Chapter 1

Introduction

1.1 To Calculate or Predict?

Until recently advances in medicinal and pharmaceutical chemistry depended on a trial and error process aided by intuition. Though the properties that would indicate a certain molecule as a drug candidate were known, it was not really feasible to investigate large numbers of molecules for these types of properties. Of course, the nature of these properties would be represented by structural features of a molecule and thus examination of certain motifs provided a direction for experimental investigations.

The problem with this approach is that it does not always lead to an understanding of why a molecule behaves as a drug against its target or why it does so. Furthermore, given a series of compounds it is not always feasible to investigate experimentally which members of the series would be more potent or less toxic. As a result, though medicinal chemistry has resulted in a series of life saving drugs, the process has traditionally been slow and tedious, and in many cases advances have been due to serendipity rather than scientifically guided investigation.

In an ideal world one would be able to take a 3-D molecular structure and calculate the required properties. This utopian goal has a number of problems associated with it. First, what type of properties are to be calculated? Certain intrinsic physical properties can be calculated using *ab initio* quantum mechanical computation techniques. Examples include dipole moments, charges and heats of formation. Though these are certainly useful, they do not provide much insight into drug-like properties such as potency and bioavailability. In addition, for large collections of molecules, *ab initio* techniques become very time consuming. Semi-empirical quantum mechanical methods alleviate the intensive nature of these calculations, but we are still faced with the restriction on the types of properties that can be calculated. Second, the drug-like activity of a molecule is intimately related to the target it is supposed to interact with. Targets generally involve some type of protein to which the putative drug will bind. Thus when considering the activity of a drug, we cannot simply consider the properties of the drug molecule itself. That is, the nature of the interaction between the drug and target must be investigated to

understand fully the activity of a drug. However, *ab initio* and semi-empirical techniques have traditionally not been suited for the modeling of large protein systems. Though recent advances in linear scaling[1,2] and hybrid techniques[3,4] have expanded the purview of quantum mechanical methods to systems containing tens of thousands of molecules, these methods are still not efficient enough to model thousands or millions of molecular structures, and their associated targets, at a time. Third, though the interactions of a drug with its target are certainly important, the drug must be absorbed by cells and the also metabolized and excreted from the body. Thus absorption properties, the nature of the metabolites and other characteristics must also be considered. Clearly, these are very complex properties that involve interactions with a large number of cellular processes. Modeling these quantum mechanically is nearly impossible.

The above discussion illustrates two fundamental problems. It is not feasible to calculate from theory all the properties of a drug molecule that would help us understand its activity and its utility, and we want to be able to analyze large sets of molecules for these properties.

Why do we need to analyze large sets of molecules? The reason for this is closely tied to the nature of drug discovery in recent years. The drug discovery process is time consuming and expensive. Often it can take 10 to 15 years for a drug to reach the market from the laboratory. Given this situation, it is important that a company select the proper compound for study. Combined with the results from high throughput screens[5] and in-house libraries, this can mean having to select tens or hundreds of compounds from a collection of millions. Furthermore, the ability to generate an arbitrary number of unique chemical structures *in silico*, to create virtual libraries, supplants the actual compounds that a company might have synthesized in its physical collection. Clearly, testing each compound libraries (virtual or real) for drug-like properties is out of the question. As we have seen above, calculating properties for collections of this size is either not feasible or impossible. The question thus comes down to this: how can we calculate arbitrary properties of hundreds of thousands of molecules rapidly and accurately? The short answer is that we avoid the calculation step completely and instead *predict* a property of a set molecules based on a model derived from the measured values of that property for a small subset.

1.2 Origins of QSAR

The predictive approach is essentially a statistical methodology and is known as the development of quantitative structure-activity relationship (QSAR)and quantitative structure-property relationship (QSPR) models, first described by Hansch[6,7] and Free et al.[8] In general, the term QSPR refers to the case where we are considering physical properties and QSAR refers to the situation where we are considering biological activities. However, in this work the term QSAR is used to include both cases.

Though Hansch was the first worker to define the term QSAR, A.F.A Cros, in 1863, had noted that the toxicity of alcohols in mammals increased with the decrease in water solubility.[9] Workers in the 1890's noted that toxicity of organic compounds depended on their lipophilicity. The precursor to QSAR models were linear free energy relationships such as the Hammett equation,[10] which was originally defined as a relationship between the electronic properties of acids (and bases) and their disassociation constants and reactivity. The equation is defined as

$$\log \frac{K}{K_0} = \rho \log \frac{K'}{K'_0} \tag{1.1}$$

where K and K' represent the dissociation constants for a set of substituted aromatic acids and K_0 and K'_0 are the constants for the unsubstituted acids. ρ is the slope of the best fit line from the model fitted to the observed constants. The term $\log(K'/K'_0)$ is denoted by σ and describes the substituents.

Hansch originally tried to develop QSAR models using the Hammett σ parameter but this did not lead to good results. He thus considered other parameters such as the lipophilicity and molecular size as represented by molar refractivity.

The essence of the QSAR methodology is thus developing a relationship between an observed property and structural features of a molecule. By considering a set of molecules, a predictive model is developed that can then be used to predict the activity of other molecules. The key words here are "structural features". The approach depends on being able to represent the structure of a molecule in numerical form. This is in contrast to the use of empirical parameters (σ) in the case of linear free energy relationships. The numerical representations of molecules are termed *descriptors*, and a wide variety of descriptors can be calculated. These include simple forms such as molecular weight and atom counts or more complex types such as partition coefficients and surface-property descriptors. Given a set of descriptors, a QSAR model can be built

by defining a relationship between these descriptors (also known as the independent variables) and the observed property (termed the dependent variable). The first QSAR models, developed by Hansch, specified linear relationships. Even now, linear models are widely used owing to their simplicity and ease of development. However, developments in the field of statistics have produced many new methods of building predictive models. These include nonlinear regression techniques and algorithmic techniques.[11] Other fields such as pattern recognition and machine learning have also developed methods that have been used successfully in QSAR modeling. These include neural networks and subsequent variants,[12,13] decision trees[14,15] and so on. Clearly, progress in the field of QSAR modeling is closely tied to developments in a number of fields including statistics, computer science and mathematics.

The process of QSAR modeling is summarized in Fig. 1.1. The diagram stresses the fact that a QSAR model is an alternate stepwise route to the calculation of molecular properties. That is, the direct calculation of molecular properties is generally not feasible. In addition, if we do not understand the nature of interactions that a molecule undergoes in expressing its activity, accurate calculation of its properties is impossible. Thus, we proceed by an indirect route, which we term the QSAR pipeline, in which we represent a molecule in a computer understandable format, distill molecular features by calculating molecular descriptors and then build predictive models. The important feature of the QSAR modeling process is that it *predicts* molecular properties rather than *calculating* them. This fact raises a number of issues such as the validity of predictions. Another important aspect is the nature of the information that is input to the model. That is, what types of descriptors should the model use, given that we can calculate thousands of them. This is the problem of feature selection. Finally, the model predicts molecular properties based on information present in the dataset that it has encoded. It is thus important to be able to extract the encoded information from the model, and this is the topic of interpretability. These issues are discussed later in this thesis.

Though the above discussion has focused on drug molecules, the QSAR methodology is certainly not restricted to these types of molecules. In fact a QSAR model can be built to predict any type of physical property of biological activity, given a set of observations and molecular structures. Examples of the prediction of physical properties include boiling points,[16–18] aqueous solubility,[19,20] glass transition temperatures[21] and ion mobility.[22] In the area of biological activities QSAR models have been developed to predict genotoxicity,[23–25] carcinogenicity[26,27] and mutagenicity.[28,29] Furthermore, the use of QSAR models is not restricted to their role in screening large libraries of compounds.

In some cases a series of compounds may be synthesized and assayed. The development of a QSAR model for these compounds would provide the synthetic chemist some idea of what types of compounds could be synthesized to exhibit better activity. In other cases, the structural features highlighted by a QSAR model can provide insight into the mode of action of a drug molecule, which might be otherwise difficult to ascertain by experimental means.

1.3 QSAR Methodologies

QSAR methodologies can be broadly divided into three groups. First, 2-D methodologies do not consider the 3-D structure of a molecule directly. Instead, the molecule is represented by a set of molecular descriptors, numerical values characterizing various aspects of molecular structure. Together with the observed activity, a predictive model is built. It should be noted, that even though some descriptors are based on 3-D coordinates, the method as a whole considers only the observed property and the descriptors, and hence is 2-D in nature. The ADAPT software suite implements the 2-D QSAR methodology.

The second type of methodology is 3-D in nature and is exemplified by the CoMFA[30] approach. In this case, the 3-D structure of the molecule is the object of study. The molecule is aligned on a grid and various properties are evaluated at a set of grid points. Clearly, this type of approach has many advantages over the more simplistic 2-D methodology. The fact that the molecule is studied directly in three dimensions, rather than being mapped to two, allows for a clearer view of the interactions between the molecule and its target that play a role in the observed activity. However it does require accurate alignments and only considers a single conformation of a molecule.

The 4-D QSAR methodology is an extension of the 3-D QSAR methodology developed by Hopfinger et al.[31] which considers conformational information as the fourth dimension. Similar to the CoMFA method, 4-D QSAR starts of by defining a set of grid points on which molecular properties will be evaluated. In addition to the grid points, the method performs conformational ensemble sampling and uses the information obtained to evaluate grid cell occupancies. These occupancies are then used to evaluate *interaction pharmacophore elements* (IPE's). The IPE's together with the molecular properties are then used to develop a predictive model.

This work focuses on the 2-D QSAR methodology and presents investigations carried out on certain steps of the model building process. Compared to the 3-D and 4-D methodologies described above the 2-D approach has a number of advantages. First, owing to the variety of molecular descriptors available, optimized coordinates are not always required. In fact, connectivity information (in the form of SMILES strings or an adjacency matrix) alone, can be used to develop QSAR models. As a result models using these types of descriptors (termed topological descriptors) can be built rapidly for very large sets of molecules. However, these types of descriptors are in general quite abstract and so if the model is to be analyzed to extract information regarding structure-property trends, other, more physically meaningful descriptors will generally be required. Second, this approach avoids the alignment step and thus can be used in the absence of experimental information regarding the binding of a molecule to its target.

The downside to the 2-D QSAR methodology is that it does not provide a detailed answer to a number of questions regarding a molecule's activity. That is, by representing structural information in the form of descriptors, aspects of a molecules activity such as its absorption properties or degradability are hidden by a layer of abstraction or not addressed at all. Thus a molecule might be observed to have low activity. A 2-D model may not be able to indicate whether this is due to its inability to bind to the target or whether this is due to its inability to cross the cell membrane. The point is that, in a 2-D QSAR model, a lot of information about various aspects of a molecule's activity are combined together and are not always individually apparent. Though interpretation methods for linear QSAR models exist, they are obviously restricted to the information encoded by the descriptors in the model. This means that though 2-D QSAR models are certainly very useful, especially for screening purposes, they should be used in conjunction with other types of models to fully understand the role that various structural features play in determining the activity of a molecule.

2-D QSAR models can also be divided into two distinct groups, namely, qualitative and quantitative models. The former type of model, also known as classificatory models, consider a categorical dependent variable. That is, the observed property for each observation is represented by a label, such as toxic or non-toxic. Thus, if a dataset is available for which an assay has been carried out indicating whether a given molecule is carcinogenic or not, a 2-D qualitative model can be built that will predict whether a molecule , not belonging to the set, is carcinogenic or not. These types of models are not restricted to yes/no problems and datasets with multiple classes (say, active, moderately active and inactive) can be modeled. The second type of 2-D QSAR models

are referred to as quantitative (or regression) models. The function of these types of models is to predict a numerical value for a property, for example, boiling points or IC_{50} values. At the same time it should be pointed out that even when the observed property for a dataset is numeric in nature, it can be studied using qualitative models. This is generally achieved by selecting a break point in the range of the observed values and placing molecules whose property is above the break point in one class and the remaining molecules in another class. With these class assignments, a classificatory model can then be built. This thesis focuses on the development of regression models.

An important part of QSAR modeling is the use of software to create structures, calculate descriptors and build predictive models. A number of commercial packages provide QSAR modeling facilities, and examples include Cerius2[32] from Accelrys and Strike[33] from Schrodinger. These packages provide a comprehensive environment that is linked to chemical databases and a variety of cheminformatics functionality and as a result, encompass the whole process of model building and data analysis. Some examples of freely available programs include PowerMV[34] and the ADAPT system described in this thesis. Other programs tend to focus on specific aspects of the QSAR model building process. For example, a number of programs are available to calculate descriptors. Examples include Dragon,[35] JOELib[36] and Codessa.[37] Some programs focus on calculating a set of properties that can indicate the drug likeness of a molecule, such as metabolite types, bioavailability and so on. An example of such a program is QikProp developed by Jorgensen et al.[38–40] It is obvious that a fundamental component of QSAR modeling is the statistical analysis of chemical information. Thus, a number of statistical packages can be used to perform QSAR modeling such as SAS, Splus and R.[41] One problem with these environments is that they are geared towards statistics. As a result, having access to chemical functionality from within these statistical environments is attractive. An example of this type of environment is the combination of R and the Chemistry Development Kit (CDK)[42] described by Guha[43] allowing the user to have access to the full statistical capabilities of R as well as the cheminformatics capabilities of the CDK.

1.4 An Outline

This section briefly outlines the various topics considered in this thesis. Chapter 2 introduces the modeling techniques that are used in this work. Though a detailed presentation of the various algorithms and models that are used in QSAR modeling would take up a whole book, the chapter describes the broad classes of models and algorithms

employed in this work and focuses in the theoretical principles of some specific methods. Chapter 3 then gives a detailed description of the general QSAR methodology that is employed in the various studies presented in this work. Subsequent chapters represent investigations and applications that have been carried out on specific steps of the QSAR model building process.

Chapter 4 focuses on the set selection step. This step in the QSAR pipeline divides the original dataset into subsets which are collectively known as QSAR sets. These subsets are then used to build and test the QSAR model. A set selection procedure is developed to create representative sets for the purpose of building and testing QSAR models using a self-organized map.[44] The assumption underlying this method is that if the features of the dataset are proportionately represented in the subsets used to build and test a QSAR model, the resultant model should exhibit better predictive ability and should be more reliable, than models built with sets selected by random selection which does not necessarily represent different features proportionately.

Chapter 5 then focuses on the validation step of the QSAR pipeline and describes a technique that was developed to be able to ascertain the reliability of a QSAR when asked to predict properties of compounds that it has never seen. The validation of QSAR models, over and above the traditional methods, using scrambling tests and an external prediction set, is an important topic. The ability to obtain a measure of confidence in the predictions of a QSAR model is very important when such models are used to process incoming data from high throughput screens or when used by a bench chemist to decide whether to invest time and effort on the characterization of a new lead. Some model types do allow confidence measures to be calculated, but these are generally specific to the model type. The method described in Chapter 5 presents a much more general approach to this problem, applicable to any type of quantitative model.

The next four chapters focus on the topic of interpretability. Chapters 6 and 7 describe the development and interpretation of linear regression QSAR models. Chapter 6 presents a study of a set of artemisinin analogs that were designed for their anti-malarial activity. Both linear and nonlinear models are developed and the former is subsequently interpreted using the PLS technique. Chapter 7 describes a study of a set of PDGFR inhibitors, which are of interest owing to their ability to interfere with cell signal transduction mechanisms and are therefore of interest as anti-cancer drugs. As before, the study develops linear and nonlinear models and presents an interpretation of the linear model. In addition,a random forest model is developed to investigate the importance of the descriptors used in the study and in specific models.

The focus of interpretation techniques in the field of 2-D QSAR modeling has generally been restricted to the interpretation of linear regression models. In some cases, neural network models have been interpreted in a broad manner. Chapters 8 and 9 describe methods that were developed to interpret neural network models. Chapter 8 describes a simple method to provide a quantitative measure of descriptor importance in a neural network. The method is based on a sensitivity analysis of the model and is similar in nature to the descriptor importance measure that is available for random forest models. However, this method is similar to other approaches to the interpretation of neural networks since it only provides information about which descriptor is the most important for the model's predictive ability. It does not provide any insight into the nature of the correlation between the input to the network and the output from the network. A method to extract detailed information regarding the structure-property relationships encoded in the weights and biases of a trained neural network models is described in Chapter 9. This method is inspired by the PLS interpretation technique for linear models. The method simplifies the neural network and considers the hidden neurons of the network in a manner analogous to the latent variables of the PLS interpretation. In addition, plots analogous to the score plots of the PLS technique are presented. Combining the visual information provided by the score plots together with the analysis of the weights and biases, the method presented is able to provide a detailed view of the correlations between the input descriptors and the predicted property. The method thus provides for neural network models, what the PLS method has provided for linear regression models. Namely, an in-depth, compound-wise dissection of the structure-property trends encoded in the respective models

Finally, Chapter 10 summarizes the results of the studies presented in this work and concludes by highlighting the contributions of this thesis to the field of QSAR modeling.

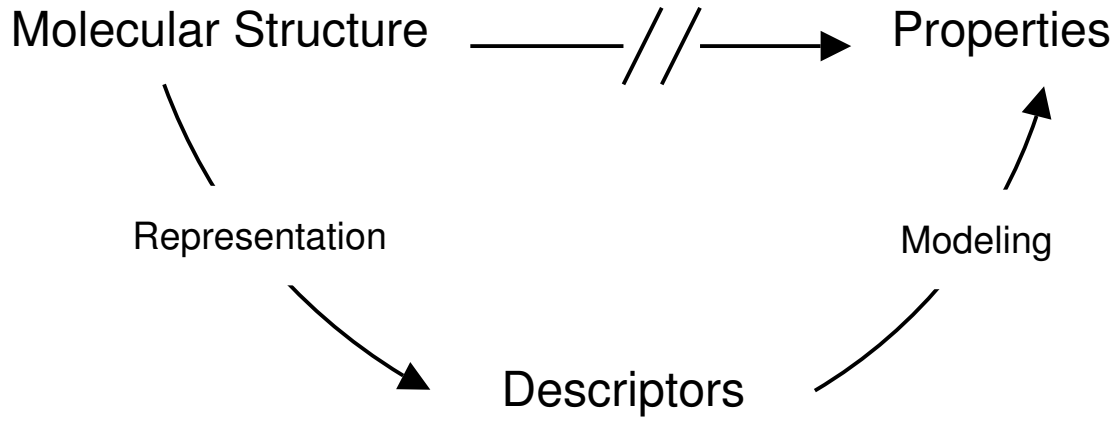

Fig. 1.1. A flowchart showing the steps involved in predicting molecular properties or activities from molecular structure

References

[1] Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. New Method for Direct Linear-Scaling Calculation of Electron Density of Proteins. *J. Phys. Chem. A* **2005,** *109,* 2–5.

[2] Dixon, S.; Merz, K. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996,** *104,* 6643–6649.

[3] Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions;* Wiley: New York, 1991.

[4] Clementi, E. *Computational Aspects for Large Chemical Systems;* Springer: New York, 1980.

[5] Hertzberg, R.; Pope, A. High-Throughput Screening: New Technology For the 21st Century. *Curr. Opin. Chem. Biol.* **2000,** *4,* 445–451.

[6] Hansch, C. A Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969,** *2,* 232–239.

[7] Hansch, C.; Fujita, T. $\epsilon-\sigma-\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964,** *86,* 1616–1626.

[8] Free, S. M. J.; Wilson, J. W. A Mathematical Contribution to Structure Activity Studies. *J. Med. Chem.* **1964,** *7,* 395–399.

[9] Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990,** *68,* 20–23.

[10] Hammett, L. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937,** *59,* 93–103.

[11] Breiman, L. Statiscal Modeling: Two Cultures. *Statistical Science* **2001,** *16,* 199–231.

[12] Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self Organizing Neural Network. *J. Am. Chem. Soc.* **1997,** *119,* 4033–4042.

[13] Espinosa, G.; Arenas, A.; Giralt, F. An Integrated SOM Fuzzy ARTMAP Neural System for the Evaluation of Toxicity. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 343–359.

[14] Schuurmann, G.; Aptula, A. O.; Kuhne, R.; Ebert, R. Stepwise Discrimination between Four Modes of Toxic Action of Phenols in the *Tetrahymena pyriformis* Assay. *Chem. Res. Tox.* **2003,** *16,* 974–987.

[15] Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 837–841.

[16] Goll, E.; Jurs, P. Prediction of the Normal Boiling Points of Organic Compounds From Molecular Structures with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 974–983.

[17] Rucker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 2070–2076.

[18] Ehresmann, B.; de Groot, M. J.; Alex, A.; Clark, T. New Molecular Descriptors Based on Local Properties at the Molecular Surface and a Boiling-Point Model Derived from Them. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 658–668.

[19] Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 837–841.

[20] Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 429-434.

[21] Mattioni, B. E.; Jurs, P. C. Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 232–240.

[22] Mosier, P.; Counterman, A.; Jurs, P.; Clemmer, D. Prediction of Peptide Ion Collision Cross Sections from Topological Molecular Structure and Amino Acid Parameters. *Anal. Chem.* **2002,** *74,* 1460–1370.

[23] Mosier, P. D.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Thiophene Derivatives from Molecular Structure. *Chem. Res. Toxicol.* **2003,** *16,* 721–732.

[24] McElroy, N. R.; Thompson, E. D.; Jurs, P. C. Classification of Diverse Organic Compounds That Induce Chromosomal Aberrations in Chinese Hamster Cells. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 2111–2119.

[25] Mattioni, B. E.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting To Generate a Model Ensemble. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 949–963.

[26] Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. Toward an Optimal Procedure for PC-ANN Model Building: Prediction of the Carcinogenic Activity of a Large Set of Drugs. *J. Chem. Inf. Model* **2005,** *45,* 190–199.

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

[28] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 671–678.

[29] Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005,** *48,* 312–320.

[30] Cramer III, R.; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Protiens. *J. Am. Chem. Soc.* **1988,** *110,* 5959–5967.

[31] Hopfinger, A.; Wang, S.; Tokarski, J.; Baiqiang, J.; Albuquerque, M.; Madhav, P.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J .Am. Chem. Soc.* **1997,** *119,* 10509–10524.

[32] Accelrys, "Cerius2", `http://www.accelrys.com/cerius2`,.

[33] Schrödinger, "Strike", `http://www.schrodinger.com/Products/strike.html`,.

[34] Liu, K.; Feng, J.; Young, S. PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. *J. Chem. Inf. Model* **2005,** *45,* 515–522.

[35] Todeschini, R.; Consonni, V.; Pavan, M. "DRAGON", 2005.

[36] Wegner, J. "JOELib", `http://joelib.sf.net`, 2005.

[37] Semichem, Inc., "Codessa", `http://www.semichem.com/codessa/index.shtml`,.

[38] Schrödinger, "QikProp", `http://www.schrodinger.com/Products/qikprop.html`,.

[39] Duffy, E.; Jorgensen, W. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000,** *122,* 2878–2888.

[40] Jorgensen, W.; Duffy, E. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000,** *10,* 1155–1158.

[41] R Development Core Team, "R: A Language and Environment For Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, 2004 ISBN 3-900051-07-0.

[42] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 493–500.

[43] Guha, R. Using the CDK as a Backend to R. *CDK News* **2005,** *2,* 2–6.

[44] Kohonen, T. *Self Organizing Maps;* volume 30 of *Springer Series in Information Sciences* Springer: Espoo, Finland, 1994.