenly split the substances and at the same time minimize the log GI_{50} range of the children nodes. Weighting factors for $S_{1,k}$ and $S_{2,k}$ were systematically determined by spawning several decision trees and varying the weights. The final weighting scheme was the one resulting in the highest fit R^2 based on correlating the mean of all log GI_{50} values of the leaf and experimental log GI_{50} values from our training set.

$$S = \min_{k}[S_{1,k} + 5.0 \times S_{2,k}] \tag{8}$$

Substances were randomly divided into ten subsets of nearly equal size. The predictive binary decision tree was trained on 90% of the substances and validated on the remaining 10% in an attempt to maximize two relevant metrics: R^2 and the binning of cytotoxic (mean log $GI_{50} < -5.0$) and non-cytotoxic (mean log $GI_{50} > -5.0$) substances. Cross validation was performed on different training and test sets based on the random classification of the substances. A final validation was performed to ensure that the predictive nature of this procedure was not an artifact by randomly assigning the mean log GI_{50} values to different substances within the training set and then rebuilding the decision tree. There was essentially no log GI_{50} correlation ($R^2 < 0.02$) between the predicted values for training set and these randomly assigned values.

4.2.6 Prediction using Least Squares Fit

A more accurate prediction model was devised by randomly dividing the 4284 substances having complete toxicity profiles into equal sized training and test sets. The experimental log GI_{50} values for each cell line were then correlated to the mean log GI_{50}. Starting with the cell line whose log GI_{50} values have the highest R^2 with the mean log GI_{50} values, we applied forward stepwise linear regression with the constraint that no more than one cell line from each class can be used. The final model was validated on the test set. A second validation was performed using the remaining 38650 substances with incomplete toxicity profiles.

4.3 Results and Discussion

We chose to adopt this strategy of imputing missing data with mean values to maintain the largest possible set of substances for mining. In order to validate that imputing the mean log GI_{50} did not inadvertently skew the overall dataset, PCA was performed on four different datasets: the complete set of 4824 substances having complete toxicity profiles, the imputed set consisting of all 43474 substances with the imputed mean log GI_{50} for the missing values respective to each substance, the nonimputed dataset consisting of all 43474 substances with missing values, and a random dataset modeled after the complete set having 4824 entries each with 59 random data points ranging from [–9, –4]. Figure 4.4 compares the first four principal components, accounting for over 92% of the log GI_{50} variance for all datasets. There is extremely good correlation between the components of the eigenvectors for these principal components relating to all datasets except for the one with randomly assigned data. As expected, the correlation of principal components between the datasets continues to degrade when examining the lower order components. Since the components of the eigenvectors are strongly correlated, we can assume that the imputed mean values did not severely impair the quality of the dataset. A similar analysis, comparing the principal components of the set of substances having complete toxicity profiles and those having greater than or equal to 90% complete toxicity profiles, showed an even closer fit to the non-imputed dataset depicted in Figure 4.4 than any of the other datasets.

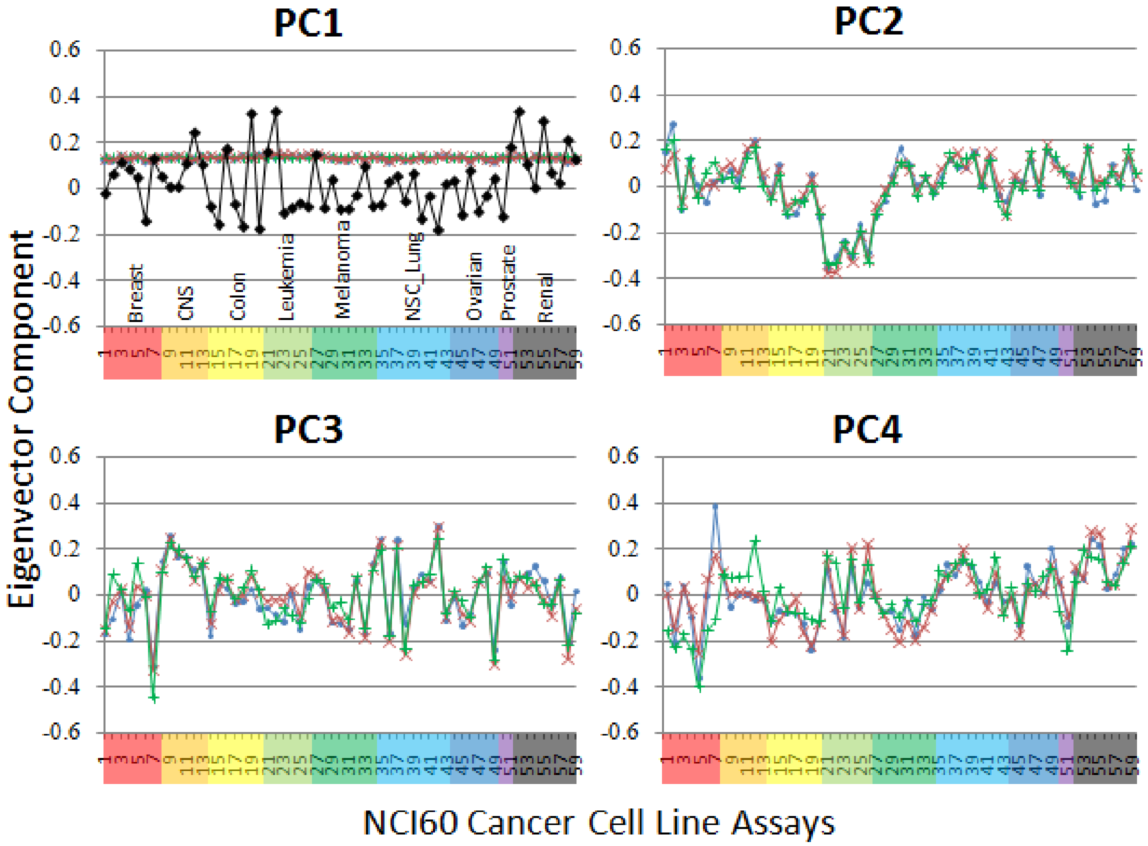


Figure 4.4. The x-axis represents the 59 cancer cell line assays and has been color coded and organized as described in Figure 4.1. The y-axis represents the components of the eigenvectors for each respective principal component. (—◆—) represents random toxicity profiles for 4824 entries, each cell within the range [–4, –9], plotted only for PC1; (—●—) are from the 4824 substances having complete profiles; (—×—) are from the 43474 substances using imputed data for missing values; (—+—) are from the 43474 substance ignoring missing values.

It was interesting to note that PC1 accounts for over 89% of the log GI_{50} variance for all the datasets with real experimental values, while PC1 for the dataset having randomly generated log GI_{50} values only accounted for 2% of the variance. Aside from the random dataset, all the components of the eigenvector for PC1 were found to be approximately equal in magnitude (even more so in the case of the non-imputed dataset). Given the high explanatory power of this component, we see that many substances tend to be uniformly toxic across all the NCI60 assays. Even when the 176 and 6203 substances having only threshold log GI_{50} values are eliminated from the respective datasets of complete toxicity profiles and the one including imputed values for all substances, PC1 still accounts for over 88% of the total log GI_{50} variance. Since this pattern dominates in both the imputed and complete datasets, while the random dataset deviates and explains very little, we can conclude that this is not an artifact of our procedure. Since PC1 corresponds to uniform log GI_{50} across all cell lines, we can artificially remove this component by mean centering each substance's toxicity profile and performing PCA once again. Indeed, PC2 in Figure 4.4 from the first analysis becomes the first principal component in the new analysis. In the original PCA, PC2 explained approximately 1.2% of the log GI_{50} variance. As the first principal component in the new analysis, it explained 12% of the log GI_{50} variance and maintained the eigenvector components from the first analysis. In any case, the result that most compounds show uniform toxicity (high or low) across all cell lines is hardly surprising, but it leads to our least squares model that greatly reduces the effort required for screening compounds.

The PC2 corresponds to a rather uniform level of toxicity across most cancer cell lines with the majority of its components between –0.17 and 0.17. It is very interesting to note that all six leukemia cell lines (RPMI_8226, SR, CCRF_CEM, K_562, MOLT_4, HL_60(TB)) have eigenvector components less than –0.2. For the complete subset, CNS_SNB_75 and Breast_HS578T have component values greater than 0.2. The increased absolute value of the eigenvector components corresponds to certain cancer cell lines, indicating that there are groups of substances for which these cell lines are either more sensitive or more resistant. PC2 is responsible for just over 1% of the total variance over all cancer cell lines. We attempted to relate the values of the eigenvector components to the doubling times of these outstanding cancer cell lines and noted that the leukemia cell lines doubling times (19.6–33.5 hours) are approximately half that of CNS_SNB_75 (62.8 hours) and Breast-HS578T (53.8 hours).[12] Unfortunately, this trend does not follow for the remainder of the cell lines with eigenvector values close to –0.2 and 0.2. The outstanding feature of PC2 is that it identified all the leukemia cell lines.

We speculate that PC2 may have identified the leukemia cell lines due to the fact that a greater portion of their surface areas is available to exogenous substances. The leukemia cell lines are grown in vials suspended in solution, whereas the other cancer cell lines are grown on plates and require attachment to the plate wall, reducing the exposed surface area. Therefore, it is plausible that the leukemia cancer cells are more susceptible to toxic substances due to increased surface area exposure.

In order to substantiate this claim, we first identified the substances most responsible for PC2 using MOE's PCA tool. We found that over 4000 substances with $|z_{i2}| \geq 1.0$ and removed them from the complete dataset in order to examine the eigenvectors produced by PCA in their absence. By removing the substances with $|z_{i2}| \geq 1.0$, we were able to reduce the predominance of this feature tenfold. Instead of being the second principal component and accounting for over 1% of the total log GI_{50} variance, the leukemia cancer cell lines were identified in the ninth principal component along with several other cancer cell lines with absolute values of eigenvector components greater than 0.2 and accounting for less than 0.1% of the total log GI_{50} variance. Furthermore, there are 782 substances with $|z_{i2}| \geq 2.0$. Notably, less than 5% of the data was imputed with 313 substances having no imputed data whatsoever. 543 of these substances, 228 with no imputed values, have fairly uniform increased cytotoxicity against the leukemia cell lines compared to the other 53 cancer cell lines. The remaining 239 substances, 75 with no imputed values, were shown to be slightly less cytotoxic to the leukemia cell lines on average. The more sensitive nature of the leukemia cell lines is illustrated by the scatter plot in Figure 4.5 between the mean log GI_{50} for the leukemia cell lines and the mean log GI_{50} for the non-leukemia cell lines for the 543 substances of interest. The plot shows that there is approximately 1.8 orders of magnitude difference between the respective mean log GI_{50} of the leukemia and other cell lines. Also, the mean log GI_{50} distributions of the leukemia cell lines are shifted to significantly higher levels of toxicity than the distributions of the non-leukemia cell lines as seen in Figures 4.6a and b. As a side note, only 33 of the 782 substances most responsible for PC2, $|z_{i2}| \geq 2.0$, were uniformly more toxic to the leukemia cell lines than all others of the NCI60. Even through the substances responsible for PC2 do not exhibit the highest levels of cytotoxicity for all leukemia cell lines, they do show a tendency to be uniformly more toxic for the majority of the leukemia cell lines.

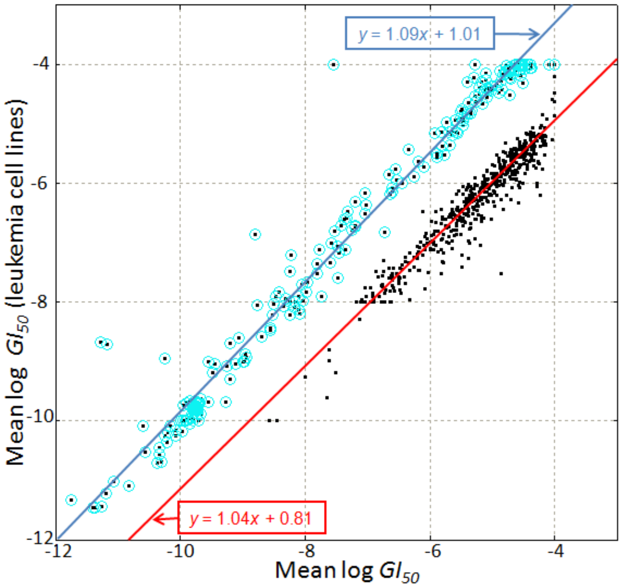


Figure 4.5. (1) The fit line through 543 substances determined by PC2 (▪), which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (2) The fit line for 239 substances determined by PC2 (⊙), which are responsible for decreased cytotoxicity of the leukemia cell lines compared to the non-leukemia cancer cell lines.

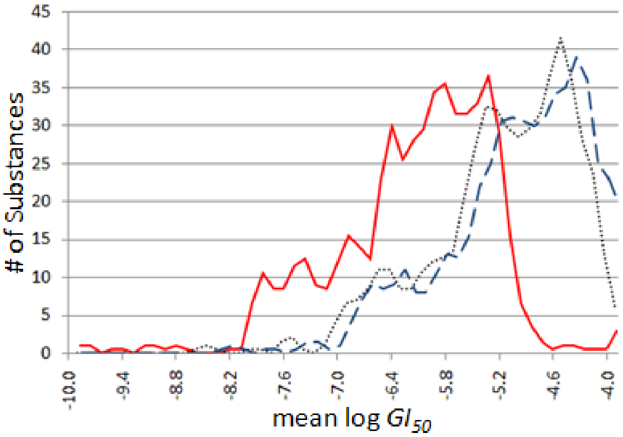


Figure 4.6a. This histogram depicts the 543 substances determined by PC2 which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (——), (·········), and (— —) represent the mean log GI_{50} of the leukemia cell lines, all cell lines, and the non-leukemia cancer cell lines, respectively.

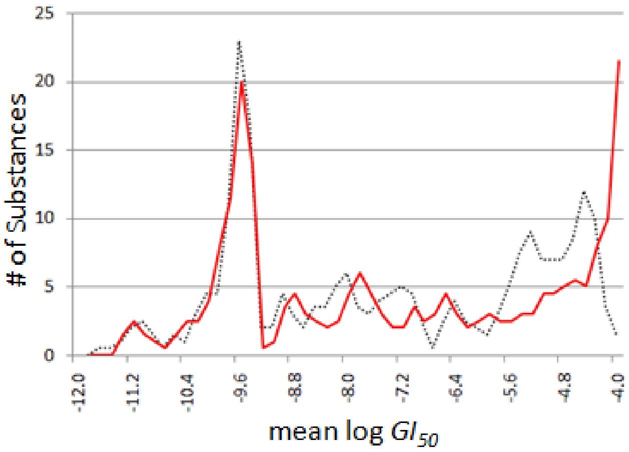


Figure 4.6b. This histogram depicts the 239 substances determined by PC2 which are most responsible for decreased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (——) and (········) represent the mean log GI_{50} of the leukemia cell lines and all cell lines, respectively.

We examined the layout of the eigenvector components for the latter principal components looking for outstanding features similar to that of PC2. While no other classes were uniformly identified having all cancer cell lines with the absolute value of the eigenvector components greater than 0.2, we did find several principal components that identified a few cancer cell lines with the absolute value of the eigenvector components significantly greater than 0.2. The two principal components with few outstanding eigenvector components accounting for the largest log GI_{50} variance are depicted in Figure 4.7.

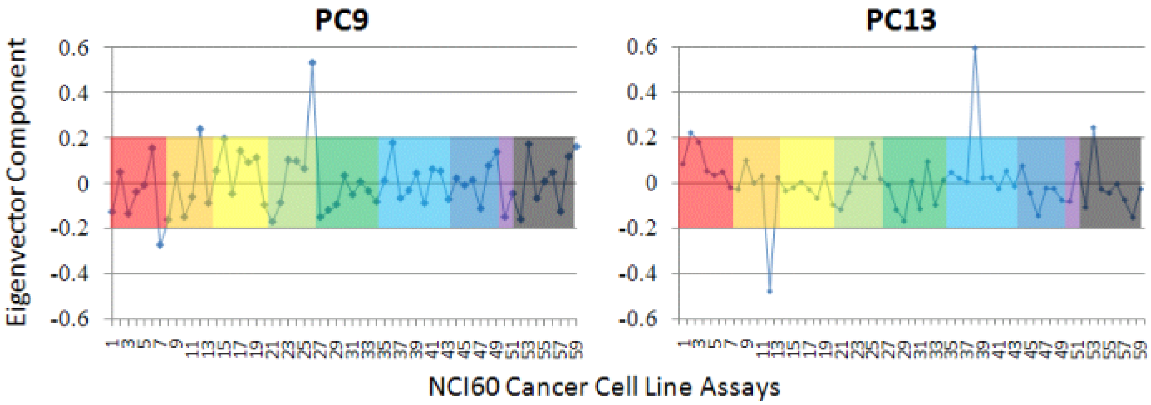


Figure 4.7. PC9 specifically identifies the Leukemia_SR cancer cell line related eigenvector component. PC13 identifies NSC_Lung_Hop_92 and CSN_SNB_75. The axis and color coding scheme are described in Figure 4.1.

Of 600 substances found to be responsible for PC9 with $|z_{i9}| \geq 1.0$, only 45 showed specific cytotoxicity against the Leukemia_SR cell line, i.e. the concentrations of substances required to inhibit cell growth for the Leukemia_SR cell line were two to four orders of magnitude lower than the concentration necessary for the remainder of the NCI60 cell lines. We found that most of these substances are structurally dissimilar. Only one pair, the phosphonium molecules, show significant structural similarity with a Tanimoto similarity coefficient $S_{tan} = 0.92$ based on the MACCS keys, as depicted in Figure 4.8a. Figures 4.8b and c illustrate two other molecules and their respective highest scoring (most similar) match within this set of molecules. In the game of fingerprint based similarity searching, one typically does not consider molecules with a $S_{tan} < 0.70$ as structurally similar. Most of the 45 substances exhibit multi-cyclic ring systems with aromatic components, and we have identified a few purine and pyrimidine derivatives. Figure 4.9 depicts a histogram of the highest S_{tan} for each of the 45 substances when compared to the other respective substances in the dataset. The mean S_{tan} for each molecule and its most structurally similar pair of the set of 45 substances was 0.55.

Having established that these substances are structurally dissimilar yet share similar toxicity profiles, we conjecture that these substances do not share a common mechanism of action against the Leukemia_SR cell line, which is contrary to the belief that substances sharing the same toxicity profiles also follow the same mechanism of activity.[20] We are not implying that toxicity profiles cannot be used for predicting mechanism of activity, but rather we believe that there may be several viable Leukemia_SR cell specific targets which when activated lead to cell death based on different mechanisms of action.

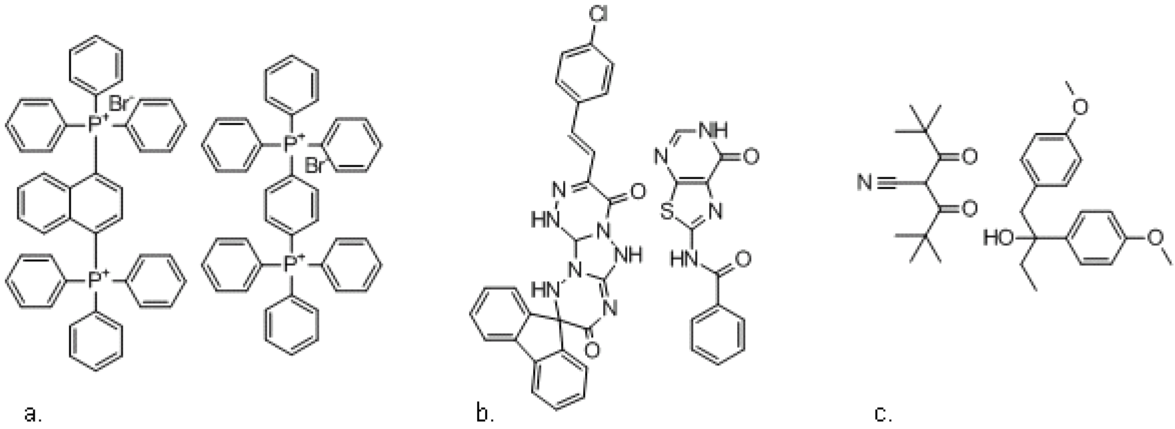


Figure 4.8. Comparison of molecules from the 45 substances having specific toxicity for the Leukemia_SR cancer cell line. The left molecule of each pair has the greatest similarity to the right one. (a) S_{tan} = 0.92, (b) S_{tan} = 0.57, (c) S_{tan} = 0.24.

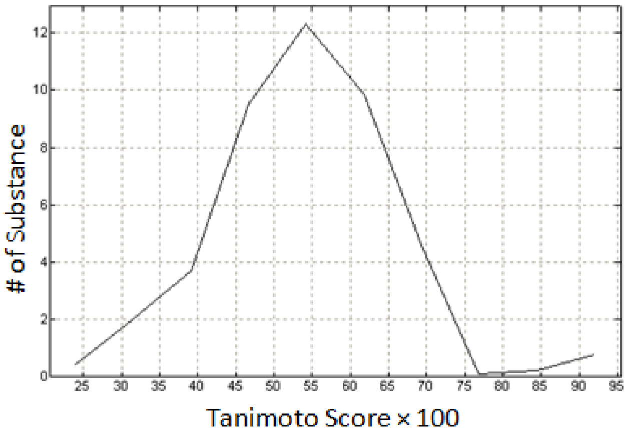


Figure 4.9. Histogram of the highest S_{tan} derived by comparing pairs of substances using the MACCS fingerprint on the set of 45 substances specifically targeting the Leukemia_SR cancer cell line. Mean = 0.55, σ = 0.13, min = 0.24, max = 0.92.

PC13 identifies the Lung Hop_92 and CNS_SNB_75 cell lines. Even fewer substances are responsible for this principal component, as it is responsible for less than 0.3% of the total system variance. The analysis of the substances responsible for toxicity profiles matching the landscape of PC13, while similar to that of PC9, can neither confirm nor reject our hypothesis that multiple mechanisms of actions may be in play when examining cytotoxic substances with similar toxicity profiles.

Several groups have recently published their ability to predict cytotoxicity.[3,4,5,6,7] Table 4.1 contains a summary of the methods used and results obtained, including the results in our study. Our method used the MACCS keys, paying attention to narrowing the range of log GI_{50} in all children nodes of our decision tree. While we were unable to achieve reliable results in a leave some out cross-validation study using only the substances in the dataset with complete toxicity profiles, we did derive a predictive model that achieved R^2 = 0.53 and RMSE = 0.71, using the dataset having imputed values for less than 10% of the cancer cell lines. See Figure 4.10. While this number is not outstanding, it allows for improved discrimination between toxic vs. non-toxic substances. Furthermore, we achieve similar results when using different 90:10 training:test splits of the dataset. When taking log $GI_{50} = -5$ as the cutoff, we were able to correctly classify 83% of all substances and 82% of the cytotoxic substances in our test set. With a log GI_{50} cutoff $= -6$ we correctly classified 93% of all substances in our test set, but only 72% of those taken to be cytotoxic. See Table 4.2 for the results presented as a confusion matrix.

Table 4.1. Comparative summary of recent cytotoxicity study results

Method	**Data Set**[a]	**Training Set Size**	**Test Set Size**	$\mathbf{R^2}$	**log $\mathbf{GI_{50}}$ Cutoff**	**Classified Correct(%)**[b]	**Improve**[c]
Neural Network[3]	19 libraries	8298	2000	N/A	N/A	73[d]	N/A
QSPR[4]	NCI60	27000	N/A	0.67	-6	N/A	× 10
Proteomic Profiling[5]	NCI60[e]	2/3	1/3	N/A	N/A	60	N/A
QSAR[6]	37 cpds.[f]	4/5	1/5	0.73	N/A	N/A	N/A
Random Forest[7]	NCI60	≤ 42723	≤ 42723	N/A	-5	76[g]	N/A
Decision Tree	NCI60	21763	2015	0.54	-5	83	× 4
					-6	93	× 10
Least Squares Fit	NCI60	2412	2412	0.99	-5	96.7	× 2
					-6	99.1	× 5
			38650	0.99	-5	97.9	× 6
					-6	99.8	× 31

[a] The NCI60 is evolving, i.e. newer versions are larger and have more complete log GI_{50} toxicity profiles. [b] Substances with log GI_{50} greater than or equal to the cutoff value are considered non-toxic, while those with log GI_{50} values less than the cutoff value are considered toxic. All percentages were rounded down. [c] Improvement over random selection is based on the respective cutoff value. [d] 20% of all substances were unclassified and assumed to be incorrect. [e] The dataset consisted of 118 drugs from the NCI60. [f] The dataset consisted of 37 naphthoquinone ester derivatives. [g] 76% is the average of correctly classified substances for the individual cancer cell lines, not the mean log GI_{50}.

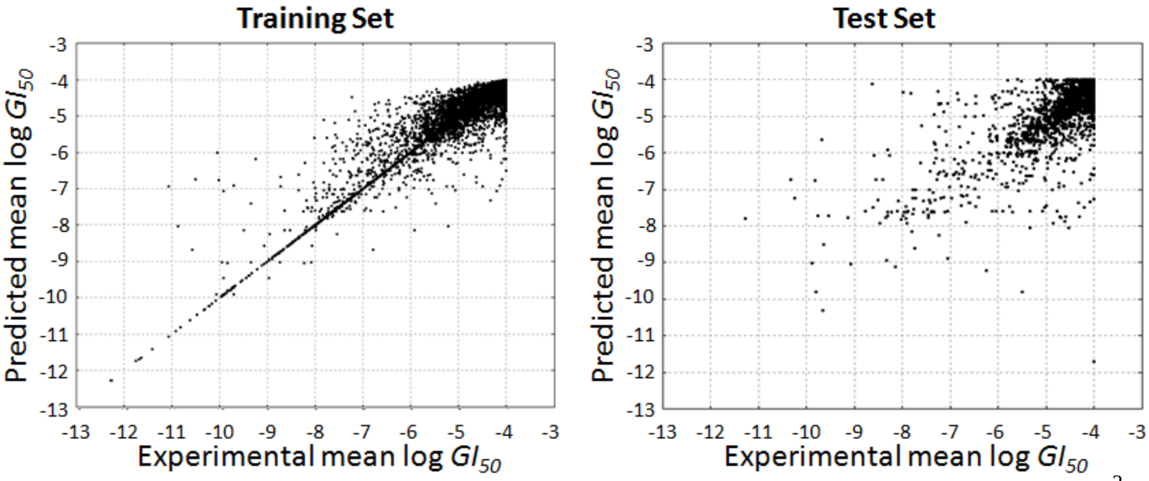


Figure 4.10. Prediction results using a decision tree. (a) The training set had a fit $R^2 = 0.9284$ with RMSE = 0.2285. (b) The test set exhibited predictive $R^2 = 0.5432$ with RMSE = 0.7119.

Table 4.2. Decision tree results for cutoff log GI_{50} = –5.0

Confusion Matrix	Cutoff log GI_{50} = –5.0			Cutoff log GI_{50} = –6.0		
	Pred. non-toxic	**Pred. Toxic**	**% Correct**	**Pred. non-toxic**	**Pred. Toxic**	**% Correct**
Exp. non-toxic	1248	252	83.2	1744	97	94.7
Exp. Toxic	90	425	82.5	49	125	71.8

Finally, with least squares fitting we achieved excellent prediction using equal sized randomly selected training and test sets using the 4824 substances with complete toxicity profiles, i.e. no data was imputed. While still skewed towards log GI_{50} = –4.0, this dataset was more uniformly distributed than the dataset used when training and testing our decision tree. If one considers log GI_{50} = –5.0 as the cutoff value between toxic and non-toxic substances, then the dataset is evenly distributed. Random selection was performed by first sorting the substances by ascending mean log GI_{50} and then dealing them one-by-one into the training and test set. This ensured that there was an even distribution between the two sets. We identified the Ovarian_OVCAR-8 cancer cell line as having the best fit to the mean log GI_{50}. Applying forward stepwise linear regression and allowing only one cancer cell line from each class to be included in the model yielded eq 9.

$$
\begin{aligned}
\text{mean log } GI_{50} = &-0.08234 \\
&+0.06395 \times Breast\_HS\_578T \\
&+0.13046 \times CNS\_U251 \\
&+0.08476 \times Colon\_HCC\_2998 \\
&+0.09574 \times Leukemia\_K\_562 \\
&+0.13151 \times Melanoma\_M14 \\
&+0.14644 \times NSCLung\_NCI\_H23 \\
&+0.09082 \times O\,\mathrm{var}\,ian\_SK\_OV\_4 \\
&+0.10767 \times Prostate\_PC\_3 \\
&+0.12995 \times Renal\_CAKI\_1
\end{aligned} \qquad (9)
$$

When applied to the test set, 96.7% and 99.1% assignment accuracy was obtained. The predictive R^2 was found to be greater than 0.99 and RMSE less than or equal to 0.09. See Figures 4.11a and b. While this method is superbly predictive for mean log GI_{50}, it requires that assay data be available for 9 of the 59 available NCI60 cancer cell lines. One surprising observation shown in Table 4.1 based on the test set of 2412 substances with complete toxicity profiles was the relative lack of improvement over random selection when compared to our decision tree model. This may be attributed to the distributions of the test sets. In the test set described for the decision tree there was approximately 25% and 9% chance that a randomly selected substance would be toxic based on the respective log GI_{50} cutoffs of –5.0 and –6.0. With the least squares fit derived through stepwise linear regression, the respective chances are 50% and 18%. Based on the dataset distribution, the second method can only improve half as much as the first. Thus, the quality of prediction should never be based on improvement over random selection alone. To verify the robustness of this method, we also used the least squares fit model to predict the mean log GI_{50} for the remaining 38650 substances with toxicity profiles containing imputed data. We found that the correlation coefficient was little changed and RMSE improved to 0.06. This was expected due to the incidences of imputed mean log GI_{50} values for all missing data. See Figure 4.11c. Here significant improvement over random selection was seen. This was due to the change in ratio of toxic:non-toxic substances. The ratio for the test set having 2412 substances with the more complete toxicity profiles was 1200:1212 and 428:1984 for the respective log GI_{50} cutoff values of –5 and –6. Whereas the ratios for the test set containing having 38650 substances with imputed data were 5268:33382 and 1215:37435 for same respective log GI_{50} cutoff values. The ratio of toxic:non-toxic is inversely proportional to the improvement over random. When considering the improvement over random selection, one must also consider the classification ratio, as the maximum improvement over an evenly divided set 1:1 (50% chance to correctly classify any substance) is twofold, as was the case with our unimputed test set. In our case with 1215:37435 toxic:non-toxic substances it is possible to improve the identification of toxic substances by ~32 fold (38650/1215).

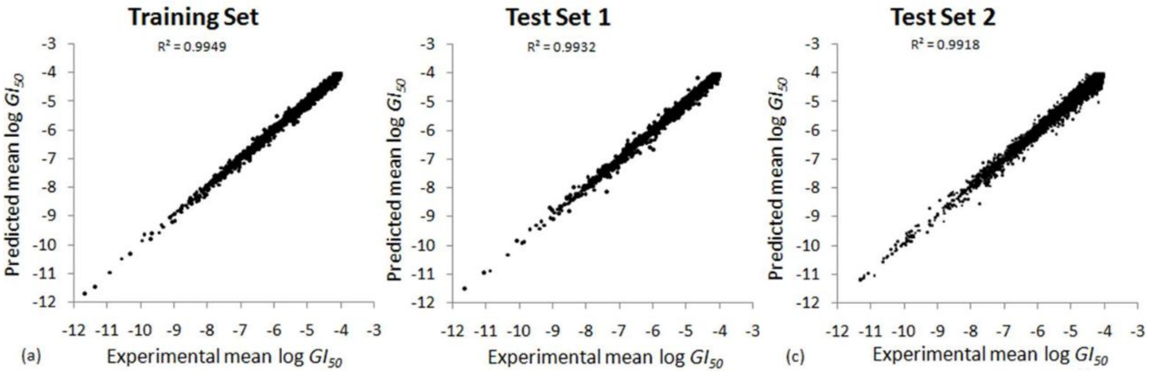


Figure 4.11. Prediction results using the least squares fit and forward stepwise regression. (a) The training set consisted of 2412 complete toxicity profiles and had a fit R^2 = 0.9949 and RMSE = 0.07855. (b) Test set 1 consisted of 2412 complete toxicity profiles and had R^2 = 0.9932 and RMSE = 0.09000. (c) Test set 2 consisted of 38650 incomplete toxicity profiles and had R^2 = 0.9918 and RMSE = 0.06195. The toxicity profiles used in (a) and (b) contained no imputed data, whereas the toxicity profiles used to determine (c) had imputed data.

4.4 Conclusions

Refining chemical datasets can facilitate the process of drug development by helping to minimize the high attrition due to poor ADMET during the clinical phases.[21] The largest problem with current public domain chemical and biological activity data is lack of curation procedure and QC. We have shown that even in the cases where curated datasets are available, one must carefully evaluate the data in order to ensure the greatest accuracy for data mining purposes.

In this work we have preprocessed and validated the quality of a large reliable subset of the NCI60 for data mining toxicity profiles. Here we have used PCA to validate the use of larger training sets by demonstrating that PC1 was not an artifact due to imputing the mean log GI_{50} and further establishing a correlation between the components of the eigenvectors for the principal components responsible for 92% of the total log GI_{50} variance. The same steps can be used to refine screening data from multiple assays, whether they include a subset of the NCI60 assays, a combination of the NCI60 and other biological screens, or any selection of HTS drawing their activity data from a common set of substances. Further examination of the PCA results led to interesting deductions regarding the general landscape of a dataset. In this case over 89% of the total system variance relates to the generally cytotoxic nature of substances, and the leukemia cell lines behaved differently. PC2, responsible for ~1.2% of the log GI_{50} variance, identifies the leukemia cell lines. Finally, it is possible to extract the substances responsible for the principal components in order to mine for similarities or differences. Here PC9 and PC13 indicate that multiple mechanisms may share similar toxicity profiles based on the NCI60.

Based on our analysis of the substances responsible for PC9 and the latter principal components, we have found the chemical structures to be of such diversity that it would be impossible to derive the underlying QSARs based on the limited size of the dataset and the likelihood that the substances' cytotoxic natures are due to different mechanisms of action. QSPR investigations may provide valuable insights regarding the compounds responsible for these principal components. Identifying QSARs within the larger groups of substances responsible for PC1 and PC2 based on structurally similar subsets of these substances may also be possible. However, as it stands both QSAR and QSPR investigations are beyond the scope of this work.

We derived two predictive methods: one using a binary decision tree, the other using forward stepwise linear regression and least squares fit. In our study, greater than tenfold enrichment over random selection can be expected for substances with mean log $GI_{50} < -6$, using both of our methods based on our subset of the NCI60. The least squares model further offers a very accurate method for determining the level of general cytoxicity for substances that need not be limited to the NCI dataset. While it has shown improved results over past cytotoxicity prediction methods, there is one caveat, that future predictions on substances outside of the NCI60 cannot be performed completely *in silico*. This pitfall is also our major discovery, i.e. only 9 of the 59 available NCI60 cancer cell lines are required to implement our model. While selectivity is one of the main goals with antitumor agents, our method can flag nonselective substances with significantly low mean log GI_{50} values for early elimination. Selective substances will not be flagged as they have higher mean log GI_{50} values for the majority of cancer cell line assays, deemphasizing the increased toxicity of selective substance for only a few cancer cell lines. The goal is to flag compounds as early as possible for elimination in the drug discovery pipeline in order to save time and money through streamlining the toxicity detection system while decreasing the overall demand on chemical and biological resources.

Being able to more accurately identify non-specific cytotoxins brings to light two new questions relevant to the Food and Drug Administration's approval of current and new drug substances. (1) Are non-cytotoxic molecules more "drug-like" for drugs that are not meant to be anticancer agents? (2) Are drug molecules less cytotoxic than non-drugs? Answering these questions might help to further access the overall generalizabilty of our methods and minimize the attrition rates of new potential drugs in the later stages of drug development pipeline.

This work has been published and has the following reference: Lee, A. C.; Shedden, K.; Rosania, G. R.; Crippen, G. M. Data Mining the NCI60 to Predict Generalized Cytotoxicity. *J. Chem. Inf. Model.* **2008**, *48*, 1379–1388.

4.5 References

(1) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nature Rev. Drug Discov.* **2002**, *1*, 882–894.

(2) Acton, G. Toxicogenomics and Predictive Toxicology: Market and Business Outlook, **2004**. http://www.vivogroup.com/reports.html (accessed Feb 29, 2008), http://www.the-infoshop.com/study/cd25153_toxicogenomics.html (accessed Feb 29, 2008).

(3) Molnár, L.; Keserű, G. M.; Papp, Á.; Lőrincz, Z.; Ambrus, G.; Darvas, F. A neural network based classification scheme for cytotoxicity predictions: Validation on 30,000 compounds. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1037–1039.

(4) Huang, R.; Wallqvist, A.; Covell, D. G. Assessment of in Vitro Activities in the National Cancer Institute's Anticancer Screen with Respect to Chemical Structure, Target Specificity, and Mechanism of Action. *J. Med. Chem.* **2006**, *49*, 1964–1979.

(5) Ma, Y.; Ding, Z.; Qian, Y.; Shi, X.; Castranova, V.; Harner, E. J.; Guo, L. Predicting Cancer Drug Response by Proteomic Profiling. *Clin. Cancer Res.* **2006**, *12*(15), 4583–4589.

(6) Saíz-Urra, L.; Maykel, P. G.; Tiejeira, M. 2D-autocorrelation descriptors for predicting cytotoxicity of napthoquinone ester derivatives against oral human epidermoid carcinoma. *Bioorg. Med. Chem.* **2007**, *15*, 3565–3571.

(7) Guha, R. Flexible Web Service Infrastructure for the Development and Deployment of Predictive Models. *J. Chem. Inf. Mod.* **2008**, *48*, 456–464.

(8) Berkowitz, B.A.; Sachs, G. Life Cycle of a Blockbuster Drug. *Mol. Interventions* **2002**, *2*, 6–11.

(9) NCBI: PubChem Project: National Center for Biotechnology Information, Bethesda, MD 2008. http://pubchem.ncbi.nlm.nih.gov/ (accessed Jan 8, 2008).

(10) NIH Roadmap: Molecular Libraries and Initative: National Institutes of Health, Bethesda, MD 2005. http://nihroadmap.nih.gov/molecularlibraries/ (accessed Jan 8, 2008).

(11) Austin, CP; Brady, L.S.; Insel, T.R.; Collins, F.S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.

(12) DTP: Developmental Therapeutics Program NCI/NIH. http://dtp.nci.nih.gov/ (accessed Jan 8, 2008).

(13) Richard, A.; Gold, L.S.; Nicklaus, M. Chemical Structure Indexing of Toxicity Data on the Internet: Moving Toward a Flat World. *Curr. Opin. Drug Discov. Devel.* **2006**, *9*, 314–325.

(14) Ekwall, B. Screening of toxic compounds in mammalian cell cultures. *Ann. N.Y. Acad. Sci.* **1983**, *407*, 64–77.

(15) Rowan, A.N.; Berlin, A.; Becking, G.C.; Ekwall, B.; Fernicola, N.; Friedrich, J.; Gournar, M.I.; Kaloyanova, F.; Krishna Murti, C.R.; Ordonez, B.; Sanockij, I.V.; Stammati, A.L. In *Short-term Toxicity Tests for Non-genotoxic Effects*; Bourdeau, P.; Somers, E.; Richardson, G.M.; Hickman, J.R., Eds.; John Wiley & Sons Ltd: New York, 1990, Chapter 2, pp 7–9.

(16) Wang, H.; Klinginsmith, J; Dong, X.; Lee, A.C.; Guha, R.; Wu, Y.; Crippen, G.; Wild, D. Chemical Data Mining of the NCI Human Tumor Cell Line Database *J. Chem. Inf. Mod.* **2007**, *6*, 2063–2076.

(17) *MOE: Molecular Operating Environment ver. 2007.0902*: Chemical Computing Group, Montreal, Quebec, Canada **2007**.

(18) Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 244–255.

(19) Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman and Hall/CRC; New York; 1990; Vol. 43, Chapter 6, pp 166.

(20) Zaharevitz, D.; Holbeck, S.; Bowerman, C.; Svetlik, P. COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graph. Mod.* **2002**, *20*, 297–303.

(21) Van de Waterbeemd, H.; Gifford, E. ADMET in *Silico* Modeling: Towards Prediction Paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204.

Chapter 5

pK_a Prediction of Monoprotic Small Molecules the SMARTS Way

5.1 Introduction

Realizing favorable absorption, distribution, metabolism, elimination and toxicity profiles is a necessity due to the high attrition rate of lead compounds in drug development today. The ability to accurately predict bioavailability can help save time and money during the screening and optimization processes. As several robust programs already exist for predicting log*P*, we have turned our attention to the fast and robust prediction of pK_a for small molecules. Using curated data from the Beilstein Database and Lange's Handbook of Chemistry, we have created a decision tree based on a novel set of SMARTS strings that can accurately predict the pK_a for monoprotic compounds with R^2 of 0.94 and root mean squared error of 0.68. Leave-some-out (10%) cross-validation achieved Q^2 of 0.91 and root mean squared error of 0.80.

Intense focus is being placed on the quick and accurate prediction of physicochemical properties, driven in particular by the pharmaceutical industry and the need to identify lead compounds with favorable absorption, distribution, metabolism, elimination and toxicity (ADMET). It is well known that pK_a, and in particular the ionic state of a molecule at physiological pH, affects pharmacokinetics and pharmacodynamics. Bioavailability measures, often characterized by the octanol/water partition coefficient log$P$ and Lipinski's rule-of-five,[1] now include pK_a so as to determine the pH-dependent distribution coefficient, log$D$.

Aqueous solubility, lipophilicity and amphiphilicity all contribute to intestinal absorption and membrane permeability, and are at least partially determined by pK_a.[2] In order to be absorbed, orally administered drugs must first dissolve in the gastro-intestinal (GI) fluids. The ionic state of a molecule can be affected during passage through the GI tract, due to environmental pH in the stomach (pH 1-3), duodenum (pH 5-7), and jejunum and ileum (pH ~8).[3] Furthermore, the majority of drugs administered orally are ionizable at physiological pH levels.[4,5] Significant dissolution enhancement has been observed when the buffer maintains the pH near or above the pK_a of the dissolving drug.[6] Therefore, adjusting the pK_a of a drug is of particular interest when dissolution has been found to be the rate-limiting step in the process of absorption, especially when dealing with drugs having poor water solublity.

Ionizable groups also affect the ability of a drug to interact with a target. It has been shown that pK_a influences the rate and site of metabolism of drugs by CYP1A2, a metabolic enzyme.[7] As strong electrostatic and hydrogen bonding interactions are key contributors to the overall free energies of binding,[8] pK_a can be critical for binding potency at the target. Moreover, based on a study of compounds targeting the human Ether-a-go-go Related Gene (hERG) potassium channel, selectivity can be influenced by controlling pK_a.[9]

Issues of toxicity are also related directly to a drug's pK_a at physiological pH, such as cardiovascular toxicity due to the lengthening of time between the start of the Q wave and the end of the T wave in the heart's electrical cycle or QT prolongation, resulting from the blockade of the hERG potassium channel.[10] Another common toxicity issue that can be affected by the ionic state of a drug molecule is the potential liability for phospholipidosis, an adverse drug reaction that occurs with many cationic amphiphilic drugs.[11] All in all, approximately 50% of all drug failures have been attributed to poor ADME, and in 2003 approximately $8 billion was lost due to the inability to predict the toxic nature of a substance during the early stages of drug development.[12,13,14]

The main advantage of *in silico* pK_a prediction is that physical samples are not needed. Even when one considers newer methods of high throughput pK_a screening, there are two limiting factors: the costs and time associated with obtaining or synthesizing the compounds of interest. Hence, there is a need for a quick, accurate and robust model for pK_a prediction for large as well as small compound libraries.

Much research has been done in the area of pK_a prediction. Seminal publications[15,16] on the prediction of pK_a base their model on linear free energy relationships (LFER), applying the Hammett equation. LFER models are still commonly used and are also implemented in popular commercially available software packages, such as SPARC.[17,18,19] One of the most common techniques used in pK_a prediction is quantitative structure activity/property relationships (QSAR/QSPR) deriving their fit equations from partial least squares (PLS) or multiple linear regression (MLR).[5,18,20,21,22,23,24,25] Other methods include neural networks,[26] quantum mechanical continuum solvation models,[27,28,29] and anti-connectivity models.[30,31] It has often been the case that a model was based on a relatively small set of experimental data for a specific ionizable group, such as carboxylic acids.[18,22,25,26,27,29,30,31] Others have tackled the problem of chemical diversity by devising and combining multiple models, each applied to a relatively small set of compounds when compared to the complete set of experimental data. [23,26] Here the overall combination of models is more robust at handling novel chemical structures, but the individual training sets may suffer from a lack of chemical diversity due to their small size. This may allow for a good fit on the training sets, but has the potential disadvantage of leaving little freedom for cross validation. Furthermore, there is a danger of cherry picking or manually selecting compounds that are well represented in the training set for the test set. In the methods section we describe how our single model is derived and applied to a large and diverse training set of monoprotic molecules. We also describe how the training and test sets were randomly selected from diverse clusters of compounds in order to ensure that no cherry picking occurred and that chemical space was fairly represented by all training and test sets for both the final model and in cross validation.

5.2 Methods

All calculations were performed using the Molecular Operating Environment (MOE)[32] on a Dell Precision 380 workstation utilizing Red Hat Linux Enterprise version 4.0.

5.2.1 Preliminary Studies. *1.1 Data Acquisition.* Data was obtained from Lange's Handbook of Chemistry 15th Edition[33] and the Beilstein (2007/04) database via the Molecular Design Limited (MDL) CrossFire Commander.[34] In all cases possible we applied the following filters in our curation of monoprotic compounds where the ionizable site was an oxygen, nitrogen or sulfur atom: titration performed in H_2O at 23–27° C, ionic strength (the molar concentration of all ions present in a solution) less than or equal to 0.1 M, and pK_a range from –1.74 (H_3O^+) to 15.7 (H_2O). We also accepted multiprotic compounds, such as *o*-nitrophenol, where only one ionizable site had pK_a within our accepted pK_a range.

1.1.1 The Lange dataset provided International Union of Pure and Applied Chemistry (IUPAC) nomenclature and pK_a values for 2162 compounds with up to six ionizable sites. We accepted 700 of the monoprotic compounds: 417 carboxylic acids and alcohols, 14 thioacids and thiols, and 269 compounds having ionizable nitrogens. IUPAC nomenclature for each of these compounds was translated into simplified molecular input line entry specification (SMILES) strings both manually and using ChemDraw Ultra 10.[35]

1.1.2 The Beilstein database provided us with 10334 unique substances. These substances were washed to remove any salts; thus only the major component with the largest number of bonded atoms was retained. After applying our filters, we accepted 1577 monoprotic compounds: 898 carboxylic acids and alcohols, 33 thioacids and thiols, and 642 molecules having ionizable nitrogens. The following relevant fields were downloaded in structure data format (SDF):[36] Beilstein registry number (BRN), molecular structure, dissociation exponent (DE) or pK_a, dissociation group (DE.GRP), dissociation temperature (DE.T), dissociation solvent (DE.S), dissociation method (DE.MET), dissociation type (DE.TYP) and dissociation comments (DE.COM).

1.1.3 Many published pK_a values for the same compound were often found to exist. This was partly due to the large overlap between Lange and Beilstein. Beilstein also contained multiple instances with literature references where more than one pK_a value was published for a single compound. The two datasets including their duplicate entries were merged into a single MOE database. In the cases where the pK_a values for a specific compound varied less than 0.5 units, the mean was accepted as the experimental pK_a. In all instances where the variation was greater than 0.5 units, the literature reference was checked to explain the discrepancy, which was often due to the use of a solvent other than H_2O. Only three compounds were accepted that had variations greater than 0.5. All had variation less than 0.7 units.

The resulting dataset contained 1881 unique monoprotic compounds, involving 1088, 33 and 760 ionizable oxygen, sulfur and nitrogen atoms, respectively. 2086 experimental pK_a values were found for the 558 compounds having multiple literature references. The RMSE for the duplicate experimental pK_a values from literature and our accepted experimental values after curation was 0.08. Figure 5.1 portrays the pK_a distributions of the ionizable atom types of interest.

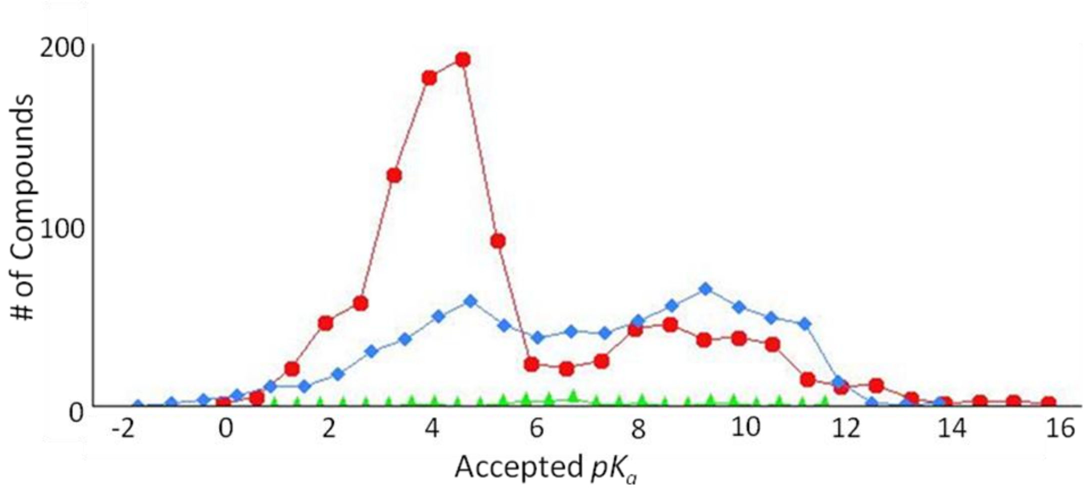


Figure 5.1. Each pK_a distribution is based on 25 bins for the monoprotic compounds having one of the three ionizable atom types considered here: oxygen (—●—), nitrogen (—◆—) and sulfur (—▲—).

1.2 Clustering molecules to obtain equally diverse training and test sets. For cross validation we separated our compounds into ten groups of equal size and having similar chemical diversity. To do this, we built a decision tree using the 166 Molecular ACCess System (MACCS) keys from MDL and requiring the leaf nodes to contain a minimum of ten compounds. The root of the tree consists of all the monoprotic compounds. Branching decisions are made by the MACCS key which most evenly (or ideally) divided the compounds at any given node and terminates when the node cannot be split into two children each containing at least ten compounds. See section 1.4 for a description of the ideal fingerprint.

Equations 1-4 describe how decisions are made at each node. $\mathbf{h}_\mathrm{k}$ and $\hat{\mathbf{h}}_\mathrm{k}$ are binary vectors with length equal to the number of compounds *m* at a node with $\mathbf{h}_\mathrm{k}$ representing the hit profile and $\hat{\mathbf{h}}_\mathrm{k}$ the inverse hit profile with respect to MACCS key *k*. $|\mathbf{h}_k|$ and $|\widehat{\mathbf{h}}_k|$ are the number of compounds hit and missed, respectively. See eqs 1-3.

$$|\mathbf{h}_k| = \mathbf{h}_k \bullet \mathbf{h}_k = \sum_{j=1}^{m} h_{kj} \tag{1}$$

$$|\widehat{\mathbf{h}}_k| = \widehat{\mathbf{h}}_k \bullet \widehat{\mathbf{h}}_k = \sum_{j=1}^{m} \widehat{h}_{kj} \tag{2}$$

$$m = |\mathbf{h}_k| + |\widehat{\mathbf{h}}_k| \tag{3}$$

Eq 4 defines the idealness score, I_k, for MACCS key *k* where $0.5 \leq I_k \leq 1.0$. A value of 0.5 indicates that the compounds at a particular node are evenly divided into children nodes, whereas a score of 1.0 indicates that one child node contains all of the compounds and the other contains none. The compounds at each leaf were randomly divided into ten bins, each consisting of approximately 10% of the monoprotic compounds. Nine of the bins were used for training purposes while the tenth was set aside as a test set for our final model.

$$I_k = \frac{\max[|\mathbf{h}_k|, |\widehat{\mathbf{h}}_k|]}{m}, \text{ for } k = 1,\ldots,166 \tag{4}$$

1.3 Analyzing the molecular diversity of the training set using principal components analysis (PCA) and the MACCS keys. PCA was applied to access the molecular diversity of the 1693 compounds. We formed a binary matrix consisting of 1693 rows (the training set compounds) and 166 columns (the MACCS keys) with every entry being either a '1' or a '0', corresponding to the respective presence or absence of each MACCS key. Of the 166 MACCS keys, 152 are found in our training set. PCA showed that 62 principal components accounted for 90% of the total system variance. Furthermore, 144 principal components were required to account for 100% rounded to the nearest thousandth of the total system variance. A non-trivial number of principal components are required to explain the vast majority of the system variance; hence, the training set has been shown to be widely diverse and applicable for the methodology implemented in developing our model.

1.4 Tailoring an ideal fingerprint with predictive power using SMARTS descriptors. 1.4.1 An ideal fingerprint is one in which the descriptors are mutually independent and each evenly divides a dataset of *n* compounds into those having vs. not having the descriptor. Here we use MOE SMiles ARbitrary Target Specification (SMARTS) strings as descriptors, as described in Table 5.1. Note, while MOE SMARTS strings are based on the concepts of Daylight SMARTS strings[37] and are for the most part the same, there are some differences. For example MOE uses the SMARTS string [i] to denote any π bonding atom, but this is not used in Daylight SMARTS.

Table 5.1. Example MOE SMARTS strings

SMARTS String	Definition
[OH]aaa[#X]	Non-carbon aliphatic atom, [#X], meta to a hydroxyl, [OH]
a[#G7]	Any aromatic atom with a Halogen substituent, [#G7]
a[C;!i]	Any aromatic atom connected to a non π-bonded carbon atom
[OH][A;r]=[A;r][A;!r]=O	Hydroxyl covalently bonded to a non-aromatic ring atom, which shares a double bond with an adjacent aliphatic ring atom that is single bonded to a non-ring atom sharing a double bond with oxygen.
[OH]A[Ov2]	Hydroxyl bonded to any aliphatic atom sharing a single bond to an oxygen that is explicitly bonded to two non-hydrogen atoms.
A[N+]=O	Aliphatic atom covalently bonded to the positively charged nitrogen of a nitroso group.

The strings described above were specifically selected to help readers unfamiliar with SMARTS notation better understand the SMARTS strings presented in Table 5.2.

The ideal fingerprint has length n_d, where n_d is the minimum number of descriptors to uniquely identify all n compounds in the training set.

$$n_d = \lceil \log_2 n \rceil \tag{5}$$

Such a fingerprint uses the minimum number of descriptors possible to uniquely identify each compound in a specific dataset, that is, the fingerprint tends to be ideal only for the dataset on which it was based.

1.4.2 Each n_d-digit binary number representing the occurrence/absence of descriptors can not only be used to identify the respective unique compound, but also may correlate with physicochemical properties.

Similarity measures, such as the Tanimoto score described in eq 6,[38] can then be applied to associate such physicochemical properties with structurally similar molecules.

Let $\mathbf{f}_i$ and $\mathbf{f}_j$ be binary vectors of equal length where each vector component represents the presence or absence of some qualitative descriptor as 1 or 0, respectively. Then the Tanimoto similarity score $s_{i,j}$ is the ratio of intersection and the union of the two sets of qualitative descriptors. If no descriptors from vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ match, then $s_{i,j} = 0$. If exactly the same descriptors are present in $\mathbf{f}_i$ and $\mathbf{f}_j$, then $s_{i,j} = 1$. Finally, if some, but not all, of the descriptors in $\mathbf{f}_i$ match those in $\mathbf{f}_j$, then $s_{i,j}$ will be some rational value between 0 and 1. The closer $s_{i,j}$ is to 1, the more similar are the vectors $\mathbf{f_i}$ and $\mathbf{f_j}$. In our case, pairs of molecules with sufficiently high $s_{i,j}$, based on an ideal or close to ideal fingerprint, are expected to have similar pK_a values.

$$s_{i,j} = \frac{\mathbf{f_i} \bullet \mathbf{f_j}}{|\mathbf{f_i}| + |\mathbf{f_j}| - \mathbf{f_i} \bullet \mathbf{f_j}} \tag{6}$$

1.5 Concept validation using molecular similarity measures. In order to validate our concept that pK_a can be predicted based on the structural similarity between two molecules using SMARTS strings as descriptors, we applied it to our data set of 403 monoprotic alcohols.

1.5.1 202 monoprotic alcohols from five of the ten bins were selected as the training set, leaving the remaining 201 alcohols as the test set.

1.5.2 The 50 SMARTS string descriptors described by Table 5.2 were manually created and uniquely identified the molecules in the alcohol training set. An ideal fingerprint with 50 descriptors should be able to uniquely identify 2^{50} molecules. This is gross overfitting in our case, as $2^{50} - 202 = {\sim}10^{15}$ binary profiles correspond to no molecule. In our case, the ideal set of descriptors would contain only 8 SMARTS strings, the minimum number of descriptors to uniquely identify all 202 alcohols.

Table 5.2. Hand selected SMARTS string fingerprint that uniquely identifies the alcohol training set

	SMARTS string subsets		
Type	A	B	C
Aromatic Centers	[OH]a, [OH]aa[#X], [OH]aaa[#X], [OH]aaaa[#X]	[OH]a(a[#X])a[#X], [OH]a(a[#X])aa[#X]	[OH]a(aa[#X])aa[#X], [OH]a[#7+]
Aromatic Modifiers	a[#G7], a[Ov2], aC=C	aC=O, a[Sv2] , a[#8+]	a[N+]=O, a[#7+](C)C, a[S+], a[S+2], aC=Cc[n+], aC=C[N+](=O), aC(C)(C)C, a[C;!i], a[P+]
Aliphatic Centers	[OH]A=A, [OH]A[O-], [OH]AAA=O, [OH][A;r](=[A;r])[A;r]=O, [OH][A;r]=[A;r][A]=O	[OH]C, [OH]A=S, [OH][A;r]=[A;r][A;!r]=O	[OH]A[#X], [OH]AA[#X], [OH]AA([#X])[#X], [OH]AAA[#X], [OH]AAAA[#X], [OH]A(A[#X])A[#X], [OH]A[Ov2], [OH]A[Sv2]
Aliphatic Modifiers	A[#G7], A[N+]=O, [#7+](~*)~*	[Sv2] , [n+]	[S+], [S+2], C=C, C=Cc[n+], C=C[N+]=O, [P+]

The SMARTS string subsets A+B+C make up a fingerprint that uniquely identifies the 202 monoprotic alcohols in the training set, $R^2 = 1.0$. Subsets A and A+B make predictive fingerprints which respectively yielded $R^2 = 0.75$ and 0.86 on the test set of monoprotic alcohols.

1.5.3 Apply reverse stepwise regression. Using the 50 SMARTS string descriptors (fingerprint components) as indicator variables, reverse stepwise regression on pK_a was applied in an attempt to minimize the number of SMARTS strings, as it was extremely unlikely that a set of ideal SMARTS string descriptors existed. First the fingerprint of all molecules in the training set was taken.

1.5.3.1 Drop the descriptor that least affects the fit R^2 of the training set. The predicted pK_a for each compound is assigned based on the average pK_a of the compounds sharing the highest $s_{i,j}$. Then the Pearson correlation coefficient (R^2) is calculated using the experimental vs. predicted pK_a values. Deleting any one of the SMARTS strings can decrease the R^2, so we drop the string that causes the smallest decrease. In the case of a tie, drop the least ideal SMARTS string. Retain the dropped SMARTS string in a pool for later re-evaluation steps.

1.5.3.2 Check pool containing dropped SMARTS strings for possible improvements to R^2 for the training set. If the discard pool contains more than one SMARTS string descriptor, then check whether re-adding any one descriptor from the discard pool except for the last dropped SMARTS string can increase the R^2 of the current model. If any such descriptors exist and their inclusion yields an R^2 greater than or equal to 0.90, re-add the SMARTS string that increases the R^2 of the model the most.

1.5.3.3 Check threshold values. Repeat the stepwise regression as described in steps 1.5.3.1 – 1.5.3.2 until the R^2 falls below 0.90.

1.5.3.4 Re-add the most recently dropped SMARTS string. In this manner, we obtained a best estimate for the most ideal predictive fingerprint that yields a correlation greater than or equal to 0.90 based on the training set.

1.5.4 Validation. Step 1.5.3 was repeated twice using threshold R^2 = 0.95 and 0.90 for the alcohols training set only. For all other investigations the threshold R^2 = 0.90 was used. The descriptors of the two predictive fingerprints that were identified are shown in Table 5.2. The first, having the lowest R^2 greater than 0.95, consisted of the 25 SMARTS string descriptors in subsets A and B. The second, having lowest R^2 greater than 0.90, used only the 15 SMARTS string descriptors in subset A. Applying the predictive fingerprints described by the combination of subsets A and B and of subset A alone to our test set provided concept validation and yielded R^2 = 0.86 and 0.75, respectively.

1.6 Expanding and refining predictive SMARTS descriptors. Our goal is to identify very general SMARTS string descriptors, which are both as ideal as possible and differentiate groups of compounds by pK_a. Selecting a discriminating SMARTS string can often be a non-trivial task, so as not to make it too specific.

1.6.1 After applying the fingerprint consisting of only 15 SMARTS strings described in Table 5.2 (subset **A**), we examined the set of compounds where all members had the same fingerprint value but large experimental pK_a variance. In this case, the largest experimental pK_a variance is at the predicted pK_a of 6.48. Figure 5.2 demonstrates how outlier molecules were identified and the predictive fingerprint was modified.

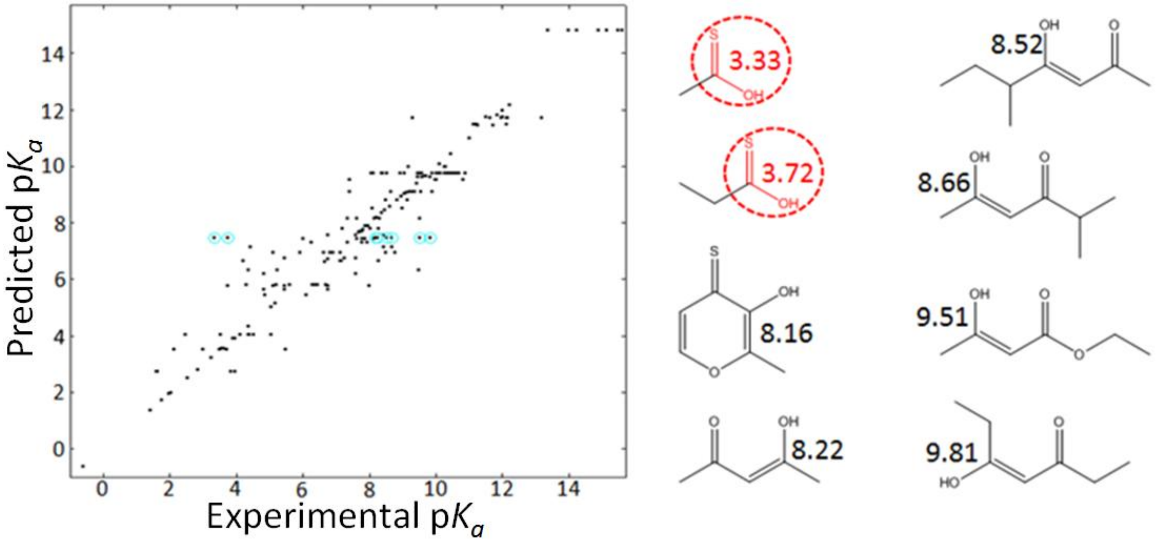


Figure 5.2. The scatter plot on the left depicts experimental vs. predicted pK_a values for 202 monoprotic alcohols from the training set using the fingerprint defined by SMARTS string subset A, having an $R^2 = 0.90$. The molecules sharing the same fingerprint profile with largest experimental pK_a variance are identified by the points circled in blue (⊙). The molecular structure and pK_a of the highlighted points on the scatter plot are shown on the right. Highlighted and encircled in red, the fragment represented by the SMARTS string, [#8][i]=[#16], differentiates the molecules with low pK_a from those with high pK_a.

1.6.2 We then generalized existing SMARTS descriptors by substituting more general atom and bond types in our existing descriptors. If no such modification could be identified, we added new descriptors until the R^2 became greater than 0.96 and allowed no predicted value to deviate by more than 1.0 pK_a units from the experimental value. Figure 5.2 depicts the thioacid moiety as the clear differentiating factor, separating the molecules into two smaller sets while minimizing the pK_a variance. When added to the fingerprint, the SMARTS string [#8][i]~[#16], representing an oxygen atom single bonded to a π bonding atom that shares some bond with a sulfur atom (inclusive of the thioacid), the R^2 rose from 0.90 to 0.92.

1.6.3 After the R^2 was improved to greater than 0.96 and deviation for all predicted values was less than 1.0, stepwise regression was performed again to identify the descriptors with least influence on the training set. These descriptors tend to be the most specific and are present in only a few molecules or almost all the molecules in the training set. Each set of molecules identified by such a descriptor was then analyzed with the intent of replacing the overly general or overly specific SMARTS string with a more ideal SMARTS descriptor that split the dataset more evenly as well as successfully differentiated the set of molecules.

1.6.4 At this point the training set was broadened with 346 (~50%) of the monoprotic carboxylic acids, which were taken from the same bins as the monoprotic alcohols. The fingerprint was then modified to include the general structural differences between the alcohols and the carboxylic acids, namely the carbonyl group between the ionizable oxygen and the remainder of the molecules. This was done by adding SMARTS descriptors similar to those used in the fingerprint for the alcohol training set, but inserting a carbonyl moiety in parentheses next to the ionizable oxygen, represented by [OH]. For example, the motif representing a non-carbon meta substitution of a phenol-like substance, [OH]aaa[#X], becomes [OH]C(=O)aaa[#X]. For our convenience, each SMARTS descriptor containing an ionizable atom places the ionizable moiety first in the string. Our process has led us to pay less attention to the general aromatic SMARTS ‘a’, representing any aromatic atom, and incorporate ‘[i]’, representing a π bonded atom which matches both aromatic atoms as well as those with conjugated bonds. This generalized SMARTS string, [OH][i](=O)~[i]~[i]~[i][#X], will identify both meta substituted phenyl rings and some conjugated carboxylic acids. Following the logic from steps 1.4 – 1.6.4 the set of SMARTS strings was modified and refined. Next, 16 monoprotic molecules containing ionizable sulfur atoms were added to the ionizable oxygen training set and the SMARTS fingerprint was refined accordingly. The ionizable site of this entire class of monoprotic molecules can be described by the SMARTS string [#G6!H0], which identifies a group 6 atom from the periodic table bonded to at least one hydrogen.

1.6.5 Far more diversity is evidenced among the monoprotic molecules containing ionizable nitrogens which are identified by the SMARTS string [#7!H0]. To simplify matters, this set of compounds was divided into three groups: primary, secondary and tertiary amines, and treated in the same manner as the molecules identified by [#G6!H0].

1.6.5.1 Starting from scratch using molecules with ionizable nitrogens drawn from the same bins from which the alcohols, carboxylic acids, thiols and thioacids were obtained, a new SMARTS string fingerprint was trained on a random selection from the primary ionizable nitrogen dataset, containing both amines and amides. The secondary and tertiary ionizable nitrogen datasets were sequentially added and trained. It is of note that we were unable to derive sufficiently good predictive fingerprints for these groups of compounds using only 50% of the bins.

1.6.5.2 In order to achieve predictive fingerprints that surpassed $R^2 = 0.75$ for the test set of all ionizable nitrogens from the unused bins, we needed to expand our training set to include 80% of the ionizable nitrogens. Here the training set was expanded by including the ionizable nitrogen containing compounds from three of the unused bins of equally diverse compounds based on the MACCS keys.

1.7 Combining fingerprints trained for [#G6!H0] and [#7!H0] monoprotic molecules. It was obvious that there was some overlap between the descriptors (aliphatic and aromatic modifiers) from our predictive fingerprint trained on the compounds containing ionizable oxygen and sulfur atoms and the predictive fingerprint trained on the compounds containing ionizable nitrogen atoms. Based on the size of the combined fingerprints, it was likely that we were overfitting the data. However, this was expected as we were unable to identify any descriptor that was truly ideal.

1.8 Obtain basis SMARTS string fingerprint. Our final training set was formed using 90% (9 bins) of our monoprotic data, including the compounds from the aforementioned 8 bins. None of the curated SMARTS strings were derived from compounds in the tenth bin, our test set. We once again performed stepwise regression, identifying the most ideal basis set of descriptors, thus allowing for the manual discovery of other novel and very general descriptors based on the outliers. As before, the fingerprint was refined until R^2 exceeded 0.96 with all calculated values deviating less than 1.0 from the experimental values. Details of the analysis and refinement of the SMARTS strings are shown in Table 5.3.

Table 5.3. SMARTS string refinement process

Group	Training set size	# bins used	# of SMARTS pre-regression	# of SMARTS post-regression[a]	# of refined SMARTS[b]	Fit R^2
[OH] and !C(=O)[OH]	202	5	50	15	21	0.96
[OH]	548	5	132	32	122	0.99
[#G6!H0]	564	5	139	35	123	0.99
[#7!H0]	606	8	145	59	130	0.96
All	1693	9	284	140	256	0.96
All (final)	1693	9			262 [c]	0.95[c]

[a] Stepwise regression, reducing the number of SMARTS strings was performed until the fit R^2 was reduced to 0.90. [b] Refinement, as discussed in section 1.6, was performed to improve the fit R^2 to exceed 0.96. [c] The final refinements to the pool of SMARTS strings were made using the decision tree as discussed in section 2.2.

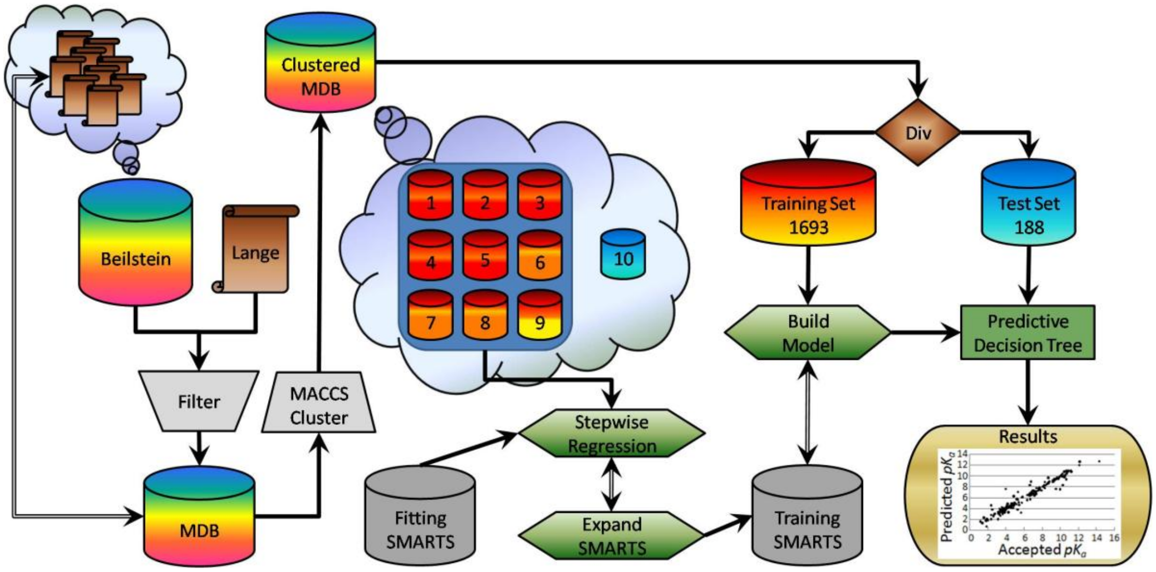


Figure 5.3. Workflow used to achieve the SMARTS pK_a decision tree. Clouds represent database contents and the brown scrolls are literature references.

5.2.2 Training and Validating the Final Model. Figure 5.3 illustrates the generalized workflow used to develop the final model as described in steps 1.1-2.2.

2.1 Growing the predictive decision tree. Our predictive decision tree is based on two factors: diversity and accuracy. Diversity is represented by the ideal nature of our SMARTS fingerprint. Using eq 4, we calculated a new diversity score, this time allowing a minimum of two molecules at each leaf node in the decision tree using our ideal SMARTS strings instead of the MACCS keys used when selecting our training and test sets. An element of accuracy is also included in making each branching decision. Eq 7 describes the accuracy score, A_k, which is intended to help minimize the pK_a range at each node in the tree. For both eqs 4 and 7, the index k now refers to the descriptors in our SMARTS fingerprint. Binary hit vectors $\mathbf{h_k}$ and $\hat{\mathbf{h}}_\mathbf{k}$, are now based on the SMARTS fingerprint rather than the MACCS keys. m refers to the number of compounds at the parent node. r_k and $\hat{r}_k$ are the pK_a ranges at the children nodes for SMARTS string k. The values of the resulting score range from (0,1.0]. In this case, values less than 1.0 represent child nodes with smaller pK_a ranges than the parent node. Values closer to zero reflect child nodes with narrower pK_a ranges.

$$A_k = \frac{r_k \left|\mathbf{h}_k\right| + \hat{r}_k \left|\widehat{\mathbf{h}}_k\right|}{\max_k [r_k, \hat{r}_k] \times m} \tag{7}$$

Eqs 4 and 7 have been normalized such that the lower values of the respective scores are desirable. Eq 8 describes how the final decision is made at each node and considers the ability of a particular SMARTS string to evenly split the substances and at the same time minimize the pK_a range of the children nodes. Weighting factors for I_k and A_k were systematically determined by spawning several decision trees and varying the weights. The final weighting scheme was the one resulting in the highest fit R^2 based on correlating the mean of all pK_a values of the leaf and the experimental pK_a values from our training set.

$$S = \min_k [I_k + 2.5 \times A_k] \tag{8}$$

2.2 Refining the predictive decision tree to establish the final model. New SMARTS strings were added based on outliers found when fitting the training set. The final SMARTS string fingerprint based on our training set consisted of 262 SMARTS strings, 139 of which were used in creating the decision tree, which gave calculated pK_as for the training set that fit the experimental values with R^2 greater than 0.95 and root mean squared error (RMSE) less than 0.65.

This method is a refinement of the one used in our previous prediction of generalized cytotoxicity where the MACCS keys were used to make decisions on a randomly partitioned training and test set.[39] Here manually derived SMARTS strings are used to make decisions and the MACCS keys were used to uniformly partition the data into the training and test sets for cross validation purposes.

5.2.3 Cross Validation and pK_a Shuffling Studies. In order to perform proper cross validation and shuffling studies, one should redo all steps in the method. Unfortunately, the method for obtaining the pool of SMARTS strings used in the decision tree is complex and immensely time consuming. We have chosen to perform leave-some-out (10%) cross validation and pK_a shuffling studies using the final pool of SMARTS strings. In the cross validation procedure, the entire dataset of 1881 molecules was used to form all possible combinations of nine of the diverse bins as training sets and the remaining bin in each situation as the test set.

pK_a shuffling studies were performed to evaluate the likelihood of overfitting the model. An attempt was made to repeat the entire procedure using only the alcohols. After shuffling the pK_a of training set compounds, reverse stepwise regression was performed as in step 1.5. Using the same 50 SMARTS strings, the R^2 for the training set was less than 0.46, so we were unable to continue the SMARTS string reduction procedure. This alone demonstrates that the 50 SMARTS strings being used could recognize bogus data, rather than fit it. Consequently, the shuffling test was only applied to the final step of model development, leaving the laboriously selected SMARTS strings fixed.

5.3 Results and Discussion

The decision tree is 13 levels deep using 139 SMARTS strings with 1527 nodes including 741 terminal nodes containing two or more of the 1693 molecule training set. The decision pathway depends on the presence or absence of the SMARTS string descriptors from the pool and the equations provided in the methods section. It is clear that any SMARTS string can be used at any level in the decision tree and is often the case with general motifs, such as long carbon chains or linear strings of atoms. Examining the first five levels of the decision tree, it is readily evident at least to some degree that the main ionizable sites are separated into groups first. Starting at the root (node 1) the first decision is always whether the molecule contains a carboxylic or thioic acid (node 2), or not (node 3). This is immediately followed at node 2 by differentiating carboxylic (node 4) from thioic acids (node 5). At node 3, aliphatic charged conjugate acids (node 6) are differentiated from other amines and alcohols (node 7). Ortho, meta, and para substitutions and any combination thereof are decided in the middle levels of the tree along with other motifs relating the positions of substituents relative to the ionizable site in aliphatic molecules. The decision tree also tends to identify the more specific substituents closer to or at leaf nodes, such as the halogens, the methyl group, as well as specialized branches and bonding configurations. As 1527 nodes exist in the decision tree and only 139 SMARTS strings are used, it is clear that the same SMARTS strings are being reused along different decision pathways. Therefore, it is the case that SMARTS strings determining the leaf nodes also occur in the middle of the decision tree, but never in the same decision pathway.

Originally the test set contained 188 molecules, but two were dropped as their tautomeric forms lead to loss of the proton from a carbon atom, rather than the expected ionizable site. It is worth mentioning that the SMARTS pK_a prediction for these molecules had errors of only 0.2 and 0.91. A third molecule was dropped because of ambiguity in converting its name to a structure. The decision tree, including predictions and ranges for each node, is provided in the supplementary materials.

The decision tree, utilizing 139 SMARTS strings, was created from a pool of 262 SMARTS string descriptors. In order to deal with the primary concern of overfitting the model, we retrained the decision tree 100 times allowing the SMARTS descriptors to be selected from the original pool while randomly shuffling the pK_a values within the training set. The R^2 of the original predictive model was shown to lie over 28 standard deviations above the mean R^2 of the 100 models with randomized data, while the RMSE of the original model was over 422 standard deviations below the mean of the respective RMSEs obtained from the randomized data models. The statistics comparing the R^2 and RMSE of the final predictive model and the models with randomized pK_a data is shown in Table 5.4. Based on these results, it is clear that we have not overfit the model.

Table 5.4. Overfitting test: accepted model vs. 100 models with randomized pK_a data

Model	Statistic	R^2	RMSE
100 Randomized	Mean	0.4830	2.1382
	Std. dev.	0.0162	0.0334
	Mode	0.4909	2.1220
	Min	0.4388	2.0721
	Max	0.5145	2.2382
Accepted		0.9548	0.6512

In order to validate that the decision tree model (SMARTS pK_a) was producing satisfactory predictions, a comprehensive literature survey was made. Furthermore, benchmarking on our test set was performed with SPARC, MARVIN, ACD/I-Labs v

8.03 and ADME Boxes. Details of the survey and benchmark are provided in Table 5.5, which includes descriptions and statistics for training sets as well as cross validation and or external test sets when provided. In summary for the various methods and the respective training sets, the R^2 ranged from 0.81 to 0.99 and the RMSE was less than 1.0. R^2 for leave-one-out or leave-some-out (10% or 20%) cross validation studies ranged from 0.78 to 0.92 with RMSE typically less than 1.0. Finally, R^2 for the external datasets ranged from 0.69 to 0.99. High values for the Pearson correlation coefficient and the lowest RMSE are expected on all the training set predictions (fits) with some falloff on predictions for cross validation and external test sets.

While SMARTS pK_a does not produce the highest fit scores, R^2 of 0.95 is very respectable. Two of the reasons our method does not outperform other methods are (a) the limited size of our complete dataset and (b) the fact that the SMARTS strings were not inclusive of any increased molecular diversity found in the 185 compound test set. Furthermore, this method is not designed to produce a fit having $R^2 = 0.99$ due to the pK_a average taken at each terminal node. By decreasing the node size from two molecules to one, both the SMARTS fingerprint and the MACCS keys were able to produce a fit with $R^2 = 0.99$. Finally, the only reason the MACCS keys were unable to produce a fit of 1.0 with the node size reduced to one was because we did not eliminate stereochemistry or E/Z conformers from the dataset. Neither our SMARTS strings nor the MACCS keys were able to distinguish between molecules having these conformational differences. Altogether 152 of the 166 MACCS keys and 156 of the 262 SMARTS strings were used in this exercise. This is valuable information regarding the chemical diversity of the training set and the potential of our SMARTS strings to divide the compounds. In an ideal situation, only 11 MACCS keys or SMARTS descriptors would be required to build a tree having 2^{11} nodes which would uniquely identify each of the 1693 training compounds.

SMARTS pK_a also performed well in cross validation with $Q^2 = 0.91$ and RMSE = 0.80. As the SMARTS strings were manually created (labor intensive) specific to the training set, we consistently retrained the cross validation models with the same pool of SMARTS from step 2.2.

Table 5.5. Literature survey of pK_a calculators with benchmarking using SPARC, MARVIN, ACD/I-Labs v8.03 and ADME Boxes

			Training Set			**Test Set**			**External Test Set**		
Method	**Ref.**	**Class**	***n***	***R^2***	***RMSE***	***n***	***Q^2***	***RMSE***	***n***	***R^2***	***RMSE***
QSPR/PLS	20	all subclasses							25	0.95[a]	0.78[a]
		acids	625	0.98	0.405	10%	0.86	1.04			
		bases	412	0.99	0.298	10%	0.87	1.12			
QSPR/PLS	23	33 subclasses	24617						39	0.80	0.90
		acidic nitrogen	421	0.97	0.41	20%	0.87	0.41			
		6 member N-heterocyclic bases	947	0.93	0.60	20%	0.85	0.86			
QSPR/PLS	24		49			49	0.86		23	0.77	
QSPR/MLR	21		15	0.97	0.12				3	0.99[a]	0.10[a]
QSPR/MLR	22	all subclasses									
		carboxylic acids	1122	0.81	0.42[b]	20%	0.81	0.43[b]			
		alcohols	288	0.82	0.76[b]	20%	0.81	0.78[b]			
QSPR/MLR	25	aromatic acids	74						33	0.99	0.27
QSPR/LFER	17	monoprotic oxy acids	135	0.99	0.455				14		0.471
Continuum Solvation	27	carboxylic acids							16	0.69	0.72
Anti-Connectivity	31		31			31	0.87	0.463			

			Training Set			Test Set			External Test Set		
Method	**Ref.**	**Class**	***n***	***R^2***	***RMSE***	***n***	***Q^2***	***RMSE***	***n***	***R^2***	***RMSE***
Neural Network (ChemSilico)	26	12 classes	>16000						665	0.83	
		primary amine	1100	0.95		20%	0.92				
		tertiary amines	870	0.92		811	0.80				
		monoprotic acids	1640	0.95		1640	0.88				
		aromatic nitrogen	1480	0.92		1367	0.80				
		alcohols	1302	0.88		1302	0.85				
Semiempirical/PLS (Novartis In-House)	5	all			0.48				350		0.81
		alcohols	202	0.87	0.58		0.80				
		amines	1403	0.89	0.49		0.84				
		anilines	311	0.90	0.49		0.78				
		carboxylic acids	681	0.90	0.34		0.86				
		imines	84	0.98	0.55		0.88				
		pyridines	397	0.95	0.58		0.86				
		pyrimidines	91	0.95	0.43		0.87				
Statistical Thermo. Quantum Solvation (COSMO-RS)	28	Bases	43	0.98	0.56[b]				58		0.66
	29	Acids	64	0.98	0.49[b]						
QSAR, LEFR (SPARC)	18		2500	0.99	0.36[b]				4338	0.99	0.37[b]
	19	Pfizer dataset[c]							123	0.92	0.78[b]
	40	Pfizer internal dataset[d]							537	0.80	1.05[b]
									185[c]	0.84	1.15
MARVIN	41,42					208[c]	0.98	0.38[b]	185[c]	0.88	1.03
ACD/I-Lab v8.03	43		>31000						185[c]	0.90	0.93
ADME Boxes	44								185[c]	0.93	0.69
SMARTS pK_a			1693	0.95	0.65	10%	0.91	0.80	185	0.94	0.68
	45,46								112	0.77	1.59

			Training Set			Test Set			External Test Set		
Method	**Ref.**	**Class**	***n***	***R^2***	***RMSE***	***n***	***Q^2***	***RMSE***	***n***	***R^2***	***RMSE***
									107[e]	0.89[e]	1.04[e]
Consensus[f]									185	0.96	0.60

In all training sets *n* refers to the number of pK_a measurements; in the test sets *n* refers to the number of pK_a measurements or percentage of the training set. [a] External set statistics were calculated from data presented in the referenced material. [b] Standard deviation. [c] It is unknown whether these molecules were used in the training set. [d] These molecules were unlikely to be found in the SPARC training set. [e] Results recorded after removing significant outliers from the referenced materials. [f] The consensus model used predictions from SPARC, MARVIN, ACD/I-Lab 8.03 and SMARTS pK_a.

Finally, SMARTS pK_a performed exceptionally well on the external test set, which was randomly selected from the uniformly diverse MACCS descriptor space described in the Methods section. The predictive R^2 was 0.94 and RMSE = 0.68. In fact, we outperformed some respected online pK_a calculators based on three standards: robustness, overall accuracy and fewest outliers.

Both SMARTS pK_a and ChemAxon's MARVIN pK_a calculator plug-in were able to provide predictions for all compounds, while the other methods missed some. Note that SPARC was not designed to handle molecules containing selenium or silicon, which resulted in the missing prediction values. If one were to substitute oxygen for selenium and carbon for silicon, SPARC predicted the pK_a with error less than 1.0 in most of the instances. SMILES for all compounds and their predicted pK_a values from SMARTS pK_a, SPARC, MARVIN, Advanced Chemistry Development (ACD)/I-Labs Web service ACD/ pK_a 8.03 and ADME Boxes are available in the supplementary materials. A consensus model, having R^2 = 0.96 and RMSE = 0.60, was derived using the mean of the three predictions with the smallest pK_a discrepancies from each of the aforementioned five methods. Statistics for all methods are provided in Table 5.5 and depicted in Figure 5.4. Table 5.6 summarizes the overall accuracy of the prediction methods as well as the consensus based on the number of compounds predicted within five ranges of error and missed predictions.

Table 5.6. Evaluation of test set prediction errors: SMARTS pK_a vs. on-line calculators

pK_a error	***SPARC***	***MARVIN***	***ACD***	***ADME Boxes***	***SMARTS pK_a***	***Consensus***
[0,1]	145	154	165	156	168	174
(1,2]	26	23	10	14	12	7
(2,3]	7	4	5	5	4	3
(3,4]	1	1	1	1	1	1
(4,∞)	1	3	3	0	0	0
Miss	5	0	1	9	0	0

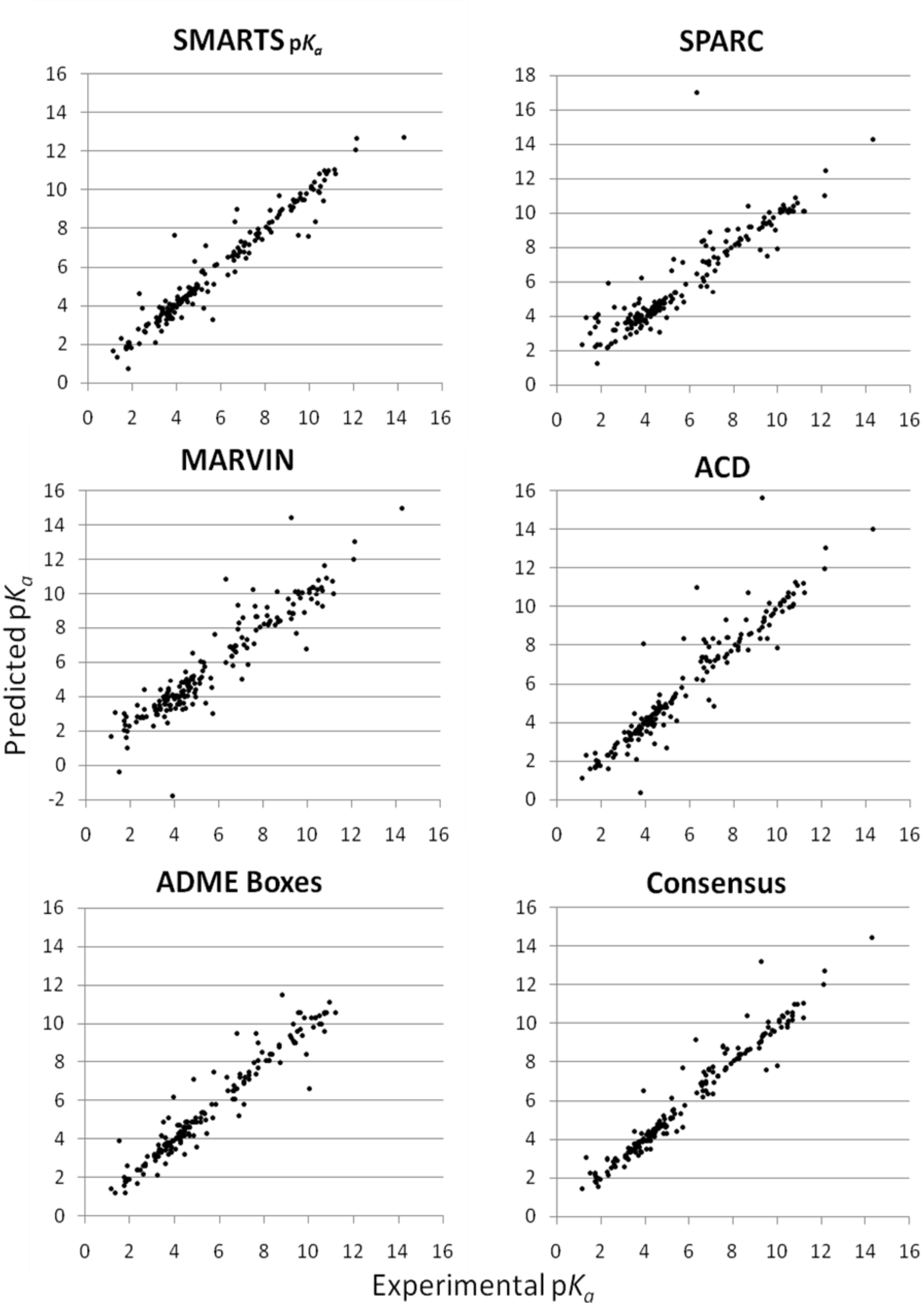


Figure 5.4. Scatterplots of the experimental vs. predicted pK_a values for the 185 compounds test set with SMARTS pK_a, SPARC, MARVIN and the ACD/I-Labs Web service ACD/ pK_a 8.03 and ADME Boxes.

A second external test set consisting of 112 compounds satisfying our filters from more recent literature was identified.[45,46] Note, the primary pK_a for the majority of these molecules has been previously calculated in a comparison study which included all of our benchmarking toolkits.[45] Applying SMARTS pK_a to this dataset resulted in an R^2 of 0.77 with RMSE of 1.59. The results are shown in figure 5.5. Removing the largest five outliers improves the statistics to 0.89 and 1.04, respectively. It follows that the chemical space occupied by the outlier molecules was poorly represented in the training set.

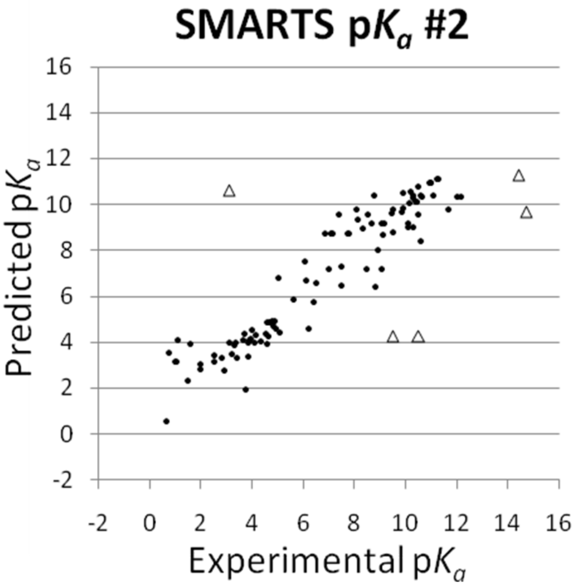


Figure 5.5. Scatterplot of the experimental vs. predicted pK_a values for 112 compounds outside the training and first test sets. Outliers with predicted values differing more than 3 pK_a units are depicted as open triangles. Three of the outliers are secondary amines covalently bonded to either carbonyl or sulfonyl moieties. The others include triethenylamine and an aliphatic halogenated alcohol.

The following is an example of a prediction made using our model. First, we selected a compound from the external test set, 4-(benzyloxy)benzoic acid. At the root of the decision tree the range of pK_a values of all compounds in the training set is 17.32. The predicted value is the mean pK_a of these compounds, 5.91 rounded to the nearest hundredth. The SMARTS string from our pool of SMARTS strings which optimally splits these compounds into two sets according to the scoring function defined in eq 8 is [#G6H]C(=O), identifying a Group 6 atom attached to both a lone hydrogen and a carbonyl. The SMARTS string positively identifies 4-(benzyloxy)benzoic acid, and another SMARTS string is selected from the pool which best splits this subset of compounds. The decisions leading to a final pK_a prediction for 4-(benzyloxy)benzoic acid are described in Table 5.7. Notice how the pK_a range at each child node remains the same or decreases, providing an overall estimate of accuracy based on the compounds sharing the terminal node.

Table 5.7. pK_a prediction for 4-(benzyloxy)benzoic acid using the final model

Node #	SMARTS String	*is* Identified[a]	*has* Child[b]	pK_a of Node	pK_a Range
1	[#G6H]C(=O)	Yes	Yes	5.91	17.32
2	[OH][i](=O)*(~*)~*	Yes	Yes	3.19	5.96
4	a[#X]	Yes	Yes	3.53	5.88
8	*~*~*~*~*~*~*~*~*~*	Yes	Yes	3.22	3.99
16	[i][#G6v2]	Yes	Yes	3.17	3.99
32	[O][i]~[i]~[i]~[i]~[i]~[i]~[i]~A	No	Yes	3.64	2.92
65	[OH][i]~[i](~*)~*	Yes	Yes	4.06	1.51
130	[OH][i](=O)[i]~[i]~[i]~[i]~[i]-A	No	Yes	4.34	0.69
261	[CH3]	No	Yes	4.51	0.26
507			No	4.47	0.08

[a] Yes/No refers to the presence/absence of the SMARTS string in 4-(benzyloxy)benzoic acid. [b] Yes refers to non-terminal nodes having children, while no refers to the terminal or leaf node.

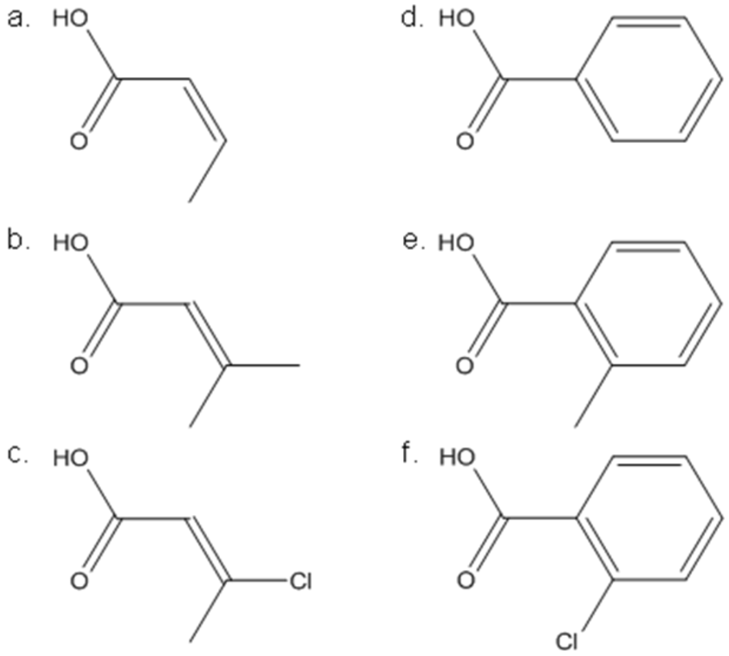


Figure 5.6. Close aromatic and aliphatic analogs with pK_a values (experimental:predicted) using the SMARTS pK_a method. a. but-2-enoic acid (4.68:4.56), b. 3-methylbut-2-enoic acid (5.12:5.14), c. 3-chlorobut-2-enoic acid (4.02:3.93), d. benzoic acid (4.21:3.83), e. 2-methylbenzoic acid (3.91:3.74), f. 2-chlorobenzoic acid (2.94:2.84)

SMARTS pK_a is capable of addressing issues of resolution for both aromatic and aliphatic analogs, as well as being able to discriminate between small differences in chemical structure. See examples shown in figure 6. With over 700 terminal nodes in the decision tree, there is sufficient differentiating power for fine grain predictions over the pK_a range of interest. Most terminal nodes contain only two compounds, but in the cases where a decision pathway terminates at a node with more than two compounds, there appears to be lessened sensitivity to small structural differences. This must be taken with the proverbial "grain of salt," as the pK_a range of all terminal nodes is minimized by the method, such that structural modifications with less influence on the pK_a tend to group together. SMARTS pK_a is capable of high resolution caused my minor structural changes that can have a significant influence on the pK_a and tends to ignore structural changes that have minimal impact on the experimental pK_a.

The SMARTS pK_a method has two other advantages: it is fast and it provides a pK_a range for each prediction based on the maximum and minimum pK_a of the molecules at each terminal node. Speed benchmarks show that it would take less than one hour to predict the pK_a for one million compounds. Also, when providing predictions for the external test set, 81% of all experimental values fell within the intervals of prediction (IoPs) determined by the terminal nodes. The width of an IoP is the maximum minus the minimum pK_a of the molecules at a terminal node, while the predicted value is the mean of the pK_as and falls somewhere between the maximum and minimum values. By extending the maximum and minimum values of each terminal node by only 0.3 pK_a units (i.e. increasing the pK_a range by 0.6), this confidence interval for an accurate prediction is increased from 81% to 100%. See Table 5.8 for a description of the test set compounds that fell inside the IoP widths.

Table 5.8. Analysis of molecules falling within IoPs

IoP width[a]	Test Set Compounds	Experimental values within IoP
(0,1]	130	99
(1,2]	34	39
(2,3]	14	14
(3,4]	3	3
(4,9)	4	4

[a] Interval of prediction width is in pK_a units. Extending the IoPs by 0.6 increases confidence from 81% to 100%.

5.4 Conclusions

As a measure of strength of acidity or basicity, pK_a is a major factor in chemical reactions and biological interaction of all compounds. It is relevant to physicochemical properties, such as aqueous solubility and log*D*, as well as ADME.[47] These facts, along with the immense and ever growing number of known and theorized chemical entities, make pK_a a major focal point in the drug discovery pipeline. Hence we need to continue creating new predictive models and improving the efficiency and accuracy of existing prediction methods.

Here we have presented a new predictive model for the pK_a of monoprotic compounds. Having obtained 1881 unique monoprotic compounds with their pK_a from Lange's 15th Handbook of Chemistry and the Beilstein Database, we used a novel set of SMARTS strings derived from a training set of 1693 monoprotic compounds to create a decision tree where the leaf nodes provide pK_a predictions and IoP, based on two or more training compounds identified by the respective leaf nodes. Leave-some-out (10%) cross validation study shows a respectable Q^2 of 0.91 and RMSE of 0.80, while an external test set has $R^2 = 0.94$ and RMSE = 0.68. Based on an overall comparison to methods described in literature and our own benchmark comparison, SMARTS pK_a outperforms previous models which have been trained on larger datasets as measured by the Pearson correlation coefficient, RMSE, and for having the fewest and least outstanding outliers.

One major difference between many prediction methods and SMARTS pK_a is that only one training set was used to derive our model, whereas other methods have combined many models specific to ionizable site types to produce their final prediction utility. It is far easier to overfit a model based on a small training set, especially when using qualitative descriptors. This leads to the major flaws inherent in any prediction method: the quality and amount of data available from which to train the model. By not breaking up the training set into subsets based on different ionizable site type, we were able to maintain the largest possible training set. Stepwise regression and the final decision tree led to the identification of 139 optimized SMARTS string descriptors. To prove that we had not overfit our model, we retrained the model 100 times, randomizing the pK_a prior to each retraining, and showed that the R^2 and RMSE of the randomized models were significantly worse than the non-randomized model.

Combining information from multiple datasets can increase the size and diversity of the training set reducing the prediction error.[48] Furthermore, with so much data being held proprietary, consensus models can lead to improved prediction accuracy and reduced overall errors. When considering the *in silico* evaluation of physicochemical properties of molecules in the pre and post screening stages of drug development, it is advisable to simultaneously examine multiple predictive models, as they are typically based on different training sets.

Issues of breadth and expanding the chemical space of our model are currently being addressed. We are now curating the data for over 10000 unique mono- and polyprotic molecules from Beilstein and Lange. The intent is to first expand the model to predict primary pK_a followed by a more comprehensive model capable of handling polyprotic molecules. Finally, we intend to combine SMARTS pK_a with MOE's Slog*P* to produce a utility for the prediction of log*D*.

This work was presented at the 40th Central Regional American Chemical Society meeting in Columbus Ohio in June of 2008 and has been published under the following reference: Lee, A. C.; Yu, J.-y.; Crippen, G. M. pK_a Prediction of Monoprotic Small Molecules the SMARTS Way. *J. Chem. Inf. Model.* **2008**, *48,* 2043–2053.

5.5 References

(1) Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.

(2) Avdeef, A. In *Absorption and Drug Development: Solubility, Permeability, and Charge State*; John Wiley & Sons: Hoboken, New Jersey, 2003, Chapter 1, pp 15–17.

(3) Hoener, B.A.; Benet, L.Z. In *Modern Pharmaceutics*; Banker, G.S.; Rhodes, C.T., Ed; Mercel Dekker Inc.: New York, 1990; pp 142–180.

(4) Wells, J.I. In *Pharmaceutical Preformulation*; Ellis Horwood Ltd.: New York, 1988, p 25.

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

(6) Chakrabarti, S.; Southard, M. Control of Poorly Soluble Drug Dissolution in Conditions Simulating the Gastrointestinal Tract Flow. 1. Effect of Tablet Geometry in Buffered Medium. *J. Pharm. Sci.* **1996**, *85*, 313–319.

(7) Upthagrove, A.L.; Nelson, W.L. Importance of Amine pK_a and distribution coefficient in the metabolism of fluorinated propanolol analogs: metabolism by CYP1A2. *Drug Metab. Dispos.* **2001**, *29*, 1377–1388.

(8) Oprea, T.I.; Marshall, G.R. Receptor-Based Prediction of Binding Affinities. In *Perspectives in Drug Discovery and Design*; Kubinyi, H., Folkers, G., Martin, Y.C., Eds; Kluwer/ESCOM: Great Britain, 1998; Vol 9-11.; pp 35–61.

(9) Alberati, D.; Hainzl, D.; Jolidon, S.; Krafft, E.A., Kurt, A.; Maier, A.; Pinnard, E.; Thomas, A.W.; Zimmerli, D. Predicting and Tuning Physiochemical Properties in Lead Optimization: Amine Basicities. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4311–4315.

(10) Jamieson, C.; Moir, E.M.; Rankovic, Z.; Wishart, G. Medicinal Chemistry of hERG Optimizations: Highlights and Hang-Ups. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(11) Fischer, H.; Kansy, M.; Bur, D. CAFCA: a Novel Tool for the Calculation of Amphiphilic Properties of Charged Drug Molecules. *Chima* **2000**, *54*, 640–645.

(12) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.

(13) Acton, G. Toxicogenomics and Predictive Toxicology: Market and Business Outlook: Vivo Group, Concord, MA, **2004**. http://www.vivogroup.com/reports.html, http://www.the-infoshop.com/study/cd25153_toxicogenomics.html (accessed Mar 14, 2008).

(14) Caldwell, G. W. Compound optimization in early- and late-phase drug discovery. Acceptable pharmacokinetic properties utilizing combined physicochemical, in vitro and in vivo screens. *Curr. Opin. Drug Discov.* **2000**, *3*, 30–41.

(15) Clark, J.; Perrin, D.D. Prediction of the Strengths of Organic Bases. *Q. Rev. Chem. Soc.* **1964**, *18*, 295–320.

(16) Perrin, D.D.; Dempsey, B.; Serjeant, E.P. pK_a *Prediction for Organic Acids and Bases*; Chapman & Hall: London, 1981.

(17) Dixon, S.L.; Jurs, P.C. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14*, 1460–1467.

(18) Hilal, S.H.; Karickhoff, S.W. A Rigorous Test for SPARC's Chemical Reactivity Models: Estimation of More Than 4300 Ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.

(19) Lee, P.H.; Ayyampalayam, S.N.; Carreira, L.A.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. In Silico Prediction of Ionization Constants of Drugs. *Mol. Pharm.* **2007**, *4*, 498–512.

(20) Xing, L.; Glen, R.C.; Clark, R.D. Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(21) Soriano, E.; Cerdán, S.; Ballestros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theo)* **2004**, *684*, 121–128.

(22) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.

(23) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(24) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values. *J. Chem. Inf. Model.* **2007**, *47*, 547-556.

(25) Ghasemi, J.; Saaidpour, S.; Brown, S.D. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct.* **2007**, *805*, 27–32.

(26) http://www.chemsilico.com/CS_prpKa/PKAexp.html (accessed Mar 11, 2008)

(27) Schüürmann, G.; Cossi, M.; Barone, V.; Tomasi, J. Prediction of the pK_a of Carboxylic Acids Using the ab Initio Continuum-Solvation Model PCM-UAHF, *J. Phys. Chem.* **1998**, *102*, 6706–6712.

(28) Eckert, F.; Klamt, A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19.

(29) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M.E. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pK_a Scale. *J. Phys. Chem. A* **2003**, *107*, 9830–9386.

(30) Pompe, M. Variable connectivity index as a tool for solving the 'anti-connectivity' problem. *Chem. Phys. Lett.* **2005**, *404*, 296–299.

(31) Pompe, M.; Randić, M. Variable Connectivity Model for Determination of pK_a Values for Selected Organic Acids. *Acta. Chim. Slov.* **2007**, *54*, 605–610.

(32) *MOE: Molecular Operating Environment*, version 2007.0902: Chemical Computing Group; Montreal, Quebec, Canada 2007.

(33) Dean, J.A. In *Lange's Handbook of Chemistry,* 15th ed.; McGraw-Hill: New York, 1999; Chapter 8, pp 8.24–8.72. http://www.knovel.com (accessed Apr 2007).

(34) *MDL CrossFire commander*, version 7; Elsevier MDL: San Leandro, CA, 2007.

(35) ChemDraw Ultra, version 10; CambridgeSoft: Cambridge, MA, 2007.

(36) Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 244–255.

(37) Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35 (accessed July 8, 2008).

(38) Rogers, D.J.; Tanimoto, T.T. A Computer Program for Classifying Plants. *Science,* **1960**, *132*, 1115–1118.

(39) Lee, A.C.; Shedden, K.; Rosania, G.R.; Crippen, G.M. Data Mining the NCI60 to Predict Generalized Cytotoxicity. *J. Chem. Inf. Comp. Sci.* **2008** (in press).

(40) SPARC Performs Automated Reasoning in Chemistry v4.2. http://ibmlc2.chem.uga.edu/sparc/ (accessed May 7, 2008).

(41) Szegezdi, J.; Csizmadia, F. New method for pK_a estimation. Proceedings of the eCheminformatics 2003 - Virtual Conference and Poster Session, Zeiningen, Switzerland, 2003; Hardy, B., Ed.; Douglas Connect: Zeiningen, Switzerland, 2003.

(42) ChemAxon. Marvin and Calculator Demo. http://www.chemaxon.com/demosite/marvin/index.html (accessed May 7, 2008).

(43) Advanced Chemistry Development ACD/Labs Online (I-Lab). http://www.acdlabs.com/ilab/ (accessed May 7, 2008).

(44) ADME/Tox WEB. http://pharma-algorithms.com/webboxes/ (accessed July 9, 2008),

(45) Dearden J.C.; Cronin, M.T.D., Lappin, D.C. A comparison of commercially available software for the prediction of pK_a. *J. Pharm. Pharmacol.* **2007**, Suppl. 1, A7.

(46) Meloun, M.; Bordovská, S. Benchmarking and validating algorithms that estimate pK_a values of drugs based on their molecular structures. *Anal. Bioanal. Chem*. **2007**, 389, 1267–1281.

(47) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155.

(48) Tetko, I.V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D.C.; Poda, G.I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today*. **2006**, *15/16*, 700–707.

Chapter 6

Future Directions:

Removing the Human Element from Physicochemical Property Prediction

6.1 Introduction

Decision trees have been widely used for data mining, organization, and classification. They are particularly useful for identifying simple relationships between variables, which can easily go unnoticed using other analytical techniques capable of modeling nonlinear data. Furthermore, decision trees are well suited to data mining tasks where little *a priori* knowledge is available and can be used to make predictions, even without a theoretical basis.

There are two types of decision trees: classification and regression. While classification trees are useful for grouping types of outcomes and providing qualitative responses, regression trees can give a quantitative response.

Recently, we investigated the predictive properties of regression trees for cytotoxicity and pK_a. Decisions were made based on presence or absence of molecular structures encoded as SMARTS strings. The selection and modification of generalized SMARTS strings used in training the pK_a model was largely manual, but the resulting regression tree was well-balanced and did not require the traditional post-pruning crossvalidation involved with optimization. Here we describe how the subjective selection of SMARTS strings could be replaced by the automatic generation of generalized SMARTS strings from the training compounds. As an example, aqueous solubility is suggested.

The seminal work by Breiman on classification and regression trees (CART) laid the foundation for CART modeling,[1] but optimizing tree structure using stepwise least squares regression was pioneered by Morgan and Sonquist in 1963 with their program Automatic Interaction Detection (AID).[2] The main differences between the AID and CART models are the pruning and estimation processes. While the AID model allows for limited lookahead, the CART model does not place restrictions on the number of values a variable can take, allows for variable combinations, the handling of missing data, assessment of variable importance, and subsampling. In this work we will describe a method for predicting aqueous solubility that shares some of the advantages of both models by adapting regression trees to the domain of chemical structural fragments represented by SMARTS strings.[3]

There are several advantages to using decision trees in predictive modeling. (1) Decision trees are simple, consisting of nodes and branches. At each node, a yes/no question is posed and the response effectively organizes the respective data. In any classification or regression scheme, the root node of the decision tree represents all data that needs to be organized, the decision path is the set of nodes and branches pertaining to a particular piece of data, and the leaf or terminal nodes contain the decisions/responses, as dictated by the model. (2) Decision trees are easy to interpret. Since the data is organized based on a series of <yes>/<no>, <is>/<is not>, or <has>/<has not> responses to some qualitative property, any decision path can be visualized, explaining how the final result is achieved. This is not possible with neural networks, which act as black boxes, or with linear regression, which provides a simple linear equation consisting of weighted descriptors. In general, it is easy to comprehend decision tree models with only a simple explanation, as they are all based on Boolean logic. (3) Decision trees are capable of handling both quantitative and qualitative data. (4) They are robust, well suited for nonlinear data modeling, capable of ignoring insignificant or unnecessary descriptors, and always provide a result. Other models may fail to provide a result based on limited parameterization. (5) Like any other model, decision trees can be validated using statistical tests. Noting that crossvalidation has been criticized,[4] it is always best to test new models on data that is external to the training set. (6) Most QSPR models, such as linear regression, are unable to return the exact property value when evaluating dependent variables from the training set. On the other hand, decision trees easily return lookup values, as long as the descriptor pool used to build the regression trees can completely grow the tree to uniquely identifying every dependent value in the training set. (8) Finally, decision trees are capable of providing predictions for vast amounts of data in a very short period of time.

Like all empirical prediction methods, decision trees have limitations. (1) Training an optimal decision tree is known to be an NP-complete problem, where NP stands for non-deterministic polynomial time. That is, a globally optimal solution to the problem exists, but the permutations of decision trees for a particular problem can be exponentially large and it may not be feasible to investigate every possible solution. Greedy algorithms are commonly used in order to more efficiently identify an acceptable solution by optimizing decisions locally for each node instead of the entire collection of nodes. As greedy algorithms are used to improve computational efficiency, there is no guarantee that the globally optimal decision tree will be produced.[5] Our implementation of this approach has several stages that lead to a relatively good and balanced decision tree. (2) Another concern is overfitting.[6] As with any empirical method, there is the danger of creating a model which does not generalize the data well. When using a large set of independent variables (descriptors) to model a dependent variable, it is quite possible that the quality of the predictions using a decision tree model or any other empirical model will reflect chance correlations. While it is possible to validate whether or not a predictive model yields chance correlations,[7] there are steps one can take to help avoid overfitting, such as pruning. Pruning is a way to mitigate the possibility of chance correlations due to overfitting. This is accomplished by reducing the number of nodes (decisions considered) in the decision tree. Two types of pruning exist: pre-pruning and post-pruning. Pre-pruning (forward pruning) avoids the generation of non-significant branches by halting the decision making process when no significant performance enhancement is gained by making further decisions. Common criteria for halting the model development along a specific decision pathway include reaching a predefined maximum path length, requiring some number of compounds greater than one at each node, and reaching a node where the property variance of the dependent variable is acceptably low. Post-pruning (backward pruning) is accomplished by using crossvalidation to identify non-significant branches for removal. Implementing crossvalidation can significantly reduce the explanatory power of a model. This is especially true when 10 or fewer folds are considered, as withholding molecules from the training set limits the scope of the potential descriptor space. Our regression trees only consider pre-pruning based on node size and property variations. (3) Another issue is the Boolean exclusive-or (XOR) problem. Considering properties *A* and *B*, *A* XOR *B* is false whenever both properties are true or both are false. *A* XOR *B* is true when only one of *A* or *B* is true, but not both. The XOR can easily be modeled by a decision tree, requiring extra decisions and potentially unnecessary complication of the model, but the property is hard to express using only a single node. See figure 6.1. The problem arises due to combinatorial issues when considering multiple properties at a single node.

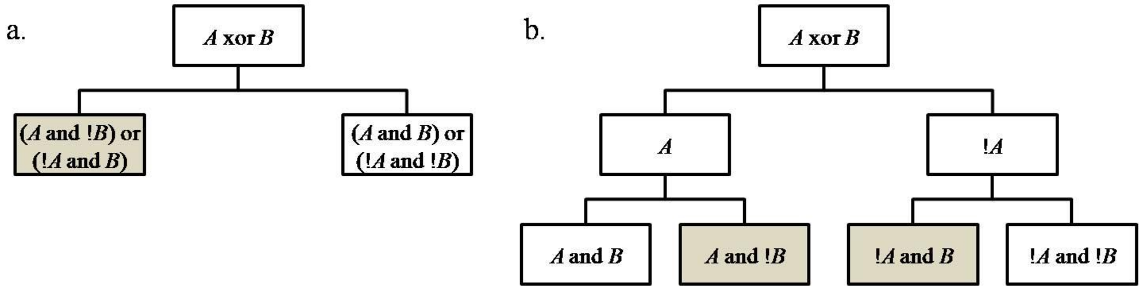


Figure 6.1. a. The XOR decision made at a single node. **b.** XOR processed by two separate decisions complicates the decision tree by increasing the number decisions made and doubling the number of decision sub-pathways from the parent node.

It is easy to see how a decision tree can become large. Approaches to solve this problem can involve changing the representation of the problem domain or using machine learning algorithms based on more expensive representations that involve statistical relational learning or inductive logic programming.[8] To help resolve this issue, our model relies on pre-pruning techniques and the use of generalized SMARTS strings. (4) Perhaps the greatest problem for all empirically based prediction methodologies, including regression trees, is parameter selection. The object is to find the minimal set of parameters that provides the maximal correlation between observed and predicted values for new data. These parameters are not always intuitive. Pearlman designed the Diverse Solutions (DVS) software package to handle this sort of cheminformatics diversity assessment. DVS is capable of defining a diverse and well distributed chemical space by identifying the best subset of BCUT descriptors, based on the least correlated eigenvalues of the Burden matrices.[9] Using this subset of descriptors it is possible to select representative subsets of compounds from a large library, to compare the diversity of two or more compound libraries, and as a means to fill the "diversity voids" in one compound collection with compounds from other molecular libraries.[10] In the methods section, we describe how automated machine learning techniques can be used to deal with these issues by extracting and modifying fragment based descriptors directly from the training set instead of relying on a predefined set of calculable chemical descriptors.

6.2 Methods

6.2.1 SMARTS and SMILES

We have chosen to use molecular fragments as the descriptors for each decision in our regression trees, as it is possible to explore every possible fragment which can be extracted from a molecular training. The Molecular fragments are represented by SMARTS strings. SMARTS, an extension of Daylight Inc.'s SMILES (Simplified Molecular Input Line Entry System),[11] is a language used for describing molecular patterns.[12] SMARTS strings extend the capabilities of simple fragmental based decisions as the atom types are generalizable. We used the SMARTS representations packaged within MOE, the Molecular Operating Environment.[13] MOE SMARTS include unique atom type variations not included in other versions, such as the *pi*-bonding atom type [i] and an atom type representing the sum of explicit bond orders: [v<*n*>].

The flexibility of the SMARTS language is very attractive from a chemical data modeling point of view. The SMARTS language can express some XOR relations as seen in figure 6.2. The descriptive power and general applicability to any molecular data set has led us to use SMARTS strings in our regression trees. Gepp and Hutter used decision trees to predict the likelihood that a molecule would induce Torsde de Pointes, or QT prolongation.[14] They identified a SMARTS string that proved “to be the most significant descriptor in the decision tree approach from which guidelines for the design of safe compounds are suggested.”

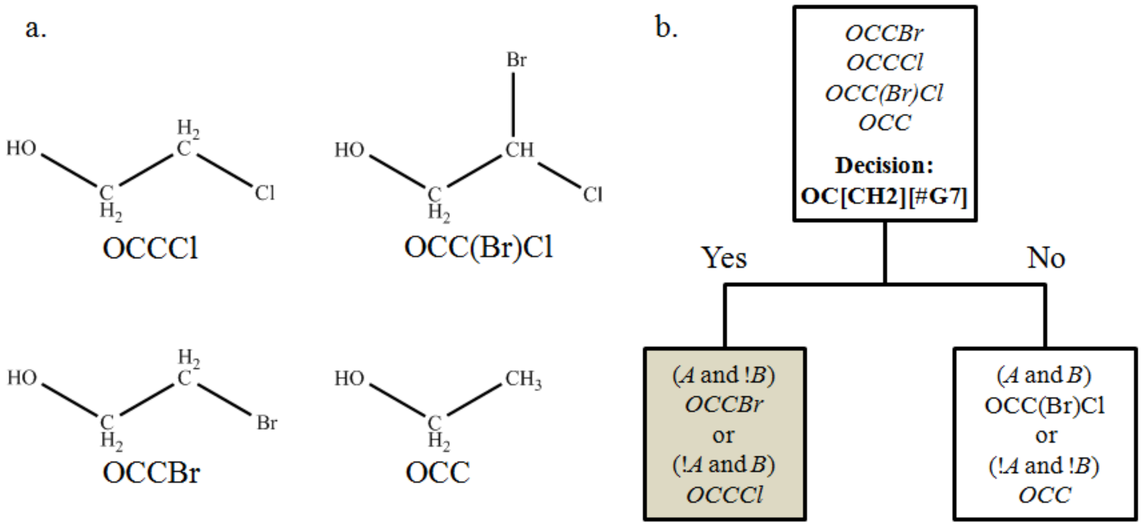


Figure 6.2. a. Ethanol and halogenated derivatives and their SMILES string representations. **b.** An example of a decision tree that can express an XOR relation with a single SMARTS string. Let *A* be 2-bromoethanol and *B* be 2-chloroethanol. Then OC[CH2][#G7] requires the 2-position carbon to have one halogen substituent and two hydrogens.

6.2.2 Data Selection

Even when dealing with small molecular datasets, a huge number of permutations can be expected. Here we suggest training and test sets for model building and assessment of regression trees. First, a competition recently advertised in the Journal of Chemical Information and Modeling offered a 101 small molecule data set for training, and another 32 molecules external to the training set for testing purposes.[15] Experimental data was provided for the training set only, while the assessment of predictions for the test set was done by the authors. Second, we have chosen to use solubility data from the PhysProp database[16] for those compounds that have experimental data for aqueous solubility at 25 °C, Henry's Law constant, and vapor pressure. Henry's law constant (K_H) is related to aqueous solubility and vapor pressure by the following equation, where p is the vapor pressure (gas outside of the solution) and c is aqueous solubility (gas in the solution):

$$K_H = p / c \tag{1}$$

The reason we chose this subset of the solubility data stems from the fact that Henry's Law constant and vapor pressure have previously been found to accurately model aqueous solubility.[17] We have randomly selected 334 molecules to be used as the training set and 58 molecules as the test set. Finally, a third training set was selected from a combination of the Huuskonen[18] and the Delaney[19] data sets where 1495 molecules were accepted as a training set and 167 were set aside for external testing. Prior to selecting the training and test sets, we performed a gross curation of the data, only accepting molecules with experimental data that could be verified by consensus among the data sources or by literature evaluation in Beilstein. The mean of the consensus log*S* values were taken as the accepted experimental value among these four datasets. Values were accepted as long as two or more experimental measurements were in agreement within 0.5 log units.

6.2.3 Training

6.2.3.1 Comprehensive Fragmentation to Obtain Chemical Descriptors

In order to sample the complete descriptor space, we fragmented each molecule in the training set. Every possible fragment of each molecule was considered and recorded as a SMARTS string along with the respective total number of hits against the training set molecules. In order to express molecular fragments as SMARTS strings, we needed to create an SVL script to preserve the aromatic atom types. Scientific vector language (SVL) is a high level language integrated in MOE which expands the usability of its toolkits and facilitates model development. Using SVL, one is able to extract the SMILES string for any molecule actively displayed in the molecule builder. Problems arise when breaking aromatic rings, as the SMILES representation for the bonded aromatic atoms will be converted into alternating single and double bonds with aliphatic atom types. If one were to use this SMILES representation as a SMARTS string and screen the molecular database, it would not hit the molecule from which it was derived. To correct this problem, common atom types were replaced with rare earth atom types. The atom type change serves two purposes. First, it forces all bond types to SP3 hybridization in MOE. Second, when extracting the SMILES representation, it allows for simple text parsing, so that the rare earth atom types can easily be changed back to their representative lower case aromatic (SMARTS) atom types after atom deletions have occurred. For example, if a molecule contained an aromatic ring represented by the SMILES o1cncc(OC)c1, those atoms would be forced to change into [Yb]1[Er][Tm][Er][Er](OC)[Er]1. At this point one or more of the ring atoms could be deleted. After the deletion, the SMILES string is then extracted in its canonical form, so that redundant fragments could be immediately eliminated from future consideration. For example, the following four SMILES strings are all correct representations of ethanol: OCC, C(O)C, C(C)O, and CCO. By forcing the canonical SMILES form of ethanol (OCC), duplicates can be efficiently identified and removed from consideration. Once uniqueness has been verified for the extracted SMILES string, it can be parsed to yield the correct SMARTS representation by converting any rare earth atom types to their respective organic aromatic atom types.

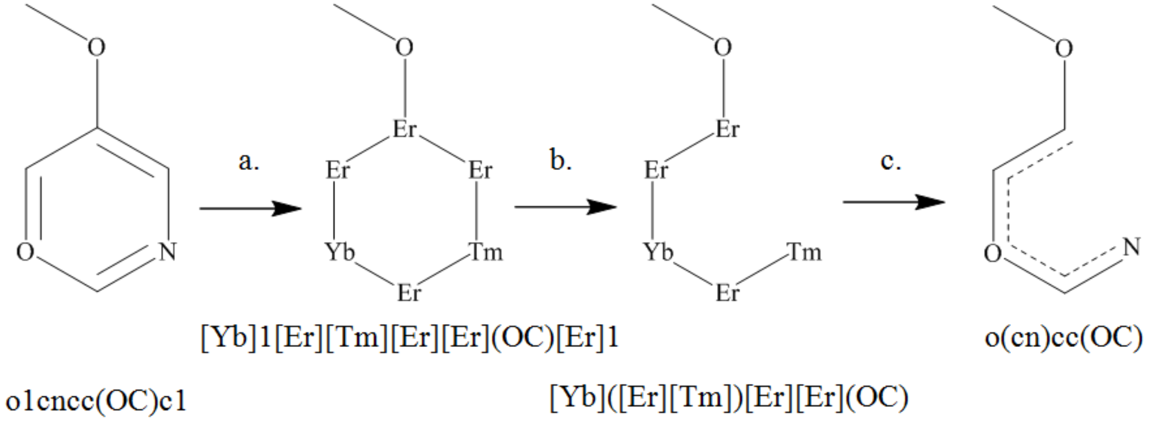


Figure 6.3. How to extract aromatic SMARTS strings in MOE. Step a: convert aromatic atoms to the lanthanide series. Step b: delete random atom(s) yielding a molecular fragment. Step c.: reconvert [Yb] → o, [Er] → c, and [Tm] → n.

6.2.3.2 Screen Training Set Molecules to Eliminate Redundancies

Once the set of molecular fragments was obtained, it can be used to screen the molecular training set. This serves two purposes. First to eliminate any SMARTS strings with redundant hit profiles. Second, to identify the strings with close to 'ideal' hit profiles. A SMARTS string with an 'ideal' hit profile would evenly split the training set into two groups: one with molecules containing the SMARTS string and the other having molecules that do not contain the molecular fragment represented by the SMARTS string.

6.2.3.3 Identify Predictive SMARTS String Descriptors

The number of SMARTS strings is derived from the fragmentation of a database of small molecules exponentially large. Fragmenting benzene (c1ccccc1) gives the five SMARTS strings: c, cc, ccc, cccc, ccccc. Fragmenting phenol gives the SMARTS strings from benzene plus: c1ccccc1, O, Oc, Occ, Occc, Occcc, Occccc, O(c)c, O(c)cc, O(c)cc, O(c)ccc, O(c)cccc, O(cc)cc, O(cc)ccc. The number of fragments obtained from a molecule depends on atom count, branching, configurational constraints, and atom types. Increasing any one of these factors can lead to an exponential increase in the number of potential fragments. See table 6.1. One observation that is readily evident from the comparison of set size to number of fragments for each respective set is that the Contest training set has significantly higher diversity in fragment space than does the PhysProp training. This is due to the higher atom count of the molecules in the Contest training set. It seems that considering fragment count and overlap would be a novel way to assess the similarity between chemical libraries.

Table 6.1. Relationship between potential fragment count and database size.

Source	Set Size	# of Fragments	Mean MW	MW sd.
Contest[a]	101	~ 0.3 million	285.8	90.1
PhysProp[b]	334	~ 0.1 million	159.8	99.8
H & D[c]	1495	~ 1.8 million	206.4	100.8

[a] Training set published in the JCIM contest. [b] Subset of the PhysProp database containing approximately 70% of the molecules having data for aqueous solubility, vapor pressure, and Henry's Law constant. [c] Combination of the Huuskonen and Delaney datasets.

For any molecular database containing an independent set of molecules, a set of SMARTS strings can be derived that uniquely identifies every molecule in the database. Considering the overwhelming number of descriptors produced by fragmentation, it is easy to envision many combinations of descriptors that perfectly fit any training set. As previously discussed in Chapter 2, a model that can perfectly fit a set of training data, tells nothing of that model's ability to predict new data. Therefore, it is important to identify a small subset of descriptors that can most accurately explain the model. Using regression trees, the minimum number of descriptors required to perfectly fit a set of data is $\log_2 N$ rounded up, where N is the number of molecules in the training set.

Furthermore, several non-overlapping groups of descriptors fitting this description may be identified for any particular training set.

With regression trees, identifying the optimal set of descriptors can be accomplished by applying logical reduction techniques:

a. Groups of descriptors uniquely identifying the molecules of the training set need to be mined from the training set fragments. Note, over 10 gigabytes (GB) of memory is required to maintain the hit profiles for each fragment obtained from training set 2. This exceeds the 4 GB of available RAM in our workstation, furthermore, it is also beyond the 8 GB capacity of MOE databases. Therefore, a significant amount of redundant molecular screening would be required for data sets of more than a few hundred molecules having molecular weight less than 500.

b. Apply backward stepwise regression, as seen in section 5.2 sub-steps 1.5 – 1.5.4, to the optimal descriptor set.

    i. If more than one optimal descriptor set exists, the resultant descriptors from the stepwise regression can be combined and backward stepwise regression can once again be performed to identify the SMARTS strings that have the most explanatory power for the particular training set being modeled.

6.2.3.4 Refine and Expand Predictive SMARTS String Descriptors

a. Predict the dependent variable based on molecular similarity, using the method described in section 5.2 sub-steps 1.4 – 1.5, where the optimal descriptor set is used for fingerprint based predictions.

b. Identify new predictive SMARTS strings. This step is very similar to that described in section 5.2 sub-steps 1.6 – 1.6.4 and will lead to the development of a refined set of SMARTS strings which fit the model well. While the new SMARTS string additions will be less ‘ideal’ than the current pool of SMARTS strings, they will be more ideal for the molecules identified by a specific fingerprint. The newly identified SMARTS strings will serve to reduce the experimental property variance and improve the overall fit of the model, while minimizing the number of new additions to the pool of SMARTS strings considered in the regression tree.

   i. Identify groups of training set molecules predicted to have the same value for the dependent variable, but exhibit a large variance for the experimental value.

   ii. Generate all potential molecular fragments for this subset of molecules.

   iii. Identify the molecular fragment which would improve the r^2 statistic if included in the fingerprint.

   iv. Repeat steps i.–iii. until the r^2 is above some specified value (0.96) with no significant outliers (greater than one log unit). These are only suggested values based on our previous model for predicting pK_a.

6.2.3.5 Grow the Regression Tree

a. Follow the process discussed in section 5.2.2 sub-section 2.1 to grow the initial decision tree and identify the best weight for the ‘accuracy’ factor, which is responsible for optimizing the fit of the training set.

6.2.3.6 Fine Tune the Regression Tree by Generalizing SMARTS String Descriptors

a. Identify the leaf node with the highest experimental property variance.

b. Perform reverse stepwise regression from leaf to root.

   i. Fragment the SMARTS string responsible for the decision at the parent node.

   ii. Randomly substitute one or more generalized MOE atom types for the specific atom types in each SMARTS string fragment and rescore each decision according to the methodology used to grow the regression tree.

   iii. Add the SMARTS string that results in the best score to the pool of SMARTS strings considered when growing the regression tree. If the score does not improve, repeat step ii. allowing a maximum of three iterations. If no improvements in the score can be identified, randomly select one generalized SMARTS strings having the same score as the original to the pool of SMARTS.

   iv. Retrain the regression tree.

   v. Predict the dependent variable for all training molecules. As a SMARTS string may be used more than once to process a decision in the regression tree, the addition of new more general descriptors can affect more than one decision pathway. Therefore, if the r^2 and RMSE improve or remain the same, maintain only the SMARTS strings used in the training process for further consideration, otherwise disregard the new additions.

vi. Repeat this procedure for the next higher parent node until the root is reached.

vii. Repeat this procedure until each decision pathway has been completely examined and generalized or some threshold variance (within 0.4 log units) for the experimental dependent variable at each leaf is reached or cannot be improved upon.

6.2.3.7 Test for Overfitting Using Dependent Variable Randomizations

a. Randomize the dependent variables for the molecules in the training set.

b. Retrain the regression tree.

c. Predict the dependent variable for the training set molecules.

d. Record the r^2 and RMSE statistics.

e. Repeat randomization test 100 times, and assess the randomization results. The r^2 and RMSE statistics for the non-randomized model need to be respectively higher and lower than the randomized model in order to show that the predicted values are not due to chance correlations and overfitting.

6.2.3.8 Validation: Test on External Data

6.3 Conclusions

One way to improve our methodology might be to consider decision graphs. Using decision trees, all decision pathways start at the root node and terminate at some leaf by way of the Boolean AND operator. Decision graphs that can solve the XOR problem would use the Boolean OR operator to join two or more pathways using Minimum Message Length (MML).[20] Furthermore, decision graphs have been dynamically trained, where new attributes are allowed to occur in different places within the graph,[21] much the same as in our regression trees. By generalizing the coding scheme, models with fewer leaves and improved accuracy may be possible. It appears that generalizing the SMARTS strings has afforded us some of the advantages inherent in decision graph models, namely the ability to minimize the number of decisions and descriptors used.

Another well documented methodology that may prove useful with our fragment based analysis is that of bagging predictors.[22] Here simplicity and interpretability are sacrificed for accuracy. Bagging predictors would lead to the derivation of multiple aqueous solubility predictors, which in the end would be used together as an aggregate predictor. The combined model would average over the multiple versions created by making bootstrap replicates of the training set and using these as the new training sets. According to Breiman, “bagging goes a long way toward making a silk purse out of a sow’s ear.” However, thus far it is our experience that SMARTS trained regression trees are capable of outperforming similar regression forest methodologies, albeit different sets of descriptors were considered.[23]

In this chapter we have described how the human element can be removed from the derivation of regression trees capable of physicochemical property prediction, provided that a well curated and diverse molecular data set with experimental property values exists. We suggest that the methodology be applied to aqueous solubility data, as this has been the focus of recent attention in literature. Several data sources (PhysProp, literature, and Beilstein) have been mined to facilitate data curation. Furthermore, a recent competition was held by the Journal of Chemical Information and Modeling.[24] Here 100 molecules with experimental measurements for aqueous solubility were presented as training data, and the participants were expected to send in predictions for 32 other small molecules for which the experimental data was not divulged. While we feel that it is unlikely that any empirical model can achieve good predictions based on such a small training set, this provides a means for comparison of the accuracy of our regression tree model to that of other robust methodologies.

6.4 References

(1) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. In *Classification and Regression Trees*, 1st ed.; Wadsworth Inc.: Belmont, California, 1984.

(2) Morgan, J. N.; Sonquist, J. A. Problems in the analysis of survey data and a proposal. *J. Amer. Stat. Assn.,***1963**, *58*, 415–434.

(3) Daylight Chemical Information Systems Inc. http://www.daylight.com/ dayhtml/doc/theory/theory.smarts.html (accessed Jun 25, 2009).

(4) Golbraikh, A.; Tropsha, A. Beware *q*2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

(5) Hyafil, L.; Rivest R. L. Constructing Optimal Binary Decision Trees is NP-complete. *Inf. Proc. Lett.*, **1976**, *5(1)*, 15–17.

(6) Brammer, M. Avoiding overfitting of decision trees. In *Principles of Data Mining.* 1st ed.; Mackie, I., Ed.; Springer: London, England, 2007; pp 119–134.

(7) Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1166–1068.

(8) In *Inductive Logic Programming.* 1st ed.; Carbonell, J. G.; Siekmann, J., Eds.; Springer: New York, 2003.

(9) Burden, F. R. *J.* Molecular Identification Number for Substructure Searches. *Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.

(10) Pearlman, R.S., Smith, K.M. Novel Software Tools for Chemical Diversity. *Perspectives in Drug Discovery and Design*, **1998**, 9/10/11, 339–353.

(11) Daylight Chemical Information Systems Inc. http://www.daylight.com/smiles/ (accessed Jun 25, 2009).

(12) Daylight Chemical Information Systems Inc. http://www.daylight.com/ dayhtml/doc/theory/theory.smarts.html (accessed Jun 25, 2009).

(13) *MOE: Molecular Operating Environment*, version 2008.10: Chemical Computing Group; Montreal, Quebec, Canada 2008.

(14) Gepp, M. M.; Hutter, M. C. Determination of hERG channel blockers using a decision tree. *Bioorg. Med. Chem.* **2006**, *14*, 5325–5332.

(15) Llinás, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.

(16) *PhysProp*, SRC Inc.: Syracuse, NY, 2009. http://www.srcinc.com/what-we-do/product.aspx?id=133 (accessed on Mar 9th 2009).

(17) Paasivirta, J.; Sinkkonen, S.; Mikkelson, P.; Rantio, T.; Wania, F. Estimation of Vapor Pressures, Solubilities and Henry's Law Constants of Selected Persistent Organic Pollutants as Functions of Temperature. *Chemosphere* **1999**, *39,* 811–832.

(18) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.

(19) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.

(20) Oliver, J. J. *Decision Graphs – An Extension of Decision Trees;* Technical Report 92/173; Monash University: Victoria, Australia, 1992.

(21) Tan, P. J.; Dowe, D. L.; MML Inference of Decision Graphs with Multi-way Joins, In *AI 2002: Advances in Artificial Intelligence*, 1st ed.; Mckay, B.; Slaney J., Eds.; Springer: New York, 2002; pp. *131*-142.

(22) Breiman, L. *Bagging Predictors;* Technical Report 421; Department of Statistics, University of California: Berkeley, CA 1994.

(23) Lee, A. C.; Shedden, K. S.; Rosania, G. R.; Crippen, G. M. Data Mining the NCI60 to Predict Generalized Cytotoxicity. *J. Chem. Inf. Model.* **2008**, *48*, 1379–1388.

(24) Llinás, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.

Chapter 7

Conclusions

There is a great need for fast, accurate, and robust models for physicochemical property predictions. To answer this need, we have introduced a methodology capable of a uniform structure based analysis, which can model and self validate for any property. We have been able to demonstrate high accuracy in the case of monoprotic pK_a prediction. To date, we have not seen any other automated application capable of self training that considers the entire set of molecular substructures in a molecular training set. Our automated regression tree stores all potential unique molecular fragments as SMARTS strings, and selects the optimal subset for model building purposes. Furthermore, our tool has taken machine learning to another level by enabling the automated generalization of molecular fragments to fine tune the final regression tree.

One major difference between many prediction methods and our regression tree based methodology is that only one training set was used to derive each model, whereas other methods have combined many models specific to congeneric series of molecules in order to produce their final prediction utility. It is far easier to overfit a model based on a small training set, especially when using qualitative descriptors. Our methodology allows us to maintain the largest possible training set by not breaking it into subsets based on chemical classes. While it may appear that regression trees use an abnormally large number of descriptors, one needs to take into consideration that a decision path uses far fewer descriptors (on the order of $\log_2 N$, where N is the number of molecules in the training set) than are used to derive the entire tree. Furthermore, we have applied steps similar to Topliss evaluation of identifying chance correlations in QSARs to our regression trees for cytotoxicity, pK_a, and aqueous solubility to verify that overfitting has not occurred.

The limiting factors for every experimental methodology are and will always be time and money. Whether it be the costs and time associated with obtaining or synthesizing the molecules of interest or a simple titration, it is not physically possible to explore and experimentally quantify even a small portion of chemical space. Domains are increased and bottlenecks are widened by computational resources, but limitations remain. Even if a single physicochemical property, such as aqueous solubility, could be calculated for the theoretical pool of 10^{60} small molecules with a supercomputer consisting of 10,000 1THz (~10^{12} Hz) CPUs, it would still take on the order of 10^{44} seconds to perform calculations for all the molecules. Finally, this completely disregards the much slower speed of storage, required for the characterization of chemical space and the multitude of conceivable descriptors which could be calculated. As in drug discovery, we must focus our attention on a target, framing the problem at hand.

The main advantage of *in silico* prediction is that physical samples are not needed. Still, new compounds need to be synthesized for experimental evaluation of physicochemical properties to better understand chemical space and expand the diversity of the molecules available to update existing models and develop new prediction methods.

As emphasized in chapter 2, the root of all good empirically based methodologies for physical property predictions is the data. Refining chemical datasets can facilitate the process of drug development by helping to minimize the high attrition rate due to poor ADMET during the clinical phases. The largest problems with current public domain physicochemical chemical property and biological activity data are the lack of curation procedure and quality control. We have shown that even in the cases where curated datasets are available, one must carefully evaluate the data in order to ensure the greatest accuracy for data mining purposes.

Quantity, chemical diversity and quality are the key components for all molecular training sets. Combining information from multiple datasets can increase the size and diversity of the training set and potentially reduce the prediction error. To this effect, there needs to be a public collaboration beyond the scope of Beilstein and the Chemical Abstracts Service to create and curate a universal database of all existing experimental data. Curation is a tremendous issue when considering the scope of literature dealing with data redundancy, procedural variations, and human error. Moreover, guidelines need to be established, such that the data for each property and future experimental measurements are made and recorded in uniform format, such as the extrapolation of pK_as to zero ionic strength for H_2O. Big Pharma can aid the cause by agreeing to provide benchmark scores on external data, but there needs to be some information sharing regarding the chemical diversity of their internal test set. A test based on 1000 molecules which are part of a congeneric series may tell nothing about the robustness and accuracy of a particular model, as poor predictions for a congeneric series of molecules are likely to be caused by a lack of chemical representation in the training set. Finally, test data needs to be diverse and external to the training set, otherwise no true evaluation can be made.

In chapters 3–5 simple preprocessing and curation procedures were discussed. Curation involved identifying conflicting experimental results, eliminating gross outliers, and significant literature searching. Principal component analysis was used as a means to validate the dimensionality of a large subset of the NCI60 both with and without imputed values. The same steps can be used to refine and validate screening data from multiple assays, whether they be a subset of the NCI60, a combination of the NCI60 and other biological screens, or any selection of HTS drawing their activity data from a common set of substances.

Finally, there is also a need for a real benchmark so that a true comparison of methodologies can be made. We have also explored the use of consensus models in relation to pK_a prediction and shown how they can lead to improved prediction accuracy and reduced overall errors. However, we understand that this improvement may be due to the inclusion of test set data in the training sets in some models with undisclosed training data. For the time being, we must set aside the notion that consensus models for physicochemical property predictions are more accurate, as this has yet to be experimentally determined. Until a group of predictive models for a particular property are retrained on exactly the same data, there is no way to ascertain which model will perform the best for the respective property. If we can reach the stage where several models can be trained and tested on the same respective data sets, then and only then can consensus models truly be evaluated.

One final thought: automated regression trees may be well suited to the task at hand. Our prediction utility for monoprotic small molecules is currently available on the Chemical Computing Group's SVL Exchange website, and we predict that the methodology will be appearing in industrial applications in the near future.

.