

**CONCEPTS TO INTERFERE WITH PROTEIN-PROTEIN  
COMPLEX FORMATIONS:  
DATA ANALYSIS, STRUCTURAL EVIDENCE AND STRATEGIES  
FOR FINDING SMALL MOLECULE MODULATORS**

DISSERTATION

ZUR

ERLANGUNG DES DOKTORGRADES

DER NATURWISSENSCHAFTEN

(DR. RER. NAT.)

dem

Fachbereich Pharmazie der

PHILIPPS-UNIVERSITÄT MARBURG

vorgelegt von

**Peter Block**

aus Bergisch Gladbach

Marburg an der Lahn, im November 2005



Vom Fachbereich Pharmazie der PHILIPPS-UNIVERSITÄT MARBURG als Dissertation

angenommen am: 15.12.2005

Erstgutachter: Prof. Dr. G. Klebe

Zweitgutachter: Prof. Dr. E. Hüllermeier

Tag der mündlichen Prüfung: 16.12.2005

## **I would like to express my gratitude to**

PROF. DR. GERHARD KLEBE for being a patient supervisor and for supporting this work with ideas, criticism, and encouragement.

PROF. DR. EYKE HÜLLERMEIER for collaboration in the EPIC project and guidance in the field of Machine Learning.

DR. CHRISTOPH SOTRIFFER for helpful discussions and hints during the entire PhD and for the very accurate reviewing of the manuscript. The teamwork in the development of AFFINDB was outstanding.

JURI PÄRN for the enjoyable collaboration in the EPIC project.

the group of PROF. DR. ALFRED WITTINGHOFER, namely DR. ALEXANDER WOLF, DR. CHRISTIAN JELICH-OTTMANN and DR. MICHAEL WEYAND for the collaboration in the 14-3-3 project and their never-ending patience with the *in vitro* testing of compounds.

my room mate HANS VELEC for long and helpful discussions about the projects and the teamwork in the development of *visual* DrugScore.

CHRISTOF GERLACH and MATTHIAS ZENTGRAF for assistance in various screening applications and NILS WESKAMP for the help with Cavbase.

DR. ANDREAS BERGNER for support and helpful discussions, especially regarding the EPIC project.

ANGELA SCHOLZ for invaluable administrative work.

my PARENTS for their love and for their support during my entire PhD.

my girlfriend SARINA for her love, her support, and her never-ending patience.

# TABLE OF CONTENTS

Introduction.....	8
<b>PHYSICOCHEMICAL DESCRIPTORS TO DISCRIMINATE PROTEIN-PROTEIN INTERACTIONS IN PERMANENT AND TRANSIENT COMPLEXES SELECTED BY MEANS OF MACHINE LEARNING ALGORITHMS</b>	
Introduction.....	11
Theory and Methods .....	18
Protein-Protein Interfaces.....	18
ACV with ACE & Sybyl Atom Type Notation (ACV <sup>ACE</sup> , ACV <sup>Sybyl</sup> ).....	20
DrugScore Contact Vectors (DCV) .....	21
SFCscore Descriptor Vectors (SDV).....	22
Machine Learning Algorithms .....	22
Support Vector Machines (SVM).....	24
Decision Trees (C4.5).....	25
K Nearest Neighbors (KNN).....	27
Naïve Bayes (NB).....	27
Feature Selection Approaches .....	29
Filter Approach.....	29
Wrapper Approach.....	30
Genetic Algorithms.....	31
Feature Analysis .....	32
Results and Discussion .....	33
Classification of Monomer versus Homodimer Complexes .....	33
Classification of Folding Complexes versus Recognition Complexes .....	35
Feature Analysis with Genetic Algorithms .....	39
Summary and Conclusions .....	49
References.....	51

# TABLE OF CONTENTS

## STRATEGIES TO SEARCH AND DESIGN STABILIZERS OF PROTEIN-PROTEIN INTERACTIONS: A FEASIBILITY STUDY

Introduction.....	55
Materials and Methods.....	57
Data Analysis and Tools for Virtual Screening .....	57
Stabilizers of Protein-Protein Interactions.....	58
Screening for Novel Targets .....	58
The H <sup>+</sup> -ATPase/14-3-3 System.....	60
The Fusicoccin Binding Site .....	61
Water and the Fusicoccin Binding Pocket.....	63
Virtual Screening Campaigns.....	65
Preprocessing of the Candidate Molecules.....	67
Docking Fusicoccin .....	68
“Hot Spot” Analysis .....	69
FTrees .....	70
Unity Database Search .....	72
Scoring.....	75
Pharmacophore Post-Filtering .....	75
Visual Inspection .....	77
Docking with Gold and AutoDock.....	77
Visual DrugScore.....	78
Results and Discussion .....	78
Virtual Screening for Stabilizers of the H <sup>+</sup> -ATPase/14-3-3 Interaction .....	86
Summary and Conclusions .....	100
References.....	102

# TABLE OF CONTENTS

## **AFFINDB: A FREELY ACCESSIBLE DATABASE OF AFFINITIES FOR PROTEIN-LIGAND COMPLEXES FROM THE PDB**

Introduction.....	111
Methods.....	113
Database Architecture .....	113
Data Collection and Database Content .....	113
Database Access .....	118
Discussion.....	119
References.....	123

### **SUMMARY**

Zusammenfassung.....	127
----------------------	-----

### **APPENDIX**

Publications Arising from this Work.....	132
Articles .....	132
Posters .....	132
Awards.....	133
Curriculum Vitae.....	134

## INTRODUCTION

“LIFE IS CONTROLLED BY OVER 50.000 PROTEIN-PROTEIN INTERACTIONS.”

(ANDREW HAMILTON, YALE UNIVERSITY, MARCH 2005, INTERNATIONAL WORKSHOP NAD3 IN RAUISCHHOLZHAUSEN, GERMANY)

**P**rotein-Protein interactions are playing a crucial role in virtually any biological system. Over the past 10 year great efforts has been made to find molecules, which modulate such interactions, since the modification of protein-protein interactions promise a valuable target. The market for the area of protein-protein interactions is expected to reach over \$50 billion by year 2010.

**I**t has been assumed that all proteins in a cell are forming an extended network with non-covalent interactions continuously forming and dissociating. Thus, the detection of specific protein-protein interactions and the determination of their affinity are of pivotal interest. Therefore, the observation of protein-protein interactions generated a multiplicity of methodologies, in particular yeast two-hybrid systems, phage display, and BIAcore. Additionally, lots of databases has been created to gather the huge amount of cumulating data in order to enlighten networks of interacting proteins in a general or target-specific way, such as the Database of Interacting Proteins (DIP) or the Mammalian Protein-Protein Interaction database (MPPI). Furthermore, different tools has been developed to extract information from the given protein-protein interactions and their complexes. Consequently, it has been shown, that the driving forces for such interactions include electrostatic forces, hydrogen bonds, van der waals forces and hydrophobic effects. It has been shown, that hydrophobic effects drive protein-protein interactions, whereas hydrogen bonds and electrostatic interactions govern the specificity of the interface.

**T**ransient associations between proteins spread across wide range of biological processes, which includes signal transduction, antibody against antigen reactions, hormone-receptor binding, and repairing actions by chaperones. In contrast, permanent



protein-protein complexes are essential in areas where the stability or function are defined by a multimeric state. Those complexes, for instance, are located in muscle fibres or participate at the formation of intracellular structures (e.g. microtubules). The importance of protein-protein interactions can be imposingly illustrated by the mode of action of G Protein Coupled Receptors (GPCRs), for instance the  $\beta$ -adrenergic receptor. The heterotrimeric so-called G protein is attached to the cytoplasmic part of the transmembrane located receptor. It is composed out of two permanently associated subunits,  $G\beta$  and  $G\gamma$ , and the GTP-binding subunit  $G\alpha$ . Upon the binding of an agonist hormone (in this case adrenaline) to the extracellular face of the GPCR, the intracellular part of the receptor is triggered to change its conformation. Thereby, the subunit  $G\alpha$  gets induced to exchange the bound GDP into energetically more active GTP and dissociates from the still assembled  $G\beta$  and  $G\gamma$  subunits. The activated  $G\alpha$  subunit is known to subsequently activate the enzyme adenylyl cyclase, which is also anchored within the intracellular membrane. Accordingly, the production of cyclic AMP (cAMP), which is often termed as second messenger, is promoted. Further on, cAMP-dependent kinases may get stimulated by cAMP and precipitate a cascade of subsequent reactions, which are also mediated by protein-protein interactions.

As we have seen in this example, the current state, whether protein-protein complexes have permanent character or tend to dissociate, can be triggered, either by conformational changes and/or electrostatic effects (phosphorylation by kinases). Therefore, it is of basic interest to understand the physicochemical properties, which leads protein complexes to disassemble. Accordingly, the specific modulation of desired protein-protein interactions allows the manipulation of distinct metabolic pathways and may finally lead to complete new classes of drugs. Thus, the distinction between permanent and transient complexes and the extraction of their discriminating features are of extraordinary interest, as we have seen for the G-protein complex ( $G\alpha$  and  $G\beta/G\gamma$ ). Furthermore, strategies to modulate protein-protein interactions in a specific way are highly desirable.

Therefore, this work focus on the one hand side on the extraction of discriminating features between different protein-protein complexes. This is achieved by applying algorithms from the area of Machine Learning in combination with Feature Selection methods. We discriminate between permanent and transient protein-protein complexes and developed an approach to quantify the relevance of the descriptors, which were applied for the classification. This work also focus on strategies to search and design stabilizers of protein-protein interactions. The approach of stabilizing protein-protein interactions appears as a promising alternative to the classical approach of inhibiting those. Finally, this work describes a database, which was developed with the goal to collect affinity data for given crystal structures of protein-ligand complexes. The database holds over 730 affinity values, handcurated from literature, for protein complexes of the Protein Databank (PDB) and is a valuable source for deriving regression-based scoring functions.

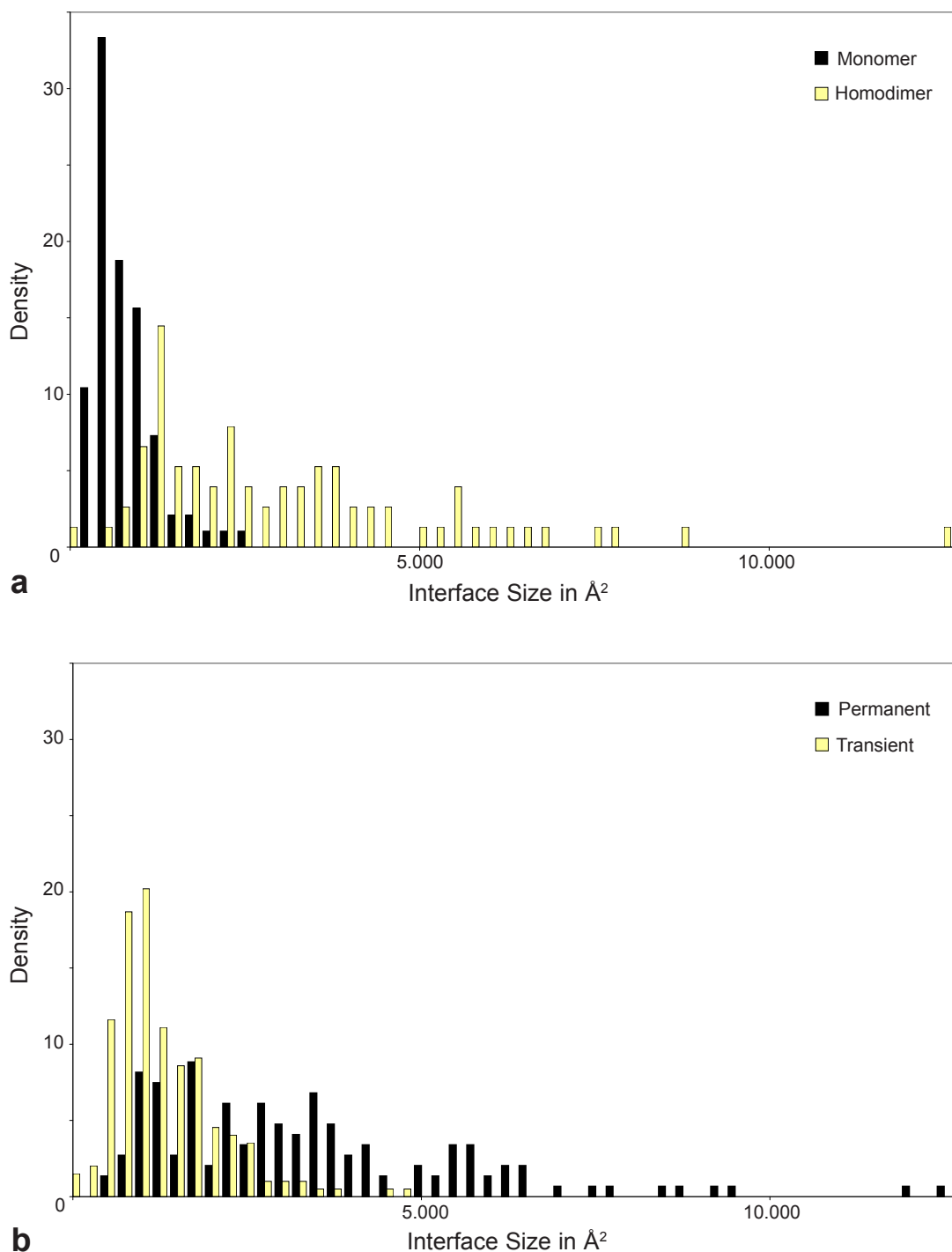
# PHYSICOCHEMICAL DESCRIPTORS TO DISCRIMINATE PROTEIN- PROTEIN INTERACTIONS IN PERMANENT AND TRANSIENT COMPLEXES SELECTED BY MEANS OF MACHINE LEARNING ALGORITHMS

## INTRODUCTION

In recent years substantial efforts have been dedicated to the investigation of protein-protein interactions since they play a crucial role in nearly every biological system. Essential protein-protein interactions have proven to be highly specific and diverse at the same time (Sheinerman *et al.*, 2000; Nooren & Thornton, 2003; Liddington, 2004). The ability to modulate protein-protein interactions could, therefore, be of interest in drug research and possibly provides access to new classes of targets. Accordingly, specific modulation of selected protein-protein interactions may lead to completely new classes of drugs. Protein-protein interaction complexes may belong to different kinds of categories: permanent or transient; biological or crystallographic. Permanent complexes are naturally formed during protein biosynthesis, thus they are also known as *folding* complexes, whereas transient complexes are those that form temporarily. Especially the mutual recognition of proteins via comparatively large surface interfaces plays a key role in most signaling pathways (e.g. G-Protein Coupled Receptors) or in hormone-receptor binding. Therefore, this type of complexes is also known as *recognition* complexes. Some authors have referred to these classes also as *two-state* (permanent) or *three-state* (transient) complexes (Tsai *et al.*, 1997a). On the other hand, crystallographers are often faced with question whether a protein-protein complex in a crystal structure is the biologically functional form or a crystallographically enforced crystal contact which is biologically not relevant (*crystal contact*). According to prior publications we adopt the definition of crystal-contact complexes as *monomer* structure, whereas biological complexes are termed *homodimers* (Ponstingl *et al.*, 2000).

Despite the considerable importance of protein-protein interactions, there is only little knowledge about the mechanisms resulting in stable protein complexes that dissociate under certain physiological conditions. To obtain better insights into these mechanisms, closer investigations at the atomic level are required. Descriptors accounting for the nature of protein complex formation are highly desired, primarily to classify various protein complexes. In recent years, several reviews (Chakrabarti & Janin, 2002; Jones & Thornton, 1996; Lo Conte *et al.*, 1999) have summarized the general physicochemical properties of protein complexes, aimed at the discovery of appropriate descriptors to distinguish different types of protein complexes. It has been shown that the average amino acid propensity in the interfaces of permanent complexes is comparable to that observed in the core of proteins, whereas transient complexes are rather composed as a mixture of core and solvent-exposed amino acids (Tsai *et al.*, 1997b). This is due to the fact that transient complexes need to be soluble when dissociated. Permanent complexes are normally stable complexes under physiological conditions, thus their interfaces show the hydrophobic character similar to the interior of an average globular protein (Lo Conte *et al.*, 1999, Jones & Thornton, 1996). Based on a statistical analysis, Tsai *et al.* (Tsai *et al.*, 1997b) also found that protein-protein interfaces are more hydrophobic than the exterior of proteins, yet more polar than the interiors of proteins. This may be due to the fact that this intermediate character is needed to promote association of the individual monomers. Lo Conte *et al.* calculated a similar packing density in the center of protein-protein interfaces in comparison to the interior of proteins (Lo Conte *et al.*, 1999). The driving force to stabilize protein-protein complexes is considered to be the hydrophobic effect (Dill, 1990), although it is not supposed to be the driving force in the association process (Sheinerman *et al.*, 2000).

Furthermore, permanent complexes tend to have larger and more twisted interfaces than transient ones. Especially the size of the interface in non-physiological complexes (e.g. crystal contacts) is a major discriminant to functional homodimer complexes, since crystal contacts have an average interface size in the range of 550 ( $\pm$  270) Å<sup>2</sup>, whereas



**FIGURE 1.** Distribution of protein-protein interfaces sizes in Å². **(a)** The distribution of the homodimer complexes (yellow) are wide spread from about 200 to 14.000 Å², whereas monomer crystal contact complexes (black) show a small distribution from about 200 to 3.000 Å². **(b)** Permanent complexes (black) show a wide spread distribution from about 200 to 14.000 Å², whereas transient complexes (yellow) show smaller distribution from about 200 to 5.500 Å².

physiologically stable homodimers normally exceed a size of  $> 1900 (\pm 1200) \text{ \AA}^2$  (Fig. 1). Nevertheless, it has to be kept in mind that biological systems such as protein complexes do not obey in all cases a straight-forward classification in either permanent or transient complexes. While splitting permanent and transient complexes in terms of the amino acid propensities across their interfaces, we are faced with the fact that a large proportion of protein complexes satisfies both criteria and would fall into both categories. Despite of all the knowledge and understanding about the properties of protein complexes today, it is not possible to easily classify them in terms of simple sequence-based descriptors.

This study focuses on the classification of crystal contact complexes versus functional homodimer complexes, as well as permanent versus transient complexes. We aim to establish a classification on a physicochemical basis by representing the protein-protein interfaces by their physical and chemical properties at the atomic level. The first study on the classification of protein-protein complexes was introduced by Ponstingl *et al.* (Ponstingl *et al.*, 2000) using statistical atom pair potentials to distinguish functional contacts from artificial contacts enforced by the crystallographic packing environment (crystal contacts). Ponstingl *et al.* were able to classify a dataset with 96 monomers and 76 homodimers with a success rate of 88.9%. This rate indicates a small but significant improvement over the pure measure of accessible surface area (ASA) buried by the interface-contributing atoms, which leads to a prediction rate of 84.6% (Ponstingl *et al.*, 2000). The continuing work of Valdar *et al.* (Valdar & Thornton, 2001) enhanced the prediction accuracy to 92% by applying sequence conservation properties. To establish the correlation they also used an optimizing artificial neural network. Bahardur *et al.* dissected specific and non-specific protein-protein interfaces by applying statistical parameters. The combination of non-polar interface area, fraction of buried interface atoms and residue propensity score of the interfaces resulted in ~94% prediction accuracy (correctly classified complexes) using a dataset comprising 188 monomers (crystal contacts) and 122 homodimers (Bahadur *et al.*, 2004). Mintseris *et al.* (Mintseris & Weng, 2003) introduced the concept of Atomic Contact Vectors (ACV) for taking advantage of

the power of methods in multivariate analysis and pattern recognition. They evaluated the dataset initially used by Ponstingl *et al.* with *Quadratic Fisher Discriminant* (QFD) and *Kernel Discriminant Analysis* (KDA) and achieved an impressive prediction accuracy of 93%. Mintseris *et al.* were the first to classify protein complexes in terms of being either permanent or transient in biological systems, which corresponds to differentiation into permanent or transient complexes. They compiled a dataset of 345 structures comprising 147 permanent and 198 transient complexes with a maximum of 25% sequence identity according to pairwise BLAST all-against-all sequence comparisons. By applying KDA, they achieved the remarkable prediction accuracy of 91.0% in a *leave-one-out* cross-validation. The method QFD fell back with only 80% correctly predicted complexes.

<b>Monomers</b>										
16PK	1AHQ	1AYI	1BN8	1DFF	1FSU	1MB1	1PJR	1VJW	2CY3	2RN2
1A0K	1AKO	1AYL	1BP1	1DJX	1GCI	1MDT	1PMI	1XGS	2END	3CMS
1A19	1AKZ	1BC2	1BRY	1DMR	1IAE	1MH1	1PPO	1YGE	2FGF	3DFR
1A6Q	1AM6	1BE0	1BU1	1EMA	1INP	1MPG	1PS1	1ZIN	2GPR	3SIL
1A80	1AMJ	1BEA	1BWZ	1ESF	1IPS	1NP4	1RGP	232L	2HEX	5CP4
1AAY	1AOH	1BG0	1C3D	1ESO	1KFS	1NUC	1RHS	2ABX	2IHL	8PAZ
1AF7	1AUA	1BGC	1CKI	1FDR	1KPT	1OPS	1TON	2ACY	2MBR	
1AFK	1AUN	1BKZ	1CKM	1FEH	1KWA	1PDA	1UCH	2ATJ	2MHR	
1AH7	1AVP	1BMB	1CTJ	1FLP	1LRV	1PGS	1URO	2BLS	2PTH	
<b>Homomers</b>										
1A3C	1AOM	1BUO	1CZJ	1ICW	1KPF	1OAC	1RPO	1TRK	2CCY	3SDH
1AD3	1AOR	1CG2	1DAA	1IMB	1LYN	1OPY	1SES	1TYS	2ILK	3SSI
1AF5	1AQ6	1CHM	1FIP	1ISA	1MJL	1OTP	1SLT	1UBY	2RSP	4KBP
1AFW	1AUO	1CMB	1FRO	1ISO	1MKA	1PGT	1SMN	1UTG	2TCT	5CSM
1AJS	1BAM	1CP2	1GVP	1JHG	1MOQ	1PRE	1SMT	1VLB	2TGI	5TMP
1ALK	1BIF	1CSH	1HJR	1JSG	1NOX	1PUC	1SOX	1WGJ	3GRS	9WGA
1AMK	1BSR	1CTT	1HSS	1KBA	1NSY	1RFB	1TOX	1XSO	3PGH	

TABLE 1. PDB codes of dataset A, listed by their class.

**Permanent**

1A3C	1AUO	1CMB	1EUD	1GVP	1I7B	1KFU	1PGT	1RPO	1WGJ	3SDH
1A6D	1AW8	1CP2	1EXB	1H2R	1IAK	1KPF	1PHN	1RTH	1XSO	3SSI
1A9X	1B4U	1CSH	1FCD	1HCN	1ICW	1KVD	1PRE	1SES	2AHJ	4KBP
1AD3	1B5F	1CTT	1FIP	1HFE	1IHF	1LUC	1PUC	1SKY	2CCY	4MON
1AF5	1B7Y	1CZJ	1FM2	1HGE	1IMB	1LYN	1QDU	1SLT	2ILK	5CSM
1AFW	1BAM	1D09	1FRO	1HJR	1IRD	1MJL	1QGW	1SMN	2LTN	5TMP
1AHJ	1BIF	1D2V	1FS0	1HLR	1ISA	1MKA	1QH1	1SMT	2RSP	9WGA
1AJS	1BMV	1DAA	1FXW	1HR6	1ISO	1MOQ	1QLA	1SOX	2TCT	
1ALK	1BSR	1DCE	1G72	1HSS	1JHG	1NOX	1QLA	1SPP	2TGI	
1AMK	1BUO	1DJ7	1G8J	1HXM	1JK0	1NSY	1QOP	1TOX	3GRS	
1AOM	1CCW	1E9Z	1G8K	1HZZ	1JRO	1OAC	1QS0	1TRK	3GTU	
1AOR	1CD1	1EFV	1GK9	1I1Q	1JSG	1OPY	1QTN	1TYS	3PCG	
1AQ6	1CG2	1EG9	1G03	1I3R	1JV2	1OTP	1REQ	1UBY	3PGH	
1AUI	1CHM	1EP3	1GOT	1I4F	1KBA	1PAU	1RFB	1UTG	3RUB	

**Transient**

1A2K	1B6C	1CMX	1DS6	1EV2	1FQV	1HJA	1IM3	1K90	1RRP	2BTF
1A2Y	1BDJ	1CN4	1DTD	1F02	1FSK	1HWG	1IM9	1K90	1SGP	2HMI
1A4Y	1BGX	1CSE	1DU3	1F34	1FYH	1HX1	1IOD	1KAC	1SLU	2JEL
1ACB	1BKD	1CXZ	1DX5	1F3V	1G0Y	1HYR	1IQ5	1LFD	1STF	2PCC
1ADQ	1BLX	1CZ8	1E0F	1F51	1G4Y	1HYS	1IQD	1LPB	1T7P	2SIC
1AHW	1BML	1D2Z	1E44	1F5Q	1G73	1I1R	1ITB	1MKX	1TBR	2VIR
1AIP	1BQQ	1D4V	1E6J	1F60	1G9M	1I2M	1J7V	1NRN	1TMQ	3YGS
1AK4	1BTH	1D5M	1E96	1F7Z	1G9M	1I4D	1JDH	1NSN	1TNR	4SGB
1APM	1BUH	1D6R	1EAY	1F93	1GCQ	1I40	1JDP	1OSP	1TX4	
1ARO	1BVN	1DE4	1EBD	1FAK	1GH6	1I5K	1JHL	1QA9	1UGH	
1ATN	1BZQ	1DEE	1EBP	1FAK	1GL0	1I7W	1JIW	1QAV	1VPP	
1AVA	1C1Y	1DEV	1EFN	1FC2	1GL1	1I8L	1JLT	1QBK	1VRK	
1AVG	1C4Z	1DF9	1EFU	1FE8	1GL4	1IAR	1JMA	1QGR	1WEJ	
1AVW	1CA0	1DF9	1EGJ	1FJ1	1GOT	1IB1	1JPS	1QKZ	1WQ1	
1AXI	1CD9	1DHK	1EJA	1FLT	1HCF	1IBR	1JRH	1QMZ	1WWW	
1AY7	1CHO	1DKG	1EMV	1FNS	1HE1	1ICF	1JTD	1Q00	1XDT	
1AZZ	1CIC	1DN1	1E08	1FOE	1HE8	1ID5	1JTG	1Q03	1YCS	
1B0N	1CLV	1DPJ	1ES7	1FQ1	1HEZ	1IHS	1JTP	1QTY	1ZBD	
1B41	1CM4	1DQJ	1EUV	1FQK	1HIA	1IIS	1K4C	1RLB	2BTC	

**TABLE 2.** PDB codes of dataset B, listed by their class.



The work of Ofra *et al.* (Ofra & Rost, 2003) did not rely primarily on derived information to classify different protein complexes, but introduced an information theory-based analysis approach by combining information from different databases. By means of their comprehensive data set they were able to apply simple statistical methods to find significant correlations between the amino acid distributions in protein-protein interfaces and the protein complex type. Ofra *et al.* distinguish six types of protein-protein complexes: (1) interfaces within one structural domain, (2) interfaces between different domains within one chain, (3) interfaces between permanently interacting identical chains, (4) interfaces between transiently interacting identical protein chains, (5) interfaces between permanently interacting different protein chains and (6) interfaces between different transiently interacting protein chains.

In the present work we focus on the classification between crystal contact complexes and functional homodimer complexes using the dataset of Ponstingl *et al.* with 172 structures (Ponstingl *et al.*, 2000) (dataset A, Tab. 1), as well as on the classification between permanent and transient complexes by means of the dataset compiled by Mintseris *et al.* (Mintseris & Weng, 2003) (dataset B, Tab. 2). The classification is carried out by algorithms from the field of machine learning. We compare the prediction accuracy with four different algorithms that can be considered for a wide spectrum of approaches in the field: Support Vector Machines (SVM), Decision Trees (C4.5), K Nearest Neighbors (KNN) and Naïve Bayes (NB).

Although SVM is a relatively new approach (Schoelkopf, 2002), there are already several studies in literature which use this algorithm to predict properties of protein-protein interaction from primary sequence data (Bock & Gough, 2001), the location of protein-protein binding-sites (Bradford & Westhead, 2005, Lo *et al.*, 2005) or protein-protein interface residues (Yan *et al.*, 2004). SVM and Decision Trees were implemented to predict beta-sheet recognition in protein-protein complexes (Siepen *et al.*, 2003) and other machine learning algorithms such as artificial neural networks have also been

utilized to predict protein-protein interaction sites (Fariselli *et al.*, 2002, Zhou & Shan, 2001). SVM and the combination of KNN with Genetic Algorithms were furthermore applied to classify molecules by “kinase inhibitor-likeliness” recently (Briem & Guenther, 2005).

In this study, the aforementioned classification algorithms have been combined with different feature selection methods, namely: Filter and Wrapper as well as Genetic Algorithms. This combination enables us to extract those descriptors from the protein complexes that are supposed to best discriminate among the different types of complexes. As data representation to describe the protein-protein complexes we adapted on the one hand the ACVs with the atom type notation ACE (Zhang *et al.*, 1997) from Mintseris *et al.* (ACV<sup>ACE</sup>), on the other hand we created new alternative representations: Atomic Contact Vectors using Sybyl atom type notation (ACV<sup>Sybyl</sup>), DrugScore Potential Vectors (DCV), and SFCscore Descriptor Vectors (SDV). In particular the latter concentrate on physicochemical properties across the interface.

## THEORY AND METHODS

### PROTEIN-PROTEIN INTERFACES

The classification of monomers and homodimers is based on the Ponstingl *et al.* dataset with 172 structures comprising 96 crystal contact complexes and 76 functional homodimers. The classification of permanent and transient complexes is based on the Mintseris *et al.* dataset with 345 structures comprising 147 permanent and 198 transient complexes. The atom coordinates of all structures can be obtained freely from the Protein Data Bank (PDB) operated by the Research Collaboratory for Structural Biology (Berman *et al.*, 2000). Protein structures in the PDB are always deposited as their asymmetric unit, which does not necessarily correspond to the biological unit of the protein. In fact, monomers often represent the entire asymmetric unit in crystal structures, especially in

the homomeric complexes. Since our work relies on protein complexes, we first had to generate the crystal packing to obtain the biologically functional unit of all proteins. The program SYBYL (see ref.: SYBYL 7.0) was used to apply the symmetry operations, deposited in the PDB data file, on the monomers. Every chain in each dataset shows a maximal sequence identity of 25% with other entries, a commonly accepted level to consider protein chains as non-homologous. All structures were reviewed against the original publication with respect of being either a monomer or a homodimer complex, or a permanent or transient complex (Ponstingl *et al.*, 2000, Mintseris & Weng, 2003). Most protein-protein complexes are composed of two chains, clearly defining the protein-protein interface, whereas some protein-protein complexes are constructed of three or even four chains. In the latter case, we had to take into account to which part of the biologically relevant complex each chain belongs. Technically chains belonging to one “side” of the interface were considered as one single chain. An accurate definition of the protein-protein interface is essential since in the following we consider only the atoms contributing to the interface and derive information about atom-atom contacts across the interface entirely on the basis of this assignment. All atom coordinates and additional meta information were stored in a database for easy and convenient access and data management. All protein-protein complexes were inspected visually to assess the relevance of the assignments of the complexed protein chains.

The data representation of protein complexes is a crucial step in using machine learners for classification and feature selection, since the prediction accuracy is strongly dependent on the quality of the input data. Furthermore, a sophisticated data representation is mandatory for a meaningful discussion of the selected features. Therefore, all atoms contributing to the protein-protein interface have to be determined. Two different methods have been described in literature for this task: In the first one, all atoms from one side of the interface contacting an atom on the other side within a given distance cutoff are classified as being part of the interface. The cutoff is set to 6 Å, assuming attractive interatomic contacts up to this distance. This method is an accurate

and fast way to retrieve all interface atoms. In the second, all atoms changing their accessible surface area (ASA) upon *in silico* complex formation (Hubbard & Thornton, 1993, Sanner *et al.*, 1996) are considered ( $\Delta$ ASA). From a theoretical point of view the latter method appears very relevant, but all atoms which are already completely buried in the uncomplexed state are neglected since obviously these atoms cannot change their ASA upon complex formation. Furthermore, different results are obtained when using different algorithms to calculate ASAs. Nevertheless, comparative evaluations applying the distance cutoff method and the  $\Delta$ ASA method show very similar results with respect to the assigned interface atoms (data not shown). Accordingly, we decided to work with the distance cutoff method particularly for reasons of comparability with results from other publications.

#### ACV WITH ACE & SYBYL ATOM TYPE NOTATION ( $ACV^{ACE}$ , $ACV^{SYBYL}$ )

Encouraged by the convincing results achieved with Atomic Contact Vectors (ACV) to represent protein-protein interfaces (Mintseris & Weng, 2003), we decided to adopt this approach. ACVs are simple vectors holding counts of occurrences of atom-atom contacts across the protein-protein interface, categorized with respect to the contact type. Only atoms in the protein-protein interface were considered for the generation of the ACVs. For comparative purposes, we followed the approach of Mintseris *et al.* and created ACVs with the atom type notation of Atomic Contact Energy (ACE) (Zhang *et al.*, 1997). In the following we call these vectors  $ACV^{ACE}$ . The atoms of the 20 proteinogenic amino acids are described by 18 different ACE atom types. This leads to a vector size of 171.

Encouraged by the convincing results achieved with Atomic Contact Vectors (ACV) to represent protein-protein interfaces (Mintseris & Weng, 2003), we decided to adopt this approach. ACVs are simple linear vectors holding counts of occurrences of atom-atom contacts across the protein-protein interface, categorized with respect to the contact type. Only atoms in the protein-protein interface were considered for the generation of

the ACVs. For comparative purposes, we followed the approach of Mintseris *et al.* and created ACVs with the atom type notation of Atomic Contact Energy (ACE) (Zhang *et al.*, 1997). In the following we call these vectors  $ACV^{ACE}$ . The atoms of the 20 *proteinogenic* amino acids can be described by 18 different ACE atom types. This leads to a vector size of 171 attributes:  $\binom{18}{2} + 18$ . On the other hand, we created ACVs based on the widely used Sybyl atom type notation, derived from the TRIPOS FORCE FIELD (Clark *et al.*, 1989). The atoms of the proteinogenic amino acids can be represented by 12 different Sybyl atom types, which leads to 78 different atom pairs:  $\binom{12}{2} + 12$ . These vectors are named  $ACV^{Sybyl}$ .

### DRUGSCORE CONTACT VECTORS (DCV)

The original implementation of ACVs simply counts the raw occurrences of specific atom-atom contacts across the protein-protein interface. This approach, however, does not consider that, depending on their mutual distance, contact pairs can differ dramatically with respect to their contribution to binding affinity. This fact is taken into account by knowledge-based scoring functions, which use distance-dependent atom-atom pair potentials extracted from distributions in comprehensive structural databases.

Statistical pair-potentials were derived from the two datasets in analogy to the DrugScore scoring function (Gohlke *et al.*, 2000). In contrast to the original DrugScore implementation, where pair potentials were derived from protein-ligand complexes, here pair potentials were compiled using protein-protein complexes. The distance cutoff threshold was maintained at 6 Å. Instead of the raw contact counts in case of the original ACV, we derived scored atom-atom contacts in the DCVs. DrugScore potentials were only collected for Sybyl atom types corresponding to the original atom type list used in the genuine DrugScore implementation. The scores for every atom-atom contact type in the protein-protein interface is summed up to serve as a descriptor.

### SFCSCORE DESCRIPTOR VECTORS (SDV)

SFCscore has recently been developed as an empirical regression-based scoring function based on descriptors for protein-ligand complexes (Sottriffer *et al.*, *in preparation*). SFCscore provides descriptors for hydrogen bonding, ring interactions, rotatable bonds, and hydrophobic complementarity, as well as a variety of other surface measures. The SFCscore library was attached to the EPIC library in order to subject this set of descriptors also to the characterization of protein-protein complexes. We generated a set of 63 SFCscore descriptors for each chain in every protein-protein interface of the dataset. These descriptors served as input vectors for the machine learning algorithms.

Not all of the considered descriptors are of relevance for the protein-protein interface classification, e.g. the metal scores. Since there are no metals in the interfaces of the protein-protein complexes selected for this study, the score is always 0. Attributes adopting the same value across all vectors are not considered by the machine learners accordingly, such irrelevant descriptors do not interfere with the results of the classification. They rather serve as reference attributes for the following feature selection process.

### MACHINE LEARNING ALGORITHMS

In this work we are dealing with supervised learning, i.e. the learning system is provided with  $n$  labeled patterns. Every pattern is a vector of  $m$  features

$$x_i = (a_1, a_2, \dots, a_m) \in \text{dom}(A_1) \times \dots \times \text{dom}(A_m) \quad (1)$$

in this work  $\text{dom}(A_i) = N$ ,  $i = 1, \dots, m$ . Consequently, a pattern is a point in the input space. Furthermore, for the training set  $S$  every pattern is labeled with a class  $y_i \in Y \subset N$ .

$$S = \{(x_i, y_i) | x_i \in X, y_i \in Y, i = 1, \dots, n\} \quad (2)$$

The elements of  $S$  will be referred to as examples. We assume that we have an independent and identically distributed set of examples, i.e. the data are generated from an unknown but fixed probability distribution:

$$Pr(S) = \prod_{i=1}^n Pr(x_i, y_i) = \prod_{i=1}^n Pr(x_i) Pr(y_i|x_i) \quad (3)$$

The goal in supervised learning is to compute a hypothesis from the examples which is able to correctly classify a new and previously unseen patterns  $(x,y)$  that are also taken from  $Pr(x,y)$ . Accordingly, we try to find a functional relationship between patterns and labels:

$$h: X \rightarrow Y \quad (4)$$

We use  $h$  in order to classify new patterns on a hypothetical basis. The generalization performance of the hypothesis  $h$  is measured in terms of its classification accuracy, i.e., the more accurate new patterns are classified by  $h$ , the better this hypothesis generalizes beyond the observed data. More specifically, the classification accuracy of  $h$  can be measured by means for a loss function:

$$l: Y \times Y \rightarrow [0, \infty] \quad (5)$$

with  $l(y, y) = 0$  for all  $y \in Y$ . For discrete sets  $Y$  we can use:

$$l(y, h(x)) \begin{cases} 0: h(x) = y \\ 1: h(x) \neq y \end{cases} \quad (6)$$

and for real valued sets of  $Y$ :

$$l(y, h(x)) = (h(x) - y)^2 \quad (7)$$

The first loss function is referred to as *0/1-loss*, the second as *mean square error*. It is important to notice that minimizing the training error or empirical risk

$$R_{emp}[h] = \frac{1}{n} \sum_{i=1}^n l(y_i, h(x_i)) \quad (8)$$

does not imply a small risk averaged over test examples extracted from the underlying distribution  $\Pr(x,y)$ . Risk is defined as:

$$R[h] = \int_{x \times y} l(y, h(x)) dPr(x, y) \quad (9)$$

*Over-fitting* occurs if  $R_{emp}$  is very small or 0 and  $R$  is very large. In this case, the system is able to classify the training set very well, but performs poorly in classifying new and unseen patterns. Since the underlying distribution  $\Pr(x,y)$  is not known we only can estimate  $R$  from the training set. One of the simplest is *k-fold* cross-validation with the special case of *leave-one-out* cross-validation. In *k-fold* cross-validation the training set is divided into  $k$  (approximately) equal-sized subsets. The learning algorithm is trained with  $k-1$  subsets and the omitted subset is used to estimate the error. This procedure is repeated  $n$  times. In the case that  $k$  equals the sample or training set size, it is called *leave-one-out* cross-validation.

## SUPPORT VECTOR MACHINES (SVM)

A relatively new approach in machine learning is provided by Support Vector Machines (SVM) (Schoelkopf, 2002; Cortez & Vapnik, 1995). SVMs try to separate data by means of optimally assigned hyperplanes. The basis of SVMs are linear functions or hyperplanes. From the set of possible hyperplanes, SVMs select the one which has the maximum margin. It can be shown that SVMs follow the so-called principle of structural risk minimization (Vapnik, 1998). To deal with data that is not linearly separable in the original input space, the idea of SVMs is to map the data into a higher dimensional



*feature space*. A linear hyperplane in this *feature space* corresponds to a non-linear decision boundary in the original input space. The mapping from the latter to the former is accomplished by means of so-called *kernel* functions. It is important to mention, however, that this mapping is never computed explicitly. Instead, the optimization problem of finding a hyperplane with maximal margin in the feature space can be solved implicitly, a property that makes SVMs computationally tractable for high-dimensional or even infinite-dimensional *feature spaces*. We applied the RBF kernel, since it produced the best results in test scenarios.

We implemented the freely available library LIBSVM (Chang & Lin, 2001) into EPIC. LIBSVM delivers a Python interface, which simplified the implementation in the Python-based EPIC library and speeds up the classification since no I/O operations are necessary.

#### **DECISION TREES (C4.5)**

**D**ecision Trees generate a tree structure which can be exploited to classify data. Since tree representations are well readable and their classification performance is powerful, Decision Trees are a very popular machine learner. Decision Trees are composed of three elements: nodes, edges, and leafs, where nodes without children are termed leafs. Nodes represent tests with respect to a feature. For every possible answer of the test, there is an edge to a child node. Every leaf represents a classification. A pattern can be classified by testing the pattern according to a property in every node starting with the root. Based on the value of the feature in the pattern, one child of the node is chosen. If the child is a leaf, the pattern becomes the class that is represented by this leaf. To create a Decision Tree the set of examples  $S$  will be recursively separated on the basis of a feature (Algorithm 1: Line 5-11). If all examples belong to the same class, the separation ends and the resulting leaf is returned (Algorithm 1: Line 2 and 3).

```

1   Decision Tree(S,A)
2   if all examples belong to the same class c then
3       return leaf marked with class c
4   else
5       Select a feature  $A_i \in A$ 
6       create node r with label  $A_i$ 
7       for all  $v_1, \dots, v_m \in \text{dom}(A_i)$  do
8            $S_{v_i} = \{x \in S \mid p(A_i(x), v_i)\}$ 
9            $B_{v_i} = \text{Decision tree}(S_{v_i}, A \setminus \{A_i\})$ 
10          add  $B_{v_i}$  as child to r with test  $p(A_i(x), v_i)$ 
11          end for
12          return r
13  end if

```

**ALGORITHM 1.** General algorithm for generating decision trees.  $S$  is a set of (classified) examples,  $A$  is the set of features and  $p$  is a predicate.

**Decision tree learning** is a heuristic approach that aims at inducing simple models, i.e., small trees, since such models are supposed to generalize better than complex models. The complexity of a decision tree strongly depends on the choice of suitable splitting attributes in line 6 of the algorithm. In this connection, potential candidate features are evaluated in terms of an information measure such as, e.g., the information gain which is based on the well-known Shannon entropy, and the “most informative” attribute is selected. To avoid *over-fitting* it is useful to “prune” the trees either in the building process or in a post-processing step. We used the original C4.5 algorithm of Quinlan (Quinlan, 1993), which is freely available on the Internet. Since this code is written in C, it was easy to write a wrapper application to implement it in the Python-based EPIC library. This speeds up the classification by the factor of 10, since no speed-reducing I/O operations are required.

**K NEAREST NEIGHBORS (KNN)**

**KNN** classifiers (Dasarathy, 1991) approximate the probability distribution over  $Y$  at the point  $x$  of the input space by the relative frequency of the class in the neighborhood of  $x$ . Thus, KNN relies on the assumption of locally constant class probabilities. In the case of a  $0/1$  loss function, classification simply amounts to predicting the most prevalent among the classes of the  $k$  nearest neighbors of  $x$  (majority voting). This basic estimation principle has been extended in various directions. A straightforward and intuitively reasonable idea, for example, is to weight the neighbors according to their closeness to  $x$ .

Needless to say, the choice of the neighborhood size  $k$  has a strong influence on the performance of KNN classification, and there are different methods for optimizing this parameter. Besides, a suitable distance measure must be defined on the input space  $X$ . If  $X$  corresponds to the  $m$ -dimensional Euclidean space (or a subset thereof), as in our application, one typically employs the Euclidean distance. As a local estimation method, KNN is known to have problems in high-dimensional input spaces. Thus, feature selection and feature weighting is of critical importance for KNN methods (Wettschereck *et al.*, 1997). We applied KNN with  $k=5$ , which showed best results in test scenarios. The KNN algorithm was implemented in Python into the EPIC library.

**NAÏVE BAYES (NB)**

Using the well-known Bayes' rule, the probability of observing class  $y$ , given an input vector  $x=(a_1, \dots, a_m)$ , can be expressed as

$$P(y|a_1, \dots, a_m) = P(a_1, \dots, a_m|y) \times P(y) / P(a_1, \dots, a_m) \quad (10)$$

In principle, (10) could be used to estimate the probability for every class  $y$  and, hence, to derive the most probable prediction (note that the denominator in (10) is a constant factor that does not depend on  $y$  and can hence be ignored.) However, since the conditional probabilities  $P(a_1, \dots, a_m | y)$  as well as the (prior) probabilities  $P(y)$  are usually unknown, they must be estimated from the data given. This is problematic, since the number of these probabilities can be huge (note that conditional probabilities must be estimated for every feature combination, i.e., for each potential input pattern). The Naïve Bayes classifier makes the simplifying assumption that the attributes are conditionally independent given the class. Thus,  $P(a_1, \dots, a_m | y)$  simplifies to  $P(a_1 | y) \times \dots \times P(a_m | y)$ , thereby reducing the number of probabilities to be estimated dramatically.

Needless to say, the Naïve Bayes assumption of conditional independence will usually not be satisfied in practice and at best provides a good approximation. In particular, this assumption is obviously incorrect in the applications considered in this work. (For example, the number of atom-atom contacts for the pairs  $(a,b)$  and  $(b,c)$  is definitely not independent.) Nevertheless, the Naïve Bayes classifier has proven to perform rather well over a wide range of practical classification problems. This can partly be explained by the fact that classification is quite robust toward variations of the class probabilities and, hence, toward imprecise probability estimations. In fact, note that the classification remains correct as long as the true class receives the highest probability, even if the estimated probabilities themselves are not very accurate. If the attribute  $A_i$  has a very large range and not many examples are given, some of the  $P(a_i | y)$  in  $P(a_1 | y) \times \dots \times P(a_m | y)$  can be nearly by 0 or more worse they are 0. To avoid this the so called  $m$ -estimate and discretisation can be used (Mitchell, 1997). A not very sophisticated discretisation showed not very much impact to the classification and therefore no discretisation was used. Due to very poor results without  $m$ -estimate the Naïve Base classifier was only used with  $m$ -estimate. The Naïve Bayes algorithm was implemented in Python into the EPIC library.

## FEATURE SELECTION APPROACHES

Similar to the human brain a machine learner can increase its learning performance by filtering out irrelevant features. As a welcome side-effect, the classification speed increases by focusing on less features. Feature selection can be viewed as a combinatorial problem: Given a set of  $m$  features, a subset of  $k \ll m$  features has to be found that minimizes the risk  $R$ . Since there are  $2^m$  possible subsets, an exhaustive search in space of all feature subsets is not feasible in most cases. Several methods have been developed to overcome this problem. In literature there are two most commonly applied approaches described to solve this problem: Filter and Wrapper approaches (Blum, 1997). In the Filter approach, feature selection is performed as a preprocessing step to the actual learning algorithm. The preprocessing step estimates general characteristics of the training set to select the most promising features and to discard all other features. Wrappers utilize the machine learning algorithm itself to extract the discriminating features. In this work, we used both a RELIEF F Filter approach and a *backward elimination* Wrapper. Furthermore, we implemented a Genetic Algorithm (GA) in Python into EPIC to overcome the combinatorial problem of evaluating huge numbers of combinations and extracting the discriminating features from the input data.

### FILTER APPROACH

RELIEF F, a Filter approach, is an enhancement of the former RELIEF feature selection method. The basic idea behind RELIEF F is the assumption that the selected features have dissimilar values for samples with different labels and similar values for samples with the same labels. RELIEF F follows this concept by assigning a quality to every feature. In order to compute the RELIEF F quality of every feature,  $d$  examples from  $S$  with  $R \subseteq S, |R|=d \ll |S|$  are randomly chosen. The quality for feature  $A_i$  is then:

$$q(A_i) = \sum_{r \in R} \sum_{h(s) \neq h(r)} \left( \frac{Pr(h(s))}{1 - Pr(h(r))} \frac{d_{rs,i}}{k} \right) - \sum_{r \in R} \sum_{h(s) = h(r)} \frac{d_{rs,i}}{k} \quad (11)$$

where  $k$  is used to overcome the problem of background noise in the data distribution and can be selected arbitrarily (but is supposed to be not too small), where  $d_{rs,i}$  is the distance between vector  $r$  and  $s$  at the position  $i$ . Filter approaches try to find the relevant features in a preprocessing step without considering the machine learning algorithm itself. This may lead to a selection of a feature subset which subsequently enhances the classification accuracy of the learning algorithm. An advantage of Filter approaches is their speed.

### WRAPPER APPROACH

Wrapper approaches utilize the learning algorithm itself to choose the discriminating features. For every feature subset, the prediction accuracy is computed using the learning algorithm itself. Since it is not feasible to explore the full space of all  $2^m$  feature subsets, several methods have been developed for searching this space in a heuristic (greedy) manner, thereby gaining efficiency at the cost of (possibly) losing optimality. There are mainly two kinds of Wrapper approaches: The so called *forward selection* and the *backward elimination*. The forward selection starts with a feature set of only one single attribute and successively adds further attributes which improve the classification accuracy. The *backward elimination* starts with the full set of all features and successively removes attributes which reduce the classification accuracy. We implemented the *backward elimination* in EPIC and started with the full set of features. In every step, the feature which decreased the prediction accuracy most was discarded in the next step. In avoidance of local minima the algorithm was applied until only one attribute was left and the set of attributes with the best classification accuracy was used. The classification accuracy was estimated by a *leave-one-out* cross-validation.

## GENETIC ALGORITHMS

Genetic Algorithms (GA) belong to a family of search algorithms which are inspired by the concept of natural evolution. These algorithms encode a potential solution for a specific problem in a chromosome-like structure and apply recombination (*crossover*) and *mutation* operators to these structures. Recombination can be understood as copying parts of the chromosomes and *mutation* can be regarded as flipping „bits“ on the chromosomes. From a computational point of view, the chromosomes are bitstrings representing *0s* and *1s* only.

An essential idea of any GA is the application of the “*survival of the fittest*“ principle. I.e., the generated individuals compete with each other, and those *individuals* scoring best with respect to the given target problem are given a higher chance to *reproduce*. Better solutions are measured according to a function  $\phi$  which will be referred to as *fitness* function. Thus, using GAs implies that a *fitness* function appropriate for the problem can be defined. Usually GAs are utilized to find the optimum of the function  $\phi$ . In this study,  $\phi$  was considered as prediction accuracy of *leave-one-out* cross validation. Traditional techniques such as gradient descent are appropriate for local optimization problems. However, this kind of optimization technique is not applicable if the function has many local optima or, even worse, if the gradient cannot be computed. Major advantage of GAs is in fact that little knowledge is required about the function to be optimized and their applicability to large scale problems. Both holds for the present problem of extracting features from our input vectors: we only have little knowledge about the underlying function of the problem and the search space is very large ( $2^{171}$  in case of  $ACV^{ACE}$  vectors).

GAs use a random walk through a highly exploitative search space encoded by discrete entities, i.e. the transition from one state to another in the search space is probabilistic and not deterministic. We implemented a so-called Simple Genetic Algorithm in EPIC,

written in Python. After preliminary tests, a population size of 30, a *crossover* rate of 75%, and a *mutation* rate of 5% was chosen for all studies. The offspring is created via two-point *crossover* between the mates and *mutation* was performed by single bit substitutions (point *mutations*). As termination criterion we defined convergence with no further changes over 10 generations or a 100% prediction quality.

## FEATURE ANALYSIS

The GA optimized feature selection, which is receiving the best results, was chosen for the analysis of the extracted features. It is not relevant to consider the chromosome of the fittest individual, since this would only correspond to a snap-shot during the optimization process. To obtain clear trends, we analyzed around  $2 \cdot 10^5$  chromosomes evaluated during the GA process. The more often a feature was considered (i.e. appeared in the chromosome) of a given *fitness*, the higher it was ranked as discriminating feature by calculating the frequency of occurrence in percentage with respect to this feature. Redundant descriptors, such as the metal contact score in SDV (always showing the value of 0), are supposed to result in a 50% probability since metal contacts are irrelevant for the classification. By plotting the calculated frequency of occurrence for each feature against the correspondingly obtained prediction accuracy, an estimate for its relevance is obtained represented by the gradient (or slope)  $m$  of the line of best fit. I.e., features becoming more relevant during the GA process show  $m > 0$ , irrelevant features  $m = 0$ , and features reducing the prediction accuracy  $m < 0$ . Thus,  $m$  is a measure for the significance of every feature contributing to the classification process; in the following, referred to as EPICscore.

The analysis of GA generated feature subsets was only performed for the combination of the C4.5 machine learner together with the SDV data representation, since C4.5 is according to our results the algorithm with the best discriminating power and the SDVs are the descriptors most amenable to a physicochemical interpretation.



## RESULTS AND DISCUSSION

We investigated both dataset A and B and classified them with four different machine learning algorithms: Support Vector Machines (SVM), Decision Trees C4.5, K Nearest Neighbors (KNN) and Naïve Bayes (NB). We utilized three different feature selection methods: Filter, Wrapper, and Genetic Algorithm.

### CLASSIFICATION OF MONOMER VERSUS HOMODIMER COMPLEXES

We used the dataset of Ponstingl *et al.*, who classified this dataset with a prediction accuracy of up to 88.9%. They used a *bootstrap re-sampling* validation using statistical potentials, based on atom-atom contact frequencies across the interface (Mintseris & Weng, 2003). We performed *leave-one-out* cross-validations for all combinations of the data representation and applied machine learning algorithms with and without feature selection algorithms. Also, we run *k-fold* cross-validations to assess the robustness of the predictions. The cross-validation results using this dataset (Tab. 3) show that all machine learning algorithms are capable to discriminate crystal packing enforced monomer contacts from functional homodimer contacts. Without any feature selection, the C4.5 algorithm shows a prediction accuracy between 89.0% for ACV<sup>Sybyl</sup> and 87.8% for DPV in the *leave-one-out* cross-validation (Tab. 3a) and between 90.9% ( $\pm 5.9\%$ ) for SDV and 84.1% ( $\pm 3.4\%$ ) for DPV in the *k-fold* cross-validation, which was performed with  $k=2$  and repeated 20 times (Tab. 3b). The KNN algorithm holds the best prediction accuracy for a *leave-one-out* cross-validation without any feature selection. In combination with ACV<sup>Sybyl</sup> input vectors, 91.3% correctly predicted complexes are revealed. This is also true for the *k-fold* cross-validation with 89.5% ( $\pm 0.8\%$ ) for the same combination. But also the ACV<sup>ACEs</sup> and the DPVs are scoring well, with 87.2% and 89.5% in case of the *leave-one-out* respectively, and 86.2% ( $\pm 2.2\%$ ) and 85.6% ( $\pm 2.5\%$ ) in case of the *k-fold* cross-validation. Only the SDV data representation is dropping back with 84.9% and 76.2% ( $\pm 5.0\%$ ), respectively. The SVM algorithm leads

to prediction accuracy between 88.4% and 86.0% for the *leave-one-out* and 86.2% ( $\pm 1.8\%$ ) to 84.3% ( $\pm 2.0\%$ ) for the *k-fold* cross-validation. The NB algorithm shows only 82.6% to 84.3% for the *leave-one-out* and 80.0% ( $\pm 3.7\%$ ) to 61.1% ( $\pm 10.2\%$ ) for the *k-fold* cross validation. This is by no means a satisfying result, since Ponstingl *et al.* reported already 84% prediction accuracy simply based on the size of the protein-protein interface. The poorly performing NB algorithm is probably due to the fact that the assumption of conditional independence is not satisfied in our study, since, for instance, the number of atomic contacts is strongly interdependent.

Utilizing the feature selection algorithms in combination with *leave-one-out* cross-validations, we were able to extract the most discriminating features by enhancing the prediction rate simultaneously. RELIEF F is the computationally least expensive algorithm. By discarding those features from the vector having the RELIEF F quality  $q \leq 0$  the vectors were reduced from 171 to 108 (ACV<sup>ACE</sup>), 78 to 63 (ACV<sup>Sybyl</sup>), 78 to 56 (DPV) and 63 to 54 (SDV) attributes. The results show that the gain in prediction accuracy is too small to extract meaningful information out of the selected features (Tab. 3c). The average enhancement in prediction accuracy is at most around 2%, in some cases we even observe a small loss in prediction accuracy. RELIEF F seems not to be a reliable feature selection algorithm for the given problem.

The *backward elimination* Wrapper approach is computationally quite expensive, since we have to run at most  $\binom{n}{2} + n$  cross-validations, where  $n$  is the size of the given attributes, i.e. 14706 for ACV<sup>ACE</sup>, 3003 for ACV<sup>Sybyl</sup> and DPV, and 1953 for SDV. The Wrapper performs better than the RELIEF F approach by enhancing the prediction rate on average by about 3%. However, the Wrapper benefits in comparison to the Filter approach since it is optimally tailored towards the applied machine learner. The backwards elimination pruned the input vectors dramatically, e.g. the ACV<sup>ACE</sup> attributes from 171 to 47. The highest prediction accuracy with the Wrapper was achieved with the combination C4.5/SDV resulting in 93.0%. NB still performs worst of the given algorithms with only

Method	ML	ACV <sup>ACE</sup>		ACV <sup>Sybyl</sup>		DPV		SDV	
Loo cv	C4.5	88.4%		89.0%		87.8%		88.4%	
	SVM	87.2%		88.4%		87.2%		86.0%	
	NB	84.3%		83.1%		82.6%		82.6%	
	KNN	87.2%		91.3%		89.5%		84.9%	
k-fold cv k=2 n=20	C4.5	87.8%	± 2.2%	84.7%	± 4.4%	84.1%	± 3.4%	90.9%	± 5.9%
	SVM	85.1%	± 1.2%	84.3%	± 2.0%	86.2%	± 1.8%	84.9%	± 2.1%
	NB	79.1%	± 4.4%	80.0%	± 3.7%	78.2%	± 3.9%	61.1%	± 10.2%
	KNN	86.2%	± 2.2%	89.5%	± 0.8%	85.6%	± 2.5%	76.2%	± 5.0%
Filter	C4.5	88.4%	108	89.0%	63	89.5%	56	89.5%	54
	SVM	87.2%	108	88.4%	63	86.6%	56	86.0%	54
	NB	84.9%	108	83.7%	63	82.6%	56	80.6%	54
	KNN	88.5%	108	91.9%	63	89.5%	56	83.1%	54
Wrapper	C4.5	91.3%	47	89.5%	41	91.9%	51	93.0%	36
	SVM	88.4%	74	88.4%	56	89.0%	44	87.2%	41
	NB	85.5%	102	84.0%	63	84.9%	57	86.0%	49
	KNN	90.1%	52	91.9%	44	89.5%	42	89.5%	43
GA	C4.5	94.8%	26	91.3%	32	93.0%	32	94.2%	24
	SVM	91.3%	34	90.1%	40	91.9%	32	91.9%	31
	NB	85.6%	30	85.6%	31	86.6%	29	83.1%	34
	KNN	90.7%	31	94.2%	24	92.4%	28	91.9%	29

**TABLE 3.** Results for the classification of dataset A. (a) *Leave-one-out* cross-validation (*Loo*), (b) *k-fold* cross-validation (*k-fold cv*) with  $k=2$  on average over  $n=20$  iterations, (c) *Leave-one-out* cross-validation with Filter RELIEF F, (d) *Leave-one-out* cross-validation with Wrapper, (e) *Leave-one-out* cross-validation with Genetic Algorithm.

Method	ML	ACV <sup>ACE</sup>		ACV <sup>Sybyl</sup>		DPV		SDV	
Loo cv	C4.5	<b>78.3%</b>		<b>77.7%</b>		<b>78.6%</b>		<b>77.7%</b>	
	SVM	<b>82.1%</b>		<b>81.8%</b>		<b>81.8%</b>		<b>80.3%</b>	
	NB	<b>80.0%</b>		<b>73.2%</b>		<b>70.5%</b>		<b>75.4%</b>	
	KNN	<b>81.7%</b>		<b>79.4%</b>		<b>80.0%</b>		<b>77.4%</b>	
k-fold cv k=2 n=20	C4.5	<b>76.3%</b>	± 2.2%	<b>75.4%</b>	± 3.1%	<b>76.0%</b>	± 3.1%	<b>75.4%</b>	± 2.9%
	SVM	<b>79.2%</b>	± 1.9%	<b>78.3%</b>	± 1.1%	<b>79.2%</b>	± 2.4%	<b>78.3%</b>	± 1.6%
	NB	<b>77.7%</b>	± 3.3%	<b>74.0%</b>	± 2.3%	<b>70.7%</b>	± 4.8%	<b>73.2%</b>	± 5.3%
	KNN	<b>79.5%</b>	± 2.1%	<b>78.6%</b>	± 1.9%	<b>78.0%</b>	± 2.5%	<b>76.2%</b>	± 3.9%
Filter	C4.5	<b>79.2%</b>	106	<b>81.2%</b>	58	<b>78.6%</b>	48	<b>80.0%</b>	52
	SVM	<b>82.1%</b>	106	<b>80.3%</b>	58	<b>80.0%</b>	48	<b>79.2%</b>	52
	NB	<b>79.2%</b>	106	<b>76.2%</b>	58	<b>68.5%</b>	48	<b>67.0%</b>	52
	KNN	<b>80.3%</b>	106	<b>80.0%</b>	58	<b>80.0%</b>	48	<b>77.4%</b>	52
Wrapper	C4.5	<b>91.0%</b>	32	<b>87.0%</b>	52	<b>87.0%</b>	51	<b>86.1%</b>	38
	SVM	<b>87.0%</b>	67	<b>88.4%</b>	61	<b>89.0%</b>	46	<b>88.4%</b>	35
	NB	<b>85.4%</b>	77	<b>80.6%</b>	56	<b>81.2%</b>	52	<b>83.2%</b>	41
	KNN	<b>82.7%</b>	68	<b>85.2%</b>	53	<b>84.3%</b>	51	<b>84.6%</b>	39
GA	C4.5	<b>93.6%</b>	62	<b>91.0%</b>	37	<b>91.3%</b>	41	<b>91.0%</b>	34
	SVM	<b>91.3%</b>	56	<b>89.3%</b>	40	<b>90.1%</b>	32	<b>89.6%</b>	36
	NB	<b>87.8%</b>	76	<b>85.5%</b>	44	<b>84.9%</b>	51	<b>86.7%</b>	38
	KNN	<b>84.6%</b>	68	<b>85.2%</b>	48	<b>85.8%</b>	51	<b>86.1%</b>	39

**TABLE 4.** Results for the classification of dataset B. (a) *Leave-one-out* cross-validation (*Loo*), (b) *k-fold* cross-validation (*k-fold cv*) with  $k=2$  on average over  $n=20$  iterations, (c) *Leave-one-out* cross-validation with Filter RELIEF F, (d) *Leave-one-out* cross-validation with Wrapper, (e) *Leave-one-out* cross-validation with Genetic Algorithm.

82.0% (SDV) to 85.5% (DPV) correctly predicted complexes. Both KNN and SVM predict around 90% of the considered complexes correctly, rather independently of the data representation used (Tab. 3d).

Finally, the Genetic Algorithm approach was utilized to extract discriminating features from the dataset. GAs are computationally very expensive, thus we performed all runs on a cluster with 25 CPUs (Intel® Pentium® 4, 2.8 GHz, 1GB RAM) in parallel to reduce the elapse time. With this approach we observed the highest prediction quality, using the C4.5 machine learning algorithm, in combination with the ACV<sup>ACE</sup>. Here we achieved remarkable 94.8% correctly assigned complexes, using only 26 of the original 171 attributes. C4.5 also performs better than 90%, using any other data representations: 94.2% (SDV, 25 attributes), 93.0% (DPV, 51 attributes) and 91.3% (ACV<sup>Sybyl</sup>, 41 attributes). KNN and SVM also predict in the range of 90% to 94% using the different data representations, simultaneously reducing the input features remarkably. Again the NB algorithm cannot convince, revealing at most 86.6% prediction accuracy in case of the DPVs (Tab. 3e).

To show the robustness of the selected features, we performed again a *k-fold* cross-validation with  $k=2$  and  $n=100$  for the selected features of the GA. Since the performance of machine learning algorithms decreases with a smaller data basis, we expected slightly worse results for the *k-fold* cross-validation. Indeed the SDV drops back to 92.7% ( $\pm 1.8\%$ ), ACV<sup>ACE</sup> 92.3% ( $\pm 2.1\%$ ), DPV 90.4% ( $\pm 1.4\%$ ) and ACV<sup>Sybyl</sup> 89.4% ( $\pm 2.8\%$ ). Nevertheless, the extracted features are robust against over-fitting.

## CLASSIFICATION OF FOLDING COMPLEXES VERSUS RECOGNITION COMPLEXES

Mintseris *et al.* reported a 76% prediction rate simply focusing on the sizes of the interfaces. This underlines the importance of the size of the contact surface as factor for distinguishing permanent from transient complexes. Since we do not normalize our input

data, this factor is considered implicitly in the input vectors. In comparison to KDA used by Mintseris *et al.* (91%) with  $ACV^{ACE}$  input vectors, the methods applied here cannot achieve the same quality in prediction accuracy. However, in combination with the feature selection methods Wrapper and GA, we were able to classify with the same or even higher accuracy. Possibly Mintseris *et al.* also used some kind of feature selection methods.

The cross-validation for this dataset B does not show as robust results as obtained for dataset A (Tab. 4). The best *leave-one-out* cross-validation results were achieved with the SVM algorithm with 82.1% correctly predicted complexes in case of the  $ACV^{ACE}$  input vectors. The other data representations fall into the same range showing 81.8% ( $ACV^{Sybyl}$  and DPV) and 80.3% for the SDV. KNN performs well in combination with the  $ACV^{ACE}$  input vector with 81.7% prediction rate, 80.0% with DPV, 79.4% with  $ACV^{Sybyl}$ , but only 77.4% with the SDVs. The well performing C4.5 with respect to dataset A, does not accomplish the expectation on this dataset: Around 78% were achieved, independently of the given input vector. The NB algorithm again performs worst across the considered machine learners: From 80.0% ( $ACV^{ACE}$ ) to only 70.5% (DPV) (Tab. 4a). Again we performed *k-fold* cross-validations with  $k=2$  and  $n=20$  with no significant loss in accuracy (Tab. 4b).

We applied the RELIEF F feature selection algorithm with  $q>0$  to remove irrelevant attributes from the input vectors. RELIEF F reduces the  $ACV^{ACE}$  to 106,  $ACV^{Sybyl}$  to 58, DPV to 48 and the SDV to 52 attributes, but it does increase prediction accuracy only in case of the C4.5 algorithm by about 2% on average. RELIEF F stagnates or even decreases in prediction accuracy with all other used algorithms (Tab. 4c).

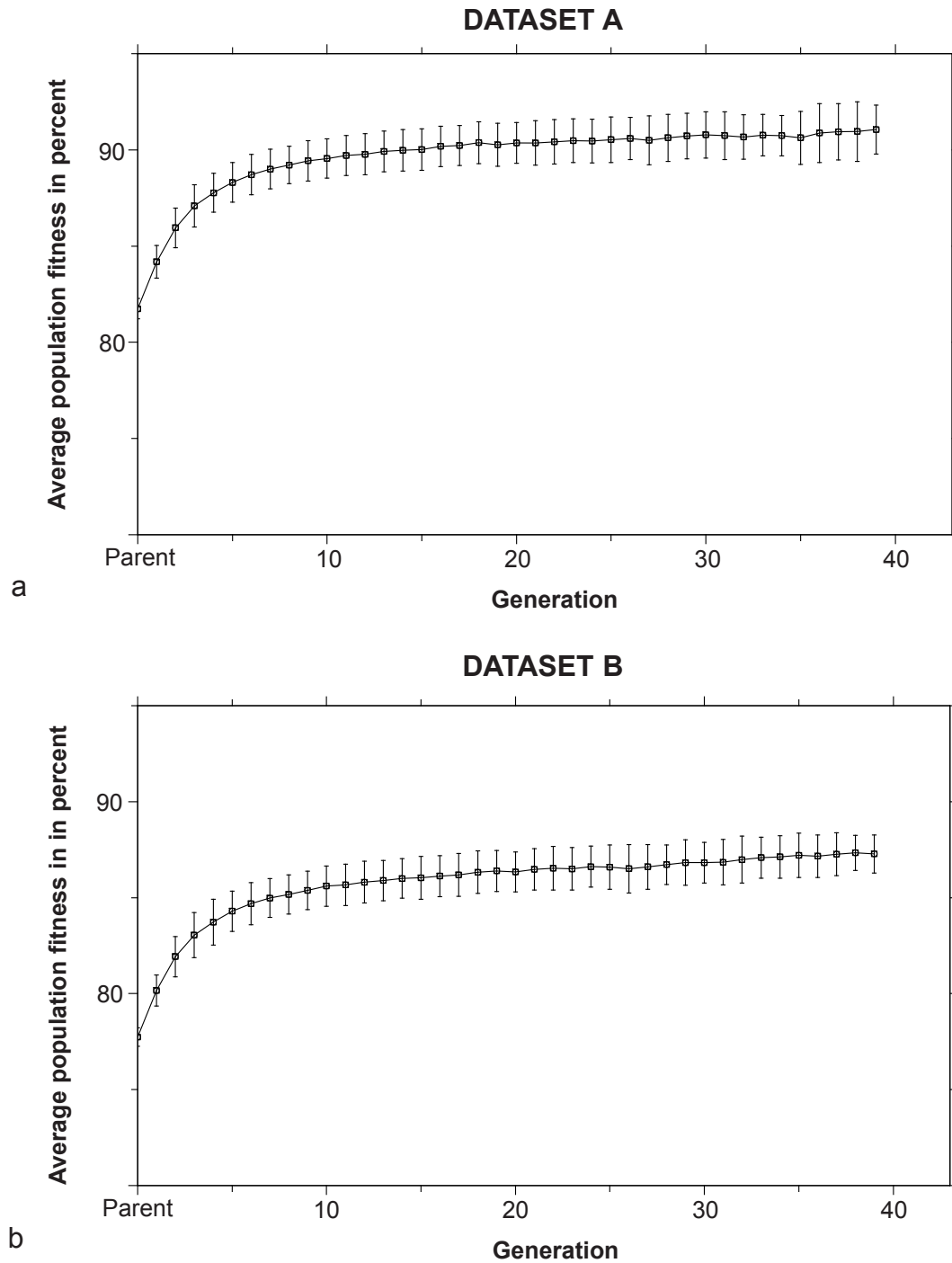
In contrast, the Wrapper approach leads to an enormous increase in prediction accuracy: by pruning the  $ACV^{ACE}$  data vectors from 171 attributes to only 32, we achieve 91.0% correctly predicted complexes with C4.5. Also SVM (87.0%), NB (85.4%) and KNN (82.7%) show better results using  $ACV^{ACE}$ . The other data representations also increase in prediction accuracy after removing non-relevant attributes from the vectors (Tab. 4d).

Finally, we used the GA approach on this dataset. After pruning the input vectors the best algorithm together with a GA is the C4.5 algorithm, although this machine learner did not show exceptionally convincing results without the pruning. Independently of the selected data representation, we achieve prediction accuracies beyond 90%, cumulating at 93.6% for  $ACV^{ACE}$ . Also the SVM results in around 90% correctly predicted complexes. KNN and NB do not exceed the 90% barrier, but enhance the prediction accuracy in all cases (Tab. 4e).

A *k-fold* cross-validation with the extracted attributes,  $k=2$  and  $n=100$ , achieved a prediction accuracy of 91.5% ( $\pm 1.6\%$ ) for  $ACV^{ACE}/C4.5$  and 90.1% ( $\pm 0.9\%$ ) for  $SDV/C4.5$ . This shows that the selected features are robust with respect to possible over-fitting.

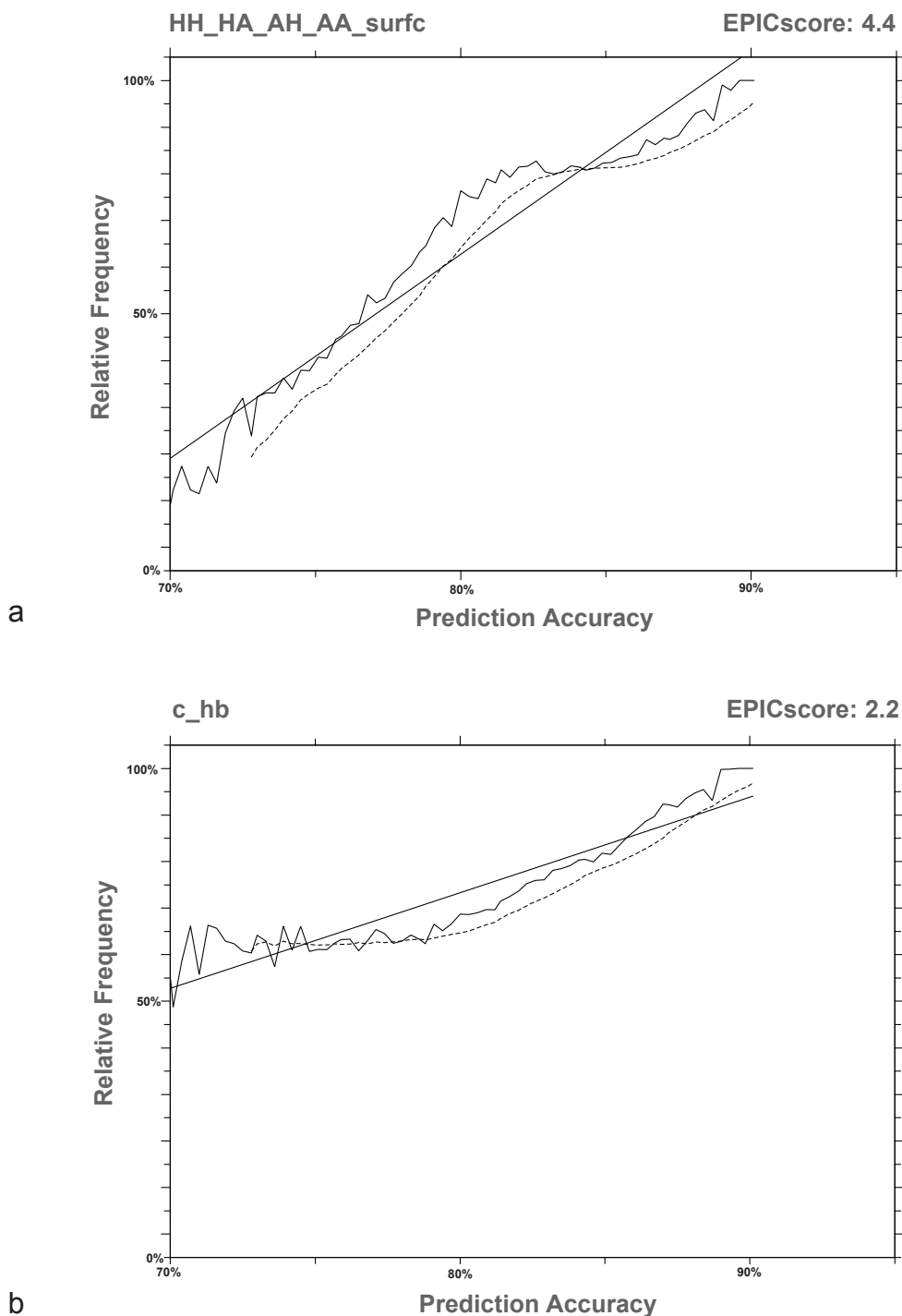
## FEATURE ANALYSIS WITH GENETIC ALGORITHMS

First, we analyzed the GA selected features of classification for dataset A. We dissected around  $1.7 \cdot 10^5$  chromosomes from 326 GA runs (Fig. 2a) and plotted the tendencies of relevance for selected descriptors (of example in Fig. 3 and 4). As expected, irrelevant features such as the metal contact descriptor act as control value and show a gradient approximately equal to 0. The most discriminating feature of the classification procedure are the contacts of hydrophobic and/or aromatic atoms located in the protein-protein interfaces (HH\_HA\_AH\_AA\_surfc). This descriptor is a prime example for demonstrating a decreasing prediction accuracy by removing and

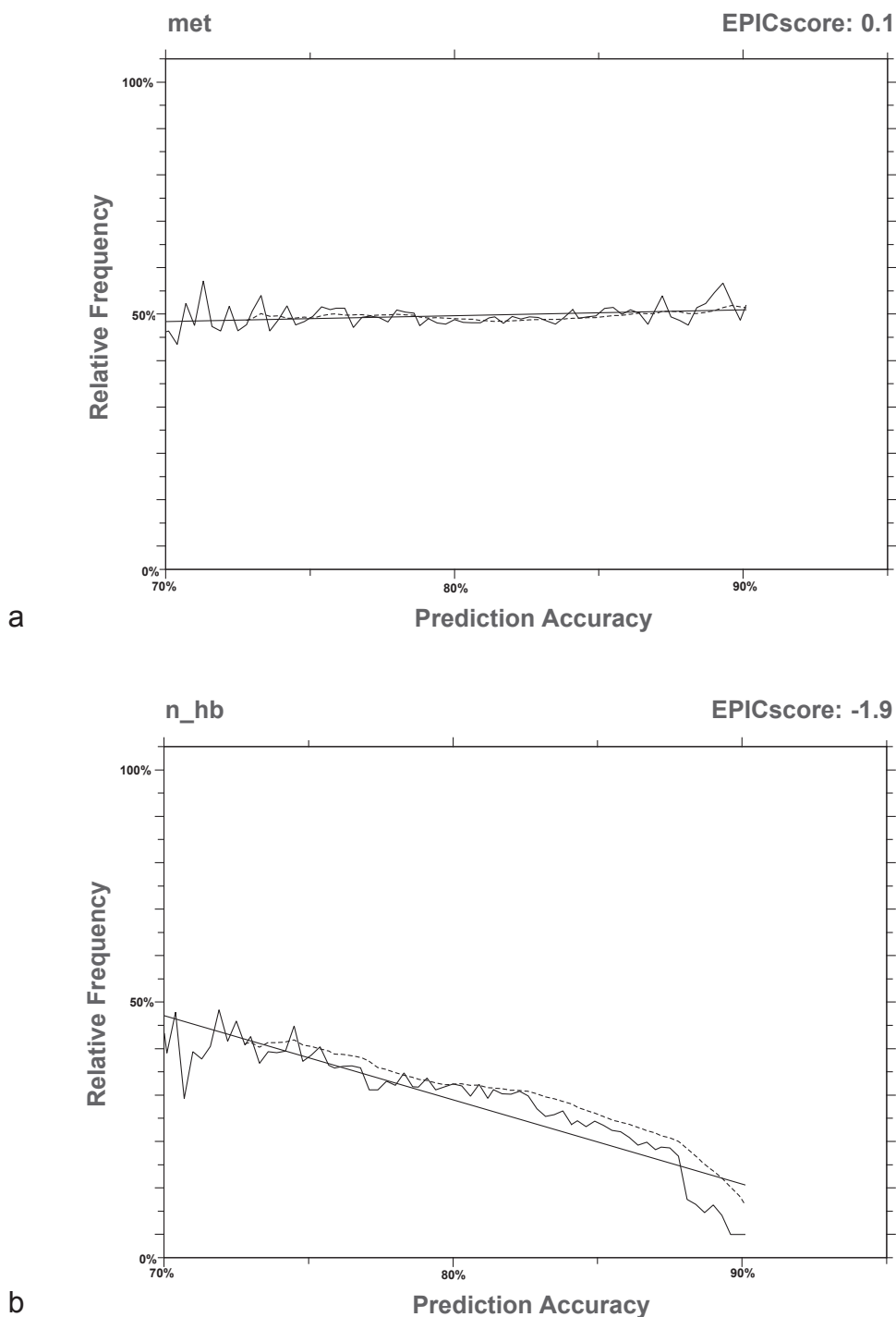


**FIGURE 2.** (a) Genetic Algorithm feature analysis for SDV data representation with C4.5 machine learner for the classification of dataset A. Each data point represents the population *fitness* averaged over 326 GA runs in dependency to the generation. (b) Genetic Algorithm feature analysis for SDV data representation with C4.5 machine learner for the classification of dataset B. Each data point represents the average population *fitness* averaged over 350 GA runs in dependency to the generation.





**FIGURE 3.** Analyzed individuals from GA feature analysis plotted by prediction accuracy (x-axis) against relative frequency of occurrence (y-axis) for the SDV descriptors of dataset B. **(a)** HH\_HA\_AH\_AA\_surfc represents contacts of hydrophobic and/or aromatic atoms in **(b)** c\_hb represents charged hydrogen bonds across the protein-protein interface. The straight line (slope of the curves) represents the importance of the selected feature (EPICscore). The dashed line represents the moving average over 10 values.



**FIGURE 4.** Analyzed individuals from GA feature analysis plotted by prediction accuracy (x-axis) against relative frequency of occurrence (y-axis) for the SDV descriptors of dataset B. **(a)** met serves as reference descriptor (see text) **(b)** n\_hb represents non-charged hydrogen bonds across the protein-protein interface. The straight line (slope of the curves) represents the importance of the selected feature (EPICscore). The dashed line represents the moving average over 10 values.

an increasing prediction accuracy by retaining it, consequently achieving the highest EPICscore with 4.2 (Tab. 5). Furthermore, the pure hydrophobic/hydrophobic atom contacts (SURFC\_HYDROPHOBIC\_HYDROPHOBIC) and the polar/hydrophobic atom contacts (SURFC\_POLAR\_HYDROPHOBIC) in the protein-protein interfaces are potent discriminating features with an EPICscore of 2.9 and 2.5, respectively. The score for charged hydrogen bonds (c\_hb) across the protein-protein interfaces also shows a clear tendency to be an important descriptor (EPICscore: 2.9). Also the ring-ring interactions show discriminating power with an EPICscore of 1.6. In contrast to the latter ones, descriptors such as the total buried surface (TotBurSurf, EPICscore: -2.7), the hydrogen bond score (HBScore, EPICscore: -2.2), or the score of non-charged hydrogen

DATASET A		DATASET B	
4.2	HH_HA_AH_AA_surfc	4.4	HH_HA_AH_AA_surfc
2.9	c_hb	2.6	SURFC_HYDROPHOBIC_HYDROPHOBIC
2.9	SURFC_HYDROPHOBIC_HYDROPHOBIC	2.5	SURFC_POLAR_HYDROPHOBIC
2.5	SURFC_POLAR_HYDROPHOBIC	2.2	c_hb
1.6	RRScore	0.9	AroBurSurf
1.1	NHBonds	0.7	NHBonds
0.6	AroBurSurf	0.3	RRScore
0.0	SURFC_AROMATIC_AROMATIC	0.1	met
-0.1	BURCP	0.0	SURFC_POLAR_POLAR
-0.1	met	-0.2	PolBurSurf
-0.3	SURFC_POLAR_POLAR	-0.4	PH_HP_PA_AP_surfc
-0.3	PolBurSurf	-0.4	PH_HP_surfc
-0.7	PH_HP_surfc	-0.7	SURFC_AROMATIC_AROMATIC
-1.2	HydBurSurf	-0.7	HydBurSurf
-1.2	AHPDI	-1.6	AHPDI
-1.7	PH_HP_PA_AP_surfc	-1.7	BURCP
-2.0	n_hb	-1.9	n_hb
-2.2	HBScore	-2.1	HBScore
-2.7	TotBurSurf	-2.8	TotBurSurf

**TABLE 5.** EPICscores of selected SFCscore features sorted by their relevance, determined by GA feature analysis for crystal contact versus homodimer classification (dataset A) and permanent versus transient classification (dataset B).

bonds (n\_hb, EPICscore: -2.0) are detrimental for the prediction accuracy. The curves clearly show an increasing prediction accuracy upon removing these latter descriptors from the classification process.

Subsequently, we investigated  $2 \cdot 10^6$  chromosomes from 350 GA runs from the classification of dataset B (Fig. 2b). We find results similar to the analysis of dataset A; however, the individual values are varying slightly and, thus, their rank order is some degree altered. Nevertheless, HH\_HA\_AH\_AA\_surfc (EPICscore: 4.4, Fig. 3a), SURFC\_HYDROPHOBIC\_HYDROPHOBIC (2.6), SURFC\_POLAR\_HYDROPHOBIC (2.5) and c\_hb (2.2) are still the descriptors with the most discriminating power, and alike n\_hb (-1.9), HBScore (-2.1) and TotBurSurf (-2.8) remain to be the descriptors perturbing the classification at most. Major differences to dataset A arise with respect to the consideration of the descriptors RRScore (EPICscore: 0.3 for dataset B, whereas 1.6 for dataset A) and BURCP (-1.7 for dataset B, whereas -0.1 for dataset A). RRScore represents the score of ring-ring interactions, whereas BURCP is the percentage of buried carbon atoms.

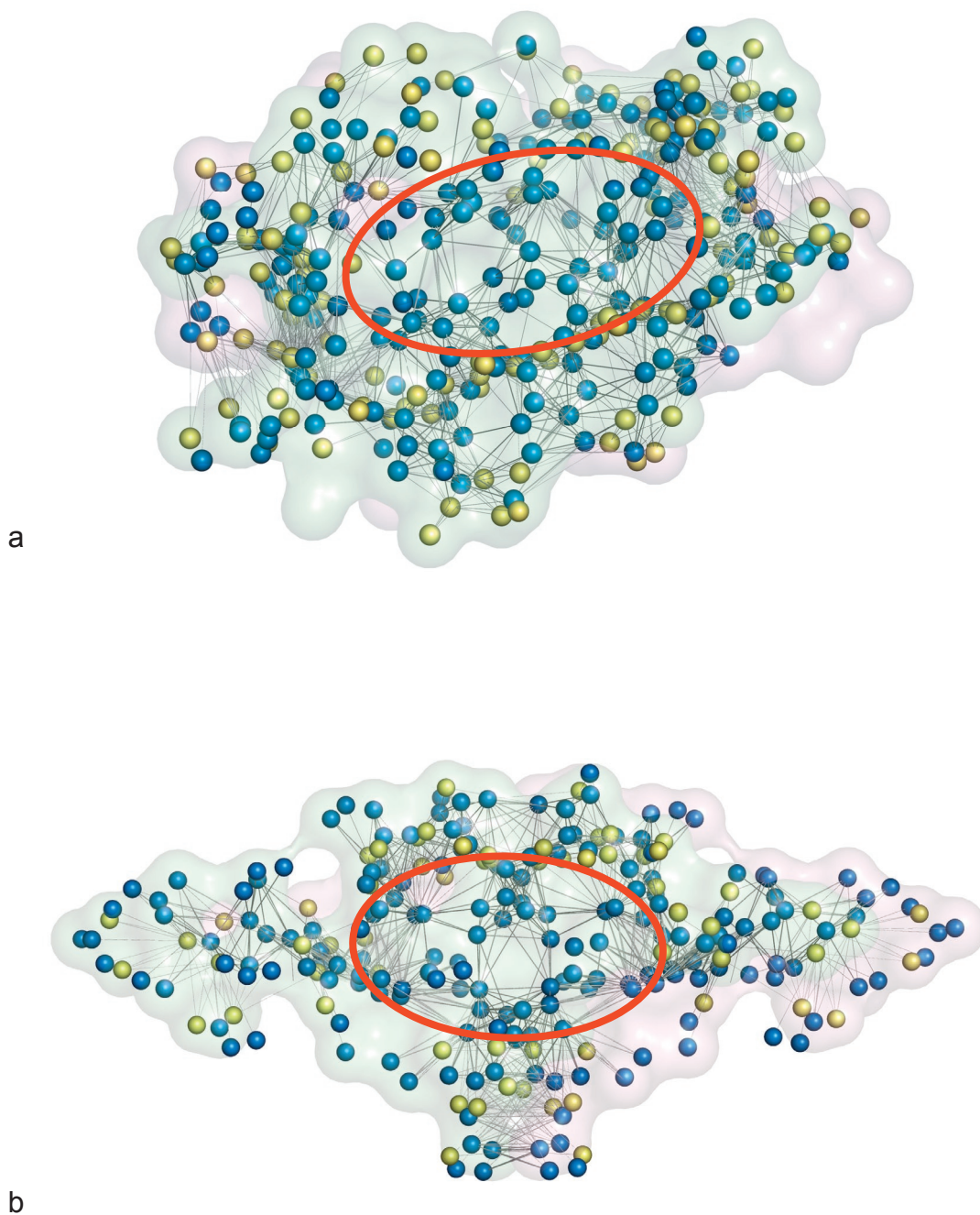
In both datasets, the contributions of hydrophobic, aromatic, polar surface patches and charged hydrogen bonds, respectively, are playing a crucial role for the discrimination between the given classes of complexes. As shown in previous studies (Larsen *et al.*, 1998), the contribution of hydrophobic (also aromatic) and polar surface patches show significant differences across the considered complex classes. Since permanent protein-protein complexes, for instance, almost never dissociate under physiological conditions, their amino acid composition across the interface differs from that of transient complexes, which are exposing their interfaces temporarily to the solvent. The same holds true for the analysis of complexes enforced by crystal contacts and functional homodimers, since in the soluted state the surfaces of crystal-contact complexes are solvent exposed. By visualizing the contribution of the extracted descriptors such as hydrophobic (and aromatic) contacts across the protein-protein interface, we are able to recognize the differences between the considered complex classes. Additionally, by means of the

implemented scoring function, we are able to quantify the differences (EPICscore). By visual inspection of the extracted features one can frequently observe that homodimer and permanent complexes often exhibit a hydrophobic core in the protein-protein interface surrounded by a rather hydrophilic corona (Fig. 5a and 5b), whereas transient complexes show evenly distributed contacts, as, according to their functional properties, these must expose their protein-protein interfaces temporarily to solvent (Fig. 6a and 6b). The present approach does not allow to quantify these geometric features and their occurrence frequency directly, but the extraction and the high weighting of the corresponding descriptors is a clear indicator that such physicochemical properties are important for the properties of the considered protein complexes.

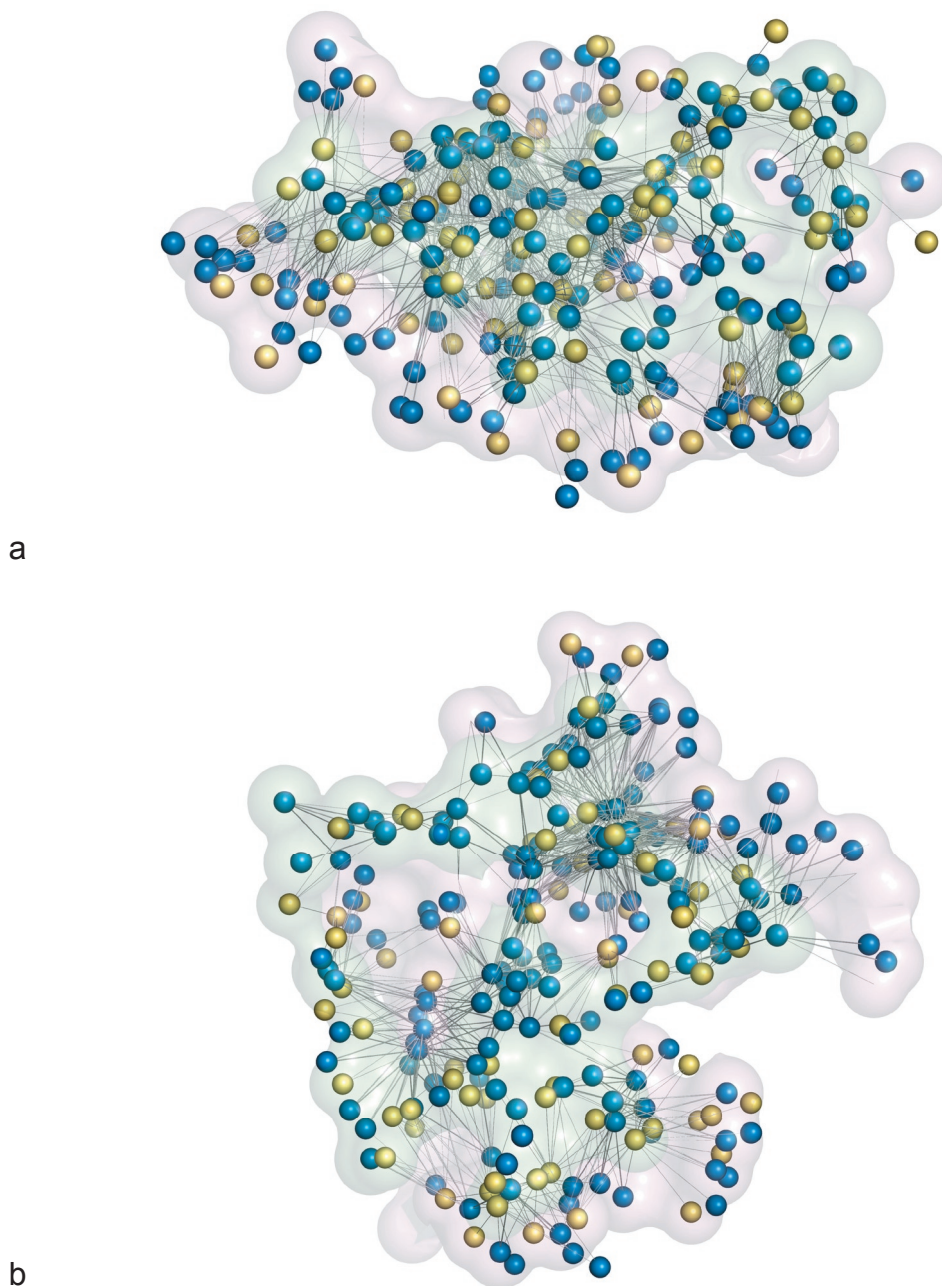
Also, the charged hydrogen bonds (*c\_hb*) are scored to be a major discriminant between both types of datasets. Hydrogen bonds were evaluated and ranked following the approach of Wang *et al.* (Wang *et al.*, 2002), based on the distance and angle ranges between the interacting atoms. Closer analysis of all 3644 charged hydrogen bonds present in the interfaces of dataset B showed two significantly distinctive features: (a) permanent complexes exhibit one charged hydrogen bond per 60 interface atoms, whereas in transient complexes one charged hydrogen bond is observed per only 55 interface atoms on average. (b) Charged hydrogen bonds in permanent complexes are on average scored higher (0.51) in comparison to their transient counterpart (0.46). Possibly this observation is due to the fact that permanent complexes are geometrically better optimized to form enhanced hydrogen bonds with more appropriate distances and angles. In contrast to the charged hydrogen bonds, the 3215 non-charged hydrogen bonds in dataset B show only one hydrogen bond per 65 atoms. The average score of these hydrogen bonds, admittedly, shows deviating values for permanent and transient complexes of 0.56 and 0.52, respectively. Nevertheless, the scoring of non-charged hydrogen bonds as implemented in SFCscore does not emerge as a relevant feature for discrimination. It is worthwhile

to mention that the average distance between the interacting atoms forming hydrogen bonds is almost exactly 3.0 Å (measured between the heavy atoms) for both permanent and transient complexes independent whether charges are involved or not.

The RRScore descriptor, considering various ring-ring interactions, is a discriminating feature in case of dataset A (EPICscore: 1.6). Rings were considered to be interacting if the center of the first ring falls below a distance threshold of 6.5 Å to the center of the second. The RRScore is calculated based on distance and the angle ranges between both interacting rings. We investigated all 223 ring-ring interactions in dataset A and 640 ring-ring interactions in dataset B, and indeed, the average RRScore in dataset A for crystal contacts ( $0.86 \pm 0.45$ ) differs from the score of the functional homodimers ( $0.78 \pm 0.31$ ), whereas in dataset B both the permanent and the transient complexes show an average score of  $0.78 \pm 0.31$ . Consequently, the EPICscore indicates the discriminative power of RRScore in dataset B to be low (0.3), and this descriptors is considered as more or less insignificant in feature selection. Undoubtedly, lower scores can be expected for the crystal-contact complexes, since they can be assumed as weakly interacting, which is supported by the fact that rather high B-values are found for the involved residues. Nevertheless, we believe that the highly scored ring-ring interactions found for these examples arise from the fact that only a low number of observations (only 48 in 96 complexes) is given with high standard deviations ( $\pm 0.45$ ). On the other hand, the 76 functional homodimer complexes show on average 2.3 ring-ring-interactions per complex, whereas the highest occurrence frequency for permanent complexes with an average of 2.6 per complex; transient complexes result in only 1.2 per complex. Thus, the RRScore descriptor is of discriminating power for dataset A, but not significant with respect to the sparsely populated data basis.



**FIGURE 5.** Representation of protein-protein interfaces, displaying hydrophobic (blue) and polar contacts (yellow). The direction of each contact is displayed by thin lines. **(a)** Permanent complex PDB 1icw, interleukin-8 mutant. Often observed hydrophobic cores (red) with surrounding hydrophilic rims in permanent complexes. **(b)** Permanent complex PDB 1jsg, 14tcl1, an oncogene product involved in t-cell prolic lymphocytic leukemia.



**FIGURE 6.** Representation of protein-protein interfaces, displaying hydrophobic (blue) and polar contacts (yellow). The direction of each contact is displayed by thin lines. **(a)** Transient complex PDB 1d6r, cancer chemopreventive Bowman-Birk inhibitor in ternary complex with bovine trypsin. **(b)** Transient complex PDB 1jdp, atrial natriuretic peptide clearance receptor in complex with C-type natriuretic peptide. Even distributed hydrophobic and hydrophilic contacts in transient complexes.



## SUMMARY AND CONCLUSIONS

Machine learning algorithms are generally appropriate to successfully classify protein-protein complexes. We used two different datasets: complexes being reinforced by the local environment (crystal contacts) versus physiologically functional homodimers, and permanent versus transient complexes. The predictive power and accuracy depends on the type of data representation, the applied machine learning algorithm, and the used feature selection algorithm. The best predictive power has been achieved using the Decision Tree algorithm C4.5, the weakest correlation resulted from the simple Naïve Bayes algorithm, independent of the chosen data representation. The feature selection method RELIEF F was not convincing, as it pruned features without enhancing the overall prediction rate significantly. In contrast, Wrapper and GA show superior performance by correctly predicting complexes without suffering from over-fitting as proven by repeated *k-fold* cross-validations. We were able to classify the Ponstingl dataset (A) based on 96 crystal contact and 76 functional homodimer complexes with an accuracy of 94.8% and 94.1% using a GA optimized C4.5 algorithm in combination with ACV<sup>ACE</sup> or SDV data representation, respectively. The classification of the Mintseris dataset (B), comprising 147 permanent and 198 transient complexes, resulted in 93.6% correctly predicted complexes by using the GA optimized C4.5 machine learner in combination with ACV<sup>ACE</sup> vectors, which were also used by Mintseris. For the SDV data representation, the same analysis reveals 91.0% correct predictions.

Furthermore, the analysis of the GA feature selection method gave us valuable insights into the understanding of protein-protein interactions at the atomic level. By analyzing each single chromosome evaluated during the GA procedure, we calculated a figure-of-merit for each attribute (EPICscore). This was performed with the SDV data representation in combination with the C4.5 algorithm, since particularly SDVs are straight-forward to interpret in physicochemical terms. We are able to elucidate the different nature of hydrophobic and polar surface patches present in protein-protein

interfaces. Visual inspection of the features shows that homodimer and permanent complexes often show hydrophobic cores surrounded by a corona of polar residues. Also the number and geometry of charged and non-charged hydrogen bonds, formed across the interface, have discriminative power which is perhaps not obvious at first glance.

**W**e developed the integrated toolkit EPIC composed by several machine learning algorithms, feature selection methods, visualization tools linked to database and storage facilities. Thus, on the one hand we are able to distinguish very reliably between different types of protein-protein complexes, while on the other we can extract discriminating features thus improving the understanding of protein-protein interactions at the atomic level. However, with respect to the interpretation of the extracted features, one has to note that the classification of protein complexes based on multidimensional input vectors which provide an intercorrelated result. Thus, the significance of one single highlighted feature must always be discussed in its context and should not be over-interpreted. On the

**REFERENCES**

- Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. (2004) A Dissection Of Specific And Non-Specific Protein-Protein Interfaces *J Mol Biol*, **336**, 943-955.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. (2000) The Protein Data Bank *Nucleic Acids Res*, **28**, 235-242.
- Blum, L. P. (1997) Selection Of Relevant Features And Examples In Machine Learning *Artificial Intelligence*, **97**, 245-271.
- Bock, J. R.; Gough, D. A. (2001) Predicting Protein-Protein Interactions From Primary Structure *Bioinformatics*, **17**, 455-460.
- Bradford, J. R.; Westhead, D. R. (2005) Improved Prediction Of Protein-Protein Binding Sites Using A Support Vector Machines Approach *Bioinformatics*, **21**, 1487-1494.
- Briem, H.; Guenther, J. (2005) Classifying „Kinase Inhibitor-Likeness“ By Using Machine-Learning Methods *Chembiochem*, **6**, 558-566.
- Chakrabarti, P.; Janin, J. (2002) Dissecting Protein-Protein Recognition Sites *Proteins*, **47**, 334-343.
- Chang, C.C.; Lin, C.J. (2001) A Library For Support Vector Machines.
- Clark, M; Cramer, R. D.; Van Opdenbosch, N. (1989) Validation Of The General Purpose Tripos 5.2 Force Field *J Comp Chem*, **10**, 982-1012.
- Cortez, C., Vapnik, V. (1995) Support Vector Networks Machine Learning, **20**, 273-279.

- Dasarathy, B. V. (ed.) (1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, California.
- Dill, K. A. (1990) The Meaning Of Hydrophobicity *Science*, **250**, 297-298.
- Fariselli, P.; Pazos, F.; Valencia, A.; Casadio, R. (2002) Prediction Of Protein-Protein Interaction Sites In Heterocomplexes With Neural Networks *Eur J Biochem*, **269**, 1356-1361.
- Gohlke, H.; Hendlich, M.; Klebe, G. (2000) Knowledge-Based Scoring Function To Predict Protein-Ligand Interactions *J Mol Biol*, **295**, 337-356.
- Hubbard, S.; Thornton, J. (1993) NACCESS: Computer Program, Department Of Biochemistry And Molecular Biology, University College London.
- Jones, S.; Thornton, J. M. (1996) Principles Of Protein-Protein Interactions *Proc Natl Acad Sci U S A*, **93**, 13-20.
- Liddington, R. C. (2004) Structural Basis Of Protein-Protein Interactions *Methods Mol Biol*, **261**, 3-14.
- Larsen, T. A.; Olson, A. J.; Goodsell, D. S. (1998) Morphology Of Protein-Protein Interfaces *Structure*, **6**, 421-427.
- Lo, S. L.; Cai, C. Z.; Chen, Y. Z.; Chung, M. C. M. (2005) Effect Of Training Datasets On Support Vector Machine Prediction Of Protein-Protein Interactions *Proteomics*, **5**, 876-884.
- Lo Conte, L.; Chothia, C.; Janin, J. (1999) The Atomic Structure Of Protein-Protein Recognition Sites *J Mol Biol*, **285**, 2177-2198.
- Mintseris, J.; Weng, Z. (2003) Atomic Contact Vectors In Protein-Protein Recognition *Proteins*, **53**, 629-639.

- Mitchell, T. M. (1997) *Machine Learning*, McGraw-Hill.
- Nooren, I. M. A.; Thornton, J. M. (2003) Diversity Of Protein-Protein Interactions *EMBO J*, **22**, 3486-3492.
- Ofran, Y.; Rost, B. (2003) Analysing Six Types Of Protein-Protein Interfaces *J Mol Biol*, **325**, 377-387.
- Ponstingl, H.; Henrick, K.; Thornton, J. M. (2000) Discriminating Between Homodimeric And Monomeric Proteins In The Crystalline State *Proteins*, **41**, 47-57.
- Quinlan, J. R. (1993) *C4.5: Programs For Machine Learning* Morgan Kaufmann, Los Altos.
- Sanner, M. F.; Olson, A. J.; Spehner, J. C. (1996) Reduced Surface: An Efficient Way To Compute Molecular Surfaces *Biopolymers*, **38**, 305-320.
- Schoelkopf, B.; Smola, A. J. (2002) *Learning with Kernels* MIT Press, Cambridge, MA.
- Sheinerman, F. B.; Norel, R.; Honig, B. (2000) Electrostatic Aspects Of Protein-Protein Interactions *Curr Opin Struct Biol*, **10**, 153-159.
- Siepen, J. A.; Radford, S. E.; Westhead, D. R. (2003) Beta Edge Strands In Protein Structure Prediction And Aggregation *Protein Sci*, **12**, 2348-2359.
- Sotriffer, C.; Sanschagrin, P.; Klebe, G. *In preparation*.
- Tsai, C. J.; Xu, D.; Nussinov, R. (1997a) Structural Motifs At Protein-Protein Interfaces: Protein Cores Versus Two-State And Three-State Model Complexes *Protein Sci*, **6**, 1793-1805.

- Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. (1997b) Studies Of Protein-Protein Interfaces: A Statistical Analysis Of The Hydrophobic Effect *Protein Sci*, **6**, 53-64.
- SYBYL 7.0; Tripos Inc.: 1699 South Hanley Rd., St. Louis, Missouri, 63144.
- Valdar, W. S.; Thornton, J. M. (2001) Protein-Protein Interfaces: Analysis Of Amino Acid Conservation In Homodimers *Proteins*, **42**, 108-124.
- Vapnik, V. N. (1998) Statistical Learning Theory, John Wiley & Sons.
- Wang, R.; Lai, L.; Wang, S. (2002) Further Development And Validation Of Empirical Scoring Functions For Structure-Based Binding Affinity Prediction *J Comput Aided Mol Des*, **16**, 11-26.
- Wettschereck, D.; Aha, D. W.; Mohri, T. (1997) A Review And Empirical Comparison Of Feature Weighting Methods For A Class Of Lazy Learning Algorithms. *AI Review*, **11**, 273-314.
- Yan, C.; Dobbs, D.; Honavar, V. A (2004) Two-Stage Classifier For Identification Of Protein-Protein Interface Residues *Bioinformatics*, **20** Suppl 1, I371-I378.
- Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. (1997) Determination Of Atomic Desolvation Energies From The Structures Of Crystallized Proteins *J Mol Biol*, **267**, 707-726.
- Zhou, H. X.; Shan, Y. (2001) Prediction Of Protein Interaction Sites From Sequence Profile And Residue Neighbor List *Proteins*, **44**, 336-343.

## STRATEGIES TO SEARCH AND DESIGN STABILIZERS OF PROTEIN-PROTEIN INTERACTIONS: A FEASIBILITY STUDY

### INTRODUCTION

Major part of the presently known small molecule drugs are either enzyme inhibitors, allosteric effectors or receptor agonists or antagonists. They replace natural substrates or endogenous ligands mostly in deeply buried stringent binding pockets. Accordingly, the presently applied drug design tools are all methodologically focused on the competitive replacement of such ligands or substrate portions by appropriate lead structures that occupy a similar region of the deeply buried binding pocket. However, functional regulation of a biological system can also be achieved via interference with protein-protein interactions. In many processes, activation of protein function requires coactivation or cross-talk via the assembly of several protein components, e.g. by formation of a multidomain complex. A large number of signal transduction cascades, e.g. transferring information across the cell membrane, operate via the formation of protein-protein complexes. Accordingly, interference with this recognition process by means of small molecule drugs would – in principle – allow developing drugs of entirely new mode of action. However, inhibiting the interaction between two proteins is challenging in many respects, first of all due to unfavorable thermodynamic considerations. Usually such recognition complex interfaces are rather flat and featureless, accordingly it will be very difficult to bind a small molecule to such surfaces. Thus upon association of both proteins, the small molecule would have to compete at such surfaces with the interface formation. Considering the gain in entropy due to the hydrophobic effect and solvation/desolvation upon protein-protein complexation would simply “wash-off” any ligand associating with the surface of the protein. Therefore, it is not surprising that biological processes under physiological conditions often modulate the interaction between proteins allosterically. In such a case,

a ligand binds to a cavity remote from the protein-protein interface and modulates the complex formation via electrostatic effects or conformational rearrangements in one or both of the interacting proteins.

Amazingly, the opposite of inhibiting protein-protein interactions, namely stabilizing them, is a widely unstudied principle today. Such concepts require either allosteric stabilization or binding of a small molecule to a crevice at the rim of protein-protein interfaces thus fasten both macromolecular portions tighter together. Nevertheless, stabilizing protein complex formation may also result in the desired effect, for instance, by delaying a signal in a transduction cascade. Furthermore, the approach of stabilizing protein-protein interactions has other advantages compared to inhibition: On the one hand, similar to protein-ligand interactions, the thermodynamic properties are potentially favorable. On the other hand, established principles and tools from small molecule structure-based drug design can be applied. It has been estimated that life is controlled by over 50000 protein-protein interactions (Yin & Hamilton, 2005) there is supposedly a sizable number of targets forming a *druggable* cavity at the boundary of the protein complex. To get a rough idea about the potential occurrence frequency of such cavities, we analyzed a dataset of 198 so-called protein-protein recognition complexes which are known to be transient on a particular time scale. We extracted all cavities falling next to the margin of each protein-protein complex and inspected these cavities visually for putative *druggability*. If such cavities are frequently given and considering the increasing speed by which novel structures of protein-protein complexes are determined either by X-ray crystallography, NMR and other methods (Berman *et al.*, 2000), this may possibly open the floodgate to a new realm of targets for structure-based drug design.

Stimulated by the fact that rather frequently rim-exposed cavities are found at the protein-protein interfaces of potentially dissociating complexes we picked the example of a plant H<sup>+</sup>-ATPase forming a complex with a 14-3-3 protein. This complex is stabilized by the natural product Fusicoccin that binds into a small cavity composed by both protein



domains. In this contribution we report on strategies and novel tools developed to perform a virtual screening campaign to discover novel leads that could successfully stabilize a protein-protein interaction.

## MATERIALS AND METHODS

### DATA ANALYSIS AND TOOLS FOR VIRTUAL SCREENING

Inhibiting protein-protein interactions is a well studied field in protein science today and many successful examples have been summarized in recent reviews (Berg, 2003; Yin & Hamilton, 2005; Zhao & Chmielewski, 2005; Arkin, 2005). However, the major part of these inhibitors are from the area of monoclonal antibodies (Waldmann, 1993), miniature proteins (Martin *et al.*, 1994; Nord *et al.*, 1997), functional oligopeptides (Schneider *et al.*, 1995) or peptidomimetics (Moss *et al.*, 1996). Most of these inhibitors are far from being *druglike*. Nevertheless, there have also small and some *druglike* molecules been developed that inhibit protein-protein interactions efficiently, for instance, the interaction between (1) p53 and MDM2 (Fasan *et al.*, 2004; Vassilev *et al.*, 2004), (2) Bcl-xL and Bcl-2 (Baell *et al.*, 2002; Wang *et al.*, 2000), both playing a crucial role in cell apoptosis (Wang *et al.*, 2003; Reed, 1997), and (3) IL-2 and the  $\alpha$  subunit of its receptor (IL-2R $\alpha$ ) (Nguyen & Wells, 2003), (4) CTLA4 and B7-2 (Green *et al.*, 2003), both important for T-cell proliferation (Greenfield *et al.*, 1998). Despite these outstanding successes, the mentioned protein-protein interactions addressed in these examples by small molecules exhibit special features that unlikely allow to generalize them into a common inhibition strategy. This is due to the fact, that along the protein-protein interface one of the macromolecular interaction partners is either rather small and/or protrudes deeply into a buried binding pocket that forms across the protein-protein interface.

### STABILIZERS OF PROTEIN-PROTEIN INTERACTIONS

Many small molecules are known to be capable to induce the interaction between different proteins, e.g. the dimerization of receptors (Qureshi *et al.*, 1999; Tian *et al.*, 1998). Furthermore, the phenomenon of stabilizing an already formed protein-protein interaction has been observed both in physiological processes and as mode of action of some drug molecules. For example, the immunosuppressive drug Rapamycin (also known as Sirolimus), a triene macrolide antibiotic, which also exerts anti-fungal, anti-inflammatory, and anti-tumor effects, acts as a protein-protein interface stabilizer. After binding to the receptor protein FKBP12, the complex binds to mTOR (mammalian Target Of Rapamycin). This prevents the subsequent interaction of mTOR with other target proteins in the signaling pathway (Choi *et al.*, 1996). The immunosuppressant Tacrolimus operates following to a related principle by tightening up two different proteins, which otherwise do not show measurable affinity for each other (Griffith *et al.*, 1995). Further prominent stabilizers are (1) Brefeldin A, a small molecule which stabilizes the transient ternary complex between Arf-GDP and its guanine nucleotide exchange factor (Peyroche *et al.*, 1999; Chardin & McCormick, 1999); (2) Taxol, a diterpenoid, acting as effective anti-cancer drug by stabilizing microtubules and preventing their de- and repolymerization (Jennewein *et al.*, 2001); (3) Forskolin, a diterpene, by stabilizing subunits of the adenylylcyclase (Tesmer *et al.*, 1997); (4) Kirromycin and fusidic acid, by interfering with the peptide transfer in the protein biosynthesis of prokaryotic ribosomes by stabilization processes (Tesmer *et al.*, 1997; Agrawal & Frank, 1999; Ramakrishnan, 2002) .

### SCREENING FOR NOVEL TARGETS

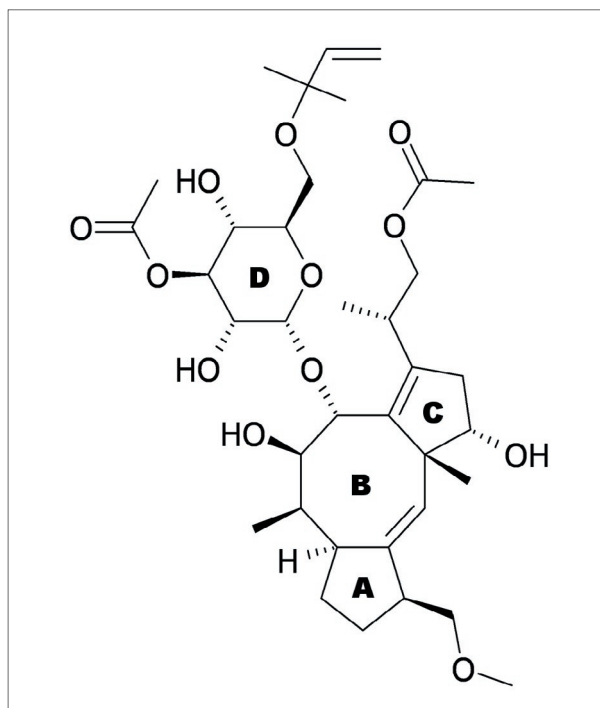
As pictured above, special prerequisites have to be given that a protein-protein interaction can be inhibited by small molecule binding. Accordingly, as an alternative it appears more attractive to address cavities or crevices formed along the rim of an interface to modulate

the stability of the protein-protein complex formation. To obtain first insights into the occurrence of such cavities at the margin of protein-protein interfaces, we screened a dataset of 198 recognition complexes reported by Mintseris *et al.* (Mintseris *et al.*, 2003). We particularly considered recognition complexes, since they are known to be temporarily labile and to dissociate under dynamic conditions whereas folding complexes have permanent character avoiding separation. To detect and extract appropriate cavities we applied Relibase+ (Guenther *et al.*, 2003; Hendlich *et al.*, 2003) in combination with its modular extension Cavbase (Schmitt *et al.*, 2002), by use of the implemented Ligsite algorithm (Hendlich *et al.*, 1997). We screened the entire dataset and detected about 380 rim-exposed cavities which were constructed by more than one chain. To prove our assumption, that rim-exposed cavities show properties comparable to those of enzyme cavities, which due to their function accommodate ligands, we compiled a dataset of 2379 enzymes with a maximum of 25% mutual sequence identity according to the recent PDB SELECT list (Hobohm *et al.*, 1992; Hobohm & Sander, 1994). The filtering was accomplished by mapping the PDB code to the PDBsum database (Laskowski *et al.*, 1997, 2005) and extracting only structures with assigned E.C. number. Subsequently, we considered only pockets with a bound ligand comprising more than six heavy atoms. This procedure resulted in 243 enzymes with 636 cavities (in case of multimeric enzymes complexes the number of cavities was counted for each monomer.) We compared the volume, the relative hydrophilicity, and the buriedness of the pocket contributing atoms between enzyme and rim-exposed cavities. The relative hydrophilicity was determined by the ratio of exposed hydrophilic and hydrophobic properties (coded by the assigned pseudocenters in Cavbase, for details see Schmitt *et al.*, 2002) in the binding pockets. The buriedness and pocket volume are descriptors calculated by Ligsite. Here pronounced buriedness is reflected by a large number of deeply buried cavity atoms.

Cavbase detects common subsets of pseudocenters in pockets to be compared and after superimposition a similarity score is calculated by evaluating the amount of matching surface properties (Schmitt *et al.*, 2002). To estimate on a possibly given *druggability* of rim-exposed binding pockets we evaluate the similarity with different enzyme pockets.

### THE H<sup>+</sup>-ATPASE/14-3-3 SYSTEM

In this study, we screened for alternative stabilizers of the interaction between plant H<sup>+</sup>-ATPase and 14-3-3 protein. The ATPase builds up an electrochemical proton gradient across the plasma membrane, which is important for maintaining of the cell turgor (Morsomme & Boutry, 2000). The C-terminus of the proton-pump as well as the N-terminus is located within the cytoplasm of the cell and acts as an intrasteric inhibitor (Kuhlbrandt *et al.*, 2002). The autoinhibition force is subsequently alleviated upon complexation with the 14-3-3 protein (Fuglsang *et al.*, 1999; Svennelid *et al.*, 1999; Maudoux *et al.*, 2000). The latter proteins are highly conserved molecules regulating various physiological processes. They are known to bind in a sequence-specific and phosphorylation-dependent manner to their targets (Sehnke *et al.*, 2002; Tzivion & Avruch, 2002; Yaffe, 2002). However, the stabilization of a given protein-protein complex results in an irreversible activation of the proton-pump, followed by an irreversible opening of the stomatal pores finally cumulating in the wilting of the plant (Ballio *et al.*, 1964). This effect is stimulated by the binding of the fungal phytotoxin Fusicoccin (FC, Fig. 1), which boosts the weak interaction between the ATPase and 14-3-3 protein by nearly a factor of 100 (Wurtele *et al.*, 2003). Although the fungus *Fusicoccum amygdali*, which is producing the diterpene Fusicoccin, is host specific, isolated Fusicoccin exerts its action in almost any higher plant (Marre *et al.*, 1979). Accordingly, Fusicoccin is a total herbicide, which turns it into an interesting agrochemical. However, from a synthetic point of view, Fusicoccin is a complex molecule. Nevertheless, any molecule, which stabilizes this protein-protein complex and is accessible to simple synthesis could be highly attractive. In a search for alternative stabilizers, a large number of Fusicoccin derivatives have been

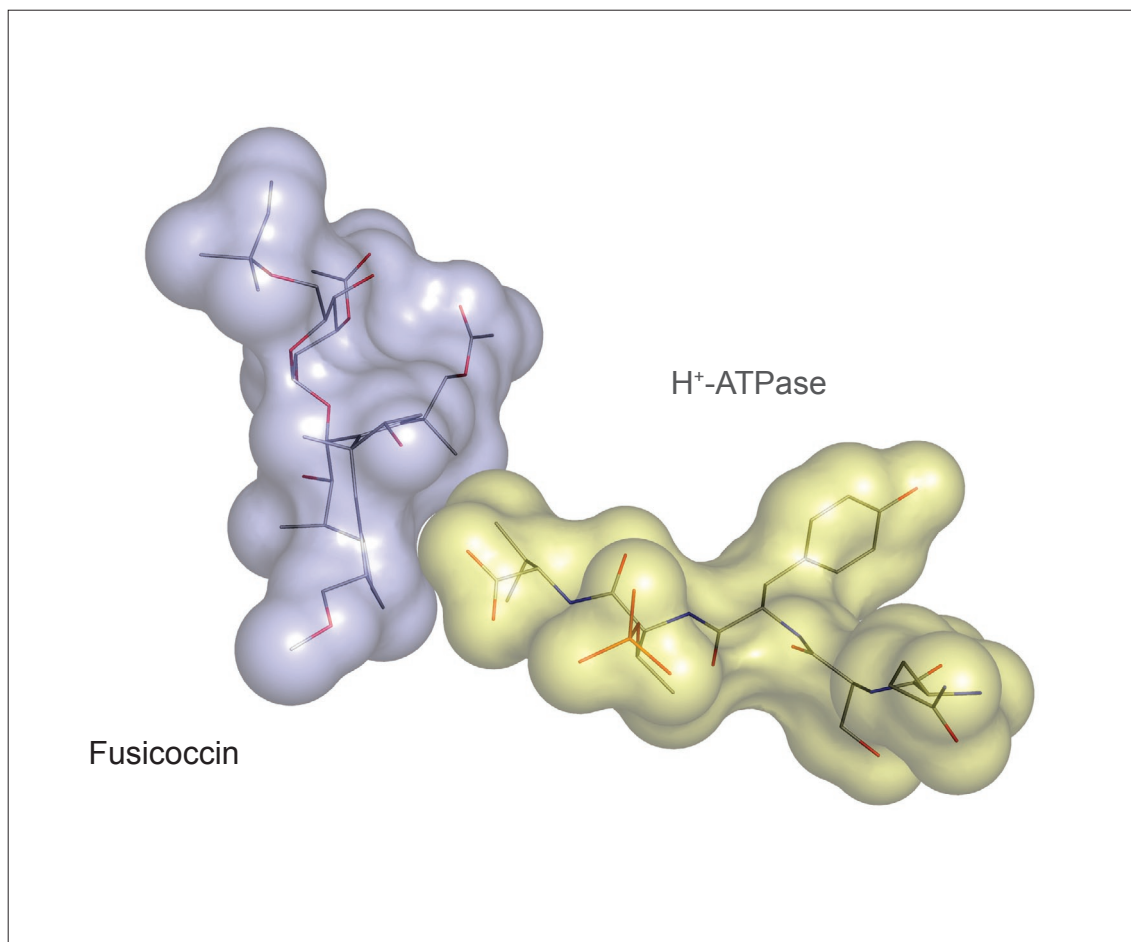


**FIGURE 1.** The structure of the fungal toxin Fusicoccin (FC).

synthesized as summarized by Ballio (Ballio *et al.*, 1979). It turned out, that already minor modifications of the FC ring system strongly impairs biological activity and the potency to stabilize the given protein-protein complex. This stimulated us to search for novel scaffolds, which potentially mimic the mode of action of Fusicoccin.

### THE FUSICOCCIN BINDING SITE

Major part of the Fusicoccin binding site is composed by the amino acids of the 14-3-3 protein; only the C-terminal Val5 of the H<sup>+</sup>-ATPase contributes to the surface of the binding pocket. It has been shown that the binding affinity of Fusicoccin to the 14 3 3 protein is very weak as long as not complexed to the ATPase (K<sub>i</sub>: 66 μM) (Wurtele *et al.*, 2003). This is due to the fact that upon complexation with the H<sup>+</sup>-ATPase the FC binding site widens up by a small but crucial extent: The hydrophobic sub-pocket, comprised by Ile175 and Ile226, is extended by the hydrophobic side chain



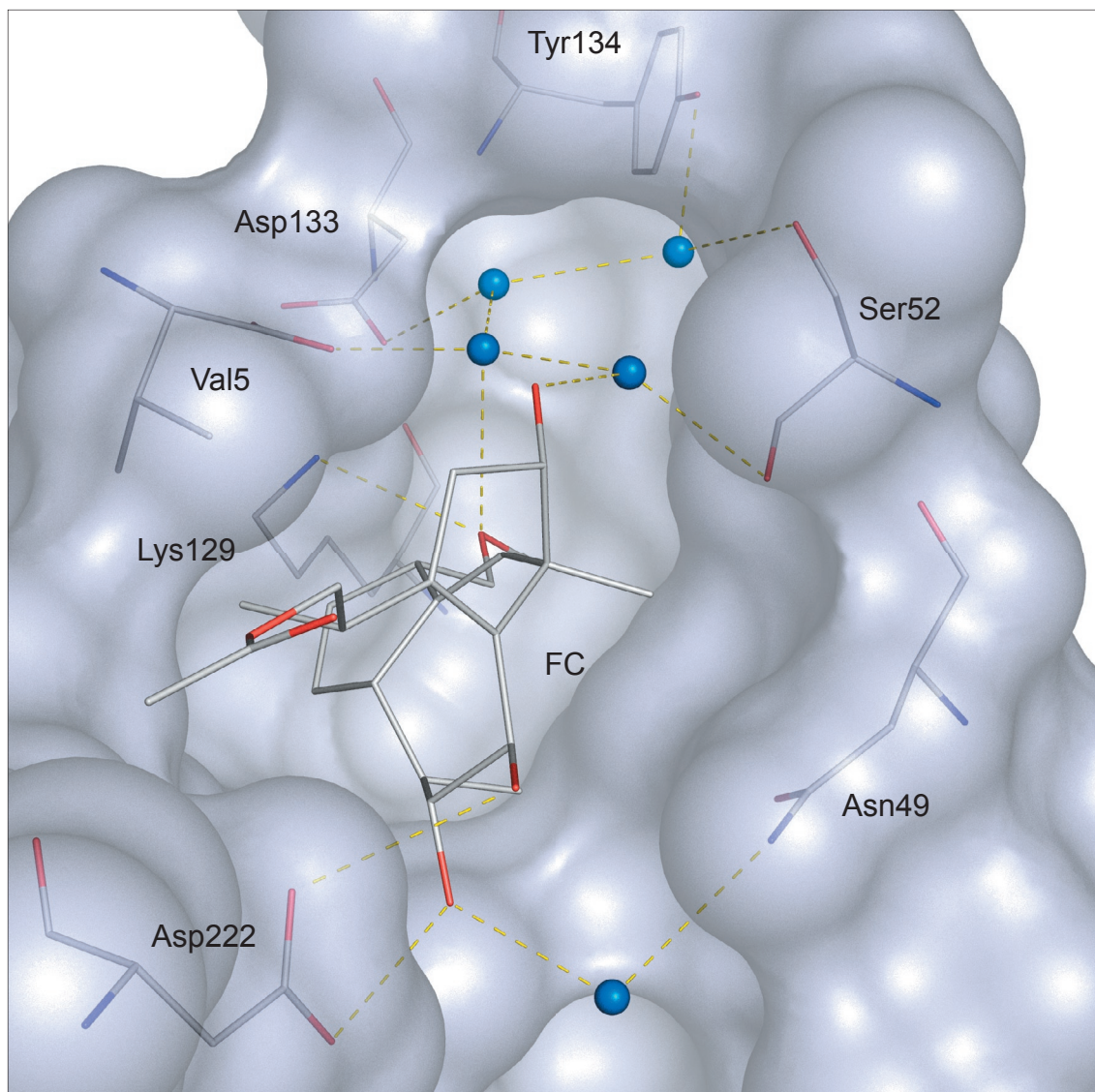
**FIGURE 2.** Fusicoccin (blue) and the pentapeptide representing the H<sup>+</sup>-ATPase (yellow) in stick and surface representation. The hydrophobic ring system (ring A, B and C) are wrapped around the C-terminal Val5 of the H<sup>+</sup>-ATPase. The tight binding of Fusicoccin in the 14-3-3 binding site and the hydrophobic stacking to the H<sup>+</sup>-ATPase results in a nearly 100-fold increased affinity between 14-3-3 and H<sup>+</sup>-ATPase.

of the C-terminal amino acid (Val5 in case of PDB 1o9f). This side chain gets virtually clasped by the carbocyclic framework of Fusicoccin (Fig. 2), resulting in an extensive hydrophobic contact upon binding. The hydrophobic environment is completed by Val53 and Phe126 that are located toward the “rear” to Fusicoccin. Furthermore, the unprotonated carboxylic C-terminus of the H<sup>+</sup>-ATPase is supposed by forming a salt bridge with protonated Lys129, which adopts a *gauche* conformation. In consequence the positively charged amino group of Lys129 is (weak) hydrogen bonded to the ether function of Fusicoccin (Fig. 3). Furthermore, FC creates tight contacts to the 14-3-3 protein to enhance the protein-protein interaction. The alcohol function at ring B of Fusicoccin forms a hydrogen bond to the most likely unprotonated carbocyclic acid

of Asp222 (Fig. 3) and a further water-mediated hydrogen bond to Asn49. In total, Fusicoccin exhibits nearly perfect shape complementarity to the target protein (Fig. 4). In conclusion, the stabilizing effect of FC binding to the 14-3-3/H<sup>+</sup>-ATPase complex can be explained (1) by the close interaction of FC with the C-terminal amino acid of the H<sup>+</sup>-ATPase, (2) by FC's shape complementarity to the target protein, and (3) by the mainly water-mediated hydrogen-bond network.

#### WATER AND THE FUSICOCCIN BINDING POCKET

A closer examination of the crystal structure (PDB 1o9f) reveals that FC does not completely fill the binding pocket. A sub-pocket, exclusively composed by amino acids of the 14-3-3 protein and located next to the tip of the C-terminus of the H<sup>+</sup>-ATPase, hosts four water molecules (Fig. 3). Possibly this sub-pocket exists because other H<sup>+</sup>-ATPase occupy this region by the extension of one amino acid at their C-terminus. The water molecules form a tight hydrogen-bond network with the amino acids of the 14-3-3 protein and among each other. Also FC is involved in this network by building a tight hydrogen bond of 2.51 Å via its alcohol function at ring C to one of the water molecules, and a loose hydrogen bond to the ether-attached ring A (3.36 Å). Since ethers are relatively weak acceptors, and this function is supposedly already involved in a hydrogen bond to the positively charged amino group of Lys129, an additional contact to the latter water molecule appears unlikely. Nevertheless, the replacement of these water molecules upon ligand binding may result in a favorable, entropy driven binding. In our design attempts to discover alternative molecular skeletons to stabilize the present protein-protein interaction we assume that this replacement contributes significantly to binding. Nevertheless, mimicking the overall bowl-shape of FC by a small, *druglike* molecule appears a rather challenging task.

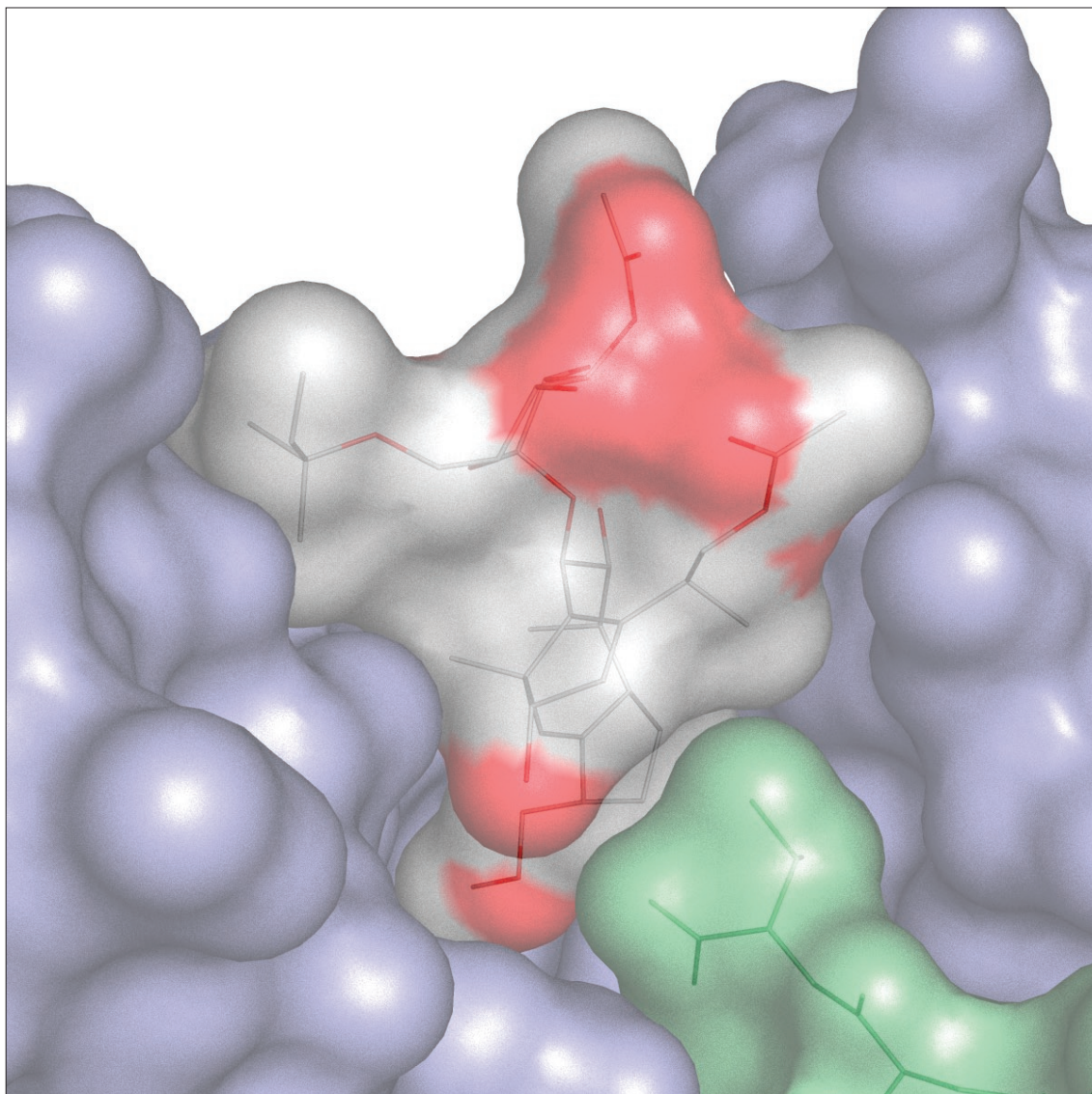


**FIGURE 3.** A hydrogen-bond network is formed by four water molecules and Fusicoccin in a sub-pocket of the binding site. Further interactions are formed between the ligand and the H<sup>+</sup>-ATPase protein chain. Additional amino acids involved in the binding are Val5, Ser52, Asp133 and Tyr134. In addition, Fusicoccin participates with its alcohol function at ring C and its ether group in forming hydrogen bonds to the 14-3-3 protein. The alcohol function at ring B of Fusicoccin forms hydrogen bonds to Asp222 and a water-mediated hydrogen bond to Asn49.



## VIRTUAL SCREENING CAMPAIGNS

Two strategies can be followed in a virtual screening campaign: forwards and backwards filtering of hits obtained by docking. In forward filtering, various criteria are used to reduce the initial screening sample of candidate molecules from some millions to several hundred to thousand most promising ones to be docked. In backwards filtering all entries from the screening sample are docked to the target and filter criteria are subsequently applied to rank the generated docking solutions. Since flexible docking is computationally the most demanding step, forward filtering requires less resources. Usually filter criteria are based on some preconceived knowledge about the target or the chemical structures of known actives. They exploit similarity considerations with either a protein-based pharmacophore hypothesis or the properties of bound ligands. Applying such filters in too stringent fashion could possibly result in a biased search discarding unexpected and novel chemistry at early filtering steps. In the past, we have been quite successful in a hierarchical forward filtering strategy (Brenk *et al.*, 2003; Grueneberg *et al.*, 2002; Evers & Klebe, 2004). However, in these studies we selected well-established enzyme targets and took reference to some known *druglike* ligands. The crevice formed along the H<sup>+</sup>-ATPase/14-3-3 protein interface appears as a novel type of binding pocket and the accommodated FC shows most likely not a high molecular similarity with entries in the candidate library. Under these given restraints we decided to follow a rather unbiased backward filtering approach starting with large scale docking of candidate ligands to the binding pocket, followed by a post-filtering taking reference to sophisticated 3D-pharmacophores. The advantage of this strategy is that different pharmacophore hypotheses can be rapidly applied to all generated docking solution. Furthermore, this allows us easily to select the number of compounds we want to consider for further in-depth searches. Nevertheless, this strategy involves the demanding task to keep track of a huge amount of data generated during the virtual screening run. Accordingly, elaborate tools to extract the relevant information are desirable. In consequence, we developed SCREENINGDB, a software suite written in Python and C with



**FIGURE 4.** Pronounced shape complementarity of Fusicoccin (white, red) in the binding pocket formed by the 14-3-3 protein (blue) and the pentapeptide representing the H<sup>+</sup>-ATPase (green).

an integrated MySQL database. SCREENINGDB stores all relevant information produced by the tools FLEXX, AutoDock, GOLD, or FTREES along with the CORINA-generated 3D models. The core of SCREENINGDB comprises the compound library holding about two million candidate molecules from various commercial compound suppliers (Tab. 1). It stores primary molecule information such as name and origin (database), a mol2- (Clark *et al.*, 1989) and SMILES-format (Weininger, 1988) representation, as well as derived

information such as molecular mass, number of total or heavy atoms, or number of rotatable bonds. The latter is calculated using the SFCscore library (Sottriffer *et al.*, *in preparation*), and considers amide and ester groups as rigid.

### PREPROCESSING OF THE CANDIDATE MOLECULES

We extracted all candidate molecules from commercial compound suppliers, listed in Tab. 1, and processed them in a unique fashion by BABEL (see ref. BABEL) and CORINA 3.1 (Gasteiger *et al.*, 1990). CORINA was used to assign Sybyl atom types

Database	No. of Compounds
ACD	239929
AMBINTER	571309
ASINEX BBLOCKS	4108
ASINEX GOLD	227273
ASINEX PLATINUM	113617
BIONET HTS	41333
CLAB	70343
IBS	287883
LEADQUEST	81201
MAYBRIDGE	59485
MAYBRIDGE HITFINDER	15997
SIGMA ALDRICH	5744
SPECS	169293
SPECS NATURAL	669
<b>Total</b>	<b>1940184</b>

**TABLE 1.** Selected databases of commercially available compounds, which were used for the virtual screening campaigns. In total, nearly two million of molecules are covered.

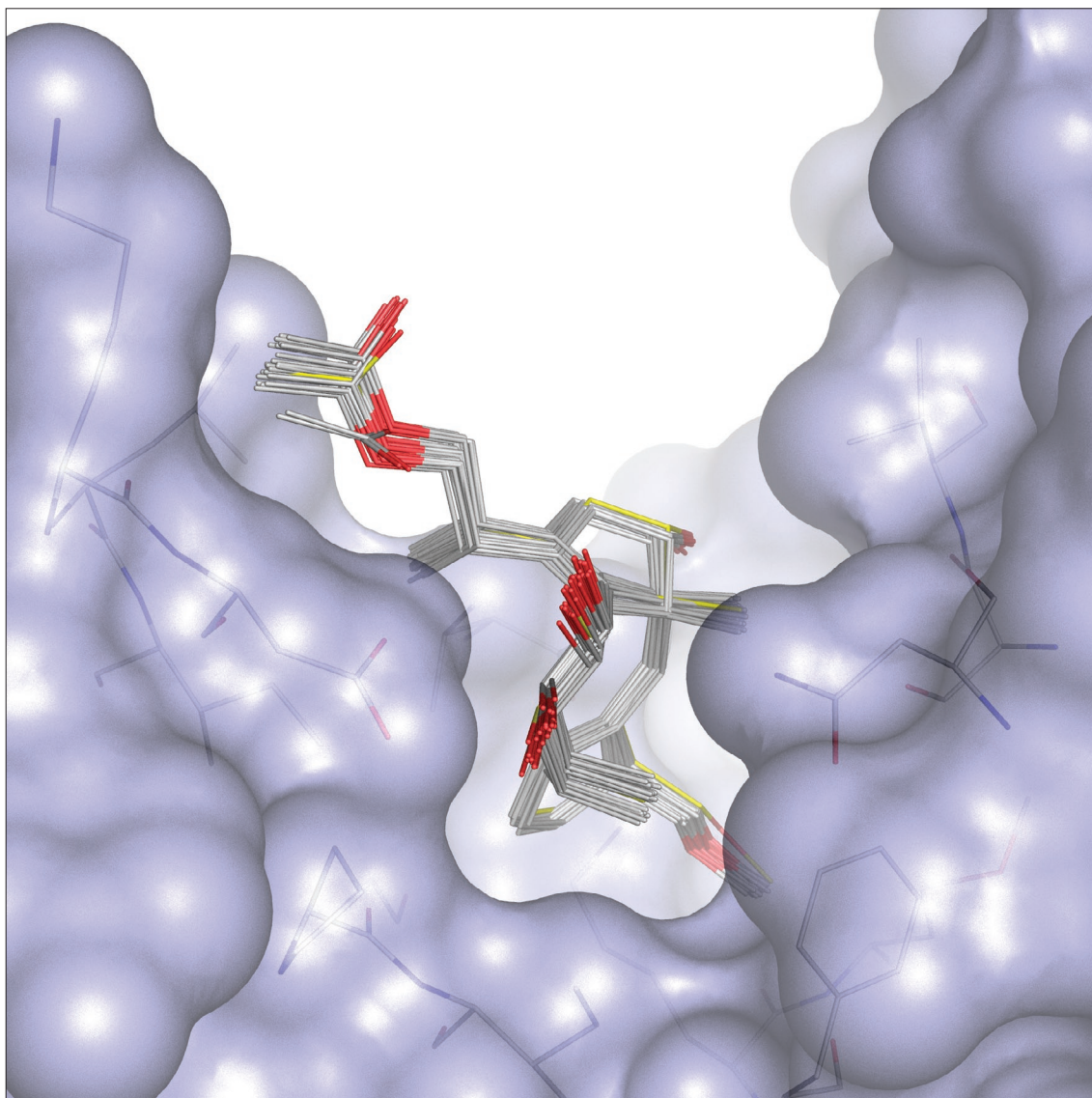
(`corina -d newtypes`) to each entry. Since CORINA requires molecules without hydrogens for the atom type assignment, we removed all hydrogen atoms with BABEL (`babel -d`). Candidate molecules stored as salts were stripped and neutralized. In case of a tautomeric or stereoisomeric molecule, the database given isomer was preserved. Finally, *in house* software corrected atom types in moieties such as amidines and guanidines, not properly handled by BABEL nor CORINA. Canonical SMILES were generated with the FROWNS library (Kelly, WWW). Additionally, we developed a command-line driven tool to access the database. For example, to retrieve all structures within the compound library, possessing a molecular weight between 250 and 450 Da, a maximum number of six rotatable bonds together with a carboxylic acid moiety requires only a few seconds. The tool allows the user to save the result in either mol2 format, as input for the subsequent docking campaign, or as SMILES annotation.

#### **DOCKING FUSICOCCIN**

To access the reliability of the docking tools FLEXX 2.0 (Rarey *et al.*, 1996), GOLD 2.2 (Verdonk *et al.*, 2003), and AUTODOCK 3.0 (Murray *et al.*, 1998), Fusicoccin was redocked into the binding pocket of the H<sup>+</sup>-ATPase/14-3-3 complex. Since the sugar moiety of Fusicoccin is not involved in protein binding (Ballio *et al.*, 1971, 1981), we only docked the aglycon. All three programs are capable to place the aglycon with a near-native binding mode into the binding pocket. GOLD generated the top 30 docking poses out of 50 attempts within 0.4 Å rms deviation to the native binding mode of FC (Fig. 5), whereas 98 out of 100 docking poses of AUTODOCK were found within 1.1 Å to the crystal structure. FLEXX placed 10 out of 30 docking solutions better than 2.0 Å, which is a commonly accepted threshold for near-native binding modes.

**“HOT SPOT” ANALYSIS**

We calculated “hot spots” in the Fusicoccin binding site of the H<sup>+</sup>-ATPase/14-3-3 complex using Superstar (Verdonk *et al.*, 1999) and DrugScore (Gohlke *et al.*, 2000). The binding pocket, accommodating Fusicoccin appears rather hydrophobic as observed by the fields for the methyl-carbon probe of Superstar and the C.2, C.3, C.ar probe of DrugScore.

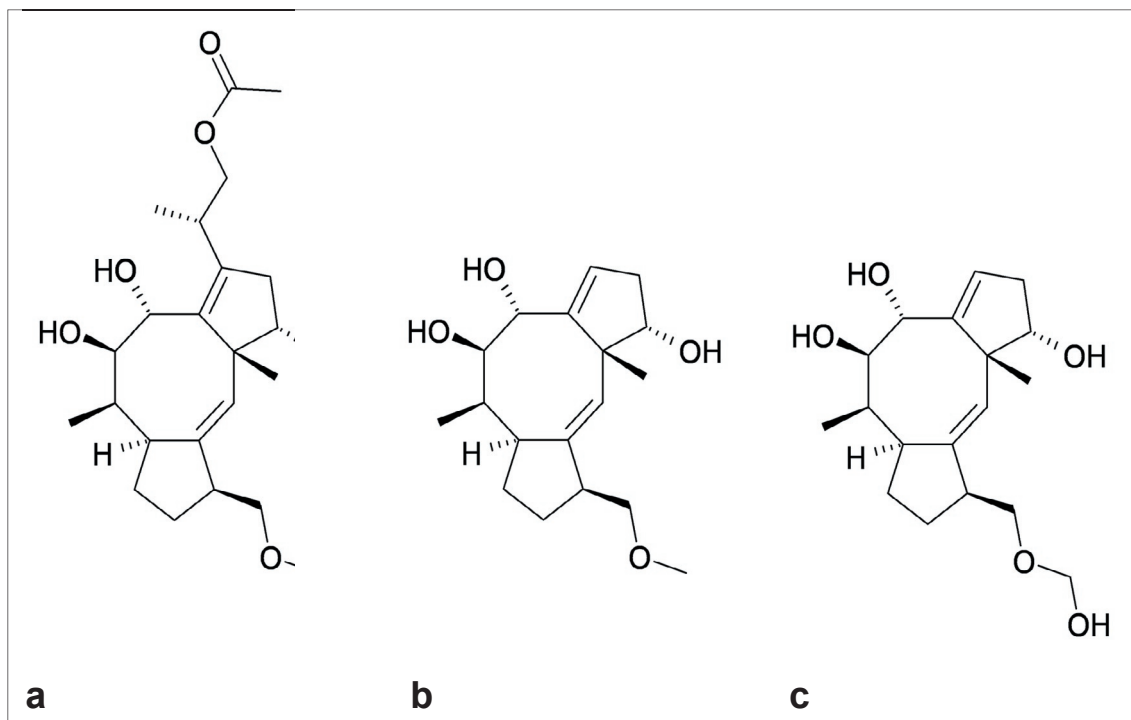


**FIGURE 5.** Redocked Fusicoccin aglycon. The native binding mode (yellow) superimposed with the 30 top-ranked docking solutions of GOLD 2.2. They show up to 0.4 Å rms deviation to the crystallographically determined binding mode.

They indicate extensive hydrophobic character of the pocket. It is mainly located in the region where the carbocyclic skeleton of Fusicoccin is found in the crystal structure. The carbonyl oxygen and the O.2 probe, respectively, represent generic hydrogen-bond acceptors; they indicate a “hot spot” next to Fusicoccin’s ether function. Additionally, the sub-pocket filled by water is suggested as favorable acceptor region, in agreement with crystallographic evidence. An alcohol oxygen and an O.3 probe, respectively, which represents a generic hydrogen bond donor, propose a donor functionality in a region that is actually occupied by the alcohol group at ring B of Fusicoccin and by the waters in the sub-pocket. Accordingly, the generated pharmacophore hypotheses corresponds well with the adopted binding pose of the reference ligand Fusicoccin.

### FTREES

For feature-based ligand similarity searching, we incorporated FTREES (Rarey & Dixon, 1998) into SCREENINGDB. Physicochemical properties are assigned to the functional groups in each candidate molecule together with their topographical location and this information is translated into a so-called Feature Tree. The optimal similarity of two Feature Tree entries is calculated by an alignment score, similarly as for the alignment of protein sequences. The similarity scores are scaled to values between 1 (identical) and 0 (completely different). In literature (Rarey & Stahl, 2001), the degree of similarity has been discussed with respect to the Feature Tree scoring and values between 0.9 and 0.95 suggest molecules with highly analogous properties to the reference. However, they still potentially exhibit different scaffolds (so-called *scaffold-hopping*). FTREES was applied to retrieve candidate molecules with FTree-similarity to Fusicoccin which was used as reference molecule (Fig. 1). The most similar molecules found in the screening database exhibited a score of only 0.84, thus clearly beyond the above-mentioned range of high similarity. This rather low score underlines the unique chemical structure of FC which has to be assumed as rather remote from being *druglike* (Oprea *et al.*, 2001), as all entries assembled in the database were requested to obey such properties. To increase



**FIGURE 6.** FTREES reference molecules: (a) Fusicoccin aglycon, (b) Fusicoccin aglycon without ester moiety, and (c) Fusicoccin aglycon without ester moiety and additional methylene hydroxy function, which was added as a “decoy” to address the water sub-pocket.

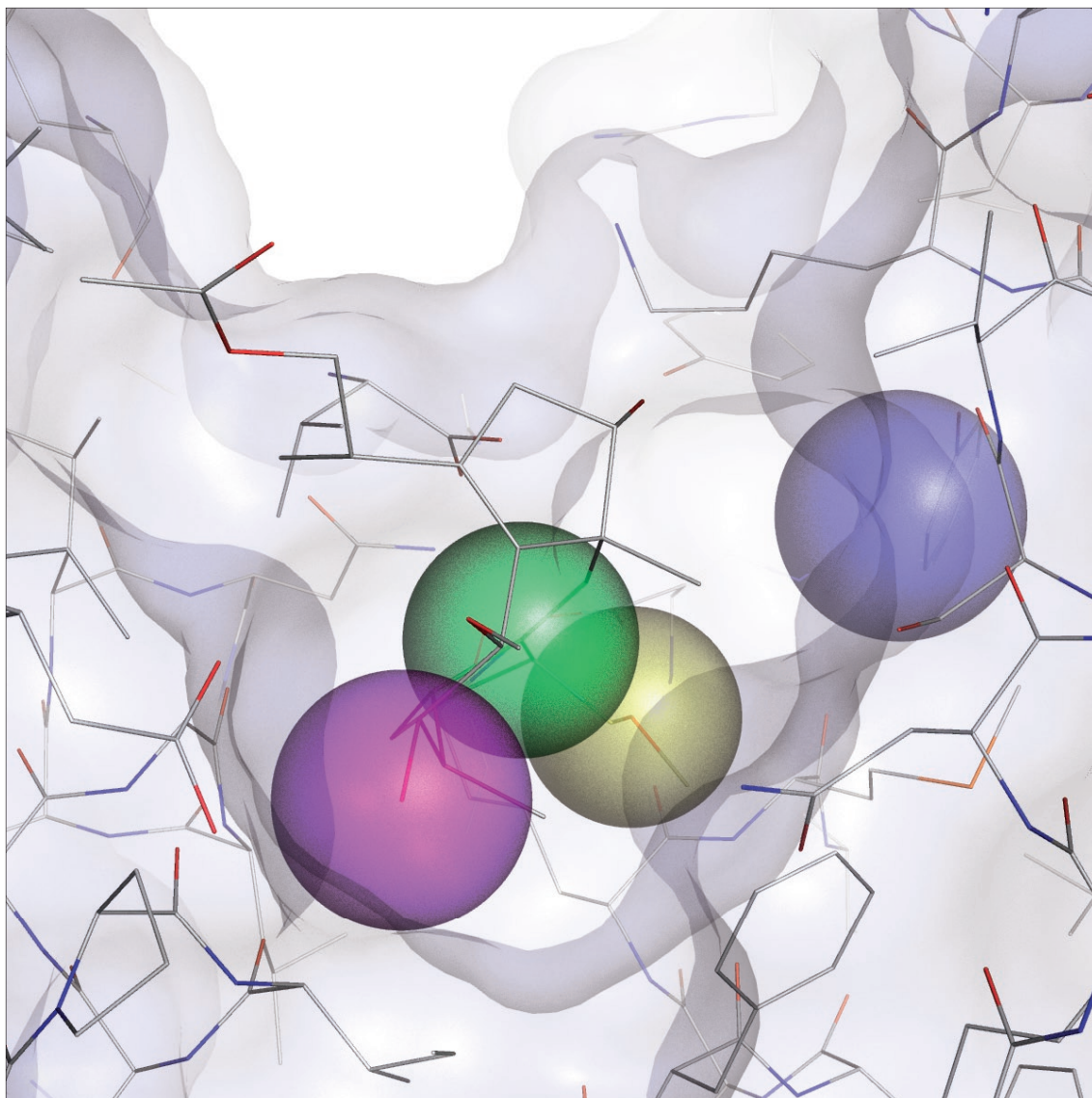
the probability to achieve higher similarity scores in particular since FTREES does not distinguish between molecular portions determinant for binding (e.g. carbocyclic framework) or exhibiting decoration only relevant for e.g. solubility as the sugar, we artificially reduced FC to its basic skeleton. Firstly, we deleted the sugar moiety (Fig. 6a), then the ester portion (Fig. 6b), since it is not involved in essential contacts to the proteins. Finally, to find molecules addressing the water sub-pocket, we extended the ether group, attached to ring A, by a methylenhydroxy function (Fig. 6c). Such doing allows matching candidates with similarity scores of up to 0.9. A set of 15.481 unique compounds with the highest similarity scores was extracted from SCREENINGDB for the subsequent docking analysis.

## UNITY DATABASE SEARCH

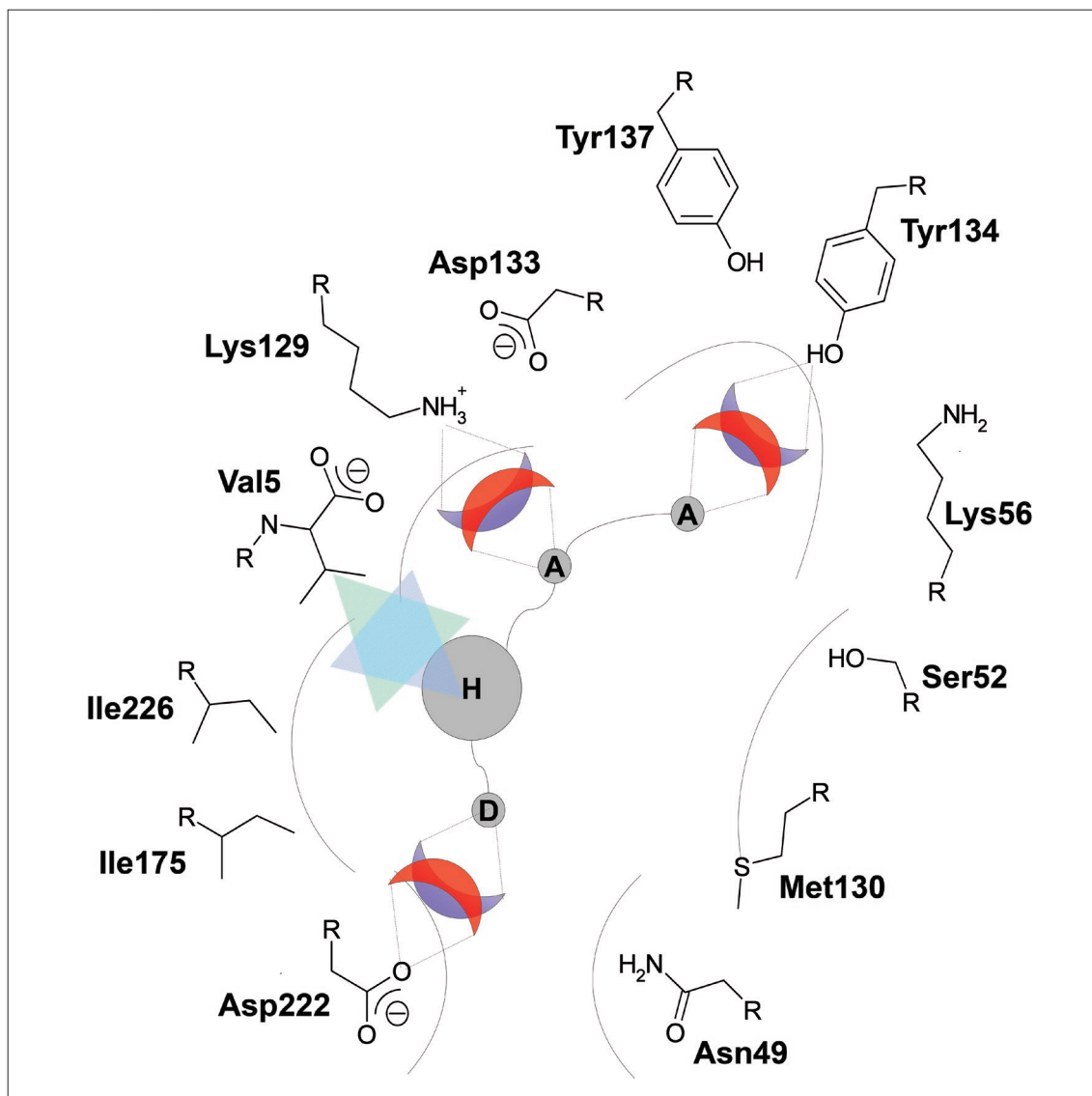
For the present target only Fusicoccin is known to stabilize this protein-protein interaction. This limited information stimulated us to choose a slightly modified screening strategy to previous attempts of Brenk *et al.* (Brenk *et al.*, 2003), Grueneberg *et al.* (Grueneberg *et al.*, 2002), or Evers *et al.* (Evers & Klebe, 2004). Similarly, UNITY (Martin *et al.*, 1992) was used to initially filter on simple criteria. However, deviating from the previous screening campaigns sophisticated 3D-pharmacophore filters has not been applied for forward filtering. Instead we decided in the present case with a much less defined concept about the chemistry to be matched by putative stabilizers to apply large-scale docking and to filter the generated docking solutions backwards by considering a variety of different 3D-pharmacophore filters.

Only molecules with up to 6 rotatable bonds and a molecular mass between 250 and 500 Da were considered. Furthermore, highly flexible molecules were discarded, since they supposedly show reduced binding affinity due to entropic losses and increase the complexity for the subsequently applied flexible docking. In total, 35% of the initial database were discarded by these filters. Subsequently, a simple topological filter was applied according to the pharmacophore hypothesis given in Fig. 7 (as scheme: Fig. 8). The central hydrophobic “hot spot” had to be matched by a hydrophobic ring position (green sphere, matching ring A of FC) along with one of the following features: (a) a donor and/or acceptor function had to be placed into the water sub-pocket pointing to Tyr134 (blue sphere, coinciding the water molecules in PDB 1o9f), (b) a donor function addressing the carboxylic acid of Asp222 (magenta sphere, located at the hydroxy oxygen of ring B of FC), or (c) an acceptor functionality complementing the charged amino group of Lys129 (yellow sphere, located at the ether oxygen attached to ring A of FC). Using this pharmacophore UNITY reduced the initial set of compounds to about 10%. Together with the molecules retrieved by FTREES search a total screening set





**FIGURE 7:** Pharmacophore hypotheses. All docking poses were filtered combinations out of these pharmacophoric constraints. Green: Central lipophilic constraint; Yellow: Acceptor constraint; Magenta: Donor constraint; Blue: Donor/Acceptor (“doneptor”) constraint.



**FIGURE 8.** Pharmacophore scheme according to the „hot spot“-analysis found for the Fusicoccin binding pocket. Important residues (Val5, Lys129, Tyr134, and Asp222) are displayed with the interaction properties explicitly requested in the different pharmacophore hypothesis. Together with the schematically sketched ligand the requested interaction types are indicated as acceptor (A), donor (D), „doneptor“ (X), and hydrophobic (H) respectively.

of 161.171 compounds was subjected to docking. As docking tools FLEXX 2.0, GOLD and AUTODOCK were used running on a 14 AMD Opteron™ 1.8 GHz computer cluster. The obtained results were deposited in SCREENINGDB.

## SCORING

The original scores of FLEXX, GOLD and AUTODOCK are automatically stored in SCREENINGDB. Additionally, we implemented external scoring functions such as DrugScore<sup>PDB</sup> (Gohlke *et al.*, 2000) and DrugScore<sup>CSD</sup> (Velec *et al.*, 2005) in SCREENINGDB, along with routines sorting according to these scores. Storage of different scores in the database provides the opportunity to build “secondary scores”, such as “consensus scores” or “cubic root scores” (score divided by the cubic root of the number of its heavy atoms to down-weight scores of increasingly large molecules) or to combine scores with other data as the number of rotatable bonds. We also stored the calculated burial of solvent accessible surface area ( $\Delta$ ASA) upon docking for each binding site residue using NACCESS (Hubbard & Thornton, 1993).

## PHARMACOPHORE POST-FILTERING

For post-filtering of the generated docking poses, SCREENINGDB provides an interface to apply pharmacophore characteristics. The presence of pharmacophoric features has only been considered in terms of spheres although other shapes such as cones or boxes have been implemented. The attributes assigned to such a sphere are its interaction-type, spatial location, and radius. The interactions are classified by the following types: DONOR, ACCEPTOR, DONEPTOR (either donor or acceptors, e.g. OH groups), EXCLUDED VOLUME or SPATIAL. As the names of the interaction types implies, DONOR, ACCEPTOR and DONEPTOR evaluate the docking solutions with respect to their match with atoms capable to form hydrogen bonds as donor, acceptor or donor/acceptor atoms, respectively, regarding the assigned spatial constraints. The property of

an atom being hydrogen-bond donor, acceptor or both is defined by their Sybyl atom type (Clark *et al.*, 1989), where donor atoms are O.3, N.am, N.4, N.pl3, acceptor atoms are N.1, N.ar, O.co2, O.2, O.3 and “doneptor” atoms are N.2, N.3, and O.3. EXCLUDED\_VOLUME constraints define areas where no atom is allowed to penetrate into, whereas SPATIAL explicitly demands pronounced spatial occupancy. A particular pharmacophore model can be composed by any number and type of constraint, nevertheless SCREENINGDB stores the result of each matched constraint individually. Thus, if we decide at a later stage to filter according to a modified pharmacophore hypothesis, the already evaluated constraints can be easily re-examined. This will speed-up any subsequent filtering steps by re-using the already processed constraints. The retrieval of candidate ligands passing individual filters is rendered automatically by cross-referencing docking solutions with matching pharmacophore constraints. In addition, SCREENINGDB gives access to all derived molecular properties such as the molecular mass, the number of rotatable bonds, etc.

SCREENINGDB is able to filter approximately 25.000 docking solutions per minute on currently available processors (Intel® Pentium® 4, 3.0 GHz, 1 GB RAM). The computational demands for pharmacophore filtering increases with the size of the candidate ligands and the complexity of the applied pharmacophore query. In the analysis, each molecule proceeds through the following steps: (1) retrieve compound data from database, (2) decompress data, (3) filter according to pharmacophore hypothesis (4) deposit results in the database. The speed limiting step is data retrieval from the database. Although the decompression step is computationally expensive too, storing uncompressed files in the database (a) slows down the data retrieval process and (b) blows up the storage requirement of the database dramatically. Nevertheless, benchmarking showed comparable results for handling compressed or uncompressed data, accordingly we decided to store compressed data by default. SCREENINGDB requires about 8 GB disk space, holding about  $2 \cdot 10^5$  compounds and  $5 \cdot 10^5$  docking solutions. Storing for each docking solution only the transformation matrix with the internal

torsional changes with respect to the reference geometry of each entry would reduce the data volume by about 25%, however additional computing time would be required to regenerate the coordinates of the docking solutions.

### VISUAL INSPECTION

We filtered about  $5 \cdot 10^5$  docking poses with respect to different pharmacophore constraints and ranked the pharmacophore matching solutions according to various scoring schemes such as DrugScore<sup>PDB</sup>, DrugScore<sup>CSD</sup> and the difference in the solvent accessible surface ( $\Delta$ ASA). The selected compounds placed into rank-ordered hit lists were prepared for visual inspection in the binding pocket. Although scoring functions try to capture all present knowledge about binding poses in algorithmic fashion our experience shows that due to the multifactorial correlation of binding mode with binding energetics the visual assessment through the eye of an experienced expert is of utmost importance. Such analyses focus on generated binding modes with respect to conformational distortions, contact surface complementarity, putative involvement of interstitial water molecules or remaining unoccupied voids along the protein-ligand interface. Accordingly, we visually inspected about 500 of the top-scored hits and examined their multiple docking poses. If looking promising, they were selected for the more elaborate docking using AUTODOCK and GOLD.

### DOCKING WITH GOLD AND AUTODOCK

Both GOLD 2.2 and AUTODOCK 3.0 apply a genetic algorithm (GA) search strategy to optimize the docking geometry within the binding site. Since GAs are computationally rather expensive, their application in large-scale docking approaches is not advisable. Nevertheless, reliable scoring requires the generation of accurate, near-native binding poses. Both programs succeeded to redock Fusicoicin correctly. Accordingly, we decided to apply these docking programs for a fine-tune docking whereas the faster FLEXX was

used for the large-scale docking. GOLD was run with elaborate parameter settings, using a population size of 100, *mutation* rate of 5%, *crossover* rate of 95%, and  $10^5$  evaluation cycles. AUTODOCK was calibrated with a grid spacing of 0.25 Å, population size of 100, *mutation* rate of 2%, *crossover* rate of 80%, and  $5 \cdot 10^5$  energy evaluations. We generated 30 docking geometries with each docking program. All docking solutions were processed in the same manner following the above-described protocol and stored in SCREENINGDB.

### VISUAL DRUGSCORE

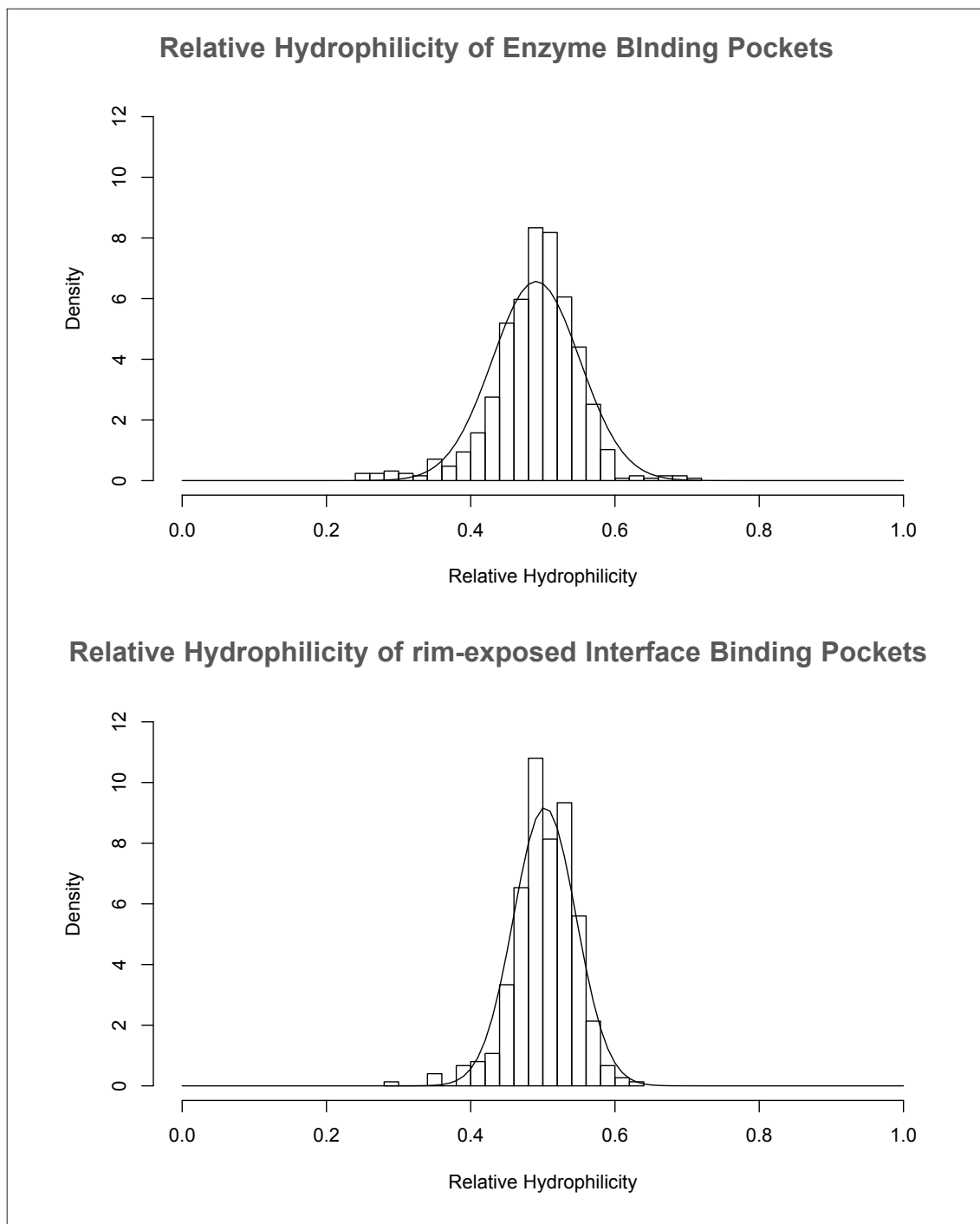
The final visual inspection was assisted by the newly developed graphical evaluation tool *visual* DrugScore. The total score computed by DrugScore is the sum over the individual contributions considering all contacts between a ligand and the surrounding protein. Nevertheless, this total score can be easily decomposed into per-atom score contributions and displayed by assigning these contributions graphically to the contacting atoms. We integrated *visual* DrugScore into the PyMOL molecular visualization system (DeLano *et al.*, 2002). The per-atom score contributions are translated into scaled spheres, that are mapped onto coordinates of the contacting ligand and protein atoms. Favorably interacting atoms are indicated by blue spheres, whereas unfavorable interactions are represented in red. The visualization tool is very supportive for the individual inspection of the docking solutions and provides valuable insight into protein-ligand interactions with respect to ligand portions contacting the protein target.

## RESULTS AND DISCUSSION

The analysis of rim-exposed and enzyme pockets resulted in very similar property contributions with respect to each of the considered descriptors (Fig. 9-11). The relative hydrophilicity of both kinds of pockets show normal distribution with a mean at about 0.5. The relative hydrophilicity of five enzymes (Tab. 2), which have proven to possess well-*druggable* binding pockets, show values between 0.39 (Cyclooxygenase II) and

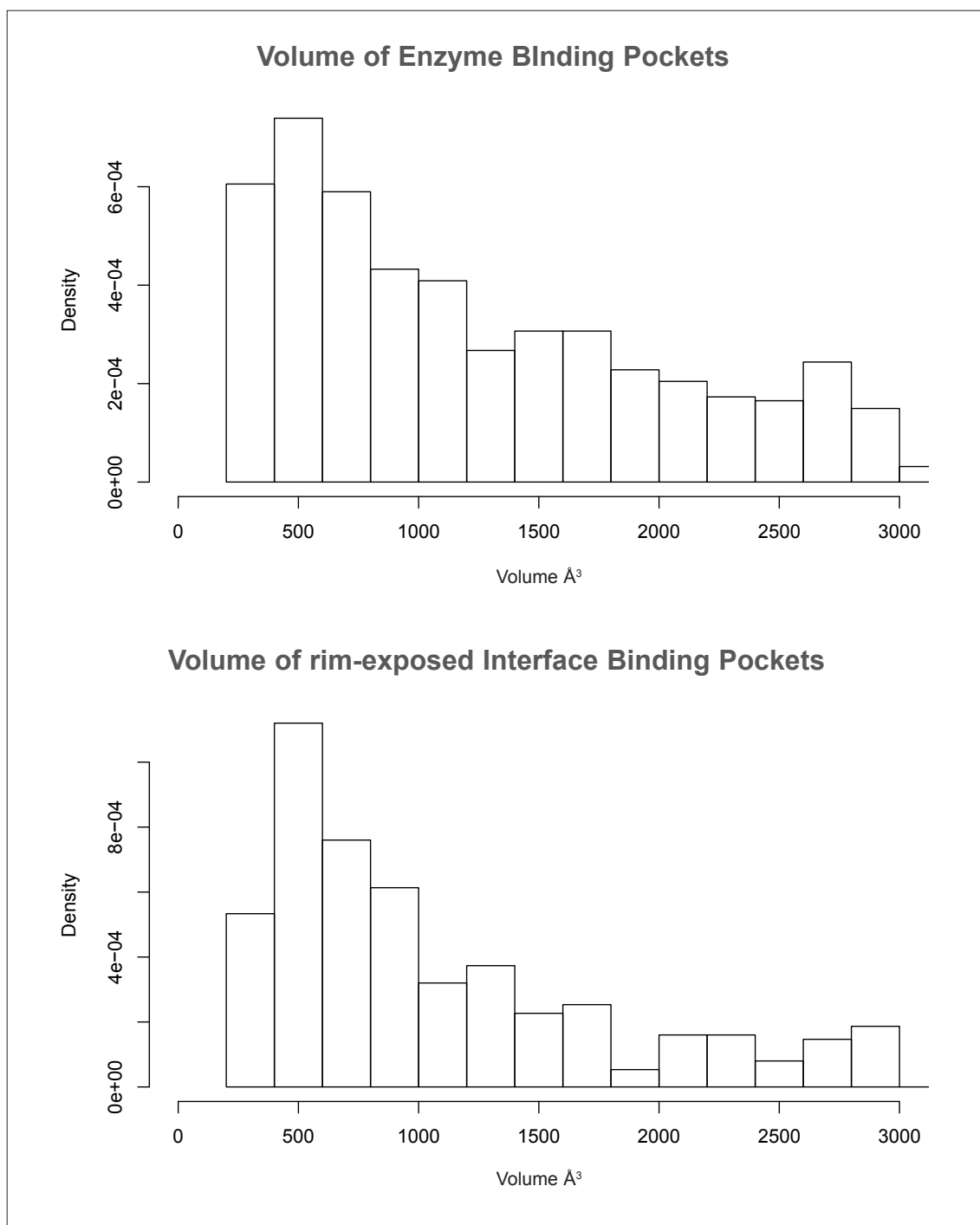
0.52 (Dihydrofolate Reductase). Although there might be other *druggable* pockets with a relative hydrophilicity beyond this range, a balanced distribution between hydrophobic and hydrophilic properties within the binding site appears to be essential for ligand binding in general. In contrast, the volume of the selected binding pockets show a widely spread distribution ranging from  $< 500 \text{ \AA}^3$  to  $> 3000 \text{ \AA}^3$ . Nevertheless, both the rim-exposed and the enzyme pockets show an accumulated incidence between 500 and  $800 \text{ \AA}^3$ . The five selected enzyme representatives range from  $658 \text{ \AA}^3$  (Carbonic Anhydrase II) to  $2037 \text{ \AA}^3$  (Dihydrofolate Reductase). However, even though a rather spacious binding pocket is given for Dihydrofolate Reductase, this enzyme can be inhibited by small molecules such as Methotrexate or Trimethoprim. The distribution of atom buriedness in the cavity also show similarity between both types of pockets with maximum values between 5 and 6. Acetylcholinesterase deviates from this (Tab. 2), as it exhibits most of the binding-site atoms with a buriedness value beyond 6. This is due to a rather long tube-shaped binding pocket (Fig. 12). As expected, the selected representatives of *druggable* enzymes show most of their binding-site atoms rather deeply buried. In summary, the similarity between rim-exposed cavities in protein-protein complexes and enzyme binding pockets is suggested as rather high in terms of the considered descriptors.

Subsequently, a mutual comparison between the pockets from the different sets using Cavbase has been accomplished. To our experience, Cavbase detects reasonable similarity if in both pockets about 8 to 20 pseudocenters are matched in common. Values beyond this range either match particularly pronounced similarity between homologous structures, or the low scores do not indicate any significant correspondence. The comparison between rim-exposed cavities and enzymes pockets resulted in a huge amount of similarity matches with scores in the outlined range (of example Fig. 13 and 14). Moreover, this indicates, that many rim-exposed pockets appear to expose similar properties compared to enzyme pockets. Nevertheless, this is only a rough indication for their putative *druggability*.

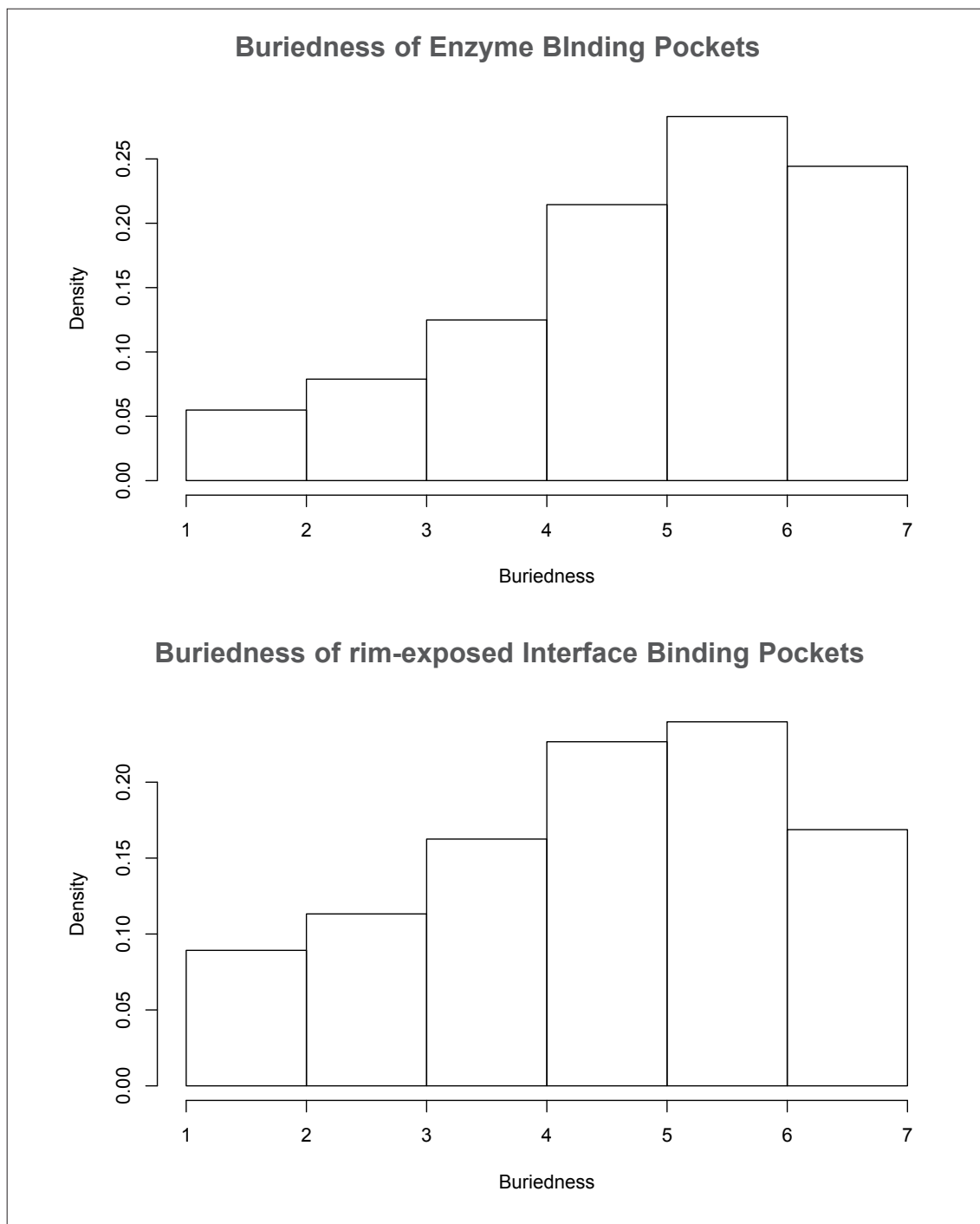


**FIGURE 9.** Both enzyme and rim-exposed cavities show normal distribution with a mean at about 0.5 for relative hydrophilicity.

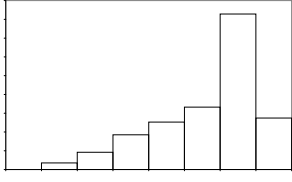
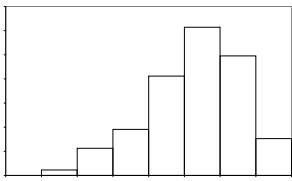
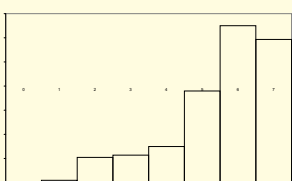
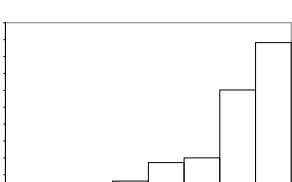
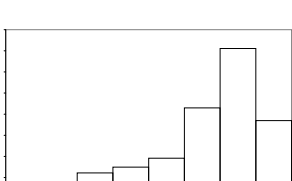
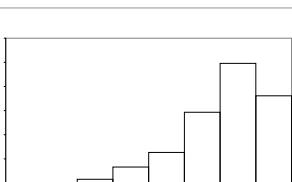




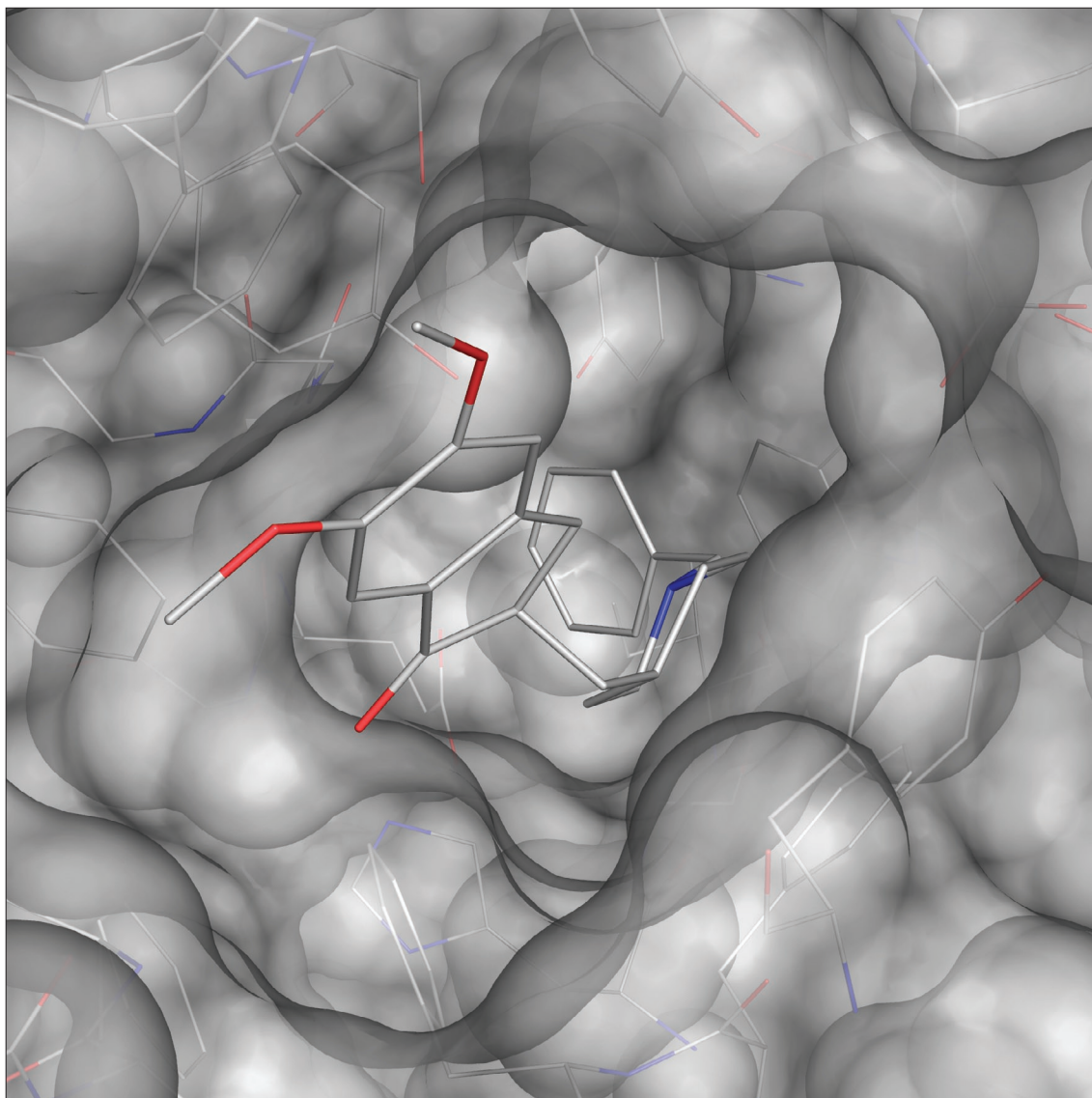
**FIGURE 10.** Both enzyme and rim-exposed cavities are spread over a wide range of volumes, but most cavities fall into a range between 500 and 1500 Å<sup>3</sup>.



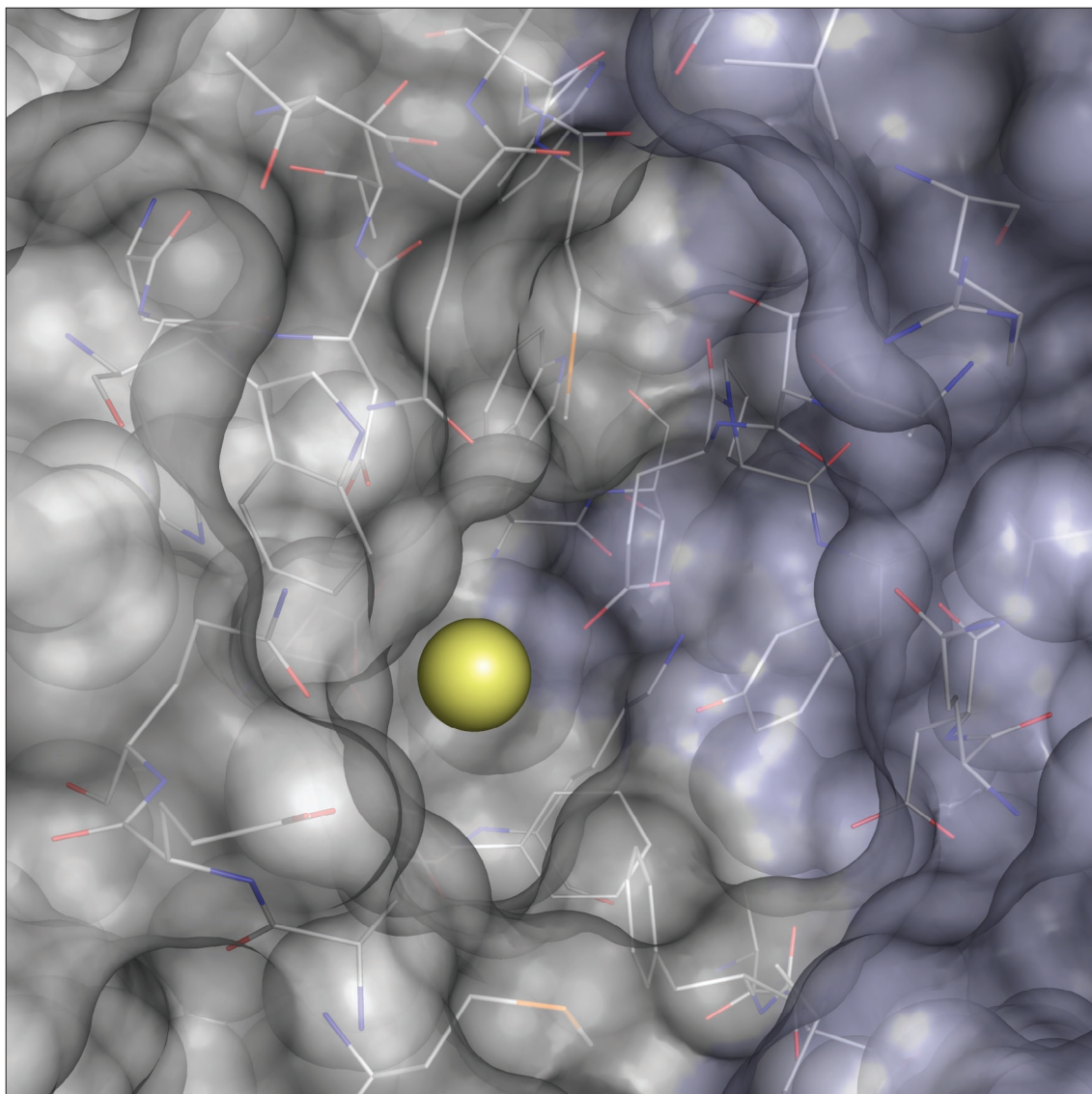
**FIGURE 11.** Both enzyme and rim-exposed cavities show similar distributions in the buriedness of their atoms.

PDB CODE	ENZYME	RELATIVE HYDROPHILICITY	VOLUME	BURIEDNESS
1cil	Carbonic Anhydrase II	0.41	658 Å <sup>3</sup>	
4cox	Cyclooxygenase II	0.39	742 Å <sup>3</sup>	
1o9f	H <sup>+</sup> -ATPase/14-3-3	0.52	874 Å <sup>3</sup>	
1eve	Acetylcholinesterase	0.47	1259 Å <sup>3</sup>	
1hxw	HIV Protease	0.50	1282 Å <sup>3</sup>	
1dds	Dihydrofolate Reductase	0.52	2037 Å <sup>3</sup>	

**TABLE 2.** Relative hydrophilicity, volume, and atom buriedness of known drug binding enzymes from the PDB. Additionally, Fusicoccin binding pocket of the H<sup>+</sup>-ATPase/14-3-3 complex (yellow). All pockets show a balanced ratio of hydrophilic and hydrophobic properties in the binding site. The pocket volume is spread over a wide range between 658 and 2037 Å<sup>3</sup>. The buriedness of the binding site atoms show similar distributions. Acetylcholinesterase deviates from this, as it exhibits most of the binding-site atoms with a buriedness value beyond 6.



**FIGURE 12.** Tube-shaped binding pocket of Acetylcholinesterase (PDB 1eve) in complex with the anti-Alzheimer drug Donepezil (ARICEPT®). The enzyme shows a pronounced number of very deeply buried atoms in its binding pocket, significantly higher than the average.



**FIGURE 13.** Rim-exposed cavity of the protein-protein complex between Barley  $\alpha$ -amylase (gray) with its endogenous protein inhibitor BASI (blue) and complexed calcium ion (yellow) (PDB 1ava). The relative hydrophilicity of 0.52 and the volume of 897  $\text{\AA}^3$  suggest this pocket putatively as *druggable*.

Considering some crude statistics based on the analyzed data set of transient complexes, more than 380 rim-exposed cavities can be detected, which suggests on average the presence of more than one putative cavity per complex that could be possibly addressed by a small molecule ligand. To assess the feasibility of such a strategy in each individual case would require rigorous estimates about the *druggability* of these sites as they define the properties to be met by putative interface stabilizers. In a recent evaluation Hajduk *et al.* (Hajduk *et al.*, 2005) suggest a set of discriminative descriptors assess the *druggability* of typical, well-buried binding sites. This study explicitly point out the multifactorial nature of the rules determining *druggability*. Apparently, it appears difficult to estimate whether the same type of descriptors are applicable to classify rim-exposed cavities or crevices at protein-protein interfaces. Nevertheless, their frequent occurrence stimulated us to embark into an elaborate virtual screening campaign using the H<sup>+</sup>-ATPase/14-3-3 interface, addressed by Fusicocin, as a case example.

## VIRTUAL SCREENING FOR STABILIZERS OF THE H<sup>+</sup>-ATPase/14-3-3 INTERACTION

Screening for alternative stabilizers of the H<sup>+</sup>-ATPase/14-3-3 interaction is challenging, since only Fusicocin is presently known to stabilize this protein-protein interaction. For our search we applied established structure-based drug design tools, since the H<sup>+</sup>-ATPase/14-3-3 complex exhibits a deep binding pocket, however it exposes only (1) a few amino acids qualified to form a directional hydrogen bond to a putative ligand and (2) this ligand has to address both, the H<sup>+</sup>-ATPase and 14-3-3 protein, with a sufficiently large hydrophobic surface portion. Therefore, we decided to screen a large data sample by docking to avoid early discard of potential hits and novel chemistry by applying too stringent filters based on preconceived pharmacophore information. Instead, we defined a variety of pharmacophore constraints, suggested by the target, that had to be matched by the ligands retrieved in the initial screening run. Accordingly, about 5·10<sup>5</sup> docking poses had to be filtered according to these setups.

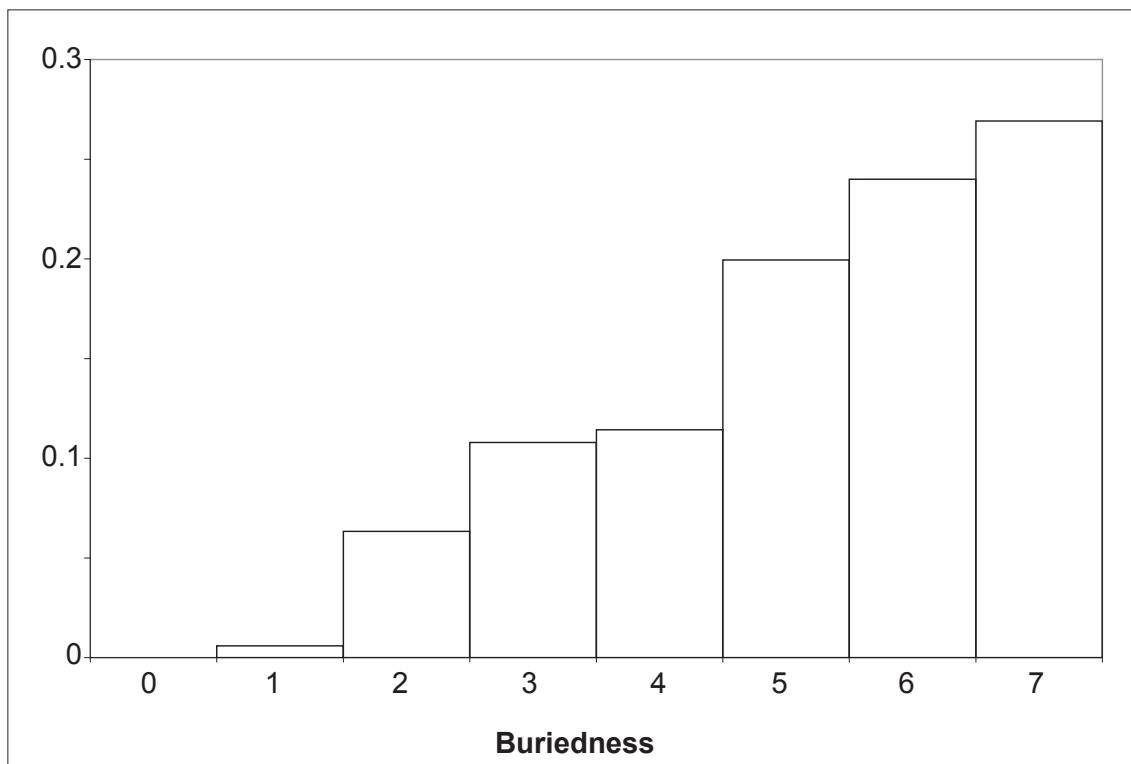


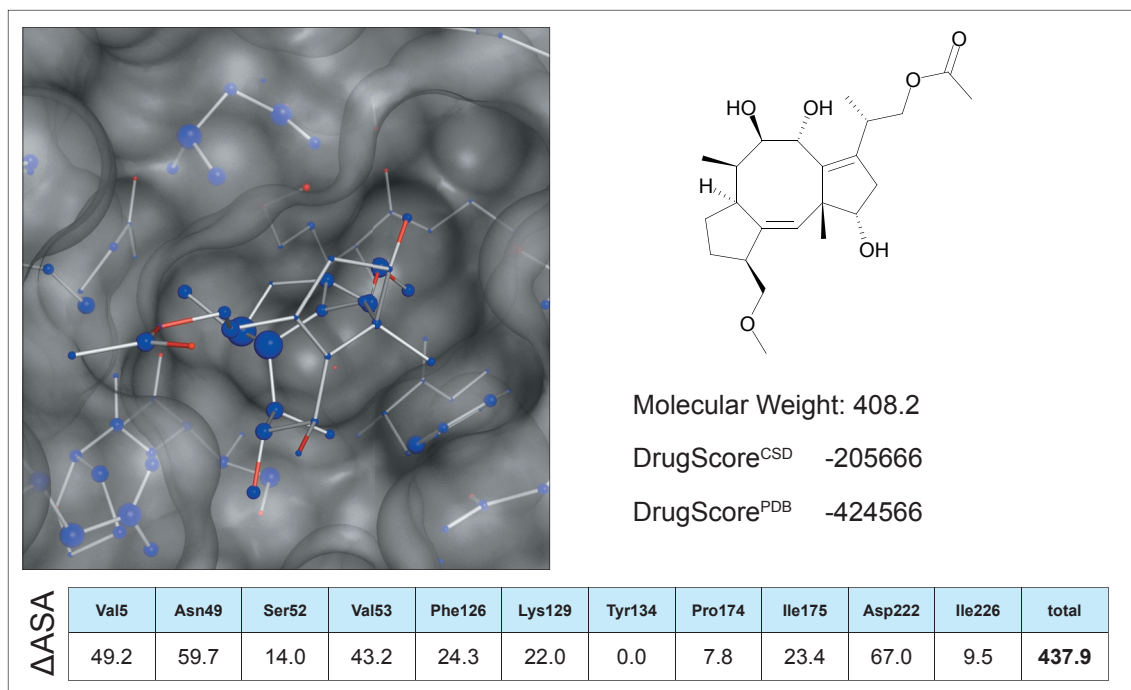
FIGURE 14. Distribution of buriedness of binding pocket atoms in PDB 1ava.

Our first screening was tailored to achieve filling of the water sub-pocket along with satisfying a reasonably large hydrophobic contact to the side chain of Val5. Furthermore, formation of a hydrogen bond to Lys129 and Asp222 was requested. All docking poses were filtered according to the pharmacophoric constraints given in Fig. 7. In total, we selected eight molecules for *in vitro* testing (Fig. 16-23), of which five exhibit a related scaffold. All selected hits suggested reasonable docking geometries, well satisfying the desired interaction pattern and achieving convincing scores. Similar binding modes were suggested by the different docking programs. Furthermore, we applied *visual* DrugScore in combination with DrugScore<sup>CSD</sup> potentials to elucidate the score contributions of each docking pose. All selected molecules show favorable interaction patterns (Fig. 16-23). Four of the selected compounds exhibit a terminal amide group (332884-29-4, Fig. 18; 604741-06-2, Fig. 19; 606116-94-3, Fig. 20 and 587012-99-5, Fig. 23) accommodate the previously water-filled sub-pocket (blue sphere in the pharmacophore hypothesis,

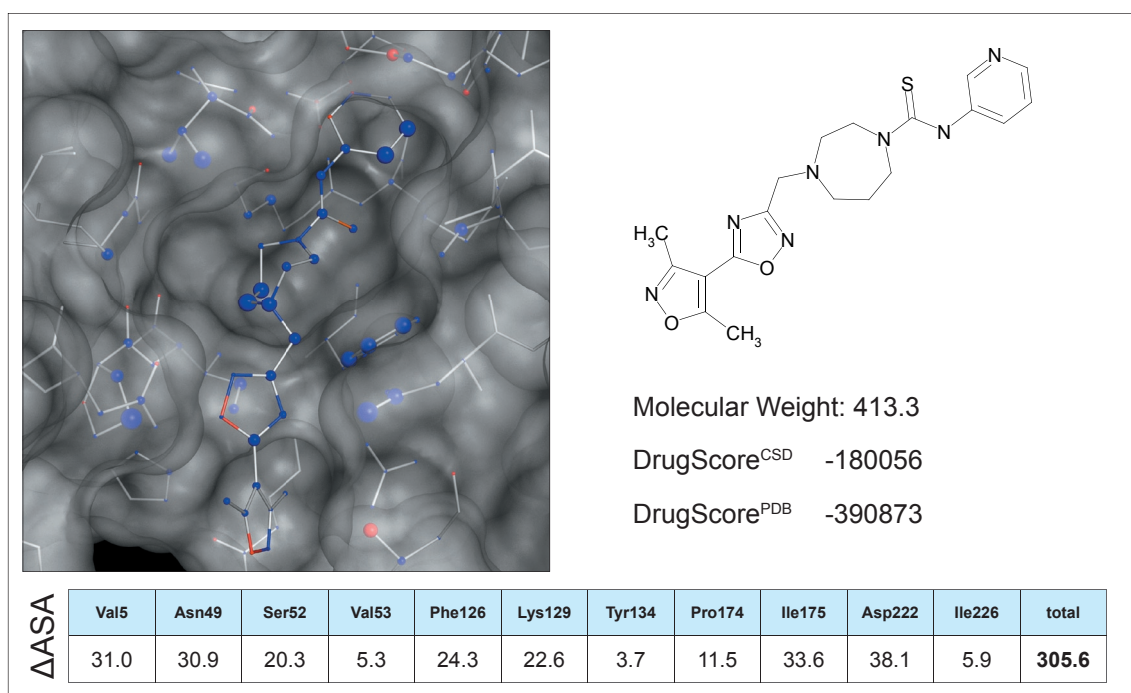
Fig. 7). Our design was tailored to address the hydroxy group of Tyr134 in this sub-pocket. Particularly, the ambivalent amide group can adopt an adequate conformation to form a hydrogen bond. The other selected compounds show either a carboxylic (606117-05-9, Fig. 21), or hydrazidic group (385376-34-1, Fig. 22), or they exhibit ring systems with appropriate acceptor functionalities (1518-10922, Fig. 16 and 405279-55-2, Fig. 17) to interact with Tyr134. All moieties selected to address the water sub-pocket, display attractive per-atom score contributions indicated by *visual* DrugScore. The acceptor functionality of the ether oxygen in Fusicoccin, which forms a hydrogen bond to Lys129, is mimicked by either a heterocyclic ether oxygen (405279-55-2, Fig. 17; 332884-29-4, Fig. 18; 604741-06-2, Fig. 19; 606116-94-3, Fig. 20 and 606117-05-9, Fig. 21), a heterocyclic nitrogen (1518-10922, Fig. 16 and 587012-99-5, Fig. 23), or the likely negatively charged tetrazol (385376-34-1, Fig. 22). The hydrophobic contact to Val5 is formed by either carbocyclic ring portions or other uncharged sidechain decoration, which experience high scoring with DrugScore. However, for all selected compounds the contacts to Val5 appear suboptimal, compared to the available contacting  $\Delta$ ASA experienced by Fusicoccin (Tab. 3).

**D**espite of this promising match with the requested pharmacophore and convincing local scoring, none of the selected compounds showed a permanent stabilizing effect in either the fluorescence or BIAcore assay, respectively, whereas Fusicoccin exhibits intensive signals (data not shown). Obviously, the tested ligands do not bind sufficiently strong to the H<sup>+</sup>-ATPase and supposedly do not capture strong enough interactions with Val5. Prime focus of our initial screen to address the water sub-pocket, which appeared as well-suited for small *druglike* molecules. However, to achieve a net contribution to binding affinity, the total inventory of energetic contributions required to replace the waters has to be considered. Depending on the enthalpic loss and entropic gain of this replacement, it can be detrimental to binding. Possibly, a stronger adhesion of the ligand to the C-terminal hydrophobic side chain (Val5) of the H<sup>+</sup>-ATPase seems to be indispensable for the stabilization particularly since this interaction seems to be exhaustively exploited by the





**FIGURE 15.** Fusicoccin aglycon: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 16.** 1518-10922: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

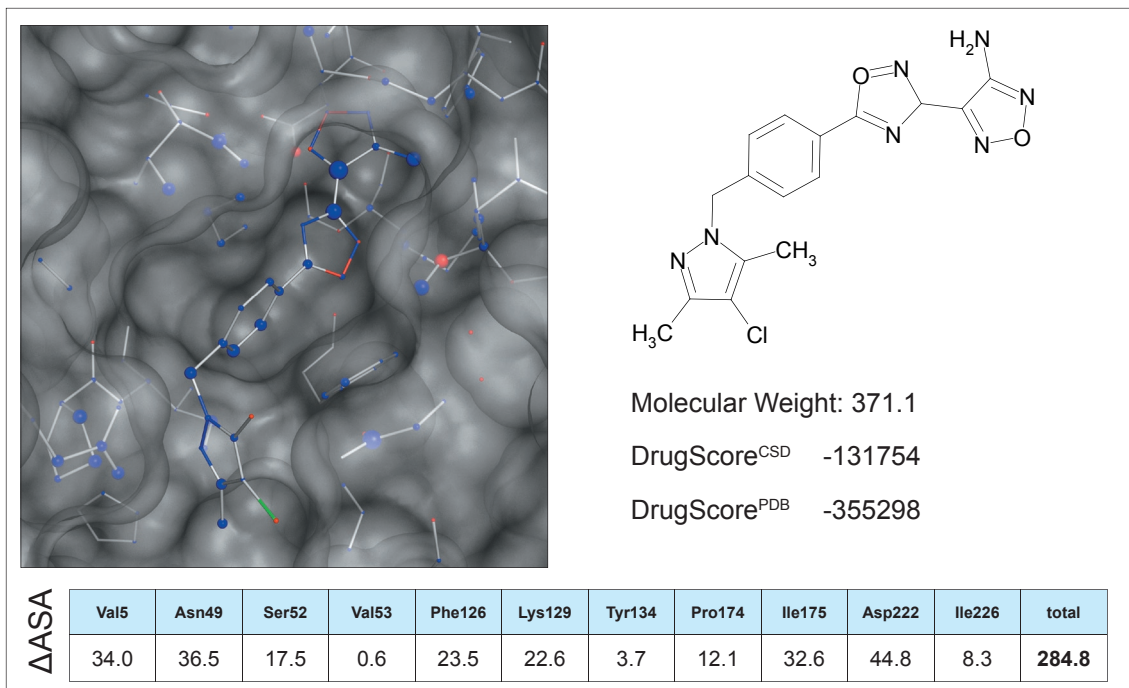


FIGURE 17. 405279-55-2: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

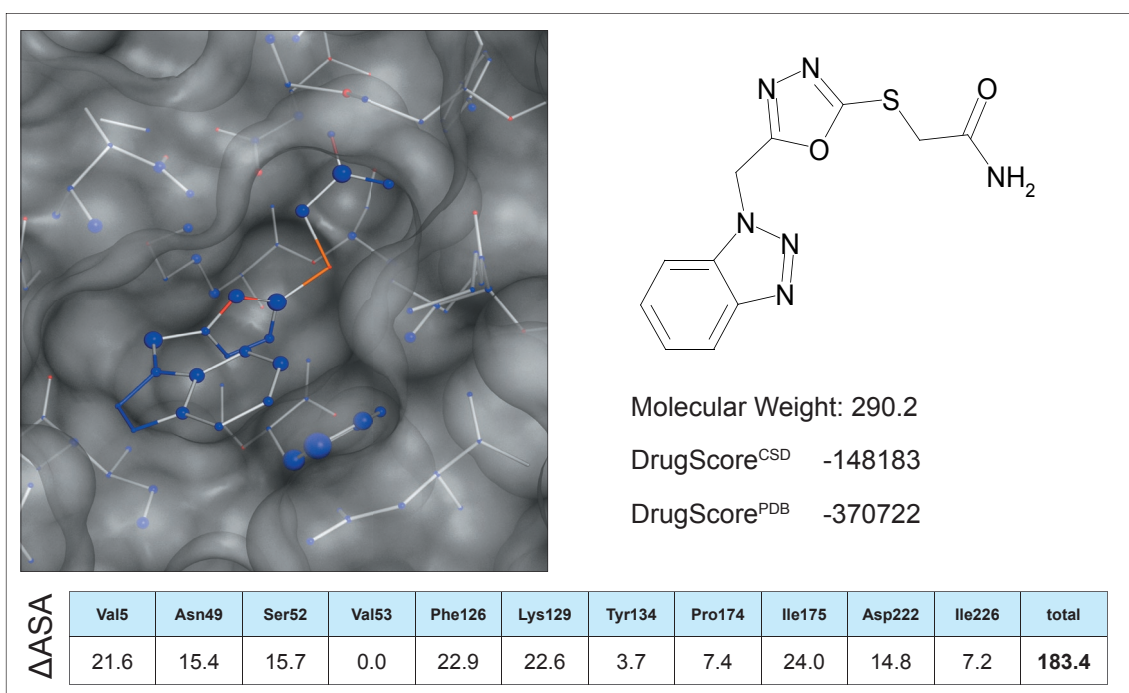
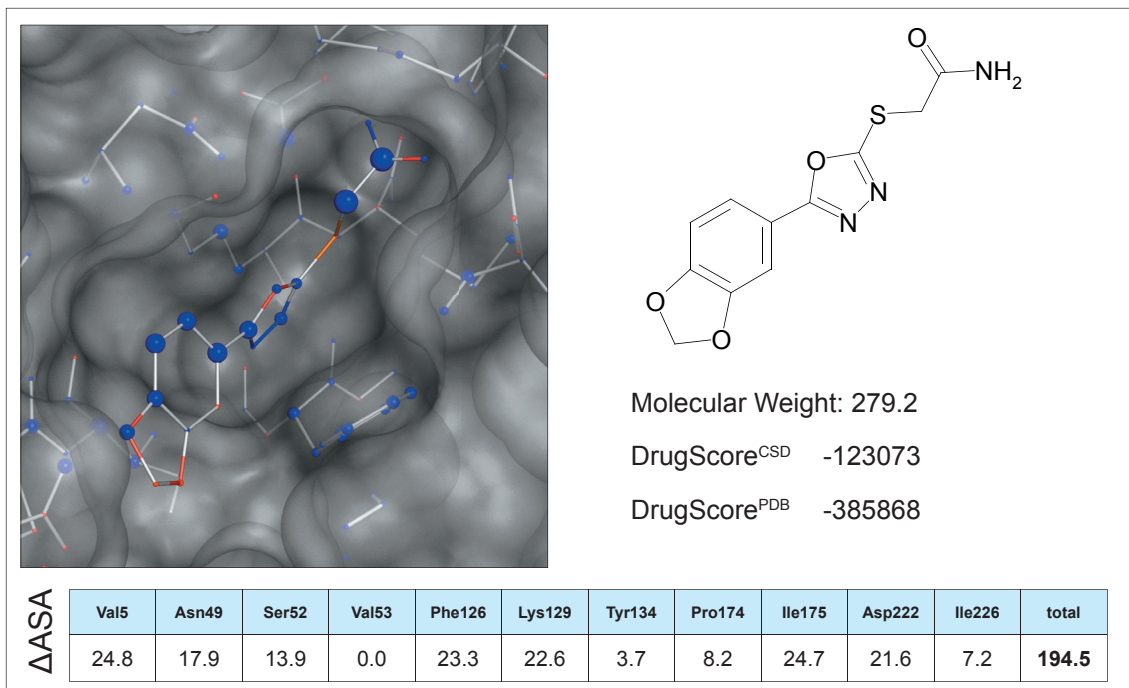
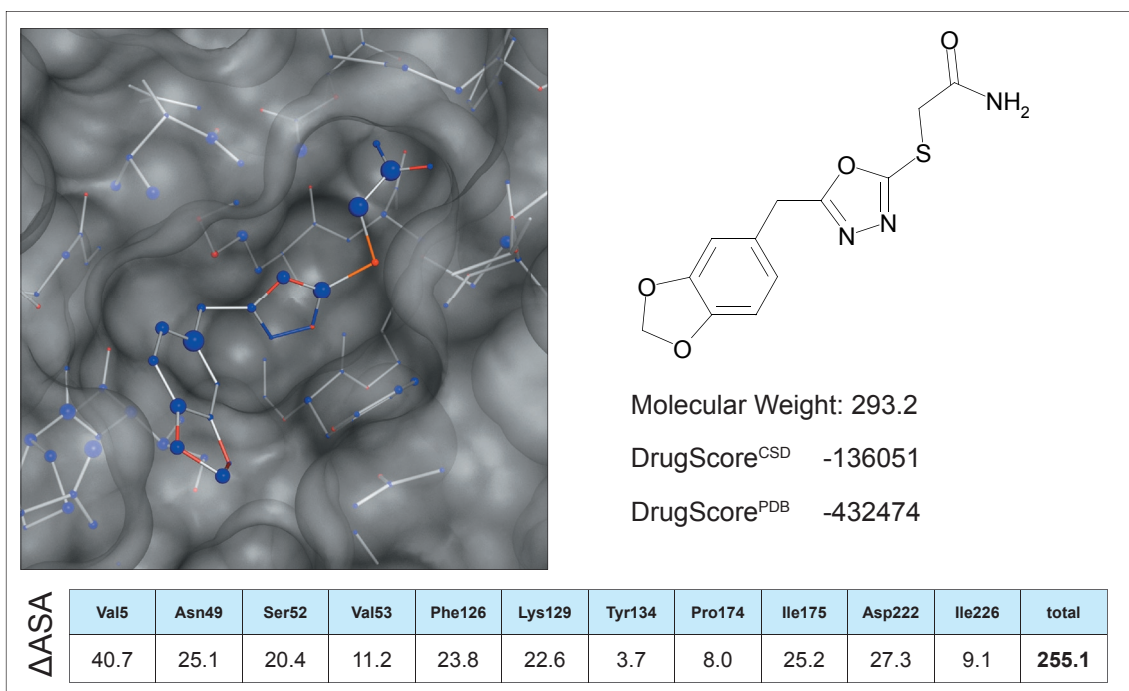


FIGURE 18. 332884-29-4: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 19.** 604741-06-2: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 20.** 606116-94-3: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

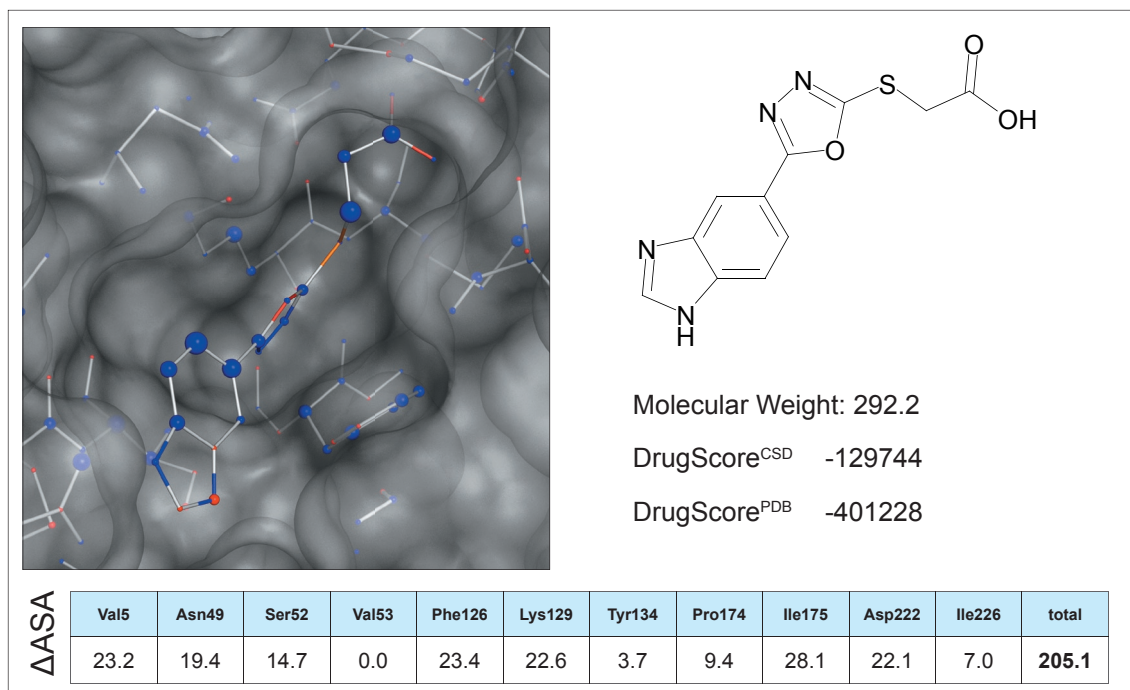


FIGURE 21. 606117-05-9: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

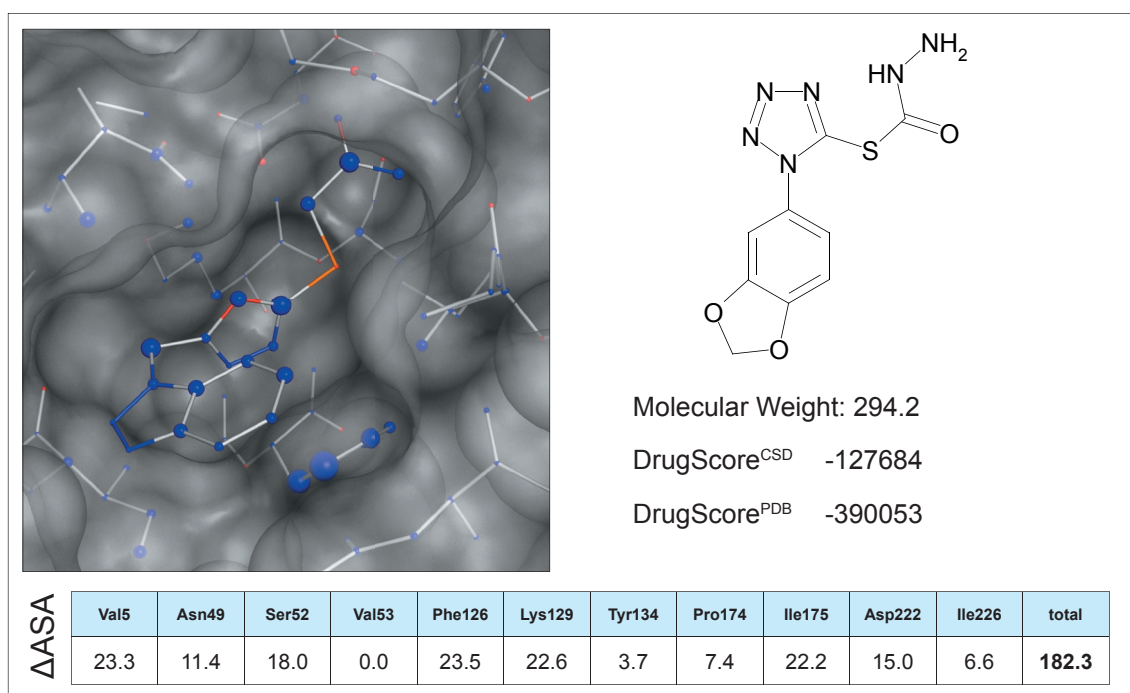
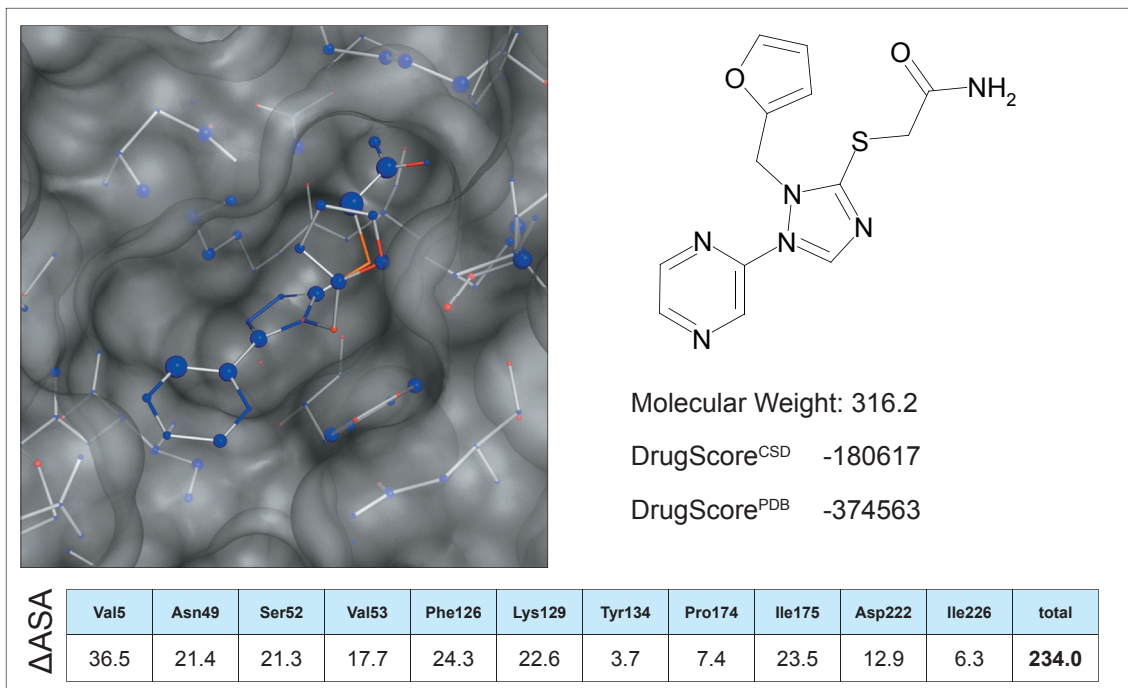
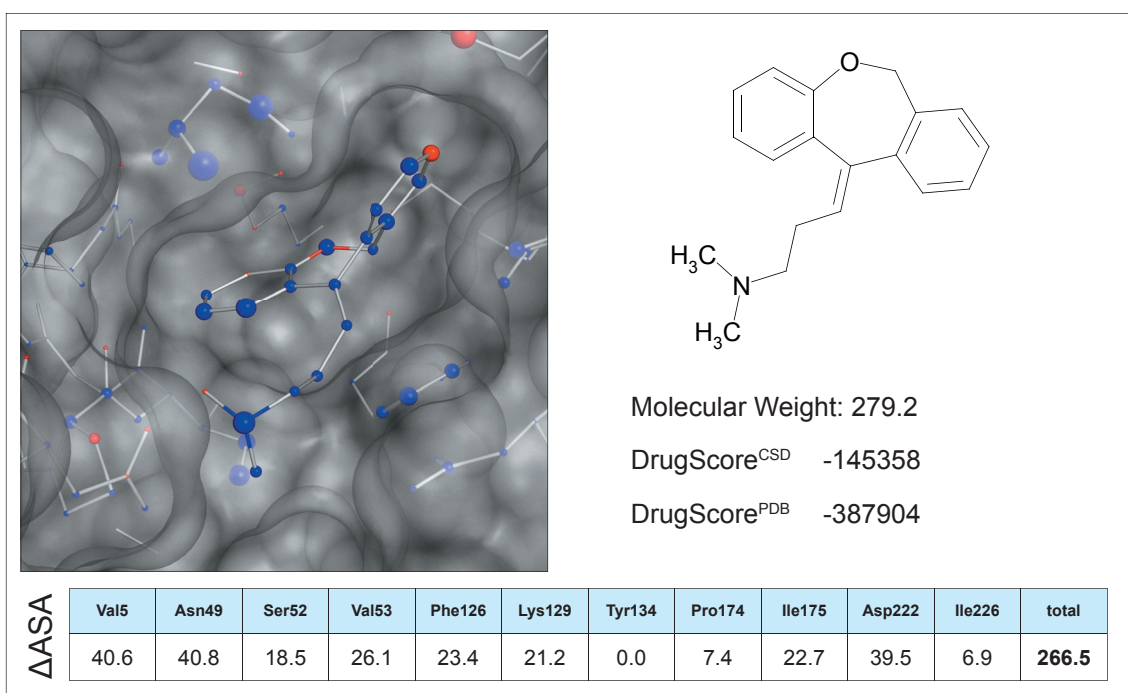


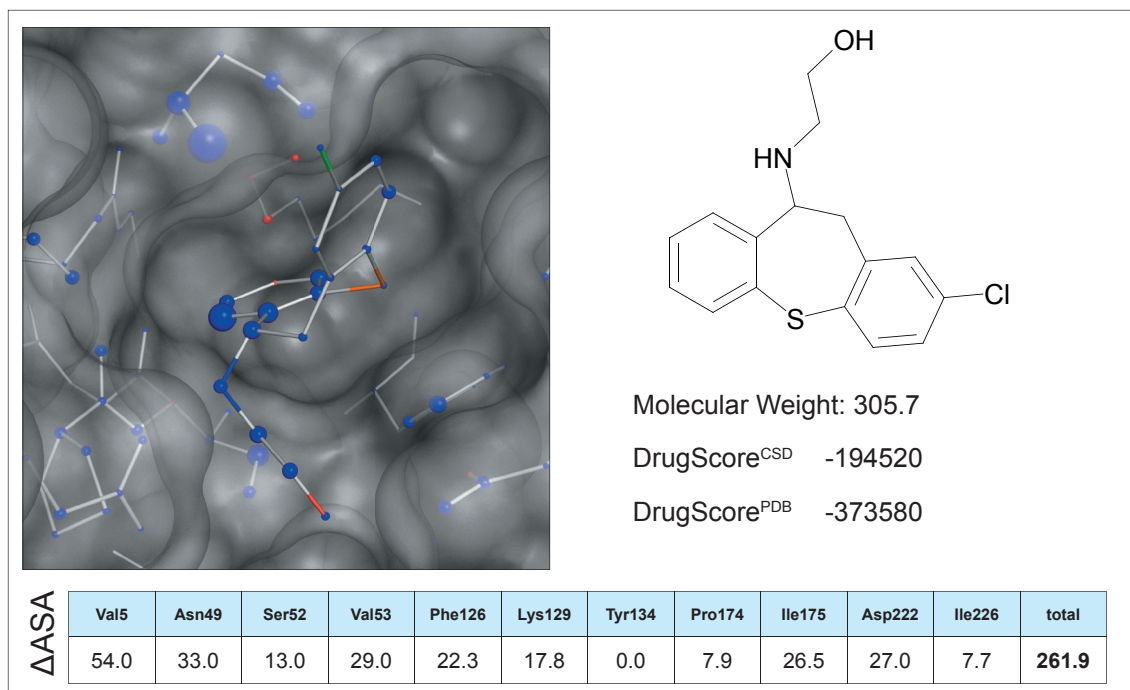
FIGURE 22. 385376-34-1: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



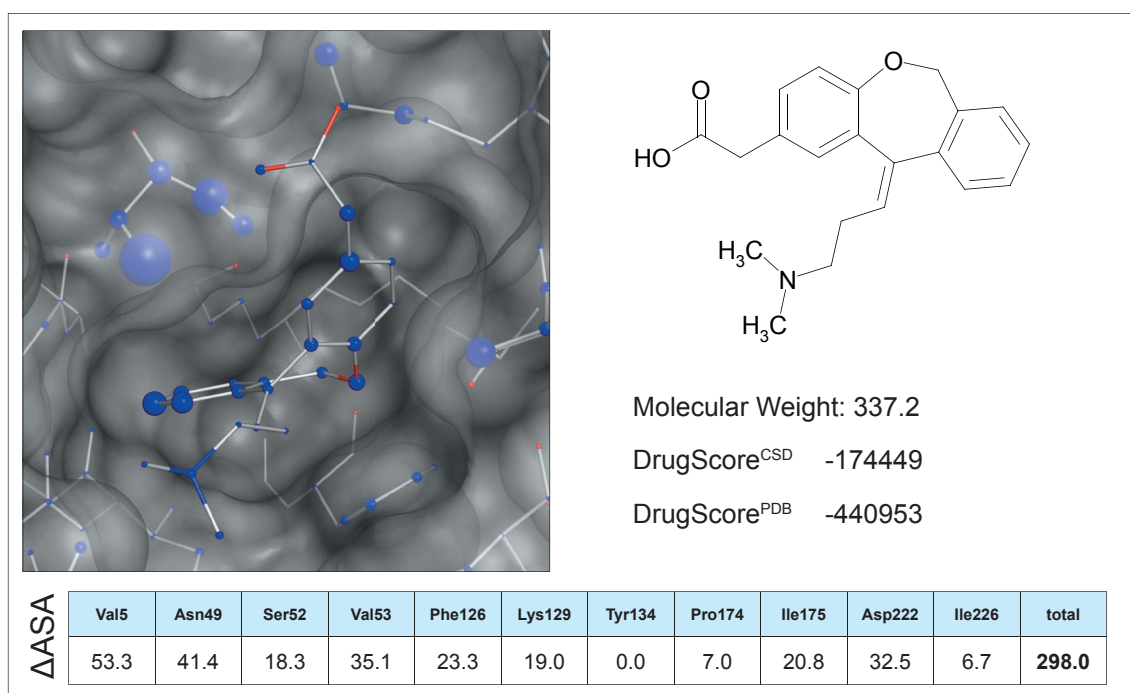
**FIGURE 23.** 587012-99-5: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 24.** 1229-29-4: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 25.** 134073-67-9: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



**FIGURE 26.** 140462-76-6: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

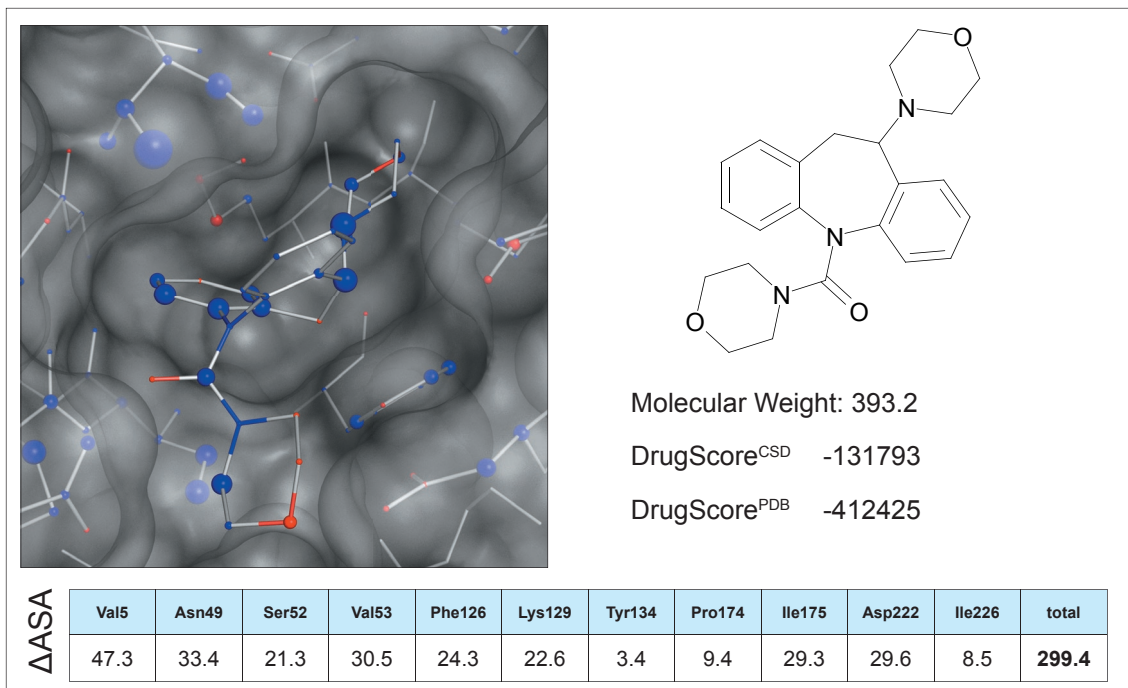


FIGURE 27. 335393-81-2: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

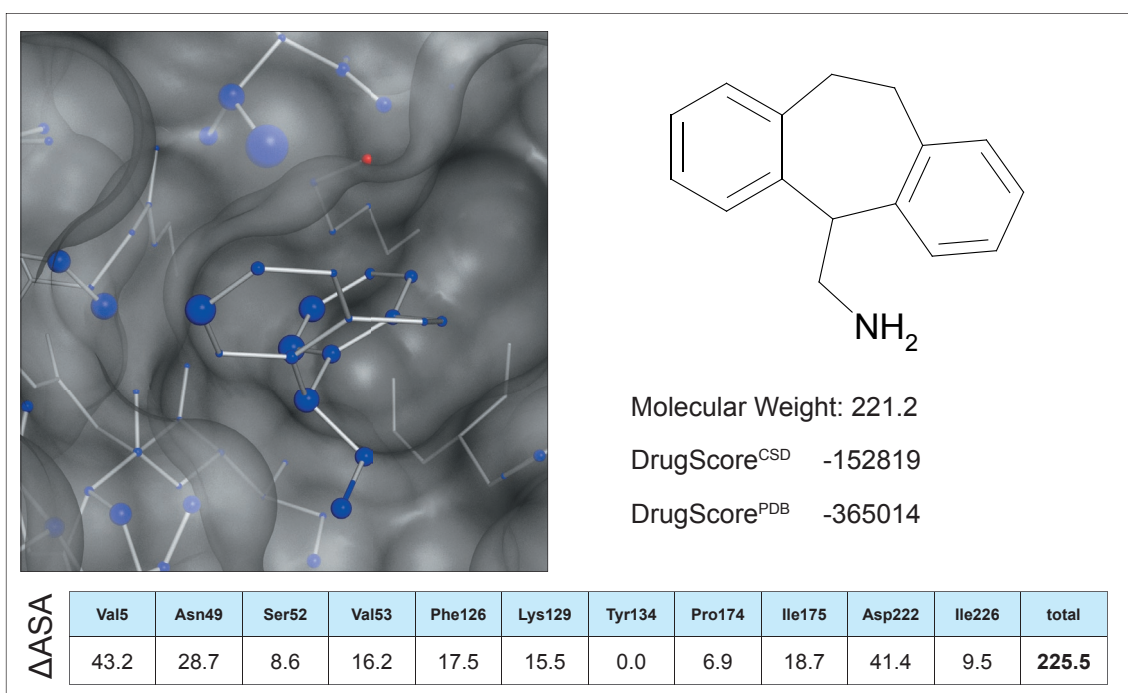


FIGURE 28. 380635-96-1: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

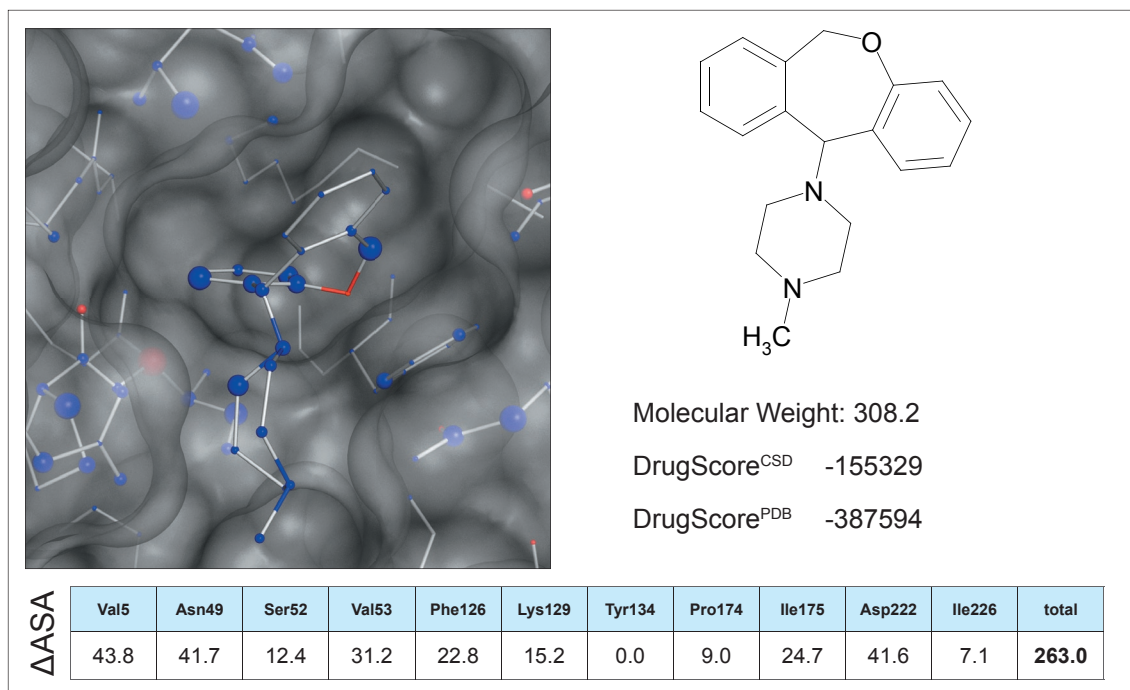


FIGURE 29. 47192-66-5: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

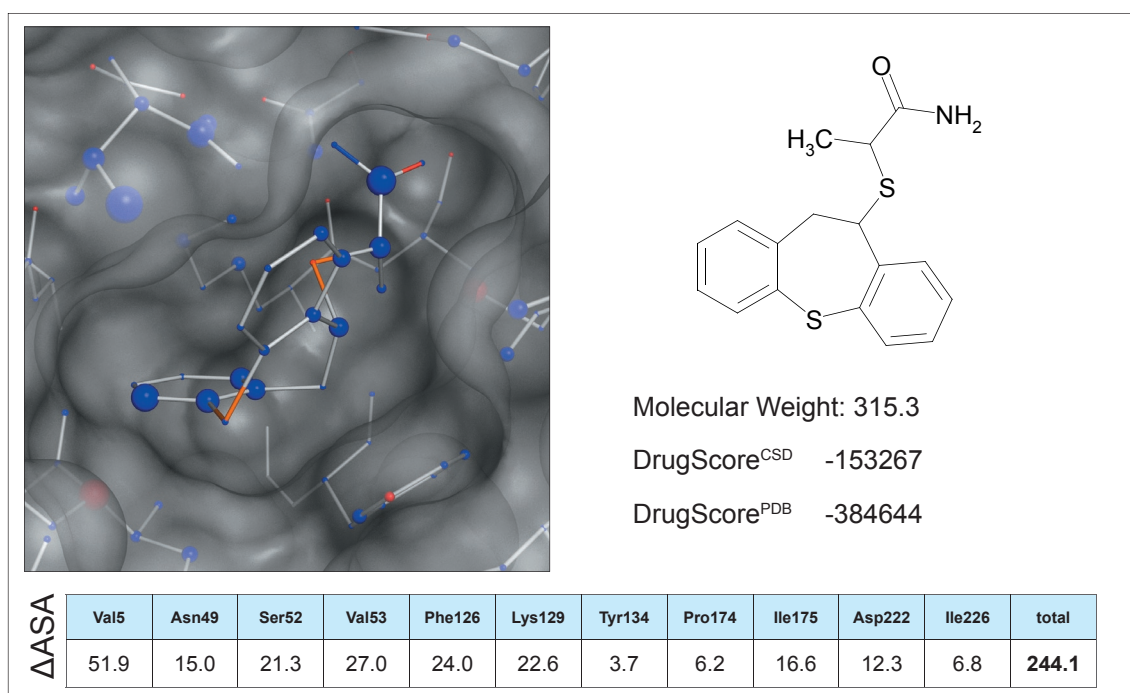


FIGURE 30. 720674-11-3: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.



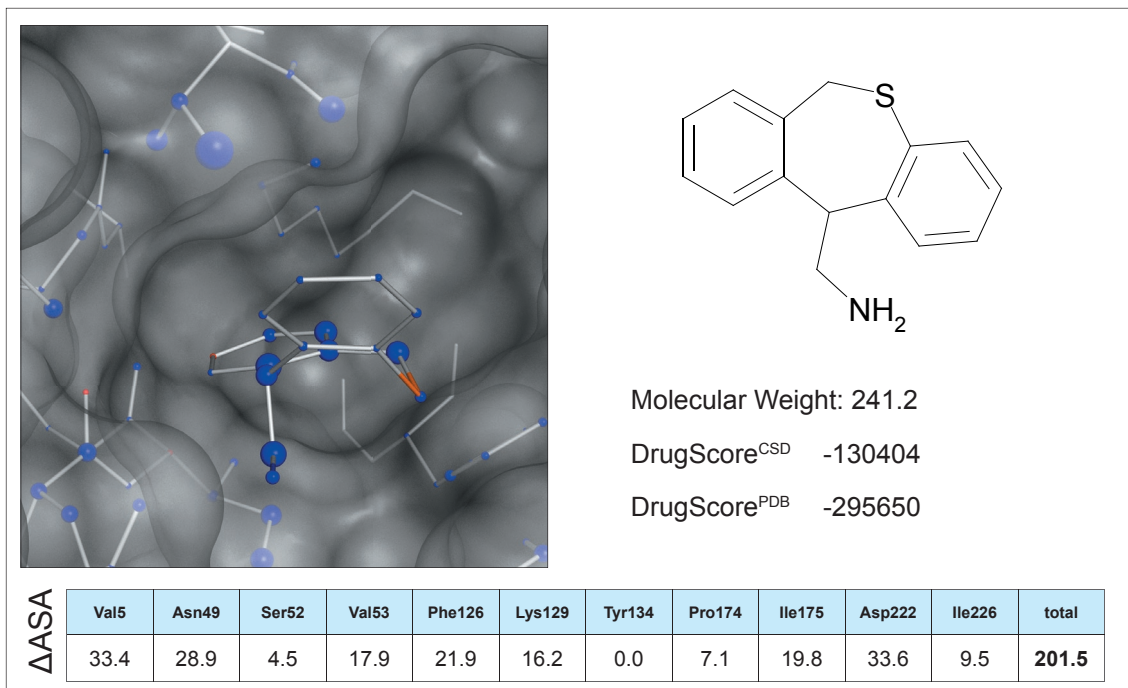


FIGURE 31. 82394-01-2: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

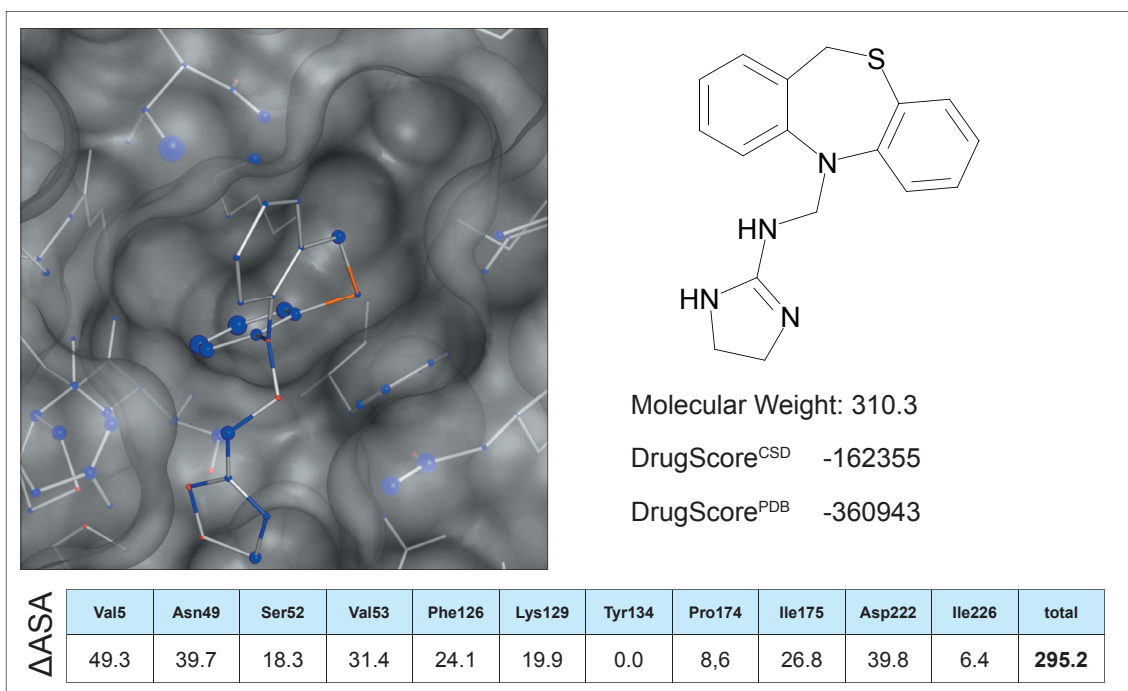


FIGURE 32. 844882-80-0: *visual* DrugScore representation with DrugScore<sup>CSD</sup> potentials, DrugScore<sup>CSD</sup> and DrugScore<sup>PDB</sup> scores and the loss on solvent accessible surface ( $\Delta$ ASA, in  $\text{\AA}^2$ ) upon complexation for selected residues in the binding site.

unique architecture of Fusicoccin. Fusicoccin shows attractive per-atom DrugScore<sup>CSD</sup> contributions throughout almost the entire molecule (Fig. 15). In particular, the ring scaffold and the ether attached to ring A experience high local scores. The burial of solvent accessible surface ( $\Delta$ ASA) assigned to FC and removed upon binding is 437.9 Å<sup>2</sup> (49.2 Å<sup>2</sup> are contributed to Val5). It appears as if spatial arrangement of the three fused rings evolved by Fusicoccin is ideally suited to fit the target. The curved ring skeleton wraps around the side chain of Val5. Therefore, prospective-looking molecular skeletons satisfying these requirements should obey a similar ring system curvature. The fused ring systems found in tricyclic antidepressants such as Doxepin exhibit a bent butterfly-type arrangement and may represent a promising architecture to wrap around the targeted Val5 sidechain. Following this idea, we filtered SCREENINGDB and SCIFINDER SCHOLAR for candidates with tricyclic ring moieties and docked them using FLEXX and GOLD. We discovered nine compounds, properly wrapping around Val5 and addressing at least one of the pharmacophore constraint defined in Fig. 7. The best scored docking solutions, with respect to DrugScore<sup>CSD</sup>, are shown together with *visual* DrugScore in Fig. 24-32. In this campaign, only FLEXX generated convincing docking solutions, since GOLD places the three rings inside the water sub-pocket and these poses are as less favorable.

In contrast to the first screening, only two of the selected compounds occupy the water sub-pocket (335393-81-2, Fig. 27 and 720674-11-3, Fig. 30), but all expose their fused-ring system in a way to wrap around Val5. As the large blue spheres in *visual* DrugScore indicate, well-scored interactions to Val5 are experienced, in particular for the compounds 140462-72-6 (Fig. 26) and 335393-81-2 (Fig. 27). A donor functionality to address Asp222 is present as amino group (1229-29-4, Fig. 24; 134073-67-9, Fig. 25; 140462-76-6, Fig. 26; 380635-96-1, Fig. 28 and 82394-01-2, Fig. 31) or guanidino group (844882-80-0, Fig. 32). Doxepin (1229-29-4, Fig. 24) may even mimic Fusicoccin's ether as it exposes an acceptor functionality via its ether oxygen in the seven-membered ring. Compared to Fusicoccin, the compounds selected in the second screen exhibit lower DrugScore<sup>PDB</sup> or DrugScore<sup>CSD</sup> scores. This is due to their lower molecular weight and

ID	Val5	Asn49	Ser52	Val53	Phe126	Lys129	Tyr134	Pro174	Ile175	Asp222	Ile226	total
<b>Fusicoccin</b>	49.2	59.7	14.0	43.2	24.3	22.0	0.0	7.8	23.4	67.0	9.5	<b>437.9</b>
1518-10922	31.0	30.9	20.3	5.3	24.3	22.6	3.7	11.5	33.6	38.1	5.9	305.6
405279-55-2	34.0	36.5	17.5	0.6	23.5	22.6	3.7	12.1	32.6	44.8	8.3	284.8
332884-29-4	21.6	15.4	15.7	0.0	22.9	22.6	3.7	7.4	24.0	14.8	7.2	183.4
604741-06-2	24.8	17.9	13.9	0.0	23.3	22.6	3.7	8.2	24.7	21.6	7.2	194.5
606116-94-3	40.7	25.1	20.4	11.2	23.8	22.6	3.7	8.0	25.2	27.3	9.1	255.1
606117-05-9	23.2	19.4	14.7	0.0	23.4	22.6	3.7	9.4	28.1	22.1	7.0	205.1
385376-34-1	23.3	11.4	18.0	0.0	23.5	22.6	3.7	7.4	22.2	15.0	6.6	182.3
587012-99-5	36.5	21.4	21.3	17.7	24.3	22.6	3.7	7.4	23.5	12.9	6.3	234.0
1229-29-4	40.6	40.8	18.5	26.1	23.4	21.2	0.0	7.4	22.7	39.5	6.9	266.5
134073-67-9	54.0	33.0	13.0	29.0	22.3	17.8	0.0	7.9	26.5	27.0	7.7	261.9
140462-76-6	53.3	41.4	18.3	35.1	23.3	19.0	0.0	7.0	20.8	32.5	6.7	298.0
335393-81-2	47.3	33.4	21.3	30.5	24.3	22.6	3.4	9.4	29.3	29.6	8.5	299.4
380635-96-1	43.2	28.7	8.6	16.2	17.5	15.5	0.0	6.9	18.7	41.4	9.5	225.5
47192-66-5	43.8	41.7	12.4	31.2	22.8	15.2	0.0	9.0	24.7	41.6	7.1	263.0
720674-11-3	51.9	15.0	21.3	27.0	24.0	22.6	3.7	6.2	16.6	12.3	6.8	244.1
82394-01-2	33.4	28.9	4.5	17.9	21.9	16.2	0.0	7.1	19.8	33.6	9.5	201.5
844882-80-0	49.3	39.7	18.3	31.4	24.1	19.9	0.0	8.6	26.8	39.8	6.4	295.2

**TABLE 3.** The loss on solvent accessible surface area ( $\Delta$ ASA) upon binding of the generated docking geometries for the most important residues within the Fusicoccin binding pocket of. The values of Fusicoccin are with yellow background.

therefore the fewer contacts formed with the protein-protein complex. Nevertheless, the ring-systems wrapping around Val5 are consistently high-scored (cf. *visual* DrugScore). Considering solely the DrugScore<sup>CSD</sup> contribution of the three ring-system, e.g. in Doxepin, a better score is achieved (-122270) compared to the fused ring skeleton of Fusicoccin (-107606). The additional decorations of the Fusicoccin aglycon improves its total score nearly a factor of two. In contrast, the sidechains at Doxepin increase scoring by approximately 10%. Consistently, the  $\Delta$ ASA burial achieve in the docking poses with

respect to the pivotal Val5 contact (35-55 Å<sup>2</sup>) fall into the same range as of Fusicoccin (49.2 Å<sup>2</sup>), although the total ΔASA burial for these ligands (200-300 Å<sup>2</sup>) is significantly smaller than for Fusicoccin (437.9 Å<sup>2</sup>, Tab. 3). In conclusion, the suggested fused ring-systems might represent a promising scaffolds for lead stabilizing the H<sup>+</sup>-ATPase/14-3-3 complex.

## SUMMARY AND CONCLUSIONS

**M**odulating protein-protein interactions by small molecular compounds is a challenging task. The approach to stabilize such interactions appears very promising, considering that several molecules, including known drugs, follow this mode of action. The crystal structure of the protein-protein complex of H<sup>+</sup>-ATPase and 14-3-3 in complex with bound phytotoxin Fusicoccin has been used for a feasibility study to discover novel stabilizers using structure-based virtual screening. Nearly two millions of commercially available compounds have been screened. A versatile combination of several standard screening protocols together with tools to analyze, cluster and classify vast of screening results has been implemented into SCREENINGDB. Furthermore, *visual* DrugScore has been developed as powerful graphic tool to visualize per atom contributions to the protein-ligand interactions captured in a generated docking solution.

**T**he first virtual screening campaign has been focused to fill a previously hydrated water sub-pocket, together with a reasonably large hydrophobic contact to the crucial side chain of Val5. Additionally, the formation of a hydrogen bond to Lys129 and/or Asp222 were requested. A set of eight compounds was selected for a subsequent *in vitro* testing based on their pharmacophore matching and achieved DrugScore rankings. However, none of the *in vitro* tested compounds showed detectable stabilization of the H<sup>+</sup>-ATPase/14-3-3 complex. This might be due to the fact, that the crucial interaction of Fusicoccin, which wraps around the side chain of Val5, is not sufficiently well mimicked by the selected screening hits. Accordingly, the second screening was focused on the

retrieval of compounds forming a more intimate contact to Val5. The discovered hits of this screen consistently show a fused system of three rings, which wraps convincingly well around Val5 comparable to Fusicoccin. Considering the well-scored ring skeletons, e.g. in Doxepin, such scaffolds potentially represent a promising architecture for leads to stabilize the H<sup>+</sup>-ATPase/14-3-3 contact.

In light of the results of this feasibility study, the general concept to use small molecules to stabilize protein-protein interactions appears quite tempting. First of all, there is a considerable number of small molecules known (including some drugs), which actually stabilize protein-protein interactions. Furthermore, the thermodynamic prerequisites for the binding of a small molecule to a rim-exposed interface cavity appear better achievable and more likely favorable than the binding to a large, rather featureless protein-protein interface. This concept is supported by the fact that a sizable number of protein-protein complexes exhibit one or more rim-exposed pockets spanning the interface. The latter finding results from the analysis of a dataset of 198 transient protein-protein recognition complexes, which exhibit such cavities. Even though, recent studies underline the difficulty estimate the putative *druggability* of a given pocket simply based on structural descriptors, we applied the Cavbase concept compare rim-exposed cavities observed across protein-protein interfaces with binding pockets of enzymes which have been successfully subjected to drug development programs. Interestingly, both types of cavities achieve pronounced similarity in terms of the distribution of hydrophilicity, the volume, and buriedness of pocket-forming atoms. Furthermore, we used the Cavbase approach to retrieve similar cavities from both sets of pockets. The frequently found correspondence substantiates the hypothesis of *druggable* cavities at the margin of protein-protein complexes. We believe, the idea of targeting rim-exposed pockets spanning across protein-protein interfaces may change the paradigm to search and design potential modulators of protein-protein complex formation.

**REFERENCES**

- Agrawal, R. K.; Frank, J. (1999) Structural Studies Of The Translational Apparatu *Curr Opin Struct Biol*, **9**, 215-221.
- Arkin, M. (2005) Protein-Protein Interactions And Cancer: Small Molecules Going In For The Kill *Curr Opin Chem Biol* , **9**, 317-324.
- Baell, J. B.; Huang, D. C. S. (2002) Prospects For Targeting The Bcl-2 Family Of Proteins To Develop Novel Cytotoxic Drugs *Biochem Pharmacol*, **64**, 851-863.
- Ballio, A.; Chain, E. B.; De Leo, P.; Erlanger, B. F.; Mauri, M.; Tonolo, A. (1964) Fusicocin: A New Wilting Toxin Produced By *Fusicoccum Amygdali* Del. *Nature*, **203**, 297.
- Ballio, A.; Bottalico, A.; Framondino, M.; Graniti, A.; Randazzo, G. (1971) Fusicocin: Structure-Phytotoxicity Relationship *Phytopathol Mediterr*, **10**, 26-32.
- Ballio, A.; Casinovi, C. G.; Framondino, M.; Marino, G.; Nota, G.; Santurbano, B. (1979) A New Cerebroside From *Fusicoccum Amygdali* Del *Biochim Biophys Acta*, **573**, 51-60.
- Ballio, A.; De Michelis, M. I.; Lado, P.; Randazzo, G. (1981) Fusicocin Structure-Activity Relationships: Stimulation Of Growth By *Cell* Enlargement And Promotion Of Seed Germination *Physiol Plant*, **52**, 471-475.
- Berg, T. (2003) Modulation Of Protein-Protein Interactions With Small Organic Molecules *Angew Chem Int Ed Engl*, **42**, 2462-2481.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

- Brenk, R.; Naerum, L.; Gradler, U.; Gerber, H.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. (2003) Virtual Screening For Submicromolar Leads Of tRNA-Guanine Transglycosylase Based On A New Unexpected Binding Mode Detected By Crystal Structure Analysis *J Med Chem*, **46**, 1133-1143.
- Chardin, P.; McCormick, F. (1999) Brefeldin A: The Advantage Of Being Uncompetitive *Cell*, **97**, 153-155.
- Choi, J.; Chen, J.; Schreiber, S. L.; Clardy, J. (1996) Structure Of The Fkbp12-Rapamycin Complex Interacting With The Binding Domain Of Human Frap *Science*, **273**, 239-242.
- Clark, M.; Cramer, R. D.; Van Opdenbosch, N. (1989) Validation Of The General Purpose TRIPOS 5.2 Force Field *J. Comp. Chem*, **10**, 982-1012.
- DeLano, W. (2002) The PyMOL User's Manual.
- Evers, A.; Klebe, G. (2004) Successful Virtual Screening For A Submicromolar Antagonist Of The Neurokinin-1 Receptor Based On A Ligand-Supported Homology Model *J Med Chem*, **47**, 5381-5392.
- Fasan, R.; Dias, R. L. A.; Moehle, K.; Zerbe, O.; Vrijbloed, J. W.; Obrecht, D.; Robinson, J. A. (2004) Using A Beta-Hairpin To Mimic An Alpha-Helix: Cyclic Peptidomimetic Inhibitors Of The P53-Hdm2 Protein-Protein Interaction *Angew Chem Int Ed Engl*, **43**, 2109-2112.
- Fuglsang, A. T.; Visconti, S.; Drumm, K.; Jahn, T.; Stensballe, A.; Mattei, B.; Jensen, O. N.; Aducci, P.; Palmgren, M. G. (1999) Binding Of 14-3-3 Protein To The Plasma Membrane H<sup>(+)</sup>-ATPase Aha2 Involves The Three C-Terminal Residues Tyr(946)-Thr-Val And Requires Phosphorylation Of Thr(947) *J Biol Chem*, **274**, 36774-36780.

- Gasteiger, J.; Rudolph, C.; Sadowski, J. (1990) Automatic Generation Of 3D-Atomic Coordinates For Organic Molecules *Tetrahedron Comp. Method*, **3**, 537-547.
- Gohlke, H.; Hendlich, M.; Klebe, G. (2000) Knowledge-Based Scoring Function To Predict Protein-Ligand Interactions *J Mol Biol*, **295**, 337-356.
- Green, N. J.; Xiang, J.; Chen, J.; Chen, L.; Davies, A. M.; Erbe, D.; Tam, S.; Tobin, J. F. (2003) Structure-Activity Studies Of A Series Of Dipyrzolo[3,4-B:3',4'-D]Pyridin-3-Ones Binding To The Immune Regulatory Protein B7.1 *Bioorg Med Chem*, **11**, 2991-3013.
- Greenfield, E. A.; Nguyen, K. A.; Kuchroo, V. K. (1998) Cd28/B7 Costimulation: A Review *Crit Rev Immunol*, **18**, 389-418.
- Griffith, J. P.; Kim, J. L.; Kim, E. E.; Sintchak, M. D.; Thomson, J. A.; Fitzgibbon, M. J.; Fleming, M. A.; Caron, P. R.; Hsiao, K.; Navia, M. A. (1995) X-Ray Structure Of Calcineurin Inhibited By The Immunophilin-Immunosuppressant Fkbp12-Fk506 Complex *Cell*, **82**, 507-522.
- Grueneberg, S.; Stubbs, M. T.; Klebe, G. (2002) Successful Virtual Screening For Novel Inhibitors Of Human Carbonic Anhydrase: Strategy And Experimental Confirmation *J Med Chem*, **45**, 3588-3602.
- Guenther, J.; Bergner, A.; Hendlich, M.; Klebe, G. (2003) Utilising Structural Knowledge In Drug Design Strategies: Applications Using Relibase *J Mol Biol*, **326**, 621-636.
- Hajduk, P. J.; Huth, J. R.; Fesik, S. W. (2005) *Druggability* Indices For Protein Targets Derived From Nmr-Based Screening Data *J Med Chem*, **48**, 2518-2525.
- Hendlich, M.; Rippmann, F.; Barnickel, G. (1997) Ligsite: Automatic And Efficient Detection Of Potential Small Molecule-Binding Sites In Proteins *J Mol Graph Model*, **15**, 359-363.



- Hendlich, M.; Bergner, A.; Guenther, J.; Klebe, G. (2003) Relibase: Design And Development Of A Database For Comprehensive Analysis Of Protein-Ligand Interactions *J Mol Biol*, **326**, 607-620.
- Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. (1992) Selection Of Representative Protein Data Sets *Protein Sci*, **1**, 409-417.
- Hobohm, U.; Sander, C. (1994) Enlarged Representative Set Of Protein Structures *Protein Sci*, **3**, 522-524.
- Hubbard, S.; Thornton, J. (1993) NACCESS: Computer Program, Department Of Biochemistry And Molecular Biology, University College London.
- Jennewein, S.; Croteau, R. (2001) Taxol: Biosynthesis, Molecular Genetics, And Biotechnological Applications *Appl Microbiol Biotechnol*, **57**, 13-19.
- Kelly, B. (WWW) FROWNS Chemoinformatic System, <http://frowns.sourceforge.net>.
- Kuhlbrandt, W.; Zeelen, J.; Dietrich, J. (2002) Structure, Mechanism, And Regulation Of The Neurospora Plasma Membrane H<sup>+</sup>-ATPase *Science*, **297**, 1692-1696.
- Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. (1997) PDBsum: A Web-Based Database Of Summaries And Analyses Of All Pdb Structures *Trends Biochem Sci*, **22**, 488-490.
- Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. (2005) PDBsum More: New Summaries And Analyses Of The Known 3D Structures Of Proteins And Nucleic Acids *Nucleic Acids Res*, **33**, D266-268.

- Maudoux, O.; Batoko, H.; Oecking, C.; Gevaert, K.; Vandekerckhove, J.; Boutry, M.; Morsomme, P. (2000) A Plant Plasma Membrane H<sup>+</sup>-ATPase Expressed In Yeast Is Activated By Phosphorylation At Its Penultimate Residue And Binding Of 14-3-3 Regulatory Proteins In The Absence Of Fusicoccin *J Biol Chem*, **275**, 17762-17770.
- Marre, E. (1979) Fusicoccin: A Tool In Plant Physiology *Annu. Rev. Plant Physiol*, **30**, 273-312.
- Martin, F.; Toniatti, C.; Salvati, A. L.; Venturini, S.; Ciliberto, G.; Cortese, R.; Sollazzo, M. (1994) The Affinity-Selection Of A Minibody Polypeptide Inhibitor Of Human Interleukin-6 *EMBO J*, **13**, 5303-5309.
- Martin, Y. C. (1992) 3D Database Searching In Drug Design *J Med Chem*, **35**, 2145-2154.
- Mintseris, J.; Weng, Z. (2003) Atomic Contact Vectors In Protein-Protein Recognition *Proteins*, **53**, 629-639.
- Moss, N.; Beaulieu, P.; Duceppe, J.-S.; Ferland, J.-M.; Garneau, M.; Gauthier, J.; Ghiron, A.; Goulet, S.; Guse, I.; Jaramillo, J.; Llinas-Brunet, M.; Malenfant, E.; Plante, R.; Poirier, M.; Soucy, F.; Wernic, D.; Yoakim, V.; Déziel, R. (1996) Peptidomimetic Inhibitors Of Herpes Simplex Virus Ribonucleotide Reductase With Improved *In Vivo* Antiviral Activity *J Med Chem*, **39**, 4173-4180
- Morsomme, P.; Boutry, M. (2000) The Plant Plasma Membrane H<sup>(+)</sup>-ATPase: Structure, Function And Regulation *Biochim Biophys Acta*, **1465**, 1-16.
- Murray, C. W.; Auton, T. R.; Eldridge, M. D. (1998) Empirical Scoring Functions. Ii. The Testing Of An Empirical Scoring Function For The Prediction Of Ligand-Receptor Binding Affinities And The Use Of Bayesian Regression To Improve The Quality Of The Model *J Comput Aided Mol Des*, **12**, 503-519.

- Nguyen, J. T.; Wells, J. A. (2003) Direct Activation Of The Apoptosis Machinery As A Mechanism To Target Cancer *Cells Proc Natl Acad Sci U S A*, **100**, 7533-7538.
- Nord, K.; Gunneriusson, E.; Ringdahl, J.; Stahl, S.; Uhlen, M.; Nygren, P. A. (1997) Binding Proteins Selected From Combinatorial Libraries Of An Alpha-Helical Bacterial Receptor Domain *Nat Biotechnol*, **15**, 772-777.
- Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. (2001) Is There A Difference Between Leads And Drugs? A Historical Perspective *J Chem Inf Comput Sci*, **41**, 1308-1315.
- Peyroche, A.; Antonny, B.; Robineau, S.; Acker, J.; Cherfils, J.; Jackson, C. L. (1999) Brefeldin A Acts To Stabilize An Abortive Arf-Gdp-Sec7 Domain Protein Complex: Involvement Of Specific Residues Of The Sec7 Domain *Mol Cell*, **3**, 275-285.
- Qureshi, S. A.; Kim, R. M.; Konteatis, Z.; Biazzo, D. E.; Motamedi, H.; Rodrigues, R.; Boice, J. A.; Calaycay, J. R.; Bednarek, M. A.; Griffin, P.; Gao, Y. D.; Chapman, K.; Mark, D. F. (1999) Mimicry Of Erythropoietin By A Nonpeptide Molecule *Proc Natl Acad Sci U S A* **96**, 12156-12161.
- Ramakrishnan, V. (2002) Ribosome Structure And The Mechanism Of Translation *Cell*, **108**, 557-572.
- Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. (1996) A Fast Flexible Docking Method Using An Incremental Construction Algorithm *J Mol Biol*, **261**, 470-489.
- Rarey, M.; Dixon, J. S. (1998) Feature Trees: A New Molecular Similarity Measure Based On Tree Matching *J Comput Aided Mol Des*, **12**, 471-490.
- Rarey, M.; Stahl, M. (2001) Similarity Searching In Large Combinatorial Chemistry Spaces *J Comput Aided Mol Des*, **15**, 497-520.

- Reed, J. C. (1997) Double Identity For Proteins Of The Bcl-2 Family *Nature*, **387**, 773-776.
- Schmitt, S.; Kuhn, D.; Klebe, G. (2002) A New Method To Detect Related Function Among Proteins Independent Of Sequence And Fold Homology *J Mol Biol*, **323**, 387-406.
- Schneider, D. J.; Feigon, J.; Hostomsky, Z.; Gold, L. (1995) High-Affinity Ssdna Inhibitors Of The Reverse Transcriptase Of Type 1 Human Immunodeficiency Virus *Biochemistry*, **34**, 9599-9610.
- Sehnke, P. C.; Rosenquist, M.; Alsterfjord, M.; DeLille, J.; Sommarin, M.; Larsson, C.; Ferl, R. J. (2002) Evolution And Isoform Specificity Of Plant 14-3-3 Proteins *Plant Mol Biol*, **50**, 1011-1018.
- Sotriffer, C.; Sanschagrín, P.; Klebe, G. *in preparation*.
- Svennelid, F.; Olsson, A.; Piotrowski, M.; Rosenquist, M.; Ottman, C.; Larsson, C.; Oecking, C.; Sommarin, M. (1999) Phosphorylation Of Thr-948 At The C Terminus Of The Plasma Membrane H<sup>(+)</sup>-ATPase Creates A Binding Site For The Regulatory 14-3-3 Protein *Plant Cell*, **11**, 2379-2391.
- Tesmer, J. J.; Sunahara, R. K.; Gilman, A. G.; Sprang, S. R. (1997) Crystal Structure Of The Catalytic Domains Of Adenylyl Cyclase In A Complex With G $\alpha$ . *Gtpgammas Science*, **278**, 1907-1916.
- Tian, S. S.; Lamb, P.; King, A. G.; Miller, S. G.; Kessler, L.; Luengo, J. I.; Averill, L.; Johnson, R. K.; Gleason, J. G.; Pelus, L. M.; Dillon, S. B.; Rosen, J. (1998) A Small, Nonpeptidyl Mimic Of Granulocyte-Colony-Stimulating Factor, *Science*, **281**, 257-259.
- Tzivion, G.; Avruch, J. (2002) 14-3-3 Proteins: Active Cofactors In Cellular Regulation By Serine/Threonine Phosphorylation *J Biol Chem*, **277**, 3061-3064.

- Vassilev, L. T.; Vu, B. T.; Graves, B.; Carvajal, D.; Podlaski, F.; Filipovic, Z.; Kong, N.; Kammlott, U.; Lukacs, C.; Klein, C.; Fotouhi, N.; Liu, E. A. (2004) *In Vivo* Activation Of The P53 Pathway By Small-Molecule Antagonists Of Mdm2 *Science*, **303**, 844-848.
- Veleg, F. G. H.; Gohlke, H.; Klebe, G. (2005) DrugScore<sup>CSD</sup> Knowledge-Based Scoring Function Derived From Small Molecule Crystal Data With Superior Recognition Rate Of Near-Native Ligand Poses And Better Affinity Prediction *J Med Chem*, **48**, 6296-6303.
- Verdonk, M. L.; Cole, J. C.; Taylor, R. (1999) Superstar: A Knowledge-Based Approach For Identifying Interaction Sites In Proteins *J Mol Biol*, **289**, 1093-1108.
- Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. (2003) Improved Protein-Ligand Docking Using Gold *Proteins*, **52**, 609-623.
- Waldmann, T. A. (1993) The Il-2/Il-2 Receptor System: A Target For Rational Immune Intervention Trends *Pharmacol Sci*, **14**, 159-164.
- Wang, J. L.; Liu, D.; Zhang, Z. J.; Shan, S.; Han, X.; Srinivasula, S. M.; Croce, C. M.; Alnemri, E. S.; Huang, Z. (2000) Structure-Based Discovery Of An Organic Compound That Binds Bcl-2 Protein And Induces Apoptosis Of Tumor *Cells Proc Natl Acad Sci U S A*, **97**, 7124-7129.
- Wang, W.; Takimoto, R.; Rastinejad, F.; El-Deiry, W. S. (2003) Stabilization Of P53 By Cp-31398 Inhibits Ubiquitination Without Altering Phosphorylation At Serine 15 Or 20 Or Mdm2 Binding *Mol Cell Biol*, **23**, 2171-2181.
- Weininger, D. (1988) Smiles, A Chemical Language And Information System. 1. Introduction To Methodology And Encoding Rules *J Chem Inf Comput Sci*, **28**, 31-36.

- Wurtele, M.; Jelich-Ottmann, C.; Wittinghofer, A.; Oecking, C. (2003) Structural View Of A Fungal Toxin Acting On A 14-3-3 Regulatory Complex *EMBO J*, **22**, 987-994.
- Yaffe, M. B. (2002) How Do 14-3-3 Proteins Work? Gatekeeper Phosphorylation And The Molecular Anvil Hypothesis *FEBS Lett*, **513**, 53-57.
- Yin, H.; Hamilton, A. D. (2005) Strategies For Targeting Protein-Protein Interactions With Synthetic Agents *Angew Chem Int Ed Engl*, **44**, 4130-4163.
- Zhao, L.; Chmielewski, J. (2005) Inhibiting Protein-Protein Interactions Using Designed Molecules *Curr Opin Struct Biol*, **15**, 31-34.

## **AFFINDB: A FREELY ACCESSIBLE DATABASE OF AFFINITIES FOR PROTEIN-LIGAND COMPLEXES FROM THE PDB**

### **INTRODUCTION**

Understanding the energetics of biomolecular recognition is of paramount importance for a large variety of biomedical and biotechnological disciplines. One of the most prominent examples is given by structure-based drug design where the three-dimensional structure of a target macromolecule (most frequently a protein) is used to identify, design, or optimize small-molecule ligands which bind tightly to the target. Obviously, such design can only be successful if the structural requirements for energetically favorable interactions and high-affinity binding are known. Much of the current knowledge has been gained from comparative analyses of different complex structures and their affinities (Klebe & Boehm, 1996; Babine & Bender 1997). These analyses, however, were normally restricted to rather small sets of data, and the understanding of protein-ligand recognition is still far from being complete, as illustrated by the recurring surprises during projects of molecular design (Müller *et al.*, 2002; Lange *et al.*, 2003; Brenk *et al.*, 2003; Specker *et al.*, 2005). Clearly, more data are instrumental to increase the knowledge about protein-ligand interactions and to improve not only the qualitative understanding, but also the quantitative tools for estimating affinities from complex structures, such as empirical, regression-based scoring functions (Wang *et al.*, 1988; Wang *et al.*, 2002; Boehm, 1994, 1998; Head *et al.*, 1996; Eldridge *et al.*, 1997).

Structural data of protein-ligand complexes are available to a large and rapidly increasing extent through the Protein Data Bank PDB (Berman *et al.*, 2000). This database, however, is a general resource for biomacromolecular structures that had not particularly been designed for protein-ligand complexes. Accordingly, secondary databases such as Relibase (Guenther *et al.*, 2003, Hendlich *et al.*, 2003) or PDBsum (Laskowski *et al.* 1997, 2005) have been developed which provide more convenient access to specific

information about protein-ligand complexes (e.g., search functions for ligand structures; analysis tools for interaction patterns in Relibase). Unfortunately, neither the secondary *structural* databases nor the PDB contain any information about the binding energetics of the corresponding complex, since this information is not required to be included upon submission of structural coordinates to the PDB. On the other hand, some databases exist that collect binding data for enzymes, receptors, or protein-ligand complexes in general, such as BindingDB (Chen & Gilson, 2001) or KiBank (Zhang *et al.*, 2004), but these, in turn, are not limited to complexes with available structure and do not provide a direct link to the available 3D structure of a given complex with measured affinity.

**G**iven the obvious need for databases that establish the missing link between structural information from the PDB and the rather sparse and widely distributed affinity data, we started to develop AffinDB, a database of affinity values collected from the scientific literature for protein-ligand complexes of known structure. Originally intended as a simple tabular collection of affinity values related to PDB codes for in-house use only, the project has grown over time, both with respect to data content and database management, such that it has ultimately been made available to the public as a potentially valuable new resource, despite the recent appearance of other databases of similar scope, most notably PDBbind (Wang *et al.*, 2004, 2005) and Binding MOAD (Hu *et al.*, 2005). In the following, we briefly describe the database architecture and content of AffinDB, as well as the data collection procedure, give a succinct introduction to possibilities for accessing the data through the user interface, and discuss differences and similarities to other databases.



## METHODS

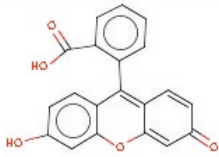
### DATABASE ARCHITECTURE

AffinDB is based on a MySQL (4.0.24) backend machine. The web interface is written in PHP (4.3.10). The database is designed to provide supplementary information for PDB structures of protein-ligand complexes. Accordingly, AffinDB is structured by PDB code, which is the primary reference for all data. Basic PDB meta information about the protein is available for every PDB structure, basic ligand information is provided for ligands with more than five non-hydrogen atoms. Ligand entries of complexes are stored only once in AffinDB, i.e. in case of multiple occurrences of the same ligand in different structures, a pointer to the reference ligand molecule is used. Affinity data and related information are always associated with a specific ligand of a specific PDB structure.

### DATA COLLECTION AND DATABASE CONTENT

The database core is constituted by basic meta information about all PDB structures. To obtain these data, a helper-application was generated with a Python-based Relibase toolkit and the data were retrieved from Relibase+ (Guenther *et al.*, 2003, Hendlich *et al.*, 2003). A further preprocessing step served to store only ligands with more than five non-hydrogen atoms in AffinDB, using a unique and consistently created name for these molecules. The PDB meta information provided for every entry includes the name of the protein or protein class (as given in the header information of the PDB file), the E.C. number (for enzymes), the protein source, the resolution of the crystal structure, and the name of the authors who determined the structure. In addition, for each PDB code links to the following external databases were added: PDB (Berman *et al.*, 2000), Relibase (Guenther *et al.*, 2003, Hendlich *et al.*, 2003), MSD (Boutselakis *et al.*, 2003), SCOP (Murzin *et al.*, 1995), and PDBsum (Laskowski *et al.* 1997, 2005).

The ligand entries consist of the chemical name (as provided in the PDB file), the molecular weight, and the SMILES code (Weininger, 1988) as basic information. In addition, a 2D molecule drawing of the ligand structure is included. These drawings

Affinity information for <b>FLU_PDB1FLR</b>	
<b>Chemical Name:</b> Fluorescein	
<b>Molecular Weight:</b> 331.3	
<b>Affinity value</b>	
<b>Experimental Conditions</b>	
Method	fluorescence quenching and polarization
Temperature	293
pH	6.8
Buffer	50 mM sodium phosphate
Reagents/Additives	n.a.
<b>Literature</b>	
Author	Gibson, A. L. et al.
Journal/Source	<a href="#">Proteins, 3, (1988), 155</a>
Literature primary	yes
<b>Comment</b>	
Method described by Herron in "Fluorescein Hapten: An Immunological Probe", Voss, E. W., Jr, ed Boca Raton, Florida: CRC Press, (1984), 49. - NOTE: With 40% MPD a very different value has been measured!! Additional values at different temperatures and pH 8 are reported in Fig. 3 of the publication.	

**FIGURE 1:** Affinity information window for one of the the affinity entries for PDB complex 1flr.

are generated for every unique ligand with MARVIN 3.5.7 (MARVIN, WWW). BABEL 1.6 (BABEL, WWW), CORINA (3.1) (Gasteiger *et al.*, 1998), and *in house* software was used to harmonize the format of the ligands in order to obtain best results from MARVIN. This automated procedure provided correct drawings for most of the ligands. A small proportion which could either not be drawn by MARVIN or gave distorted pictures had to be postprocessed by hand. Titratable functional groups are always shown in their neutral state, independent of any actual protonation state.

The protein and ligand information described so far is shown by AffinDB regardless whether affinity data is already available for the PDB entry or not. The main purpose of AffinDB, however, is to provide affinity information. Affinity data are exclusively extracted from the scientific literature. Both „primary“ and „secondary“ references

**AffinDB V1.0.1 - Affinity Database For Protein-Ligand Complexes** 3 user(s) online

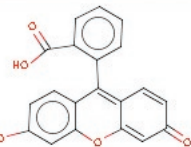
**AffinDB**  
Affinity database for protein-ligand complexes

Home Search Browse Login Contact Help  
Marburg, Friday, September 09th, 2005

**PDB Code: 1flr**

Entry	1flr
Name	Immunoglobulin
E.C. Number	
Source	Mol Id: 1 Organism Scientific: Mus Musculus Organism Common: Mouse Strain: Balb/C Variant: Balb/Cv Cell Line: 4-4-20 Murine-Murine Hybridoma Organ: Spleen Cell: Lymphocyte-Plasma Cell
Resolution	1.85 Å
Author	M. Whitlow
External Databases	<a href="#">PDB</a> <a href="#">Relibase</a> <a href="#">MSD</a> <a href="#">SCOP</a> <a href="#">PDBsum</a>

**Ligands**

Entry	FLU_PDB1FLR																		
Chemical Name	Fluorescein																		
Ligand	 MW: 331.3 Download: <a href="#">SMILES</a>																		
Affinity	<table border="1"> <thead> <tr><th>Affinity</th><th>Author</th><th>Year</th></tr> </thead> <tbody> <tr><td><input type="checkbox"/> <math>\Delta G</math>: -13.94 kcal/mol</td><td><a href="#">Gibson, A. L. et al.</a></td><td>1988 <a href="#">Details</a></td></tr> <tr><td><input type="checkbox"/> <math>\Delta G</math>: -10.98 kcal/mol</td><td><a href="#">Gibson, A. L. et al.</a></td><td>1988 <a href="#">Details</a></td></tr> <tr><td><input type="checkbox"/> Kd: 0.01 nM</td><td><a href="#">Boder, E. T. et al.</a></td><td>2000 <a href="#">Details</a></td></tr> <tr><td><input type="checkbox"/> Kd: 0.31 nM</td><td><a href="#">Boder, E. T. et al.</a></td><td>2000 <a href="#">Details</a></td></tr> <tr><td><input type="checkbox"/> Kd: 0.7 nM <math>\pm</math> 0.3</td><td><a href="#">Boder, E. T. et al.</a></td><td>2000 <a href="#">Details</a></td></tr> </tbody> </table>	Affinity	Author	Year	<input type="checkbox"/> $\Delta G$ : -13.94 kcal/mol	<a href="#">Gibson, A. L. et al.</a>	1988 <a href="#">Details</a>	<input type="checkbox"/> $\Delta G$ : -10.98 kcal/mol	<a href="#">Gibson, A. L. et al.</a>	1988 <a href="#">Details</a>	<input type="checkbox"/> Kd: 0.01 nM	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>	<input type="checkbox"/> Kd: 0.31 nM	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>	<input type="checkbox"/> Kd: 0.7 nM $\pm$ 0.3	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>
Affinity	Author	Year																	
<input type="checkbox"/> $\Delta G$ : -13.94 kcal/mol	<a href="#">Gibson, A. L. et al.</a>	1988 <a href="#">Details</a>																	
<input type="checkbox"/> $\Delta G$ : -10.98 kcal/mol	<a href="#">Gibson, A. L. et al.</a>	1988 <a href="#">Details</a>																	
<input type="checkbox"/> Kd: 0.01 nM	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>																	
<input type="checkbox"/> Kd: 0.31 nM	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>																	
<input type="checkbox"/> Kd: 0.7 nM $\pm$ 0.3	<a href="#">Boder, E. T. et al.</a>	2000 <a href="#">Details</a>																	

Show details for selection

AffinDB V1.0.1, © Peter Block, Christoph Sotriffer, Gerhard Klebe 2002-2005


**FIGURE 2:** Main window showing a PDB entry with affinity data in AffinDB. PDB complex 1flr is used as an example. Five different affinity values measured in different studies and under different conditions are available for this complex. The left navigation bar provides fast access to all functionalities of AffinDB.

are taken into account. A primary reference is a paper describing the original work of the affinity measurement for the corresponding protein-ligand complex. A secondary reference, instead, is any other paper that reports an affinity value for a PDB complex; this may include publications with compilations of affinity data for the development of scoring functions or similar purposes. In a secondary paper, the affinity value for a PDB complex is often only cited, without specifying further experimental details. – So far, more than 740 affinity values covering over 470 PDB complexes could be collected and stored in AffinDB (cf. Discussion).

The input of affinity data into AffinDB is implemented in the form of a wizard. After entering the desired PDB code, AffinDB provides a list of all the ligands of the PDB entry in combination with a 2D molecule drawing and a hint whether affinity information is already stored for the given ligand. After choosing the desired ligand, the user is requested to enter the affinity information. Upon submission of the data, simple checks of the data integrity are performed and the entry is flagged for review by the database curators. Only after a database curator has checked these data, they are released for public access in AffinDB.

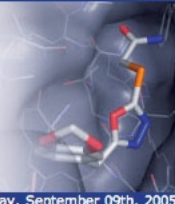
The binding affinity is thermodynamically quantified as free energy of binding  $\Delta G_{\text{bind}}$  or as equilibrium constant (for association:  $K_a$ ; for dissociation:  $K_d$ ) for the reversible equilibrium reaction between protein P and ligand L to form the protein-ligand complex PL:  $P + L \leftrightarrow PL$ .  $\Delta G_{\text{bind}}$  and the equilibrium constants are related by the equation:  $\Delta G_{\text{bind}} = -RT \ln K_a = RT \ln K_d$ , where T is the temperature (in Kelvin) and R is the ideal gas constant ( $8.314 \text{ Jmol}^{-1}\text{K}^{-1}$ ). For enzyme inhibitors, affinities are more frequently quantified in terms of parameters derived from kinetic assays. This may either be the inhibition constant  $K_i$  (which to a first approximation may be considered as a  $K_d$  for the enzyme-inhibitor complex, thus  $\Delta G_{\text{bind}} = RT \ln K_i$ ) or the  $IC_{50}$  value, which is the inhibitor concentration leading to 50% inhibition of the enzymatic activity. In AffinDB,

AffinDB V1.0.1 - Affinity Database For Protein-Ligand Complexes 3 user(s) online



## AffinDB

Affinity database for protein-ligand complexes



[Home](#)   [Search](#)   [Browse](#)   [Login](#)   [Contact](#)   [Help](#)

Marburg, Friday, September 09th, 2005

[Home](#)

[Search](#)

[Browse](#)

[Contact](#)

[Help](#)

---

[Login](#)

[Register](#)

---

[Test DrugScore Online](#)

---

---

[Affinity data: 746](#)


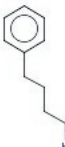
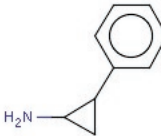
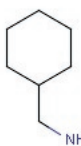
[Covered PDBs: 474](#)

---

[Latest Entries](#)

[Save as csv](#)

**Found affinity information (7 entries)**

Ligand	MW	PDB	Affinity	pH	Author	Year
 <small>MW: 121.18</small>		<a href="#">1tnj</a>	Ki: 11.00 mM pKi: 1.96 <a href="#">Details</a>	8.0	Kurinov I.V. et al. <a href="#">PUBMED</a>	1994
 <small>MW: 150.25</small>		<a href="#">1tnj</a>	Ki: 20.00 mM pKi: 1.70 <a href="#">Details</a>	8.0	Kurinov I.V. et al. <a href="#">PUBMED</a>	1994
 <small>MW: 134.2</small>		<a href="#">1tnj</a>	Ki: 13.30 mM pKi: 1.88 <a href="#">Details</a>	8.0	Kurinov I.V. et al. <a href="#">PUBMED</a>	1994
 <small>MW: 114.21</small>		<a href="#">1tng</a>	Ki: 1.17 mM pKi: 2.93 <a href="#">Details</a>	8.0	Kurinov I.V. et al. <a href="#">PUBMED</a>	1994

**FIGURE 3.** Tabular report, showing part of the search results produced by a query for affinity data published by a specific author (“Kurinov”) in a primary reference.

the affinity value is stored in the same form as published in the specified reference, i.e., without any conversion of type or unit. If available, experimental uncertainties or error margins are saved as well.

Along with the affinity value itself, also the experimental method and conditions of the affinity measurement are stored in AffinDB, if specified in the corresponding reference. This information is provided to the user through a separate “affinity information window” (Fig. 1), which can be opened by activating the “Details” link next to the affinity entry in the main window (Fig. 2). The method by which the affinity was determined is characterized by a keyword or a brief statement. Temperature and pH value at which the measurements were carried out are stored separately. In addition, the buffer and any other significant reagents or additives present in the solution are reported. For the literature reference itself, the name of the first author, the title of the journal, as well as volume, year, and first page of the publication are stored. The reference is linked to the corresponding PubMed entry, which provides direct access to the abstract. A flag indicates whether the reference is of primary or secondary type. Finally, comments and additional valuable information regarding the method, the reference, the structure, or the affinity value itself are saved in a separate data field.

## **DATABASE ACCESS**

The database is freely accessible at <http://www.agklebe.de/affinity>. Data can be retrieved via the PDB code, by defining specific search queries using the affinity search form, or simply by browsing.

Upon specifying a PDB code in the data entry field on the left navigation bar of the main window (cf. Fig. 2), the summary information for the PDB entry is shown. If an affinity value for the ligand is available in the database, it is displayed below the ligand structure, along with the first author and the year of the publication which reports this value. If additional affinity values are available from other references, these are displayed as well, each in a separate line (cf. Fig. 2). Further details can be requested

for each affinity entry. Searching for a specific PDB entry with the affinity search form (accessible through the „Search“ link in the left navigation bar) yields only a result if an affinity value is already associated with the corresponding PDB entry.

In the affinity search form, a variety of queries for affinity data and related information can be defined. It is possible to search for affinities of a certain range of magnitude and for measurements carried out at a specific temperature and/or pH range. Affinities for certain enzyme classes or PDB codes may be retrieved, as well as affinities for ligands of a certain molecular-weight range. Also the affinity values published by a certain author or within a specified time frame can be requested, and the retrieved affinity values may be limited to those obtained from primary literature sources.

AffinDB generates tabular reports for displaying affinity search results and for browsing through the database (Fig. 3). The format of the tables consists of six columns providing the drawing of the ligand structure; the PDB code (linked to the summary information for the PDB entry), the affinity value in the originally reported form as well as converted to the negative base-10 logarithm (i.e., as  $pK_i$ ,  $pK_d$ ,  $pIC_{50}$  (relative to the standard concentration of 1 mol/l)); the pH value of the measurement, the first author of the publication (with a link to the PubMed entry); and the year of the publication. Tables reporting search results can be saved as „csv“ file, which is an ASCII file with semicolon-separated columns and one affinity entry per line.

## DISCUSSION

AffinDB has been designed to provide fast and easy access to affinity data. The popular MySQL backend was chosen as database machine, since MySQL offers a speed-optimized SQL engine. Using the scripting language PHP, special care was taken to

generate a clearly structured layout which enables fast and easy navigation. Since all data can be accessed and retrieved directly via the web browser, the user does not have to install any special software to work with AffinDB.

Using current PC hardware (CPU: AMD Athlon™ XP 2400+), AffinDB executes search queries in less than 0.1 seconds, fairly independent of the complexity of the query. The representation of the tabular report including the 2D molecule drawings needs between 0.1 and 5 seconds, depending on the number of hits (5 seconds if all entries are retrieved). The representation of a PDB entry takes up to 0.2 seconds, depending on the number of affinity data available for that entry. These values reflect only a server-side benchmarking. Obviously, the real speed also depends on the client-side hardware, the Internet connection and the browser.

Data collection for AffinDB is a very time-consuming process which can hardly be automated since scientifically educated readers are required to critically extract the relevant data from the scientific literature. In contrast to other databases (cf. below), we decided to include all affinity data found during literature research for a given PDB complex. Multiple affinity entries may, thus, be available for certain structures. These may reflect measurements with different methods or under different experimental conditions (e.g., PDB 1flr; cf. Fig. 1), or it may be due to additional reports from secondary references, which allows the user to trace back in which context the corresponding complex and its affinity have already been used. Only purely redundant data are not included (e.g., if the value is reported in the same paper as  $K_d$  and  $\Delta G$  derived thereof).

The current coverage of more than 470 PDB structures derives from a priority selection made upon constructing the database. The initial basis was formed by compilations of affinity values from secondary references concerning empirical scoring functions. Due to discrepancies among some of the values and to obtain more detailed information, primary references were also retrieved for part of this initial set. Subsequently, the database was



augmented by seeking affinity data for PDB complexes of different data sets, such as a docking test set of validated structures (Nissink *et al.*, 2002) or data sets for certain target classes (e.g., carbonic anhydrases, trypsin-like serine proteases). Furthermore, published data from our own laboratory were also directly included.

AffinDB is a valuable resource for anyone interested in correlating structural data with binding energetics and complements other databases of similar subject, specifically the Ligand-Protein Database LPDB (Roche *et al.*, 2001), the Protein-Ligand Database PLD (Puvanendrapillai & Mitchell, 2003), PDBbind (Wang *et al.*, 2004, 2005), and Binding MOAD (Hu *et al.*, 2005). LPDB is a compilation of 262 PDB complexes with affinity data. Since it also provides scoring values, docked ligand poses („decoys“), and ligand files setup for docking, LPDB is primarily intended to serve as a data set for testing and developing docking and scoring methods. It does neither provide details nor references for the affinity values. The same is true for PLD, which contains 485 complexes and experimental binding energies for 344 of them. PLD can be searched by using a variety of single search criteria, but no combined search queries are possible. Like AffinDB it is freely accessible to anybody over the internet, whereas PDBbind and Binding MOAD require a registration before granting academic users a free login account. The latter two databases offer by far the largest amount of affinity values, both covering well beyond 1700 complexes in their latest updates. Details about the affinity measurement and experimental conditions, however, are not included, which is an information provided by AffinDB for data retrieved from primary references. In summary, although there is certainly some overlap among the structure-affinity databases recently arisen from independent efforts, there are clear differences in focus, design, and content, rendering each database on its own and in mutual combination an indispensable tool for the scientific community as long as affinities are not reported by the PDB and/or no common repository for biomolecular affinity data exists.

AffinDB encourages users to contribute data and submit references to papers with affinity data for PDB complexes. After registering for upload, an input form can be accessed which facilitates the submission of all relevant data in a clear format. Data submitted by users do not directly enter the database, but must first undergo revision by the database curators. This should ensure high fidelity of the affinity data collected from literature and reported by AffinDB.

**REFERENCES**

- BABEL (WWW) Application To Interconvert Between Many File Formats Used In Molecular Modeling And Computational Chemistry: <http://www.eyesopen.com/babel>.
- Babine, R.; Bender, S. (1997) Molecular Recognition Of Protein-Ligand Complexes: Applications To Drug Design. *Chem Rev*, **97**, 1359-1472.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
- Bohm, H. J. (1994) The Development Of A Simple Empirical Scoring Function To Estimate The Binding Constant For A Protein-Lligand Complex Of Known Three-Dimensional Structure. *J Comput Aided Mol Des*, **8**, 243-256.
- Bohm, H. J. (1998) Prediction Of Binding Constants Of Protein Ligands: A Fast Method For The Prioritization Of Hits Obtained From De Novo Design or 3D Database Search Programs. *J Comput Aided Mol Des*, **12**, 309-323.
- Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Golovin, A.; Henrick, K.; Hussain, A.; Ionides, J.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Oldfield, T.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, J.; Tagari, M.; Tate, J.; Tromm, S.; Velankar, S.; Vranken, W. (2003) E-MSD: The European *Bioinformatics* Institute Macromolecular Structure Database. *Nucleic Acids Res*, **31**, 458-462.
- Brenk, R.; Naerum, L.; Gradler, U.; Gerber, H.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. (2003) Virtual Screening For Submicromolar Leads Of tRNA-Guanine Transglycosylase Based On A New Unexpected Binding Mode Detected By Crystal Structure Analysis. *J Med Chem*, **46**, 1133-1143.

- Chen, X.; Liu, M.; Gilson, M. K. (2001) BindingDB: A Web-Accessible Molecular Recognition Database. *Comb Chem High Throughput Screen*, **4**, 719-725.
- Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. (1997) Empirical Scoring Functions: I. The Development Of A Fast Empirical Scoring Function To Estimate The Binding Affinity Of Ligands In Receptor Complexes. *J Comput Aided Mol Des*, **11**, 425-445.
- Gasteiger, J.; Rudolph, C.; Sadowski, J. (1990) Automatic Generation Of 3D-atomic Coordinates For Organic Molecules. *Tetrahedron Comp Method*, **3**, 537-547.
- Guenther, J.; Bergner, A.; Hendlich, M.; Klebe, G. (2003) Utilising Structural Knowledge In Drug Design Strategies: Applications using Relibase. *J Mol Biol*, **326**, 621-636.
- Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. (1996) VALIDATE: A New Method For The Receptor-Based Prediction Of Binding Affinities Of Novel Ligands. *J Am Chem Soc*, **118**, 3959-3969.
- Hendlich, M.; Bergner, A.; Guenther, J.; Klebe, G. (2003) Relibase: design And development Of A Database For Comprehensive Analysis Of Protein-Ligand Interactions. *J Mol Biol*, **326**, 607-620.
- Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. (2005) Binding MOAD (Mother Of All Databases). *Proteins*, **60**, 333-340.
- Klebe, G.; Bohm, H. J. (1996) What Can We Learn From Molecular Recognition In Protein-Ligand Complexes For The Design Of New Drugs? *Angew Chem Int Ed Engl*, **35**, 2588-2614.
- Lange, U. E. W.; Baucke, D.; Hornberger, W.; Mack, H.; Seitz, W.; Hoffken, H. W. (2003) D-Phe-Pro-Arg Type Thrombin Inhibitors: Unexpected Selectivity By Modification Of The P1 Moiety. *Bioorg Med Chem Lett*, **13**, 2029-2033.

- Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. (1997) PDBsum: A Web-Nased Database Of Summaries And Analyses Of All PDB Structures. *Trends Biochem Sci*, **22**, 488-490.
- Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. (2005) PDBsum More: Nnew Summaries And Analyses Of The known 3D Structures Of Proteins And Nucleic Acids. *Nucleic Acids Res*, **33**, D266-8.
- MARVIN (WWW) Applications To Draw Chemical Structures: <http://www.chemaxon.com/marvin>
- Mueller, M. M.; Sperl, S.; Sturzebecher, J.; Bode, W.; Moroder, L. (2002) (R)-3-Amidinophenylalanine-derived Inhibitors Of Factor Xa with A Novel Active-Site Binding Mmode. *Biol Chem*, **383**, 1185-1191.
- Murzin, A .G.; Brenner, S. E.; Hubbard, T.; Chothia, C. (1995) SCOP: A structural classification Of Proteins Database For The Investigation Of Sequences And Structures. *J Mol Biol*, **247**, 536-540.
- Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. (2002) A New Test Set For Validating Predictions Of Protein-Ligand Interaction. *Proteins*, **49**, 457-471.
- Puvanendrapillai, D.; Mitchell, J. B. O. (2003) L/D Protein Ligand Database (PLD): Additional Understanding Of The Nature And Specificity Of Protein-Ligand Complexes. *Bioinformatics*, **19**, 1856-1857.
- Roche, O.; Kiyama, R.; Brooks, C. L. 3rd (2001) Ligand-Protein Database: Linking Protein-Ligand Complex Structures To Binding Data. *J Med Chem*, **44**, 3592-3598.

- Specker, E.; Bottcher, J.; Lilie, H.; Heine, A.; Schoop, A.; Muller, G.; Griebenow, N.; Klebe, G. (2005) An Old Target Revisited: Two New Privileged Skeletons And An Unexpected Binding Mode For HIV-Protease Inhibitors. *Angew Chem Int Ed Engl*, **44**, 3140-3144.
- Wang, R.; Liu, L.; Lai, L.; Tang, Y. (1988) SCORE: A New Empirical Method For Wstimating The Binding Affinity Of A Protein-Ligand Complex. *J Mol Model*, **4**, 379-394.
- Wang, R.; Lai, L.; Wang, S. (2002) Further Development And Validation Of Empirical Scoring Functions For Structure-Based Binding Affinity Prediction. *J Comput Aided Mol Des*, **16**, 11-26.
- Wang, R.; Fang, X.; Lu, Y.; Wang, S. (2004) The PDBbind Database: Collection Of Binding Affinities For Protein-Ligand Complexes With Known Three-Dimensional Structures. *J Med Chem*, **47**, 2977-2980.
- Wang, R.; Fang, X.; Lu, Y.; Yang, C.Y.; Wang, S. (2005) The PDBbind Database: Methodologies And Updates. *J Med Chem*, **48**, 4111-4119
- Weininger, D. (1988) SMILES, A Chemical Language And Information System. 1. Introduction To Methodology And Encoding Rules. *J Chem Inf Comput Sci*, **28**, 31-36.
- Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. (2004) Development Of KiBank, A Database Supporting Structure-Based Drug Design. *Comput Biol Chem*, **28**, 401-407.

## ZUSAMMENFASSUNG

Protein-Protein Interaktionen spielen in nahezu jedem biologischen Organismus eine essentielle Rolle, beispielsweise in der Signal-Transduktion, der DNA-Synthese, dem Aufbau intramolekularer Strukturen (z.B. Mikrotubuli) oder der Ausbildung des aktiven Zentrums von Enzymen (z.B. HIV-Protease). Physiologisch bedeutsame Protein-Protein Interaktionen weisen eine hohe Spezifität auf und werden so zu einem äußerst interessanten Target für die pharmazeutische Forschung. Eine zielgerichtete und spezifische Modulierung dieser Interaktionen könnte eines Tages zu einer völlig neuartigen Klasse von Arzneistoffen führen, deshalb ist es von großer Bedeutung ein breites Verständnis von Protein-Protein Interaktionen zu erlangen. Für ein rationales, strukturbasiertes Design von potentiellen Arzneistoffen ist außerdem die Aufklärung von Deskriptoren auf atomarer Ebene von essentieller Natur. So ist es beispielsweise bis heute nur unzureichend möglich, Protein-Protein Komplexe zu beschreiben im Hinblick auf ihre Eigenschaften, ob es sich um so genannte „permanente“ oder „transiente“ Komplexe handelt. Unter „transienten“ Protein-Protein Komplexen versteht man solche Komplexe, die unter bestimmten physiologischen Bedingungen dissoziieren können. Sie übernehmen häufig die Rolle von Signal-Transduktoren; so z.B. in G-Protein gekoppelten Proteinen. „Permanente“ Komplexe sind hingegen so fest miteinander assoziiert, dass sie unter physiologischen Bedingungen dauerhafte Kontakte eingehen.

In dieser Arbeit wird das Programmpaket EPIC (Epic Protein Interface Classification) vorgestellt, das die Prozessierung und Klassifizierung von Protein-Protein Komplexen mit Algorithmen aus dem Bereich des Maschinellen Lernens (ML) ermöglicht. Es wird die Vorhersagequalität von vier verschiedenen ML Algorithmen verglichen: Support Vector Maschinen (SVM), C4.5 Entscheidungsbäume, K-Nächste-Nachbarn (KNN) und Näive Bayes (NB). Für die Extraktion relevanter Deskriptoren lassen sich diese Algorithmen mit so genannten Feature-Selektionsverfahren kombinieren, wie

beispielsweise Filter- bzw. Wrapper-Methode und Genetischen Algorithmen. Die Kombination von C4.5 Entscheidungsbäumen und Genetischen Algorithmen konnte einen Datensatz von 345 Protein-Protein Komplexen (147 „permanente“ und 198 „transiente“ Komplexe) in einer so genannten „*Leave-One-Out* Cross-Validierung“ zu 93,6% richtig vorhersagen. Des Weiteren wurde eine Klassifizierung von so genannten Protein-Protein Kristallkontakten (Interaktionen, die allein durch kristallographische Packungseffekte erzwungen werden) gegenüber funktionellen Protein-Protein Komplexen durchgeführt. Ein Datensatz von 172 Protein-Protein Komplexen (76 funktionelle Komplexe gegenüber 96 Kristallkontakten) konnte in einer „*Leave-One-Out* Cross-Validierung“ zu 94,8% richtig klassifiziert werden.

Mit Hilfe der Auswertung und Optimierung anhand Genetischer Algorithmen ließ sich ein Verfahren entwickeln, das es ermöglicht, eine quantitative Aussage über die Relevanz einzelner Deskriptoren zu treffen. Dazu werden alle so genannten Individuen des Genetischen Algorithmus evaluiert und die relative Häufigkeit der einzelnen Deskriptoren ins Verhältnis zur der Vorhersagerate gesetzt. Dadurch lassen sich für alle Deskriptoren Tendenzen über ihre Relevanz ableiten. Durch diese Analyse konnte gezeigt werden, dass beispielsweise das Verhältnis von hydrophober zu hydrophiler Oberfläche zwischen den Protein-Protein Komplexen eine für die Diskriminierung entscheidende Rolle spielt. Eine genauere Betrachtung der Protein-Protein Komplexe zeigte, dass die Kontaktflächen von permanenten Komplexen häufig ein hydrophobes Zentrum aufweisen, welches von einem Ring mit hydrophilen Atomkontakten umgeben ist. Auf der anderen Seite zeigen flüchtige Komplexe eine weitgehend gleichmäßige Verteilung hydrophiler und hydrophober Atomkontakte. Dieses Phänomen lässt sich vermutlich darauf zurückführen, dass flüchtige Komplexe ihre Kontaktoberfläche zeitweise dem Lösungsmittel aussetzen und durch eine gleichmäßige Verteilung hydrophiler Gruppen eine bessere Solvatisierung erfahren.



Der zweite Teil dieser Arbeit konzentriert sich auf die Suche und das Design von Stabilisatoren für Protein-Protein Interaktionen. Obwohl ein Großteil der heute eingesetzten Arzneistoffe allosterisch fungierende Effektoren bzw. Agonisten oder Antagonisten unterschiedlichster Rezeptoren sind, ist die funktionelle Regulierung biologischer Systeme prinzipiell auch durch die Modulierung von Protein-Protein Interaktionen möglich. Der Forschungsschwerpunkt zum Erreichen einer solchen Modulierung lag in den letzten Jahren eindeutig im Design von Inhibitoren, die kompetitiv die Ausbildung des Protein-Protein Kontakts stören. Dazu wurden verschiedene Strategien entwickelt, beispielsweise die Entwicklung von Miniatur-Proteinen, Oligopeptiden oder Peptidomimetika. Das ehrgeizige Ziel kleine, arzneistoffähnliche Moleküle zu entwerfen, die in der Lage sind Protein-Protein Interaktionen zu inhibieren, führte allerdings bis heute nur in Einzelfällen zu Erfolg. Dies ist unter anderem dadurch zu erklären, dass die Proteinoberfläche in der Kontaktfläche von Protein-Protein Komplexen häufig sehr flach ausgebildet ist und somit die zur Ausbildung des Protein-Protein Kontakts kompetitive Bindung von kleinen Molekülen erschwert. Aus thermodynamischer Sicht kann ein kleines Molekül, das nur schwach an die Kontaktoberfläche des Proteins bindet, nur unzureichend ein Protein kompetitiv verdrängen.

Eine Modulierung von Protein-Protein Interaktionen muss allerdings nicht zwangsläufig durch eine kompetitive Inhibierung erfolgen. Mittels einer gezielten Stabilisierung, bei der kleine Moleküle im Randbereich der Proteinkontaktfläche eines Protein-Protein Komplexes binden, kann man ebenfalls die gewünschte Modulierung erzielen. Durch eine ausgeprägte Wechselwirkung des Liganden zu beiden Proteinen des Komplexes kann es folglich zu einer verzögerten Dissoziierung und somit zu einer Modulierung der Protein-Protein Interaktion kommen. Dieses Phänomen wird eindrucksvoll durch die Bindung von Fusicoccin, einem diterpenoidem Phytotoxin, welches die Interaktion zwischen einer pflanzlichen  $H^+$ -ATPase und einem 14-3-3 Protein um nahezu den Faktor 100 verstärkt, beschrieben. Diese Stabilisierung führt zu einer dauerhaften Aktivierung

der Protonenpumpe und bewirkt letztendlich ein Welken der Pflanze. Beeindruckend ist dabei die relativ schwache Bindungsaffinität des Fusicoccins an den Proteinkomplex (66  $\mu\text{M}$ ).

Auf der Suche nach niedermolekularen Verbindungen, die ebenso wie Fusicoccin den Protein-Protein Komplex stabilisieren, wurden verschiedene Datenbanken mit käuflich erwerbbaaren Molekülen durchmustert. Die Anzahl der ca. 2 Millionen verfügbaren Moleküle ließ sich durch verschiedene Filterschritte reduzieren, deren Komplexität schrittweise erhöht wurde. Auf diese Weise konnte die Anzahl auf ca. 160000 Kandidatenmoleküle eingeschränkt werden. Mit verschiedenen Dockingprogrammen wurden diese in die Bindetasche des  $\text{H}^+$ -ATPase/14-3-3-Komplexes eingepasst. Die zahlreichen generierten Dockingposen wurden in einer Datenbank gespeichert und im Folgenden anhand geeigneter Pharmakophorfilter selektiert. Dazu ließen sich Methoden entwickeln, die effizient mit großen Datenmengen umgehen können. Diejenigen Moleküle mit pharmakophorerfüllenden Eigenschaften und Geometrien wurden anschließend mit verschiedenen Bewertungsfunktionen evaluiert. Für einen intuitiven Einblick in die Beiträge einzelner Atome des Liganden zu dessen Gesamtbewertung, wurde eine etablierte Bewertungsfunktion in ihrer Funktionalität erweitert. Mit Hilfe der erweiterten Bewertungsfunktion wurden schließlich verschiedene Moleküle für eine *in vitro* Testung ausgewählt.

Des Weiteren konnte an einem Datensatz mit 198 Protein-Protein Komplexen gezeigt, dass nahezu alle der untersuchten Komplexe taschenförmige Vertiefungen im Randbereich ihrer Kontaktfläche aufweisen. Eine nähere Betrachtung der Taschen zeigt, dass einige eine ähnliche Gestalt zu Bindetaschen in globulären Proteinen aufweisen, die bekanntermaßen kleine Moleküle binden. Diese Erkenntnis lässt vermuten, dass auch Bindetaschen im Randbereich von Protein-Protein Komplexen ein vielversprechendes Target für die Bindung kleiner Moleküle darstellen. Eine Modulierung von Protein-

Protein Interaktionen im Sinne einer Stabilisierung durch niedermolekulare Verbindungen wie im Falle des Fusicoccins erscheint somit als interessante Alternative zur Inhibierung solcher Interaktionen.

Ein weiterer Teil dieser Arbeit entstand im Verlaufe eines Projektes zur Entwicklung einer verbesserten Bewertungsfunktion *in silico* generierter Dockingposen, wie sie bei der Auswahl möglicher Liganden zum Binden in die Fusicoccin Bindetasche des H<sup>+</sup>-ATPase/14-3-3-Komplexes eingesetzt wurden. Für die Entwicklung empirischer Bewertungsfunktionen, bei denen vorhergesagte gegenüber gemessenen Affinitäten regressionsbasiert korreliert werden, sind große und diverse Datensätze von Protein-Ligand Kristallkomplexen und deren ermittelter Affinität essentiell. In diesem Zusammenhang konnte die webbasierte Datenbank AffinDB entwickelt werden. Sie umfasst inzwischen über 730 gemessene Affinitäten allgemein zugänglicher Protein-Ligand Kristallstrukturen aus der Protein Data Bank (PDB). AffinDB ist im Internet unter <http://www.agklebe.de/affinity> frei verfügbar.

## APPENDIX

### PUBLICATIONS ARISING FROM THIS WORK

#### ARTICLES

Block, P., Paern, J., Huellermeier, E., Sanschagrin, P., Sotriffer C. A., Klebe, G. PHYSICOCHEMICAL DESCRIPTORS TO DISCRIMINATE PROTEIN-PROTEIN INTERACTIONS IN PERMANENT AND TRANSIENT COMPLEXES SELECTED BY MEANS OF MACHINE LEARNING ALGORITHMS. (*expected 2006*).

Block, P., Weskamp N., Klebe, G. STRATEGIES TO SEARCH AND DESIGN STABILIZERS OF PROTEIN-PROTEIN INTERACTIONS: A FEASIBILITY STUDY. (*expected 2006*).

Block, P., Sotriffer, C. A., Dramburg, I., Klebe, G. (2006) AFFINDB: A FREELY ACCESSIBLE DATABASE OF AFFINITIES FOR PROTEIN-LIGAND COMPLEXES FROM THE PDB. *Nuclear Acid Research*, Database Issue 2006. (*in press*)

#### POSTERS

THE TEMPTATION OF HIGH-THROUGHPUT DOCKING. POSSIBLE STRATEGIES AND THE DEVELOPMENT OF REQUIRED TOOLS. (2005) International Workshop New Approaches in Drug Design & Discovery, Rauschholzhausen, Germany.

MODULATING PROTEIN-PROTEIN INTERACTIONS BY SMALL LIGANDS. (2004) 18. CIC-Workshop der GDCh, Boppard, Germany.

**AWARDS**

AffinDB has been awarded with the first prize of the web award 2005 of the MOLECULAR GRAPHICS AND MODELLING SOCIETY, German Section (May 2005, Erlangen, Germany).

# CURRICULUM VITAE

**Peter Block**

**DAY OF BIRTH** October 21, 1974

**PLACE OF BIRTH** Bergisch Gladbach, Germany

## DISSERTATION

Prof. Dr. Gerhard Klebe 06.2002 - today  
*Research group for drug design &  
X-ray crystallography*  
PHILIPPS UNIVERSITY MARBURG (Germany)

## DIPLOMA STUDIES

Prof. Dr. Gerhard Klebe 06.2001 - 05.2002  
PHILIPPS UNIVERSITY MARBURG (Germany)

## PRACTICAL TRAINING

GOETHE APOTHEKE (Pharmacy), Linden (Germany) 05.2000 - 10.2000  
CAMBRIDGE CRYSTALLOGRAPHIC DATA CENTRE (U.K.) 11.2000 - 05.2001

## STUDIES OF PHARMACY

PHILIPPS UNIVERSITY MARBURG (Germany) 04.1996 - 04.2000

## CIVIL SERVICE

HERTHA-VON-DIERGARDT-HAUS (Old people's home) 08.1994 - 10.1995  
Leverkusen (Germany)

## ABITUR (A-LEVEL)

LISE-MEITNER-GYMNASIUM (Grammar School) 06.1994  
Leverkusen (Germany)

# ERKLÄRUNG

Ich versichere, dass ich meine Dissertation

**CONCEPTS TO INTERFERE WITH PROTEIN-PROTEIN COMPLEX FORMATIONS:**

**DATA ANALYSIS, STRUCTURAL EVIDENCE AND STRATEGIES**

**FOR FINDING SMALL MOLECULE MODULATORS**

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlich Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszweckn gedient.

Marburg, den 16. Dezember 2005

