Emerging Chemical Patterns for Virtual Screening and Knowledge Discovery

Dissertation zur
Erlangung des Doktorgrades (Dr. rer. nat.) der
Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Jens Horst Auer
aus Mayen

Bonn
November 2008

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Andreas Weber
Tag der Promotion: 21.01.2009
Erscheinungsjahr 2009

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn unter
`http://hss.ulb.uni-bonn.de/diss_online` elektronisch publiziert.

Abstract

The adaptation and evaluation of contemporary data mining methods to chemical and biological problems is one of major areas of research in chemoinformatics. Currently, large databases containing millions of small organic compounds are publicly available, and the need for advanced methods to analyze these data increases. Most methods used in chemoinformatics, e.g. quantitative structure-activity relationship (QSAR) modeling, decision trees and similarity searching, depend on the availability of large high-quality training data sets. However, in biological settings, the availability of these training sets is rather limited. This is especially true for early stages of drug discovery projects where typically only few active molecules are available. The ability of chemoinformatic methods to generalize from small training sets and accurately predict compound properties such as activity, ADME or toxicity is thus crucially important. Additionally, biological data such as results from high-throughput screening (HTS) campaigns is heavily biased towards inactive compounds. This bias presents an additional challenge for the adaptation of data mining methods and distinguishes chemoinformatics data from the standard benchmark scenarios in the data mining community.

Even if a highly accurate classifier would be available, it is still necessary to evaluate the predictions experimentally. These experiments are both costly and time-consuming and the need to optimize resources has driven the development of integrated screening protocols which try to minimize experimental efforts but still reaching high hit rates of active compounds. This integration, termed "sequential screening" benefits from the complementary nature of experimental HTS and computational virtual screening (VS) methods.

In this thesis, a current data mining framework based on class-specific nominal combinations of attributes (emerging patterns) is adapted to chemoinformatic problems and thoroughly evaluated. Combining emerging pattern methodology and the well-known notion of chemical descriptors, emerging chemical patterns (ECP) are defined as class-specific descriptor value range combinations. Each pattern can be thought of as a region in chemical space which is dominated by compounds from one class only. Based on chemical patterns, several experiments are presented which evaluate the performance of pattern-based knowledge mining, property prediction, compound ranking and sequential screening. ECP-based classification is implemented and evaluated on four activity classes for the prediction of compound potency levels. Compared to decision trees and a Bayesian

binary QSAR method, ECP-based classification produces high accuracy in positive and negative classes even on the basis of very small training set, a result especially valuable to chemoinformatic problems.

The simple nature of ECPs as class-specific descriptor value range combinations makes them easily interpretable. This is used to related ECPs to changes in the interaction network of protein-ligand complexes when the binding conformation is replaced by a computer-modeled conformation in a knowledge mining experiment. ECPs capture well-known energetic differences between binding and energy-minimized conformations and additionally present new insight into these differences on a class level analysis.

Finally, the integration of ECPs and HTS is evaluated in simulated lead-optimization and sequential screening experiments. The high accuracy on very small training sets is exploited to design an iterative simulated lead optimization experiment based on experimental evaluation of randomly selected small training sets. In each iteration, all compounds predicted to be weakly active are removed and the remaining compound set is enriched with highly potent compounds. On this basis, a simulated sequential screening experiment shows that ECP-based ranking recovers 19% of available compounds while reducing the "experimental" effort to 0.2%. These findings illustrate the potential of sequential screening protocols and hopefully increase the popularity of this relatively new methodology.

Acknowledgments

I am grateful to my supervisor Prof. Dr. Jürgen Bajorath for his guidance and his support during the work on this thesis. I would like to thank Prof. Dr. Andreas Weber for being so kind as to be my second referee. The preparation of this thesis would not have been possible without the advice and help of all my colleagues at the B-IT. They contributed to a wonderful working atmosphere and created many unforgetable moments. Special thanks goes to Ingo Vogt with whom I shared the office for the last three and a half years. He has been a vital partner for scientific and private discussions. I owe many thanks to Martin Vogt and Lisa Peltason for proof-reading. My scientific work has also benefited from collaborations with Eugen Lounkine, Ingo Vogt and Hany E. Ahmed.

Finally, I would like to thank my mother and my family for their support during the last years, especially my girlfriend Heike Woyk who also provided valuable hints for the language and style of my thesis.

List of abbreviations

ADME	Absorption, Distribution, Metabolism, and Excretion
BIN	Binary Qunatitaive Structure-Activity Relationship
BZR	Benzodiazepine Receptor
DHFR	Dihydrofolate Reductase
DT	Decision Tree
ECP	Emerging Chemical Pattern
EP	Emerging Pattern
GSK3	Glycogen Synthase Kinase-3 Inhibitors
HIVPROT	HIV Protease Inhibitors
HTS	High-Throughput Screening
JECP	Jumping Emerging Chemical Pattern
JEP	Jumping Emerging Pattern
IC_{50}	Half Maximal Inhibitory Concentration
MOE	Molecular Operating Environment
QSAR	Quantitative Structure-Activity Relationship
RMSD	Root Mean Square Deviation
TDZD	Thiadiazolidinone
TPSA	Total Polar Surface Area
VS	Virtual Screening

Contents

1 Introduction **1**

1.1 Related work 3

1.2 Research topics 4

1.3 Thesis Outline 7

2 Methodology **9**

2.1 Chemical Descriptors 9

2.1.1 Descriptor Set 1: 61 Type I and II Descriptors 11

2.1.2 Descriptor Set 2: Type III Descriptors 11

2.1.3 Descriptor Discretization 12

2.2 Mining Emerging Patterns 14

2.2.1 Mining Patterns Related to the Rule of Five 14

2.2.2 Formal Definition of Emerging Chemical Patterns 15

2.2.3 Mining Algorithms 18

2.2.4 Classification 19

2.3 Emerging Chemical Patterns 19

3 ECP Data Mining and Classification **23**

3.1 Data Mining for Conformational Differences 23

3.1.1 Data Set 25

3.1.2 Methodology 28

3.1.3 Differences Between Modeled and Bioactive Conformations 29

3.1.4 Class-based Pattern Mining 30

3.1.5 Structural Interpretation of ECPs 32

3.2 ECP Classification 43

3.2.1 Data Set 43

3.2.2 Evaluation 46

3.3 Discussion 53

4 Integration of Virtual and High-Throughput Screening **57**
4.1 Simulated Lead Optimization 59
4.2 Simulated Sequential Screening 61
4.2.1 Screening Calculations 64
4.2.2 Pattern Distribution and Composition 68
4.2.3 Screening Performance 69
4.2.4 Dynamic Interaction of VS and HTS 72
4.3 Discussion 77

5 Summary and Conclusion **79**

A Chemical Descriptors **83**
A.1 61 Uncorrelated Descriptors 83
A.2 3D Descriptors for Conformation Anaylsis 85

B Simulated Lead Optimization **87**

C Simulated Sequential Screening **95**

D Patterns for Conformation Analysis **97**

Bibliography **105**

List of Figures

2.1 Most Expressive Jumping Emerging Pattern 17
2.2 Classification Using JEPs 20

3.1 Exemplary Bioactive and Modeled Conformations 27
3.2 Adenosine Deaminase Binding Site with Bioactive Ligand Conformations 34
3.3 Adenosine Deaminase Protein-Ligand Interactions 35
3.4 Ribonuclease A Binding Site with Bioactive Ligand Conformation 37
3.5 Ribonuclease A Protein-Ligand Interactions 39
3.6 Trypsin Binding Site with Bioactive Conformation 41
3.7 Trypsin Protein-Ligand Interactions 41
3.8 Potency Distribution within Compound Activity Classes 44
3.9 Representative Structures for ECP Classification 45
3.10 Top ten Compounds of an individual GSK3 trial 49

4.1 Sequential Screening 58
4.2 Simulated Lead Optimization 60
4.3 Results for Simulated Lead Optimization 62
4.4 Structures of the 12 Most Active DHFR Inhibitors 65
4.5 Most Prominent Descriptors for Sequential Screening 69
4.6 Average Recovery Rates Over 100 Independent Screening Trials . . 71
4.7 Hit Rate for Three Simulated Sequential Screening Runs 72

B.1 Decision Tree Simulated Lead Optimization 87
B.2 Binary QSAR Simulated Lead Optimization 91

C.1 32 DHFR Inhibitors 95

List of Tables

2.1 Classification of Molecular Descriptors 10
2.2 Artificial Sample Data Set 14
2.3 Emerging Patterns for Table 2.2 15

3.1 Inhibitors, Modeled Conformers, and Discriminatory Patterns . . . 26
3.2 ECPs Discriminating Binding from Modeled Conformations 30
3.3 Discriminatory Patterns for Adenosine Deaminase Inhibitors . . . 36
3.4 Discriminatory Patterns for Ribonuclease A Inhibitors 38
3.5 Discriminatory Patterns for Trypsin Inhibitors 40
3.6 Compound Classes and Potency Levels 44
3.7 JECPs for Activity Class GSK3 47
3.8 Comparison of Classification Methods 52
3.9 Prediction Accuracy for Very Small Training Sets 53

4.1 Average ECP Results Over 100 Independent Screening Trials . . . 71
4.2 Top 10 ECP Trials for Selection Sets of 10 Database Compounds . 73
4.3 Top 10 ECP Trials for Selection Sets of 100 Database Compounds 74
4.4 Top 10 ECP Trials for Selection Sets of 500 Database Compounds 75

A.1 61 Uncorrelated Type I and II Descriptors 83
A.2 Type III Descriptor Set 85

D.1 Discriminatory Patterns for Acetylcholine Esterase Inhibitors . . . 97
D.2 Discriminatory Patterns for Carbonic anhydrase Inhibitors 98
D.3 Discriminatory Patterns for Carboxypeptidase Inhibitors 98
D.4 Discriminatory Patterns for Cyclin-dependent Kinase Inhibitors . . 99
D.5 Discriminatory Patterns for Elastase Inhibitors 99
D.6 Discriminatory Patterns for Endothiapepsin Inhibitors 100
D.7 Discriminatory Patterns for Factor Xa Inhibitors 100
D.8 Discriminatory Patterns for FK506 Binding Protein Inhibitors . . . 101
D.9 Discriminatory Patterns for HIV Protease Inhibitors 101
D.10 Discriminatory Patterns for Plasminogen Activator Inhibitors . . . 102

D.11 Discriminatory Patterns for Protein Tyrosine Phosphatase 1b Inhibitors 102
D.12 Discriminatory Patterns for Protocatechuate3,4-dioxygenase Inhibitors 103
D.13 Discriminatory Patterns for Thermolysin Inhibitors 103
D.14 Discriminatory Patterns for Thrombin Inhibitors 104
D.15 Discriminatory Patterns for Tyrosine Kinase Inhibitors 104

1 Introduction

Machine learning algorithms are widely used in chemoinformatics for the prediction of compound properties, including physicochemical properties like logP or biological activity. Adapted algorithms include partitioning methods (Rusinko et al., 1999; Stahura and Bajorath, 2003), clustering algorithms (Feher and Schmidt, 2003; Tamura et al., 2002), neural networks (Keserü et al., 2000; Sadowski, 2000), Bayesian models (Labute, 1999; Vogt et al., 2007), decision trees (Rusinko et al., 1999; Stockfisch, 2003) and kernel based methods (Geppert et al., 2008; Harper et al., 2001; Jorissen and Gilson, 2005).

The recent increase in available data for pharmaceutical or biological purposes drives the need for efficient data mining tools to explore, manipulate and analyze these data sets. Commercial data sources, e.g. the MDL drug data report database[1] or the Wombat database (Olah et al., 2004) and public efforts such as PubChem[2] or the ZINC database (Irwin and Shoichet, 2005) collect large amounts of data about organic compounds. PubChem now contains 40 million records about 19 million unique structures including information from more than 1000 biological assays, and the ZINC database contains ∼8 million compounds in total and ∼5 million drug-like molecules. Libraries like the MDDR or the Wombat library have also introduced a qualitative advancement in their collection of verified information about compounds from sources like patent information or journal publications. However, the fraction of experimentally tested compounds remains extremely small compared to the estimated number of possible organic molecules (Kirkpatrick and Ellis (2004) cite an estimate of more than 10^{60} potential small organic molecules) even though high-throughput screening (HTS) technology has made dramatic advances in the last years and enabled the automatic screening of millions of compounds for desired properties in a short time. Faced with this large amount of data, the application of contemporary data mining algorithms is becoming one of the fundamental topics in chemoinformatic research. Traditionally, chemoinformatics has borrowed algorithms such as clustering or partitioning

[1] MDL Drug Data Report from MDL Information Systems, San Leandro, CA, USA. Available at `http://www.mdl.com/products/knowledge/drug_data_report/`

[2] `http://pubchem.ncbi.nlm.nih.gov`

from the data mining community and invented new approaches such as similarity searching based on molecular fingerprints. Many of these algorithms are used to select small subsets from large compound libraries which have a high probability to show desired properties, e.g. activity against a target.

This thesis presents the adaptation of a recently developed pattern mining framework to chemoinformatic problems. It adapts the emerging pattern (EP) mining algorithms to molecular data by using discretized molecular descriptors as attributes. EP mining finds class-specific combinations of attributes with high frequency in a home class but low occurence in a background data set. In this way, class-specific knowledge about molecules will be encoded as combinations of descriptor value ranges. Each combination can be thought of as a class-specific chemical descriptor space dominated by compounds from one class only. Applications of the EP mining framework include molecular classification, ranking and knowledge mining on conformational data sets.

Besides introducing a new data mining method into the chemoinformatics research area, special attention is paid to possible efficient integration of computational and experimental methods. Both research areas, computational chemistry or chemoinformatics and the experimentally HTS technology, have made dramatic advances in the last years. It is now possible to automatically screen millions of compounds in reasonable time. However, an undirected brute-force search is very likely to waste valuable resources by inspecting many unpromising compounds. Computational methods are in a sense orthogonal to experimental methods. With computational means, it is possible to screen even larger compound libraries in short time and to access compounds which are not available for experimental testing. Traditionally, both methods are used exclusively (experimental testing is of course also used to validate virtual screening (VS) results). It has been proposed that an integration of computational and experimental efforts leads to an improvement in screening methodology by reducing the experimental effort while achieving significant hit rates of active compounds (Bajorath, 2002). Although proposed several years ago, the integrated "sequential screening" paradigm is far from being implemented in standard drug discovery projects and lacks evaluation studies. The second part of this thesis is concerned with two "experiments" exploring the potential of sequential screening.

1.1 Related work

The application of data mining algorithms to search for hidden knowledge in the form of combinations of features has a long history in chemoinformatics and related fields such as bioinformatics. Given the long tradition of string algorithms in bioinformatics, it is not surprising that many prominent examples are actually string mining methods. Frequent substring mining approaches are used to find repetitive structures such as common subsequences in proteins. These algorithms can be efficiently implemented using suffix trees (Gusfield, 1997). Recently, Fischer et al. (2006) have introduced linear-time algorithms to find emerging substrings, i.e. substrings frequent in one set but rare in another set of strings.

In chemoinformatics, pattern mining is used in a more general sense than the common data mining concept of frequent pattern mining. Most applications of pattern mining algorithms focus on the discovery of frequent or common substructures of sets of molecules. Nicolaou and Pattichis (2006) review the most important algorithms for molecular substructure mining. All these algorithms compute frequent substructures for a set of molecules. However, they do not take class information, e.g. activity, into account and thus cannot be used to compute common substructures which are specific for sets of active compounds in contrast to a background database. Ting and Bailey (2006) present an application of a hypergraph based algorithm which is also used to compute emerging patterns to compute class-specific emerging subgraph patterns. Common to all graph-mining algorithms is that they are computationally expensive and only applicable for small data sets.

Related to common substructure mining, fragment based approaches are sometimes used to find common fragments in sets of compounds. In contrast to substructure mining, fragment-based methods typically use a set of predefined fragments to screen a database of molecules. A simple example of a fragment-based approach is the MACCS structural keys fingerprint which assigns to each molecule a bit vector accounting for presence or absence of 166 structural fragments. Substructure mining and most fragment-based approaches do not consider combinations of substructures, but are only concerned with the identification of frequent single structural entities. Fragment-based methods are sometimes used to find combinations of class-specific fragments, but they usually use only small combinations of fragments (Lounkine et al., 2008). Sometimes, statistical sampling (Lameijer et al., 2006) is used to cut down computation time and escape the problem of exponential growth of possible fragment combinations. Both methods

further restrict the number of fragments by breaking molecules into chemically reasonable parts of ring systems, linkers and side chains. Lameijer et al. (2006) then count co-occurrence of fragments in a stochastic experiment without including a background database. The experiments described by Lounkine et al. (2008) are closely related to emerging pattern mining since they show how formal concept analysis can be used to interactively construct queries on sets of fragments computed from compounds with associated selectivity data. Initial evaluations have shown that typical fragment combinations consist of only a handful of fragments and thus the search is restricted to use only four fragments at most. It is shown that these combinations of fragments are highly specific for activity classes and thus using a data mining approach to compute class-specific descriptor value combinations seems a promising and reasonable idea.

In a way, similarity searching algorithms like DynaMAD (Eckert and Bajorath, 2006) and CA-DynaMAD (Vogt and Bajorath, 2008) are similar to emerging pattern mining. CA-DynaMAD heuristically computes class-specific combinations of descriptor value ranges which are then used for assessing the probability of test compounds for being active against the same target as the molecules in the training set. Both algorithms first rank class-specific descriptor value ranges[3] based on the number of (inactive) database compounds matched by this value range. For classification, they iteratively eliminate database compounds which do not match the top-ranking descriptor value ranges, traversing the descriptors in descending order. The implementation differs, but both methods result in a small set of potentially active compounds and a corresponding descriptor value range set which includes all descriptors used in the iterations. This descriptor set can be interpreted as a class-specific chemical reference space. However, these algorithms do not use data mining methods to explore possible combinations of descriptors, but heuristically select the next descriptors on the basis of their score.

1.2 Research topics

This thesis investigates how current pattern mining methods can be used in drug discovery research. More specifically, it is concerned with the application of pattern mining technologies computing class-specific combinations of discrete attributes to chemical knowledge mining, classification and ranking. The adaptation of data mining algorithms is also an interesting experiment from a computer

[3]The min. and max. values for a descriptor in the active training set are used as the class-specific descriptor value range (min, max).

science point of view. Data sets such as results from biological screening campaigns are different from the standard benchmark sets used in the data mining community. First, most screening data sets contain intrinsic errors. Experimental screens are often influenced by side effects in an unpredictable way. A data mining methodology must be able to handle error-prone noisy data. Another challenge for data mining algorithms results from the knowledge distribution of active and inactive compounds. Usually, the number of active compounds is much smaller than the number of inactives with differences of several orders of magnitude. This is especially true for the early stages of drug design projects where typically only a handful of known active compounds are available. A data mining method must be able to generalize well from such a small number of active template compounds, but still be specific enough to minimize the false-positive rate. Even the simplest classification possible, which would predict all compounds to be inactive regardless of their properties, would yield a high accuracy! A useful chemoinformatic method must be aware of this intrinsic bias and predict activity (or other properties) with high accuracy for both active and inactive cases. In addition to these raw class distribution problems, chemoinformatics data sets often have an additional structural bias. Most data sets are constructed during compound optimization projects by chemical modification and experimental testing. Thus, they contain series of analogue structures which are very similar in their constitution and features. This places an additional bias in the data set towards compounds with similar structure and makes generalization to structurally different compounds hard to accomplish.

The adaptation of the EP framework poses three questions which are central for this thesis:

Question 1: How can pattern mining-based knowledge be rationalized at the molecular level?

One of the benefits of pattern-based knowledge mining is the easy interpretation of the computed knowledge in form of class-specific combinations of attributes. One of the earliest applications of pattern-mining has been the analysis of customer shopping baskets (Agrawal et al., 1993) and the resulting patterns could be rationalized by typical customer behavior, e.g. people tend to buy milk and bread in combination. In chemoinformatics and drug discovery, researchers are mostly concerned with structural features of ligands and their interaction with their target protein. Successful pattern mining should lead to patterns which can

well be rationalized chemically by relating the descriptors included in the pattern to changes in the interactions upon formation of the protein-ligand complex.

Question 2: How does a pattern-based classifier perform on biological data?

This question is directly related to the adaptation of the pattern mining framework as a chemoinformatic tool. Chemical compounds must be represented in a way that can be used by pattern mining algorithms, meaning transformation of chemical information into nominal attributes.

In chemoinformatics, the focus of computational methods is often on early discovery of active compounds instead of maximizing recovery rates. This is often done by ranking large databases (compound libraries) and testing only a small fraction of the top-ranked compounds. The standard method here is similarity searching, commonly done by using bit-string representations of compounds known as fingerprints (Auer and Bajorath, 2008). However, fingerprints are not class-specific and the creation of class-specific similarity search methods is one of the major research goals. The adaptation of a pattern-based data mining method as a ranking tool would be a major contribution to similarity searching methods.

Question 3: Is it possible to integrate computational and experimental methods in an efficient way such that both methods benefit from the integration?

Virtual Screening and experimental screening are complementary methods having specific strengths and weaknesses (Bajorath, 2002). A tight integration of both methods into a single screening protocol is expected to dramatically reduce the experimental effort while keeping high and early recovery (Bleicher et al., 2003). How this integration can be established is still a question of research.

The proposed integration is relatively new and only few experiments have been reported yet which analyze the results of such a method. Most of these experiments solely analyze the performance in terms of active compound recovery but do not look at the dynamic interplay of computational and experimental screening. The availability of many HTS data sets released by PubChem makes it possible to simplify integration experiments to pure computational experiments by replacing the HTS screening phase with a simulated step based on real HTS data. In this way, statistically valid experiments can be designed which can then be thoroughly evaluated.

1.3 Thesis Outline

Chapter 2 introduces the fundamental concepts of ECP mining. EPs have been introduced recently in the data mining field by Dong et al. (1999a) and are adapted to chemoinformatics by representing molecules using molecular descriptors. Since ECP mining algorithms only work on discrete data (opposite to the continuous nature of most chemical descriptors), the chapter also describes standard discretization techniques to transform continuous descriptors into sets of discrete, non-overlapping value-ranges. The chapter includes an introduction to EP mining using the well-known example of Lipinski's Rule of Five (Lipinski et al., 2001) to characterize drug-like compounds and a formal definition of these patterns as combinations of class-specific descriptor value ranges. It also gives a short description of the algorithms used in the experiments.

Chapter 3 evaluates the emerging chemical pattern approach as a tool for chemoinformatic knowledge mining and property prediction. It shows how ECPs are used to analyze and rationalize differences between bioactive and modeled ligand conformations, which answers the first research question. Compounds from 18 target classes for which crystallographic data of the binding conformations are available are subjected to a stochastic energy minimization. Binding and computed conformations are joined into a database which is then mined for differentiating ECPs. The computed patterns are rationalized on a molecular basis by analyzing differences in the interactions of the protein-ligand complexes found in the crystallographic data and a hypothetical protein-ligand complex built by superposing the minimized conformation on the crystallographic binding conformation. Patterns described specific differences in the conformations leading to different interaction patterns.

The second research question is addressed by exploring the possibilities of molecular classification of active vs. inactive compounds using ECP-based classifiers. ECP-based classification using 61 descriptors derived from 1D or 2D information performs equally well as established binary classification methods. It is further pointed out that the exploration of the exponential space of possible descriptor value range combinations leads to classifiers which allow highly accurate classification when the training sets are as small as six compounds (three compounds per class).

Chapter 4 focuses on the possible integration of HTS and VS methods and shows that ECPs can be used to perform a class-directed similarity search which selects

compounds with similar biological properties as given by a training set, thus addressing the last research question. Two examples of iterative computational and experimental screening protocols are described. The first simulates a possible iterative lead-optimization experiment which optimizes a set of compounds with respect to potency. Sequential screening is then shown to be an efficient methodology integrating computational and experimental screening methods.

Chapter 5 summarizes and discusses the results. The findings in the previous chapters are discussed with respect to the three research questions stated above. Important aspects of each experiment are related to each of the research questions and the implications are discussed. Finally, the thesis is summarized in short.

2 Methodology

Chemoinformatics has always been concerned with the application of new data mining algorithms to analyse chemical and biological data and predict properties of molecules. In this thesis, a recent pattern-based approach is adapted to chemoinformatics by using molecular data in the form of chemical descriptors. Details of the data mining method and the adaptation to chemistry are described in the remaining parts of this chapter. It starts with the introduction of chemical descriptors as the basis for molecular representation. After introducing and formally defining pattern-mining, emerging chemical patterns (ECP) are defined as class-specific combinations of descriptor value ranges.

2.1 Chemical Descriptors

Molecules are represented as entries in databases, associated with numeric attributes which encode molecular properties. These properties are generally called molecular or structure descriptors. Terfloth (2003) defines structure descriptors as "a mathematical representation of a molecule resulting from a procedure transforming the structural information encoded within a symbolic representation of a molecule". Simple examples of such a mathematical representation are molecular weight or formal charge which can be easily computed from the chemical formula of a compound. These descriptors are easy to compute, but might not provide useful information about the compounds. Many researchers have defined chemical descriptors for different tasks with different levels of information and complexity. The Dragon software[1] currently implements more than 3200 different descriptors. The Handbook of Molecular Descriptors (Todeschini and Consonni, 2000) lists details and definitions for these descriptors. Table 2.1 divides these descriptors into five categories. Categories I-III relate to the molecular representation needed to compute the descriptor value. Category IV includes descriptors which make use of experimental data, e.g. measured binding affinities and category V includes complex descriptors which are derived from combinations of other descriptors. Given the limited amount of available experimental data, it is obvious that descriptors

[1] TALETE srl, `http://www.talete.mi.it/dragon.htm`

Table 2.1: Classification of molecular descriptors. Molecular descriptors are partitioned into five classes based on the molecular information needed for their computation (Auer and Bajorath, 2008).

Type	Derived from	Examples
I	Global (bulk) molecular properties	(estimated) logP(o/w), atom counts
II	2D structure (molecular graph)	structural keys, connectivity indices
III	3D structure	surface properties, radius of gyration
IV	Biological properties	affinity fingerprints
V	Combination of descriptors	BCUT

of category IV are usually not available although using such data is expected to improve the pure computational methods.

Since the biological activity of compounds is strongly influenced by 3D properties such as complementary molecular shape of ligands and binding pockets, it would be expected that 3D descriptors are generally more sensitive to biological activity than simpler representations. However, this is not necessarily the case. Numerous studies indicate that 2D information is often sufficient to produce accurate results (Brown and Martin, 1996; Xue and Bajorath, 2002). In addition, 2D representations are typically much less prone to errors than calculated 3D representations. This is due to the lack of experimentally determined 3D conformations of bound ligands and the insufficiency of computational methods to model these conformations. However, when 3D information is available, type-III descriptors provide much information about compounds and are certainly high-quality features worth to be used.

The different experiments presented in this thesis use two different sets of descriptors. First, a knowledge-mining experiment is described which uses type III descriptors to investigate differences between different conformations of compounds, namely conformations in binding mode and computationally generated conformations. A second set of experiments is concerned with classification and ranking of compounds based on their activity. These experiments utilize only 2D information in the form of type I and II descriptors. Both descriptor sets are described in the next two sections. The complete lists of descriptors, including a short description of each descriptor and reference information is listed in tables A.1 and A.2 in appendix A.

2.1.1 Descriptor Set 1: 61 Type I and II Descriptors

The first set of descriptors consists of 61 uncorrelated type I and II descriptors selected with minimized pairwise correlation and maximized information content in a database of 1.34 million compounds (Xue et al., 2003). This set contains eleven type I descriptors: logP(o/w), abbreviated as logP, total polar surface area TPSA and 9 atom counts, e.g. the number of fluorine (a_nF) or chlorine atoms (a_nCl). Most of the remaining 50 descriptors are type II descriptors computed from the connectivity information encoded in the molecular graph. This set contains four bond counts, four connectivity indices and adjacency matrix descriptors, eight pharmacophore descriptors, 17 subdivided surface area descriptors and 17 partial charge descriptors. The minimization of pairwise correlation effects has also positive effects on the run time of the computations because the complexity of data mining algorithms depends on the dimensionality of the data. The list of all 61 descriptors is given in appendix table A.1.

2.1.2 Descriptor Set 2: Type III Descriptors

The second descriptor set consists of a total of 67 conformation-dependent (type III) descriptors available in the molecular operating environment (MOE) (Chemical Computing Group Inc., version 2007.09). It includes a variety of type III descriptors belonging to five categories: energy, shape, and charge distribution descriptors, molecular surface properties, and volume-dependent descriptors. A subset of these descriptors including, for example, heat of formation, ionization potential, and highest-occupied molecular orbital (HOMO) or lowest unoccupied molecular orbital (LUMO) energies was calculated through MOE's interface with MOPAC (Stewart, 1990) and the semi-empirical AM1 (Dewar et al., 1985), PM3 (Dewar and Thiel, 1977) and MNDO (Stewart, 1989) methods. Partial charge and potential energy descriptors were calculated with a MOE-internal modified version of the Merck molecular force field (MMFF94) (Halgren, 1996a,b,c,d; Halgren and Nachbar, 1996). The MOE implementation treats conjugated nitrogens as planar atoms instead of tetrahedral atoms. The radius of gyration (rgyr) serves as a measure of compactness of a molecule and the principal moments of inertia (pmi, pmiX, pmiY, pmiZ) account for mass distribution. Surface descriptors include, for example, the total solvent-accessible surface area (ASA) and positively or negatively charged (ASA+ and ASA-), hydrophobic (ASA_H) or polar (ASA_P) surface areas. In addition, descriptors depending on the van der Waals volume (vol) are also included, such as molecular density (dens) that is calculated

by dividing molecular weight by vol. The calculation of some of these type III descriptors, such as MOPAC descriptors, is computationally expensive. However, all descriptors need to be calculated only once for each data set. The list of all 67 3D descriptors is given in table A.2.

All descriptor calculations are done using MOE. Prior to descriptor calculation, the compounds are "washed" with MOE removing solvents and salts and assigning reasonable protonation states and partial charges. The default settings of MOE are applied.

2.1.3 Descriptor Discretization

Traditionally, pattern mining algorithms use nominal attributes instead of continuous value attributes. Most molecular descriptors are, however, continuous in nature. One way to handle continuous attributes with pattern mining algorithms is to transform the value range of an attribute into bins and use these bins as new discrete attributes. The way how these bins are derived depends on the discretization algorithm. These algorithms can be divided into supervised and unsupervised algorithms, where algorithms from the first category use the training data to take knowledge about class and value distribution of the continuous data into account. Unsupervised algorithms divide the value range of continuous attributes independent of the class and value distribution of the training data set into bins (Dougherty et al., 1995).

Unsupervised Discretization algorithms include simple *equal-interval* binning where a descriptor's value range is divided into a fixed set of intervals all spanning the same value range. However, equal-interval binning often leads to uneven distributions of the data into the bins, because attribute are often irregularly distributed in their value range (Witten and Frank, 2005). An equal-interval discretization produces many sparsely populated (or even empty) bins and a small number of heavily populated bins, weakening the attribute's information content. A better unsupervised discretization technique is to derive the bins in a way that all bins contain the same number of attributes, irrespective of their class value. This *equal-frequency* binning produces a fixed number of equally populated bins. However, since it is still an unsupervised method which does not use the class distribution of the training data available, the resulting binning scheme can also lead to a loss of information, e.g. when the equal-frequency constraint forces a split point between two bins in a way that the first bin is pure, i.e. containing only compounds from one class, but the second bin starts with a small number

of instances from the one class and then contains only instances of another class. In this case, giving up the equal-frequency constraint to produce two pure bins would be favorable.

Supervised Discretization techniques make use of the class distribution of the training instances, e.g. by using statistical error measures such as the χ^2 test (Kerber, 1992; Kohavi and Sahami, 1996) or information theory (Fayyad and Irani, 1992). The information theory based discretization from Fayyad and Irani (1992) has shown good results in several applications of EP mining (Li and Wong, 2002a,b; Ramamohanarao and Bailey, 2003) and is subsequently used as the standard discretization algorithm.

The information theory based discretization algorithm utilizes information entropy as a measure of pureness for a possible splitting point. In this way, it is similar to the attribute splitting criterion used in decision tree learning (Quinlan, 1993) and extends this idea to a recursive algorithm to successively split a continuous attribute into bins until a stopping criterion is reached. The stopping criterion is based on the minimum description length (MDL) principle and ensures that the number of bins stays reasonably small. The method starts by examination the information entropy of all possible splitting points[2]. The class information entropy induced by a splitting point T which split an attribute A and training data S into two intervals S_1 and S_2 is defined as

$$\mathrm{E}(A,T,S) = \frac{|S_1|}{|S|}\mathrm{Ent}(S_1) + \frac{|S_2|}{|S|}\mathrm{Ent}(S_2),$$

with $\mathrm{Ent}(S) = -\Sigma_{i=1}^{|C|} P(C_i,S)\log(P(C_i,S))$ being defined as the entropy of a dataset S with n classes $C = \{C_1,\ldots,C_n\}$. $P(C_i,S)$ gives the proportion of instances in S with class C_i. After computing the information entropy of each splitting point, the one with the minimum class information entropy is selected and the algorithm recursively discretizes both sides of the induced partitioning.

An additional benefit of the information entropy based discretization is that the MDL principle stops the recursion in the first iteration if no good splitting point can be found and thus returns a binning with one bin spanning the whole value range of $(-\infty,\infty)$. Such attributes provide no information for classification and can thus be removed prior to pattern mining, reducing the dimensionality of the pattern space and computation time.

[2] A possible splitting point is located only in the middle between two subsequent values from the training set.

Table 2.2: A sample data set with ten compounds and four calculated descriptors: molecular weight (MW), logP(octanol/water) (logP), the number of hydrogen bond acceptors (HB-acc) and donors (HB-don). Each descriptor is discretized into two intervals.

	Class	MW		logP		HB-acc		HB-don	
		$[0, 500)$	$[500, \infty)$	$(-\infty, 5]$	$(5, \infty)$	$[0, 10]$	$(10, \infty)$	$[0, 5]$	$(5, \infty)$
1	active	×		×		×			×
2	active	×		×		×			×
3	active	×			×		×	×	
4	active		×	×		×		×	
5	inactive	×		×			×	×	
6	inactive		×	×		×			×
7	inactive	×			×	×			×
8	inactive	×		×		×			×
9	inactive	×			×	×			×
10	inactive	×		×			×		×

2.2 Mining Emerging Patterns

The following sections introduce the concepts and notions of EP mining using a toy data set and then formally define the relevant concepts.

2.2.1 Mining Patterns Related to the Rule of Five

The formal concept of emerging patterns will be introduced in the following using the well known example of Lipinski's rule of five (Lipinski et al., 2001). For this purpose, consider the model data set reported in table 2.2 that consists of 10 compounds and four descriptors: molecular weight (MW), logP(octanol/water), the number of hydrogen bond acceptors (HB-acc), and the number of hydrogen bond donors (HB-don). The descriptors have been divided into two intervals. Four of the hypothetical compounds are active and the remaining six are inactive.

The example data shows that a single descriptor interval is not sufficient to describe the difference between active and inactive compounds. However, using combinations of some of the descriptors might lead to a discriminatory descriptor value range selection. In this example, it would be possible to test all combinations by hand, but for large numbers of descriptors and compounds, due to the exponential growth in possible combinations, a computational method that searches the space of all descriptor value range combinations for class-specific patterns is employed. A pattern is simply any combination of descriptor value ranges, e.g. the

Table 2.3: Emerging chemical patterns for active compounds computed from the sample data set in Table 2.2

Growth	Support_{active}	$\text{Support}_{inactive}$	Pattern
3	0.5	0.17	{HB-don:[0.00,5.00]}
3	0.5	0.17	{MW:[0.00,500.00), logP:$(-\infty$,5.00], HB-acc:[0.00,10.00]}

single descriptor value range {MW:[0.00,500.00]} or a more complex combination of descriptors such as, for example, molecular weight with the number of hydrogen bond donors: {MW:[0.00,500.00], HB-acc:[0.00,10.00]}. Patterns that are specific for the active compounds should only rarely occur in inactive ones. How often a pattern p occurs in a data set D is measured by the support $\text{supp}_D(p)$ as the percentage of compounds that match the pattern. The support of the pattern {HB-don:[0.00,5.00]} in the sample data is $\frac{2}{4}$ for the active class and $\frac{1}{6}$ in the inactive class. A class-specific pattern can be defined as a pattern where the fraction of both supports, called growth rate, is larger than a defined threshold. The pattern p has a growth rate $\text{growth}_{active,inactive}(p) = 3$. A class-specific pattern is called emerging pattern (Dong et al., 1999a). From all possible EPs, those are especially interesting which are of smallest cardinality. The previously shown pattern p could be changed by adding an additional descriptor value range, thereby altering the supports in both classes and thus the growth rate. However, adding an additional descriptor value range also makes the pattern more selective and probably reduces the number of matched compounds for the new pattern. Each compound that is matched by the modified pattern is also matched by the original pattern, and thus p subsumes the more specific pattern. In other words, the pattern p is more general than any pattern that can be constructed by adding one or more additional descriptor value ranges. Table 2.3 shows the EPs specific for the active compounds. Given these patterns, one would conclude that active compounds have a molecular weight lower than 500 Da, a logP(octanol/water) below five, not more than 10 hydrogen bond acceptors, and not more than five hydrogen bond donors.

2.2.2 Formal Definition of Emerging Chemical Patterns

The concept of emerging patterns will be defined formally in the following sections. Pattern mining is concerned with data sets of sets instances, each being a subset

of a fixed set I of items $I = \{i_1, \ldots, i_n\}$. A pattern (or item set) is simply a combination of these items:

Definition 2.2.1 *A pattern P is a combination of items found in a data set: $P \subseteq I$.*

A pattern matches an instance d of a data set S if each item of the pattern is also part of that instance, i.e. $P \subseteq d$.

Given two data sets S_1 and S_2, the *support* and *growth rate* of a pattern P can be defined.

Definition 2.2.2 *The support* supp *of a pattern P in a data set S is the percentage of instances matched by P:*

$$\text{supp}_S(P) = \frac{|\{d|d \in S \wedge P \subseteq d\}|}{|S|}$$

For two data sets S_1 and S_2, the growth rate growth *of a pattern P is defined as the fraction of support in both data sets:*

$$\text{growth}_{S_1,S_2}(P) = \begin{cases} \frac{\text{supp}_{S_1}(P)}{\text{supp}_{S_2}(P)}, & \text{if } \text{supp}_{S_2}(P) > 0 \\ \infty, & \text{else} \end{cases}$$

The above definition does not define a growth rate for patterns with zero support in the home class S_1. However, these patterns are not of interest for data mining and thus attributes with zero support in S_1 are removed before pattern mining. A pattern which is rare in one class, e.g. S_2, but common to its home class S_1 is of particular interest from a data and knowledge mining perspective. These patterns are called emerging patterns. Patterns with infinite growth rate (patterns which occur only in one class but not in the other) are called jumping emerging patterns (Dong et al., 1999a):

Definition 2.2.3 *Given a threshold $t \geq 0$, a pattern P with a growth rate* $\text{growth}_{S_1,S_2}(P) \geq t$ *is a* t-emerging pattern. *A ∞-emerging pattern is called* jumping emerging pattern (JEP).

The set of all t-emerging pattern for a data set grows exponentially with the number of possible items. To reduce the number of patterns for a knowledge mining experiment, only patterns which balance specificity and generality are retained and the remaining patterns are pruned. A pattern maximizes generality if it matches as many instances in its home class data set as possible. This is

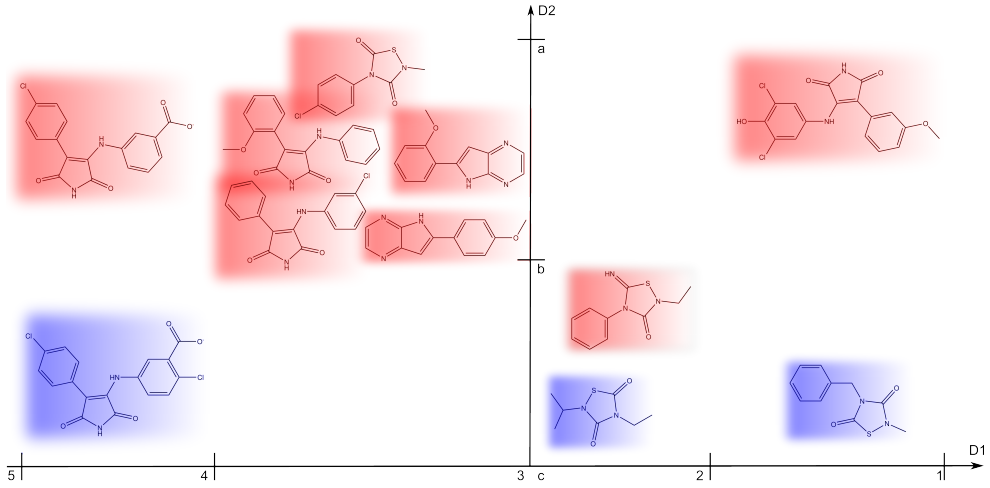


Figure 2.1: Most expressive jumping emerging pattern. A small data set of 11 compounds from two classes (red and blue) is projected into a space of two attributes D1 and D2. Each attribute is discretized into two intervals. $\{\text{D1:}(1,2), \text{D2:}(b,c)\}$ and $\{\text{D2:}(a,b)\}$ are most expressive jumping emerging patterns.

done by selecting patterns with minimum cardinality. Increasing the cardinality of a t-emerging pattern P by adding an additional item increases the number of restrictions an instance must fulfill to be matched and thus is likely to decrease the number of matched instances in the home class of P. Specificity is maximized by t-emerging patterns with large growth rate. Of special interest in this regard are jumping emerging patterns because they provide the sharpest distinction between two classes of data. From the set of all jumping emerging patterns, most expressive jumping emerging patterns, as defined by Li et al. (2001) are especially interesting:

Definition 2.2.4 *A jumping emerging pattern P, computed from two data sets S_1 and S_2 is most expressive if and only if*

1. *Each proper subset of P is no longer a jumping emerging pattern:*

$$\forall p \subset P.\text{growth}_{S_1,S_2}(p) \neq \infty$$

2. *Each proper superset of P has smaller support in P's home class S_1:*

$$\forall p \supset P.\text{supp}_{S_1}(P) > \text{supp}_{S_1}(p)$$

Figure 2.1 illustrates the concept of the most expressive emerging pattern using a small set of eleven compounds labeled by color with two classes (red and blue). The pattern $\{D1\text{:}(1,2), D2\text{:}(b,c)\}$ is a JEP of minimum cardinality. Both subsets

$\{D1{:}(1,2)\}$ and $\{D2{:}(b,c)\}$ match compounds from both classes and are thus not a JEP anymore. Since it already uses all possible descriptors in the data set, it also fulfills the second requirement for most expressive JEPs. Another example is the pattern $\{D2{:}(a,b)\}$. This pattern is a JEP for the class of red compounds. Having a cardinality of 1, it easily fulfills the first requirement of most expressive JEPs. If the pattern is extended by a adding second attribute, the support decreases. As an example, consider the extended pattern $\{D2{:}(a,b), D1{:}(3,4))\}$, where the support decreases from $\frac{7}{8}$ to $\frac{5}{8}$. Thus, any larger pattern built from $\{D2{:}(a,b)\}$ has lower support in the red class and $\{D2{:}(a,b)\}$ is a most expressive JEP.

2.2.3 Mining Algorithms

The development of algorithms for mining emerging patterns is a field of active research and has produced a number of algorithms with different properties. Many algorithms use so-called borders to store large sets of patterns in a compact way and use the border-diff algorithm (Dong et al., 1999b) to compute the set of all jumping emerging patterns of two data sets. The efficiency of the border-diff operation depends on the dimensionality of the positive and negative examples as well as on the number of negative examples. Various approaches exist which try to optimize the use of border-diff to improve the efficiency of emerging pattern mining, including tree-based approaches (Bailey et al., 2002) and a hypergraph-based approach (Bailey et al., 2003). These algorithms all work in a divide-and-conquer approach and try optimize the usage of the border-diff operation. Other approaches for mining emerging patterns include tree-based approaches which use search trees to explore the space of possible patterns (Zhang et al., 2000b), sometimes using additional pruning techniques to remove uninteresting patterns early in the search process (Fan and Ramamohanarao, 2003). A recent implementation uses zero-suppressed binary decision diagrams and is shown to be one of the fastest algorithms available (Loekito and Bailey, 2006).

Although a number of fast algorithms are available, the problem of mining emerging patterns remains computationally hard. It is shown to be MAX-SNP–hard (Wang et al., 2005a). The MAX-SNP complexity class contains graph-theoretical prolems which can be described by existential second-order logic. It has been shown that fixed-ratio approximation schemes for these problems are also NP-hard (Papadimitriou and Yannakakis, 1988).

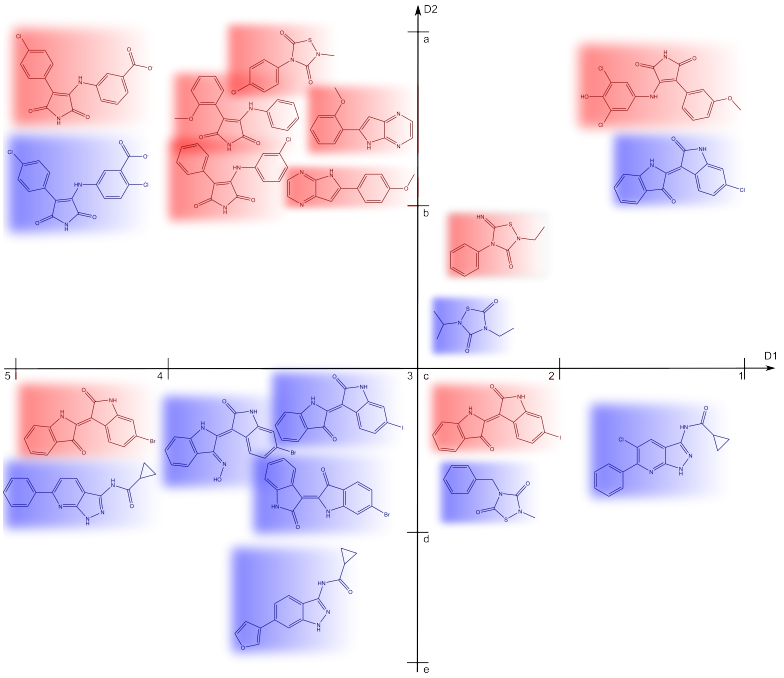
2.2.4 Classification

Emerging Patterns can be used to construct high-performance classifiers. Ramamohanarao and Bailey (2003) provide an overview of most of the algorithms. For classification, a set of emerging patterns is first computed for each class. If the data set contains more than two classes, a round robin approach (Fürnkranz, 2002) can be used. Round-robin computes patterns for each class by using all other classes as a merged background class. A test compound is then classified based on which set of patterns it resembles more. The actual implementation of the decision function differs for each algorithm, but most algorithms aggregate the support of all patterns matching the test instance for each class and then assign the class with largest aggregated support (Li et al., 2001, 2004; Zhang et al., 2000a).

A general outline of the classification procedure is shown in figure 2.2. First the training data set, which consists of 20 compounds divided into two classes by color code and two attributes, is mined for JEPs. After an optional pruning step, some of these JEPs are stored for classification of unknown test compounds. Two test compounds are shown. The accumulated support is computed for each test compound by checking which JEP matches the test compound. In this example, the first test compound matches two JEPs from the red class and only one JEP from the blue class. The second test compound matches one JEP from each class. Based on the sum of support for these patterns, the first test compound is predicted to belong to the red class, while the second test compound is predicted to belong to the blue class.

2.3 Emerging Chemical Patterns

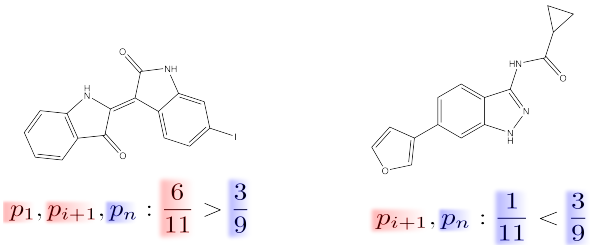
Emerging pattern mining provides a powerful technique to extract important knowledge from data sets and use this knowledge to analyze data or to predict properties of new, unknown test data. It is straightforward to combine emerging pattern mining with molecular data represented as molecular descriptors. Since pattern mining only works for discrete attributes, the continuous descriptors first have to be transformed into discrete attributes by means of discretization algorithms. The supervised, information theory based discretization technique has been shown to yield good results in several applications of EP based classification studies (Li et al., 2004; Li and Wong, 2002a; Ramamohanarao and Bailey, 2003) and is the preferred method used in the experiments described in the following chapters. However, the unsupervised discretization methods are also evaluated



(a) 20 compounds projected into a two-dimensional descriptor space. Red and blue indicate the compound class.

Pattern	Support
p_1={D1:[3,4), D2:[a,b)}	5/11
...	
p_i={D1:[3,d), D2:[d,e)}	1/9
p_{i+1}={D1:[1,2), D2:[c,d)}	1/11
...	
p_n={D1:[3,4), D2:[c,d)}	3/9

(b) Emerging chemical patterns computed from the data set in (a).



(c) Classification of two unknown test compounds using the ECPs stored in table (b).

Figure 2.2: Outline of the classification procedure using a JEP based classifier. A data set of 20 compounds, divided into two classes, is projected into a chemical reference space of two descriptors D1 and D2. From this space, ECPs are mined and stored in a table. For classification of unknown test compounds, each pattern is tested wether is matches the test compound and the supports of matching patterns are summed. Two example test compounds are shown in (c). The first compounds matches patterns p_1, p_{i+1} and p_n. The cummulated support is 6/11 for the red class and 3/9 for the blue class. Thus, this compound is predicted to belong to the red class. The second test compound matches patterns p_{i+1} and p_n and has cummulated supports of 1/11 and 3/9 for the red and blue class, respectively. It is labeled to belong to the blue class.

in the classification experiments described in section 3.2, and a simple binning scheme based on statistical properties is used in the simulated sequential screening experiment described in section 4.2.1.

Emerging chemical patterns (ECP) are based on combining chemical information in the form of chemical descriptors and pattern mining. A *specific descriptor value range* is written as a pair $D : (min, max)$, where D denotes the descriptor name and min and max define the lower and upper bound of the value range. Round parenthesis and square brackets are used as to distinguish between open and closed intervals. A compound matches a descriptor value range if the corresponding descriptor value for that compounds lies between the lower and upper boundaries of the interval.

Definition 2.3.1 *A* chemical pattern (CP) cp *is a combination of* $1 \leq i \leq n$ *descriptor value ranges, where each descriptor* D_i *occurs only in one descriptor value range:*

$$\text{cp} = \{D_1 : (min_1, max_1), \ldots, D_n : (min_n, max_n)\}.$$

A set of descriptor value ranges S *is matched by a chemical pattern* cp *if and only if each descriptor value range of* cp *is also present in* S*:* $\text{cp} \subseteq S$.

A chemical pattern cp *is a* t*-emerging chemical pattern (ECP) for two data sets* D_1 *and* D_2 *if it has a growth rate larger than* t*:*

$$\text{growth}_{D_1,D_2}(\text{cp}) \geq t.$$

Based upon the concept of most-expressive jumping emerging patterns, *jumping emerging chemical patterns (JECP)* (Auer and Bajorath, 2006) are defined as most-expressive jumping emerging patterns computed from discretized chemical compound data represented as chemical descriptors. As described in section 2.2.4, classification is based on most-expressive JECPs only.

Two mining algorithms are used to extract ECPs from chemical data sets. For the virtual screening experiments and the simulated sequential screening in chapter 4, a hypergraph based algorithm developed by Bailey et al. (2003) was used which computes all JECPs for a data set of two classes. Classification is done by accumulating the supports of all patterns for a class which are present in the test compound's descriptor data. The knowledge mining experiment on 3D conformations in section 3.1 utilizes the zero-suppressed binary decision diagram based algorithm from Loekito and Bailey (2006) to compute all t-emerging patterns in-

stead of only jumping emerging patterns. Given the fact that the biological data here is highly imbalanced (a few active compound are usually compared to large compound databases), accumulated supports are normalized by dividing it by the maximum possible accumulated support for each class.

3 ECP Data Mining and Classification

A variety of machine learning methods are used in chemoinforamtics research. Typical applications include analyzing biological data sets, e.g. results of HTS campaigns, or predicting properties based on training sets with experimentally measured biological data.

This chapter describes two experiments used to validate and explore the potential of the ECP methodology in chemoinformatics. It is first shown that ECPs capture class-specific features in a high-resolution manner even if the underlying compounds are structurally highly similar and differ only in their 3D conformation. Computed patterns are validated on a molecular level to show the validity of the ECP mining approach. In a second experiment, ECPs are used to construct accurate classifiers on the basis of very small training sets. ECP-based classification is evaluated using four different compound sets. The prediction accuracy based on training sets of different size is compared to two established classification methods, namely a decision tree (DT) implementation and a Bayesian-based binary QSAR (BIN) classification technique, both implemented in MOE.

3.1 Data Mining for Conformational Differences

One feature of patterns is their simplicity as combinations of class-specific descriptor value ranges and thus the possibility to interpret and relate patterns to features of the molecules. knowledge mining for differences between two or more compound sets by analyzing biological data is one of the major applications of data mining algorithms in chemo- and bioinformatics. EPs have already been shown to extract useful knowledge out of gene expression data (Li and Wong, 2002a,b). These findings motivated an experiment where the aim was to find distinguishing patterns that reflect the differences of experimentally determined binding (bioactive) conformations of ligands compared to computationally predicted conformations. Sadowski (2003) reviews many methods for computing binding conformations, starting from early algorithms to compute the conforma-

tions of six-membered rings to state-of-the-art methods like Corina[1] or Omega[2]. A deeper understanding of the differences between computed and bioactive binding conformations would greatly increase the quality of 3D structure generation programs. Additional knowledge could be used to filter out conformations which are not similar to bioactive conformations or could directly be incorporated into rule- and data-based methods.

Over the past decade, several studies have investigated binding conformations of known active compounds, mostly enzyme inhibitors, taken from complex crystal structures (Agrafiotis et al., 2007; Bostrøm et al., 1998; Diller and Merz, 2002; Nicklaus et al., 1995; Perola and Charifson, 2004; Stockwell and Thornton, 2006). A major focal point of these investigations has been the analysis of intramolecular strain energy of small molecules that is generally induced upon protein binding. It is well appreciated that ligands do not bind in global energy-minimum conformations to their targets because achieving a high degree of molecular complementarity within a binding or active site generally comes at the cost of steric strain. The strain energy penalty associated with the formation of protein-ligand complexes can be approximated by computational means. For example, depending on the force field used to calculate relevant energy terms, Perola and Charifson (2004) have estimated total strain energy of average small molecular ligands to be approximately 2 kcal/mol. Steric strain effects contribute to the difficulties associated with correctly predicting bioactive ligand conformations, which is often attempted by systematic conformational sampling and filtering of low energy conformers. It is therefore not surprising that strain energy and its consequences have been intensely studied.

Going beyond the analysis of strain effects, only very few studies have attempted to systematically explore differences between binding and modeled conformations. For example, in a pioneering study reported in 2002, Diller and Merz compared 65 small molecules taken from X-ray structures of protein-ligand complexes to 5000 low energy conformations. For each experimental and corresponding energy-minimized conformations, the distribution of the values of six type III (three-dimensional) descriptors was compared. These descriptors included polar and apolar solvent-accessible surface area, the radius of gyration, dipole moment, the number of intermolecular interactions, and the ratio of two principal molecular axes. It was found that binding conformations tended to have larger solvent-accessible surface area than minimized conformations because of fewer intramolec-

[1] `http://www.molecular-networks.com/software/corina`
[2] `http://www.eyesopen.com/products/applications/omega.html`

ular interactions. Binding conformations were in general also found to be less compact than energy-minimized ones.

One would hope that systematic comparisons of bioactive and modeled ligand conformations for different targets might ultimately help to identify active conformations in conformational ensembles, which is of paramount importance for reliable 3D structure generation and thereby various ligand-based drug design strategies, e.g. QSAR modeling or pharmacophore analysis. This would require to deduce target-specific rules or feature combinations that could differentiate between alternative conformations. The ability to capture even subtle differences between highly similar compounds makes knowledge mining using ECPs a promising tool to follow up on the theme of the analysis by Diller and Merz. In the experiment presented in this chapter, ECP mining is evaluated for its potential to identify compound class-specific descriptor value range patterns (i.e. signature patterns) that distinguish bioactive conformations from other low energy conformers with high accuracy. Inhibitors of 18 target proteins were studied and in each case, ECP mining identified patterns that correctly identify bioactive conformations and differentiate them from others, even if conformational differences were subtle. Furthermore, key patterns could be rationalized at the molecular level of detail by analyzing X-ray structures of enzyme-inhibitor complexes.

3.1.1 Data Set

The data set is assembled from the PDBbind database (Wang et al., 2004, 2005b), an online accessible compilation of protein-ligand complexes extracted from the Protein Data Bank (PDB, release No. 107, January 2004). Ligand selection is done from the "refined" subset of the PDBbind, which provides high-quality ligand structures selected for comparison of structure-based virtual screening methods. This subset contains only X-ray crystallography structures[3] with a resolution of at least 2.5 Å. Only binary complexes, i.e. complexes formed by one protein and one ligand molecule, with non-covalently bound ligands and known equilibrium constants were added to the set. Additionally, ligand molecules are restricted to contain only the most common organic elements C, N, O, P, S, F, Cl, Br, I, and H and to have a molecular weight less than 1000 Da. The compounds in this set were divided into activity classes based on their target protein, and 18 classes (all enzymes) were selected because these classes have a reasonable number of different ligands. Table 3.1 summarizes these 18 classes. The class size ranges

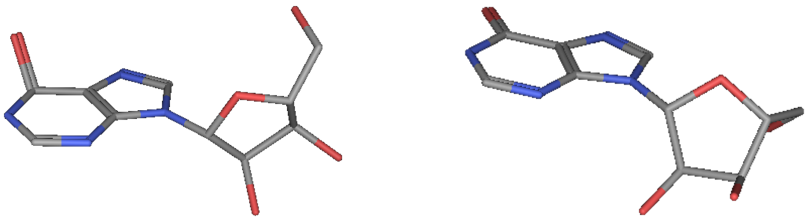
[3]NMR structures were not included since they only accounted for 39 out of 1622 candidate structures.

Table 3.1: Reported are the number of inhibitors with experimental binding conformations per class (Cmpds) and the number of modeled low energy conformers, their RMSD range relative to the corresponding experimental conformation, the average RMSD per class, and the number of emerging chemical patterns (Patterns) that discriminate between bioactive and modeled conformations.

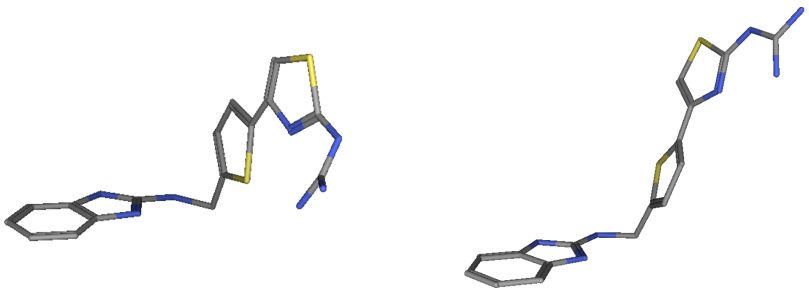
Class	Cmpds	Modelded conformers	RMSD range	RMSD average	Patterns
Acetylcholine esterase	5	16	0.69 - 5.39	2.69	15
Adenosine deaminase	15	139	1.10 - 4.18	2.45	17
Carbonic anhydrase	21	198	0.57 - 4.30	2.26	36
Carboxypeptidase	8	60	0.73 - 5.03	2.71	17
Cyclin-dependent kinase	31	247	0.70 - 4.08	2.24	7
Elastase	3	30	1.59 - 4.09	2.69	81
Endothiapepsin	6	19	3.30 - 5.87	4.80	33
Factor Xa	10	91	0.94 - 5.68	2.52	32
FK506 binding protein	6	51	1.82 - 4.23	3.06	15
HIV protease	20	142	0.98 - 6.27	3.69	8
Plasminogen activator	7	43	0.25 - 2.76	1.69	8
PT Phosphatase1b	14	67	0.30 - 4.58	2.40	17
Protocat.-dioxygenase	10	16	0.08 - 3.39	0.87	10
RibonucleaseA	9	45	0.98 - 3.39	2.12	10
Thermolysin	6	42	2.05 - 4.95	3.30	197
Thrombin	21	158	0.60 - 4.51	2.69	7
Trypsin	30	195	0.33 - 3.37	1.74	5
Tyrosine kinase	5	23	2.72 - 5.47	3.95	30

from three for elastase to 31 for cyclin-dependent kinase (CDK). On average, each class contains ∼13 conformations.

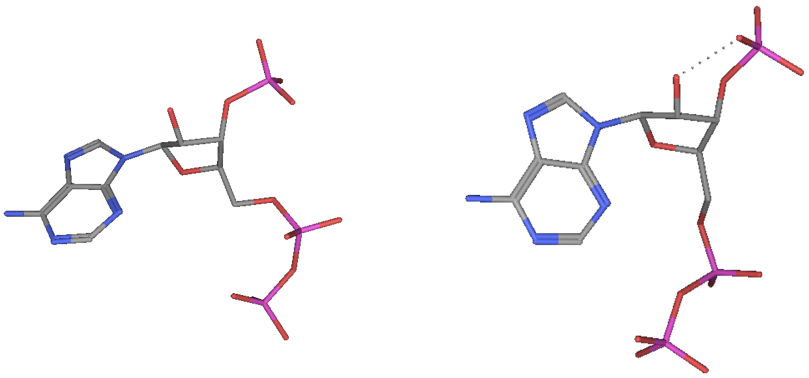
Each active compound was subjected to extensive conformational search using the Molecular Operating Environment (MOE, version 2007.09). A stochastic conformational search was carried out for 10,000 iterations by randomly rotating single bonds in test molecules. Following each iteration, the resulting conformation was energy-minimized and sampled. Cartesian minimization was carried out using MOE's MMFF94x force field until the RMS gradient of the energy function was less than $0.001\,\mathrm{kcal/(mol \times Å)}$. For each inhibitor, sampled low energy conformations were compared in order to eliminate conformations from pairs of very similar ones, applying a root mean square deviation (RMSD) threshold value of 0.1 Å. As reported in table 3.1, between 16 and 247 low energy conformations were retained per class with an average of 88 conformations. RMSD values for modeled and experimental conformations were calculated based on superposition of all non-hydrogen atoms. In 12 of 16 cases, the conformational ensembles contained conformers that were very similar to binding conformations, i.e. within



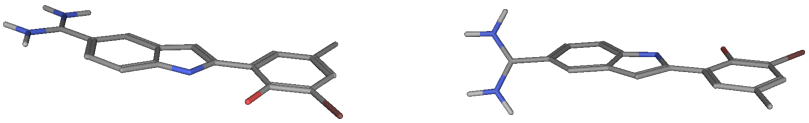
(a) Adenosine deaminase inhibitors



(b) Adenosine deaminase inhibitors



(c) Ribonuclease A inhibitors



(d) Trypsin inhibitors

Figure 3.1: Exemplary bioactive and modeled conformations. The binding conformations (on the left) and corresponding low energy conformers (right) are shown for four inhibitors of three enzymes discussed in the text. The dashed line represents an intramolecular hydrogen bond. (a) and (b) adenosine deaminase inhibitors (taken from PDBbind entries "1fkx" and "1ndv", respectively), (c) ribonuclease A inhibitor (1afk), (d) trypsin inhibitor (1o3h).

1 Å RMSD. Thus, in many cases, differences between experimental and modeled conformations were rather subtle, which can also be appreciated in figure 3.1. However, all ensembles also contained conformations that deviated from binding conformations by several Å RMSD. Most classes produced an average RMSD of around 2 Å, suggesting that modeled conformers were overall not dramatically different from binding conformations. The sampled conformations represent a spectrum of conformers provides a good basis for the analysis of the ability of ECPs to capture class-specific knowledge about the differences in bioactive conformations.

3.1.2 Methodology

Pattern Mining

The experiments described herein are based on differences in the spatial arrangements of molecules. In order to capture differences in their 3D properties, the set of type III descriptors described in section 2.1.2 was used as the basis for pattern mining. Prior to descriptor calculation, the compounds were normalized by first aligning their three principle molecular axes to the x-, y- and z-axis in descending order. Afterwards, the molecules were translated such that their center of mass matched the origin of the coordinate system. This minimized the influence of translational and rotational differences for descriptors depending on external coordinates.

For ECP mining, the descriptors were first discretized using the supervised information entropy-based discretization method. A fast implementation of ECP mining based on zero-suppressed binary decision diagrams (Loekito and Bailey, 2006) was then used to compute all ECPs matching given support thresholds in bioactive and modeled classes. For bioactive conformations, a support threshold of min. 50% was applied, meaning that patterns matching at least half of the binding conformations were calculated. For modeled conformations, a threshold of max. 10% was applied. These parameter settings ensured that each detected pattern was at least five times more frequent in bioactive than in modeled conformations. For each inhibitor set, all patterns with a growth rate of at least 10 were analyzed.

Protein-Ligand Complex Depiction

For the interpretation of key patterns, details of protein-ligand interactions in the X-ray structures of their complexes were studied and represented with the aid of two-dimensional (2D) interaction diagrams (Clark and Labute, 2007) calculated

with MOE. These diagrams provide a 2D abstraction of the interactions between a ligand and its target protein and are used extensively in section 3.1.4 to analyse the interactions of conformations. In the interaction diagrams, the active site region is delineated as an envelope. Amino acids are color-coded according to polar, negatively charged, and positively charged (pink, red, and blue, respectively) and hydrophobic (green) character. Solvent-exposed residues and ligand atoms have additional shading (light and dark blue, respectively). Hydrogen bonds are drawn as dashed donor-acceptor arrows and are colored green if they involve a protein side chain or blue if they involve backbone atoms. Green dashed lines containing aromatic ring symbols indicate donor interactions with π-electron systems or π-π interactions. Metal contacts are displayed in magenta. For comparison, modeled low energy conformers were superposed on experimental conformations of each inhibitor and interaction diagrams were also analyzed for these hypothetical complexes. Superposition was done by matching of non-hydrogen atoms in both conformations. These hypothetical complexes were used to interpret patterns in structural terms instead of pure statistical analysis.

Two experiments were done using the methodology and data set described above. First, ECPs were used to discover global differences in bioactive and energy-minimized conformations. A second set of experiments further investigated these differences on a per-class basis. Patterns were rationalized based on the interactions formed upon formation of the ligand-protein complexes.

3.1.3 Differences Between Modeled and Bioactive Conformations

Assessing the global differences between modeled and bioactive conformation has been the focus of several studies during the last years. These studies have confirmed that modeled ligands generally have more compact sphere-like conformations, while bioactive conformations tend to elongated out to form energetically favorable ligand-protein interactions, e.g. hydrogen bonds, or to compensate charge distributions. This leads to an induced strain energy in the ligand itself. Additionally, bioactive conformations tend to have a larger surface-area and less intramolecular interactions, e.g. internal hydrogen bonds (Diller and Merz, 2002; Perola and Charifson, 2004). These findings should also be reflected in the patterns computed from the set of all bioactive and inactive conformations and thus provides a useful proof-of-concept experiment for ECP knowledge mining on 3D conformations prior to analyzing descriptor patterns for individual classes.

ECPs were computed from the combined data set of all 227 bioactive conformations and 1598 low-energy conformations. The three patterns with largest growth

Table 3.2: Emerging chemical patterns for discriminating the combination of all bioactive from all modeled conformations. The growth of the most discriminatory patterns is reported. "B" stands for bioactive/binding and "M" for modeled conformations.

Growth	B [%]	B	M [%]	M	Pattern
1364.14	86	195	0	1	{E_strain:(24.46,∞]}
75.95	67	152	1	14	{E_str:(14.90,∞]}
29.22	31	71	1	17	{E:(147.95,∞]}

rate are reported in table 3.2. Only the first two patterns met the predefined support threshold levels of min. 50% support for bioactive conformations and max. 10% support for low-energy conformations. Both patterns discriminate effectively, given the magnitude of their growth rates, and clearly reflect the general introduction of ligand strain energy upon complex formation, regardless of the nature of the interactions. The strain energy pattern was dominant with a growth rate of 1364, followed by the bond stretch energy pattern having a growth rate of 76. The third pattern in table 3.2 with lower growth indicates that the total potential energy of binding conformations was generally higher compared to minimized conformations, which is of course also expected. This pattern has a support rate of only 1% in the energy-minimized conformations and is thus very specific for bioactive conformations. However, its support rate of 31% in the bioactive conformation class makes it less general than specified by the stringent threshold of 50% for the bioactive class. The overall diversity of the ligands in the data set and the high growth rate of almost 30 rationalize the inclusion of this pattern in the global analysis.

At the chosen level of support stringency, only patterns were identified that referred to strain or total energy as generally discriminating features. Patterns relating to differences in shape, e.g. compactness measures, were not found because the high ligand diversity prevented shape-related patterns from achieving a high support of at least 50%. However, the global comparison of binding and modeled conformations on the basis of emerging chemical patterns produced meaningful results highlighting the well-known strain energy penalties.

3.1.4 Class-based Pattern Mining

The global analysis has shown that ECP was able to compute reasonable and discriminating patterns to distinguish bioactive from low-energy conformations. These global patterns were simple general rules to discriminate conformations

based on energetic properties, mainly strain and bond-stretch energy penalties. A class-level analysis of computed ECPs is motivated by the high diversity of the data set and the different binding modes of the included inhibitors. Different binding modes result in different chemical properties and thus in different ECPs. However, the overall diversity makes it hard to satisfy the support threshold for bioactive conformations. A second set of experiments thus investigated the ability of ECP mining on a per-class basis. In this experiment, only conformations from one of the 18 activity classes were used as a database for ECP mining to compute class-specific ECPs discriminating bioactive from low-energy conformations.

Table 3.1 reports the number of class-specific patterns that were obtained, given the applied support stringency threshold, which ranged from five for trypsin to 197 for thermolysin inhibitors. In most cases, between approximately 10 and 30 patterns were obtained. In general, the more diverse experimental binding conformations are, the fewer widely applicable signature patterns are identified. Patterns for three representative sets are reported in tables 3.3 to 3.5 and discussed in detail in the next section including a discussion of the conformational differences in the corresponding ligand conformations and their influence on the interactions of the induced protein-ligand complex. Patterns for all remaining classes, including the four classes discussed in the following text are reported in appendix D. As expected, strain energy and related patterns were found in all cases. However, for each of the 18 inhibitor sets, different types of signature patterns also emerged. Most patterns were relatively small including only one to five descriptor value ranges, with the exception of the thermolysin set that produced patterns with up to nine descriptors. This set also generated the largest number of patterns, which is not surprising, given the fact that increasing numbers of descriptors lead to an exponential growth in the possible number of patterns. These tables show that the majority of signature patterns had infinite growth, which means that they exclusively occurred in bioactive, but not modeled reference conformations. Therefore, these patterns were highly specific. In many instances, patterns consisting of a single energy descriptor were found to be highly discriminatory. Among these were torsion and out-of-plane energy descriptors whose value ranges were indicative of in part significant distortions of inhibitors upon binding. However, patterns containing no energy term descriptors were also found in most cases and usually consisted of combinations of at least two descriptors. For example, the pattern {CASA+:(1956.91,∞], FCASA-:(1.40,3.21], pmi:(10083.21,∞]} discriminated bioactive factor Xa inhibitor conformations from inactive conformations. This pattern combines charge distributions on the solvent-accessible

surface area with the principal moment of inertia (a descriptor of general mass distribution). For ligands of the FK506 binding protein, angle bending energy and hydrophobic solvent accessible surface area displayed value range preferences for bioactive conformations, but the individual descriptors only partly discriminated between bioactive and modeled conformations. However, a pattern combining these preferred value ranges perfectly classified these conformers. Similar patterns combining energy term descriptors with surface area, charge, or shape features were found to be prominent for many classes. For example, for protein tyrosine phosphatase 1b inhibitors, the pattern {ASA_P:(210.39,282.26], E_ele:$(-11.07,\infty]$, E_tor:$(1.88,\infty]$} combines two energy descriptors with polar solvent-accessible surface area and has infinite growth rate. Other examples include ribonuclease A inhibitors where a combination of the dipole moment and electrostatic energy leads to patterns with infinite growth rate. In some cases, even relatively simple geometric descriptors were found to be highly discriminatory. As an example, for protocatechuate-3,4-dioxygenase, simple shape measures such as the extension along a molecular axis effectively distinguished between bioactive and modeled conformations. In this case, the inhibitors are small (substituted benzenes). Upon binding to the enzyme, substituents are forced out of the aromatic ring plane, due to a structural constraint caused by a bound cation, which significantly alters the shape of these small inhibitors. Energy minimization of these ligands without the presence of the cation places the substituents in an energetically favourable planar position and thus leads to smaller extension along the vertical principle axis.

Taken together, pattern analysis for individual compound classes revealed that highly discriminatory patterns of variable composition were identified in each case. These target-specific patterns accurately distinguished between bioactive and modeled conformations of inhibitors.

3.1.5 Structural Interpretation of ECPs

The quantitative analysis in the previous section has demonstrated that binding conformation sensitive patterns could be systematically identified for a variety of target sets. A further attempt to rationalize the validity of the pattern mining approach was to go beyond statistical analysis of discriminatory patterns and attempt to relate them to details of protein-ligand interactions. For ECPs, this meant to relate signature patterns to changes in the interactions between protein-ligand complexes formed by the bioactive ligand conformation and a hypothetical protein-ligand complex formed by a low-energy conformation. The low-energy

protein-ligand complexes were computed by superposition of the low-energy ligand and the bioactive conformation taken from the X-ray crystallographic data. However, key interactions were expected to differ as a consequence of conformational differences that are a prerequisite for highly discriminatory signature patterns. As shown in the previous sections, binding conformations contain several energy penalties (e.g. strain or bond-stretch energy) compared to low-energy conformations.

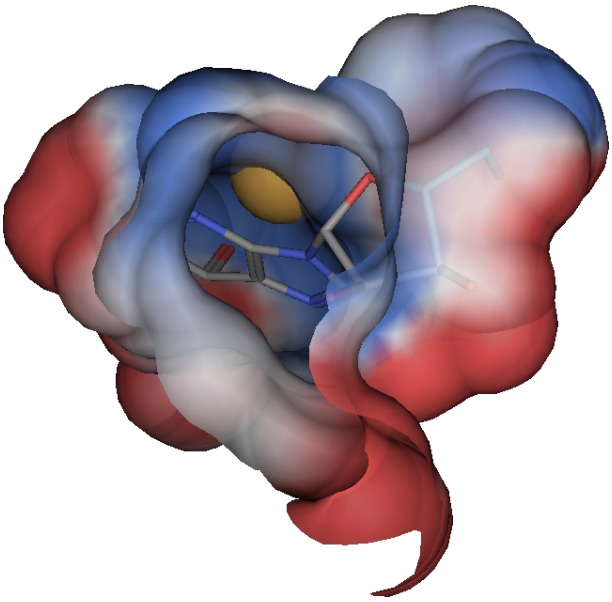
The remaining part of this section presents an analysis of three representative cases, inhibitors of adenosine deaminase, ribonuclease A, and trypsin, for which small and intuitive discriminatory patterns were identified. Figure 3.1 shows the bioactive and modeled conformations of inhibitors used to generate exemplary complexes. The examples will illustrate that modeled conformations of the inhibitors studied here could not have been used to correctly predict details of the enzyme-inhibitor interactions, even if pose information was available. However, the analysis shows that key conformational differences can be well rationalized with the aid of discriminatory descriptor patterns.

Adenosine Deaminase Inhibitors

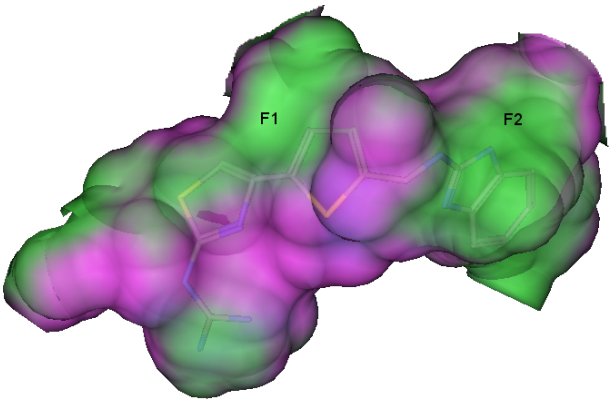
Adenosine deaminases is a metallo-enzymes that catalyzes the deamination reaction of adenosine to inosine. The active site contains a zinc cation that is involved in the activation of a water molecule during catalysis (Cristalli et al., 2001). Table 3.3 reports the signature patterns for adenosine deaminase inhibitors.

Typical strain and bond stretch energy descriptors displayed increased values for binding conformations. In addition, increased out-of-plane and angle bending energies were characteristic for the majority of binding, but for none of the modeled conformations. Ring distortions were detected in all inhibitors with available crystallographic binding conformations. Figure 3.2a shows the active site together with the bound ligand as taken from the crystallographic data and figure 3.3a shows the interactions of this complex. The hypoxanthine ring system is twisted in order to position the carbonyl oxygen above the ring plane, which results in high out-of-plane and angle bending energies. In this position, the carbonyl oxygen strongly interacts with the zinc cation and thereby inhibits the enzyme. Figure 3.3b shows the corresponding hypothetical complex. In the modeled inhibitor, the ring system is planar and not distorted, which prevents the interaction between the carbonyl oxygen and the zinc cation.

Figure 3.3c shows the crystallographic complex formed by a chemically different adenosine deaminase inhibitor that contains three rings. All three ring systems

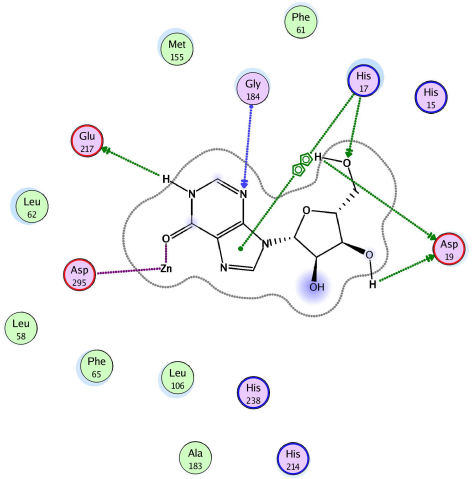


(a) Binding site and ligand conformation of the crystallographic 1fkx protein-ligand complex. The active site of adenosine deaminase contains a zinc cation (shown in yellow) that catalyzes the deamination of adenosine by activating a water molecule. The inhibitor interacts with the zinc cation, resulting in a distorted hypoxanthine ring system.

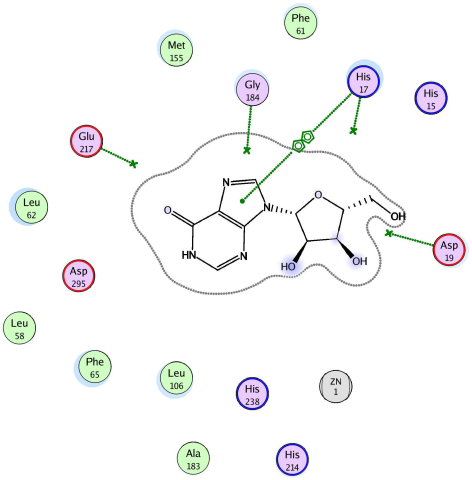


(b) Binding site and ligand conformation of the crystallographic 1ndv protein-ligand complex. The ligand occupies several hydrophobic pockets in the active site of adenosine deaminase. This requires high shape complementary and results in twisted ring conformations of the ligand.

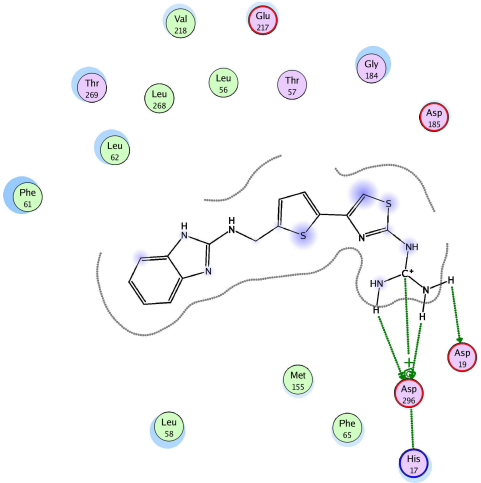
Figure 3.2: Binding sites and ligand conformations for two adenosine deaminase inhibitor conformations from crystallographic data. The color scheme in (a) visualizes the partial charge of the binding pocket. Red and blue encode the charge of the pocket, from negatively (red) to positively charged areas (blue). In (b), the color-code visualizes possible interactions of the binding pocket. Hydrophobic areas are colored in green, purple encodes possible hydrogen bonding sites and blue weakly polar areas.



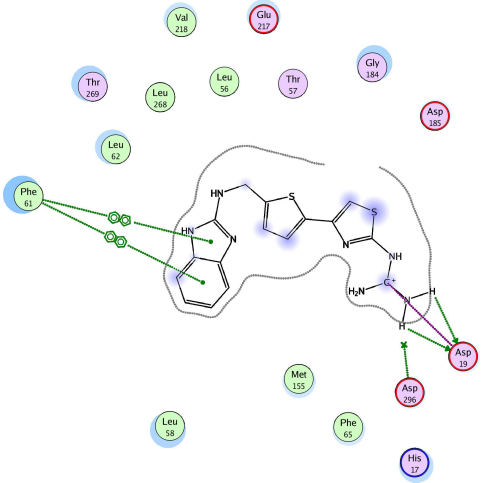
(a) X-ray structure of an adenosine deaminase inhibitor complex (1fkx). The inhibitor complexes a zinc cation that activates a water molecule during catalysis. The contact is formed through a distorted hypoxanthine ring system.



(b) The corresponding protein-ligand hypothetical protein-ligand complex with a low energy conformer. Here the hypoxanthine ring is planar, which prevents the zinc contact, and fewer interactions are formed.



(c) Structure of another adenosine deaminase-inhibitor complex (1ndv). Distorted ring systems (with large calculated out-of-plane energy penalty) match hydrophobic binding pockets and present a guanidino group for an array of salt bridge interactions.



(d) The corresponding protein-ligand hypothetical protein-ligand complex with a low energy conformer. In the modeled conformation, the rings are planar, the guanidino group is re-positioned, and the salt bridge interaction can no longer be formed.

Figure 3.3: Experimental and modeled adenosine deaminase interactions. Protein-ligand interaction plots are described in section 3.1.2. Ligand conformations are shown in figure 3.1.

Table 3.3: Discriminatory patterns for adenosine deaminase inhibitors. "B" stands for bioactive/binding and "M" for modeled conformations.

Growth	B [%]	B	M [%]	M	Pattern
∞	93	14	0	0	{E_strain:(24.46,∞]}
∞	80	12	0	0	{E_str:(14.90,∞]}
∞	53	8	0	0	{E_tor:(1.88,∞], std_dim2:(1.81,∞]}
∞	53	8	0	0	{E_tor:(1.88,∞], pmiY:(1141.06,∞]}
∞	53	8	0	0	{E_oop:(1.64,∞]}
∞	53	8	0	0	{E_ang:(17.54,∞]}
∞	100	15	0	0	{E_strain:(15.69,∞]}
∞	87	13	0	0	{E_str:(10.26,∞]}
∞	73	11	0	0	{E_ang:(13.51,∞]}
∞	53	8	0	0	{E_stb:(0.96,∞]}
138.00	100	15	1	1	{E:(55.17,∞]}
32.20	93	14	3	4	{E_oop:(0.25,∞]}
27.60	60	9	2	3	{E_tor:(1.88,∞], E_vdw:(28.30,72.89]}
24.53	53	8	2	3	{E_tor:(1.88,∞], PM3_HF:($-$85.25,69.90]}
24.53	53	8	2	3	{E_tor:(1.88,∞], MNDO_HF:($-$79.25,169.85]}

occupy hydrophobic pockets and significantly contribute to the binding affinity (Terasaka et al., 2004), as can be seen in figure 3.2b. The mode of inhibition is completely distinct from the previously discussed inhibitor in this case. It does not involve complexation of the zinc cation, as discussed above, but rather forms interactions between a guanidino group of the inhibitor and catalytic residues. In its bound conformation, all three rings of the inhibitor are twisted to varying degrees. The largest contribution to the out-of-plane energy results from a deformation of the benzimidazol-2-amino moiety where the amino substituent is moved out of the ring plane. The ring distortions are an apparent consequence of achieving a degree of shape complementarity while correctly positioning the guanidino group for strong salt bridge and hydrogen bonding interactions with aspartic acid and histidine residues. Figure 3.3d shows the corresponding hypothetical complex. The modeled ligand with planar ring systems can no longer form the strong interactions via the guanidino group and achieves overall lower shape complementarity, although the benzimidazol ring penetrates deeper into the hydrophobic F2 pocket.

Thus, in the case of adenosine deaminase, signature patterns were identified for a compound set containing chemically distinct inhibitors with different modes of action that discriminated between bioactive and modeled conformations with

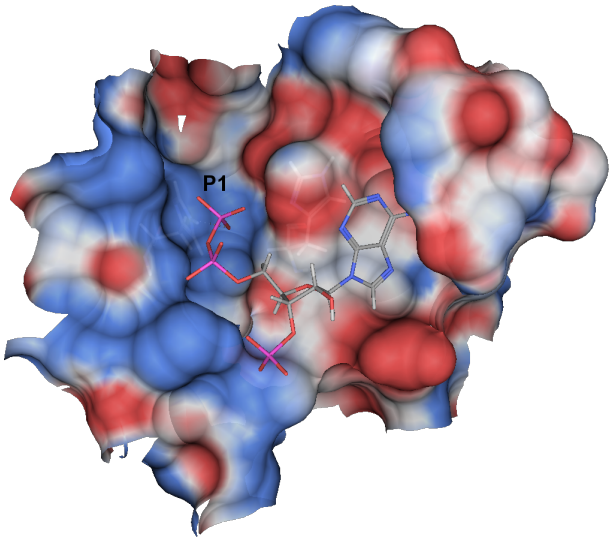


Figure 3.4: Binding site and ligand conformation of the crystallographic 1afk protein-ligand complex. The active site of ribonuclease A contains several positively charged binding pockets designed for RNA phosphate groups. These pockets are occupied by phosphate groups from the inhibitor, which results in high shape complementarity and places hard constraints on the binding conformation of ribonuclease A inhibitors. Colors encode the charge of the binding site from positive (blue) to negative (red) partial charge.

high accuracy. This was possible because the inhibitors were conformationally perturbed in similar ways upon binding, although their structures and interactions were distinct.

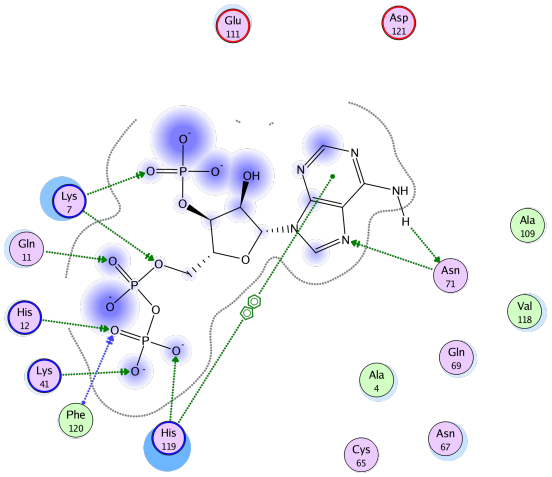
Ribonuclease A inhibitors

For ribonuclease A inhibitors, strain and bond stretch energy were not the most discriminatory patterns. As reported in table 3.4, other descriptors were found to perfectly discriminate between bioactive and modeled conformations.

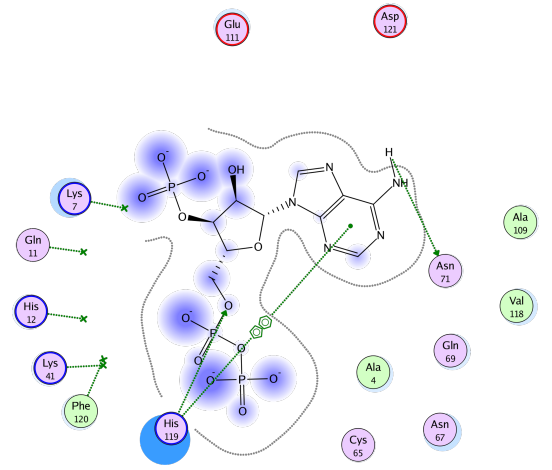
Signature patterns contained the van der Waals energy as a single descriptor as well as combinations of the dipole moment and various electrostatic energy term descriptors. The comparison of binding conformations with van der Waals energy values matching the signature pattern and the corresponding modeled conformations revealed that the increase in van der Waals energy in the low energy conformers resulted from the formation of an intramolecular hydrogen bond involving one of the phosphate groups (figure 3.1c) that was not observed in the binding conformation. Figure 3.4 shows the binding pocket of the ribonuclease A protein together with the crystallographic binding conformation of the inhibi-

Table 3.4: Discriminatory patterns for ribonuclease A inhibitors. "B" stands for bioactive/binding and "M" for modeled conformations.

Growth	B [%]	B	M [%]	M	Pattern
∞	78	7	0	0	{E_vdw:$(-\infty$,28.30]}
∞	56	5	0	0	{AM1_dipole:(17.34,∞], PM3_Eele:$(-755075.90,-335581.90]$}
∞	56	5	0	0	{AM1_dipole:(17.34,∞], MNDO_Eele:$(-788514.20,-336721.10]$}
∞	56	5	0	0	{AM1_Eele:$(-783413.80,-340732.90]$, AM1_dipole:(17.34,∞]}
∞	78	7	0	0	{E_vdw:$(-\infty$,32.76]}
∞	67	6	0	0	{E_strain:(32.31,∞]}
∞	67	6	0	0	{E_stb:$(-0.36,\infty]$}
∞	56	5	0	0	{E_str:(9.28,∞]}
∞	56	5	0	0	{E:(92.14,∞]}
29.33	67	6	2	1	{E_strain:(24.46,∞]}
6.11	56	5	9	4	{AM1_dipole:(17.34,∞], ASA+:$(-\infty$,227.75], PM3_E:$(-124393.40,-95187.02]$, dipoleZ:$(-1.24,\infty]$}
6.11	56	5	9	4	{AM1_dipole:(17.34,∞],ASA+:$(-\infty$,227.75], MNDO_E:$(-158562.00,-105111.10]$, dipoleZ:$(-1.24,\infty]$}
4.89	56	5	11	5	{dipoleX:$(-\infty,-1.59]$, glob:(0.08,0.10]}



(a) Interactions of a ribonuclease A inhibitor complex (1afk). Binding of the phosphate groups is stabilized by multiple salt bridges and hydrogen bonds.



(b) The corresponding protein-ligand hypothetical complex. Many of the key interactions are absent.

Figure 3.5: Interactions of experimental and modeled ribonuclease A conformations. Protein-interaction plots are described in section 3.1.2. Figure 3.1 shows the two ligand conformations and figure 3.4 shows the lingand in the binding site of the crystallographic complex.

tor 1afk. The phosphate groups of ribonuclease A inhibitors occupy positively charged regions in the active site that accommodate the phosphate groups of the RNA substrate (Yakovlev et al., 2006). Figure 3.5a shows the interactions of the X-ray structure of a ribonuclease A-inhibitor complex. In its binding conformation, the inhibitor positions the phosphate groups to strongly interact with the histidine and lysine residues in the phosphate binding pockets, thereby achieving charge complementarity. In the corresponding hypothetical complex shown in figure 3.5b, the phosphate groups are positioned differently and many of the interactions seen in the crystal structure can no longer be formed. Here signature patterns correctly detected a conformational artifact that led to a more compact inhibitor structure that would not have been capable of forming the electrostatic interactions within the active site.

Trypsin inhibitors

Trypsin inhibitors represent an interesting test case for pattern analysis. Diller and Merz (2002) found that bioactive trypsin inhibitor conformations were particularly difficult to distinguish from minimized conformations because binding conformations also displayed strong intramolecular interactions between ring systems. However, as shown in table 3.5, besides the general increase in strain and bond stretch energy, two discriminatory patterns with single descriptors emerged, accounting for bond stretching/angle bending cross term energy and the z-component of the principal moment of inertia.

Trypsin inhibitors matching these patterns contain a diamino methyl substituent at an indol ring (figure 3.1d). In its binding conformation, the diamino methyl substituent is co-planar with the indol ring (figure 3.6), which correctly positions the amino groups for well-defined hydrogen bonding and electrostatic interactions (figure 3.7a). This conformation has a low z-component of the principal moment of inertia because most atoms are positioned in or near the indol ring

Table 3.5: Discriminatory patterns for trypsin inhibitors. "B" stands for bioactive/binding and "M" for modeled conformations.

Growth	B [%]	B	M [%]	M	Pattern
∞	100	30	0	0	{E_strain:(5.11,∞]}
∞	63	19	0	0	{E_str:(10.03,∞]}
∞	63	19	0	0	{E_stb:(0.53,∞]}
24.25	50	15	2	4	{pmiZ:($-\infty$,16.69]}

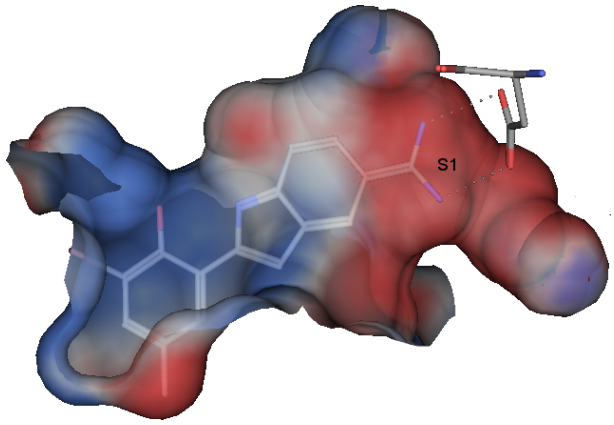
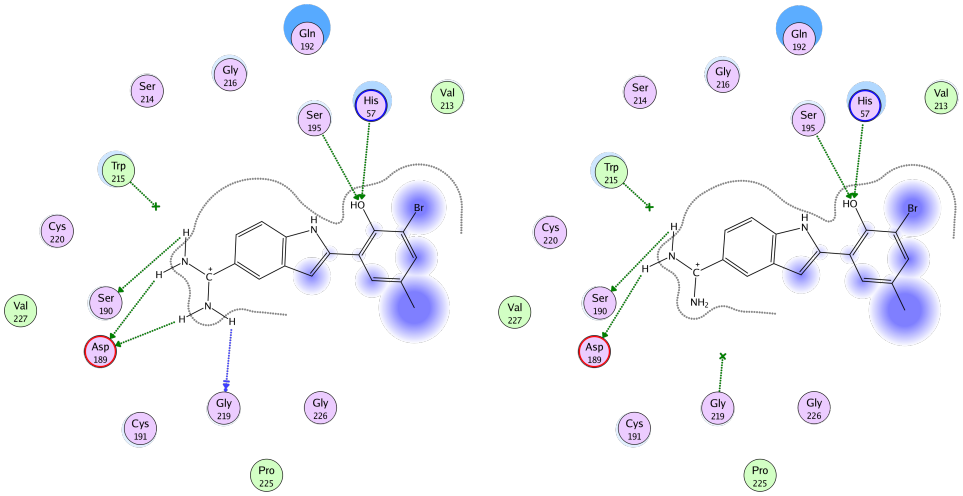


Figure 3.6: Binding site and ligand conformation of the crystallographic 1o3h protein-ligand complex. Trypsin is a protease catalyzing the hydrolysis of peptide bonds at positively charged residues. The negative asparagine in the S1 pocket recognizes these residues. The inhibitor interacts with this asparagine through axial diamino groups.



(a) Trypsin-inhibitor complex (1o3h). The diamino methyl substituent is in a strained equatorial conformation that enables the formation of multiple salt bridge and hydrogen bonding interactions.

(b) The corresponding hypothetical complex. Here the diamino methyl group is in energetically preferred axial orientation, which breaks the crystallographic interaction pattern.

Figure 3.7: Interactions of experimental and modeled trypsin conformations. Figure 3.6 shows the binding site of the crystallographic complex. Colors encode the charge of the binding site from positive (blue) to negative (red) partial charge. Protein-interaction plots are described in section 3.1.2. Ligand conformations are shown in figure 3.1.

plane of the molecule. Figure 3.7b shows the corresponding hypothetical complex. In the modeled conformation, the diamino methyl moiety is rotated such that the amino groups are orthogonal to the indol ring, which increases the value of z-component of the principal moment of inertia. However, this orientation would break the network of crystallographic interactions involving the diamino methyl substituent.

3.2 ECP Classification

In early stages of hit-to-lead or lead-optimization programs, the amount of available training data is rather limited. Additionally, biological data is often biased towards inactive compounds because only a small set of all possible chemical compounds shows desired biological properties, e.g. activity against certain targets. This situation has motivated the development of a novel classification approach that could also be applied in these situations, for example, when attempting to guide compound selection or design on the basis of only a few reference molecules. Traditionally, such efforts have been supported by qunatitative structure-activity relationship (QSAR)-type methods (Esposito et al., 2004) to quantitatively model structure-activity relationships and predict analogues having improved potency. However, QSAR methods also require high-quality training data sets containing as many compounds as possible, which are often not available when analyzing novel hits or leads.

Additionally, a high-quality computational model decreases the number of screening experiments in early stages of drug discovery and helps to focus on compounds with desired properties. Experimental screening is expensive in both time and money, constituting as much as 15% to the total research and development budget of pharma companies (Klopack, 2000). Consequently, reducing the number of screens also reduces the costs and time of drug-discovery programs. Although this chapter focuses on the prediction of potency, other properties, e.g. toxicity or ADME properties, could also be predicted, ruling out compounds in early stages which otherwise would have been included and eliminated in later stages of the drug discovery program.

3.2.1 Data Set

The experiments in this chapter use a total of four classes of compounds with experimentally measured binding affinities to different targets. These classes are assembled from public sources and contain compounds covering greatly varying potency distributions: benzodiazepine receptor (BZR) ligands, inhibitors of dihydrofolate reductase (DHFR), glycogen synthase kinase (GSK3) and HIV protease inhibitors (HIVPROT). The first two sets are taken from a QSAR study done by Sutherland et al. (2003). The other two sets have been extracted from the BindingDB database (Chen et al., 2002). In all data sets, potency is measured as half maximal inhibitory concentration (IC_{50}) values. IC_{50} is the amount of a compound needed to bind to half of the target molecules present in the experi-

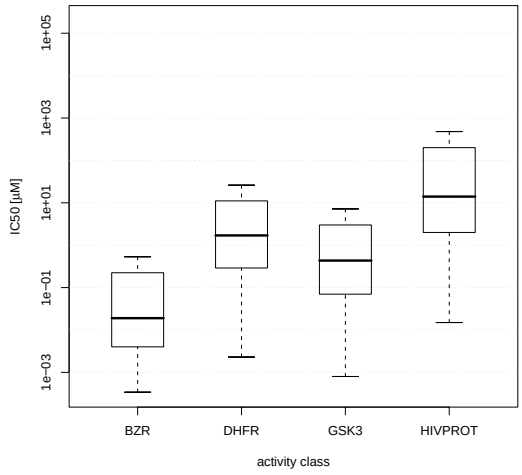


Figure 3.8: Potency distribution within compound activity classes. Shown are box plots for each activity class (see table 3.6). In these plots, boxes show the 0.75 (top) and 0.25 quartile (bottom) separated by the median (horizontal bar). The lines indicate the largest and smallest values (falling within a distance of max. 1.5 times the box size from the nearest hinge).

Table 3.6: Compound classes and potency levels. Compound potencies are reported as IC_{50} values ; "n" is the number of compounds per class and the next two columns report how many of these compounds have IC_{50} values below or above $1\,\mu$M; "max", "min", and "avg" give the highest, lowest, and average compound potency, respectively.

class	n	$\leq 1\,\mu$M	$> 1\,\mu$M	max.	min.	avg.
BZR	321	283	38	0.00034	250.00	2.02
DHFR	586	249	337	0.0023	929.00	16.67
GSK3	464	281	183	0.0008	1000.00	12.84
HIVPROT	967	821	146	0.000015	200.00	2.68

ment. The composition and potency distribution of the data set is summarized in table 3.6. All sets contain more than 300 compounds with measured potencies, but the distribution of highly potent ($IC_{50} \leq 1\,\mu$M) and weakly potent compounds ($IC_{50} > 1\,\mu$M) varies. The BZR and HIVPROT sets are dominated by compounds with potency (small IC_{50}), reflected by high average and minimum potency. Potencies in the other two sets are more equally distributed, with a bias for weakly potent compounds in DHFR and a bias towards compounds with high potency in GSK3. Figure 3.8 further illustrates the potency distribution in each class in the form of boxplots. Each class spans a range of three orders of magnitude. For DHFR, GSK3 and HIVPROT, the potency distribution is centered somewhere around the $1\,\mu$M threshold between highly and weakly active

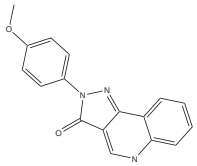
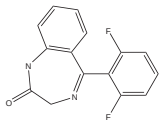
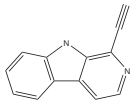
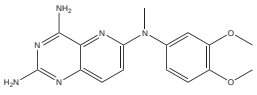
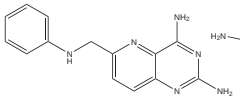
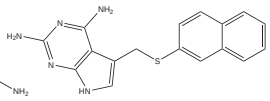
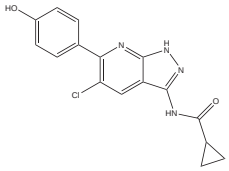
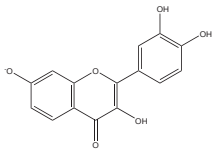
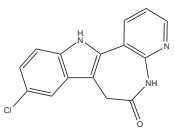
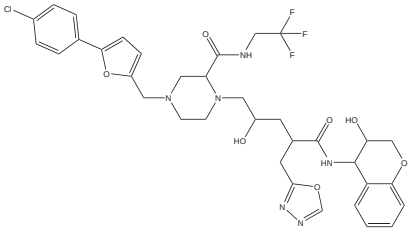
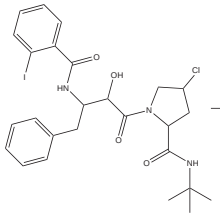
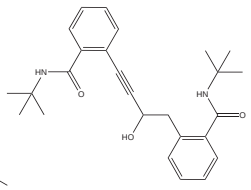
Class	Most potent	Median potent	Least potent
BZR			
DHFR			
GSK3			
HIVPROT			

Figure 3.9: Representative structures for the four compound classes used in the classification and virtual lead optimization experiments. For each activity class, the most and least potent compounds are shown together with one having an IC_{50} value equal or close to the median.

compounds, whereas the compounds in BZR show a smaller IC_{50} value and thus a higher potency.

Figure 3.9 shows representative structures for each class. It can be seen that the inter- and intra-class similarity of the classes is rather low. This diversity makes them a challenging and interesting set for the evaluation ECP classification.

3.2.2 Evaluation

In all experiments, the set of 61 uncorrelated, information rich descriptors described in section 2.1.1 is used. This set covers a wide range of descriptors, but does not utilize any 3D information of the compounds.

The first experiment is a proof-of-concept experiment that validates the accuracy of the ECP classifier compared to standard approaches in chemoinformatics. The ECP approach was applied to classify compounds belonging to either the sets with potency above or below an IC_{50} value of 1 μM and compared to binary QSAR and decision trees, as implemented in MOE. As described in section 2.2.4, classification is based on most expressive jumping ECPs (JECPs) only. For each classification, training sets of increasing size (10–50% of compounds belonging to each activity class) were selected. For binary QSAR, subsets of most suitable descriptors are selected by applying MOE contingency analysis prior to model building. For classification, a probability threshold value of 0.5 was applied for each compound to belong to the higher potency class. For decision trees, the 61 preselected descriptors available in MOE were evaluated during the tree construction process. The constructed tree was post-processed in a pruning step to remove not required attributes and increase general applicability. Two trees were constructed in a 2-fold cross-validation step and the best one was selected as the final decision tree structure. For all three methods, classifiers were trained on 500 randomly selected sets, the resulting models applied to predict the potency-dependent class label of the remaining compounds and average accuracies calculated. Initially, training sets containing 10–50% of compounds per class were used.

In a second series of calculations, classifiers derived from training sets of very small size are analyzed, only consisting of three, five, or 10 compounds per class from each potency range, corresponding to ∼0.6–6% of the compounds in the activity classes. This presents a typical lead optimization scenario where only a few active compounds are available as an information source for predictions.

Table 3.7: JECPs for activity class GSK3. JECPs with highest support are reported for a single training calculation on a total of 25% GSK3 compounds. In this example, descriptor value intervals were determined using the linear binning scheme.

(a) JECPs for $IC_{50} \leq 1\,\mu$M compounds.

support	JECP
0.37	{PEOE_VSA-3:(7.32,10.98]}
0.35	{PEOE_VSA+3:(5.94,11.88], a_nS:(−∞,0.20], vsa_don:(8.71,13.07]}
0.35	{PEOE_VSA+2:(7.94,15.89], PEOE_VSA-4:(−∞,4.72]}
0.34	{SMR_VSA2:(15.31,22.96], SlogP_VSA3:(−∞,11.06], VDistEQ:(3.19,3.44]}
0.34	{PEOE_VSA+2:(7.94,15.89], PEOE_VSA+3:(5.94,11.88]}
0.34	{PEOE_VSA+2:(7.94,15.89], vsa_don:(8.71,13.07]}
0.32	{SMR_VSA2:(15.31,22.96], SlogP_VSA3:(−∞,11.06], SlogP_VSA5:(−∞,12.28], a_nF:(−∞,0.30]}
0.32	{SlogP_VSA3:(−∞,11.07], SlogP_VSA6:(−∞,0.44], VDistEQ:(3.19,3.44], a_nS:(−∞,0.20]}
0.32	{PEOE_VSA+6:(−∞,3.84], SlogP_VSA3:(−∞,11.06], SlogP_VSA5:(−∞,12.28], VDistEQ:(3.19,3.44]}
0.32	{PEOE_VSA+3:(5.94,11.88], a_nBr:(−∞,0.30], vsa_don:(8.71,13.07]}

(b) JECPs for $IC_{50} > 1\,\mu$M compounds.

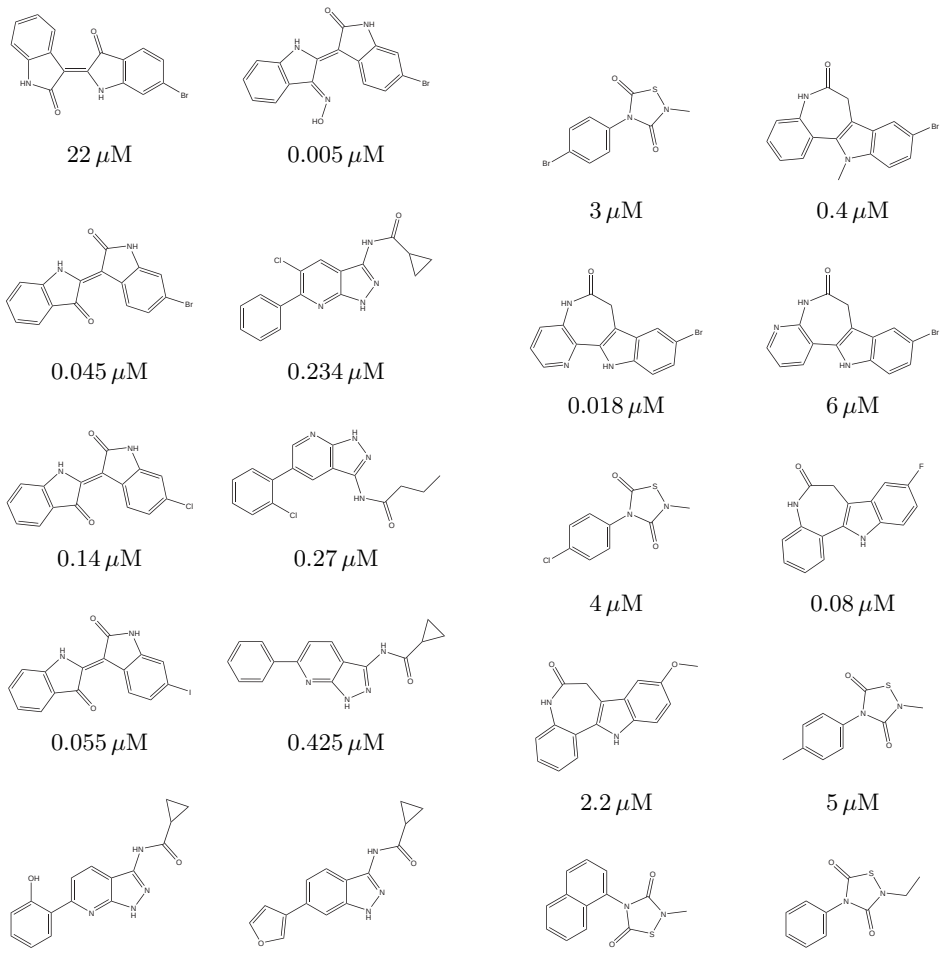
support	JECP
0.48	{SlogP_VSA0:(−∞,7.169], a_nBr:(−∞,0.3], vsa_don:(−∞,4.356]}
0.46	{a_nBr:(−∞,0.30], vsa_don:(−∞,4.36], vsa_pol:(−∞,5.17]}
0.46	{a_don:(−∞,0.60], a_nBr:(−∞,0.30]}
0.39	{SMR_VSA3:(−∞,3.35], SlogP_VSA8:(−∞,9.43], a_nBr:(−∞,0.30]}
0.39	{PEOE_VSA+3:(−∞,5.94], a_don:(−∞,0.60]}
0.39	{PEOE_VSA+3:(−∞,5.94], vsa_don:(−∞,4.34]}
0.37	{PEOE_VSA+5:(−∞,4.06], SlogP_VSA0:(−∞,7.17], SlogP_VSA8:(−∞,9.43], a_nBr:(−∞,0.30], b_triple:(−∞,0.10]}
0.37	{PEOE_VSA+4:(−∞,3.94], PEOE_VSA+5:(−∞,4.06], SlogP_VSA0:(−∞,7.17], a_nBr:(−∞,0.30]}
0.37	{SMR_VSA3:(−∞,3.35], a_nBr:(−∞,0.30],vsa_don:(−∞,4.36]}
0.37	{SlogP_VSA0:(−∞,7.17], SlogP_VSA4:(−∞,5.55], vsa_don:(−∞,4.36]}

Comparison of Classification Methods

The initial test if ECPs are able to capture the differences between nano- and micro-molar compound sets is an analysis of the JECPs generated during training from a data set containing a random sample of 25% of each nano- and micro-molar compounds of class GSK3. It has already been shown that ECPs can capture class-specific information based on type III descriptors, but this time, only 2D information is available. Table 3.7 reports the ten JECPs with highest support for both classes. In this test calculation, an unsupervised equal-width discretization scheme was applied for demonstration purposes.

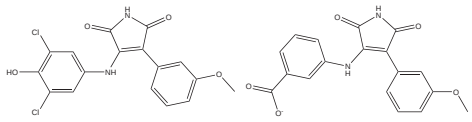
ECPs are chemically intuitive and can be easily interpreted. For example, the top JECP for the $\leq 1\,\mu$M class consists of a single descriptor range and occurs in 37% of the compounds of this class but none of the other. Another interesting JECP is the pattern {a_don:(3.6,4.2]}[4], accounting for hydrogen bond donors, having a support of 11% in the $\leq 1\,\mu$M class. In the $> 1\,\mu$M class, patterns involving descriptors counting the number of bromine atoms and hydrogen bond donors with small value ranges are prominent and occur in direct combination in the pattern {a_don:$(-\infty$,0.6], a_nBr:$(-\infty$,0.3]} with 46% support. In addition to the top 10 patterns reported in table 3.7, the number of sulfur atoms is also identified as a discriminating feature ({a_nS:(1.8,∞]} with 17% support) as is the limited polar surface area within the range 28.81–41.83, captured by the JECP {TPSA:(28.81,41.83]} with a support of 22%. Figure 3.10a shows the top GSK3 compounds predicted to belong to the $\leq 1\,\mu$M class and reveals accurate predictions; only one compound was incorrectly classified. This false-positive prediction can easily be explained when considering the training data. The numerical descriptors were divided into 10 equal-size intervals, which represents a low resolution encoding assigning exactly the same intervals to the missclassified compounds 1 and 3 in figure 3.10a. Thus, the classifier considered them identical molecules. The top 10 GSK3 compounds predicted to belong to the $> 1\,\mu$M class are shown in figure 3.10b. Here three of 10 compounds were misclassified. The ECP classifier correctly inferred that thiadiazolidinone (TDZD) derivatives have potency $> 1\,\mu$M, although the GSK3 data set only contained 29 TDZD derivatives with IC_{50} values between 2 and $100\,\mu$M. Thus, ECPs were highly discriminatory even for relatively underrepresented chemotypes. These calculations were then repeated applying information entropy based descriptor discretization. The top 10

[4]The floating-point thresholds in the descriptor value range are the result of the unsupervised discretization. The equal-width discretization algorithm has divided the value range $[0, 6]$ into ten equal-width intervals.

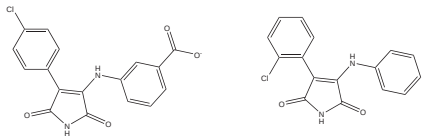


(a) unsupervised, $\leq 1\,\mu$M **(b)** unsupervised, $\leq 1\,\mu$M

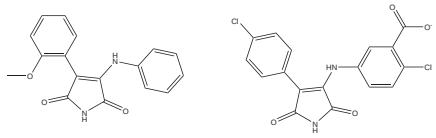
Figure 3.10: Top-ten compounds predicted to belong to the nano-molar ((a) and (c)) and micro-molar class ((b) and (d)) from a sample run of the ECP classification using a training set of 25%. (a) and (b) are the result of an unsupervised discretization of descriptors prior to JECP mining. (c) and (d) show the top ten compounds resulting from supervised (information entropy based) discretization. Potency is printed in μM.



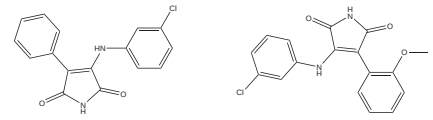
0.142 μM 0.195 μM



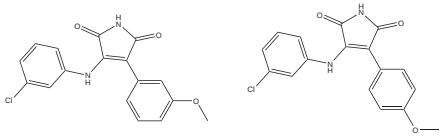
0.186 μM 0.216 μM



0.216 μM 0.109 μM

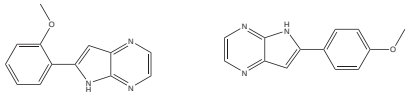


0.301 μM 0.114 μM



0.257 μM 0.156 μM

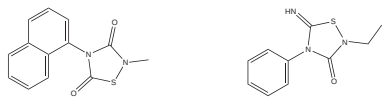
(c) supervised, $\leq 1\,\mu$M



3.3 μM 1.1 μM

3 μM 4 μM

4 μM 5 μM



3 μM 65 μM

35 μM 2 μM

(d) supervised, $\leq 1\,\mu$M

Figure 3.10: continued.

compounds for each predicted class are shown in subfigures 3.10c and 3.10d. Here the classifier achieved 100% accuracy, both for "nano-molar" and "micro-molar" compounds. These findings demonstrate the predictive value of patterns derived from descriptor settings subjected to entropy-based discretization and rationalize the usage of the supervised discretization as the primary discretization algorithm in the following comparison.

The analysis of JECPs in the last section showed that ECPs can capture class-specific knowledge and are well suited for classification of compounds based on molecular descriptors. The following paragraphs evaluate the performance of ECP-based classification (ECP) compared to a DT classifier and the BIN classification algorithm, both implemented in the MOE software package. The decision tree is constructed by a recursive procedure which adds one branch-point to the tree, similar to the supervised information-entropy based discretization method. Each branch point is a decision rule splitting the current data set into two parts based on a cut value c for one particular molecular descriptor. The descriptor and cut point c are chosen to maximize the purity of the induced partitions. The MOE implementation uses the Gini index of diversity as a measure for pureness. It is defined as $it(t) = \Sigma_x \Sigma_{y,x \neq y} p(x|t)p(y|t)$ for a node t and classes x and y. $p(x|t)$ gives the proportion of class x in node t. The procedure stops when no cut point with increasing pureness is found. The constructed decision tree can be pruned by removing nodes in the tree to increase the generalization potential. Unpruned trees are likely to be adapted to the training data, producing high accuracy on known examples, but lower accuracy on unknown test data due to overfitting. The BIN method is based on Bayesian statistics (Labute, 1999). It estimates the class probability based on estimated descriptor value distributions. Opposed to ECP and DT, it assigns a probability p for a compound to belong to the class of nano- or micro-molar compounds (with probability $1 - p$). This probability is transformed into a binary class label by a probability threshold of 0.5 for the nano-molar class. Descriptor selection for BIN classification is done prior to model creating by contingency analysis implemented in MOE.

The comparison is done on training sets of different size ranging from 10% to 50% of the available compounds. The results are reported in table 3.8. For these training sets consisting of relatively large fractions of the data sets, all three classifiers performed almost equally well, achieving overall prediction accuracy at the 80% level. In fact, it was interesting to see that the difference in performance between these methods was very small for all four activity classes. Surprisingly, performance hardly changes from 10–50%. Prediction accuracy was consistently

Table 3.8: Average prediction accuracies for five training set sizes (10–50% of each activity class) are reported for ECP, binary QSAR (BIN), and decision tree (DT) classification. The last row reports the average accuracy over all classes.

	training set size														
	10%			20%			30%			40%			50%		
class	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT
BZR	0.82	0.78	0.85	0.85	0.67	0.83	0.86	0.73	0.84	0.86	0.77	0.85	0.87	0.80	0.85
DHFR	0.64	0.66	0.62	0.67	0.69	0.65	0.67	0.70	0.67	0.67	0.70	0.68	0.67	0.71	0.68
GSK3	0.76	0.77	0.79	0.78	0.81	0.82	0.79	0.82	0.83	0.79	0.82	0.84	0.80	0.82	0.85
HIVPROT	0.89	0.78	0.84	0.89	0.84	0.85	0.89	0.86	0.86	0.89	0.87	0.87	0.89	0.88	0.86
average	0.78	0.75	0.78	0.80	0.75	0.79	0.80	0.78	0.80	0.80	0.79	0.81	0.81	0.80	0.81

lowest for DHFR (approximately 60–70%) and highest for HIVPROT (approximately 80–90%), irrespective of the methodology.

The former evaluation showed that ECP classification is comparable to standard chemoinformatics methods. The next paragraph focuses on predictions based on very small training sets of 10 or fewer compounds. It was thought that ECP should be well suited to operate under such unusual training conditions because of the high level of resolution emerging chemical patterns displayed in compound ranking. Thus, it should be possible to extract discriminatory patterns from only a few compounds. Therefore, for each activity class training sets of 10, 5, or 3 compounds were assembled from each of the $\leq 1\,\mu$M and $> 1\,\mu$M potency classes by random selection and predicted the class label of the remaining compounds. The results are presented in table 3.9. Under these challenging training conditions, ECP performed better than BIN or DT classification, in three cases still reaching approximately 80–90% average prediction accuracy for training on only three compounds. Overall, ECP performed best on three compound sets and BIN on one. For the smallest learning sets, decision tree classification lost any predictive ability, because it classified all compounds as highly potent, yielding an artificial prediction accuracy of 100% for the $\leq 1\,\mu$M class, but 0% for the $> 1\mu M$ class. For three of our classes, BIN showed slightly better results for prediction of compounds from the $\leq 1\,\mu$M class than ECP. However, for all classes, the prediction accuracy for compounds from the $> 1\,\mu$M class was clearly below random, thus indicating the presence of systematic prediction errors. Thus, BIN could not be applied in a meaningful way to these test cases when only three compounds were used for training. For training sets composed of five compounds, where apparent systematic prediction errors were absent, the prediction accuracy of ECP was

Table 3.9: Average prediction accuracies are separately reported for both potency classes and very small training sets of three, five, or 10 compounds from each class. Abbreviations are used according to table 3.8.

	training set size								
	3			5			10		
class	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT
				a) $\leq 1\,\mu$M					
BZR	0.62	0.72	1.00	0.75	0.58	0.57	0.74	0.57	0.59
DHFR	0.54	0.68	1.00	0.72	0.54	0.58	0.73	0.71	0.59
GSK3	0.57	0.74	1.00	0.80	0.64	0.68	0.82	0.51	0.69
HIVPROT	0.79	0.73	1.00	0.78	0.65	0.63	0.81	0.57	0.66
				b) $> 1\,\mu$M					
BZR	0.88	0.45	0.00	0.75	0.55	0.64	0.79	0.58	0.63
DHFR	0.75	0.39	0.00	0.55	0.55	0.50	0.59	0.44	0.50
GSK3	0.86	0.44	0.00	0.68	0.52	0.65	0.72	0.70	0.62
HIVPROT	0.57	0.45	0.00	0.61	0.52	0.59	0.65	0.62	0.63

significantly higher than for DT and BIN calculations. Thus, overall only ECP calculations displayed consistent predictive ability for very small training sets.

3.3 Discussion

This chapter established the potential of using ECPs as a data mining tool on molecular data. ECPs were calculated from a data set of crystallographic and computed energy-minimum conformations and used to systematically detect compound class-specific features that differentiate bioactive and modeled compound conformations. As a first result, strain and bond-stretch energy penalties were found in computed conformations. These finding where consistent with earlier results and can be rationalized by induced changes in the ligand conformations upon ligand binding. However, signature patterns of individual classes contained more information and accounted for class-specific differences. For each of 18 different compound classes, several highly discriminatory patterns were identified based on a pool of 67 type III (3D) descriptors. Signature patterns typically consisted of only one to three descriptor value ranges. Signature patterns could be well rationalized on the basis of available structural data and used to distinguish between hypothetical and experimentally observed protein-ligand interactions. This shows that ECP data mining is well suited for biological applications. Conformational

changes are related to changes in the interaction network of the protein-ligand complex. These differences place constraints on the binding conformations which are reflected by corresponding patterns. This directly relates to the first research question posed in the introductory chapter. For the discrimination between bioactive and modeled conformations, this question could be answered by relating key patterns to changes in the protein-ligand interaction network induced by the conformational differences reflected by the patterns.

In addition to distinguishing between experimental and modeled binding conformations, it is conceivable that descriptor pattern analysis can also be used to identify binding conformations of novel active compounds and targets for which experimental conformations of other ligands are already available. If no experimental conformations are available, binding conformations of other ligands can not be predicted. A prerequisite for the success of this analysis is that underlying protein-ligand interactions are similar. ECP knowledge mining and classification could be used to improve 3D structure generation programs in two ways. First, it is possible to amend rule-based structure generation programs by class-specific rules generated from 3D ECPs. Second, an ECP-based classifier could be used to refine a set of computed conformations by eliminating conformations which do not contain key patterns as a pattern filter. This would require that the known conformations have the same binding mode as the compound for which a new conformation is to be generated.

In the future, it is hoped that the positive results of ECP knowledge mining can be repeated on data from docking calculations. Finding discriminating patterns between bioactive conformations and conformationally different computed conformations from docking programs would be of great benefit for further development of docking algorithms. Similar to 3D structure generation, patterns could be used as filters after the docking process to remove conformational artifacts or directly be incorporated into the docking algorithm, e.g. to modify the scoring function.

In a next step after knowledge mining, a second experiment has shown that the encouraging results of ECP-classification from data mining in bioinformatics applications could be transferred to chemoinformatics applications. The experiments have shown that classification based on JECPs performs as well as BIN and DT based classification in the prediction of relative compound potencies, directly answering research question two. The ability to handle very small training sets is the distinguishing feature of ECP classification compared to established methods. Even sets containing only three to five highly active and weakly active (or inactive) compounds are sufficient to build a predictive ECP model which can then

be used to identify other potent molecules or predict the potency level of newly designed compounds. These features make the ECP approach attractive to aid in analogue design during early stages of lead optimization where molecular information is usually rather limited and where often too few molecules with different potencies are available to build conventional QSAR models. These early stages often include the design of combinatorial libraries based upon a small number of known active compounds, which are then evaluated experimentally. An ECP classifier, trained from a small set of known active compounds, could easily be incorporated into a combinatorial library design process to remove weakly active compounds, resulting in a smaller library which is then evaluated. In this way, experimental resources for testing the compounds can be saved by cherry picking and a larger region of the chemical analogue space can be inspected (Bleicher et al., 2003).

Traditionally, activity prediction is the domain of QSAR methods. Different from classical QSAR approaches, ECP does not attempt to predict actual activity values. Rather, it is trained to predict compounds to be active above or below a predefined potency threshold level and is thus conceptually more similar to binary QSAR analysis. The ECP classifier extends the feature space of the compounds from single descriptors to a space constituted by the combination of class-specific descriptor value range combinations. This leads to a higher resolution and also increases the features available for classification. This increased feature space could be one of the reasons for the superior performance on very small training sets. The general nature of the ECP classifier makes it also easy to adopt the procedure for classifying compounds not only for potency, but also for different properties, e.g. toxicity or absorption, distribution, metabolism, and excretion (ADME) properties. A predictive model early in the process of drug design would be a valuable tool for reducing the fail ratio for developing drugs. The good performance in predicting potency levels indicates that ECP could also be used as a predictive tool for other properties with high accuracy.

Recently, it has been shown that EP classification can also benefit from ensemble classification. Here, ensembles of many classifiers are trained on samples subsets of the available training data set and the output of all classifiers is combined into a single prediction. With small changes in the scoring function, Fan et al. (2006) show that the accuracy improves significantly on 14 of 27 test sets when using an ensemble of 51 EP classifiers. This suggests that ECP classification would also benefit from ensemble classification and perform significantly better than using a single ECP classifier, at the cost of additional training and classification time.

However, on very small training sets, methods such as boosting (Freund, 1990; Schapire, 1990) and bagging (Breiman, 1996) cannot be used because they are based on building ensembles of classifiers using statistical samples selected from the available training data. In light of the superior performance of ECP classification on very small training sets, this certainly is a small drawback. In addition to ensemble classification, the presented classification could also benefit from the inclusion of ECPs with finite growth rate. The conformation mining results suggest that patterns with large but finite growth rate have high information content and would provide relevant knowledge about the differences in the two data sets to the classifier and improve classification accuracy.

Despite its encouraging performance, the ECP approach also has some general limitations. First, ECP classifications are qualitative in nature, and the computed scores do not reflect absolute differences in potencies and are thus not appropriate for potency-based compound ranking. Furthermore, for large training sets, ECP calculations become computationally expensive. Mining ECPs is an NP-hard problem, meaning that the computational time grows exponentially with the number of compounds and descriptors used. Finally, ECP predictive accuracy is influenced by the degree of similarity of training compounds. If the compounds are highly similar, the probability is high that many descriptors will be discretized into only one range and thus be eliminated by the information-based discretization technique, which in turn makes it difficult to identify highly discriminatory patterns. However, this problem also reflects a unique strength of the ECP approach, its ability to derive sound predictive models on the basis of few compounds having different structural features yet similar activity. This suggests that ECP can be successfully applied to predict highly active compounds with significant structural modifications compared to learning set molecules, which also distinguishes the ECP approach from QSAR-type methodologies.

4 Integration of Virtual and High-Throughput Screening

Virtual and high-throughput screening are complementary in nature (Bajorath, 2002; Parker and Bajorath, 2006). Bleicher et al. (2003) argue that an iterative approach which combines HTS and VS strategies is superior to an individual HTS or VS setting where large compound libraries are screened without guidance or feedback from intermediate steps. Each step in an iterative approach provides rapid feedback and new information for the following step. In this way, computational and experimental resources can be adapted and focused on promising parts of the chemical space, resulting in close cooperation of computational scientists and chemists or biologists.

A formal framework integrating VS and HTS methods is provided by so-called "iterative" or "sequential" screening[1] schemes (Engels and Venkatarangan, 2001; Parker and Bajorath, 2006), briefly illustrated in figure 4.1. The underlying idea is to enrich small database subsets with novel hits and establish an iterative computational and experimental screening protocol. Sequential screening integrates VS and HTS methods by alternating iterations of computational VS and experimental HTS applications. The first iteration starts with generating a computational model from known active compounds. As VS templates, already known experimental hits are used or, alternatively, sets of known active compounds from patents or the literature. Depending on the available data, VS methods such as cluster analysis, QSAR, partitioning or similarity searching are then applied to preselect small subsets from large compound libraries for experimental evaluation. The subsets are experimentally screened and newly identified hits (or inactives) are taken into account as additional information for model refinement during subsequent rounds until a sufficiently large number of additional hits are obtained or a maximum number of experimental evaluations is reached (Bajorath, 2002).

Combining VS and HTS in sequential screening can dramatically reduce the number of compounds that need to be tested. Engels and Venkatarangan (2001) compare the results of several sequential screening campaigns. They suggest that

[1] The term "sequential screening" will be used exclusively in the following.

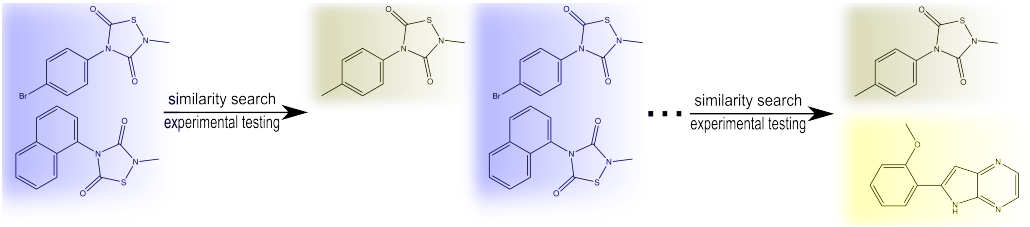


Figure 4.1: Illustration of sequential screening. Sequential screening starts with building a computational model from known active compounds. This model is then used to virtually screen a compound database to select compounds for experimental evaluation. In this example, the first similarity search finds one active compound (green) and one false-positive. In subsequent iterations, the newly discovered compounds are added to the template set and used to refine the computational model. In this way, knowledge derived from false-positive can also be included into the model-building step.

at least 50% of hits available in screening libraries can be identified by experimental testing of only 10–20% of all database compounds. Jones-Hertoz et al. (1999) used recursive partitioning and a structure-based similarity search to screen a compound collection for 14 G-protein–coupled receptors in a prospective study. Both methods are based on binary descriptors encoding structural features. They conclude that sequential screening provides a significant improvement over random selection for testing, leading to a significant reduction in HTS costs.

This chapter presents two experiment that show the potential of the integrated sequential screening scheme using an embedded ECP-based ranking or classification method. The first experiment is based on the generally good performance of ECP classifiers constructed on very small training sets. As described in section 3.2, ECP-based classifiers could be trained on very small training sets where other established methods failed. This motivated the design of an experiment simulating early stages of lead-optimization processes where the objective is to optimize sets of related compounds for activity against a given target. The experiment uses randomly selected sets of compounds, e.g. from a combinatorial library design project, to train ECP-based classifiers and to eliminate compounds with low potency in an iterative manner. The second experiment shows how ECP-based classification could be used as a ranking tool for similarity searching and simulates a sequential screening experiment using a real HTS data set.

4.1 Simulated Lead Optimization

Many drug discovery programs start with small sets of known active compounds, e.g. results from HTS campaigns, known templates from literature or patent sources. From these sets, libraries of analogues are produced which are then tested experimentally for their activity. Here, a computational method was used to guide compound selection and minimize the resources needed to produce a small set of potent compounds. Instead of experimentally testing a large number of analogues in a brute-force approach, the set is optimized in an iterative manner integrating computational and experimental efforts illustrated in figure 4.2. The idea is to start with a small random set of active compounds, e.g. HTS hits. This small set is then divided into highly and weakly potent compounds using the median potency as a threshold, and an ECP classifier is trained on this set to distinguish between the two classes. The ECP classifier is then used to remove weakly potent compounds from the remaining library, resulting in a reduced set of hopefully more potent compounds. The procedure is repeated on the reduced library for ten iterations to produce a small set of highly potent compounds.

Figure 4.3 shows the results for the four classes used in the ECP classification evaluation described in section 3.2. For each activity class and training set size, the calculations consistently reached convergence during the first eight or fewer iterations, producing selection sets of less than 10 compounds with average potency in the sub-micromolar range. Training sets of five or 10 compounds produced comparable results. During the first few iterations a sharp decline in compound numbers was observed for all classes accompanied by significant reduction in average IC_{50}. Observed potency enrichment ranged from one (GSK3) to three orders of magnitude (HIVPROT). For comparison, the same lead optimization simulation procedure was implemented for DT and BIN as classification technique, and the results are presented in appendix B. The observed potency enrichment significantly differed from ECP. As expected, the DT classifier could not be used in a meaningful way for only five training compounds because it classified the entire test set to belong to one class, either eliminating all compounds, if classified as weakly potent molecules, or retaining all compounds for the next iteration, if classified as highly potent ones. Compared to BIN, ECP produced much better potency enrichments on all classes by at least one order of magnitude for classes HIVPROT and GSK3. For these classes, binary QSAR calculations failed to produce a median potency value below 1 μM. On training sets with 10 compounds, ECP also performed better than the other methods. Only for GSK3, the decision

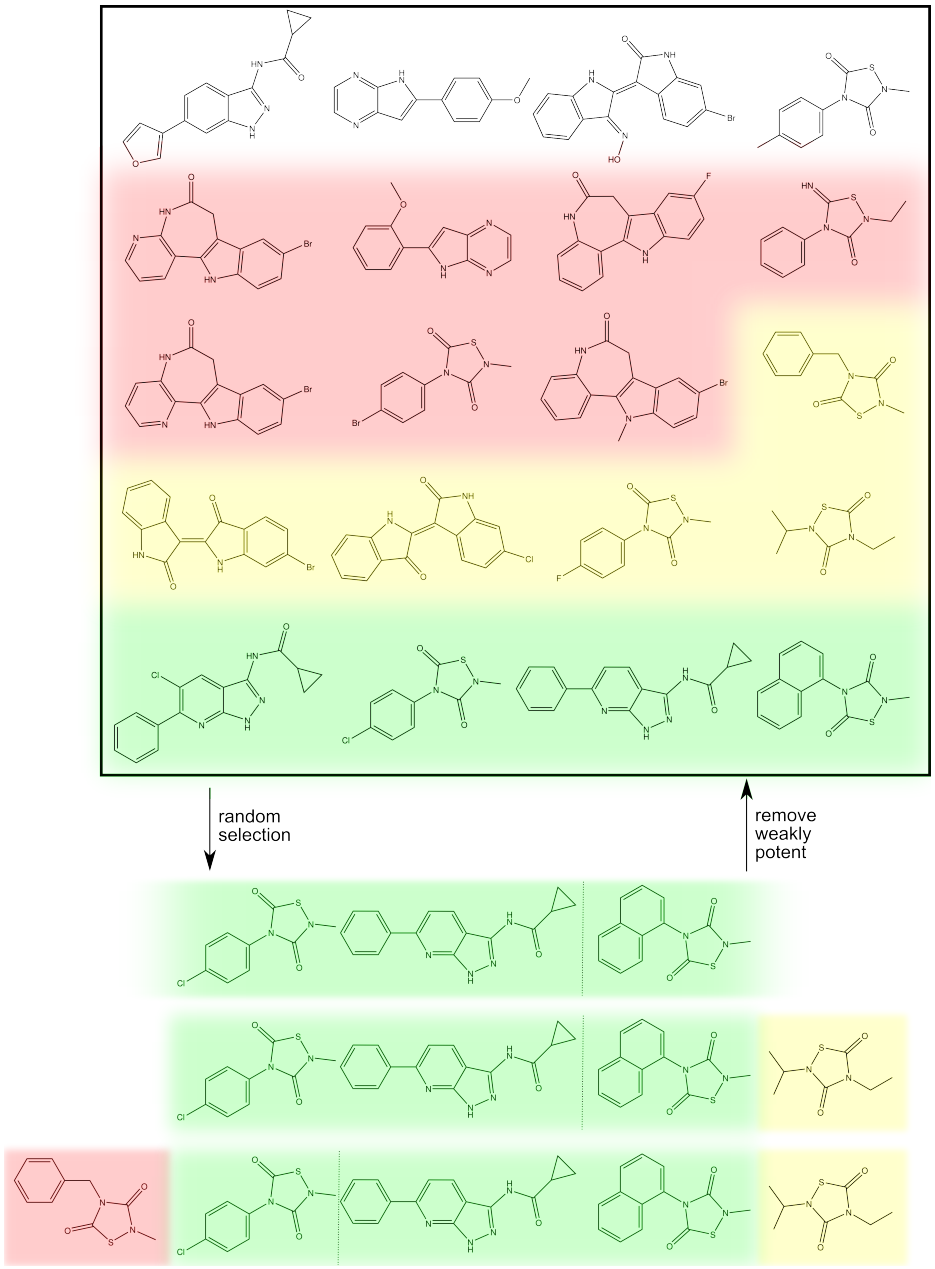


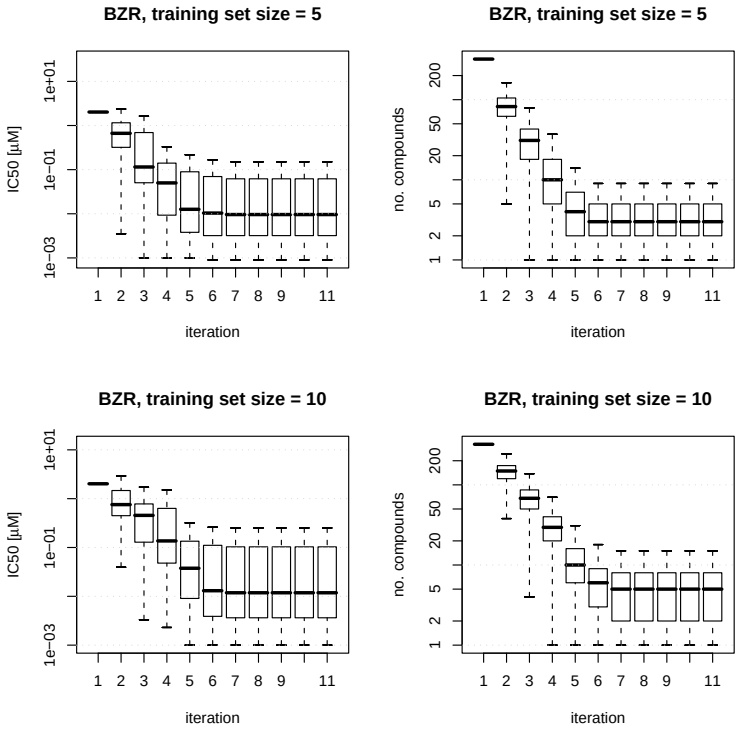
Figure 4.2: Three iterations of the simulated lead optimization process. In each iteration, a small random subset from the compound library (top) is selected and evaluated. The compound set is divided into two parts based on the median potency (dashed line). All selected compounds are used as a training set to train an ECP classifier which then removes all compounds from the library classified as weakly active. In the first iteration (green), three compounds are selected for training. These three compounds are used to train an ECP classifier which distinguishes highly potent (potency above the median) from weakly active compounds, removing one compounds from the library. In the second iteration (yellow), one compound is randomly selected and added to the training set. The training set is split in two parts and four compounds (yellow) classified as weakly active are removed. The final iteration adds one compound to the training set which turns out to be highly active. Finally, six compounds (red) are removed from the library and a set of four untested compounds remains.

tree classifier achieved a potency enrichment close to ECP. For HIVPROT, only ECP produced potencies below 1 μM, with a sharp increase in average potency from the third iteration on.

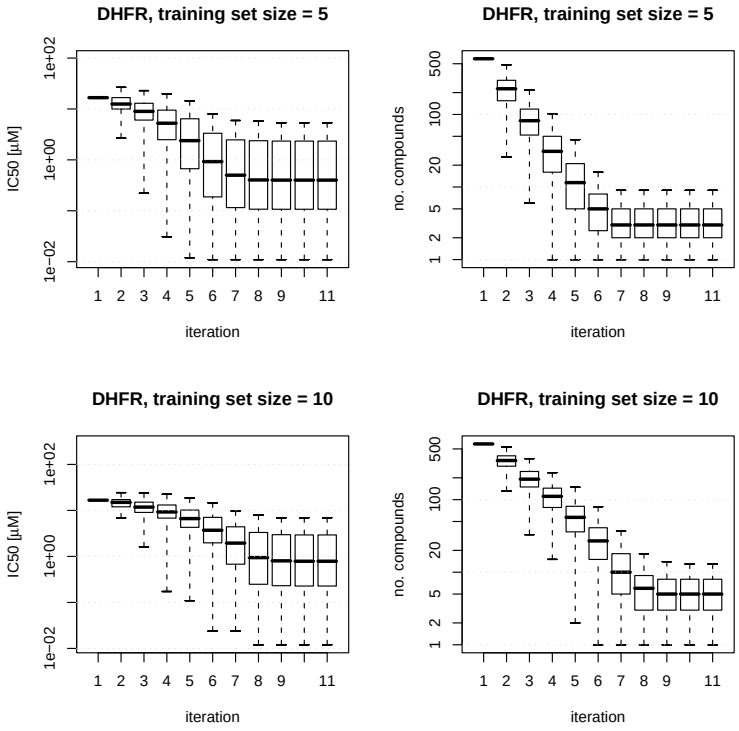
The iterative classification scheme devised for ECP led to significant enrichment of potent compounds in small selection sets, without the need for careful training set assembly. This approach should have considerable potential for the analysis of analogue or target-focused libraries (Schnur et al., 2004) that can be enumerated in-silico using different design protocols (Rose and Stevens, 2003) and the selection of preferred subsets or single molecules. Precomputed libraries focused on known active compounds can be subjected to iterative ECP classification in order to identify library subsets enriched with compounds predicted to be most potent.

4.2 Simulated Sequential Screening

Having demonstrated the potential of integrating computational ECP-based VS methods and experimental screening, it is straightforward to implement the full sequential screening protocol using ECP-based VS. To do so, the binary classification method must be transformed into a ranking method that is able to select promising compounds from large library. This section describes this transformation and an evaluation of the protocol in a simulated sequential screening experiment. The availability of real HTS data sets enables the design of studies which mimic prospective studies instead of doing retrospective benchmarking on data sets. The experimental HTS step of sequential screening can be simulated and thus the experiment is more reliable than using an artificially constructed benchmark set combining known active compounds with database compounds with unknown activity. The analysis of real HTS data has two additional intrinsic advantages. First, HTS data are generally prone to noise and experimental errors (Parker and Bajorath, 2006), which is not the case in the standard VS benchmark scenario. For example, the DHFR data set used in the experiment has been screened two times, and the number of detected inhibitors has been corrected from 62 to 32. Since these errors are an intrinsic problem of HTS methods, an experiment simulating HTS steps as an integrated part of a sequential screening campaign should also use error-prone HTS data. The second benefit is the experimental validation of the whole screening database. Database compounds are commonly regarded as inactive decoys in VS benchmarks, although large compound databases are very likely to contain compounds active against the given target. Using HTS data makes it possible to use experimentally validated inactives for training and evalu-

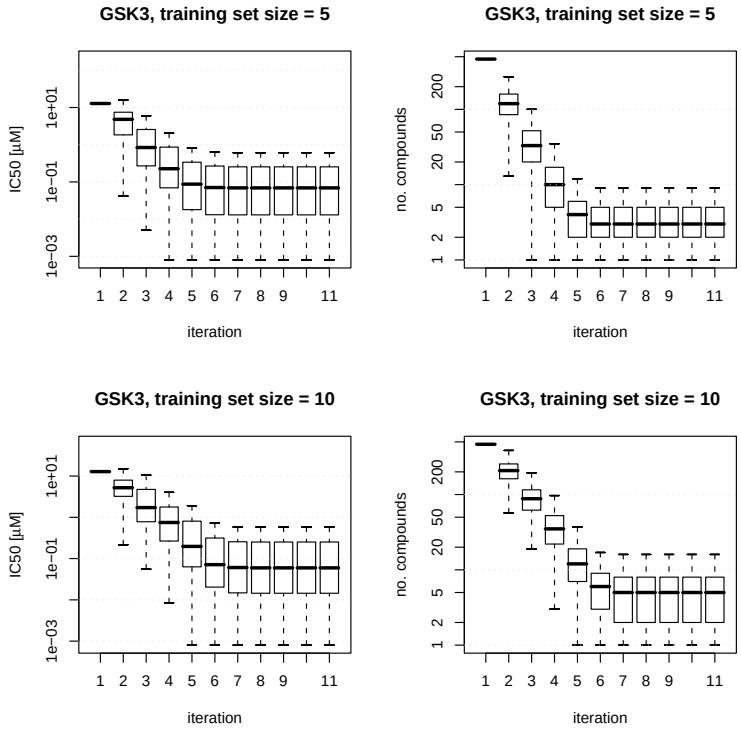


(a) BZE

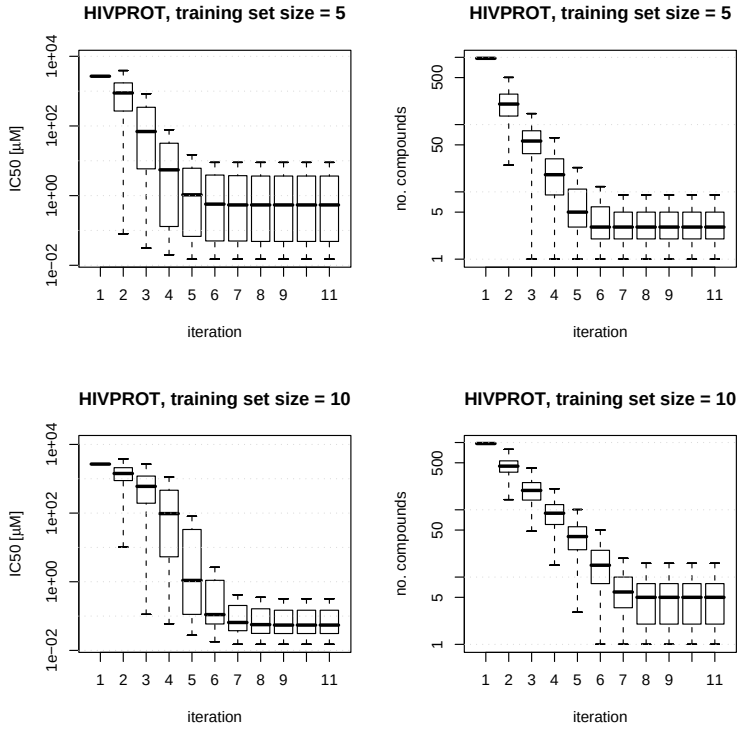


(b) DHFR

Figure 4.3: Simulated lead optimization trials. Average potencies of the compound set (left) and compound numbers (right) over 500 calculations of 10 iterations each are reported for training sets of five and 10 compounds. Potencies and compound set sizes show a rapid development to small nano-molar compound sets.



(c) GSK3



(d) HIVPROT

Figure 4.3: continued.

ation. All these aspects make simulations on HTS data more similar to practical screening applications.

4.2.1 Screening Calculations

Data Set The design of the simulated sequential screening experiment differs from the usual virtual screening benchmark scenario in several ways. First, sequential screening is concerned with the reduction of experimental efforts opposed to maximizing recovery and hit rates, which is the scenario in VS benchmarks. Because the experiment simulates the integration of computational VS and experimental HTS methods, it is important to simulate the experimental HTS step. Typical VS experiments are constructed from databases as decoys and utilize optimized literature compounds or even known drugs as potential hits. This approach has two drawbacks for simulated sequential screening. Literature hits are usually optimized with respect to pharmacological, biological, physicochemical and other properties and differ substantially from non-optimized screening hits obtained in typical HTS campaigns. Keserü and Makara (2006) estimate the similarity of HTS hit and lead compounds to only 0.61 based on MACCS keys and the Tanimoto similarity coefficient[2], indicating a rather low similarity for hits and leads. This increases the gap between VS benchmarks and HTS experiments. As the goal of this experiment is to simulate a sequential screening campaign, it is important to use a data set of real screening hits as active compounds instead of literature leads. To overcome these problems, the simulated sequential screening experiment is based on a real HTS data set consisting of 50,000 test compounds produced in the search for novel inhibitors against the well-known enzyme and drug target dihydrofolate reductase (DHFR) (Zolli-Juran et al., 2003). DHFR catalyzes the reduction of dihydrofolate to tetrahydrofolate. Inhibitors of this step are widely used as anti-infective, anti-neoplastic and anti-inflammatory drugs (Schweitzer et al., 1990), including wll-known examples such as methotrexate. DHFR is one of the best studied therapeutical targets and has been an active area of research for more than fifty years (Hitchings, 1989). Still, the development of new DHFR inhibitors continues. DHFR inhibitor development is driven by drug resistances, e.g. in tumors and malaria and the need for optimized target selectivity. Among the 50,000 compounds there were a total of 32 confirmed competitive DHFR

[2]The MACCS keyed fingerprint is a structural fingerprint representation of molecules consisting of 166 substructures. Each key is assigned one bit in a binary vector (fingerprint). Similarity between two molecules is then computed as the similarity between two fingerprints, given by the Tanimoto coefficient $\mathrm{tc}(A, B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}$ using the inner product between two vectors. The Tanimoto coefficient measures the overlap of set bits.

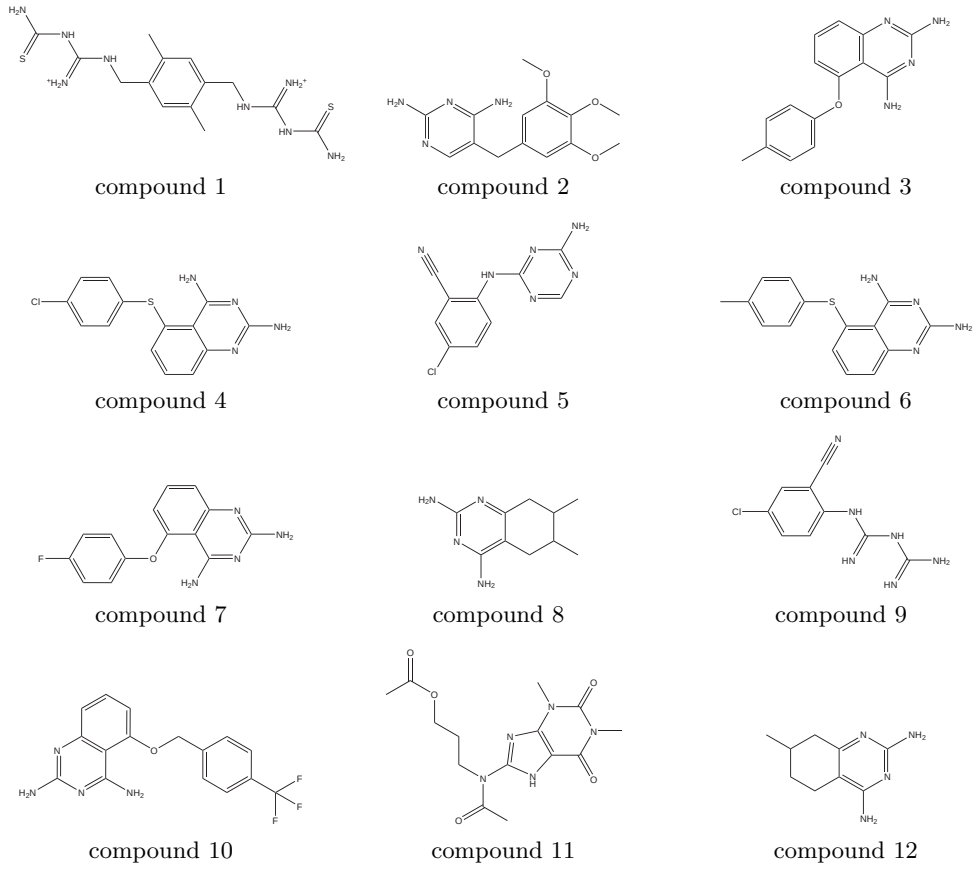


Figure 4.4: Structures of the 12 most active DHFR inhibitors. The comparison reveals the presence of different scaffold types among the HTS hits.

inhibitors with K_i values[3] ranging from 26 nM to 1 μM (Elowe et al., 2005; Zolli-Juran et al., 2003). This data set was made publicly available in the context of the first McMaster University data mining and docking competition as a training data set (Parker, 2005). Figure 4.4 shows the 12 most active inhibitors that were identified in this screen, and the structures of all inhibitors are shown in figure C.1 in appendix C. As can be seen, these compounds have diverse structures. Most of the highly active compounds shown in figure 4.4 contain a pyrimidine fragment, which is a common theme for DHFR inhibitors. From the twelve most active compounds, five (compounds 3,4,6,7 and 10) contain a quinazoline ring structure and resemble the structure of trimetrexate, an approved inhibitor of DHFR used in the treatment of pneumonia. Compound 2 is trimethoprim, an antibiotic

[3] K_i is the inhibition constant of an active compound. It gives the concentration of a ligand in a competition assay that would bind to 50% of the receptors if no competing radio-ligand were present.

drug used to treat urinary infections. The remaining active compounds are structurally diverse, including a set of seven compounds based on a methylurea linker fragment combining a 2-nitrothiophene ring with different aromatic rings. Some of the inhibitors are small fused ring systems of one aromatic heterocycle, often pyrimidine, and an unsaturated cyclohexane. Overall, the inhibitor set consists of diverse structures with different chemotypes.

Methodology The simulated sequential screening experiment uses the set of 61 uncorrelated, information-rich descriptors described in section 2.1.1. The continuous descriptors are transformed into discrete attributes by a simple discretization scheme. On small training sets, it is not possible to use the supervised information-entropy based discretization method since this method would eliminate most if not all of the descriptors. However, Eckert and Bajorath (2006) have shown that descriptor value distributions are class-selective with the majority of descriptor values found in small class-specific value ranges. These value ranges are particularly attractive for the generation of patterns. The discretization procedure utilizes the value distribution in the training set by using the mean μ and the standard deviation σ of each descriptor. Each descriptor is discretized into three bins based on the mean value: $(\mu - \sigma, \mu + \sigma)$, $(-\infty, \mu - \sigma]$ and $[\mu + \sigma, \infty)$. Descriptors which show no variation at all or a high variation are removed before discretization based on the coefficient of correlation $c_v = \frac{\sigma}{\mu}$. Descriptors without variance limit the generalization of the classifier, whereas descriptors with a coefficient of variation $c_v > 1.0$ are likely to have many irrelevant database compounds populate the activity-class specific range around the mean, which severely restricts the predictive value of such descriptors. Consequently, they are also omitted.

The sequential screening experiment goes beyond the binary ECP classification experiment described in the previous chapter. Even though ECP classification showed very high accuracy for active and inactive molecules, the uneven distribution of 32 active compounds and ∼50.000 inactive compounds induces high absolute false-positive rates, making binary classification problematic or even useless for virtual screening. To solve this problem, a ranking scheme is introduced which ranks compounds based on their accumulated support on both classes. Each compound is assigned a score which is computed by dividing the accumulated support for being active by the accumulated support for being inactive. After ranking the complete compound library, a small subset of the highest ranked compounds is selected for simulated HTS screening.

The simulated sequential screening experiment was done using the above described methods to implement the sequential screening protocol. Initially, ECP training was carried out on sets of five randomly selected DHFR inhibitors and 20 inactive compounds taken from the HTS data. Compounds were randomly selected using a random number generator such that individual runs were independent. The resulting ECP classifiers were then applied to rank all remaining compounds in the remaining HTS data set. Following the sequential screening paradigm, the top-scoring 10, 100, or 500 compounds are selected from the HTS set and examined for new hits, thereby mimicking experimental evaluation of the top-ranked compounds. From each selection set, the top-ranked 10 compounds plus all remaining hits (for selection sets of 100 and 500 molecules) were then added to the training set in order to re-build and refine the classifier for the next iteration. For each selection set size (10, 100, or 500 molecules), 100 individual trials using different training sets were carried out in order to produce a statistically relevant sample. In each case, a total of nine sequential screening iterations were carried out such that the maximum number of "tested" compounds was smaller than 10% of the entire HTS data set for the largest selection set of 500 compounds.

The experiment provides an interesting and challenging scenario to investigate whether ECP calculations are sufficiently sensitive to select the very small number of active molecules from a large number of inactive database compounds. Each run of the sequential screening protocol starts with only five active compounds, a number too small for typical classification procedures used in virtual compound screening. Since the HTS set contains 32 actives out of 50.000 compounds, the database contains only 27 possible hits.

For each selection set size, 100 independent calculations with different training sets are evaluated in order to (a) determine the top performance and potential of the ECP methodology and (b) estimate the expected performance level in iterative screening independent of the composition of learning sets. This was done because compound classification calculations are generally much influenced by compound-class specific features and training set composition (Bajorath, 2002; Parker and Bajorath, 2006). For ECP, this analysis was particularly relevant since only five active compounds were used for training, which put high weight on the characteristics and contributions of each individual molecule. The experiment provides a good test of the ability of the ECP classifier to generalize from a very small set of training samples.

The remaining part of this section presents the results of the simulated sequential screening calculations by analyzing statistical and qualitative features of the computed patterns and by evaluating the performance of the screening protocol in terms of recovery rates. Special attention is paid to the dynamic interaction between computational and (simulated) screening methods.

4.2.2 Pattern Distribution and Composition

The quality of the ECP classifier relies on its ability to derive discriminating patterns from the training data. These patterns must reflect the properties of the training data, but also be general enough for prediction of unknown test compounds. The composition of the patterns is also interesting from a knowledge mining point of view. Knowledge about the used descriptors, including their number, type, complexity and about the pattern size and distribution are of significant interest.

A quantitative analysis of the pattern composition was done based on the simulated sequential screening protocol described above. On average, 29 to 30 (of 61 available) descriptors were utilized in each ECP calculation, thus only about half the descriptor basis set. Therefore, large numbers of descriptors were not required for pattern derivation. For active compounds in the learning set, on average ∼10,700 patterns consisting of 7.5 descriptor value pairs were produced. By contrast, for inactive compounds on average only ∼170 patterns emerged with 3.3 descriptor value pairs per pattern. Thus, active compounds generated significantly more and larger patterns than inactive ones. These findings can be rationalized by considering that during learning each active compound must be distinguished from all inactive molecules and vice a versa. Since there were considerably more inactive than active compounds in each training set, active molecules required more descriptors to be distinguished from inactive compounds. Therefore, the large difference in the number of patterns between active and inactive molecules is due to the fact that the number of potential patterns grows exponentially in descriptor spaces of increasing dimensionality. It follows that the deliberately unbalanced composition of the learning sets was reflected in large differences in the numbers of “active” and “inactive” patterns, consistent with expectations. For iterative screening applications, learning sets should contain more inactive than active compounds because many more inactive molecules are available.

Given the large amount of computed patterns, an in depth analysis of the pattern composition is a demanding task. For the sake of simplicity, the analysis is limited to the discretization step of the simulated sequential screening protocol.

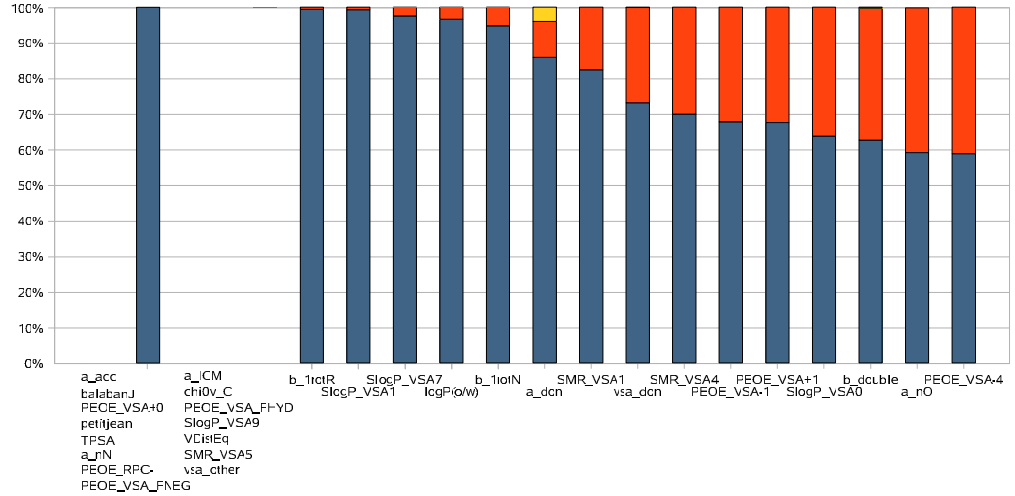


Figure 4.5: Descriptors used in more than 50% of the simulated sequential screening iterations. For each descriptor, the percentage of all iterations it has been used in (blue) and the percentage of all iterations where it has been removed due to a large coefficient of variation (red) or because of no variation (yellow) is shown. The first 15 descriptors are used in all iterations and merged into a single bar.

The discretization procedure removes descriptors with no variation and descriptors with a high coefficient of variation. Figure 4.5 shows the statistics for the 30 descriptors which are used in at least 50% of all iterations. The analysis shows that 12 out of 30 prominent descriptors are rather simple atom/bond counts, pharmacophore properties or adjacency measures. These descriptors are pure 2D (type II) descriptors. Two estimated physical property descriptors (TPSA and logP(o/w)) are included in almost every iteration. Most of the more complicated implicit 3D descriptors are removed in most iterations, with notably exceptions of descriptors relating to partial charge (PEOE_VSA_+0, PEOE_RPC-, PEOE_VSA_FNEG). The importance of molecular refractivity and the octanol/water partitioning coefficient is further emphasized by the prominence of implicit 3D descriptors relating to atomic contributions to these properties.

4.2.3 Screening Performance

The global analysis statistically estimates the performance of ECP simulated sequential screening based on 100 independent trials. Since each trial is based on different, randomly sampled selection sets, learning sets which result in classifiers with different performance in the first iterations are taken into account. In some cases, the initial classifier fails to find any active compounds, but improves by

adding more information about inactive compounds in subsequent runs. The results are reported in figure 4.6 and table 4.1. Figure 4.6 summarizes the recovery rate of 100 trials with selection set size of 10, 100 and 500 compounds. Table 4.1 additionally shows the training set size, number of virtually tested compounds and the absolute number of found active compounds. For selection sets of 10, 100, and 500 database compounds, average recovery rates of 19%, 26%, and 39% were observed, respectively. For the smallest selection set, this corresponds to the identification of approximately five hits when evaluating only 115 database compounds (∼0.2% of the HTS set) and for the largest selection set, 10 to 11 hits based on the evaluation a total of 4525 compounds (∼9%). In terms of enrichment factors, the experiments with selection set size 10 resulted in an enrichment of 83.7 compared to random selection of 115 compounds from the complete screening set. The enrichment factors for the other two experiments decreased with the number of virtually tested compounds. For selection sets of size 100, the experiments estimated a 14-fold enrichment over random selection, and for the largest selection set size of 500, only a 4.3-fold enrichment was observed.

The steady increase in average recovery rates over the nine iterations indicated that additional hits could be retrieved by adding more screening cycles. An attractive feature of iterative ECP calculations was the enrichment of hits among selection sets of only 10 compounds, where evaluation of 115 database compounds was sufficient to produce on average five hits, given initial learning sets containing only five active compounds. The high enrichment in these experiments and the small number of evaluations during each iteration render this setting an attractive candidate for extending the number of iterations. On average, selecting 10 times more database compounds gave two additional hits and testing 50 times more compounds doubled the number of recovered hits relative to the smallest selection sets. These results were well in accord with the previous findings that ECP was capable of successfully operating on the basis of very few active compounds and revealed an additional aspect, the presence of high sensitivity and specificity of ECP calculations, especially for small compound selection sets: most of the recovered active compounds are highly active. Investigating the recovery of the subset of the 12 most potent hits in the DHFR HTS set shown in figure 4.4 revealed that the final iteration for selection sets of 10, 100, and 500 compounds contained on averaged about four, five or six of these hits, respectively, showing a clear tendency to select highly active compounds from the screening data set. Thus, in small selection sets, the largest relative enrichment of potent hits was observed. Taken together, these findings indicate that ECP simulated sequen-

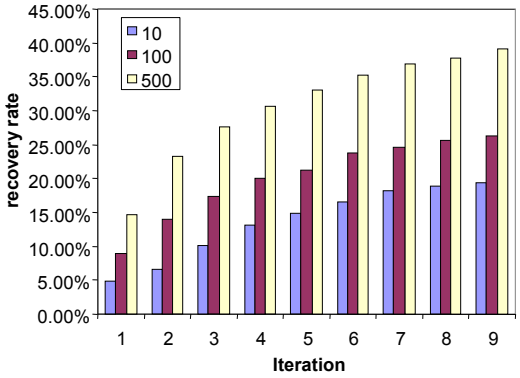


Figure 4.6: Average recovery rates over 100 independent screening trials.

Table 4.1: Average ECP results over 100 independent screening trials. Averages are reported for all nine iterations and selection sets of 10, 100, and 500 database compounds. The total number of tested database compounds ("TS") is given and "RR" provides averaged cumulative recovery rates. Averages are reported for all nine iterations and selection sets of 10, 100, and 500 database compounds. Column "It." counts the iteration, TS states the size of the test set in each iteration, "TR" the training set size, "ACT" the number of found actives and RR[%] the recovery rate of found actives in percent.

	10				100				500			
It.	TS	TR	ACT	RR[%]	TS	TR	ACT	RR[%]	TS	TR	ACT	RR[%]
1	35	35	1.32	4.9	125	36.00	2.41	8.9	525	37.40	3.94	14.6
2	45	45	1.78	6.6	225	46.84	3.78	14.0	1025	49.09	6.29	23.3
3	55	55	2.73	10.1	325	57.26	4.69	17.4	1525	60.01	7.47	27.7
4	65	65	3.53	13.1	425	67.74	5.41	20.0	2025	70.61	8.26	30.6
5	75	75	4.00	14.8	525	78.00	5.80	21.3	2525	81.24	8.94	33.1
6	85	85	4.50	16.6	625	89.00	6.40	23.7	3025	91.80	9.52	35.3
7	95	95	4.90	18.2	725	99.00	6.60	24.6	3525	102.15	9.95	36.9
8	105	105	5.10	18.9	825	109.00	6.90	25.6	4025	112.40	10.20	37.8
9	115	115	5.20	19.4	925	119.00	7.10	26.3	4525	122.71	10.60	39.1

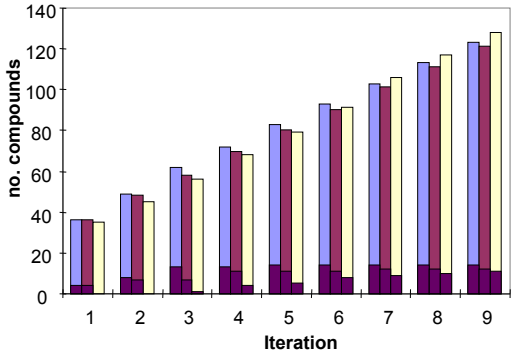


Figure 4.7: Hit rate for three simulated sequential screening trials using a selection set of 10 compounds. Each color indicates one independent trial, with dark segments in each bar showing the number of found active compounds. ECP virtual screening fails to select active compounds in the first three iterations of trial three. However, the iteratively refined classifier finds active compounds in subsequent iterations.

tial screening produced recovery rates that were at least comparable to expected results of 40% recovery while testing approximately 10% of the database (Engels and Venkatarangan, 2001) when clustering or other classification methods were used for sequential screening. However, ECP calculations already recovered approximately 20% of available hits when only about 100 of 50,000 screening set compounds were evaluated, and these hits were among the most potent ones available in the HTS set. Therefore, the application of ECP is thought to further reduce compound selection set sizes in iterative screening trials, which adds to the sequential screening paradigm.

Selection set size influenced the calculations because different numbers of active molecules were added to the training sets during each of the nine iterations. Selection sets of increasing size, from 10 to 100 and 500 compounds, typically contained an increasing number of hits that were added to the training set for the next iteration. The more active compounds are available, the better the basis for training becomes.

4.2.4 Dynamic Interaction of VS and HTS

The dynamic interaction between computational virtual screening and experimental screening methods cannot be investigated in a global statistical analysis of the screening performance of simulated sequential screening. In order to investigate if and how the ECP classifier benefits from the feedback provided by the simulated HTS experiments, individual runs of the simulated sequential screening experiment were analyzed. Tables 4.2 to 4.4 report the top ten ECP trials for

Table 4.2: Top 10 ECP trials for selection sets of 10 database compounds. "TR" reports the size of the training sets, "ACT" reports the total number of active compounds retrieved (with newly identified ones added to the training set for the next round), and "RR" reports the cumulative recovery rates of active compounds.

		Trial									
Iteration		1	2	3	4	5	6	7	8	9	10
1	TR	35	35	35	35	35	35	35	35	35	35
	ACT	3	3	3	3	3	3	3	3	3	3
	RR[%]	11.1	11.1	11.1	11.1	11.1	11.1	11.1	11.1	11.1	11.1
2	TR	45	45	45	45	45	45	45	45	45	45
	ACT	5	5	5	5	5	5	5	5	5	5
	RR[%]	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5
3	TR	55	55	55	55	55	55	55	55	55	55
	ACT	5	5	5	5	5	5	5	5	5	5
	RR[%]	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5	18.5
4	TR	65	65	65	65	65	65	65	65	65	65
	ACT	7	7	7	7	7	7	7	7	7	7
	RR[%]	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9
5	TR	75	75	75	75	75	75	75	75	75	75
	ACT	7	7	7	7	7	7	7	7	7	7
	RR[%]	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9
6	TR	85	85	85	85	85	85	85	85	85	85
	ACT	7	7	7	7	7	7	7	7	7	7
	RR[%]	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9
7	TR	95	95	95	95	95	95	95	95	95	95
	ACT	8	8	8	8	8	8	8	8	8	8
	RR[%]	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6
8	TR	105	105	105	105	105	105	105	105	105	105
	ACT	9	9	9	9	9	9	9	9	9	9
	RR[%]	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
9	TR	115	115	115	115	115	115	115	115	115	115
	ACT	9	9	9	9	9	9	9	9	9	9
	RR[%]	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3

Table 4.3: Top 10 ECP trials for selection sets of 100 database compounds. "TR" reports the size of the training sets, "ACT" reports the total number of active compounds retrieved (with newly identified ones added to the training set for the next round), and "RR" reports the cumulative recovery rates of active compounds.

		Trial									
Iteration		1	2	3	4	5	6	7	8	9	10
	TR	36	36	35	35	35	35	35	36	36	38
1	ACT	4	4	0	0	0	0	0	3	3	4
	RR[%]	14.8	14.8	0.0	0.0	0.0	0.0	0.0	11.1	11.1	22.2
	TR	49	48	45	45	45	45	45	47	46	49
2	ACT	8	7	0	0	0	0	0	10	4	7
	RR[%]	29.6	25.9	0.0	0.0	0.0	0.0	0.0	37.0	14.8	25.9
	TR	62	58	56	56	56	56	56	57	56	59
3	ACT	13	7	1	1	1	1	1	10	4	8
	RR[%]	48.1	25.9	3.7	3.7	3.7	3.7	3.7	37.0	14.8	29.6
	TR	72	70	68	68	68	68	68	68	66	69
4	ACT	13	11	4	4	4	4	4	11	4	8
	RR[%]	48.1	40.7	14.8	14.8	14.8	14.8	14.8	40.7	14.8	29.6
	TR	83	80	79	79	79	79	79	78	77	79
5	ACT	14	11	5	5	5	5	5	11	5	8
	RR[%]	51.9	40.7	18.5	18.5	18.5	18.5	18.5	40.7	18.5	29.6
	TR	93	90	91	91	91	91	91	88	88	90
6	ACT	14	11	8	8	8	8	8	11	7	9
	RR[%]	51.9	40.7	29.6	29.6	29.6	29.6	29.6	40.7	25.9	33.3
	TR	103	101	106	102	102	102	102	98	99	101
7	ACT	14	12	9	9	9	9	9	11	8	10
	RR[%]	51.9	44.4	33.3	33.3	33.3	33.3	33.3	40.7	29.6	37.0
	TR	113	111	117	113	113	113	113	108	111	112
8	ACT	14	12	10	10	10	10	10	11	11	11
	RR[%]	51.9	44.4	37.0	37.0	37.0	37.3	37.0	40.7	40.7	40.7
	TR	123	121	128	124	124	124	124	118	121	122
9	ACT	14	12	11	11	11	11	11	11	11	11
	RR[%]	51.9	44.4	40.7	40.7	40.7	40.7	40.7	40.7	40.7	40.7

Table 4.4: Top 10 ECP trials for selection sets of 500 database compounds. "TR" reports the size of the training sets, "ACT" reports the total number of active compounds retrieved (with newly identified ones added to the training set for the next round), and "RR" reports the cumulative recovery rates of active compounds.

		Trial									
Iteration		1	2	3	4	5	6	7	8	9	10
1	TR	39	38	39	37	38	41	43	41	36	41
	ACT	5	6	6	3	5	8	8	6	7	9
	RR[%]	18.5	22.2	22.2	11.1	18.5	29.6	29.6	22.2	25.9	33.3
2	TR	52	48	51	51	51	53	55	55	47	53
	ACT	9	6	9	10	8	10	10	11	9	11
	RR[%]	33.3	22.2	33.3	37.0	29.6	37.0	37.0	40.7	33.3	40.7
3	TR	62	58	62	64	64	64	68	65	57	64
	ACT	9	6	10	13	11	14	13	11	9	12
	RR[%]	33.3	2.2	37.0	48.1	40.7	51.9	48.1	40.7	33.3	44.4
4	TR	75	69	72	74	78	75	78	76	68	74
	ACT	12	7	10	13	15	15	13	12	10	13
	RR[%]	44.4	25.9	37.0	48.1	55.6	55.6	48.1	44.4	37.0	48.1
5	TR	90	80	84	85	88	85	90	86	82	84
	ACT	17	8	13	14	15	15	15	12	14	13
	RR[%]	63.0	29.6	48.1	51.9	55.6	55.6	55.6	44.4	51.9	48.1
6	TR	101	93	96	96	99	95	100	98	93	94
	ACT	18	11	15	15	16	15	15	14	15	13
	RR[%]	66.7	40.7	55.6	55.6	59.3	55.6	55.6	51.9	55.6	48.1
7	TR	111	105	106	107	109	106	110	109	103	105
	ACT	18	13	15	16	16	16	15	15	15	14
	RR[%]	66.7	48.1	55.6	59.3	59.3	59.3	55.6	55.6	55.6	51.9
8	TR	121	117	117	117	119	116	120	119	113	115
	ACT	18	16	16	16	16	16	15	15	15	14
	RR[%]	66.7	59.3	59.3	59.3	59.3	59.3	55.6	55.6	55.6	51.9
9	TR	132	128	128	127	129	126	130	129	123	126
	ACT	19	18	17	16	16	16	15	15	15	15
	RR[%]	70.4	66.7	63.0	59.3	59.3	59.3	55.6	55.6	55.6	55.6

selection sets of 10, 100, and 500 compounds, respectively. The results reflect the trends of the global performance analysis. For selection sets of only 10 database compounds, the best individual trials recovered 33% of the hits (table 4.2), which corresponded to nine of 27 available active compounds. The top ten trials showed the same retrieval characteristics because they identified the same sets of active compounds, despite different learning set composition (whereas other trials within the top 20 list produced different sets). During iterative screening, the number of cumulatively identified hits increased from three in the first trial to nine after the last. For selection sets of 100 compounds (table 4.3), the best cumulative recovery rate of individual trials was 52% corresponding to 14 of 27 hits. Here retrieval characteristics substantially differed and in some instances, individual trials produced good results, although active compounds could not be recovered during the first one or two iterations, as illustrated by the third experiment shown in figure 4.7. In these cases, the training sets were expanded by addition of 10 inactive compounds per iteration, which then led to the identification of hits. Adding the top-ten compounds even if they are misclassified inactives gives additional information to the training phase of the next ECP classifier, which can then use this information to generate patterns that became increasingly characteristic for active compounds because more inactive molecules needed to be discriminated. This dynamic interaction between computational and simulated experimental methods highlights the benefits of the integration of both methods in the sequential screening paradigm. These findings further illustrate the usefulness of unbalanced training sets if only a few hits were available for learning and because inactive compounds contain information for ECP classification. Unbalanced training sets (and test sets) pose serious problems to classification algorithms, but in this case, classification actually benefits from the introduced bias. Extending the training set for inactive compounds makes the ECP classifier more specific for the subtle differences in the descriptor value distribution and thus lowers the probability of matched patterns for inactive compounds.

Best recovery rates were obtained for selection sets of 500 compounds, as expected (table 4.4). Here the top ten trials retrieved between 15 and 19 hits, producing a top recovery rate of 70%. Thus, at top performance levels, iterative ECP calculations were capable of producing significant retrieval rates for selection sets of varying size.

4.3 Discussion

The integration of VS and HTS methods is explored in this chapter. ECP-based computational classification and ranking methods were used in two experiments to evaluate if and how iterative screening protocols could be used in drug discovery programs. First, binary ECP classifiers were used to optimize sets of known active compounds, a common topic in early stages of drug discovery programs, e.g. when optimizing HTS hits. It could be shown that ECP classification generally optimizes QSAR data sets of several hundred compounds with different activity distributions into small highly potent sets within a few iterations of alternating computational classification and simulated experimental screening. Compared to decision trees and binary QSAR, only ECP could be used reliably in this situation where the number of known actives was extremely small.

The simulated lead optimization experiment has shown the potential of sequential screening protocols. However, this study is based on binary classification and uses QSAR benchmark data sets instead of real HTS data. Consequently, the next step was to extend ECP classification into a ranking method for prioritizing compounds with respect to activity and design a sequential screening experiment. The publicly available DHFR screening data set made it possible to simulate the outcome of HTS screens and to perform a thorough statistical analysis of the methodology in simulated sequential screening demonstrating the potential of modern data mining techniques in pharmaceutical research. Although this screening paradigm is currently far from being established in pharmaceutical research, it is increasingly considered a complement or an alternative to brute force HTS (Bajorath, 2002; Engels and Venkatarangan, 2001; Parker and Bajorath, 2006). In simulated sequential screening trials, ECP calculations generated a steady increase in the recovery of active compounds and already produced multiple hits by iteratively selecting as little as 0.2% of the HTS data set. The efficiency of sequential screening strongly depends on the performance of the embedded virtual screening method. It must be able to generalize from small sets of known active (and inactive) compounds, fast enough to screen large compound collections and also be sensitive to changes in the training data during the sequential screening iterations. Each iteration provides the classifier with new knowledge and should result in a refined model, possibly correcting previous errors. The need for a sensitive method working on small training sets and the ubiquitous errors in experimental testing requires methods which generalize well from individual features of the training set and provide high-resolution models. ECP has been shown to be

easily implemented into sequential screening protocols and to yield high recovery rates after a small number of iterations.

Clearly, with 50,000 compounds, the DHFR test set studied here was smaller than many currently used HTS compound sets that are frequently an order of magnitude larger. However, the size of screening sets is not a limiting factor for ECP sequential screening and given the observed sensitivity and specificity of the calculations, there are no reasons to expect substantially different results if varying screening set size. ECP calculations are particularly attractive for sequential screening applications when complete recovery of available active compounds is not the primary goal of the screening efforts, but rather rapid recovery of novel hits. The results further support the view that iterative computational and experimental screening can streamline biological screening efforts and greatly reduce the experimental and data analysis requirements, including secondary assays to eliminate false-positives. If a practical ECP-supported sequential screening application on the DHFR set would have produced results similar to our simulations, it would have been possible to replace HTS analysis of this data set with series of low-throughput assays to identify multiple hits. These findings suggest that ECP analysis should merit further consideration in HTS data mining and sequential screening.

5 Summary and Conclusion

This chapter summarizes the findings of this thesis regarding the three questions presented in the introduction. First, emerging pattern (EP) mining has been adapted to chemoinformatics by using discretized molecular descriptors as nominal attributes. EPs based on these descriptors were formally defined as emerging chemical patterns (ECP) in chapter 2. Using this methodology, knowledge mining and property prediction capabilities of ECPs were explored in chapter 3. Finally, the integration of computational and experimental methods is investigated by the experiments presented in chapter 4. Chapter 3 answers questions one and two, and chapter 4 provides a possible answer to the last question. These answers are summarized in the remaining part of this section.

Question 1: How can pattern mining-based knowledge be rationalized at the molecular level?

Section 3.1 has shown that ECPs are able to capture even subtle differences between bioactive binding and computationally generated energy-minimized conformations. ECPs reflected well-known strain energy penalties on a global level. On a per-class level, ECPs were able to find more detailed differences in the two conformation sets. For each of the 18 classes investigated, discriminating patterns matching stringent support thresholds were computed. These patterns did not only contain energy descriptors, but also descriptors that can be related directly to differences in shape and surface properties.

Furthermore, these class-specific patterns could be rationalized on a molecular level. Binding conformations of protocatechuate-3,4-dioxygenase inhibitors can be easily discriminated from computer-generated conformations by looking at the extension along the third principle axis. A cation in the binding pocket of protocatechuate-3,4-dioxygenase forces the substituents to move out of the molecule's main plane into an energetically unfavorable position. This change is then reflected by the corresponding patterns. Additionally to this simple interpretation, conformational changes of ligands can be related to changes in the interaction network which is formed upon binding to the target protein. It could be shown for three exemplary classes that discriminating patterns can be re-

lated to important distortions in the binding conformations which are not found in energy-minimized conformations due to energetic penalties. This presents a strong advantage of the easy interpretation of ECPs and highlights their ability to capture class-specific information in biological applications.

Question 2: How does a pattern-based classifier perform on biological data?

ECP-based classifiers have been extensively evaluated on biological data. In section 3.2, ECP-based classification was shown to be at least comparable to standard binary classification methods for property prediction. Four different QSAR data sets were used to evaluate the accuracy of ECP-based classification, decision tree and binary QSAR property prediction methods. While all methods performed equally well when predicting the potency level of individual compounds, only ECP-based classification could be used on training sets as small as containing only three compounds from each class. This is of outstanding significance for early stages of pharmaceutical drug discovery processes where typically only a few active compounds are at hand. A classifier with good predictive capabilities in highly and weakly potent classes certainly is of great benefit in these early stages, e.g. when designing focused libraries based on active HTS hits.

In chapter 4, ECP-based iterative experiments have been introduced which further show the potential of pattern mining methods when processing biological data. The experiments show how the binary classification scheme can be extended to compound ranking. In pharmaceutical settings, early discovery of active compounds is more important than complete recovery. The simulated sequential screening experiment shows that ECP-based ranking rapidly finds hidden active compounds even in imbalanced data sets. Given the large size-difference in available experimentally tested compounds and the size of current compound databases, it is very important that data mining algorithms used in chemoinformatic campaigns can handle this biased data.

Question 3: Is it possible to integrate computational and experimental methods in an efficient way such that both methods benefit from the integration?

The complementary nature of computational (VS) and experimental (HTS) methods is explored in chapter 4. First, an iterative protocol was designed simulating compound selection tasks common to lead optimization stages in drug discovery.

This experiment iteratively combined ECP classification and (simulated) experimental evaluation of a training set. Starting from a small set of known active compounds, each of the four classes used in the classification evaluation in section 3.2 was optimized to small sets of highly potent compounds. After training an ECP classifier on the initial training set, all compounds predicted to be weakly active were removed from the remaining set. A small number of the remaining compounds was then added to the training set. It is shown that after a small number of iterations, small compound sets with potency below 1 μM could be produced with ECP classifiers, but not with decision tree or binary QSAR methods.

In section 4.2 the potential of VS and HTS integration has been investigated. It has been shown how the sequential screening paradigm can be implemented with ECP-based compound ranking. The availability of real HTS data enabled the design of an experiment which mimics real-world screening campaigns, but still offers the opportunity to do statistically valid evaluations. Experiments which use HTS data instead of artificially constructed benchmark data sets are more significant and closer to prospective studies. Typical benchmark data sets introduce a structural bias for active compounds which is not the case for HTS data sets. In addition to that, HTS experiments are generally error-prone and the simulation of HTS by using a stored outcome for simulation also keeps this intrinsic error.

The analysis of the simulated sequential screening calculations has shown that testing a small fraction of only 0.2% of the available compound library still recognizes 20% of the contained hits. It is very likely that the recovery rate will be improved if the sequential screening calculations are extended for more than nine iterations. This dramatic decrease in experimental effort while still keeping high recognition of active hits is expected to be of great benefit for pharmaceutical research.

Screening also benefits from the iterative nature of sequential screening protocols. Feedback provided by previous iterations enabled the ECP classifier to correct errors made in earlier runs. In this way, the classifier was able to find a high number of hits even when classification completely failed in earlier iterations.

In summary, the emerging pattern mining framework as introduced in the data mining community by Dong et al. (1999b) has been successfully adapted to chemoinformatic problems. Combining EP mining and the well-known notion of chemical descriptors, emerging chemical patterns are defined as class-specific descriptor-value range combinations. These combinations capture even subtle

differences in different compound sets and can be used for knowledge mining compound data (section 3.1) or for property prediction as binary classifiers (section 3.2). In a second part, the integration of ECP-based classification and HTS methods is evaluated in the form of simulated sequential screening. Sequential screening as described herein shows how both embedded methods can benefit from each other and justifies a tight integration of the work computational and experimental scientists in drug discovery projects.

A Chemical Descriptors

A.1 61 Uncorrelated Descriptors

Table A.1: 61 uncorrelated type I and II descriptors. Seven different types of descriptors are used: (computed) physical properties, atom or bond counts, connectivty indices, distance matrix measures, subdivided surface area properties and partial charge descriptors. The van–der–Waals surface area (VDW surface area) v_i for atom i is approximated from the connection table (Labute, 2004). For subdivided surface area descriptors, L_i denotes the contribution of atom i to the logP(octanol/water) partitioning coefficient and R_i denotes the contribution of atom i to the molar refractivity as described in Wildman and Crippen (1999). Partial charges are estimated by the partial equalization of orbital electronegativities (PEOE) method from Gasteiger and Marsili (1980).

Descriptor	Type	Definition
logP(o/w)	phys. property	Log of the octanol/water partition coefficient (Labute, 1998)
TPSA	phys. property	Polar VDW surface area (connection table approximation) (Ertl et al., 2000)
a_ICM	atom count	Atom information content (mean)
a_nBr	atom count	No. of bromine atoms
a_nCl	atom count	No. of chlorine atoms
a_nF	atom count	No. of fluorine atoms
a_nI	atom count	No. of iodine atoms
a_nN	atom count	No. of nitrogen atoms
a_nO	atom count	No. of oxygen atoms
a_nP	atom count	No. of phosphor atoms
a_nS	atom count	No. of sulfur atoms
b_1rotR	bond count	Fraction of rotatable bonds
b_1rotN	bond count	No. of rotatble bonds
b_double	bond count	No. of double bonds
b_triple	bond count	No. of triple bonds
balabanJ	distance matrix	Balaban averaged distance sum connectivity (Balaban, 1982)
petitjean	distance matrix	$(diameter - radius)/diameter$ (Petitjean, 1992)
VDistEQ	distance matrix	Vertex distance equality index

chi0v_C	connectivity index	Carbon valence connectivity index of order 0 (Hall and B.Kier, 1991; Hall and Kier, 1977).
SlogP_VSA0	subdiv. surf. area	Σv_i such that $L_i \leq -0.40$
SlogP_VSA1	subdiv. surf. area	Σv_i such that $L_i \in (-0.40, -0.20]$
SlogP_VSA2	subdiv. surf. area	Σv_i such that $L_i \in (-0.20, 0.00]$
SlogP_VSA3	subdiv. surf. area	Σv_i such that $L_i \in (0.00, 0.10]$
SlogP_VSA4	subdiv. surf. area	Σv_i such that $L_i \in (0.10, 0.15]$
SlogP_VSA5	subdiv. surf. area	Σv_i such that $L_i \in (0.15, 0.20]$
SlogP_VSA7	subdiv. surf. area	Σv_i such that $L \in (0.25, 0.30]$
SlogP_VSA8	subdiv. surf. area	Σv_i such that $L_i \in (0.30, 0.40]$
SlogP_VSA9	subdiv. surf. area	Σv_i such that $L_i > 0.40$
SMR_VSA0	subdiv. surf. area	Σv_i such that $R_i \in (0.00, 0.11]$
SMR_VSA1	subdiv. surf. area	Σv_i such that $R_i \in (0.11, 0.26]$
SMR_VSA2	subdiv. surf. area	Σv_i such that $R_i \in (0.26, 0.35]$
SMR_VSA3	subdiv. surf. area	Σv_i such that $R_i \in (0.35, 0.39]$
SMR_VSA4	subdiv. surf. area	Σv_i such that $R_i \in (0.39, 0.44]$
SMR_VSA5	subdiv. surf. area	Σv_i such that $R_i \in (0.44, 0.49]$
SMR_VSA6	subdiv. surf. area	Σv_i such that $R_i \in (0.485, 0.56]$
a_acc	pharmacophore	No. of hydrogen bond acceptor atoms
a_don	pharmacophore	No. of hydrogen bond acceptor atoms
a_base	pharmacophore	No. of basic atoms
vsa_acid	pharmacophore	VDW surface areas of acidic atoms
vsa_base	pharmacophore	VDW surface areas of basic atoms
vsa_don	pharmacophore	VDW surface areas of H-bond donors
vsa_other	pharmacophore	VDW surface area of other atoms
vsa_pol	pharmacophore	VDW surface areas of polar atoms
PEOE_RPC-	partial charge	Relative negative partial charge
PEOE_VSA_FNEG	partial charge	Fractional negative VDW surface area
PEOE_VSA_FHYD	partial charge	Fractional hydrophobic VDW surface area
PEOE_VSA+0	partial charge	Σv_i where $q_i \in [0.00, 0.05)$
PEOE_VSA+1	partial charge	Σv_i where $q_i \in [0.05, 0.10)$
PEOE_VSA+2	partial charge	Σv_i where $q_i \in [0.10, 0.15)$
PEOE_VSA+3	partial charge	Σv_i where $q_i \in [0.15, 0.20)$
PEOE_VSA+4	partial charge	Σv_i where $q_i \in [0.20, 0.25)$
PEOE_VSA+5	partial charge	Σv_i where $q_i \in [0.20, 0.30)$
PEOE_VSA+6	partial charge	Σv_i where $q_i \geq 0.30$
PEOE_VSA-0	partial charge	Σv_i where $q_i \in [-0.05, 0.00)$
PEOE_VSA-1	partial charge	Σv_i where $q_i \in [-0.10, -0.05)$
PEOE_VSA-2	partial charge	Σv_i where $q_i \in [-0.15, -0.10)$
PEOE_VSA-3	partial charge	Σv_i where $q_i \in [-0.20, -0.15)$
PEOE_VSA-4	partial charge	Σv_i where $q_i \in [-0.25, -0.20)$
PEOE_VSA-5	partial charge	Σv_i where $q_i \in [-0.30, -0.25)$
PEOE_VSA-6	partial charge	Σv_i where $q_i < -0.30$

A.2 3D Descriptors for Conformation Anaylsis

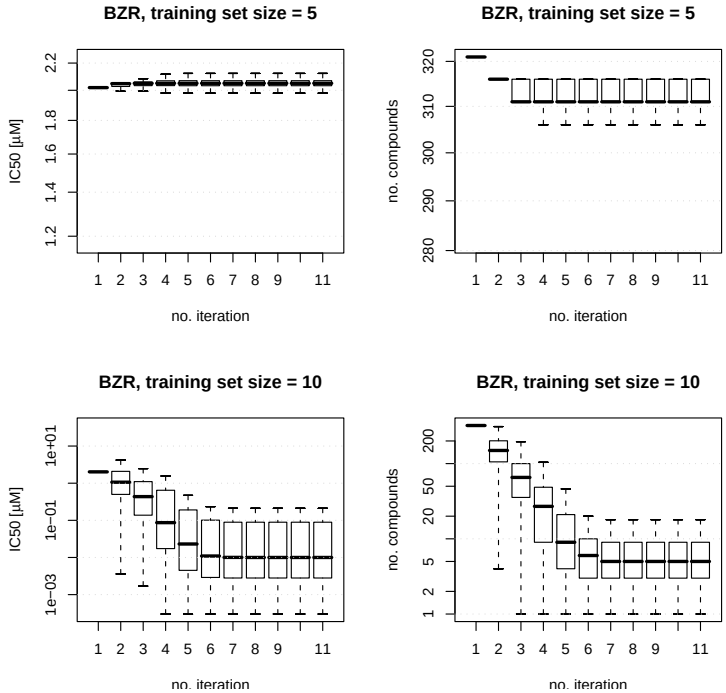
Table A.2: The set of type III (3D) descriptors used for conformation difference mining. The descriptors are calculated with MOE after normalization as described in 3.1.2. For descriptors using the AM1 (Dewar et al., 1985), PM3 (Stewart, 1989) or MNDO (Dewar and Thiel, 1977) method, MOE uses an integrated implementation of MOPAC (Stewart, 1990). The descriptors are grouped into five different classes energy, shape, charge, surface and volume depending on the chemical property they encode.

Descriptor	Type	Definition
E	energy	Potential energy
E_ang	energy	Angle bending energy component of the potential energy
E_ele	energy	Electrostatic energy component of the potential energy
E_nb	energy	Non-bonded energy component of the potential energy
E_oop	energy	Out-of-plane energy component of the potential energy
E_sol	energy	Solvation energy component of the potential energy
E_stb	energy	Bond stretching - angle bending energy component of the potential energy
E_str	energy	Bond stretch energy component of the potential energy
E_strain	energy	Strain energy component of the potential energy
E_tor	energy	Torsion energy component of the potential energy
E_vdw	energy	Van der Waals energy component of the potential energy
AM1_E	energy	Potential energy calculated with the semi-empirical AM1 method
AM1_Eele	energy	AM1 electrostatic energy component
AM1_IP	energy	AM1 ionization potential
AM1_HF	energy	AM1 heat of formation
AM1_HOMO	energy	AM1 energy of the highest occupied molecular orbital
AM1_LUMO	energy	AM1 energy of the lowest unoccupied molecular orbital
PM3_E	energy	Potential energy calculated with the semi-empirical PM3 method
PM3_Eele	energy	PM3 electrostatic energy component
PM3_IP	energy	PM3 ionization potential
PM3_HF	energy	PM3 heat of formation
PM3_HOMO	energy	PM3 energy of the highest occupied molecular orbital
PM3_LUMO	energy	PM3 energy of the lowest unoccupied molecular orbital
MNDO_E	energy	Potential energy calculated with the semi-empirical modified neglect of diatomic overlap (MNDO)
MNDO_Eele	energy	MNDO electrostatic energy
MNDO_IP	energy	MNDO ionization potential
MNDO_HF	energy	MNDO heat of formation
MNDO_HOMO	energy	MNDO energy of the highest occupied molecular orbital
MNDO_LUMO	energy	MNDO energy of the lowest unoccupied molecular orbital

glob	shape	Globularity index. Fraction of the smallest and largest Eigenvalue of the covariance matrix of atomic coordinates. Values range from 0.0 (two- or onedimensional objects) to 1.0 (perfect sphere)
std_dim1	shape	Standard deviation of coordinates along the first principal axis
std_dim2	shape	Standard deviation of coordinates along the second principal axis
std_dim3	shape	Standard deviation of coordinates along the third principal axis
rgyr	shape	Radius of gyration. Root mean square distance of atomic coordinates from the center of mass
pmi	shape	Absolute value of the principal moment of inertia
pmiX	shape	x–component of the principal moment of inertia
pmiY	shape	y–component of the principal moment of inertia
pmiZ	shape	z–component of the principal moment of inertia
dipole	charge	Absolute value of the dipole moment
AM1_dipole	charge	AM1 dipole moment
PM3_dipole	charge	PM3 dipole moment
MNDO_dipole	charge	MNDO dipole moment
dipoleX	charge	x–component of the dipole moment
dipoleY	charge	y–component of the dipole moment
dipoleZ	charge	z–component of the dipole moment
ASA	surface	Solvent-accessible surface area calculated using a probe of radius 1.4 Å
ASA+	surface	Solvent-accessible surface area of positively charged atoms
ASA-	surface	Water accessible surface area of negatively charged atoms.
ASA_H	surface	Solvent-accessible surface area of hydrophobic atoms
ASA_P	surface	Solvent-accessible surface area of polar atoms
CASA+	surface	Positively charged weighted surface area: ASA+ multiplied with the partial charge of the most positive atom (Stanton and Jurs, 1990)
CASA-	surface	Negatively charged weighted surface area: ASA multiplied with the partial charge of the most negative atom (Stanton and Jurs, 1990)
DASA	surface	Absolute value of the difference between ASA+ and ASA
DCASA	surface	Absolute value of the difference between CASA+ and CASA- (Stanton and Jurs, 1990)
FASA+	surface	Fraction of the positively charged solvent-accessible surface area: ASA+ divided by ASA
FASA-	surface	Fraction of the negatively charged solvent-accessible surface area: ASA- divided by ASA
FASA_H	surface	Fraction of the hydrophobic solvent-accessible surface area: ASA divided by ASA
FASA_P	surface	Fraction of the polar solvent-accessible surface area: ASA_P divided by ASA
FCASA+	surface	Fractional positively charged weighted surface area: CASA+ divided by ASA
FCASA-	surface	Fractional negatively charged weighted surface area: CASA- divided by ASA
VSA	surface	Van der Waals surface area
dens	volume	Density. Molecular weight divided by the van der Waals volume
vol	volume	van der Waals volume approximation

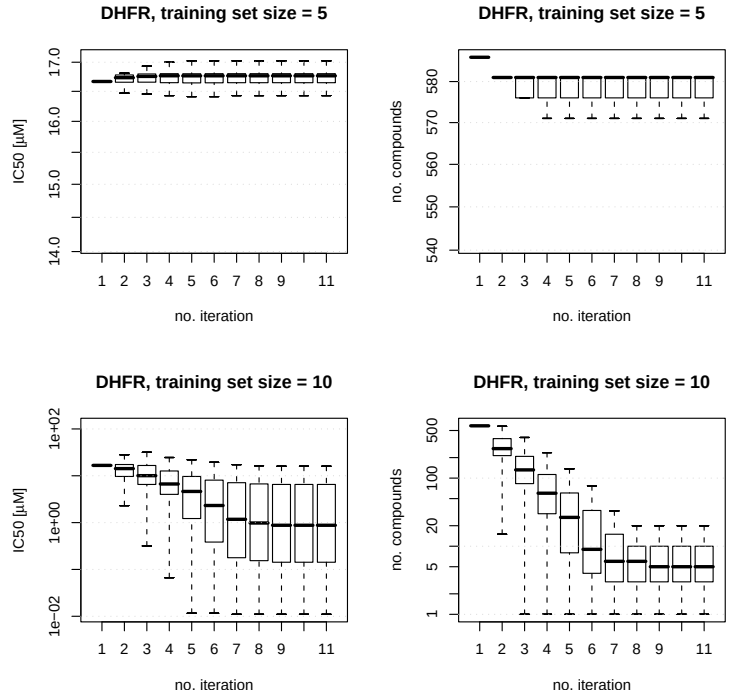
B Simulated Lead Optimization

Decision Tree

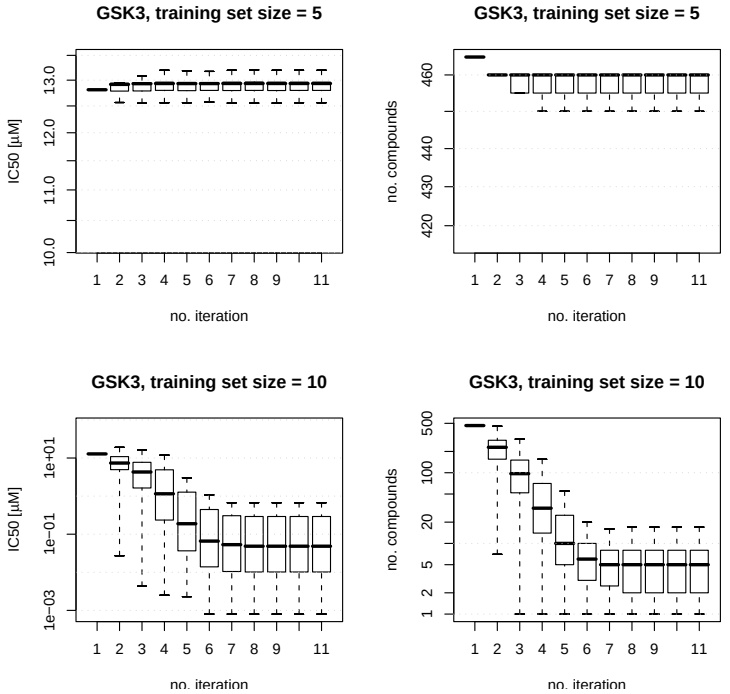


(a) Bzr

Figure B.1: Simulated lead optimization trials with decision tree classification. Average potencies (left) and compound numbers (right) over 500 calculations of 10 iterations each are reported for training sets of five and 10 compounds. Decision tree classification completely fails to produce highly potent compound sets because it is unable to handle training sets containing only five compounds.

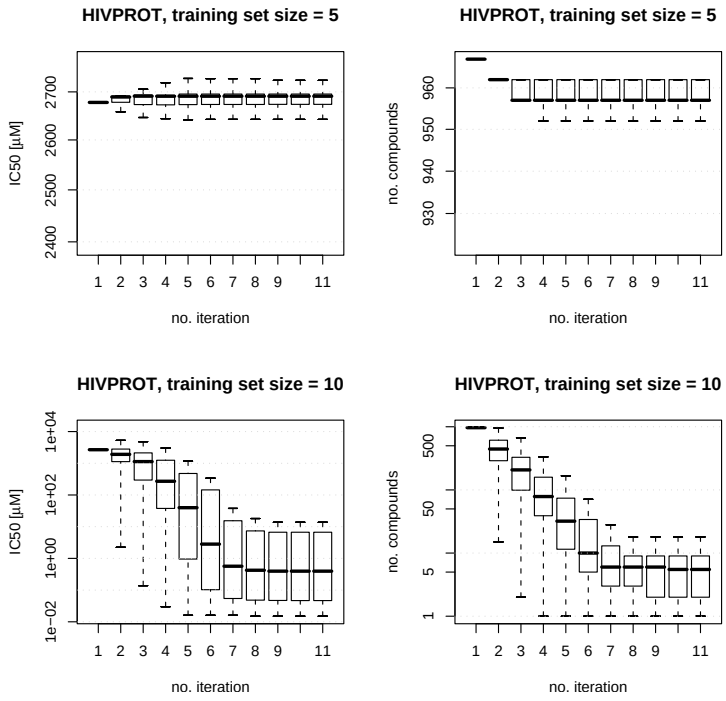


(b) DHFR



(c) GSK3

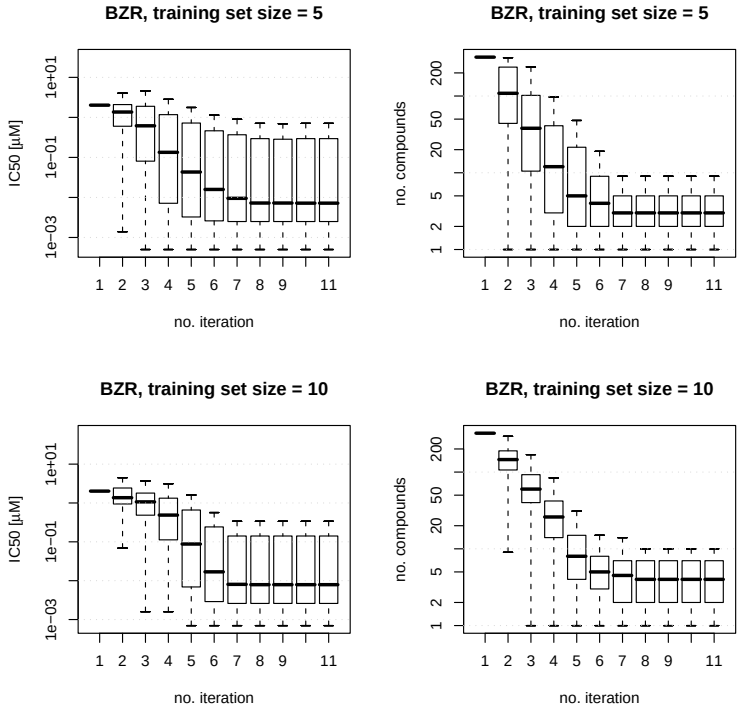
Figure B.1: continued.



(d) HIVPROT

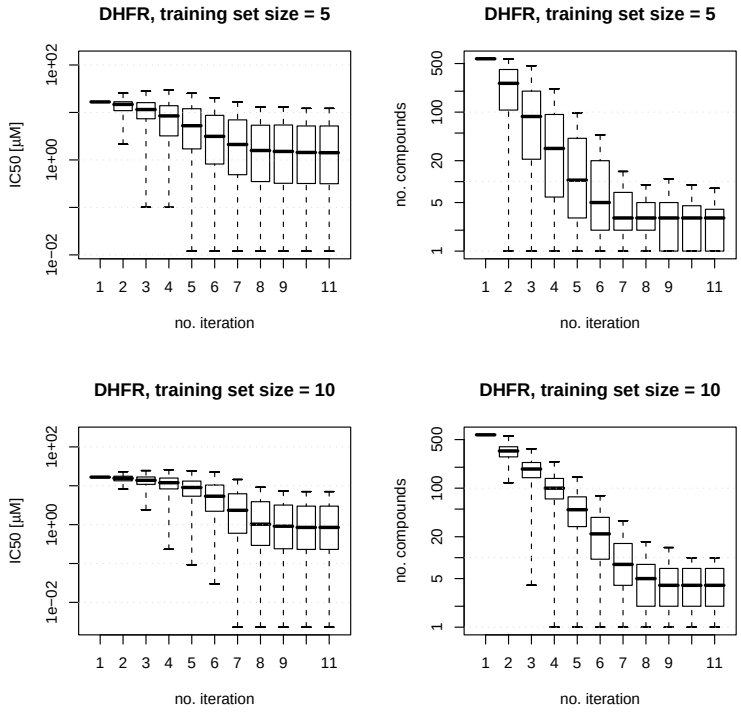
Figure B.1: continued.

Binary QSAR

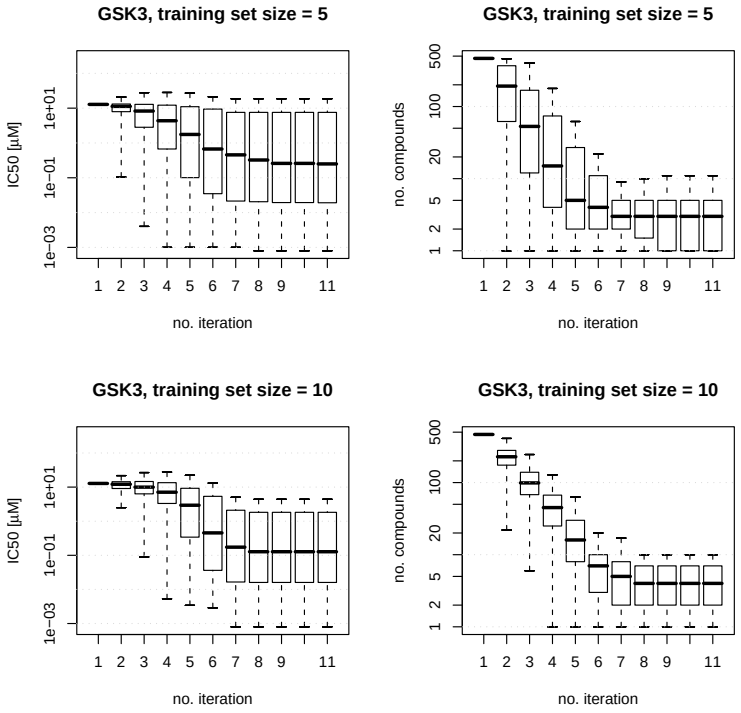


(a) Bzr

Figure B.2: Simulated lead optimization trials with binary QSAR classification. Average potencies (left) and compound numbers (right) over 500 calculations of 10 iterations each are reported for training sets of five and 10 compounds. The ECP-based experiments described in section 4.1 reach higher potency levels in few iterations. For HIVPROT, binary QSAR completely fails to produce nano molar compound sets.

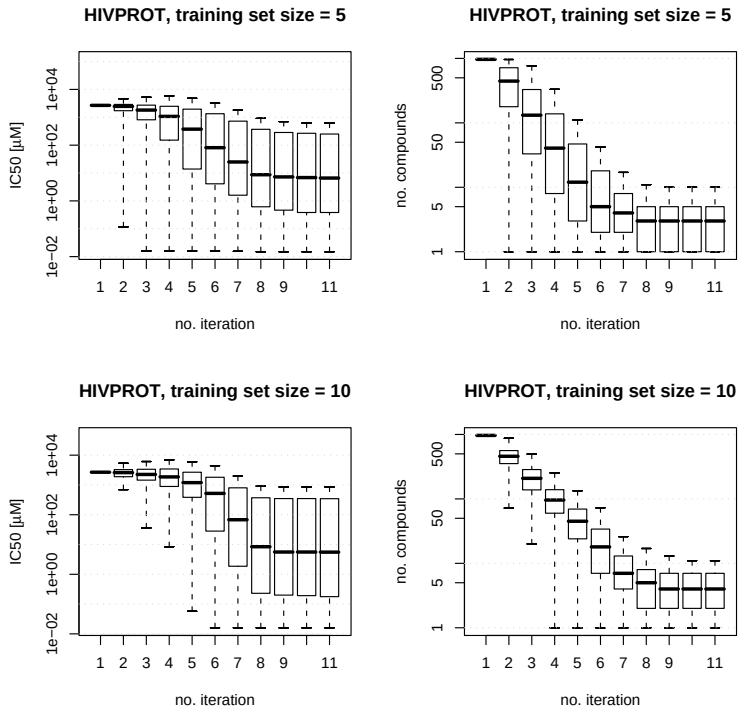


(b) DHFR



(c) GSK3

Figure B.2: continued.



(d) HIVPROT

Figure B.2: continued.

C Simulated Sequential Screening

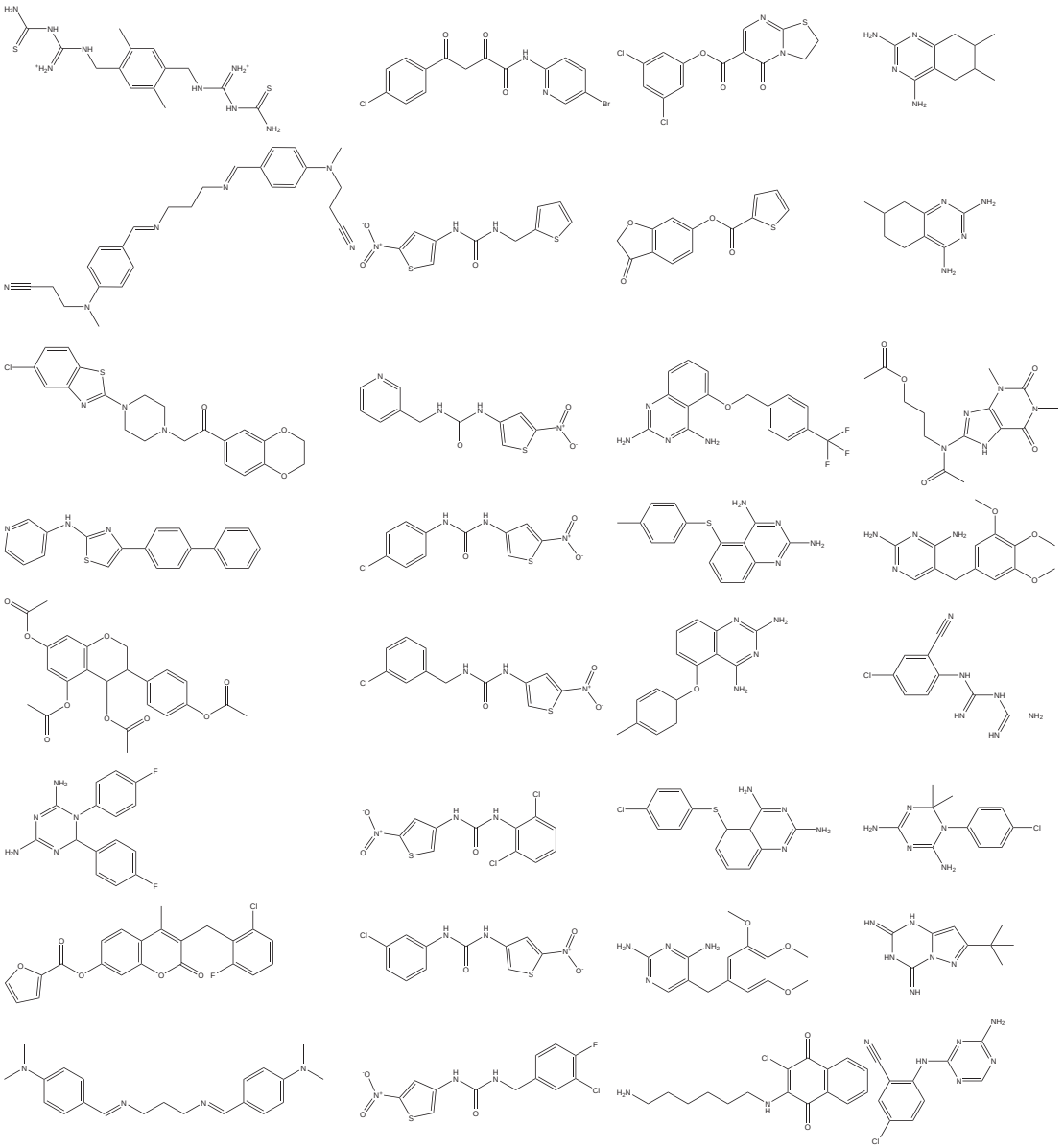


Figure C.1: Structures of all 32 DHFR inhibitors used in the simulated sequential screening experiment described in chapter 4.

D Patterns for Conformation Analysis

Emerging chemical patterns for the remaining 15 classes presented in section 3.1. In all tables, "Growth" gives the growth rate of a pattern, "B" ("B%") stands for the absolute (relative) support in the binding conformation class, "M" ("M%") stands for the absolute (relative) support in the modeled conformation data set and "Pattern" shows the pattern.

Table D.1: Discriminatory Patterns for Acetylcholine Esterase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	5	0	0	{E_strain:(9.79,∞]}
∞	80	4	0	0	{E_str:(20.31,∞]}
∞	60	3	0	0	{E_strain:(15.34,24.46]}
∞	60	3	0	0	{E_str:(14.90,∞], vol:(197.50,351.56]}
∞	60	3	0	0	{E_str:(14.90,∞], pmiY:(334.58,1141.06]}
∞	60	3	0	0	{E_str:(14.90,∞], VSA:(219.79,391.36]}
∞	60	3	0	0	{E_str:(14.90,∞], PM3_Eele:($-$755075.90,$-$335581.90]}
∞	60	3	0	0	{E_str:(14.90,∞], MNDO_Eele:($-$788514.20,$-$336721.10]}
∞	60	3	0	0	{E_str:(14.90,∞], E_vdw:(28.30,72.89]}
∞	60	3	0	0	{E_str:(14.90,∞], E_tor:($-\infty$,1.88]}
∞	60	3	0	0	{CASA+:(404.72,1956.91], E_str:(14.90,∞]}

Table D.2: Discriminatory Patterns for Carbonic anhydrase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	21	0	0	{E_strain:(12.92,∞]}
∞	90	19	0	0	{E_str:(6.05,∞]}
∞	86	18	0	0	{E_strain:(24.46,∞]}
∞	86	18	0	0	{E_stb:(−∞,−1.27]}
∞	76	16	0	0	{E_str:(14.90,∞]}
∞	76	16	0	0	{E_ang:(17.54,∞], dipole:(1.31,∞]}
∞	71	15	0	0	{E_ang:(24.53,∞]}
∞	62	13	0	0	{E_stb:(−∞,−2.63]}
∞	62	13	0	0	{E:(48.58,∞]}
121.95	62	13	1	1	{E_sol:(−20.79,−12.58], dipole:(1.31,∞]}
121.95	62	13	1	1	{E_sol:(−20.79,−12.58], E_vdw:(28.30,72.89]}
65.67	67	14	1	2	{E_ang:(17.54,∞], E_sol:(−20.79,−12.58]}
34.40	52	11	2	3	{E_sol:(−20.79,−12.58], std_dim2:(−∞,1.39]}
34.40	52	11	2	3	{E_ang:(17.54,∞], std_dim2:(−∞,1.39]}

Table D.3: Discriminatory Patterns for Carboxypeptidase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	8	0	0	{E_strain:(24.46,∞]}
∞	100	8	0	0	{E_str:(14.90,∞]}
∞	100	8	0	0	{E_strain:(28.68,∞]}
∞	100	8	0	0	{E_str:(13.71,∞]}
∞	63	5	0	0	{E_ang:(23.84,∞]}
∞	50	4	0	0	{E:(128.36,∞]}
29.50	50	4	2	1	{E_tor:(12.17,∞]}
18.44	63	5	3	2	{dipoleZ:(−∞,−1.24]}
14.75	50	4	3	2	{E_ele:(−11.07,∞], E_sol:(−∞,−177.16], FASA_P:(−∞,0.32]}
14.75	50	4	3	2	{E_ele:(−11.07,∞], E_sol:(−∞,−177.16], FASA_H:(0.68,∞]}
14.75	50	4	3	2	{E_ele:(−11.07,∞], E_nb:(42.27,136.00], MNDO_HF:(−∞,−79.25]}
14.75	50	4	3	2	{AM1_dipole:(8.17,17.34], E_ele:(−11.07,∞], rgyr:(2.97,5.32]}
12.29	63	5	5	3	{AM1_dipole:(8.17,17.34], E_ele:(−11.07,∞], FASA_P:(−∞,0.32]}
11.06	75	6	7	4	{dipoleZ:(−∞,−1.01]}

Table D.4: Discriminatory Patterns for Cyclin-dependent Kinase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	31	0	0	{E_strain:(24.46,∞]}
∞	100	31	0	0	{E_strain:(23.84,∞]}
∞	100	31	0	0	{E_str:(10.07,∞]}
∞	90	28	0	0	{E_str:(14.90,∞]}
∞	52	16	0	0	{E:(118.44,∞]}
28.57	58	18	2	5	{E:(82.08,147.95], E_nb:(−∞,42.27]}
10.38	55	17	5	13	{E:(82.08,147.95], VSA:(219.79,391.36]}
9.77	52	16	5	13	{E:(82.08,147.95], vol:(197.50,351.56]}
9.07	52	16	6	14	{E:(82.08,147.95], PM3_dipole:(−∞,9.76]}
9.07	52	16	6	14	{E:(82.08,147.95], MNDO_dipole:(−∞,8.98]}
9.07	52	16	6	14	{AM1_IP:(5.51,10.61], E:(82.08,147.95], pmi:(2429.75,10083.21]}
9.07	52	16	6	14	{AM1_HOMO:(−10.61,−5.51], E:(82.08,147.95], pmi:(2429.75,10083.21]}

Table D.5: Discriminatory Patterns for Elastase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	3	0	0	{FASA-:(0.47,∞], pmiY:(1141.06,∞], std_dim1:(3.33,∞], std_dim3:(1.21,∞]}
∞	100	3	0	0	{FASA-:(0.47,∞], pmiY:(1141.06,∞], pmiZ:(538.74,∞], std_dim1:(3.33,∞]}
∞	100	3	0	0	{FASA-:(0.47,∞], glob:(0.10,∞], pmiY:(1141.06,∞], std_dim1:(3.33,∞]}
∞	100	3	0	0	{FASA-:(0.47,∞], PM3_HF:(−∞,−197.90], std_dim1:(3.33,∞], std_dim3:(1.21,∞]}
∞	100	3	0	0	{FASA-:(0.47,∞], PM3_HF:(−∞,−197.90], pmiZ:(538.74,∞], std_dim1:(3.33,∞]}
∞	100	3	0	0	{FASA-:(0.47,∞], PM3_HF:(−∞,−197.90], glob:(0.10,∞], std_dim1:(3.33,∞]}
∞	100	3	0	0	{E_strain:(24.46,∞]}
∞	100	3	0	0	{E_ele:(−11.07,∞], FASA-:(0.47,∞], std_dim1:(3.33,∞], std_dim3:(1.21,∞]}
∞	100	3	0	0	{E_ele:(−11.07,∞], FASA-:(0.47,∞], pmiZ:(538.74,∞], std_dim1:(3.33,∞]}
∞	100	3	0	0	{E_strain:(16.91,∞]}
∞	100	3	0	0	{E_str:(8.96,∞]}
∞	100	3	0	0	{E:(85.27,∞]}
∞	67	2	0	0	{E_ele:(1.88,∞]}
∞	67	2	0	0	{E_ang:(22.32,∞]}
29.00	100	3	3	1	{PM3_dipole:(−∞,4.26]}

Table D.6: Discriminatory Patterns for Endothiapepsin Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	6	0	0	{E_strain:(24.46,∞]}
∞	100	6	0	0	{E_str:(14.90,∞]}
∞	100	6	0	0	{glob:(−∞,0.10]}
∞	100	6	0	0	{E_strain:(43.04,∞]}
∞	83	5	0	0	{E_str:(19.79,∞]}
∞	83	5	0	0	{E_stb:(−∞,−0.14]}
∞	83	5	0	0	{E_ang:(29.37,∞]}
∞	83	5	0	0	{E:(116.41,∞]}
∞	67	4	0	0	{glob:(0.01,0.08]}
∞	67	4	0	0	{E_tor:(1.88,∞], dipole:(1.31,∞]}
∞	67	4	0	0	{E_nb:(42.27,136.00], E_tor:(1.88,∞]}
∞	67	4	0	0	{E:(147.95,∞]}
∞	67	4	0	0	{std_dim1:(5.15,∞]}
∞	67	4	0	0	{rgyr:(5.78,∞]}
∞	67	4	0	0	{E_nb:(68.19,∞]}
∞	67	4	0	0	{AM1_dipole:(11.85,∞]}
∞	50	3	0	0	{dipoleX:(−∞,−1.59]}
∞	50	3	0	0	{PM3_IP:(−∞,5.59]}
∞	50	3	0	0	{PM3_HOMO:(−5.59,∞]}
∞	50	3	0	0	{pmi:(23385.78,∞]}
15.00	83	5	6	1	{std_dim3:(−∞,1.38]}
15.00	83	5	6	1	{pmiZ:(−∞,1011.65]}
15.00	83	5	6	1	{pmiX:(13024.76,∞]}

Table D.7: Discriminatory Patterns for Factor Xa Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	10	0	0	{E_strain:(24.46,∞]}
∞	100	10	0	0	{E_str:(14.90,∞]}
∞	100	10	0	0	{E_strain:(17.83,∞]}
∞	100	10	0	0	{E_str:(12.19,∞]}
∞	60	6	0	0	{CASA+:(1956.91,∞], E_ang:(17.54,∞], pmi:(10083.21,∞]}
∞	50	5	0	0	{CASA+:(1956.91,∞], FCASA-:(1.40,3.21], pmi:(10083.21,∞]}
72.00	80	8	1	1	{E_ang:(23.79,∞]}
54.00	60	6	1	1	{E_oop:(−∞,0.02]}
18.00	60	6	3	3	{ASA+:(426.07,∞], E_ang:(17.54,∞], pmi:(10083.21,∞]}
15.00	50	5	3	3	{E:(82.08,147.95], std_dim3:(1.21,∞]}
15.00	50	5	3	3	{ASA+:(426.07,∞], FCASA-:(1.40,3.21], pmi:(10083.21,∞]}
15.00	50	5	3	3	{ASA+:(426.07,∞], CASA-:(1025.33,1591.69]}
15.00	50	5	3	3	{ASA:(653.76,791.58], E_ang:(17.54,∞], PM3_E:(−124393.40,−95187.02]}

Table D.8: Discriminatory Patterns for FK506 Binding Protein Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	6	0	0	{E_strain:(3.82,∞]}
∞	67	4	0	0	{E_strain:(24.46,∞]}
∞	67	4	0	0	{ASA_H:(468.24,665.84], E_ang:(17.54,∞]}
∞	50	3	0	0	{E_ang:(17.54,∞], vol:(351.56,486.00]}
∞	50	3	0	0	{E_ang:(17.54,∞], pmi:(2429.75,10083.21]}
∞	50	3	0	0	{E_ang:(17.54,∞], VSA:(391.36,553.94]}
∞	50	3	0	0	{E_ang:(17.54,∞], PM3_Eele:(−1429154.00,−755075.90]}
∞	50	3	0	0	{E_ang:(17.54,∞], PM3_E:(−124393.40,−95187.02]}
∞	50	3	0	0	{E_ang:(17.54,∞], MNDO_Eele:(−1515367.00,−788514.20]}
∞	50	3	0	0	{E_ang:(17.54,∞], MNDO_E:(−158562.00,−105111.10]}
∞	50	3	0	0	{E_str:(11.74,∞]}

Table D.9: Discriminatory Patterns for HIV Protease Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	20	0	0	{E_strain:(24.46,∞]}
∞	100	20	0	0	{E_strain:(21.68,∞]}
∞	65	13	0	0	{E_str:(16.49,∞]}
45.83	65	13	1	2	{E_stb:(−∞,−0.12]}
32.90	70	14	2	3	{E:(147.95,∞]}
14.10	70	14	5	7	{E_str:(14.90,∞]}
10.97	70	14	6	9	{glob:(−∞,0.12]}
5.88	50	10	9	12	{ASA:(791.58,∞], ASA_H:(665.84,∞], ASA_P:(−∞,210.39], E_ang:(17.54,∞], E_tor:(1.88,∞], PM3_LUMO:(−0.74,0.54]}
5.54	55	11	10	14	{ASA_H:(665.84,∞], E_ang:(17.54,∞], E_ele:(−11.07,∞], FCASA+:(1.78,∞], std_dim1:(3.33,∞]}
5.54	55	11	10	14	{ASA_H:(665.84,∞], E_ang:(17.54,∞], E_ele:(−11.07,∞], FASA+:(0.54,∞], std_dim1:(3.33,∞]}
5.54	55	11	10	14	{ASA_H:(665.84,∞], E_ang:(17.54,∞], E_ele:(−11.07,∞], E_tor:(1.88,∞], std_dim1:(3.33,∞]}
5.54	55	11	10	14	{ASA_H:(665.84,∞], E_ang:(17.54,∞], E_ele:(−11.07,∞], E_tor:(1.88,∞], FCASA+:(1.78,∞]}
5.54	55	11	10	14	{ASA_H:(665.84,∞], E_ang:(17.54,∞], E_ele:(−11.07,∞], E_tor:(1.88,∞], FASA+:(0.54,∞]}

Table D.10: Discriminatory Patterns for Plasminogen Activator Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	7	0	0	{E_strain:(10.78,∞]}
∞	86	6	0	0	{E_strain:(24.46,∞]}
∞	71	5	0	0	{E_str:(14.90,∞]}
∞	71	5	0	0	{E_str:(11.78,∞]}
∞	71	5	0	0	{E_stb:(0.77,∞]}
∞	71	5	0	0	{E:(92.13,∞]}
12.00	57	4	5	2	{E_ang:(17.54,∞]}
14.00	100	7	7	3	{E_ang:(12.42,∞]}
6.00	57	4	10	4	{dipoleX:(−∞,−1.59], std_dim1:(2.41,3.33]}
6.00	57	4	10	4	{E_ele:(−11.07,∞], std_dim1:(2.41,3.33]}
6.00	57	4	10	4	{E_ele:(−11.07,∞], PM3_IP:(10.77,∞], dipoleX:(−∞,−1.59]}
6.00	57	4	10	4	{E_ele:(−11.07,∞], PM3_IP:(10.77,∞], PM3_dipole:(9.76,∞]}
6.00	57	4	10	4	{E_ele:(−11.07,∞], PM3_HOMO:(−∞,−10.77], dipoleX:(−∞,−1.59]}
6.00	57	4	10	4	{E_ele:(−11.07,∞], PM3_HOMO:(−∞,−10.77], PM3_dipole:(9.76,∞]}

Table D.11: Discriminatory Patterns for Protein Tyrosine Phosphatase 1b Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	14	0	0	{E_strain:(24.46,∞]}
∞	100	14	0	0	{E_strain:(26.79,∞]}
∞	93	13	0	0	{E_str:(14.90,∞]}
∞	93	13	0	0	{E_str:(16.03,∞]}
∞	57	8	0	0	{ASA_P:(210.39,282.26], E_ele:(−11.07,∞], E_tor:(1.88,∞]}
37.71	57	8	2	1	{AM1_IP:(−∞,5.51], ASA_P:(210.39,282.26], E_tor:(1.88,∞]}
37.71	57	8	2	1	{AM1_HOMO:(−5.51,∞], ASA_P:(210.39,282.26], E_tor:(1.88,∞]}
33.00	50	7	2	1	{E_ang:(17.54,∞], dipoleX:(−∞,−1.59]}
33.00	50	7	2	1	{ASA_P:(210.39,282.26], E_ang:(17.54,∞]}
25.93	79	11	3	2	{E_ang:(21.00,∞]}
21.21	64	9	3	2	{E_ele:(−11.07,∞], E_tor:(1.88,∞], dipoleX:(−∞,−1.59]}
16.50	50	7	3	2	{E_ang:(17.54,∞], FCASA-:(1.40,3.21]}

Table D.12: Discriminatory Patterns for Protocatechuate3,4-dioxygenase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	10	0	0	{E_strain:(2.55,∞]}
∞	90	9	0	0	{E_str:(2.76,∞]}
∞	80	8	0	0	{E_strain:(4.30,15.34]}
∞	70	7	0	0	{pmiZ:(0.01,1.43]}
∞	60	6	0	0	{std_dim3:(0.01,0.05]}
∞	60	6	0	0	{glob:(0.00,0.00]}
∞	60	6	0	0	{E_tor:(1.55,∞]}
∞	60	6	0	0	{E_oop:(0.09,∞]}
13.50	90	9	7	1	{E_ang:(2.73,∞], pmiY:($-\infty$,201.06]}
12.00	80	8	7	1	{E:(25.00,∞], pmiY:($-\infty$,201.06]}

Table D.13: Discriminatory Patterns for Thermolysin Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	6	0	0	{E_strain:(24.46,∞]}
∞	100	6	0	0	{E_str:(14.90,∞]}
∞	100	6	0	0	{E_strain:(42.00,∞]}
∞	100	6	0	0	{E_str:(16.05,∞]}
∞	100	6	0	0	{E:(99.01,∞]}
∞	83	5	0	0	{E_ele:(-11.07,∞], E_nb:(42.27,136.00]}
∞	83	5	0	0	{E_stb:($-\infty$,-0.45]}
∞	83	5	0	0	{E_ele:(-3.67,∞]}
∞	50	3	0	0	{E_ang:(17.54,∞], dipoleX:($-\infty$,-1.59]}
∞	50	3	0	0	{E:(147.95,∞]}
∞	50	3	0	0	{CASA-:(1025.33,1591.69], E_ele:(-11.07,∞]}
∞	50	3	0	0	{ASA-:(275.40,∞], PM3_HF:(-197.90,-85.25], PM3_dipole:(9.76,∞]}
41.00	100	6	2	1	{E_nb:(42.27,136.00], PM3_dipole:(9.76,∞]}
34.17	83	5	2	1	{E_ang:(17.54,∞], E_ele:(-11.07,∞]}
20.50	100	6	5	2	{E_nb:(46.18,∞]}

Table D.14: Discriminatory Patterns for Thrombin Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	21	0	0	{E_strain:(16.48,∞]}
∞	95	20	0	0	{E_strain:(24.46,∞]}
∞	95	20	0	0	{E_str:(9.67,∞]}
∞	62	13	0	0	{E_str:(14.90,∞]}
∞	62	13	0	0	{E:(124.47,∞]}
∞	52	11	0	0	{E:(147.95,∞]}
44.86	57	12	1	2	{E_oop:(0.97,∞]}
5.87	52	11	9	14	{E_nb:(42.27,136.00], E_tor:(1.88,∞], pmi:(2429.75,10083.21], std_dim1:(3.33,∞]}
5.14	52	11	10	16	{E_nb:(42.27,136.00], std_dim1:(3.33,∞], std_dim2:(1.81,∞], vol:(351.56,486.00]}
5.14	52	11	10	16	{E_nb:(42.27,136.00], pmiY:(1141.06,∞], std_dim1:(3.33,∞], vol:(351.56,486.00]}
5.14	52	11	10	16	{E_nb:(42.27,136.00], VSA:(391.36,553.94], dens:(−∞,1.24], glob:(0.10,∞], pmiY:(1141.06,∞], std_dim1:(3.33,∞]}
5.14	52	11	10	16	{ASA_H:(468.24,665.84], E_nb:(42.27,136.00], std_dim1:(3.33,∞], std_dim2:(1.81,∞]}
5.14	52	11	10	16	{ASA_H:(468.24,665.84], E_nb:(42.27,136.00], pmiY:(1141.06,∞], std_dim1:(3.33,∞]}

Table D.15: Discriminatory Patterns for Tyrosine Kinase Inhibitors

Growth	B [%]	B	M [%]	M	Pattern
∞	100	5	0	0	{E_strain:(24.46,∞]}
∞	100	5	0	0	{E_strain:(32.80,∞]}
∞	80	4	0	0	{E_str:(12.68,∞]}
∞	60	3	0	0	{E_str:(14.90,∞]}
∞	60	3	0	0	{E:(82.08,147.95], E_vdw:(28.30,72.89], FASA_P:(0.32,0.52], MNDO_Eele:(−∞,−1515367.00]}
∞	60	3	0	0	{E:(82.08,147.95], E_vdw:(28.30,72.89], FASA_H:(0.48,0.68], MNDO_Eele:(−∞,−1515367.00]}
∞	60	3	0	0	{dipoleZ:(−∞,−1.91]}
∞	60	3	0	0	{MNDO_IP:(0.17,∞]}
∞	60	3	0	0	{MNDO_HOMO:(−∞,−0.17]}
17.60	80	4	5	1	{E_nb:(42.27,136.00], FASA_P:(0.32,0.52]}
17.60	80	4	5	1	{E_nb:(42.27,136.00], FASA_H:(0.48,0.68]}
17.60	80	4	5	1	{ASA_P:(282.26,∞], E_nb:(42.27,136.00]}
17.60	80	4	5	1	{ASA_H:(468.24,665.84], E_nb:(42.27,136.00]}
17.60	80	4	5	1	{PM3_IP:(−0.45,∞]}
17.60	80	4	5	1	{PM3_HOMO:(−∞,0.45]}
13.20	60	3	5	1	{E_vdw:(28.30,72.89], FASA_P:(0.32,0.52], dipoleX:(−1.59,∞], rgyr:(5.32,∞]}
11.00	100	5	9	2	{E_ele:(−18.26,∞]}

Bibliography

Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007** 47, 1067–86.

Agrawal, R.; Imielinski, T.; Swami, A. N. Mining Association Rules between Sets of Items in Large Databases. In Buneman, P.; Jajodia, S., (Eds.) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.* ACM Press, New York, NY, USA, **1993** pp. 207–216.

Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006** 46, 2502–2514.

Auer, J.; Bajorath, J. Molecular Similarity Concepts and Search Calculations. In Keith, J. M., (Ed.) *Methods Mol. Biol.* vol. 453 Springer, **2008** pp. 327–347.

Bailey, J.; Manoukian, T.; Ramamohanarao, K. Fast Algorithms for Mining Emerging Patterns. In Elomaa, T.; Mannila, H.; Toivonen, H., (Eds.) *Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002. Lecture Notes in Computer Science*, vol. 2336 Springer, **2002** pp. 39–50.

Bailey, J.; Manoukian, T.; Ramamohanarao, K. A Fast Algorithm for Computing Hypergraph Transversals and its Application in Mining Emerging Patterns. In *Proceedings of the Third IEEE International Conference on Data Mining.* IEEE Computer Society, Washington, DC, USA, **2003** p. 485.

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

Balaban, A. Highly Discriminating Distance-based Topological Index. *Chem. Phys. Lett.* **1982** 89, 399–404.

Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nature Rev. Drug Discov.* **2003** 2, 369–378.

Bostrøm, J.; Norrby, P.-O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Des.* **1998** 12, 383.

Breiman, L. Bagging Predictors *Mach. Learn.* **1996** 24, 123–140.

Brown, R.; Martin, Y. Use of Structure-Activity Data To Compare Structure-based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996** 36, 572–584.

Chemical Computing Group Inc. *Molecular Operating Environment* 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada **version 2007.09**.

Chen, X.; Lin, Y.; Gilson, M. K. The Binding Database: Overview and User's Guide. *Biopolymers* **2002** 61, 127–141.

Clark, A.; Labute, P. 2D Depiction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007** 47, 1933–1944.

Cristalli, G.; Costanzi, S.; Lambertucci, C.; Lupidi, G.; Vittori, S.; Volpini, R.; Camaioni, E. Adenosine Deaminase: Functional Implications and Different Classes of Inhibitors. *Medicinal Res. Rev.* **2001** 21, 105–128.

Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. the MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977** 99, 4899–4907.

Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985** 107, 3902–3909.

Diller, D. J.; Merz, K. M. Can We Separate Active from Inactive Conformations? *J. Comput.-Aided Mol. Des.* **2002** 16, 105–12.

Dong, G.; Li, J.; Chaudhuri, S.; Madigan, D.; Fayyad, U. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM Press, New York, NY, USA, **1999** pp. 43–52.

Dong, G.; Li, J.; Zhang, X. Discovering Jumping Emerging Patterns and Experiments on Real Datasets. In Fong, J., (Ed.) *Proceedings of the Ninth International Database Conference on Heterogeneous and Internet Databases.* City University of Hong Kong Press, Hong Kong, China, **1999** pp. 155–168.

Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. In Prieditis, A.; Russel, S. J., (Eds.) *Proceedings of the Twelfth International Conference on Machine Learning.* Morgan Kaufmann, Tahoe City, CA, USA, **1995** pp. 194–202.

Eckert, H.; Bajorath, J. Determination and Mapping of Activity-Specific Descriptor Value Ranges for the Identification of Active Compounds. *J. Med. Chem.* **2006** 49, 2284–2293.

Elowe, N. H.; Blanchard, J. E.; Cechetto, J. D.; Brown, E. D. Experimental Screening of Dihydrofolate Reductase Yields a "Test Set" of 50,000 Small Molecules for a Computational Data-Mining and Docking Competition. *J. Biomol. Screen.* **2005** 10, 653–657.

Engels, M. F.; Venkatarangan, P. Smart Screening: Approaches to Efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001** 4, 275–283.

Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-based Contributions and its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000** 43, 3714–3717.

Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. *Methods Mol. Biol.* **2004** 275, 131–214.

Fan, H.; Fan, M.; Ramamohanarao, K.; Liu, M. Further Improving Emerging Pattern Based Classifiers Via Bagging. In Ng, W. K.; Kitsuregawa, M.; Li, J.; Chang, K., (Eds.) *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference.* Springer, **2006** pp. 91–96.

Fan, H.; Ramamohanarao, K. Efficiently Mining Interesting Emerging Patterns. In Dong, G.; Tang, C.; Wang, W., (Eds.) *Advances in Web-Age Information Management, 4th International Conference, WAIM 2003* Lecture Notes in Comput. Sci. Springer, **2003** pp. 189–201.

Fayyad, U. M.; Irani, K. B. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Mach. Learn.* **1992** 8, 87–102.

Feher, M.; Schmidt, J. Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignments. *J. Chem. Inf. Comput. Sci.* **2003** 43, 810–818.

Fischer; Heun; Kramer Optimal String Mining Under Frequency Constraints. In Fürnkranz, J.; Scheffer, T.; Spiliopoulou, M., (Eds.) *Tenth European Conference on Principles and Practice of Knowledge Discovery in Databases.* Lecture Notes in Comput. Sci. Springer Berlin / Heidelberg, **2006** pp. 139–150.

Freund, Y. Boosting a weak learning algorithm by majority In *Proceedings of the Third Annual Workshop on Computational Learning Theory.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, **1990** pp. 202–216.

Fürnkranz, J. Round Robin Classification. *J. Mach. Learn. Res.* **2002** 2, 721–747.

Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity – A Rapid Access to Atomic Charges *Tetrahedron* **1980** 36, 3219–3228.

Geppert, H.; Horváth, T.; Gartner, T.; Wrobel, S.; Bajorath, J. Support-Vector-Machine-based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds. *J. Chem. Inf. Model.* **2008** 48, 742–746.

Gusfield, D. *Algorithms on Strings, Trees and Sequences.* Cambirdge University Press, Cambridge, UK, **1997**.

Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996** 17, 490–519.

Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996** 17, 520–552.

Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996** 17, 553–586.

Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *J. Comput. Chem.* **1996** 17, 616–641.

Halgren, T. A.; Nachbar, R. B. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *J. Comput. Chem.* **1996** 17, 587–615.

Hall, L.; B.Kier, L. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. In Lipkowitz, K. B.; Boyd, D. B., (Eds.) *Rev. in Comput. Chem.* vol. 2 Wiley-VCH Verlag, Weinheim, **1991** pp. 367–422.

Hall, L. H.; Kier, L. B. The Nature of Structure-Activity Relationships and Their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **1977** 12, 307.

Harper, G.; Bradshaw, J.; Gittins, J.; Green, D.; Leach, A. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001** 41, 1295–1300.

Hitchings, G. H. Selective Inhibitors of Dihydrofolate Reductase. *Angew. Chem. Int. Ed. Engl.* **1989** 28, 879–885.

Irwin, J.; Shoichet, B. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005** 45, 177–182.

Jones-Hertoz, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. Use of Recursive Partitioning in the Sequential Screening of G-Protein–Coupled Receptors. *J. Pharmacol. Toxicol.* **1999** 42, 207–215.

Jorissen, R.; Gilson, M. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005** 45, 549–561.

Kerber, R. ChiMerge: Discretization of Numeric Attributes. In Swartout, W. R., (Ed.) *Proceedings of the Tenth National Conference on AI.* AAAI Press, Menlo Park, CA, USA, **1992** pp. 123–128.

Keserü, G. M.; Makara, G. M. Hit Discovery and Hit-to-Lead Approaches. *Drug Discov. Today* **2006** 11, 741–748.

Keserü, G. M.; Molnár, L.; Greiner, I. A Neural Network Based Virtual High Throughput Screening Test for the Prediction of CNS Activity. *Comb Chem High Throughput Screen.* **2000** 3, 535–40.

Kirkpatrick, P.; Ellis, C. Chemical space. *Nat. Rev. Drug Discov.* **2004** 432, 823–865.

Klopack, T. G. Balancing the Risks and the Benefits. *Drug Discov. Today* **2000** 5, 157–160.

Kohavi, R.; Sahami, M. Error-based and Entropy-based Discretization of Continuous Features. In Simoudis, E.; Han, J.; Fayyad, U., (Eds.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, USA, **1996** pp. 114–119.

Labute, P. MOE LogP(Octanol/Water) Model. **1998**. Source code in $MOE/lib/svl/quasar.svl/q_logp.svl.

Labute, P. Binary QSAR: A New Method for the Determination of Quantitative Structure Activity Relationships. *Pacific Symposium on Biocomputing* **1999** , 444–55.

Labute, P. Derivation and Applications of Molecular Descriptors Based on Approximate Surface Area In Bajorath, J., (Ed.) *Chemoinformatics. Methods Mol. Biol.*, vol. 275 Humana Press, Totowa, NJ, USA, **2004** pp. 261–278.

Lameijer, E.-W.; Kok, J. N.; Bäck, T.; Ijzerman, A. P. Mining a Chemical Database for Fragment Co-occurrence: Discovery of "Chemical Clichés". *J. Chem. Inf. Model.* **2006** 46, 553–562.

Li, J.; Dong, G.; Ramamohanarao, K. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. *Knowl. Inf. Syst.* **2001** 3, 131–145.

Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. DeEPs: A New Instance-based Lazy Discovery and Classification System. *Mach. Learn.* **2004** 54, 99–124.

Li, J.; Wong, L. Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns. *Bioinformatics* **2002** 18, 725–734.

Li, J.; Wong, L. Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns. *Bioinformatics* **2002** 18, 1406–1407.

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

Loekito, E.; Bailey, J. Fast Mining of High Dimensional Expressive Contrast Patterns using Zero-suppressed Binary Decision Diagrams. In Eliassi-Rad, T.; Ungar, L. H.; Craven, M.; Gunopulos, D., (Eds.) *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM Press, New York, NY, USA, **2006** pp. 307–316.

Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008** 51, 5342–5348.

Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995** 3, 411–428.

Nicolaou, C.; Pattichis, C. S. Molecular Substructure Mining Approaches for Computer-Aided Drug Discovery: A Review. In Fotiadis, D.; Pattichis, C., (Eds.) *Proceedings of the International Special Topic Conference on Information Technology in Biomedicine (ITAB2006).* **2006**.

Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. WOMBAT: World of Molecular Bioactivity. In Oprea, T., (Ed.) *Chemoinformatics in Drug Discovery.* Wiley-VCH, NY, USA, **2004** pp. 223–239.

Papadimitriou, C.; Yannakakis, M. Optimization, approximation, and complexity classes In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing.* ACM Press, New York, NY, USA, **1988** pp. 229–234.

Parker, C. McMaster University Data-Mining and Docking Competition: Computational Models on the Catwalk. *J. of Biomol. Screen.* **2005** 10, 647–648.

Parker, C. N.; Bajorath, J. Towards Unified Compound Screening Strategies: A Critical Evaluation of Error Sources in Experimental and Virtual High-Throughput Screening. *QSAR Comb. Sci.* **2006** 25, 1153–1161.

Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004** 47, 2499–510.

Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992** 32, 331–337.

Quinlan, J. R. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, **1993**.

Ramamohanarao, K.; Bailey, J. Discovery of Emerging Patterns and their Use in Classification. In *Australian Conference on Artificial Intelligence. Lecture Notes in Comput. Sci.*, vol. 2903 Springer, Berlin / Heidelberg, **2003** pp. 1–12.

Rose, S.; Stevens, A. Computational Design Strategies for Combinatorial Libraries. *Curr. Opin. Chem. Biol.* **2003** 7, 331–339.

Rusinko, A.; Farmen, M.; Lambert, C.; Brown, P.; Young, S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999** 39, 1017–1026.

Sadowski, J. Optimization of Chemical Libraries by Neural Networks. *Curr. Opin. Chem. Biol.* **2000** 4, 280–282.

Sadowski, J. 7.1 3D Structure Generation In Gasteiger, J., (Ed.) *Handbook of Chemoinformatics* WILEY-VCH Verlag, Weinheim, **2003** pp. 231–260.

Schapire, R. E. The Strength of Weak Learnability *Mach. Learn.* **1990** 5, 197–227.

Schnur, D.; Beno, B. R.; Good, A.; Tebben, A. Approaches to Target Class Combinatorial Library Design. *Methods Mol. Biol.* **2004** 275, 355–378.

Schweitzer, B.; Dicker, A.; Bertino, J. Dihydrofolate Reductase as a Therapeutic Target *FASEB J.* **1990** 4, 2441–2452.

Stahura, F. L.; Bajorath, J. Partitioning Methods for the Identification of Active Molecules. *Curr. Med. Chem.* **2003** 10, 707–15.

Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.* **1990** 62, 2323–2329.

Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989** 10, 209–220.

Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990** 4, 1–103.

Stockfisch, T. Partially Unified Multiple Property Recursive Partitioning (PUMP-RP): A New Method for Predicting and Understanding Drug Selectivity. *J. Chem. Inf. Comput. Sci.* **2003** 43, 1608–1613.

Stockwell, G. R.; Thornton, J. M. Conformational Diversity of Ligands Bound to Proteins. *J. Mol. Biol.* **2006** 356, 928–44.

Sutherland, J.; O'Brien, L.; Weaver, D. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003** 43, 1906–1915.

Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data Analysis of High-Throughput Screening Results: Application of Multidomain Clustering to the NCI Anti-HIV Data Set. *J. Med. Chem.* **2002** 45, 3082–93.

Terasaka, T.; Kinoshita, T.; Kuno, M.; Nakanishi, I. A Highly Potent Non-Nucleoside Adenosine Deaminase Inhibitor: Efficient Drug Discovery by Intentional Lead Hybridization. *J. Am. Chem. Soc.* **2004** 126, 34–35.

Terfloth, L. Calculation of Structure Descriptors. In Gasteiger, J.; Engel, T., (Eds.) *Chemoinformatics* Wiley-VCH Verlag, Weinheim, **2003** pp. 401–437.

Ting, R. M. H.; Bailey, J. Mining Minimal Contrast Subgraph Patterns In Ghosh, J.; Lambert, D.; Skillicorn, D. B.; Srivastava, J., (Eds.) *Proceedings of the Sixth SIAM International Conference on Data Mining.* **2006** pp. 638–642.

Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors* Wiley-VCH Verlag, Weinheim, **2000**.

Vogt, I.; Bajorath, J. Design and Exploration of Target-Selective Chemical Space Representations. *J. Chem. Inf. Model.* **2008** 48, 1389–1395.

Vogt, M.; Godden, J. W.; Bajorath, J. Bayesian Interpretation of a Distance Function for Navigating High-Dimensional Descriptor Spaces. *J. Chem. Inf. Model.* **2007** 47, 39–46.

Wang, L.; Zhao, H.; Dong, G.; Li, J. On the Complexity of Finding Emerging Patterns. *Theoret. Comput. Sci.* **2005** 335, 15–27.

Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem* **2004** 47, 2977–2980.

Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem* **2005** 48, 4111–4119.

Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999** 39, 868–873.

Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman, **2005**.

Xue, L.; Bajorath, J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *J. Chem. Inf. Comput. Sci.* **2002** 42, 757–764.

Xue, L.; Godden, J.; Stahura, F.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **2003** 43, 1151–1157.

Yakovlev, G.; Mitkevich, V.; Makarov, A. Ribonuclease Inhibitors. *Mol. Biol.* **2006** 40, 867–874.

Zhang; Dong; Ramamohanarao Information-based Classification by Aggregating Emerging Patterns. In Leung, K.-S.; Chan, L.-W.; Meng, H., (Eds.) *Intelligent Data Engineering and Automated Learning - IDEAL 2000, Second International Conference.* Lecture Notes in Comput. Sci. Springer-Verlag, London, UK, **2000** pp. 175–188.

Zhang, X.; Dong, G.; Kotagiri, R. Exploring Constraints to Efficiently Mine Emerging Patterns from Large High-Dimensional Datasets. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM Press, New York, NY, USA, **2000** pp. 310–314.

Zolli-Juran, M.; Cechetto, J. D.; Hartlen, R.; Daigle, D. M.; Brown, E. D. High Throughput Screening Identifies Novel Inhibitors of Escherichia Coli Dihydrofolate Reductase that are Competitive with Dihydrofolate. *Bioorg. Med. Chem. Lett.* **2003** 13, 2493–2496.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation „Emerging Chemical Patterns for Virtual Screening and Knowledge Discovery“ selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006** 46, 2502–2514.

Auer, J.; Bajorath, J. Distinguishing between Bioactive and Modeled Compound Conformation through Mining of Emerging Patterns. *J. Chem. Inf. Model.* **2008** 49, 1747–1753.

Auer, J.; Bajorath, J. Molecular Similarity Concepts and Search Calculations. In Keith, J. M., (Ed.) *Methods Mol. Biol.* 453 Springer, **2008** pp. 327–347.

Auer, J.; Bajorath, J. Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns. *Med. Chem.* **2008** 4, 80–90.

Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008** 51, 5342–5348.

Vogt, I.; Ahmed, H. E.; Auer, J.; Bajorath, J. Exploring Structure-Selectivity Relationships of Biogenic Amine GPCR Antagonists Using Similarity Searching and Dynamic Compound Mapping. *Mol. Diversity* **2008** 12, 25–40.

Bonn, den 17.11.2008

(Jens Auer)