Zoltán Simon

Drug Discovery by Polypharmacology-based Interaction Profiling

PhD thesis

Supervisor: András Málnási-Csizmadia DSc.

Structural Biochemistry Doctoral Program
Doctoral School in Biology

Program Leader: Prof. László Gráf DSc.
Head of the School: Prof. Anna Erdei DSc.

Eötvös Loránd University
Budapest, Hungary
2010

A tudomány és művészet hazája nem a lét, az „esse”,
hanem a lehetőség, a „posse”, s ha a létben megnyilvánul,
attól a lét lesz gazdagabb; a tudomány és művészet részéről
végtelen alázat, hogy a létben magát megnyilatkoztatni engedi,
hiszen minden alakot-öltése fogyatékos.

WEÖRES SÁNDOR

Preface

Since the first Venus statues carved by the prehistoric man, artists attempt to reach the divine and bring it from the heaven down to the earth, to the community. Similarly, science aspires to get acquainted with the phenomena of the world and to understand their nature. As no artist has ever been able to capture and artificially reproduce the ultimate Beauty, scientists have never found the Truth in its entirety. Art and science, as stated by Sándor Weöres in its work entitled *Towards completeness*, are necessarily incomplete and deficient in our world.

Nevertheless, science struggles to understand the complex nexus of things for several thousand years. Recent methodical improvements, especially the rapidly increasing computational capacity and mathematical systems theory enable us to find answers or at least approximations for complex questions in a new, complex manner. Prediction of the behavior of drug molecules in the human body is one of the most complex questions. It is not surprising that pharmacology has not yet solved the problem of capturing the entire bioactivity profiles of small-molecule compounds. Consequently, major side effects might remain hidden and endanger the patients’ health – or, on the other hand, potentially beneficent alternative uses of drugs were rarely recognized.

The early pharmacologic theories explained drug-effect associations in a mechanistic manner, implying that a drug selectively acts on a specific biological target and affects its function like a magic bullet (Ehrlich). Now the tide seems to turn: drugs are recognized as affecters of complex biological networks. Systems-based approaches are gathering larger and larger ground, bringing a holistic view into drug research and giving an opportunity to reveal the full effect profiles of drugs.

A similar shift of scientific viewpoint occurred in a much smaller but not simpler system: a protein. In Emil Fischer's time, proteins were considered as static objects whose interactions with other molecules could be described as the interaction of two complementary shapes, a key and a lock. Now, proteins ceased to be purely mechanical objects; they become dynamic, "breathing" particles. In contrast with the "molecular machinery" approach, it is now revealed that proteins show more flexibility than anything engineered by a human being. The Department of Biochemistry (and its ancestor), where my PhD work was carried out, has a great tradition of the paradigm of flexibility and the handling of complex problems, from Albert Szent-Györgyi to the present work of László Gráf [1].

In my thesis, I assess these two levels of complexity. First, I present a newly developed approach that attempts to predict hidden effects on known drugs. I also show that protein dynamics can be approached with internal viscosity, a specific measure of protein flexibility in an interdomain conformational rearrangement. The two topics are related not only in complexity but also in the similar treatment needed to process them. In pharmacological effect prediction, complex drug-protein interaction patterns and bioactivity profiles must be handled that mean an enormous amount of information. Dimension reduction and capturing of the important factors are the keys to solve the main problems of pharmacology. On the other hand, a single protein possesses an almost infinite complexity. For example, human trypsin 4 contains 216 amino acid residues. Consider only two conformers for each residue – this will result in 2^{216} possible conformations of this single protein [2]. This practically infinite conformational space makes it impossible to understand protein flexibility; dimension reduction is needed to find the smaller system in which flexibility can be studied. The activation of a human trypsin isoform, fine-tuned by a single residue at a specific position in the protein, offers us an ideal model system.

Therefore, my thesis consists of two parts. In the first part, I present the holistic approach which lead to the development of *Drug Profile Matching* method which is able to systematically capture the bioactivity profiles of drug molecules in their entirety. I give a short overview of the different branches of *in silico* pharmacology, highlighting their advantages and disadvantages. After that, I present a recent paradigm called *polypharmacology*, i.e., the observation that many drugs affect multiple targets. Polypharmacology can bring the long-awaited breakthrough in drug discovery: it is a key to understand and catch the full spectrum of pharmacological actions of a compound in the human body. I review the latest attempts to exploit polypharmacology in bioactivity prediction and protein binding site description. Then, I present our starting hypothesis that interaction profiles of drugs, even if generated *in silico*, correlate with their bioactivity profiles. I present different ways of effect prediction based on polypharmacology: a one-dimensional method and a more sophisticated, multidimensional one, the so-called Drug Profile Matching. Both approaches justify our starting hypothesis and are able to predict full effect profiles of drugs with high confidence. I point to our secondary finding that binding site geometry plays a minor role in the determination of affinity profiles in general; however, there are certain drug categories for which binding site shape is a crucial feature. I prove that *in silico* interaction profiles serve sufficient information for reliable bioactivity prediction without the consideration of the interactions of drugs with known targets *in vivo*. Results of in-house developed and independently performed *in vitro* and cell culture tests of certain effect predictions will be discussed. I summarize the recent and the possible future applications of Drug Profile Matching, i.e., drug repositioning predictions and bioactivity prediction of drug candidates, respectively.

In the second part, I describe the study of the effect of point mutations on the rate of a specific conformational rearrangement. Mutations were introduced in a hinge region playing a major role in the activation of human trypsin 4. I prove that the rate of the conformational transition of the trypsin mutants is inversely proportional to the solvent viscosity. This phenomenon is interpreted in terms of the Kramers' theory. I conclude that the rate of the conformational change during activation is determined by the internal viscosity around this hinge site and the flexibility of a protein regarding this specific conformational transition can be affected by point mutations at the hinge region. This work is the first study that points to the effects of internal friction on the energy barrier of an enzymatic transformation. Since a new methodology was needed to study enzymatic reactions in a wide temperature range, we developed and applied a novel transient kinetic equipment called heat-jump/stopped flow. Based on our recent experiments, we propose that friction compresses the complex features of an enzymatic reaction, i.e., the inherent flexibility of a protein and the roughness of the potential energy landscape, into a one-dimensional parameter, the internal viscosity.

In summary, my PhD work comprises the issues of multiparameter systems and their common methodical problem, i.e., the experimental selection of the relevant features.

Contents

Preface 1

Contents 4

Acknowledgements 8

Abbreviations 10

Glossary 11

Part I: Drug Discovery by Polypharmacology-based Interaction Profiling 13

Introduction 13

1. Drug development: an overview 13

1.1 Current scientific problems 15

1.2 Current problems from the viewpoint of the pharmacological industry 17

2. *In silico* pharmacology 19

2.1 Quantitative Structure-Activity Relationships 20

2.2 Virtual Ligand Screening 20

2.2.1 Ligand-based VLS strategies 21

2.2.2 Target-based VLS strategies 21

2.3 Virtual Affinity Profiling 24

2.3.1 Ligand-based VAP strategies 24

2.3.2 Target-based VAP strategies 28

2.4 *In silico* pharmacology: a short summary 29

3. Polypharmacology 30

4. Protein binding site description 37

5. Starting hypothesis and applied methodology 38

5.1 Risk analysis of the initial hypothesis 40

Methods 41

1. Development of the Interaction Profile database 41

1.1 Data Collection 41

1.2 Docking Preparations 41

1.3 Docking 42

1.4 Filtering, normalization and centralization 43

2. Generation of the Effect Profile matrix 43

3. Generation of PocketPicker shape descriptor matrix 44

4. One-dimensional analyses assessing the relation between binding affinity patterns and bioactivity profiles 46

4.1 IP-based Drug-Drug Similarity Calculations 46

4.2 Validation of effect prediction 46

4.2.1 Randomization of the effect dataset 47

4.2.2 Leave-one-out cross-validation 47

4.3 Effect Prediction 48

4.3.1 Neighbor-focused Prediction Method 48

4.3.2 Effect-focused Prediction Method 48

5. Multidimensional analyses 49

5.1 Principal Component Analysis 49

5.2 Canonical Correlation Analysis 50

5.3 Canonical Redundancy Analysis 51

5.4 Linear Discriminant Analysis 52

5.5 Validation 52

5.6 Receiver Operating Characteristic analysis 53

5.7 Top Hit Rate calculation 53

6. *In vitro* analyses 54

6.1 Materials 55

7. Cell culture D1, D2, α1B, α2A and ß1 assays 55

Results and Discussion 57

1. One-dimensional analyses 58

1.1 Validation 60

1.2 Effect prediction 62

1.3 Summary and further optimization possibilities 64

2. Multidimensional analyses 65

2.1 Protein diversity analysis 67

2.2 Analysis of the importance of binding site geometry 67

2.2.1 PCA of molecular affinity profiles of target proteins 67

2.2.2 PCA of the geometric characteristics of protein binding sites 69

2.2.3 Comparison of the factorial structure of molecular affinity profiles and geometric characteristics of protein binding sites 71

2.2.4 Canonical Correlation Analysis 72

2.2.5 Sensitivity Analysis: the importance of different scoring functions 77

2.3 Multidimensional IP-EP correlation 78

2.3.1 Assessing Accuracy 80

2.3.2 Leave-one-out validation 82

3. Case studies 85

3.1 Ziprasidone 86

3.2 Bromodiphenhydramine 87

3.3 Valproic acid 87

4. Experimental confirmations 88

4.1 ACE Inhibitors 88

4.2 COX Inhibitors 91

4.3 Cell culture tests 95

5. Summary 96

Part II: Internal viscosity: the role of hinge residues in trypsin activation 98

Introduction 98

Aims 101

Materials and methods 102

1. Development of the heat-jump/stopped flow 102

1.1 Heat-jump/stopped-flow apparatus 102

1.2 Calibration of the adjusted temperature pairs of the heating loop and the cuvette house 103

1.3 Dead time determination 103

2. Mutagenesis and expression of human trypsinogen 4 variants 103

3. Steady state kinetic measurements 104

3.1 Determination of k_{cat} and K_m 104

4. Transient kinetic measurements 104

4.1 Stopped flow measurements 104

4.2. Heat-jump/stopped-flow experiments 105

4.3. Determination of dependence of the rate constants on the relative external viscosity 105

4.4 Kinetic and thermodynamic analysis 105

Results 108

1. Principle and construction of the heat-jump/stopped-flow 108

1.1 Calibration of heat jump stopped flow 110

2. Internal viscosity: the role of hinge residues in trypsin activation 111

2.1 Enzymatic activity of human trypsin 4 and its mutants at position 193 111
2.2 Kinetic analysis of the conformational change during pH-jump activation 112
2.3 Thermodynamic analysis of the conformational change during pH-jump activation 112
2.3.1 Temperature dependence of the conformational transition 112
2.3.2 External viscosity dependence of the conformational transition 114
2.3.3 Relative internal viscosity values of R193A/G mutants and wild type trypsin 4 115
Discussion 117
1. Advantages of tryptophan signal detection 117
2. Evaluation of the observed rate constants 117
3. Structural consequences of the replacement of R193F with a bulkier residue 118
4. The effects of a bulkier residue at position 193 on the kinetic and thermodynamic parameters of the conformational transition 118
5. Internal viscosity of the trypsin hinge mutants 120
Conclusions 121
Appendices 124
1. List of the applied drugs 124
2. List of the applied proteins 136
3. Effect prediction parameters 140
4. Neighborhood of valproic acid 145
Summary 148
Összefoglalás (Summary in Hungarian) 149
References I-XIII

Acknowledgements

Regularly, this section should be placed at the end of the work. Nevertheless, I would like to express my gratitude here, before the presentation of the results since this thesis summarizes a five-year long working period in which a lot of people were involved.

My greatest thanks goes to my supervisor, *András Málnási-Csizmadia*. He provided enormous scientific support during these years, helping to solve the daily arising problems on this difficult field and lighting the future way of the research. Moreover, I have learnt a lot from him about project management and self-management as well. We have just begun to reap the harvest of our work; I hope we can continue it.

I am thankful to *Péter Hári* at Delta Informatika, Inc. Being a businessman, he always believed in this project, which is, I believe, crucial for achieving something important. As my principal, he always secured the freedom I needed to do my work.

I am very greatful to *Pál Czobor* at Semmelweis University. I have never encountered anyone with at least similar mentality; I have learnt a lot from the challenging discussions.

I thank *László Gráf* and *László Nyitray* for their support and the opportunity to work at the Department of Biochemistry. It is pleasant to remember the joint work with Professor Gráf not only on protein flexibility but on a non-scientific project, the editing of the 40th anniversary book of the department.

Very special thanks go to *Ágnes Peragovics* and *Margit Vigh-Smeller*. No one can wish more talented PhD students. Without their work, this project could not reach this blossoming stage. I thank *Balázs Jelinek* for his management support and the useful scientific and interesting non-scientific conversations. I am thankful to *Anna Rauscher* for her reading suggestions and her clever ideas that usually approach problems from a view that is completely different from mine. I thank *Zhenhui Yang* and *Gergely Zahoránszky-Kőhalmi* for their work at the beginning of the project. I am thankful to all members of Málnalab not mentioned before: *László Végner* (especially for his work on the *in vitro* tests), *Boglárka Várkuti, Miklós Képíró, Vitalii Stadnyk, István Lőrincz, Ilona Ozoróczyné*, and the former members: *Dániel Papp* and *Erzsébet Gere-Pászti*. I thank *Bálint Kintses* and *Máté Gyimesi* for their help since I started my work at Málnalab. I am thankful to *István Bitter*, *Gábor Csukly* and *László Tombor* at Semmelweis University for their work on the common projects.

Special thanks go to *Júlia Tóth* for our work together. I learnt a lot from her. I say thank you to *Mihály Kovács*, leader of Stocilab, for all his comments on my work and for the inspiring discussions. I am also thankful to the members of his team.

I would like to thank the informatics team at Delta Informatika Inc., led by *Zoltán Brandhuber*, for their efforts to make this project live since 2006.

I thank *Beáta Simon* and the members of Málnalab, especially *Ágnes Peragovics, Margit Vigh-Smeller* and *Anna Rauscher* for reading and correcting the first versions of my thesis.

I am thankful to my teachers, especially *Andrea Varga, Zoltán Gőgh* and *Imre Csőre*, for their slight but continuous direction toward science during my school years.

I will never forget the persistent support of my mother, my father and the whole family. My friends have also helped a lot since decades. Thank you all.

Abbreviations

Part I

ACE: Angiotensin-converting enzyme

ADME(T): Adsorption, Distribution, Metabolism, Elimination, (Toxicity)

AUC: Area Under the Curve

CCA: Canonical Correlation Analysis

COX: Cyclooxigenase

CRA: Canonical Redundancy Analysis

EP: Effect Profile

FDA: Food and Drug Administration

FPR: False Positive Rate

IP: Interaction Profile

LDA: Linear Discriminant Analysis

MAF: Molecular Affinity Fingerprint

MIF: Molecular Interaction Fingerprint

PCA: Principal Component Analysis

PD: Pharmacodynamics

PK: Pharmacokinetics

QSAR: Quantitative Structure-Activity Relationship

ROC: Receiver Operating Characteristic

SAR: Structure-Activity Relationship

TPR: True Positive Rate

VAP: Virtual Affinity Profiling

VLS: Virtual Ligand Screening

Part II

CABS: 4-(Cyclohexylamino)-1-butanesulfonic acid

DMF: dimethylformamide

HEPES: N-(2-Hydroxyethyl)piperazine-N′-(2-ethanesulfonic acid)

k_{cat}: catalytic constant

K_m: Michaelis constant

NATA: N-acetyl L-tryptophan amide

NBS: N-bromo-succinimide

Tricine: *N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine

Z-Gly-Pro-Arg-pNA: N-carbobenzyloxy-glycyl-prolyl-arginyl p-nitroanilide

Glossary

DOCKING: Computational mapping of the conformational space in order to find the most efficient conformation of two interacting molecules, e.g., a drug and a protein. The term “most efficient” is evaluated by scoring functions that estimate binding affinity, i.e., the strength of interaction between the two compounds.

POLYPHARMACOLOGY: The binding of a drug to multiple targets.

MOLECULAR INTERACTION FINGERPRINT (MIF): A vector of calculated binding free energies for a compound against a protein set, compared to a vector of reference values.

MOLECULAR AFFINITY FINGERPRINT (MAF): A vector of calculated binding free energies for a set of small molecules against a protein, compared to a vector of reference values.

INTERACTION PROFILE (IP): A series of *in silico* calculated binding affinity values for a drug against a predefinied set of proteins. This profile reflects the interaction properties of the small molecule.

EFFECT PROFILE (EP): An interpretation of the bioactivity properties of a compound. In this work, it refers to a binary fingerprint containing 181 entries for the studied 181 effects.

PRINCIPAL COMPONENT ANALYSIS (PCA): A mathematical transformation of several possibly correlated variables into a smaller number of uncorrelated variables, i.e., the principal components. Principal components explain the variance of the data set in a decreasing order.

LINEAR DISCRIMINANT ANALYSIS (LDA): A commonly used statistical approach to identify the best discriminating surfaces in the multidimensional space of feature sets that generate complex pattern classes.

CANONICAL CORRELATION ANALYSIS (CCA): A method to study the relationship between two datasets by creating derived variables that are linear composites of the original variables. Its principal goal is to simplify complex relationships, while providing some specific insights into the underlying structure of the data.

CANONICAL REDUNDANCY ANALYSIS (CRA): CRA is used to study the “overlap” between two sets of variables in terms of explained variance. This approach allows for the determination of the amount of variance (or redundancy) that the canonical components account for in their own set of variables and in the opposite set of variables.

S1 PROTEASE FAMILY: An endopeptidase protein family with a common catalytic mechanism which a serine residue is involved in.

TRUE POSITIVE RATE (TPR): The fraction of true positives out of the positives; also known as sensitivity.

FALSE POSITIVE RATE (FPR): The fraction of true negatives out of the negatives. It can be referred to as (1-specificity).

RECEIVER OPERATING CHARACTERISTIC (ROC): A TPR vs. FPR plot, referring to the classification accuracy, helps in decision making.

AREA UNDER THE CURVE (AUC): Area under a TPR vs. FPR plot, called Area Under the Curve (AUC) is the measurement of the accuracy of classification. AUC=1 means perfect classification while a random classification results in an AUC value of 0.5.

TANIMOTO SIMILARITY INDEX: Tanimoto coefficient is a commonly used measurement for the comparison of two binary fingerprints with the same length. E.g., for molecules A and B, if N_A and N_B represents the number of “1” entries for compound A and B, respectively, and N_{AB} is the number of common “1” entries in both fingerprints, Tanimoto similarity coefficient is calculated as:

$T=N_{AB}/ (N_A + N_B – N_{AB})$

Tanimoto dissimilarity is also used:

$T_D=1-T$.

Part I: Drug Discovery by Polypharmacology-based Interaction Profiling

Introduction

1. Drug development: an overview

The behavior of a drug molecule in a biological system is immanently ambiguous. From one hand, a drug or other xenobiotic affects the biological system, possessing *bioactivity* and *toxicity*, commonly referred to as *pharmacodynamic* (PD) events. On the other hand, the biological system also acts on the drug by *absorbing*, *distributing*, *metabolizing* and *excreting* it. These actions are collectively named as *pharmacokinetic* (PK) events. Here, a single protein, isolated cells or even an entire human organism can be considered as a biological system [3]. Pharmacodynamic and pharmacokinetic events are naturally and necessarily interdependent. A drug can affect the PK abilities of an organism, while absorption, distribution and elimination determine and modify the spatial and temporal distribution of PD events. The situation becomes more complex when one considers metabolism which produces compounds with their own PD and PK properties. (*A*dsorption, *d*istribution, *m*etabolism and *e*limination are often referred collectively to as ADME properties. The abbreviation is sometimes extended with a T that stands for *t*oxicity.)

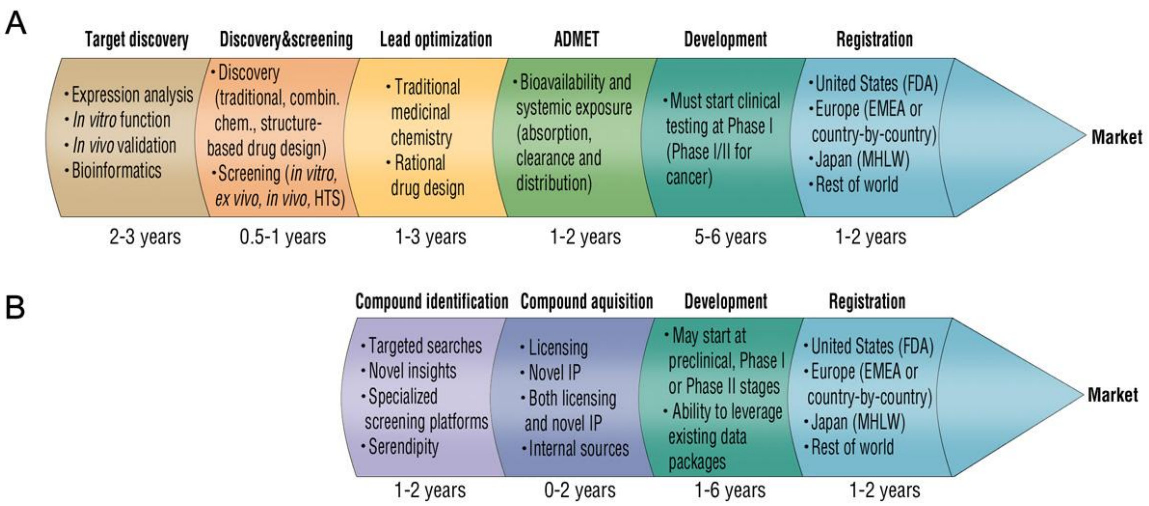


Figure 1. Drug development pipelines. **A.** Traditional drug development. **B.** Drug repositioning. (After [4], modified).

In a nutshell, the aim of drug development is to provide bioactive compounds that show beneficent PD and PK properties, in order to restrain human diseases, increase life expectancy and elongate the productive lifespan of human beings. To achieve this goal, a pipeline of drug development has been evolved during the cca. 100 years of pharmaceutical industry. Traditional or *de novo* drug discovery and development starts with target discovery, a survey for a potentially drugable protein (Figure 1a). Actually, only 600 human proteins are registered as known targets of the cca. 1,200 small-molecule drugs approved by the U.S. Food and Drug Administration (FDA) (own data). Serendipity often helps finding new targets; i.e., selective serotonin inhibitors were identified during a screening for antihistamines [5]. In the next step, compounds are screened against the target, performing a succession of *in silico, in vitro, ex vivo* and *in vivo* preclinical screening phases. The active compounds (leads) are further optimized in order to develop molecules with better PD/PK properties. After this level, clinical screening phases begin [6]. A Phase I clinical study examines the effect of the drug candidate in 20-80 human beings and assesses its effectiveness in terms of PD/PK events. If the observations are in synchrony with the PD/PK events described before, a Phase II study can be performed, involving no more than a few hundred patients. Here, efficacy of the drug in a specific therapeutic use is examined and the side effects are monitored. After a successfully completed Phase II study, the drug candidate enters Phase III in which a larger number of patients are involved. Longer-term safety and efficacy studies are carried out in this stage. If a candidate proves its applicability in a specified cure, the registration procedure begins and the drug can finally reach the market. The whole procedure takes 10-17 years and the success rate is less than 10%. Nowadays, post-marketization experiences on efficacy and adverse events (side effects) are collected and this stage is sometimes referred to as Phase IV.

The results of pharmaceutical industry are unquestionable; however, there is still room for improvement. The simplificative approaches of the past decades are passing away; making space to a newly emerging body of complex methodologies. The increasing computational capacity enables us to handle biological/medical data of a level of magnitude that was impossible even ten years ago. This fruitful tendency in informatics opened the way to complex, systems-like approaches that are closer to the organism in their complexity than the former attempts. As a mathematical systems theory states, "the scale and the complexity of the solution should match the scale and complexity of the problem" [7]. This law stands in biology, or more closely, in pharmacology as well. In order to develop appropriate cures for complex diseases like cancer, cardiovascular disorders and mental illnesses, complex models should be applied. With the blossoming number of public databases and *in silico* tools, the amount of data is not a limiting factor anymore; the question is now "what is valuable". What kind of experiments (including data mining) should be performed in order to describe the PD or PK properties of a compound and how can one extract the information needed to catch PD/PK events in their entirety? To answer these questions, we first take a look on the conventional drug development pipeline and review its actual troubles. Finally, we take a glimpse of the field of *in silico* pharmacology to see the attempts to overcome the concerns about drug development.

1.1 Current scientific problems

Generally speaking, prediction of processes or unknown parameters of complex systems, e.g. pharmacology is one of the most exciting questions in modern science. Predicting effect profiles of drugs and drug candidates is a great challenge.

Many attempts have been made to unravel the bioactivity profiles in their entirety. In this work, I will summarize the different approaches in *in silico* pharmacology, from the first quantitative structure-activity relationships to the novel techniques that try to estimate the whole bioactivity pattern of a compound with a series of calculations. In synchrony with the recent findings, effect profile of a drug is a complex feature, since a molecule entering the organism usually interacts with multiple targets as indicated by the theory of polypharmacology [8-10]. Multiple actions may be important for clinical efficacy, especially in case of complex diseases. For example, psychiatric drugs affecting several well-defined proteins have high efficacy [11]. The earlier single target-based approaches therefore might prove insufficient for identifying the full spectrum of effect profiles (EPs) [7].

Hitherto, heuristic and empirical experiences have played the principal role in identifying various effects of bioactive molecules. Some recently developed systematic prediction methods [12-14] increase the efficiency of drug development and safety control. A logical continuation of these approaches would be to relate atomic-level information with bioactivity profiles.

Our working hypothesis was that a feature set must comprise similar complexity to that of clinical effect profiles in order to yield systematic information with predictive power for the effect profiles [7]. The task was to extract the relevant information stored in complex feature sets of drug molecules in order to unravel effect profiles in their entirety. To accomplish this, in the present study an atomic-level strategy is introduced for the prediction of the effect profiles of drugs by systematic mapping of their molecular interactions. For this, the central assumption of polypharmacology is adopted and it is presumed that similar interaction profiles (IPs) of molecules are related to their similar biological actions. In order to test this assumption empirically, we generated IPs for 1,226 FDA-approved drugs by calculating their binding affinities for a set of proteins and the IPs were correlated with the EPs of all drugs. A correlation between IPs and EPs would hold out the promise for the discovery of novel effects of drugs and prediction of side effects of drug candidates in the development phase. The aims of my work are to uncover IP-EP relationships, and to derive general rules for effect prediction. Our principal findings were validated statistically and confirmed by a series of systematic, unbiased *in vitro* experiments.

We were also interested in the importance of a molecular feature, i.e. the shape of protein binding sites that build up the IPs. As a secondary aim, this question was assessed by combining IPs with geometric descriptors for each protein. We found that, generally, the geometry of the binding site is not a pivotal factor in selecting drug targets. Nonetheless, based on strong specific associations between certain IPs and specific geometric descriptors, the shapes of the binding sites do have a crucial role in virtual drug design for certain drug categories.

1.2 Current problems from the viewpoint of the pharmacological industry

Extreme cost demands of pharmacological research, combined with its relevance on the life of human beings made a strong interdependence of science and industry – an interdependence some consider to be unnatural and of hazardous economical and social importance. Presenting the industrial point of view in a PhD thesis might seem unusual, but due to this undeniable relationship, I believe there is space for these considerations in a pharmacology-related work.

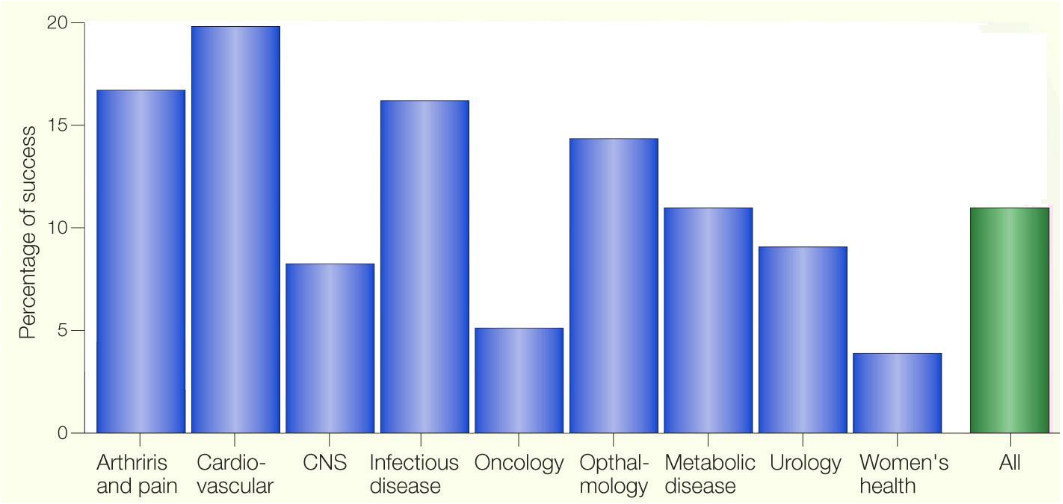


Figure 2. Success rates in clinical phases in different therapeutic categories. (After [15], modified)

The high failure rate of drug candidates [15] due to unexpected adverse reactions and lack of expected clinical efficacy have become fundamental problems of drug development. Generally, only 11% of the compounds entering to the clinical phase finish as a marketed drug (Figure 2). For some therapeutic areas, the rate is even worse (e.g. central nervous system agents and oncology). The most abundant reasons for attrition in 2000 are efficacy, toxicology and clinical safety issues (26 – 12 %). In 1991, more than 40% of the applications failed due to PK/bioavailability problems that were far less abundant in 2000 (less than 10%). This fact points to the improvements in PK property prediction in the decade while safety issues are still to be overcome. Of course, the more and more rigorous regulatory agency rules are also responsible for the different attrition reasons; however, it is clear that clinical safety must be the highest standard all time. The failure rate of drug candidates becomes more disturbing if we considered that only 5 out of 40,000 molecules tested in animals reach human testing [16].

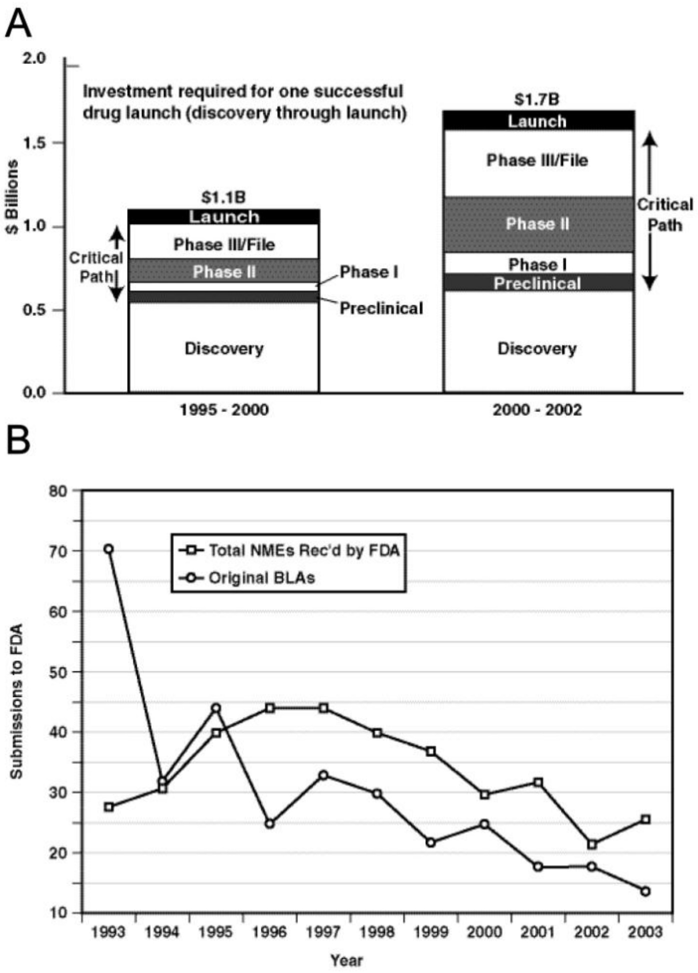


Figure 3A. Drug launch costs before and after 2000. Critical path is a term applied for clinical phases since financial needs and failure probability increases here rapidly. Panel **B** shows the number of drug applications submitted to FDA shows a decreasing tendency. NME, i.e., new medical entity is a compound with new molecular structure (instead of a modification of a previously existing drug). BLA stands for Biologics License Applications. [17]

Attached to this fact, drug development costs are continuously increasing: after 2000, the estimated investment needed per successful compound reached 1.7 billion dollars (from 1.1 billion dollars before 2000) [17] (Figure 3a). Despite the great efforts and resources spent on biomedical research, the number of new medical entities is decreasing year-by-year (Figure 3b). Moreover, a definite bias has been emerged in pharmaceutical industry from *de novo* drug discovery towards drug repositioning or repurposing, a much safer, cheaper and faster way of drug development that seeks for new therapeutic applications of existing drugs [4]. In this case, many safety and toxicity issues are known from previous tests thus repositioners have to prove the efficacy only. Considering that our knowledge is limited even for the well-studied drugs, it is not unlikely to find new uses for them. Some typical examples were sildenafil (repositioned from an antianginal agent to the treatment of male erectile dysfunction) and topiramate (from an antiepileptic agent to antiobesity) [4]. Finasteride was repositioned from the treatment of prostate enlargement to an anti-baldness agent after the discovery that its target, 5α-reductase, is involved in these distinct processes [18]. The feared thalidomide, that once caused severe fetal defects by administering as an anti-emesis agent for pregnant women, was recently rehabilited as an antileprosy drug [19].

Drug repositioning generally needs 5-8 years from discovery to marketization [4] (Figure 1b). The risk is definitely smaller compared to *de novo* drug discovery especially when the

EC_{50} value for the new effect, i.e. the concentration of therapeutic applicability, is similar to the EC_{50} value of the already approved effect. Although playing a sure game seems economically more rewarding than the hazardous *de novo* development, focusing on repositioning highly reduces the chemical space for searching and, in a larger perspective, might result in an even greater decline in drug development.

Facing this alarming situation, the U. S. Food and Drug Administration (FDA) underlined the importance of the application of fast and cost-effective *in silico* approaches [17]. As FDA expressed, the wide-spread usage of *in silico* filtering/screening methods before *in vitro / in vivo* assays might decrease the number of failed drug candidates in clinical studies ("fail fast, fail cheap"). The urgent need of the improvement of efficiency and effectiveness has also been highlighted in the 2005 Pharma Report of PriceWaterhouseCoopers and by Eli Lilly at the Drug Discovery Technology Conference in Boston [16]. Some recent, unfortunate failures of approved drugs like Vioxx (rofecoxib) (www.drugrecalls.com/vioxx.html) [20] point to the deficiencies of the animal tests i.e. undesired effects and adverse reactions might remain hidden even in thoroughly designed animal studies. Moreover, some adverse reactions cannot be detected in animals, e.g. nausea, headache and cognitive impairment. *In silico* approaches based on information collected from human applications (clinical trials, post-market data and repositioning claims) can overcome this phenomenon which also strengthens their importance. Thus, computer-aided drug discovery and development (CADDD) is able to reduce time and cost requirements of different levels of drug development, along the whole length of the development pipeline [16, 17].

2. *In silico* pharmacology

In silico pharmacology is a rapidly growing area that uses biological and medical data from different sources in order to create computational models or simulations for predictions in medicine [21]. It uses computational power to streamline drug discovery and development process and can be applied throughout the whole drug development pipeline [16]. In this section, I present an overview of the different approaches in the field of *in silico* pharmacology, following the classification system published in the excellent work by Ekins *et al* [21, 22], among others.

2.1 Quantitative Structure-Activity Relationships

Quantitative Structure-Activity Relationships (QSARs) were the first initiatives in the field of *in silico* pharmacology. Generally, a QSAR is a linear mathematical model that sets up the correlation between a set of structural properties and a desired activity/toxicity/ADME profile, i.e., a PD/PK event. A typical QSAR model uses a training set of several dozen molecules which the tuning of the parameters are done for. Then, the derived QSAR is applied on a test compound set that must be similar to the training set in terms of structural properties. The reliability of the QSAR function can be calculated from this test set.

The most widely used QSARs are *descriptor-based methods*. Descriptors are numerical representations of chemical structure. One dimensional descriptors are the most straightforward ones, e.g. molecular weight, logP (water/octanol partition coefficient), refractivity, number of rotatable bonds, number of H donors and acceptors etc. In contrast with their simplicity, a number of good correlations were set up using such descriptors, i.e. the QSAR for the possibility of passing the blood-brain barrier, one of the most exiting questions in drug development (e.g. [23, 24]). 2D descriptors are based on the topology of the molecule and describe their two-dimensional topology, e.g. the Balaban index correlates with the branchedness of the compound. 3D descriptors are less common; they can be determined from the 3D structure of the molecule thus an aligned set of 3D structures of the applied compounds are needed for 3D-QSARs.

In some cases, rules are used in QSARs instead of structural descriptors. The *rule-based methods* are driven by a large set of available data on the studied biological activity, i.e. the possible biotransformation routes of drugs [25].

2.2 Virtual Ligand Screening

Virtual Ligand Screening (VLS) is the *in silico* adaptation of High Throughput Screening (HTS), the methodology that searches among an enormous number of structurally unrelated compounds for the desired activity *in vitro* or *in vivo*. HTS techniques yielded far worse performances than anticipated before [26] thus pharmaceutical industry quickly introduced VLS into the drug development pipeline when computational capacity enabled its involvement. To perform VLS, structural information is needed for the ligands and/or for the target of interest.

2.2.1 Ligand-based VLS strategies

Ligand-based VLS methods adopt the principle that structurally similar molecules should possess similar PD/PK properties [27]. Therefore, molecules in a database are scored based on their structural similarity to the known active ligands. Chemical structures are often coded as one-dimensional structural fingerprints for easier data handling (e.g. ChemAxon fingerprint, [28]).

A special case of ligand-based VLS is pharmacophore design. The pharmacophore is, according to the extensive and rigorous definition by IUPAC, "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or block) its biological response. A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept […] can be considered as the largest common denominator shared by a set of active molecules." [29] Pharmacophore design was used in many studies [30, 31].

2.2.2 Target-based VLS strategies

Target-based screening is one of the most wide-spread *in silico* pharmacology techniques. The structural information on the selected target protein needed to perform target-based screening might originate from different sources:

1. Crystal structure obtained by X-ray crystallography or NMR.
2. If the crystal structure is unavailable, the structure of a close relative of the protein of interest can be used. By homology modeling [32], a model of the protein of interest can be fitted to the peptide backbone of the template structure and the side chains can be added and minimized in a second step.

Actually, Protein Data Bank contains more than 63,000 different three-dimensional structures of proteins [33]. However, as of 2005, 34.5% of the enzyme structures in PDB correspond to only 34 proteins and there is a clear bias towards the soluble proteins, due to methodologic issues [34]. Unfortunately, many important pharmacological target proteins are membrane receptors that are uneasy to crystallize. Solving the 3D structure of β_2-adrenergic receptor was a significant breakthrough in 2007 [35]. Not long after, several other pharmacologically relevant GPCR structures have been determined [36, 37].

Once one has an experimentally determined or a modeled structure of the target protein, a library of smaller ligands can be screened against it. This screening involves two steps: (1) to find the best conformation of the ligand in the binding site, and (2) the calculation (or estimation) of its binding affinity. The first step, i.e. mapping of the conformational space is commonly reffered to as docking while the binding affinity evaluation is called scoring. A number of docking and/or scoring programs exist nowadays; for a summary of the most widely used softwares and functions see Tables 1 and 2, respectively. Conformational space screening can be done through matching algorithm, genetic algorithm, Monte Carlo search or incremental construction. E.g., the widely used AutoDock uses a Lamarckian genetic algorithm to produce "generations" of possible compound conformations and applies the survival of the fittest (i.e., the lowest energy conformation) rule for selection between the conformers. Torsional angle data are stored in a "gene" for each molecule and crossing-overs and mutations are allowed.

Software	**Method**	**References**
AutoDock4	Genetic Algorithm	[38]
eHITs	Incremental Construction	[39, 40]
Glide 4.0	Hierarchical filters and Monte Carlo	[41-43]
GOLD 3.1	Genetic Algorithm	[44, 45]
HADDOCK	Simulated Annealing	[46]
PatchDock	Shape complementarity	[47]

Table 1. Commonly used docking softwares. (From [48], modified.)

Scoring Function	**Software examples**	**Type**	**References**
ChemScore	GOLD	empirical	[49]
GlideScore	Glide	empirical	[41, 50]
X-SCORE	standalone	empirical/consensus	[51]
AutoDock	AutoDock	force field/empirical	[52]
GoldScore	GOLD, CScore	force field	[53]
DrugScore	standalone	knowledge-based	[54]

Table 2. Widespread scoring functions. (From [48], modified.)

Scoring functions (SFs) (Table 2) can be divided into three groups: empirical functions, force field (FF)-based SFs and knowledge-based SFs. *Empirical SFs*, like X-SCORE and ChemScore, usually contain terms estimating the van der Waals, hydrogen-bonding, electrostatic and hydrophobic interactions. Freezing of rotatable bonds during binding is also taken into consideration. *FF-based scoring functions* apply modified force fields originally developed for molecular dynamics. Some FF-based functions are extended with empirical terms like AutoDock4 in which an additional entropy term describes the entropy loss during ligand binding. Analysis of crystal structures of ligand-protein complexes led to the development of *knowledge-based scoring functions*. Based on the interaction types and occurrences found in crystal structures, a score is calculated for each possible pair of atom types.

Applicability and predictive force of scoring functions are subjects of continuous debate in the scientific community [48, 55]. Today, it is accepted that docking programs dock 70-80% of the ligands correctly [56]. We must mention here that the comparison of different docking/scoring methods is inherently difficult. The applied parameters have great importance on the results and the level of knowledge of the reviewing authors might be different in case of different programs (i.e., the fine-tuning of their own-developed program may be more efficient than that of the others). The commonly used rescoring, i.e. the application of different SFs to recalculate the binding free energy of a docked conformation, is misleading without prior local optimization of the ligand pose in the force field of the new functions.

Target-based screening techniques were successfully used in many studies, including Hetenyi’s work [57] in which a subclass-specific myosin inhibitor, blebbistatin was docked to the nucleotide binding site of myosin IIa using blind docking, i.e. the whole protein surface was handled as a putative binding site, without predefined constraints [58]. (Generally, only the active site of a protein is used in docking simulations.) Even so, the inhibitor found the correct binding surface and its calculated conformation matched with the experimentally determined one.

Nevertheless, docking has its backdrops. The method chosen for mapping the conformational space and the applied scoring function has a great influence on the reliability of the results. The calculation time needed for the ligand increases rapidly with its size and the number of rotatable bonds. Handling the flexibility of the ligand and the protein is also a problem to be considered, as well as the role of structural waters [58]. A few years ago, only ligands were allowed to possess flexibility while the protein was handled as rigid body (e.g., in AutoDock3). This assumption obviously reduced the reliability of the entropic part of the binding affinity estimation. In the latest softwares, flexibility is taken into consideration for both agents. The imperfectness of scoring functions has been disputed before. And, as always, the principle of *garbage-in, garbage-out* (GIGO) also exists: results can be only as good as the input data was. A crystal structure with a low resolution or a bad homology model can ruin the screening even before its beginning. If no reliable structural information on the target is available, one should consider applying different methods.

2.3 Virtual Affinity Profiling

Virtual Affinity Profiling (VAP), the newest branch of *in silico* pharmacology, assesses the pharmacological (PD/PK) profile of a compound by considering its interactions with a series of targets. VAP techniques intrinsically adopt the theory of polypharmacology which states that drugs generally act on multiple targets. Since polypharmacology is a topic of outstanding interest both in the context of the pharmaceutical industry and this work, it will be discussed in details in a following section.

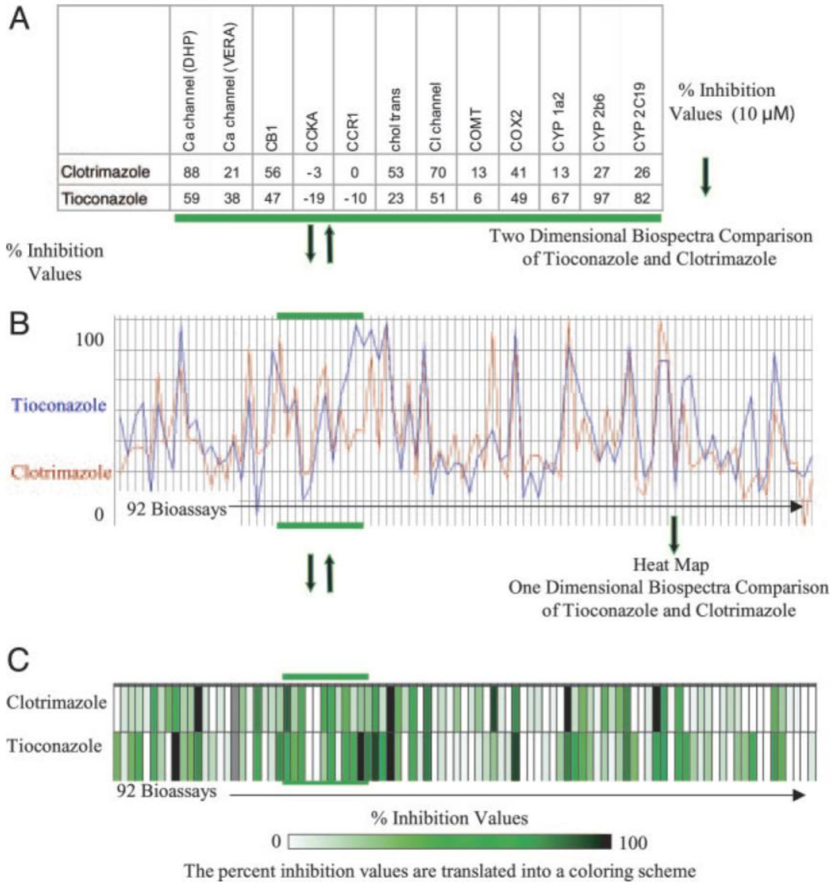
2.3.1 Ligand-based VAP strategies

One of the first initiatives in this field was PASS (Prediction of Activity Spectra for Substances) developed by Poroikov *et al* [59-61]. PASS applies the biological activity spectrum definition by Filimonov, i.e. the list of bioactivity names that are originated from the interaction of a compound and an organism [59]. Traces of SARs are observable in PASS method as it applies a set of 2D structural descriptors, called "multilevel neighborhoods of atoms", for cca. 250 000 compounds to correlate with more than 565 bioactivities [61]. The PASS training set contains 45 466 known biologically active compounds, retrieved by an extensive search in literature (as of March, 2002). PASS can be accessed online at http://195.178.207.233/PASS/ .

The main disadvantage of Poroikov's PASS system is the descriptor used: the relatively simple topological descriptor alone has low predictive power; even older QSAR studies involved more than one descriptor. On the other hand, PASS can be applied for estimating the whole bioactivity profile of a chemical compound which was seriously missing from all previous attempts.

The term “pharmacological profiling” is introduced in the work of Poulain *et al* [62]. More than 70 pharmacologically important receptors, including benzodiazepine, dopamine, serotonin and adrenergic receptors were selected and 48 compounds were *in vitro* screened against them. The obtained IC_{50} values were handled as an activity vector in “the space of pharmacological profiles”, instead of individual observations. SAR approach was adopted on these activity vectors, resulting in a more general structure-profile relationship. According to the findings, pharmacophore differences correlated well with the obtained pharmacological profiles.

A research group at Pfizer Global Research and Development introduced a technique called “biospectra analysis” in 2005 [63]. Here, a portion of the BioPrint database of Cerep [64] was used as a source of 92 *in vitro* binding assay results for 1,567 structurally diverse small-molecule compounds. This database contains percent inhibition values determined at single 10 μM ligand concentrations against a representative subset of the drugable proteome formed mainly by GPCRs, ion channels, kinases and proteinases from a number of protein superfamilies. Fliri *et al* introduced the term “biospectrum” that is the series of the *in vitro* percent inhibition values of a compound against the applied 92 proteins, handled as a continuum rather than a series of individual observations, an assumption similar to that of Poulain’s (Figure 4). The similarities between the 1,567 biospectra were assessed two different ways. The first method resulted in a similarity score compared to a reference compound. As an alternative method, hierarchical clustering was applied on the biospectra data set. Visual inspection of the generated clusters revealed a relation between biospectra similarity and molecular structure similarity between the compounds. When the biospectra of four new antifungal agents were added to the initial database, hierarchical clusterization resulted in a perfect linkage map in which all newly added molecules appeared in the cluster of molecules most similar to them [63]. This phenomenon was observed despite the target protein of many applied small molecules were absent from the protein set, including the above example of antifungals. In a later study, biospectra of dopamine agents were truncated as their respective targets were removed from the protein set and clusterization were repeated using six alternate protein set reduced in different ways [13]. The integrity of the resulting clusters did not change significantly which shows some robustness of the method; however, this robustness was never quantified nor fully discussed in their studies. It is mentioned that the applied high ligand concentration leads to “unspecific” binding, describing the binding potential of the molecule to the whole protein family instead of its used member solely.



	Ca channel (DHP)	Ca channel (VERA)	CB1	CCKA	CCR1	chol trans	Cl channel	COMT	COX2	CYP 1a2	CYP 2b6	CYP 2C19
Clotrimazole	88	21	56	-3	0	53	70	13	41	13	27	26
Tioconazole	59	38	47	-19	-10	23	51	6	49	67	97	82

Figure 4. The method of biospectra creation [63].

After connecting biospectra to chemical similarity, a relation to pharmacological effects has also been revealed in the case of dopamine agents [13]. Along this way, a relationship was determined between biospectra results of 872 medicines and their 240-bit-long binary side effect fingerprints [65]. A definite correlation (R=0.79) was observed between the two datasets.

In sum, biospectra describes (but not "identifies" as the authors state in [63]) the structure of small-molecule compounds: biospectra similarities can be translated into structural similarities. Moreover, the presence of the target proteins in the protein set is not necessary for efficient classification. Third, a definite relationship between biospectra and pharmacologic effects were revealed.

The largest intrinsic problem of the methodology is that biospectra are originated from a series of *in vitro* determined inhibition values. In a solution, drugs can interact with the whole surface of the receptor, possibly occupying different sites. As a consequence, the discriminator surface will not be uniform for the whole drug set, decreasing the comparability of the measured inhibition values. Moreover, only strong interactions can be detected by the applied *in vitro* binding assays. Weak interactions that might play an important role in the generation of drug effect profiles (see later) might remain hidden. Since a large portion of possible interactions cannot be detected, the robustness of the method is limited.

Another disadvantage is the applied hierarchical clusterization as it compresses the originally multidimensional diversity of biospectra into one dimension. Theoretically, 92 dimensions can be considered if the protein set is diverse enough to exclude dimensional interdependency. Principal component analysis (PCA) would reveal the underlying dimensional structure of biospectra data.

A third problem of biospectra approach is the weak cost-effectiveness and the incompatibility with high throughput screening techniques. Performing a series of *in vitro* binding assays for each candidate would have enormous work time and money requirements in drug development, reducing the method's applicability in the preclinical screening phases. (Again, a PCA might offer a possibility to reduce the number of tests needed to produce a biospectrum, thus decreasing the cost requirements.)

A VAP technique was presented by Hetenyi *et al* in 2003 in which 39 aromatic ligand molecules were docked to 31 different proteins using AutoDock3 docking motor with the in-built scoring function [66]. Comparison of the resulted protein-ligand complexes to the experimental results resulted in a good fit in 82% of the 1209 docking runs (RMSD less than 2 Å) and acceptable fit for the remaining 18% (between 2-3 Å). Calculated binding free energy values were transferred into a matrix in which ligands formed rows while energy values for different proteins were ordered next to each other. Then, Molecular Interaction Fingerprints (MIFs) for each molecule can be calculated as follows:

$$\mathrm{MIF}=\mathrm{E}-\mathrm{E}_{\mathrm{REF}}$$

where MIF stands for Molecular Interaction Fingerprint vector, E is the vector of calculated binding free energies on the protein set and $\mathrm{E}_{\mathrm{REF}}$ is the vector composed by experimental reference energies obtained from the protein's complexes with their original ligands, respectively. MIFs form the row vectors of the matrix shown in Figure 5 and represent the interaction properties of the ligands to a set of proteins. Similarly, Molecular Affinity Fingerprints (MAFs) can be defined for the column vectors of the matrix. (Individual columns can be considered as *in silico* target-based screenings while inspections involving the whole set of columns belong to the next group to be discussed, the *target-based* VAP methods.) As the authors proved, MIFs and MAFs are in agreement with fragment and similarity considerations.

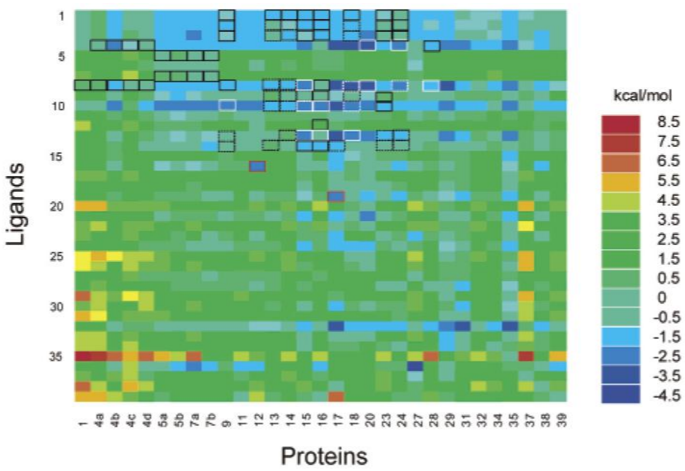
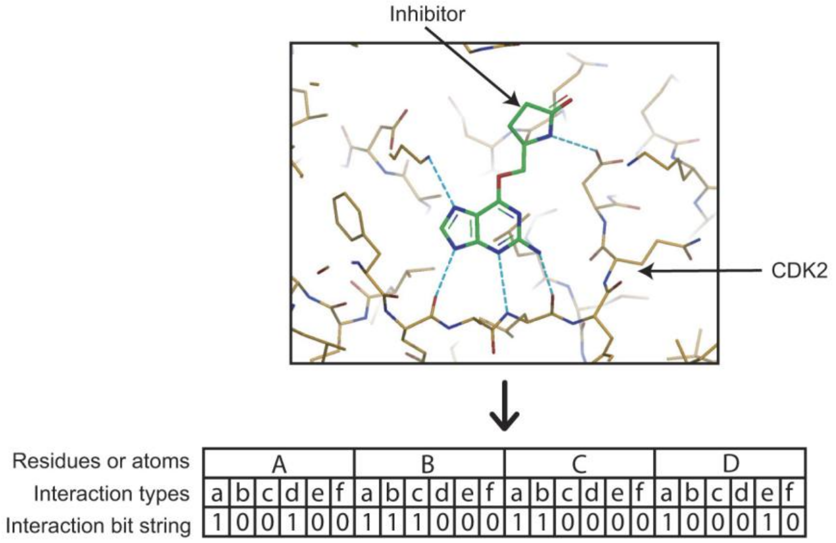


Figure 5. MIF matrix [66].

Although this study does not involve pharmacological information, its possibility is discussed: "MIFs and MAFs calculated on the basis of a larger free energy database may aid the exploration of the appropriate biochemical interaction route of any compounds, e.g. by automated comparison of their fingerprints with those of ligands of elucidated biochemistry." [66] Compared to the works by Poulain and Fliri, the uniform handling of interaction surfaces is secured by using pre-defined docking boxes in which conformational mapping and rotation of the ligand can occur while the other parts of the protein are excluded from the calculations. Being a pure *in silico* study, its cost-effectiveness is undoubtable; however, the limitations of docking and scoring mentioned above might be considered.

2.3.2 Target-based VAP strategies

The Structural Interaction Fingerprint (SIFt) method is a good example of target-based VAPs [67, 68]. SIFt is a fingerprint representation of seven types of the possible interactions for each residue in the binding site of a protein. The seven bits refer to a contact with the ligand, a main chain contact, a side chain contact, a polar interaction, a non-polar interaction, and the possibility of the residue being a hydrogen bond donor or acceptor, respectively (Figure 6). A SIFt therefore represents a single protein-ligand complex.



Residues or atoms	A						B						C						D					
Interaction types	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f
Interaction bit string	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	1	0

Figure 6. Generation of Structural Interaction Fingerprint in the case of CDK2 and its inhibitor. [67]

An improvement of the method, called pSIFt (profile-SIFt) involves information on more ligands bound to the same protein and applies weighted values for each bit [67]. SIFts and pSIFts can then be clusterized based on Tanimoto similarity scores (see Glossary) in order to compare and cluster the docked poses of ligands for visualization. Or, fingerprints can be scored according to their similarity to fingerprints representing experimentally determined complexes, in order to dissect active compounds from inactives.

Chen *et al* developed a technique called inverse docking where a ligand is screened against a large set of protein cavities, in order to reveal potential multiple targets [69]. Although the method is extremely resource-demanding, Li *et al* set up a web server for inverse docking, offering a set of binding sites of 700 proteins [70].

2.4 *In silico* pharmacology: a short summary

The choice between different applications of *in silico* pharmacology is basically driven by the level of information available on the desired pharmacological effect. If a target protein can be assigned and its structure is solved, target-based VLS can be applied. If only the structure of a close relative exists, target-based screening might be performed after homology modeling. In case of no structural information or unknown target protein, ligand-based VLS (or QSAR) might be a good choice if a training set can be built up from a sufficient number of active molecules. If a compound possessing a complex profile of desirable effects is targeted, e.g. in the case of complex diseases, VAP strategies should be chosen.

3. Polypharmacology

Ehrich’s therapeutic goal that a drug generally possesses its actions by modifying the functions of a single protein [71] defined the vision of pharmaceutical industry for decades. This reductionist approach, i.e., the creation of target-specific “magic bullets” resulted in the increasing attrition rates as discussed earlier. As it becomes clearer and clearer nowadays, the conception of magic bullets does not stand anymore. Recent evidences showed that many drugs act on multiple proteins. E.g., benzodiazepines not only affect GABA-ergic ion channels but mitochondrial receptors as well [72]. Mechanism of action of methadone involves the GPCR-type μ-opioid receptor and the ion channel NMDA receptor as well [73]. Serotonergic drugs can bind to 5-HT receptor subtypes 1,2 and 4-7 (GPCRs) as well as the ion channel 5-HT3A [11, 74]. In these cases, no sequence or structural similarity can be found between the shared targets of a compound. The promiscuous binding might be involved in therapeutic effect or might result in off-target activities leading to side effects. Roth *et al* reported that selectively non-selective (sic) drugs might prove higher efficacy in treatment of complex CNS diseases that single-target acting drugs [11]. Applying their terminology, multiple-acting “magic shotguns” like clozapine are more effective in treating schizophrenia than other agents developed in order to target the receptor that is responsible for the beneficent effects of clozapine [75, 76]. A similar observation was made in the case of antidepressants [11].

One of the most recent success stories is imatinib mesylate (Gleevec), an anticancer agent developed as a “magic bullet”, acting only on a single protein. However, it was later discovered that it also affects platelet-derived growth factor and c-kit, two other proteins involved in the drug’s high efficacy, turning it to a “magic shotgun” [7].

Valproic acid also shows high promiscuity: it affects GSK3 kinase, histone deacetylase HDAC1, GABA tranaminase prolyl oligopeptidase [77] and cyclooxygenase (COX) [78]. This small-molecule drug can therefore be applied for different diseases including epilepsy, bipolar disorder, tumors and Alzheimer’s disease. Based on its effect on HDAC1 and the fact that this protein is involved in HIV infection, valproic acid has been tested as an adjuvant to the highly active anti-retroviral therapy with great success [77].

The fact that most of the drugs affect multiple targets rather than acting on one single target, is commonly referred to as *polypharmacology* [8, 10]. As Mestres *et al* stated, a drug interacts with 6 targets in average [79]. E.g., the binding profile of some antipsychotics can be found on Figure 7 [11]. Despite the novelty of polypharmacology, the topic is highly addressed and widely discussed in scientific community. Polypharmacology, in contrast with previous approaches, brings a holistic view into drug development: drugs of the future should affect protein networks instead of focusing on single proteins. “Magic bullets” should be replaced by “magic shotguns”. According to mathematical systems theory mentioned earlier, the problem and a solution should have the same level of complexity. Consequently, complex diseases like mental illnesses and age-related issues could be cured only applying complex approaches like polypharmacology. The theory of polypharmacology has been successfully adopted for many different complex diseases already. Bolognesi *et al* reported memoquin, a multitarget-type compound against Alzheimer’s disease [80, 81]. Apsel *et al* developed dual inhibitors of tyrosine and phosphoinositide kinases [82]. Nevertheless, polypharmacology is rarely used in pharmaceutical industry; probably due to its novelty and the obvious difficulties in data evaluation [7]. An important and already applied consequence of polypharmacology is that it turned the attention of drug companies to drug repositioning by highlighting potential off-target effects of existing drugs as it was demonstrated in the case of valproic acid.

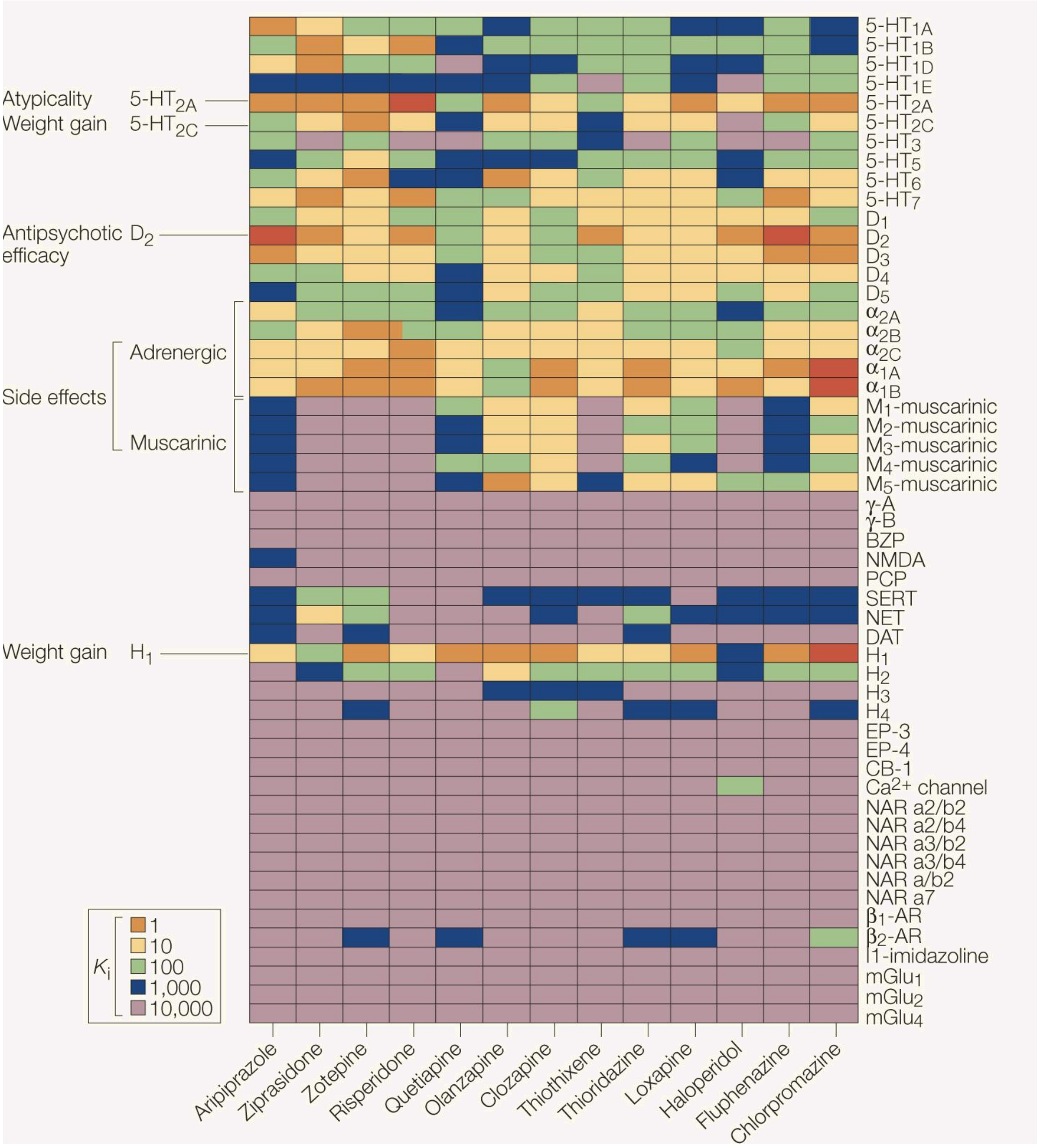


Figure 7. Promiscuous interaction profiles of certain CNS drugs. [11]

Applying polypharmacology, certain semi-blind approaches have been developed that show different levels of ability to systematically screen drugs for hidden therapeutic benefits or side effects.

Campillos *et al* reported a method based on connecting drug side effect information with the probability of sharing a target [12]. Their methodology can be found in Figure 8. Shortly, side effect information was derived from public drug labels. Then, a standardizing was performed to exclude synonym pairs and a weighing scheme was applied in order to eliminate the biasing effect of side effect correlations (i.e., nausea is often associated with vomiting).

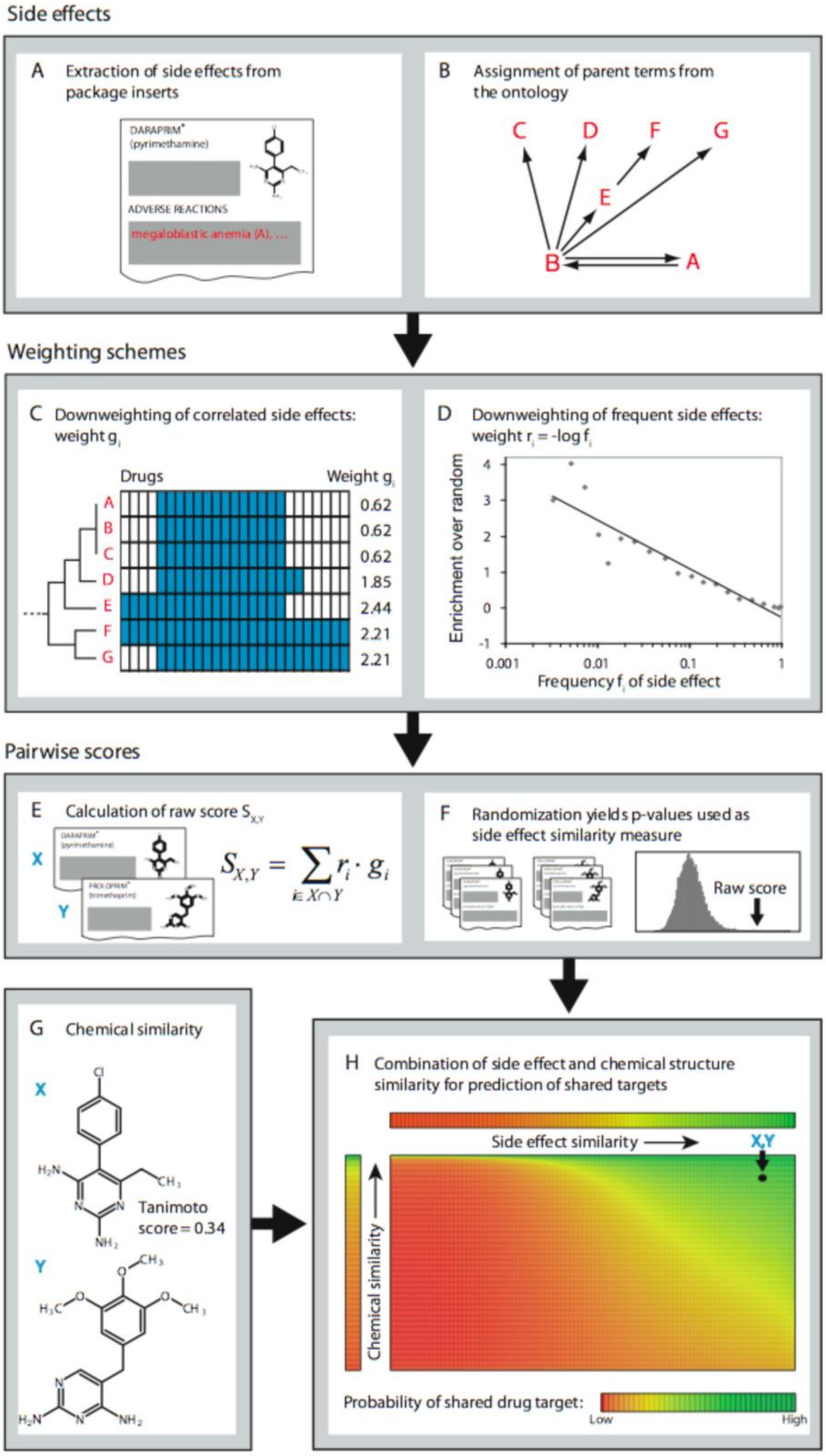

Side effects
A Extraction of side effects from package inserts
DARAPRIM®
(pyrimethamine)
ADVERSE REACTIONS
megaloblastic anemia (A), ...
B Assignment of parent terms from the ontology
C
D
F
G
E
B
A
Weighting schemes
C Downweighting of correlated side effects: weight g_i
Drugs
Weight g_i
A 0.62
B 0.62
C 0.62
D 1.85
E 2.44
F 2.21
G 2.21
D Downweighting of frequent side effects: weight r_i = -log f_i
Enrichment over random
4
3
2
1
0
-1
0.001
0.01
0.1
1
Frequency f_i of side effect
Pairwise scores
E Calculation of raw score $S_{X,Y}$
X
Y
DARAPRIM®
(pyrimethamine)
PROLOPRIM®
(trimethoprim)
$S_{X,Y} = \sum_{i \in X \cap Y} r_i \cdot g_i$
F Randomization yields p-values used as side effect similarity measure
Raw score
G Chemical similarity
X
Y
Tanimoto score = 0.34
H Combination of side effect and chemical structure similarity for prediction of shared targets
Side effect similarity
X,Y
Chemical similarity
Probability of shared drug target:
Low
High

Figure 8. Target prediction method developed by Campillos *et al*. [12]

In the next step, a side effect similarity score was calculated between each drug pair, compared to a set containing random side effects. Finally, a chemical similarity score was defined for each pair and the two similarity measurements, i.e. the side effect and chemical similarity, were combined into one probability value that describes the possibility of sharing a target between two given drugs. The functions were tested on a training set collected from Matador, DrugBank and PDSP K_i databases, containing 502 FDA-approved drugs with 4,857 known associated drug-target interactions. A clear-cut correlation was found between side effect similarity and the probability of sharing a target (Figure 9a) while a smaller level of correlation was detected for 2D Tanimoto similarity data (Figure 9b). Applying the combined function on the whole data set, 2,903 drug pairs were identified with at least 25% probability of sharing a target. After filtering out the known issues and pairs showing similar chemical structures or similar targets, 754 non-obvious drug pairs were obtained. 20 drug pairs were tested experimentally and 13 of them were validated with K_i values generally lower than 10μM. Nine predictions were also confirmed in cell assays. The found relations might have therapeutic impact, i.e. the nootropic donepezil was found to possess serotonergic activity, similarly to the antidepressant venlafaxine. Indeed, donepezil has already been proposed for testing in the treatment of depression.

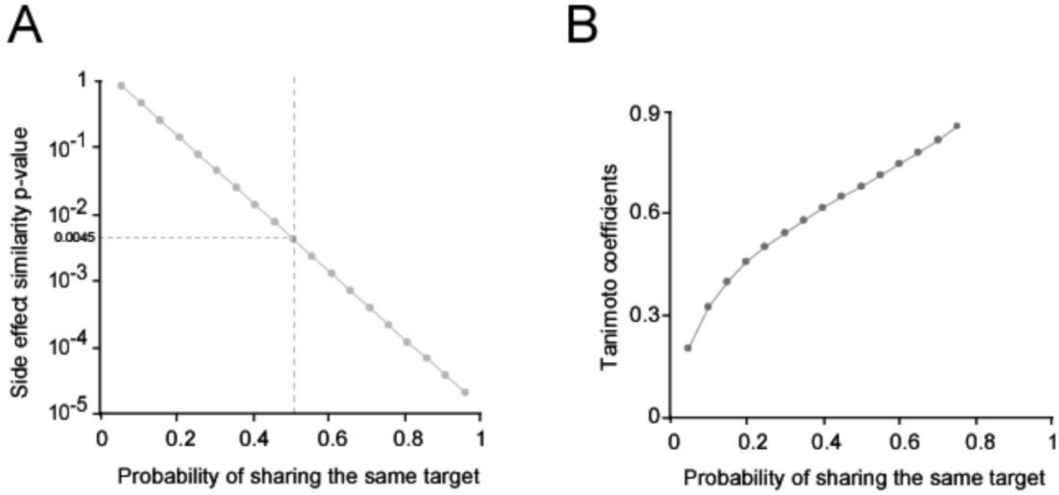
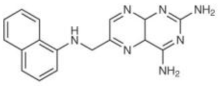
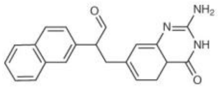
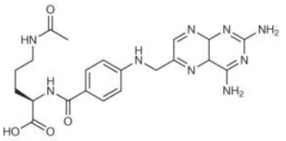
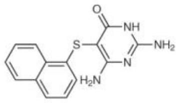


Figure 9. Target sharing probability vs. side effect similarity **(a)** and Tanimoto chemical similarity **(b)**. [12]

The authors successfully overcame the noisy nature of side effect data, thus resulting in good correlations and high predictive power. (The research group recently reported the establishment of a public online side effect database, SIDER. [83]) However, target similarity not always refers to side effect similarity, e.g. if a drug can pass the blood-brain barrier but the other cannot. Moreover, a shared target not necessarily results in shared side effects due to the possibly antagonistic effect of the other targets that masks the presence of a common target protein. Applying drug effect information instead of side effect data might produce even stronger correlations.

Keiser *et al* presented a systematic prediction method that uses ligand chemical similarity in order to clusterize 246 enzymes and receptors [14]. Protein-ligand affinity data were obtained from MDL Drug Data Report (MDDR), containing ligand sets for 246 proteins, 65,241 compounds in sum. The whole similarity half-matrix (65,241*65,241/2) was calculated using Daylight fingerprint and Tanimoto similarity coefficients. (Average similarity values were determined for the 246 ligand sets as well.) A similarity measurement called Single Ensemble Approach (SEA) was introduced to calculate an expectance value (E-value), i.e. the level of identity of the ligand sets of two activity classes, i.e. receptors: the lower the E-value between two receptors, the higher the similarity of their respective ligand sets (Figure 10a).

Rank	Activity class	E-value	Example molecule
1	DHFR inhibitor	7.07×10^{-182}	
2	Glycinamide ribonucleotide formyltransferase inhibitor	3.97×10^{-100}	
3	Folylpolyglutamate synthetase inhibitor	4.59×10^{-62}	
4	TS inhibitor	1.11×10^{-61}	

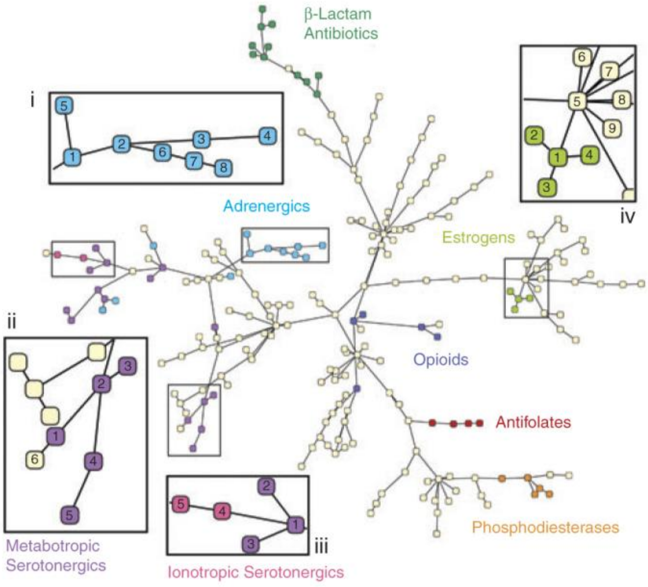


Figure 10. Panel **A** shows the activity classes resembling to dihydrofolate reductase inhibitor (DHFR) class. Panel **B**: a slice of the pharmacological space of proteins based on E-values [14].

Based on the E-values for each activity class pairs, a minimal spanning tree was formed. Although the authors used no direct biological information on the enzymes and receptors, biologically relevant classification was obtained in which several functional/pharmacological groups were separated to individual branches (Figure 10b). This method can be used for activity prediction by calculating the chemical similarity values of the query set (one molecule or a set of compounds) to the 246 representative ligand sets for the 246 studied activity classes, resulting in E-values for the probability of possessing every activity classes. In a later work, authors involved FDA-approved and investigational drugs and predicted 6,928 associations between drugs and targets [84]. After filtering out the trivialities, 3,832 predictions remained. 184 of them were selected for further investigations; 42 of them turned out to be known associations while 30 predicted interactions were tested experimentally. 23 of them was confirmed, five of them with K_i values lower than 100 nM. The physiological relevance of one prediction was confirmed in knock-out mouse. Surprisingly, many new cross-boundary targets were identified and validated. E.g., Delavirdine mesylate, a HIV-1 reverse transcriptase enzyme inhibitor turned to be a histamine H_4 receptor, a GPCR antagonist as well; or NMDAR-agent ifenprodil (ion channel) also affects a GPCR, the α_2 adrenergic receptor [84].

The significance of the method reported by Keiser *et al* relies in the fact that 2D chemical similarity values of ligands were related with common targets, bridging atomic level information of ligands to binding affinity profiles.

Another way of polypharmacology-driven drug development could be called “polypharmacophore” approach. Here, two distinct pharmacophores are linked together by a conjugate linker, or common, overlapping or highly integrated pharmacophores are applied [85].

A special application of profile-based drug development is the administration of *drug combinations*. The combination of amoxicillin and clavulanic acid (marketed in Hungary as Aktil Duo) is a widespread example. Amoxicillin inhibits cell wall synthesis while clavulanate inhibits β-lactamase, the enzyme responsible for the biotransformation and elimination of amoxicillin. Clavulanate maintains the concentration of amoxicillin in the cell wall, producing a highly efficient antibacterial drug combination. Although pharmaceutical industry tends to avoid drug combinations due to the presumably higher level of their side effects, this assumption is not necessarily true. If the simultaneously administered drugs act synergistically, the doses of the individual compounds can be dramatically reduced, resulting in larger tolerability [86]. However, PK/PD prediction of administration of drug combinations have not been overcame thus far.

4. Protein binding site description

A recent publication by Milletti and Vulpetti points to the application of protein binding site descriptors in polypharmacology prediction [87]. The authors developed a new method for binding site description, based on the shape-context-based descriptors presented by Belongie *et al*. First, they apply FlapSite for pocket detection and the found sites are extracted into a PDB file. Then, the original atom types are converted to relay similarities, e.g. “Hyd” is introduced to represent all hydrophobic atoms (C, aromatic C, S). Side chain flexibility is also handled by linking rotatable groups into a backbone carbon. Finally, 14 concentric spherical layers are defined around each atom and the occurrence of the previously defined atom types in these 14 neighborhood zones is collected into fingerprints. After the 3D alignment of binding sites, a similarity score is calculated. Based on binding site similarity, promiscuity was predicted for 17 inhibitors against 189 kinases for which *in vitro* inhibition values have been determined previously. Receiver operator characteristic (ROC, see Glossary) analysis was used to evaluate the results. This process will be presented in Methods in detail; here we mention that it is based on the rate of true positives (TPR) and false positives (FPR). A positive hit predicted by the method is accepted as “true positive” if it inhibits the given kinase with a Kd < 10 μM. The presented approach resulted in an average AUC (see Glossary) of 0.64 which belongs to a medium accuracy range and is comparable to two ligand-based chemical similarity measurements.

This work and the aforementioned target-based VAP methods point to the application of protein site information in polypharmacology predictions. Binding sites can be characterized in a numerous ways, involving topological and/or chemical information: CavBase, SiteEngine, IsoCleft, PocketPicker etc. E.g., CavBase defines donor, acceptor, aromatic, aliphatic etc. centers in the binding site, based on the amino acid constitution of the pocket [88]. PocketPicker uses only geometric information on pocket description and uses atomic level information instead of amino acid composition [89]. Although pure geometric similarity might be considered as an oversimplification, numerous studies exist that point to the efficiency of shape-based descriptors in different fields of *in silico* drug development [90]. For example, finding complementary shapes for the active site of a drugable protein is a starting point of *de novo* drug design if the target structure is previously determined [91]. Fragment positioning and molecule growth methods, together with fragment searches in cheminformatics databases produce the primary hits. (These results are evaluated further by scoring functions that consider more parameters for a better prediction of ligand-binding properties.) A method called Shape signatures describes ligand and protein binding site shapes using ray-tracing algorithm, producing one-dimensional histograms for ray-trace segment lengths [92]. Zauhar *et al* demonstrated the suitability of this method in finding shape similarities among small-molecule ligands for estrogen and serotonin receptors. Shape-based approaches have an important role in the simulation of protein-protein interactions. For example, Venkatraman *et al* reports on the development of a docking algorithm based on 3D Zernike Descriptors (i.e., 3D function representations of protein surface) that produced outstanding performance compared to other methods [93].

5. Starting hypothesis and applied methodology

Three levels of information can be distinguished in pharmacology:
Level 1: atomic-level information: chemical structure and physicochemical properties of the ligand and/or protein;
Level 2: binding affinity information: known activity values of ligands on different proteins;
Level 3: bioactivity: PK/PD, effect and side effect information.

The actions of drugs can be translated on each level. If an administered drug enters the human body, it will form strong and weak interactions with the members of the proteome, thus selecting, discriminating among them. On the other hand, the proteome also discriminates among the xenobiotics entered. The result of this two-sided interaction network is a binding pattern. This binding pattern is projected at the organism level as a bioactivity pattern, i.e., effects and side effects. From this point of view, there is no difference between effects and side effects: a desired bioactivity is referred to as “effect” while an undesired activity is grouped into the category of “side effects”. As presented before, these are loose categories, often being perturbed in drug repositioning.

The primary aim of this work is the development of a systematic *in silico* prediction method for effect prediction of existing drug molecules (and drug candidates) that is able to uncover the whole effect profile of compounds. After the careful reviewing of *in silico* pharmacology, VAP methods were selected, adopting the paradigm of polypharmacology.

Campillos *et al* proved the connection between Levels 2 and 3 while Keiser's group related Levels 1 and 2. Fliri *et al* demonstrated the connection between Levels 1 and 3 but biospectra method is inconvenient and hardly introducible into the drug development pipeline. However, the finding that similar binding pattern refers to similar bioactivities [63, 65], might be applied to *in silico* affinity profiles as well. Therefore, in this work, atomic level information of small-molecule drugs will be related to drug effect profiles, based on the milestone work by Hetenyi *et al* [66]. Since the calculation method is slightly different from the one presented in [66], the term Interaction Profile (IP) will be used henceforth in the text instead of MIF. The advantages of the docking-based IP approach are:

1. Generation of the calculated binding free energies is relatively fast and simple, only structural information is needed from the ligands.
2. Uniform treatment of interactions is secured by the uniform definition of docking boxes; consequently, the same discriminator surface can be applied throughout the ligand set. No target proteins are required for bioactivity prediction, as presented earlier, e.g. [13, 63].
3. According to our principal hypothesis that an interaction pattern refers to bioactivity pattern, a random set of proteins can be applied to the calculation of interaction patterns.
4. The hypothesis is possible to test by determining the level of correlation between the *in silico* interaction profiles and the effect profiles of drugs.
5. With the generation of a MIF-like IP matrix, MAFs are also created. Thus, the important factors in the determination of binding affinity e.g. protein site geometry can be studied on the same interaction profile data set.

Based on many studies discussed before, ligand chemical similarity calculation is a reasonable way to classify pharmacologically targeted proteins and *vice versa*. Therefore, after generating interaction profiles for the drugs, it is reasonable to change our point of view from drugs to proteins, i.e., from IPs/MIFs to MAFs, according to the terminology introduced by *Hetenyi* et al. By evaluating the protein-ligand interaction matrix from the proteins' direction, one can determine the factors that play an important role in the determination of binding affinities. Despite the promising results pointing to the possibility of biologically meaningful clusterings along shape-based and affinity fingerprinting investigations, the connection between the affinity profiles and the structural characteristics of protein binding sites still remains unclear. Thus, the secondary goal of this study is to investigate the relationship between virtual drug screening results (calculated binding free energy values) and the shape of protein binding sites. For this purpose, PocketPicker algorithm was chosen for geometric description since it is based on atomic level information which is crucial because the software will be applied to calculate similarities in docking boxes that might contain partial residues at the box perimeter. Moreover, it produces an easily comparable fingerprint to describe binding sites.

5.1 Risk analysis of the initial hypothesis

The chosen methodology has several risks as well:

1. There is no strict agreement in the scientific community about the reliability of scoring functions using during docking therefore at least two different scoring functions must be applied to ensure that our findings are not artifacts originated from the scoring. However, we mention that the purpose of the scoring function in this study is to quantify an interaction between large sets of ligands and proteins in a uniform manner.
2. Although we presume the opposite, it might turn out that the composition of the protein set is of crucial importance. This would ruin the concept of the application of *in silico* binding affinity patterns in bioactivity predictions. If target proteins are needed for sufficient predictive power, our method cannot be considered as a ligand-based VAP approach.

Methods

1. Development of the Interaction Profile (IP) database

1.1 Data collection

1,255 FDA-approved drug molecules were extracted from DrugBank database [94] as of June, 2009 (Appendix 1). A two-step selection was applied: molecules labeled "FDA-approved small molecule drug" were separated first (969 entries) and used for preliminary evaluations. This list was extended later with "FDA approved drugs" below the molecule size limit of 600 Da (286 entries). 160 proteins were collected from RCSB Protein Data Bank [33] which met the following requirements:

(1) structure contained ligand,

(2) resolution better than 2.3 Å,

(3) complete ligand binding site,

(4) primary structure was not significantly different from the wild type protein's structure.

If a structure contained water molecules involved in ligand coordination, its conformation was compared with available structures without water. If no significant difference was observed around the ligand binding site, the one with better resolution was used. (See Appendix 2 for the list of the PDB codes of the applied proteins.)

1.2 Docking preparations

Docking preparations and calculations were performed by AutoDockTools [95] and DOVIS 2.0 (DOcking-based VIrtual Screening) [96] softwares in case of one and multidimensional analyses, respectively (see later), using AutoDock3 (in case of one-dimensional analyses) and AutoDock4 docking engines. AutoDock3.0 (for one-dimensional analyses), Autodock4.0 and X-SCORE scoring functions were applied [38, 51, 66, 96, 97]. Explicit hydrogens were added to the drug molecules and optimization procedures were applied for aromatic rings and for the overall 3D structure before docking using AutoDockTools and ChemAxon JChem Base softwares (version 5.2.0, 2008) [28, 95]. AutoDock enables the definition of a box in which docking calculations are carried out. The docking box was centered to the geometrical center of the original ligand of the protein (as found in the intact PDB file); box size and grid spacing were set to 22.5 Å and 0.375 Å (default value), respectively. Protein parts outside the box were excluded from the calculations. The applied box size enables each member of the drug set to rotate freely in order to find the conformation with the lowest binding free energy without steric clashing to the box perimeter. For consistency, no further reductions in box size were applied to smaller ligands and the same box was used for geometric characterization of the binding site as well.

1.3 Docking

Each drug molecule was docked to each protein, performing 25 runs each. Binding free energies were extracted and the minima were imported to a database. Docking runs were performed on a Hewlett-Packard cluster of 104 CPUs.

Three different scoring functions were applied throughout this study. AutoDock3 was used for initial one-dimansional evaluation while AutoDock4 and X-SCORE were applied for multidimensional investigations.

For better reliability, redocking was performed instead of rescoring the previously docked conformations. Thus, three binding free energy values have been determined. First, AutoDock3 was used for an initial screening of a set of 969 FDA-approved drug molecules against 89 proteins. Then, the protein set was extended to 160 and the docking procedure was repeated applying AutoDock 4 for mapping the conformational space, using either its own scoring function or X-SCORE function, on a set consisting of 1,255 compounds. Thus, the impact of different scoring functions on the results can be assessed. In the initial phase, 969*89=86,241 dockings were performed, repeated 25 times for each drug-protein pair, docked and scored by AutoDock3. Later, 1,255*160=200,800 dockings were carried out that means 200,800*25=5,020,000 individual docking runs both for AutoDock4 and X-SCORE scoring. Lowest binding free energy values for each drug-protein pair were extracted and the minima were imported to databases (i.e., AutoDock3, AutoDock4 and X-SCORE-based results).

1.4 Filtering, normalization and centralization

Six binding pockets out of 160 produced outlying binding free energy values throughout the applied drug set. Visual inspection revealed that the grid center was misplaced in three of them (1hdo, 1mv9, 1uwh), resulting in no or only few atoms in the docking box. On the other hand, three binding pockets proved to be too narrow to dock the full drug set (1ryo, 2ibn, and 1mlw). These six proteins have been removed from the protein set and calculated binding energy values on the remaining 154 proteins were applied in further investigations. No drugs were excluded due to outlying docking results.

Normalization and centralization was performed on the IP datasets in order to transform the data to a common statistical scale, thereby ensuring that the underlying data vectors reflect the molecular interaction profiles instead of the scale parameters (such as the mean and the standard deviation) that are more sensitive to measurement errors and outlying observations. Before effect prediction, the following normalization/centralization procedure was carried out: drugs were considered as cases (1,255 rows) while proteins as variables (154 columns). Normalization and centralization were performed row-by-row for each drug as follows:

$$energy' = \frac{energy - mean}{SD}$$

Where *mean* is the mean and *SD* is the standard deviation of the docking energies for a given drug.

For the analysis of the relation between MAF data and the binding site geometry descriptor set, drugs were considered as variables (1,255 columns) and proteins as cases (154 rows). Normalization and centralization were done row-by-row for each protein.

2. Generation of the Effect Profile (EP) matrix

As mentioned above, structural and pharmacological information on 1,255 FDA-approved small-molecule drugs were extracted from DrugBank database [94]. This effect list was applied to perform a first, initial evaluation of the relationship between AutoDock3-based binding affinity and bioactivity data. Then, a list of 559 effects was formed that contains all effect entries that appeared on the drug information. Effect entries were further refined in order to eliminate initial database inconsistencies (e.g. “GABA agent” was not registered to every benzodiazepine). Structural categories often showed incompleteness, e.g. not every phenothiazine was registered as so. Since effect categories with less than 10 registered drugs contain insufficient amount of information for meaningful classification, the effect list was reduced to 181 categories. Then, a binary matrix was formed that shows the presence or absence of the studied 181 effects for each drug. Here, the appearance of an effect for a drug is marked with a "1" value and *vice versa*.

A preliminary side effect database was formed using adverse reactions data published on the official FDA labels of drugs. FDA labels were collected from www.rxlist.com. Database inconsistencies were cleared in this database as well (i.e., vomiting/emesis). This database was used for the evaluation of certain associations between MIF database and adverse reactions.

3. Generation of PocketPicker shape descriptor matrix

In order to analyze the relationship between the MAFs of the proteins and the geometry of their binding sites, we used the PocketPicker algorithm [89] to generate 420-dimensional fingerprints representing the geometrical features of the binding sites. Originally, the algorithm considers the areas of the entire protein located closely to the protein surface. This is in contrast to the docking process which aims to find the best fit of a ligand in a well defined area of the protein, i.e., in the docking box. Consequently, applying the PocketPicker algorithm on the original protein structure might lead to the detection of binding sites outside of the docking box. To ensure that the same set of atoms is involved in the MAF matrix generation and the PocketPicker description, the atoms of the given protein enclosed by the docking box defined above were extracted while preserving their original spatial coordinates. PocketPicker algorithm was applied to this set of atoms. This process assures that the PocketPicker algorithm characterizes the geometrical features of the docking box only.

In the *first step* of the process of PocketPicker fingerprint generation, the degree of buriedness of the different areas of the docking box is determined, which in turn provides information on how accessible that particular area is. A rectangular grid with 1Å mesh size is generated around the protein; each point of this grid is described as a grid probe. Over the process of scanning it is determined how many atoms are located in the surroundings of each grid probe. This is achieved by placing on each grid probe 30 so-called search rays that are distributed in a closely equidistant manner on a sphere. Each search ray is 10 Å long and has a width of 0.9 Å. The buriedness value *Bu(j)*, of the given grid probe *j* is the number of search rays that hit at least one atom. Grid probes of buriedness value in the range of 15 and 26 are recorded and classified into the following six categories: (1) category A: *Bu(j)* = 15-16, (2) category B: *Bu(j)* = 17-18, (3) category C: *Bu(j)* = 19-20, (4) category D: *Bu(j)* = 21-22, (5) category E: *Bu(j)* = 23-24, (6) category F: *Bu(j)* = 25-26.

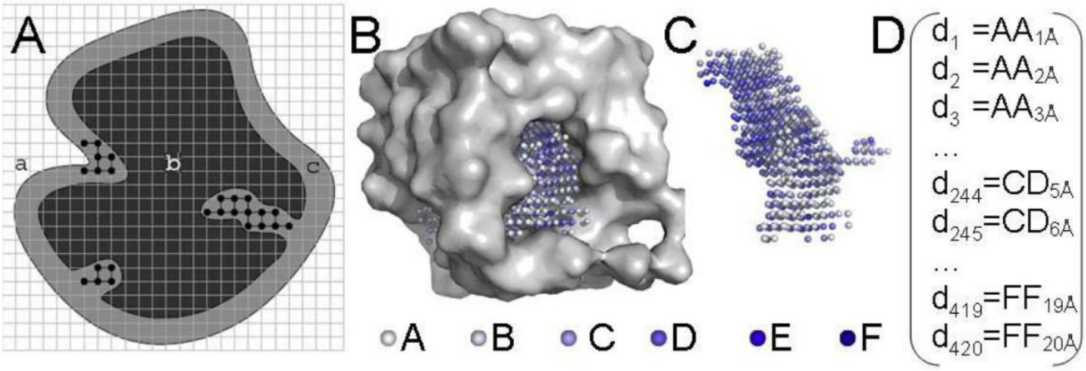


Figure 11. PocketPicker fingerprint generation. **A)** Pocket detection: a rectangular grid is generated around the protein. Grid points within the protein (black area, b) and grid points far from the protein (white area, a) are automatically excluded. Pockets consist of grid points around the protein with proper buriedness (*Bu(j)*>14). (modified based on [89]) **B)** An example of a pocket represented by small spheres colored according to their buriedness values inside a docking box (grey surface, constructed from the PDB structure 1zid) **C)** The same pocket without the protein atoms. **D)** 420 dimensional vectors are constructed representing the number of grid point pairs belonging to a given buriedness category and distance. (E.g.: the 244th dimension contains the number of cases where a grid point with buriedness category C is 5 Å away from a grid point of buriedness category D.) (Rauscher)

The PocketPicker algorithm characterizes the geometrical features of binding sites on the basis of the distribution of the distances between grid points of each buriedness category. Therefore, in the *second step* it is counted how many grid probes of the different buriedness categories can be found in a distance of 1 – 20 Å from each grid probe. Considering that there are 21 possible combinations of the six buriedness categories (e.g. A-A, A-B, A-C ... F-F), and that the distances are divided into 20 bins covering ranges of 1-20 Å, there are 21 * 20 = 420 possibilities to record the distance between a pair of grid probes of the same or different buriedness types. These possibilities give rise to the 420 components of the PocketPicker fingerprints (Figure 11). Therefore, the value of the coordinate of each component provides information on how many times it is observed that two grid points of particular buriedness types are located within a given distance from each other. The buriedness types of these two grid probes and the distance between them are exactly defined by the given component of the fingerprint.

In summary, the geometrical features covering the shape of the binding site are given by the spatial distribution of the pairs of grid probes of different buriedness types. Buriedness and distance parameters were assigned to 3 categories for further examinations. In particular, A and B type descriptors were considered as representing low; C and D medium; and E and F high buriedness levels. Distances between 1-7 Å, 8-14 Å and 15-20 Å were considered as representing low, medium and large distance values, respectively, as presented in [98].

4. One-dimensional analyses assessing the relation between binding affinity patterns and bioactivity profiles

4.1 IP-based Drug-Drug Similarity Calculations

A similarity coefficient based on the angle enclosed by two IP vectors was used to calculate the IP similarity. These vectors are determined by the docking energy values as coordinates in an 89-dimensional space created by the 89 members of the protein set. Cosine angle distance coefficient [99] was used to determine the angle between two vectors in the above described 89-dimensional space as follows:

$$d_{AB} = \arccos\left(\frac{\Sigma_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \right) \frac{180}{\pi}$$

where d_{AB} is the IP distance value between molecules A and B, a_i and b_i are the docking energy values of molecules *A* and *B* on the *i*-th protein of *n*=89, respectively.

4.2 Validation of effect prediction

Validation of the prediction of pharmacological effects of drugs was carried out on a set of 969 drug molecules approved by FDA based on the collection of pharmacological data of these molecules.

Accuracy of prediction for a given molecule (A_i) was determined by the following function:

$$A_i(k,c) = 100 * N_{si}(c) / N_{ai}$$

where N_{si} is the number of the types of effects which appear at least *c* times and can be found both in the effect list of the left out molecule and the list of the *k* most similar molecules, and

N_{ai} is the number of the effects of the left-out molecule. The total average prediction accuracy is defined as:

$$A(k,c) = \sum_{i=1}^{969} A_i(k,c) / 969$$

Confidence (C) of the validation was determined based on the comparison of the sum of accuracy values of the real and randomized (see below) datasets. Randomization was performed five times. C is calculated as follows:

$$C = (1 - A_{random} / A_{real}) * 100$$

where A_{random} is the average of average prediction accuracy values calculated from the five random datasets and A_{real} is the average prediction accuracy of the real dataset.

4.2.1 Randomization of the effect dataset

The confidence of prediction was characterized by the ratio of accuracy of prediction derived from the observed and randomized effect databases, respectively. For the generation of randomized effect databases known effects belonging to each drug according to DrugBank Database were randomized. Randomization of the effects of molecules was performed by using two constraints: (1) each molecule had to have at least one effect; (2) none of the molecules was allowed to have more than 20 effects. In a hypothetical molecule-effect matrix consisting of 969 rows (referring to molecules) and 20 columns (referring to effects) the 2,307 effect records were distributed randomly by requesting a random cell identifier for a given effect record and placing it there if the randomly chosen cell was empty. Five different random datasets were created. The distribution of the number of effects of the molecules was identical in the case of the empirically observed and the randomized datasets.

4.2.2 Leave-one-out cross-validation

Leave-one-out cross-validation (LOOCV) technique was used to test and validate the prediction method. Using this technique, the effect profile of each drug was chosen as subject of prediction at the end of a validation cycle. Thus the quantitative data characterizing the confidence of prediction refers to an average value which is calculated based on averaging the individual prediction accuracy results. The LOOCV technique was applied for the real and the five random datasets.

Confidence of the prediction increases by decreasing the number of the most similar molecules involved in prediction and also by increasing the number of required appearance of a specific effect. On the other hand, the number of predicted effects decreases simultaneously.

Consequently, in future applications of the method prediction results can be fine-tuned by varying the number of required appearances and the number of molecules involved in the calculation.

4.3 Effect Prediction

Prediction of new effects of molecules was approached by two different ways. The *neighbor-focused* prediction method investigates the overlap of the effects of the IP-neighbors while the *effect-focused* prediction method investigates the overlap of the IP-neighborhood of molecules assigned to a given effect.

4.3.1 *Neighbor-focused* Prediction Method

According to the *neighbor-focused* prediction method the prediction started with the collection of the 20 nearest IP neighbors (20-NN) of a certain molecule for which the prediction was made. Next, a list was compiled by registering all of the effects of the 20 IP neighbors, even if they appeared multiple times. It should be noted that the FDA approved effects of the studied molecule on this list were ignored. The number of appearance for each effect on this list was counted which defines the appearance parameter (*c*). If a certain effect appears at least a predefined *c* times, it is considered as a new predicted effect for the molecule. A prediction confidence value could be determined to each prediction as defined earlier.

The average prediction confidence (C_{Pav}) was calculated as follows:

$$C_{Pav}(c) = 1 - \frac{\sum_{i=1}^{969} \overline{N}_{pi.random}(c)}{\sum_{i=1}^{969} N_{pi.real}(c)}$$

where $N_{pi.real}$ is the number of the types of predicted new effects appearing at least *c* times, $\overline{N}_{pi.random}$ is the average number of the types of predicted new effects appearing at least *c* times in five different randomized databases in which all effects are randomly reallocated to the drugs.

4.3.2 *Effect-focused* Prediction Method

One might consider that the lower the number of studied neighbors (*k*) the higher the possibility to exclude effects. Thus the selection of parameter *k* should not be too restrictive.

On the other hand, a large *k* value might result in a larger number of false positive hits during effect prediction. Based on this consideration we selected 20 as a fixed value for parameter *k*.

In case of *effect-focused* prediction method, drugs assigned to each effect were collected, and their 20 most MIF-similar drugs were collected into a list. Drugs appearing multiple times on this list of most similar drugs are predicted to possess the given effect, in case their effect lists did not contain that effect originally.

5. Multidimensional analyses

The Statistical Analysis System for Windows (version 9.2; SAS Institute, Cary, NC) was used for computing Type I error probability. The alpha error level of 0.05 (two-sided) was adopted for all statistical analyses.

5.1 Principal component analysis

Factor analysis was performed on the set of Molecular Affinity Fingerprints and the structural characteristics of the protein binding pockets yielded by the PocketPicker descriptor system. The purpose of factor analyses was twofold:

(1) delineation of the basic underlying structure of the MAF and structural characteristics of the target proteins;

(2) data reduction in order to facilitate further examination of the relationship between MAF profiles and the geometric feature set.

Such a data reduction was needed for subsequent multivariate analyses in case of protein binding site analyses since the number of variables exceeded the number of cases, i.e., 154 proteins of interest (used as “cases”) are paired with 1,255 MAF variables (energy values) and 405 structural characteristics variables (geometric descriptors). 15 descriptors were omitted from the original set of 420 descriptors due to lack of variance.

For the inspection of the effect of different scoring functions on interaction profile results, 1,226 drugs were considered as “cases” or “observations” while the protein set of 154 entries served as variables.

A separate factor analysis was conducted for the X-SCORE and AutoDock4 based IP energy values, for the X-SCORE based MAF energy values and for the geometric descriptor variables, respectively. For the purpose of these analyses, we adopted the principal component method for factor extraction. The extracted factors were subjected to ORTHOMAX/PARSIMAX rotation in order to derive a simple structure for helping the interpretation. In short, principal component analysis generates new variables in an equal number to the original variables. The so-called factors or principle components are orthogonal and formed by the linear combination of the original variables that can be characterized by their weights in the linear functions, i.e., their *loadings*. Factors form a new coordinate system in which new coordinates, i.e., *scores*, are assigned to the observations. Dimension reduction is achieved by the reduction of the number of factors taken into consideration.

Variables were allocated to factors according to their highest loading; the threshold loadings of 0.4 and -0.4 were chosen to indicate saliency in all factor analyses. For the examination of the dimensionality of data based on the factor analysis (i.e., to determine the number of factors to be used in further analyses), the average variance criterion was adopted: factors were considered as significant if they explained more than the average (>1/154 = 0.65%) of the total variance individually. This threshold, which corresponds to the Kaiser-Guttman eigenvalue>1 rule [100], was chosen since it represents the variance accounted for by an individual variable by chance based on the intrinsic dimensionality of our data (i.e., the smaller of the number of cases or variables in the data). For the implementation of the factor analyses, SAS "FACTOR" procedure was applied.

5.2 Canonical correlation analysis

In canonical correlation analysis, variates in one set are formed to describe the correlation structure in a different set of variables. Therefore, canonical correlation analysis can be considered to be an extension of factor analysis for two separate sets of variables. In particular, the objective of this method is to obtain as high correlation as possible between the derived variables (here, pairs of variates or 'factors' are formed from the two sets) in variable set 1 and those in variable set 2. In other words, this technique is an optimal linear method for studying interset association: components from the two sets are extracted jointly to be maximally correlated with a component of the complementary variable set, within the constraint of orthogonality of all components except the correlated pair.

The statistically significant canonical factor pairs were examined further in order to:

(1) assess the importance of different scoring functions;

(2) match the complex pattern structures of IP and effect profile matrices;

(3) visualize the relationship between drugs and protein binding sites.

In case of the canonical correlation analysis of the association between MAFs and protein binding site descriptors, PCA factors of the MAF and the PocketPicker descriptor matrices with salient canonical loading over 0.25 or below -0.25 were collected in each canonical factor pairs. Canonical PCA loading structures were analyzed and in case of the MAF PCA factors representatives of the appeared typical drug groups were selected. In case of the PocketPicker PCA factors, salient descriptors were collected mapping the concomitant buriedness indices within the three distance levels. Proteins having salient canonical scores (over 1 and below -1) were also collected. Sign of the loadings was taken into consideration for the interpretation.

The same criteria of saliency were applied for scoring function evaluation. Canonical correlation analysis for effect prediction was performed without previous dimension reduction by PCA. We mention that a reduced drug set containing 1,226 entries was used for the evaluation of IP-EP relationship, due to incomplete effect information in 29 cases.

5.3 Canonical redundancy analysis

Canonical redundancy analysis is used to examine how much of two sets of variables “overlap” in terms of explained variance or redundancy. This approach allows the determination of the amount of variance (or redundancy) that the canonical components (factors) account for in their ‘own set’ of variables, and in the ‘opposite set’ of variables (e.g. in case of protein site analysis, how much the individual structural canonical factors explain of the total variance of the structural characteristics of the protein binding pockets and of the MAF profiles, respectively).

Canonical redundancy analysis was used to determine the overlap between two dataset pairs:

(1) the AutoDock4 and X-SCORE-based IPs;

(2) MAFs and PocketPicker structural descriptors.

In addition to the explained variance associated with the individual canonical factors, we also determined total redundancy, i.e., the total amount of explained (predicted) variance of one set of variables given the whole predictor set. We note that, unlike canonical correlation, redundancy indices are nonsymmetric; in general, by designating one variable set a predictor set, the associated redundancy of the other set differs from what it would be if the functions of the two sets were reversed. The F statistic was used for significance testing of correlations measured between canonical variate pairs.

To perform canonical correlation and redundancy analyses, we used the SAS “CANCORR” procedure.

5.4 Linear discriminant analysis

Linear discriminant analysis (LDA) is a commonly used statistical approach to find an optimal linear transformation for maximizing the between-class variance and minimizing the within-class variance, thereby identifying the best discriminating surfaces or "hyperplanes" in the multidimensional space of feature sets that generate complex pattern classes.

Based on the canonical factor pairs of IPs and effect profiles, we calculated the probability of each effect for each drug via LDA. Classification functions for each effect were determined in order to classify observations into known effect classes based on the IP canonical factors. The performance of the classification function was evaluated by estimating the drug-effect probability for each drug with regard to each effect and the rate of correct classification for all drugs with regard to all effects. In order to accomplish this, each observed IP was plugged in the classification function in order to generate the drug-effect probability matrix.

5.5 Validation

In order to evaluate robustness of the effect prediction results, i.e., the extent to which the aforementioned effect classification results would generalize to independent data, a cross-validation with the leave-one-out procedure (LOOCV) was performed. LOOCV, which is also called rotation estimation, includes N rounds of validation, where N is the number of observations in the sample (i.e., the set of 1,226 drugs). Adopting the standard LOOCV approach, one round of validation consisted of three steps:

(1) partitioning the data set into two complementary subsamples, with N-1 and 1 observations/drugs, respectively;

(2) conducting the CCA and LDA to derive the IP-based classification function using the subset with N-1 observations;

(3) computing the drug-effect probability as well as determining (predicting) effect-group membership for the set with the single observation.

The cross-validation results for each of the originally registered drugs were then combined to yield a single average estimate for each effect.

5.6 Receiver Operating Characteristic analysis

Efficacy of the effect classification functions was assessed by Receiver Operating Characteristic (ROC) analysis, i.e. determining the true positive rate (TPR) and the false positive rate (FPR) for every effect, using the classification function (determined by LDA) and a sliding cut-off parameter running from 1 to 0. Molecules are reclassified at each point, considering compounds as “positive” if they have larger possibility for an effect than the actual cut-off value and “negative” in the opposite case. Positives can be further divided into true and false positives depending on the binary value originally assigned to the given drug-effect pair i.e., if a drug had “1” in the effect profile and produced a classification value larger than the cut-off point, it will be considered as “true positive”. True and false negatives can be distinguished as well at each step. TPR and FPR are the fraction of true positives among the positives and the fraction of false positives among the negatives, respectively and are often referred to as sensitivity and (1-specificity). TPR and FPR values for each cut-off point are plotted on a two-dimensional graph called ROC curve. A completely random classification would result in an ROC curve on the diagonal of the graph, meaning that for every true positive hit, a false positive hit also falls into the classification. The better the classification, the closer the curve to the (0,1) point of the graph.

5.7 Top Hit Rate calculation

Besides ROC analysis, an alternate evaluation method called Top Hit Rate calculation was developed to assess the efficacy of effect classification. Here, the entire set of the 1,226 drugs were listed in descending order by their probability value of possessing the given effect, and the *top of the list* was cut at the number of the registered drugs to the studied effect. This top list contains registered and not registered drugs of the given effect since the not registered drugs can also gain high probability value in the multidimensional validation method and registered drugs can have low value.

Classification accuracy can be characterized with the proportion of the registered drugs in the top list. Therefore, the following Top Hit Rate value was calculated for each of the 181 effects:

$$Top\ Hit\ Rate = \frac{number\ of\ the\ registered\ drugs\ in\ the\ top\ of\ the\ list}{number\ of\ all\ registered\ drugs\ of\ the\ given\ effect}$$

Here, the number of all registered drugs of the given effect equals to the number of drugs in the top list, as discussed above.

6. *In vitro* analyses

In vitro tests were performed on a Hamilton Starlet Liquid Handling Workstation (Hamilton Robotics, Bonaduz, Switzerland). Spectroscopic measurements were carried out on BMG FluoStar Optima (Offenburg, Germany). Commercially available assay kits were used for the measurements and the robot was programmed according to the manufacturers' instructions. The selected drugs were initially tested at 500 μM concentration and certain drugs were further tested to determine the K_d values. Each data point is an average of two independent measurements.

ACE inhibition was tested using the ACE Kit-WST from Dojindo Molecular Technologies, Inc. (Kumamoto, Japan, Cat. No. A502-10). 3-hydroxybutyril-glycil-glycil-glycine is utilized as a substrate in this kit and under the actions of ACE and aminoacylase it is converted into 3-hydroxybutyric acid. In the development step it is further oxidized into acetoacetate by the action of 3-hydroxybutyrate dehydrogenase. At the same time, the cofactor, NAD^+ is converted into the reduced form NADH. During the oxidation of NADH to NAD^+ a water-soluble tetrazolium salt is reduced coupled with an electron mediator and generates a yellow formazan. Tested drugs were incubated at the given concentrations with enzyme working solution and the substrate for 60 min at 37°C. In the next step indicator working solution was added to the reactions, the plate was incubated at room temperature for 10 minutes and read at 450 nm. Captopril was used as a control for inhibition.

COX inhibition was investigated using the COX Inhibitor Screening Assay Kit from Cayman Chemical Co. (Cayman Europe, Tallinn, Estonia; Cat. No. 560131). Briefly, this enzyme immunoassay kit quantifies the inhibition of COX-1 and COX-2 activities by measuring the formation of prostanoid products from the substrate arachidonic acid. Tested drugs were preincubated at the given concentrations with enzymes COX-1 and COX-2 for 10 minutes at 37°C. Reactions were started by adding the substrate, then incubated for 2 minutes at 37°C and stopped by 1M HCl. Prostaglandin screening was performed on a 96-well microplate coated with mouse anti-rabbit IgG. COX reaction samples were mixed with an AChE-linked tracer and the antiserum then incubated for 18 hours at room temperature. The washed plate was developed by Ellman's reagent for 60 minutes and read at 400nm. Aspirin was used as a control for inhibition.

In vitro tests were carried out by Ágnes Peragovics, László Végner, Balázs Jelinek and András Málnási-Csizmadia.

6.1 Materials

Aminosalicylic acid, furosemide, monobenzone, nitrofurazone and nitroxoline were purchased from Aldrich, maraviroc from AvaChem, chlorambucil, clavulanate, ethacrynic acid, flucytosine, furazolidone, latamoxef (moxalactam), lipoic acid, nitrofurantoin, novobiocin, paclitaxel, penicillin V, phenazopyridine and tinidazole from Fluka, carbenicillin from Merck, chlormezanone and chlorphenesin from MP Biomedicals, dasatinib and tipranavir from Santa Cruz Biotechnology, acitretin, alpha-linolenic acid, aspartame, aspirin, azithromycin, captopril, estrone-sulfate, flutamide, gemfibrozil, L-carnitine, lomustine, L-proline, metronidazole, milrinone, nalidixic acid, nateglinide, nelfinavir, nilutamide, penicillin G, pyridoxal phosphate, telmisartan, ticarcillin and valproic acid from Sigma, benzyl benzoate and biotin from Sigma-Aldrich and ambenonium from Tocris Bioscience.

Predicted ACE inhibitors pentosan polysulfate, polystyrene sulfonate and udenafil were commercially not available at the time of testing. Astemizole was omitted from testing because it was withdrawn from the market in most countries.

Predicted COX inhibitors aminohippurate, amlexanox, bexarotene, phenprocoumon, procarbazine, rosoxacin, stepronin, tolcapone and valrubicin were commercially not available at the time of testing. Gentian violet and sodium lauryl sulfate were excluded from testing due to their limited clinical applicability.

7. Cell culture D_1, D_2, α_{1B}, α_{2A} and β_1 assays

Cell culture assays were performed independently by Euroscreen S. A., Brussels, Belgium, according to the company's internal protocols. Amiloride and minoxidil were tested in duplicate for agonist and antagonist activities on the human adrenergic α_{1B}, α_{2A}, β_1 and D_2 receptors using the Aequorin assay, and on the D_1 receptor using the cAMP HTRF assay. Activities were provided as percentages of total activities of the company’s reference compounds.

Results and Discussion

Considering the enormous mass and the complexity of data used in this study, Results and Discussion sections are merged in order to improve readability.

Two evaluation strategies were applied in order to study the relationship between the IP and EP datasets. The first method is a so-called *one-dimensional analysis,* a simple approach based on the one-dimensional distance between IPs of drug pairs in a multidimensional space. Its advantage is the clear, straightforward logic. To assess the applicability of *in silico* affinity data in bioactivity predictions, an *in silico* screening was performed using AutoDock3 docking and scoring for 969 FDA-approved drug molecules and a set of 89 proteins. A linear distance measurement was introduced in order to create an IP similarity matrix for the small-molecule drugs. Effect data were extracted from DrugBank without further refinement. Side effect data were collected manually for each drug from FDA drug labels. Based on IP similarity considerations, two different one-dimensional effect prediction methods were developed and validated using leave-one-out cross-validation.

After the promising initial results, the docking procedure was repeated applying the improved AutoDock4 scoring function. Results were reproduced using the more reliable X-SCORE function as well. On this data set, a *multidimensional analysis* procedure was performed. This methodology overcomes the inaccurate handling of the dimension reduction problem occurred in the previous analyses (see later in details). Here, principal component analyses were done in order to determine the dimensionality of the resulted IP data sets and the factor structure was examined. Canonical correlation and redundancy analyses were performed to determine the relation between the binding affinity values originated from the two scoring functions. Based on these results, X-SCORE data were chosen for determining the correlation between the IP data and the effect profiles. DrugBank based effect categories were refined and extended manually. Side effect data were omitted due to the high level of noise. Canonical correlations were done between the whole IP set and each binary effect patterns. Then, linear discriminant analysis was carried out for each effect in order to develop a classification function that calculates the probability that a given drug will possess the studied effect. Accuracy of the acquired classification functions was assessed by Receiver Operating Characteristic analysis. Robustness of the classification was determined by leave-one-out cross-validation for each effect. Finally, several drugs and effects were selected for deep analysis. Here, retrospective literature analyses were done to reveal the validity of “false positives” of the classification. If no data was available, *in vitro* tests were carried out for falsification or justification.

The diversity of the applied protein set was also examined to check the validity of our assumption that the protein set is diverse enough to mimic the proteome. Moreover, in order to determine the importance of binding site geometry on binding affinity results, factor structures of MAF and PocketPicker geometric descriptor data sets were assessed by PCA. Then, canonical correlation analysis and redundancy analysis was applied to calculate the level of correlation and the explained variances between the sets, respectively. Finally, the canonical factor structure was examined in order to derive clear rules describing the connection between protein affinity data and binding site shape [98].

1. One-dimensional analyses

The basis of the one-dimensional analyses is the pairwise similarity between two IPs, considered as vectors in a multidimensional space. The approach is called one-dimensional since the distance of the two IP vectors is a one-dimensional measure. The advantage of this measure is that it reflects the pattern of the binding energy values in the profile more than the actual binding affinity values. I.e., if two drugs possess the same interaction pattern but with different average binding affinity, their vector distance will be relatively small, suggesting that the two compounds are similar from a polypharmacologic point of view. (On the other hand, a single miscalculated docking energy value can cause significant error in the distance measurement.)

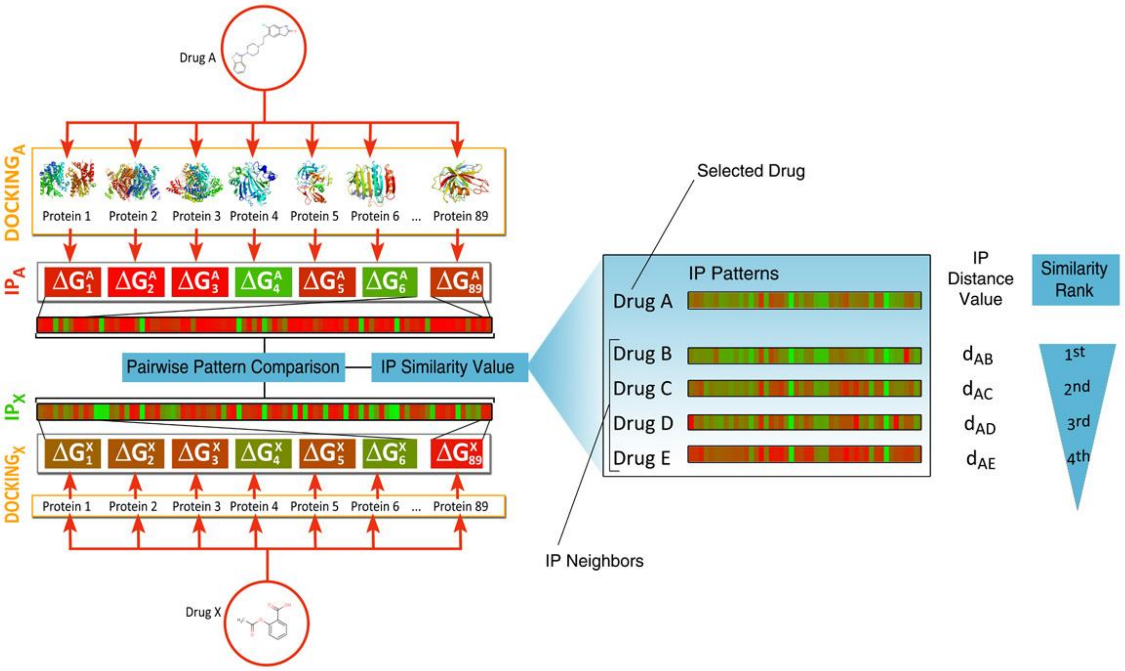


Figure 12. Summary of the IP generation and the IP similarity calculations. Drugs A and X are docked to the 89 members of the protein set. Their respective IPs (with color-coded energy values, ranging from green to red, i.e., from lower to higher binding free energy values, respectively) are compared and a pairwise similarity value is calculated. Based on these values, similarity lists are created for each drug, containing the remaining set of molecules in a decreasing order of similarity.

In these examinations, structural and pharmacological data of 969 FDA approved drug molecules were collected from the DrugBank Database. The 2,307 effect and 28,919 adverse reaction records of these drug molecules were manually categorized and organized into relational databases, resulting in an effect and an adverse reaction dataset. Interaction patterns of the drug molecules were generated with 969 * 89 = 86,241 docking runs, each repeated 25 times. The calculated lowest binding free energies of each protein-drug complex were collected and organized into a matrix. The 89 docking energies of each drug constitute the row vectors of the IP matrix. Then, we compared the IPs with each other in order to generate an IP distance matrix. Comparability of IPs is assured by the identical discriminator surface for all of the studied molecules because they were only allowed to interact with the selected surface regions of each protein. In order to quantify the similarity between two IPs we introduce an IP Distance Value (d) based on cosine angle distance indexed by the angle enclosed by two IP vectors, ranging from 0 (most similar) to 180 (least similar). It is important to note that IP Distance Value refers to the binding profile similarity irrespective of the general binding strength of certain drugs to the whole protein set. For each molecule, a similarity rank list was generated (Figure 12). The similarity rank list was built on the basis of the IP Distance Value (d): the smaller the value of d, the higher the position on the list. Different *k*-NN (*k* nearest neighbors) were generated from these lists (Figure 12).

1.1 Validation

Analysis of the relation of the IP and effect databases was performed applying the leave-one-out cross-validation technique. A selected molecule was treated as a reference and its effects were predicted based on the effects of molecules having similar IPs to this reference molecule (Figure 13a) as determined before. Every effect of this set of IP-similar molecules was compiled into a list and the number of times an effect appears on this list was counted (Figure 13a). The effects which appear at a prespecified number of times (i.e., the predicted effects) were matched to the known effects of the reference molecule. The prediction accuracy of effects is the ratio of the predicted and known effects (Figure 13b). In order to determine the confidence of the prediction, we generated random effect databases, and then performed the same prediction procedure on them. The high values of a confidence measure calculated from the ratio of the effect prediction accuracy of real and random databases throughout the whole database suggest that IPs correlate with observed effects (Figure 13c) as expected. E.g., using the first 20 neighbors of any molecule and considering effects as “real” or “true positive” if they appear at least two times, 36.7% of the registered effects can be recovered for any drug. Applying a random database, only 8% of the known effects are regained (Figure 13b). Increase of the applied neighbors will increase the accuracy values, e.g. considering double appearance for 30 neighbors, 44.1% and 13.2% of the known effects are recognized in a real and a random dataset, respectively. On the other hand, more “false positives” will appear. However, if we aggravate the appearance threshold up to at least four appearances for 20 neighbors, the accuracy values will decrease to 15% and 0.7% for the real and random datasets, respectively (Figure 13b). Although these values seem to be small, one must consider that the confidence of the prediction increases by increasing the appearance threshold. In this case, applying 20 neighbors and at least two appearances, the confidence value is 78% while it increases to 95.4% if the threshold is set to at least four appearances. To sum up, increasing the number of considered neighbors and decreasing the appearance threshold results in an increase of the ratio of the regained effects while the confidence decreases. Reducing the number of neighbors or increasing the threshold has the opposite effect: accuracy decreases but confidence increases, resulting in a lower number of regained effects but with higher confidence. These parameters (i.e., the appearance threshold and the number of neighbors) should be fine-tuned for effect prediction.

Furthermore, we validated the method by using a subset of drug molecules whose target protein is not represented in the docking protein set (809 molecules). Similar results were obtained between the validations of this drug subset and the set containing 969 drugs (Figure 13c inset). This demonstrates that the predictive power of the one-dimensional method is independent of the presence of the drug targets on the discriminator surface. Therefore, this is the first evidence that our *in silico* interaction pattern based effect prediction method is not a target-based (target-specific) approach.

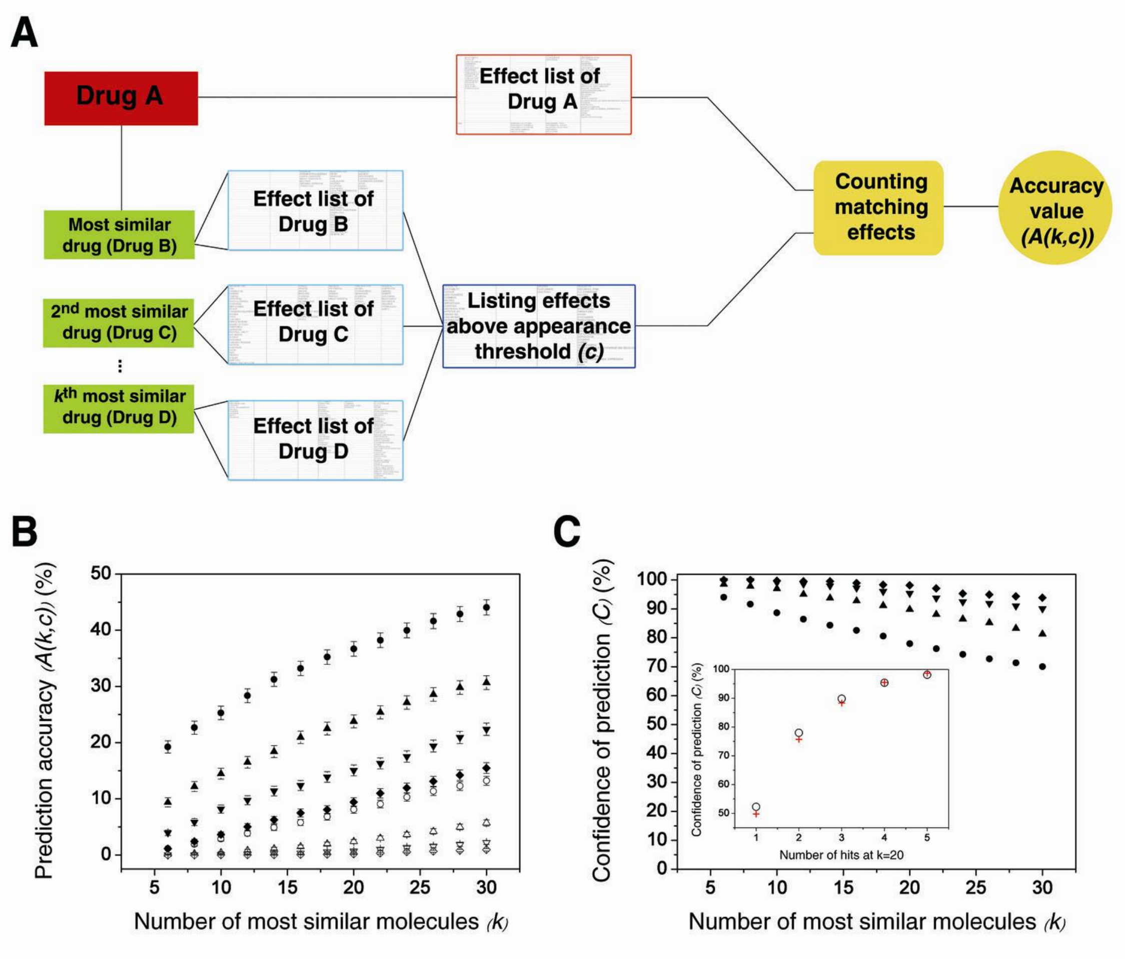


Figure 13. Validation of the effect prediction method based on IP similarity. Panel (**a**) shows the schematic summary of the validation procedure. The validation of effect prediction of the one-dimensional method was carried out by the leave-one-out cross-validation technique, using the real and five randomized datasets. The prediction model contains two variables: *k*: the number of most similar molecules to the reference molecule according to the *k*-NN queries, *c*: the effects that appeared minimum twice (circle), three (up-triangle), four (down-triangle) or five (diamond) times in the list created from summing up the effects of the most similar molecules. Solid and open symbols represent data calculated from real and randomized datasets, respectively. Accuracy of prediction for a given molecule (A_i) and total average prediction accuracy (*A(k,c)*) were determined

as described in Methods. A (*k*,1), A (*k*,2), A (*k*,3), A (*k*,4), A (*k*,5) of the real (solid symbols) and the average of A (*k*,*c*) values of the randomized datasets (open symbols) are plotted with error bars (standard error of the mean), as a function of *k* on panel **B**. Panel **C** shows the confidence values plotted as a function of *k*. Inset shows the confidence values for the 20 most similar neighbors of each drug (*k*=20), in the case of the whole (circle) and the reduced drug set (cross). The reduced set consists of molecules whose target proteins are not included in the set of 89 proteins (969 and 809 drugs, respectively).

1.2 Effect prediction

Two one-dimensional approaches have been developed to predict new effects for the approved drug molecules (Figure 14).

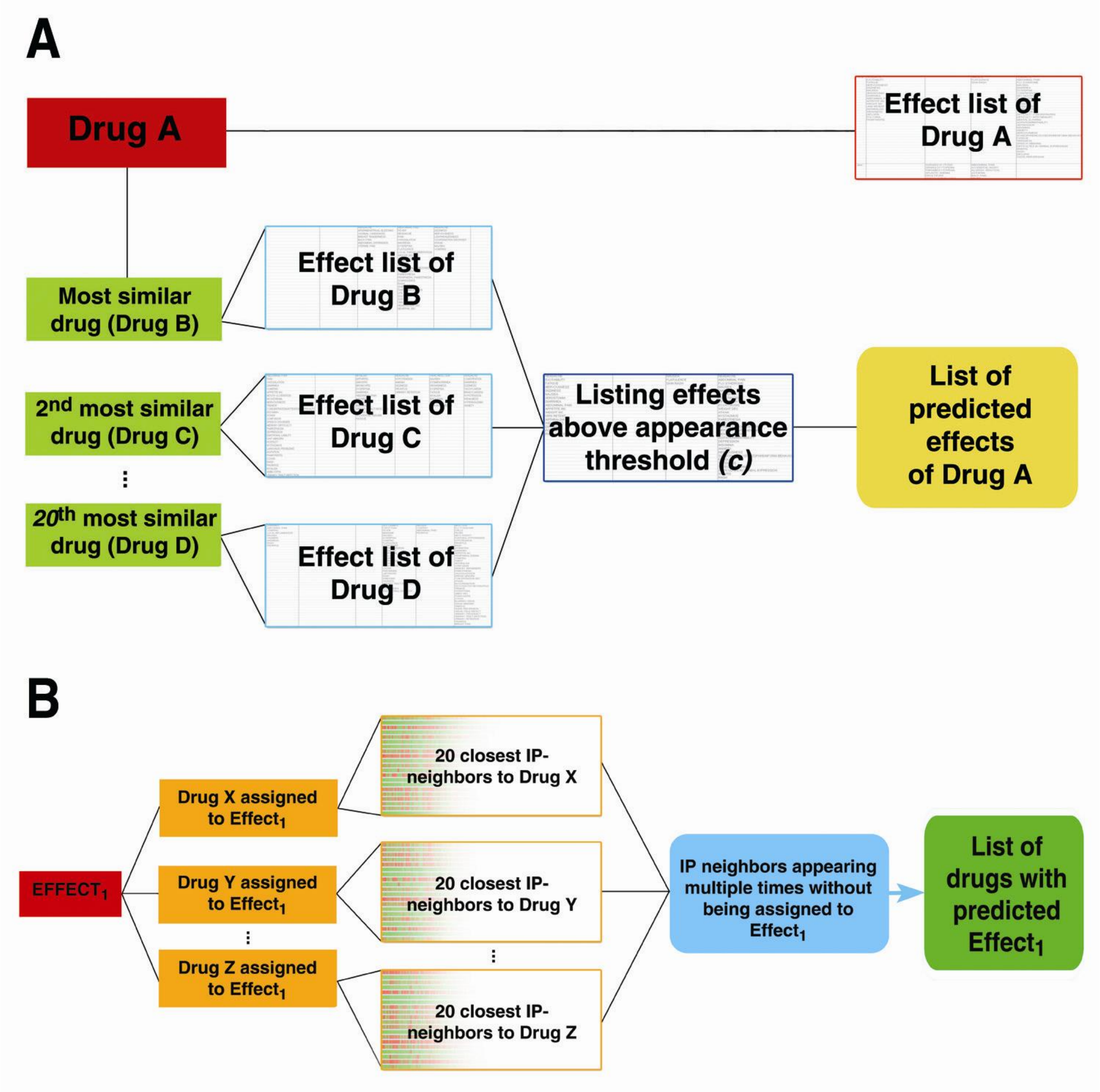

A
Drug A
Effect list of Drug A
Most similar drug (Drug B)
Effect list of Drug B
2nd most similar drug (Drug C)
Effect list of Drug C
20th most similar drug (Drug D)
Effect list of Drug D
Listing effects above appearance threshold (c)
List of predicted effects of Drug A
B
EFFECT$_1$
Drug X assigned to Effect$_1$
Drug Y assigned to Effect$_1$
Drug Z assigned to Effect$_1$
20 closest IP-neighbors to Drug X
20 closest IP-neighbors to Drug Y
20 closest IP-neighbors to Drug Z
IP neighbors appearing multiple times without being assigned to Effect$_1$
List of drugs with predicted Effect$_1$

Figure 14. Schematic summary of the two IP similarity-based prediction methods.

The first one called *neighbor-focused* prediction method (Figure 14a) applies the same procedure as the presented validation process except for that the predicted effects are not matched with the approved effects. In this case, the close IP neighbors (*k*=20) of a studied drug molecule are listed and their effects are collected. The appearance of each effect is counted (parameter c_{real}). The same process was applied on the five randomized effect databases which were used in the validation step in order to determine the average value of c_{random}. In order to determine the confidence of the prediction (C_P) on a certain drug and effect the following equation was applied:

$$C_P = 1 - \frac{c_{random}}{c_{real}}$$

The alternative, *effect-focused* prediction method is based on finding common members in the lists of IP-similar molecules of drugs associated with particular effects (Figure 14b). Here, drugs associated with a certain effect were collected. Then a list was formed from the 20 most IP-similar drugs of the collected drugs. Drugs originally associated with the studied effect were excluded from the list. Drugs which appeared multiple times in this list were indicated to have this effect.

The method was subjected to a detailed validation process in order to ensure that our approach can be used for effect prediction. With regard to the main objective of the project, i.e., to discover and confirm new effects for approved drugs, our results show that 838 and 267 new hitherto unrevealed effects were predicted with average prediction confidence values ($C_{P\ av}$) over 80% ($c \geq 5$) and 90% ($c \geq 6$), respectively (Figure 15).

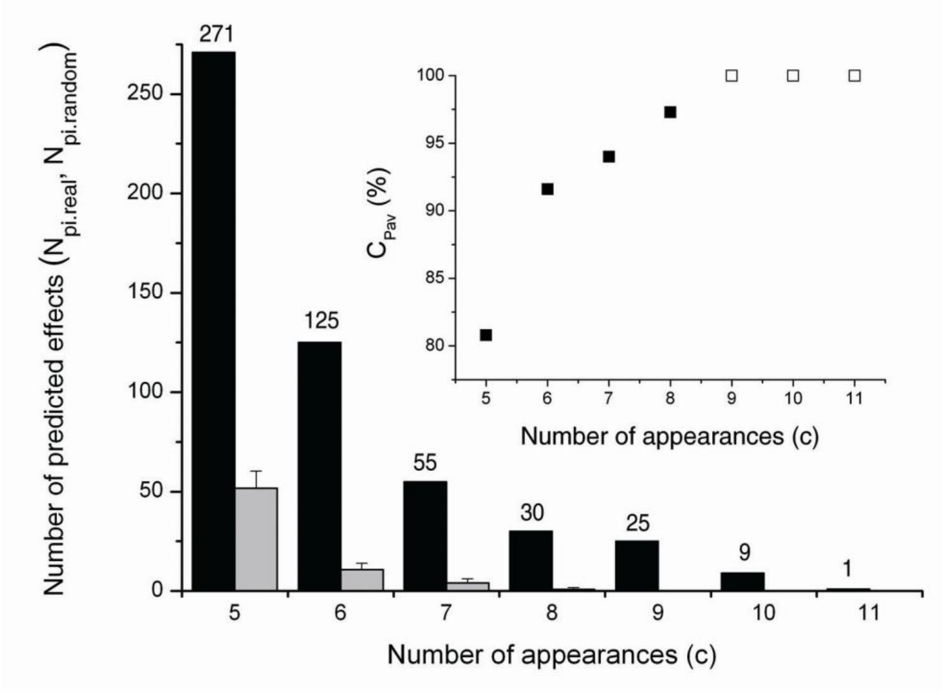


Figure 15. Distribution of predicted new effects based on the *neighbor-focused* prediction method. Inset shows the average prediction confidence (C_{Pav}) calculated as follows:

$$C_{Pav}(c) = 1 - \frac{\sum_{i=1}^{969} \overline{N}_{pi.random}(c)}{\sum_{i=1}^{969} N_{pi.real}(c)}$$

where $N_{pi.real}$ is the number of the types of predicted new effects appearing at least *c* times (black bars on the main figure), $\overline{N}_{pi.random}$ is the average number of the types of predicted new effects appearing at least *c* times in five different randomized databases (grey bars on the main figure) in which all effects are randomly reallocated to the drugs. The last three data points (open symbol) are virtually 100% because $\overline{N}_{pi.random}$ was zero in these cases.

1.3 Summary and further optimization possibilities

With the aforementioned good empirical validation results in mind, it was realized that many aspects of this initial IP-based prediction can be further improved. For example, currently available phenotypic characterizations of complex drug effect/adverse reaction profiles that underlie the prediction are incomplete. These factors, in turn, are expected to lead to suboptimal prediction. However, due to its expandability, our system can be supplemented with emerging knowledge on hitherto unknown clinical effects of marketed drugs and information on newly approved drugs. Decision rules can be optimized, and relationships between MIF-similar drugs and bioactivity profiles can be characterized better. The predictive power of the approach can be enhanced by applying statistical methods that are able to overcome the dimensionality problem of IPs. The neighborhood-based similarity lists applied here inherently reduce the multidimensional nature of the IP data, forcing multidimensional structures into a potentially misleading measurement. For example, if a drug's closest neighbors form two distinct and well-defined clusters in the space of the applied protein set with similar average distances, the neighborhood list will consist of elements almost randomly picked from the two non-related groups. The one-dimensional distance parameter is a reasonable descriptor for molecules A-B and A-C but it contains no data of the B-C distance. Consequently, the information represented by the two subgroups will be blurred. Therefore, this distance cannot be applied for "mapping" the interaction pattern space. To overcome this issue, a new multidimensional approach was introduced.

Together with the development of a more thorough effect classification/prediction system, it was decided to introduce a larger protein set that might prove higher diversity than the originally used 89 molecules. 71 new proteins were selected based on the same selection criteria, thus extending the protein set to 160 entries (although 6 of them were removed before the analyses, see Methods). We also extended the drug set with molecules below 600 Da and labeled “FDA-approved drug” in DrugBank, collecting 286 new entries, resulting in 1,255 drugs.

During the one-dimensional evaluation of the results, AutoDock4 was launched that performed better than the older version we used [101]. DOVIS, a newly presented docking manager engineered by the Biotechnology High Performance Computing Software Applications Institute (U.S. Army Medical Research and Materiel Command) enabled the parallelization of docking jobs on multiprocessor systems that resulted in one magnitude acceleration of the calculation speed. The two scoring functions implemented in DOVIS, i.e. AutoDock4 and X-SCORE, served an opportunity to compare two widely used scoring functions and to assess the importance of scoring on the results. Based on its superior properties to AutoDock3, DOVIS was chosen to perform the new docking runs needed for the detailed analysis. Considering the aforementioned backdrops of rescoring, complete redocking was carried out for producing the AutoDock4 and X-SCORE based data matrices as well.

2. Multidimensional analyses

IPs and effect profiles were generated based on structural and pharmacological information on 1,226 FDA-approved small-molecule drugs this time (Figure 17, Appendix 1). Effect profiles (EPs) were extracted from the DrugBank database and stored as a row vector for each drug with binary entries, comprising 559 effect categories that were reduced to 181 by excluding effects with a low number of registered drugs. X-SCORE and Autodock4 scoring functions were used to calculate the corresponding binding affinity values of the 1,226*154 drug-protein complexes as described in Methods. The binding affinity values were piped into the IP vectors. The IP and EP vectors were collected into matrices and used as input databases in the subsequent investigations (Figure 16).

Before proceeding with the analysis of IP-EP correlation, the extended protein set was subjected to a diversity analysis. Moreover, the importance of protein binding site geometry was also assessed. The effect of different scoring functions on the binding affinity data matrix was studied in order to choose the scoring function that is more suitable for further analyses.

Based on the results, X-SCORE scoring function was applied for multidimensional analyses assessing the IP-EP association.

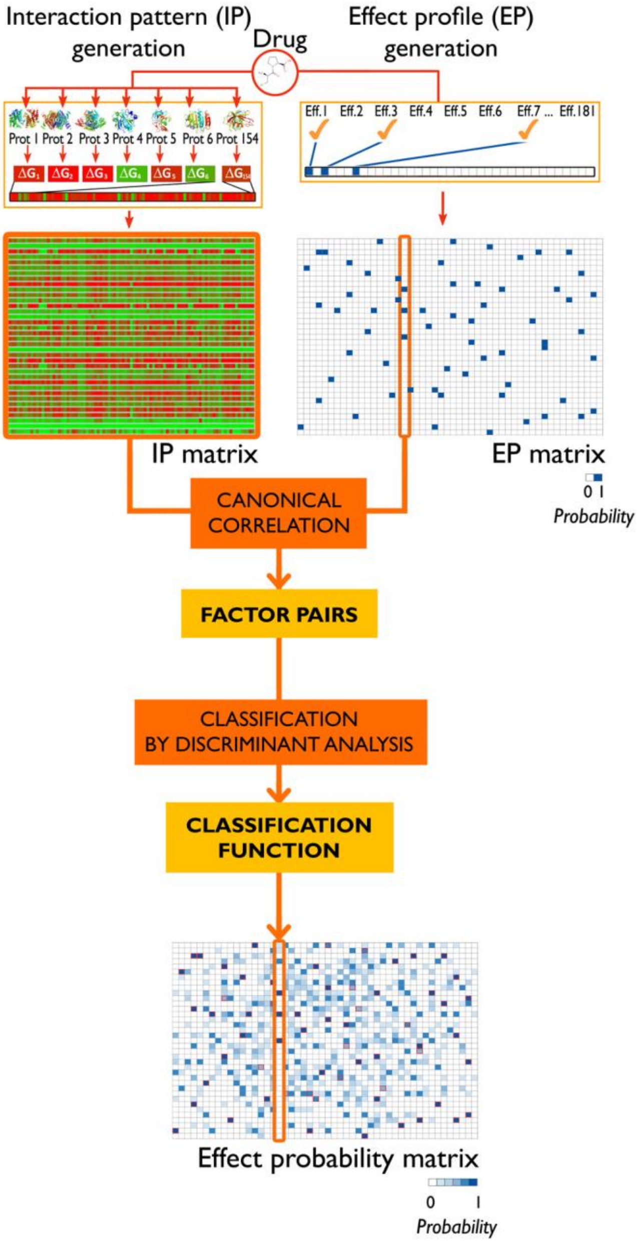


Figure 16. Graphical summary of the Drug Profile Matching method: from the atomic structures to the effect probability matrix. A drug molecule is docked to a set of 154 proteins and the calculated binding free energies are entered into a row vector, i.e. the Interaction Profile (IP). IPs of the 1,226 studied drugs form the IP matrix. The Effect Pattern (EP) matrix contains the therapeutic effects of the drugs in a binary coded form (blue and white cells represent the presence and the absence of a given effect from the 181 categories, respectively). Then, canonical correlation analysis is performed in order to generate highly correlating factor pairs that serve as the input for linear discriminant analysis. This way, classification functions are produced that yield the probability for each drug-effect pair, resulting in the effect probability matrix. Note that the values in this matrix are continuous.

2.1 Protein diversity analysis

We assume that an IP vector with a diverse set of proteins used in the present study models the interactions formed by a given drug with the human proteome. To check this assumption, the diversity of the protein set was calculated from the similarity values of the binding site geometry descriptors obtained from PocketPicker software. 95.5% (11,252 out of 11,781) of the values in the protein-protein dissimilarity half-matrix are above the dissimilarity threshold published in [89], suggesting a fairly diverse set of proteins.

2.2 Analysis of the importance of binding site geometry

To address the question whether the geometrical parameters of the protein binding sites are important in determining drug-protein binding properties [98], large data matrices were assembled from both sides, i.e. the interactions of 154 proteins and 1,255 FDA-approved small-molecule drugs were studied while protein binding site shapes were described using 405 geometrical parameters. The same set of atoms isolated from each protein and centered to the gravity center of the natural ligand was applied in docking simulations and binding site description procedure as well. The size of the docking box was set to ensure that even the largest members of the drug set have enough space for finding the lowest-energy conformation. Box sizes were not adjusted to smaller ligands, keeping consistent treatment of proteins our priority.

2.2.1 PCA of molecular affinity profiles of target proteins

As described in the Methods, PCA with ORTHOMAX/PARSIMAX rotation of the molecular affinity fingerprints was conducted in order to determine the underlying factor structure of the MAF profiles. Table 3 displays the explained variances for the first 40 factors resulted by the factor analysis.

Factor Number	**Explained Variance**	**Cumulative Explained Variance**
1	0.1816	0.1816
2	0.0768	0.2584
3	0.0574	0.3158
4	0.0382	0.3539
5	0.0322	0.3861
6	0.0309	0.4171
7	0.0247	0.4417

8	0.0236	0.4653
9	0.0197	0.4850
10	0.0181	0.5032
11	0.0169	0.5200
12	0.0164	0.5364
13	0.0147	0.5511
14	0.0139	0.5650
15	0.0127	0.5777
16	0.0123	0.5900
17	0.0118	0.6018
18	0.0113	0.6131
19	0.0107	0.6239
20	0.0105	0.6344
21	0.0100	0.6443
22	0.0089	0.6533
23	0.0087	0.6619
24	0.0082	0.6702
25	0.0080	0.6781
26	0.0078	0.6860
27	0.0073	0.6933
28	0.0070	0.7003
29	0.0069	0.7072
30	0.0068	0.7139
31	0.0064	0.7203
32	0.0063	0.7266
33	0.0061	0.7327
34	0.0059	0.7386
35	0.0058	0.7444
36	0.0056	0.7500
37	0.0053	0.7553
38	0.0052	0.7605
39	0.0051	0.7656
40	0.0050	0.7706

Table 3. Explained Variances of PCA Factors obtained from the MAF Matrix. The first 40 factors obtained from the factor analysis of the MAF profiles of 154 target proteins are displayed. 30 factors were retained in accordance with the average variance criterion (i.e., explaining individually more than 1/154=0.65% of the total variance). They explain cumulatively 71.4% of the total variance.

Overall, 30 factors explained 71.4% of the total variance of the MAF energy values and were retained for subsequent analyses. 90% of the total variance is explained by using 78 factors of the theoretically possible 154 factors with nonzero eigenvalues. To investigate the performance of the orthogonal rotation procedure in terms of achieving a simple structure, we examined the number of salient loadings for each of the individual factors retained for further analyses. Figure 17 shows the distribution of the number of salient loadings for each of the 30 retained factors. The number of salient loadings varied between 10 and 35 for the individual factors, indicating that simple structure was achieved since the rotated factors contained only a small subset of the original variables.

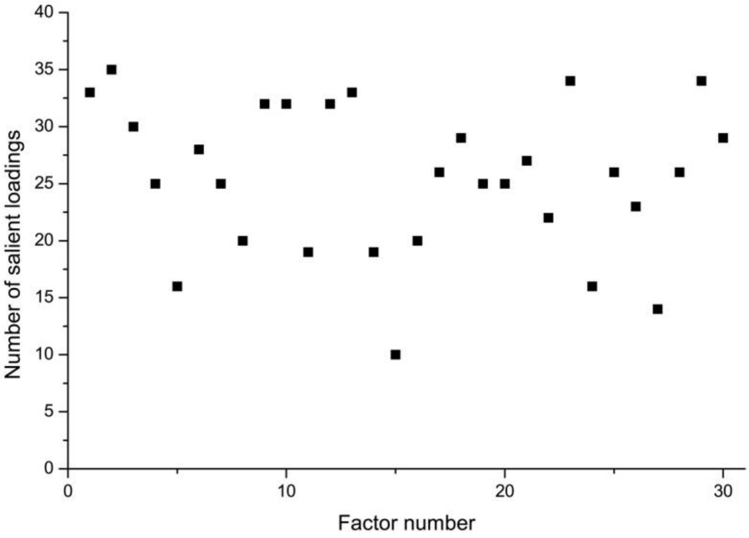


Figure 17. Number of salient loadings across the 30 PCA factors of the MAF matrix. 30 factors were obtained from the matrix of the Molecular Affinity Fingerprints (MAFs) of target proteins by principal component analysis (PCA). The number of salient loadings (i.e., loadings with a value of ≥ 0.4 or ≤ -0.4) varied between 10 and 35 for the individual factors, indicating a simple factor structure since the number of variables in the original MAF matrix was 1,255.

2.2.2 PCA of the geometric characteristics of protein binding sites

Analogous to the analysis of the MAF fingerprints, PCA analysis with ORTHOMAX/PARSIMAX rotation was performed for the full set of 405 variables comprised in the PocketPicker descriptor matrix. Explained variances for the first 40 factors resulted by the factor analysis are displayed in Table 4. Analogous to the approach adopted for the PCA analysis of the molecular affinity fingerprints of the 154 proteins, we determined the number of factors that explained at least 0.65% of the total variance individually. As indicated by Table 4, this criterion resulted in 13 factors which explained cumulatively 94.1% of the total variance. Altogether, 5 factors, respectively, explained >5% of the total variance of the geometric descriptors. Furthermore, 9 factors of the theoretically possible total of 154 factors with nonzero eigenvalues accounted for 90% of the total variance. We note that 116 factors explained 100% of the variation of the full set of PocketPicker descriptors (n=405).

Factor Number	**Explained Variance**	**Cumulative Explained Variance**
1	0.3847	0.3847
2	0.2359	0.6206
3	0.0818	0.7024
4	0.0544	0.7568
5	0.0524	0.8091
6	0.0377	0.8469
7	0.0257	0.8726
8	0.0180	0.8906
9	0.0136	0.9042
10	0.0120	0.9162
11	0.0100	0.9262
12	0.0078	0.9340
13	0.0074	0.9414
14	0.0057	0.9471
15	0.0049	0.9520
16	0.0041	0.9561
17	0.0038	0.9599
18	0.0029	0.9628
19	0.0029	0.9657
20	0.0028	0.9685
21	0.0025	0.9709
22	0.0022	0.9732
23	0.002	0.9752
24	0.0019	0.9771
25	0.0017	0.9788
26	0.0016	0.9804
27	0.0015	0.9819
28	0.0012	0.9831
29	0.0012	0.9843
30	0.0011	0.9854
31	0.0009	0.9863
32	0.0009	0.9872
33	0.0008	0.9880
34	0.0007	0.9888
35	0.0007	0.9895
36	0.0006	0.9901
37	0.0006	0.9908
38	0.0006	0.9913
39	0.0005	0.9919
40	0.0005	0.9924

Table 4. Explained Variances of PCA Factors obtained from the PocketPicker Descriptor Matrix. The first 40 factors obtained from the factor analysis of the geometric features of the binding sites of 154 target proteins are shown. 13 factors were retained in accordance with the average variance criterion (i.e., explaining > 1/154=0.65% of the total variance). Cumulatively, they explain 94.1% of the total variance.

Similar to the PCA of the MAF profiles, the performance of the orthogonal rotation procedure in achieving a simple structure was examined through the number of salient loadings for each of the individual factors. Figure 18 shows the distribution of the number of salient loadings across the first 13 factors. As shown by the Figure, it varied between 42 and 75 across the individual factors. Again, similar to the PCA analysis of the MAF profiles, such a distribution of salient loadings reflects a simple structure since the rotated factors contained only a small subset of the original variables.

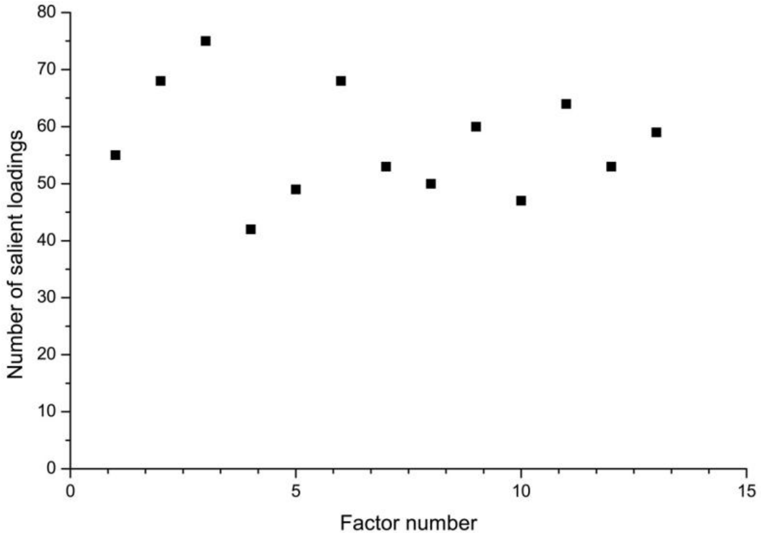


Figure 18. Number of salient loadings across the 13 PCA factors of the PocketPicker descriptor matrix. 13 factors were obtained from the matrix of geometric features of the binding sites of target proteins by PCA. The number of salient loadings (i.e., loadings with a value of $\geq$ 0.4 or $\leq$ -0.4) varied between 45 and 72 for the individual factors which reflect a simple factor structure since the original PocketPicker descriptor matrix contained 405 variables.

2.2.3 Comparison of the factorial structure of molecular affinity profiles and geometric characteristics of protein binding sites

Figure 19 displays superimposed Scree plots based on the MAF fingerprints and the PocketPicker-based geometric descriptors, respectively. As shown by the cumulative variance of MAF factors and PocketPicker factors, explained variances for the PocketPicker factors saturate much faster than for the MAFs. Accordingly, Molecular Affinity Fingerprints consisting of the 1255 energy values for each protein can be described by substantially more parameters (factors) than the set of PocketPicker descriptors. This result reflects the fact that the energy values of the drugs are more heterogeneous as compared to the geometries of the protein pockets, which can be characterized by 13 underlying geometric descriptor factors effectively (with approximately 94% of the variance explained; in contrast to the 55% of the variance explained by the same number of factors for the MAF fingerprints, see Table 3). A similar observation was made by other groups [102, 103] including Favia *et al* who studied the interactions between 27 members of a protein family and approximately 1,000 compounds including their natural ligands. They found that binding affinities vary in a wide range even within clusters of structurally similar molecules, docked to a set of structurally and evolutionary related proteins [103].

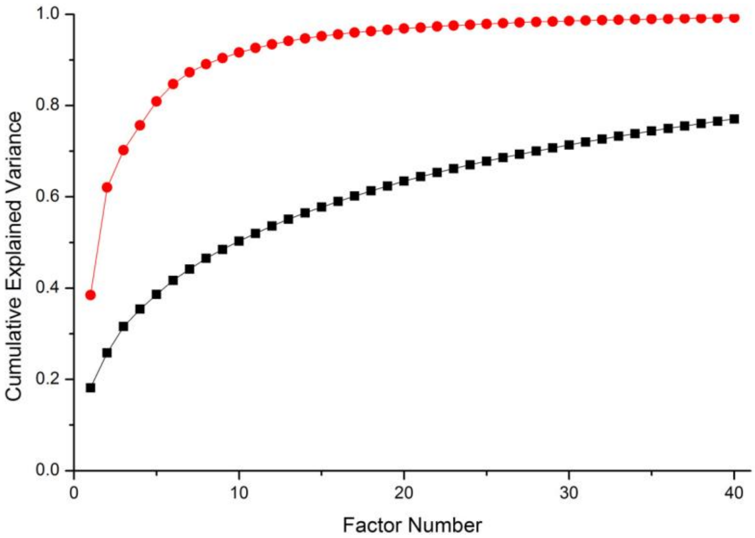


Figure 19. Superimposed Scree plots based on the MAF fingerprints and the PocketPicker descriptors. Cumulative variance explained by the PCA factors for the geometric descriptor matrix based on PocketPicker (circle) saturates much faster than the cumulative variance for the MAF profiles (square), suggesting that the MAF matrix has more complex structure. The first 40 factors of both matrices are plotted.

2.2.4 Canonical Correlation Analysis

Relationship between molecular affinity profiles of target proteins and structural properties of their respective binding sites was investigated by canonical correlation and canonical redundancy analyses. For the purpose of these analyses, factor scores from the set of 30 and 13 factors from the PCA of MAF and PocketPicker descriptors, respectively, were used as input variables.

Canonical Factor Pair	**Canonical R**	**F statistic**	**p**	**Structure of Canonical Factor Pairs**	
				MAF Factor	**PocketPicker Factor**
I.	0.87	2.17	<0.0001	6, 12 , -19	5, 8, 9, 10, 11, 12
II.	0.84	1.74	<0.0001	-7, -15, -16, 28, -30	1, 2, -12
III.	0.77	1.34	=0.0004	-8, 9, 18	-1, 2, 5, -12

Table 5. Canonical correlations and component structure for canonical factor pairs between the MAF and PocketPicker matrices. Canonical correlation analysis between the PCA factors of the MAF profiles of target proteins and the geometric characteristics of their respective binding sites indicated a statistically significant association for 3 pairs of canonical factors. PCA factors of the MAF and the PocketPicker matrices with salient canonical loading (>0.25 or < -0.25) are shown for each of these canonical factor pairs. (Negative signs indicate negative loading.)

Results of CCA indicated statistically significant multivariate relationships between the two sets (Table 5). In particular, the first 3 canonical correlations with a value of 0.87, 0.84, and 0.77, respectively, reached statistical significance. Canonical factor structure for the first three canonical factor pairs is shown in Table 5. As shown by the table, relatively small number of the underlying principal components attain saliency in the canonical factor pairs of the MAF and geometric descriptors (3, 5 and 3 for MAF; 6, 3 and 4 for PocketPicker geometric descriptors) based on the threshold loadings of 0.25 and -0.25 applied for these examinations.

Despite the close multivariate association between the two sets of variables, redundancy analysis indicated that canonical components of MAF factor fingerprints associated with the first 3 canonical correlations explained approximately 15.9% of the total variance of the geometric descriptor factor set (Table 6). Analogously, results of the canonical redundancy analysis revealed that canonical components of the corresponding PocketPicker descriptor factors (associated with the first 3 canonical correlations) explained approximately 6.9% of the total variance of the MAF factor set. In addition, the theoretically possible 13 canonical components with nonzero eigenvalue explained 13% of the total variance of the MAF factor fingerprints; the analogous value for the PocketPicker descriptor factors using 13 canonical components with nonzero eigenvalue was 100%.

Variance of the MAF Variables Explained by					
Canonical Variable Number	**Their Own Canonical Variables**		**Canonical R-Square**	**The Opposite Canonical Variables**	
	Proportion	**Cumulative Proportion**		**Proportion**	**Cumulative Proportion**
1	0.0333	0.0333	0.7638	0.0255	0.0255
2	0.0333	0.0667	0.7122	0.0237	0.0492
3	0.0333	0.1000	0.5852	0.0195	0.0687
4	0.0333	0.1333	0.4275	0.0142	0.0830
5	0.0333	0.1667	0.3403	0.0113	0.0943
6	0.0333	0.2000	0.2952	0.0098	0.1041

7	0.0333	0.2333	0.2362	0.0079	0.1120
8	0.0333	0.2667	0.1811	0.0060	0.1181
9	0.0333	0.3000	0.1238	0.0041	0.1222
10	0.0333	0.3333	0.1168	0.0039	0.1261
11	0.0333	0.3667	0.0833	0.0028	0.1288
12	0.0333	0.4000	0.0180	0.0006	0.1294
13	0.0333	0.4333	0.0129	0.0004	0.1299

Variance of the PocketPicker Variables Explained by					
Canonical Variable Number	**Their Own Canonical Variables**		**Canonical R-Square**	**The Opposite Canonical Variables**	
	Proportion	**Cumulative Proportion**		**Proportion**	**Cumulative Proportion**
1	0.0769	0.0769	0.7638	0.0588	0.0588
2	0.0769	0.1538	0.7122	0.0548	0.1135
3	0.0769	0.2308	0.5852	0.0450	0.1586
4	0.0769	0.3077	0.4275	0.0329	0.1914
5	0.0769	0.3846	0.3403	0.0262	0.2176
6	0.0769	0.4615	0.2952	0.0227	0.2403
7	0.0769	0.5385	0.2362	0.0182	0.2585
8	0.0769	0.6154	0.1811	0.0139	0.2724
9	0.0769	0.6923	0.1238	0.0095	0.2819
10	0.0769	0.7692	0.1168	0.0090	0.2909
11	0.0769	0.8462	0.0833	0.0064	0.2973
12	0.0769	0.9231	0.0180	0.0014	0.2987
13	0.0769	1.0000	0.0129	0.0010	0.2997

Table 6. Results of the Canonical Redundancy Analysis. Proportion of the variance of PCA factor sets (yielded by the MAF and the PocketPicker matrices, respectively) explained by the canonical variates obtained from the same and from the other matrix, respectively. According to the canonical correlation analysis, the first 3 canonical variables reached significance.

Salient components of the three statistically significant canonical factor pairs were examined in order to further interpret our findings.

Factor pair I contained benzodiazepines, barbiturates and morphine derivatives with high positive scores from the MAF side and a fairly homogenous distribution of PocketPicker descriptors associated with low, medium and high values of buriedness and distance (Figure 20a). There were no detectable correlations with short-distance, low-buriedness or distant, highly buried descriptors (white blocks). High negative scores were observed for several drugs including proton pump inhibitors and others that do not form any cohesive groups.

Factor pair II contained phenotiazines, tricyclic antidepressants and certain large molecules (e.g., antibiotics) with negative scores on the MAF factor side while beta-lactams and antiviral agents participated with positive scores in this factor. On the PocketPicker factor side, low and medium buriedness values, associated with low and medium distances, were observed with positive scores. Large distance descriptors in association with medium buriedness levels displayed a negative correlation.

On the MAF factor side of factor pair III, compact molecules (amino acids, tertiary amines, antihistamines) produced positive correlation, in contrast with molecules that have elongated chains which yielded negative correlation. From the PocketPicker side, medium and large buriedness and small/medium distance values obtained positive scores while small (and medium) buriedness values associated with small, medium and especially large distances had a negative correlation.

Overall, because of the abundance of medium/large buriedness and small/medium distance values, we conclude that canonical factor pair III is associated with narrow, deep binding sites. This is supported by the fact that descriptors associated with large distances and low buriedness values have negative correlation. Deep, narrow pockets are in good agreement with the shapes of the drug molecules responsible for the salients of the MAF side of canonical factor pair III since small, compact molecules have positive correlation while elongated compounds have negative correlation. Figure 20b shows the binding pockets of the proteins responsible for the salients on the PocketPicker side. These pockets correlate well with the hypothesized overall shape discussed above. Factor pair II points to medium-sized binding sites as they can be described with small/medium distance parameters and the anticorrelation of parameters coding large distances. Large molecules showed a negative correlation as well; however, this relationship is not as straightforward as in the case of factor pair III. (See Figure 20b for the binding pockets.) Due to the fact that a wide range of PocketPicker descriptors from different classes are represented in the salients of factor pair I, no specific association can be identified in this case. The reason for the suspicious appearance of different structural classes of GABA$_A$-acting pharmaceuticals – e.g. benzodiazepines, barbiturates and morphine derivatives – requires further investigation since the binding pockets of this group possess different shape properties (i.e., elongated and highly branched structures can be found here as well).

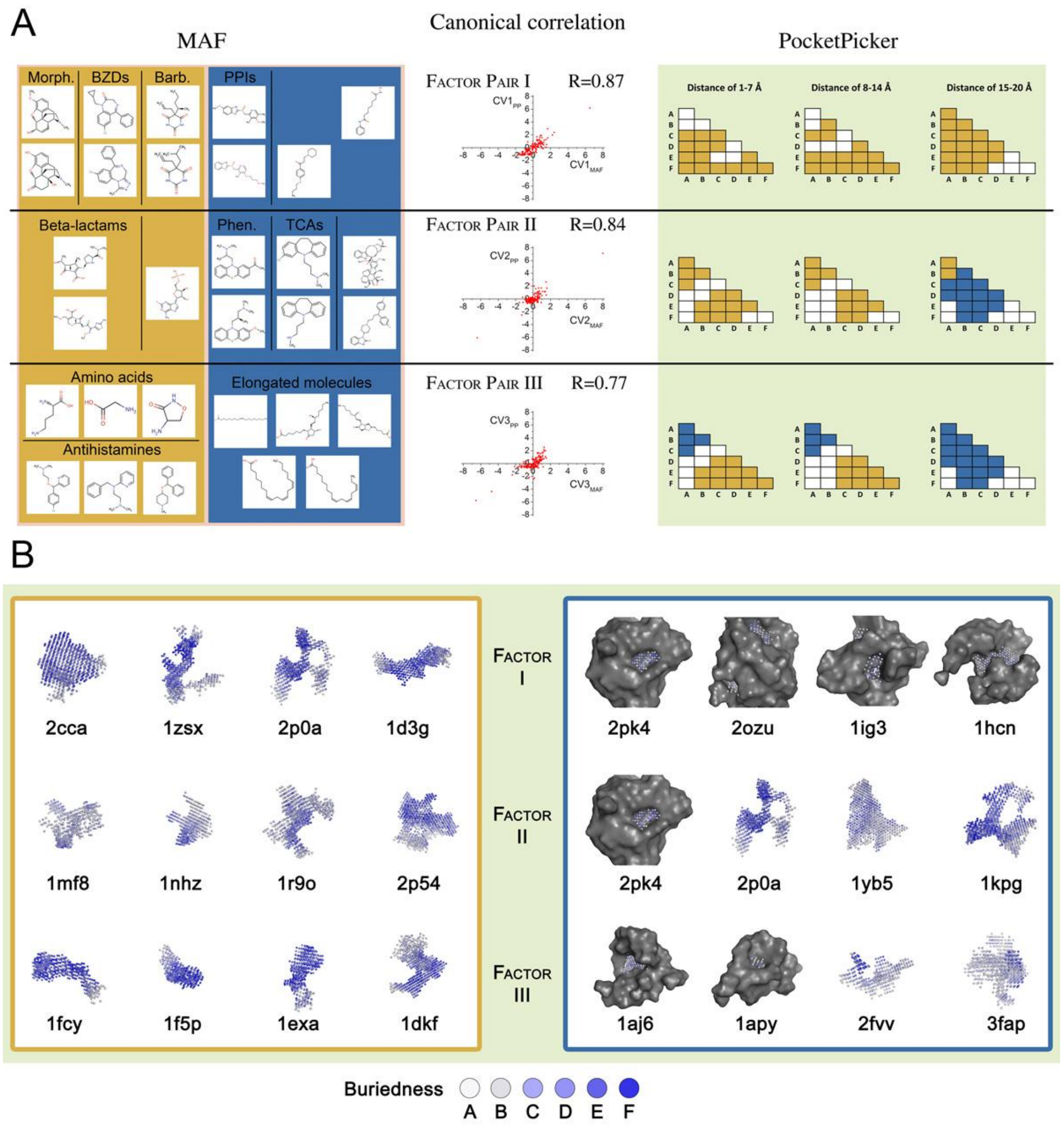


Figure 20. Visual summary of the results of canonical correlation between the MAF and PocketPicker descriptor matrices. **a**. Three statistically significant canonical factor pairs were obtained with the correlation values of 0.87, 0.84 and 0.77, respectively. Canonical correlation (R value) for each factor pair is shown in the middle part. Representative molecules for the MAF factors are shown on the left panel (orange and blue background for positive and negative salients, respectively). Distribution of PocketPicker salients is shown on the right panel. The six different buriedness levels are represented by the letters A-F, with F representing the highest level of buriedness while distance parameters were collected into three groups (1-7 Å, 8-14 Å, 15-20 Å). Orange and blue colors stand for the positive and negative salients, respectively. White blocks represent the absence of a given descriptor pair within a given distance. See text for details of analysis. Abbreviations: BZDs: benzodiazepines; Morph.: morphine derivatives; Barb.: barbiturates; PPIs: proton pump inhibitors; Phen: phenotiazines; TCAs: tricyclic antidepressants. Panel **b**: Shapes of protein binding pockets represented with high scores among the first three canonical factor pairs. Positive and negative salients are represented by orange and blue boxes. Binding site shapes are represented with colored balls positioned in a 1Å-spaced grid with deeper

blue representing a higher level of buriedness. Protein surfaces were removed for better view of the binding pockets in most cases excluding flat, surface sites e.g. 2pk4.

As shown by the figure, proteins of the positive salients of factor III have narrow, deep binding pockets while negative salients contain shallow, small pockets (1aj6, 1apy) and wide, extensive binding sites (2fvv, 3fap). Factor II proteins can be described as having binding sites of medium size and width. Based on the distribution of salient loadings of PocketPicker variables, factor I proteins do not form a coherent group. Elongated (1d3g), branching (1zsx, 2p0a) and bulky binding sites (2cca) belong to this factor.

Our examination of the relation between MAFs and the binding site shape descriptor matrix indicates that the MAF matrix has a complex structure that is correlated with the geometry of the ligand molecules and the protein itself; however, it cannot be sufficiently described by binding site shape descriptors. Binding pocket shape does not play a principal role in the determination of the affinity profiles of proteins except for few specific cases discussed above. Since the MAF profile reflects to the target specificity of ligand binding sites we can conclude that the shape of the binding site is not a key factor to select drug targets. Protein binding sites can be characterized through other more complex descriptors that take additional variables into consideration, for example electrostatic interactions [88, 104]. Our findings are in agreement with a recent study where NMDA receptor antagonists were selected from a library of 8.8 million compounds, applying different virtual screening methods i.e. 2D descriptor-based filtering, molecular docking, QSAR pharmacophore hypothesis and 3D shape search [105]. The best positive hits from each approach were further evaluated by *in vitro* tests. Comparing the four approaches, the 3D-shape-based one gave the worst hit rates while docking produced the highest number of successfully validated compounds.

From another perspective, our results suggest that the shapes of the binding sites could have an impact in virtual drug design for a few drug categories such as morphine derivatives, benzodiazepines, barbiturates and antihistamines, where they strongly correlate with the MAF profiles [98].

2.2.5 Sensitivity Analysis: the importance of different scoring functions

In order to compare the binding free energy values obtained by using AutoDock4 and X-SCORE scoring functions, principle component analyses and canonical correlation analyses were carried out on the two datasets.

Principle component analyses were performed on the IP datasets obtained by using AutoDock4 and X-SCORE scoring functions, respectively. The 1,226 variables (i.e., the drugs) were transformed into 152 factors explaining 99.92% of the total variance in case of the AutoDock4 results. X-SCORE values resulted in 154 factors, explaining 99.16% of the total variance. Consequently, the two datasets show a similar level of complexity.

Canonical correlation analysis between the two datasets revealed 14 significant factor pairs by applying the Kaiser-Gutman eigenvalue criterion. The largest and the smallest obtained correlation values were 0.97 and 0.72, respectively, suggesting that a high degree of correlation exists between the datasets based on the different scoring functions.

Consequently, the complexity of the binding free energy data originated from the two scoring functions is similar and a clear correlation can be seen between them. Based on these results, both scoring functions seem to be suitable for further analyses. However, based on the relatively high acceptance of X-SCORE in the literature [51, 106], this scoring function was chosen.

To determine the robustness of our findings on the importance of shape in binding affinity determination and to study the impact of the applied scoring function on the results, data evaluation was carried out on both datasets. We note that, in contrast to scoring functions used for evaluating docking results, the PocketPicker algorithm shows no stochasticity as it describes binding pockets in a fully reproducible manner while scoring functions are only able to find local minima on the energy landscape, depending greatly on the initial conformation and the applied parameters of searching and scoring methods [48]. Therefore we decided to evaluate the reliability of docking results but not the geometric descriptive method.

There was no significant difference between the canonical correlation analyses based on X-SCORE or AutoDock4 set and the PocketPicker descriptor matrix. Three significant factor pairs were obtained in both cases. For AutoDock4 data, canonical R values were 0.83, 0.70 and 0.66 for the three factor pairs, respectively. The canonical redundancy analyses also revealed consistency between the two approaches. The significant PocketPicker factors explain 8.54% of the variance of the AutoDock4-based MAF factor set while this factor set explains 12.5% of the variance of the PocketPicker descriptors. The results suggest that our principal findings are robust both in terms of the close association and the moderate amount of explained variance observed in the case of the original dataset. In summary, we showed that our findings may reflect from the intrinsic properties of protein binding sites and drug molecules and are not artifacts of the applied methodology. [98]

2.3 Multidimensional IP-EP correlation

In order to match the complex pattern structures of IPs and effect profiles, canonical correlation analyses were performed between them and the basic underlying factor pair that show maximal correlation was identified. We calculated the probability of each effect for each drug based on the drug's IP by linear discriminant analyses, producing a classification function for all effects. As shown by Figure 21, each observed IP was plugged in the classification function in order to generate the drug-effect probability matrix.

To avoid confusion with the previous prediction attempts, this newly introduced multidimensional method is referred to as *Drug Profile Matching* throughout this section.

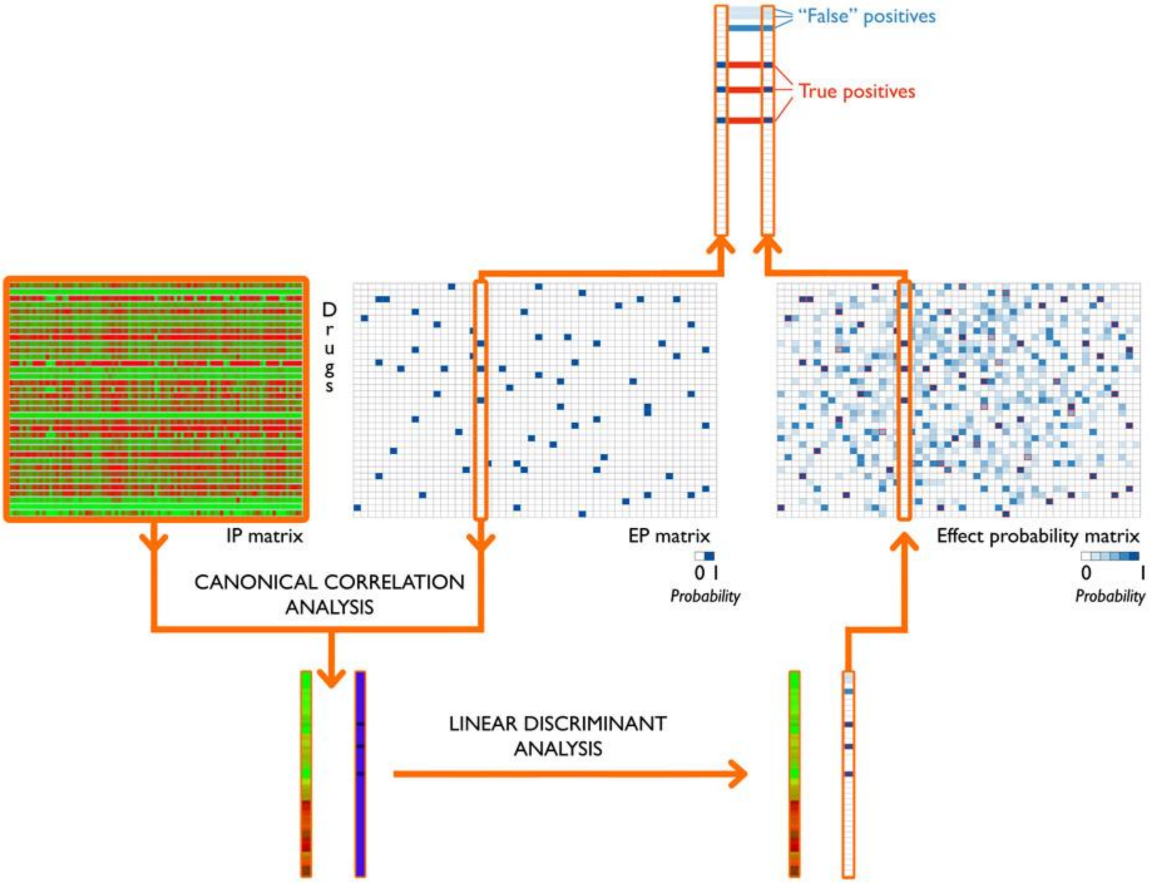


Figure 21. Mechanism of the multidimensional effect prediction. First, canonical correlation analysis is performed on the IP matrix and a selected effect category (bottom arrows), resulting in a vector pair that are the linear combinations of the original interaction matrix and effect vector, respectively. Then, linear discriminant analysis is conducted in order to generate a classification function that calculates the probability that a given drug will possess the studied effect. This analysis is repeated for the studied 181 effect, producing a recalculated effect probability matrix. (Note that probability values in this matrix are continuous.) Comparison of the original and the recalculated vector of an effect category (top arrows) reveals the true positive hits (i.e., the matches;

highlighted with red) and the "false" positives that can be considered as predictions (differences; highlighted with light blue).

2.3.1 Assessing Accuracy

Receiver Operating Characteristic (ROC) curves were examined first in order to quantitatively assess the potential clinical relevance of the drug-effect probability values (Figure 22). ROC analysis characterizes classification performance in terms of true positive rate and false positive rate of drug-effect classification. ROC curves allow the fine-tuning of the detection threshold in order to optimize for TPR and/or FPR. Area Under the ROC Curve, i.e., the AUC value characterizes classification accuracy: an AUC close to 1 indicates high-accuracy classification while a random guess classification would result in a diagonal ROC, yielding an AUC value of 0.5 (see Figure 22a for selected examples).

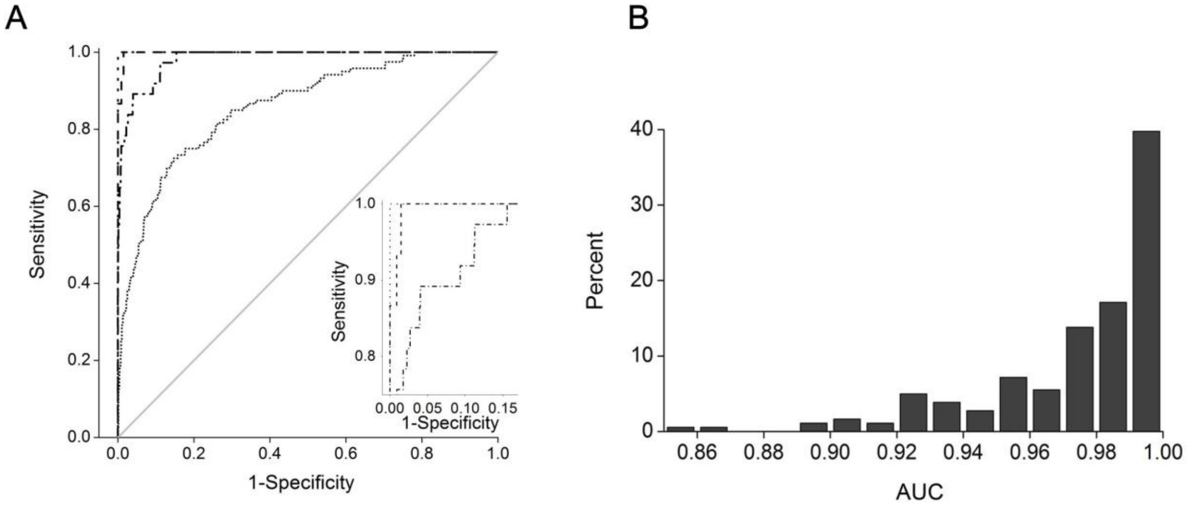


Figure 22a. Representative ROC curves. ROC curve provides a characterization of classification accuracy; here, ROCs of the "Tetracycline" (best classification), "ACE inhibitor", "COX inhibitor" and "Antineoplastic agent" (our most inefficient classification) effect categories are shown (dotted, dashed, dash-dotted and short-dotted lines, respectively). The gray diagonal line represents classification based on random guess. The inset shows an enlarged portion of the upper left region of the plot. Panel **b** shows the AUC histogram, i.e., the distribution of the Area Under the Curve (AUC) values for the studied 181 effects. Results suggest that near-perfect classification was obtained in most cases.

Figure 22b shows the distribution of the AUCs for the entire effect set. 82% of the effects resulted in an AUC value larger than 0.95, indicating that an excellent classification was obtained (see Appendix 4 for the complete list of the studied effects). Certain structure-based effect categories resulted in the best AUC values: progestins, barbiturates, sulphonylureas and tetracyclins (10-17 registered drugs, AUC=1.000 in all cases). This is reasonable and expected since these molecules share a large amount of chemotype features and the pharmacologically relevant classification for this query type has been proved by other groups (e.g. [63]). However, it is much more interesting that molecules showing less chemical similarity but sharing a target also produced outstanding AUC values, e.g. Angiotensin-converting enzyme (ACE) inhibitors (14 registered elements, AUC=0.999) and Serotonin reuptake inhibitors (21 drugs, AUC=0.997). Similarly good results were observed even for physiological effects that can be achieved by different mechanisms. For example, Antiparkinson agents can affect monoamine-oxidase, catechol-O-methyl transferase or the dopamine receptors. This category yielded an AUC of 0.971 (30 registered entries). The case is similar for Vasoconstrictors (42 drugs, AUC=0.963). The lowest AUC values were observed in the case of the two most populated effect categories, i.e., Anti-infective agents (219 registered drugs, AUC=0.869) and Antineoplastic agents (120 registered drugs, AUC=0.859). These categories summarize large groups of distinct effects regarding the mechanisms of action. Nevertheless, even they were classified with an acceptable accuracy. As a summary, we can conclude that our method performs accurate classifications even in case there is no chemical similarity between the compounds registered to a given effect.

We mention that a moderate linear correlation with an R^2 of 0.655 was observed between the number of the registered drugs to an effect and the respective AUC value: the smaller the number of the assigned drugs the higher the AUC value. It is not surprising considering the fact that in a complex system that can be described by many parameters, it is easy to find those parameters that separate a smaller group of observations from the rest of the data in the parameters' multidimensional space. Therefore, AUC solely is not sufficient for the description of the predictive power of the Drug Profile Matching method and a careful independent validation should also be performed.

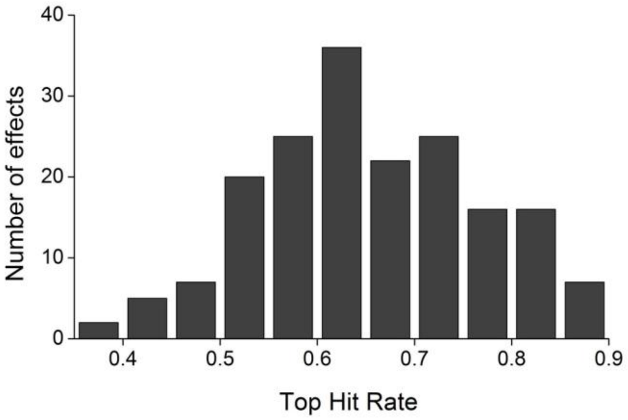


Figure 23. Distribution of the Top Hit Rate values among the 181 studied effects.

From another perspective, we introduced Top Hit Rate as an alternate accuracy measurement. Let us consider an effect ROC curve as it was based upon a list of drugs ordered by descending probability values, regardless of their original FDA effect registration. High classification accuracy value is obtained in case the registered drugs of the given effect appear on the top of the list, meaning a high TPR/FPR value. If we cut the list at the number of the registered drugs to the given effect, irrespective of the fact that true or false positives appeared on the top of the list, we can calculate the ratio of true positives in this top list, i.e., the Top Hit Rate. The distribution of Top Hit Rates can be found on Figure 23. We found that, in average, 66% of the registered drugs appear on this top list. If we consider this number, two thirds of the registered drugs are in the top 2.6% of the list, since in average 32 out of 1,226 drugs belong to an effect.

2.3.2 Leave-one-out validation

To check the validity of the effect classification obtained by Drug Profile Matching, an independent cross-validation with the leave-one-out procedure was performed, similarly to the case of the one-dimensional analysis presented earlier (Figure 24a).

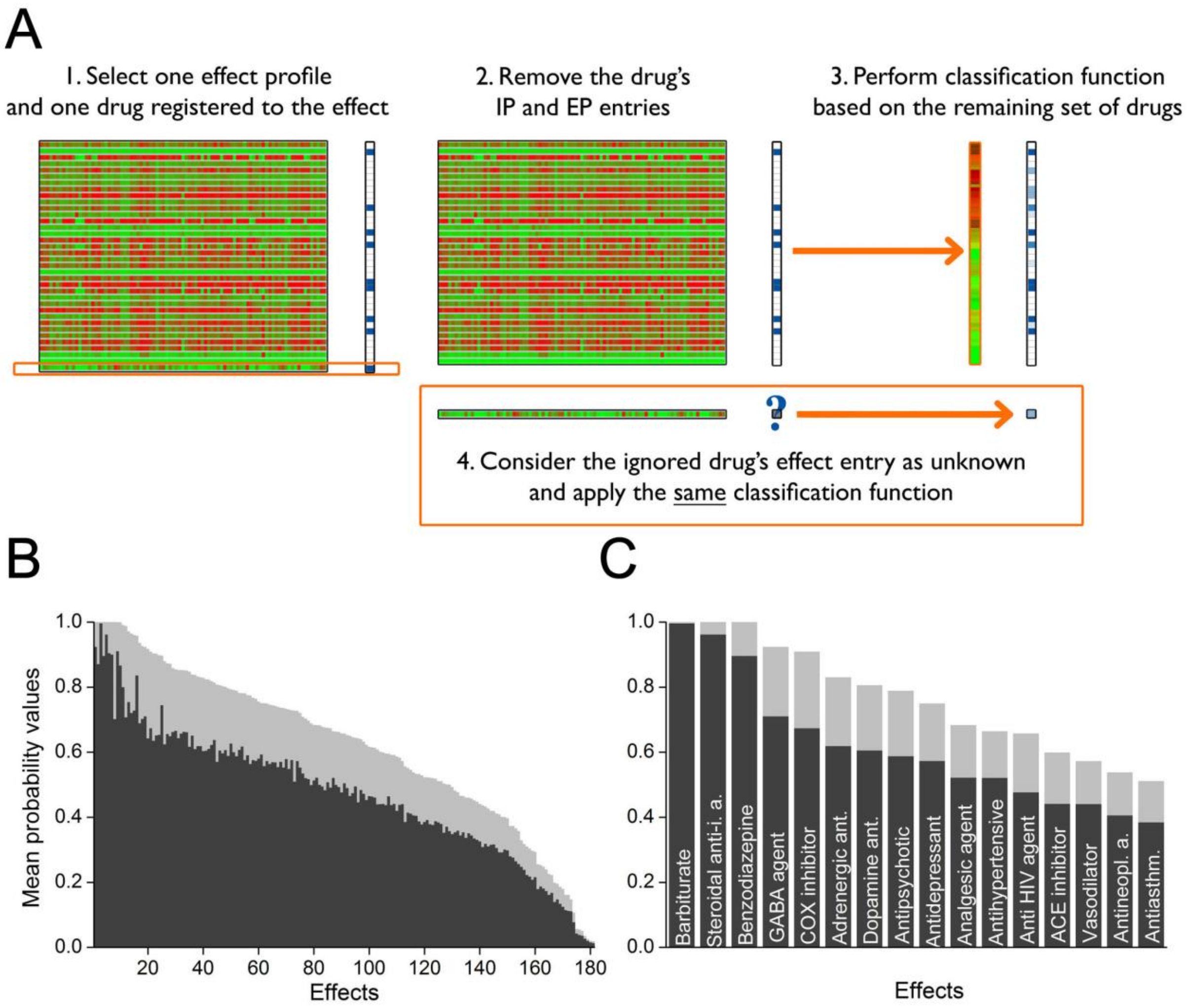


Figure 24. Leave-one-out validation of a selected effect category is schematically summarized on panel **a**. First, the IP and EP entries of a drug originally registered to the studied effect are removed. Then, the effect probability value of this drug is calculated from the classification function based on the remaining set of molecules and the same process is repeated for each drug. Panel **b** shows the mean probability values for the 181 studied effect categories, obtained from leave-one-out validation. Dark gray bars refer to the mean probability values of the whole set of drugs registered to the given effect; light gray bars represent the upper 75%, i.e. the subset performing the best 75% of the calculated probability values. The average of the probability values for all categories is 0.47 which is a high value compared to a randomized EP list which results in an average probability value of 0.026. Panel **c** presents the Mean probability values for some selected effect categories. Dark and light gray bars represent the same values as for the previous panel. Abbreviations: anti-i. a. – anti-inflammatory agent, ant. – antagonist, antineopl. a. – antineoplastic agent, antiasthm. – antiasthmatic agent.

Each drug was examined whether its registered effects can be identified based on its IP profile and a classification function derived from the IPs and effect profiles of the other molecules registered to the studied effect. For each effect we calculated a mean probability value, i.e., the mean of the calculated probabilities for the drugs registered to the given effect (Appendix

3). High obtained mean probability value indicates the method’s robustness that is the resistance of the classification system against the loss of information due to the removal of each molecule entry, one by one, when the classification rules are established during the validation. Figures 24b and c show the mean probability values for the studied 181 effects and some selected examples. The majority (51.9%) of the studied effects are validated by a mean probability value larger than 0.5. In average, this is a 25.1-fold increase compared to a random classification based on the prior probability of each of the 181 effect categories (see Appendix 3 for the values for each effect category). A closer look at the studied effects reveals even 40-80-fold enrichment values compared to a random set.

We observed that for certain effects, a small number of the registered drugs were validated with low probability, which may reflect the existence of subgroups within the effect categories. Therefore, we also present the mean probability values for the upper 75% of the drugs (Figures 24b, c, Appendix 3). We found that, applying this portion of the drugs, 71% of the effects have a mean probability value above 0.5.

If we examine the mean probabilities of different effect categories, the highest values belong to effects based on a high degree of structural similarity among their registered compounds, as expected. E.g., barbiturates, benzodiazepines and steroidal anti-inflammatory agents result in mean probability values of 0.995, 0.895 and 0.961, respectively. However, effect categories based on common target protein still show rather high mean probability values (e.g., 0.673 and 0.605 for cyclooxygenase (COX) inhibitors and dopamine antagonists, respectively), similarly to the observations taken in the previous section. Finally, clinical effect categories encompassing an extensive set of drugs with different mechanisms of action also could be characterized by fairly high mean probability values (e.g., 0.587, 0.573 and 0.520 for antipsychotics, antidepressants and antihypertensive agents, respectively) (Figure 20C, Appendix 3). Mean probability values show no dependence on the number of drugs registered to the effects.

Although many aforementioned studies underline that the presence of target proteins is unnecessary for relevant classification, and it was proved using one-dimensional analysis on our system, it can be hypothesized that the classification function of Drug Profile Matching might be affected by the presence of target proteins. In our dataset, only two known targets are present which are involved in the mechanisms of the 181 effects studied in a multidimensional way: angiotensin-converting enzyme (ACE) and cholinesterase. Examination of the respective classification functions revealed that the canonical loadings of the target proteins for the concerned effects are in the same range than those of the other proteins in the docking set: -0.18 and -0.39 for angiotensin-converting enzyme (1uze; ACE inhibitor effect) and cholinesterase (1p0p; Cholinesterase inhibitor effect), respectively. This finding indicates that the classification functions are practically unaffected by the small number of included targets and this is in a good agreement with our one-dimensional evaluation study (Figure 13c inset).

3. Case studies

Theoretically, four groups of drugs can be distinguished based on structure-activity correlations (Figure 25):

(1) Molecules which show high similarities both in their respective IPs and chemical structures.

(2) Molecules showing similar IPs but small structural similarity. These drugs form a group of structurally unrelated compounds with similar mechanism of action. Revealing such groups is of crucial importance in drug design nowadays, therefore this might become the most interesting and promising application of Drug Profile Matching method.

(3) Molecules with high structural but low IP similarity might hurt the earlier presented finding that structurally similar molecules should possess similar pharmacologic properties. In our study, no such group was identified.

(4) Finally, molecules with low structural and IP similarity show unrelated molecules with both different structures and actions.

In the next section, we introduce three case examples from group 2, revealing the pharmacologic correlations between the compounds.

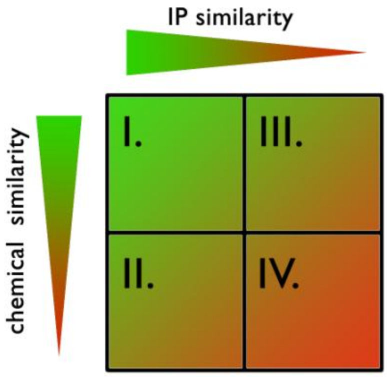


Figure 25. Four types of drugs considering structural and interaction pattern similarities.

3.1 Ziprasidone

In a manner analogous to the effect prediction, we observed that side effects can also be related to the IPs. For example, IP similarities can be used to predict the prolongation of the QT interval, a potentially fatal cardiovascular side effect. (QT is a characteristic part of ECG which contains the QRS complex reflecting to cardiac depolarization and T-wave referring to repolarization. Prolongation of QT interval is a serious side effect which might lead to fatal cardiac arrhythmias e. g. *torsades de pointes*.) Psychiatric patients are at high risk of cardiovascular disease, and therefore it is important to define the adverse reactions of psychiatric drugs which are related to QT prolongation [107]. This side effect has been described previously in case of 9 out of 13 drug molecules in the close IP neighborhood of ziprasidone, a prototypical antipsychotic agent in the FDA database that causes QT prolongation. We investigated whether the pharmacological profile (not only the effect profile) of ziprasidone can be predicted based on the profiles of its thirteen closest IP neighbor molecules. These molecules represent a structurally diverse set; their Tanimoto dissimilarity values to ziprasidone range from 0.754 to 0.973, suggesting a minimal level of chemical similarity. Table 7 summarizes the predicted and the published effects, mechanisms of action and Phase 1 metabolizing isoenzymes (representative enzymes responsible for direct decomposition of drugs in liver) of ziprasidone. We found that two out of the three effects of ziprasidone were found multiple times in the effect list of the neighboring molecules. As an extension of the one-dimensional prediction method, we checked whether the metabolism and mechanisms of actions of ziprasidone can be predicted based on these properties of the neighborhood (data were collected from DrugBank Database). All types of mechanisms and Phase 1 metabolizing isoenzymes were predicted.

	PREDICTED PROPERTY	**PUBLISHED PROPERTY**
EFFECT (DRUG CATEGORY)	Adrenerg Agents	Antipsychotics
	Anesthetics	Dopamine Antagonists
	Anti-anxiety Agents	Serotonin Antagonists
	Antiemetics	
	Antihypertensive Agents	
	Antipsychotics	
	Dopamine Antagonists	
PHASE 1 METABOLIZING ENZYME	Cytochrome P450 3A4 (CYP3A4)	Cytochrome P450 3A4 (CYP3A4)
	Cytochrome P450 2D6 (CYP2D6)	Cytochrome P450 2D6 (CYP2D6)

MECHANISM OF ACTION	Adrenergic Agents	Adrenergic Agents
	Dopamine Antagonism	Dopamine Antagonism
	Histamine Antagonism	Histamine Antagonism
	Serotonin Antagonism	Serotonin Antagonism

Table 7. Prediction of effects and mechanisms of action of ziprasidone. Predictions were based on the pharmacological properties of the thirteen most similar IP-neighbors of ziprasidone. Predicted properties are listed if they appeared on the lists of the effects, Phase 1 metabolizing enzymes and mechanisms of the neighbor molecules multiple times. Two effects out of three were predicted successfully. Almost all hits of unpublished effects are related with the mechanisms and most common adverse reactions of ziprasidone. Note that all four mechanisms of action were predicted correctly. Evidence exists that ziprasidone inhibits CYP2D6 *in vitro* [108].

3.2 Bromodiphenhydramine

The second example is bromodiphenhydramine, an antihistamine and antitussive agent. The *effect-focused* prediction method resulted in the following effect profile for this drug: Norepinephrine-Reuptake Inhibitors (NARIs), Selective Serotonin Reuptake Inhibitors (SSRIs), Tricyclic Antidepressive Agents, Histamine H1 Antagonists, Selective Serotonin Agonists, Antidepressive Agents, Amebicides, Local Anesthetics, Second-Generation Antidepressive Agents, Antitussive Agents. The antihistaminic property of bromodiphenhydramine was validated by one-dimensional and multidimensional approaches, as well as its antitussive properties. Moreover, the definite SSRI/NARI antidepressive profile we predicted was in accordance with the published effect profile of bromodiphenhydramine. Indeed, due to these effects, a structural relative of bromodiphenhydramine was used as a starting point in the development of fluoxetine [109] (Prozac) which has been one of the most popular antidepressive drugs in the world. It is important to point out that these effects of bromodiphenhydramine were successfully identified by our method despite the fact that none of the target proteins were represented on the discriminator surface.

3.3 Valproic acid

The third example is the examination of the neighborhood of the previously presented valproic acid, a well-known, promiscuous anticonvulsant and antiepileptic agent. Metronidazole, a nitroimidazole derivative is the closest neighbor of valproic acid based on IP similarity; it is used for the treatment of infections caused by anaerobic bacteria or protozoa, *Helicobacter pylori* infections in peptic ulcer disease etc. Acetylsalicylic acid, the second closest neighbor, is the best-known non-steroidal anti-inflammatory drug (NSAID) molecule on the market which has analgesic, antipyretic and antirheumatic actions. NSAIDs act as cyclooxygenase (COX) inhibitors [110]. COX converts arachidonic acid into prostaglandins which are involved in physiological (e.g. platelet aggregation) and pathological mechanisms (e.g. inflammation, fever, pain). IP neighborhood of valproic acid shows that 14 out of the closest 30 molecules (IP similarity value ≤ 4.9) belong to the group of NSAIDs (acetylsalicylic acid is the closest one). We analyzed the pharmacological profiles of the first 30 neighbors and found that the higher the degree of IP similarity the more common the effects and adverse reactions of the neighbor drugs with valproic acid. These findings for the first three molecules are summarized in Appendix 4.

4. Experimental confirmations

Using the developed one-dimensional effect prediction methods and the multidimensional classification functions, probability values were assigned for each drug-effect pair in our dataset. For many drugs, a number of unregistered effects were detected with high probability. These "false positive" hits can be indicative of hidden effects which potentially could be used for new drug effect predictions. In order to test these findings, in two selected effect categories all predictions exceeding a certain probability threshold were verified by *in vitro* tests and literature data. For two other compounds, cell culture tests were carried out to justify our predictions.

4.1 ACE Inhibitors

First, the inhibition of ACE was selected to investigate the predictive power of Drug Profile Matching, our multidimensional evaluation system. ACE inhibitors are widespread antihypertensive agents also used for the treatment of congestive heart failure and diabetic nephropathy [111, 112].

The following criteria were considered in the selection of this effect:

(1) robustness and accuracy values of classification functions,

(2) the importance of the therapeutic effects and

(3) availability of an *in vitro* test kit.

The effect category "ACE Inhibitor" is a good representative of the upper region of classification accuracy (0.998) while its robustness value belongs to the medium range (0.44). For the ROC curve, see Figure 22a. Its respective enrichment value, i.e. the mean/random mean is 36 (Appendix 4).

Here, the prediction acceptance threshold was set to a level above which all 14 originally registered drugs were classified as positive. 19 "false" positives appeared among them (Figure 26a, c, Table 8). Retrospective literature analysis revealed that for 3 of the predicted compounds, i.e., candoxatril, carvedilol and nebivolol, the effect of interest was described earlier [113-116]. L-proline, tipranavir, dasatinib, novobiocin, nelfinavir and telmisartan showed the predicted activity in the *in vitro* tests, resulting in 20-97% inhibition at 500 μM.

ACE inhibition curves were determined for the three strongest agents, i.e., telmisartan, L-proline and novobiocin. The strongest ACE inhibitory activity (K_d = 6 μM) was observed for telmisartan which is registered as an angiotensin II receptor antagonist, without mentioning it as an ACE inhibitor in the literature. The observed K_d value to ACE is comparable to the peak plasma concentration of telmisartan which is around 10μM, according to [117].

Interestingly, L-proline also produced a definite activity with a K_d of 86 μM. Visual inspection of the chemical structures of the common ACE inhibitors e.g. captopril reveals that they contain a proline moiety; however, there is no published evidence that would support the importance of this moiety in their pharmacologic actions.

The aminocoumarin antibiotic novobiocin possessed a moderate ACE inhibition (K_d = 167 μM).

Altogether, 60% of the ACE inhibitory predictions were confirmed by literature and *in vitro* tests.

DB code	Drug name	Class	Probability	Tested	Active	Inhibition % at 500 μM	Ref.
DB00722	Lisinopril	1	1.000				
DB00542	Benazepril	1	1.000				
DB00519	Trandolapril	1	1.000				
DB00966	Telmisartan	0	1.000	+	+	97,2	
DB00492	Fosinopril	1	0.999				
DB00790	Perindopril	1	0.997				
DB01340	Cilazapril	1	0.994				
DB01089	Deserpidine	1	0.993				
DB00881	Quinapril	1	0.894				
DB00691	Moexipril	1	0.875				
DB01197	Captopril	1	0.779				
DB00584	Enalapril	1	0.624				

DB01348	Spirapril	1	0.510				
DB01229	Paclitaxel	0	0.419	+	-		
DB04570	Latamoxef	0	0.410	+	-		
DB00172	L-proline	0	0.384	+	+	93,3	
DB04835	Maraviroc	0	0.369	+	-		
DB00932	Tipranavir	0	0.293	+	+	20*	
DB01254	Dasatinib	0	0.175	+	+	46,1	
DB01051	Novobiocin	0	0.157	+	+	74,5	
DB00686	Pentosan polysulfate	0	0.141	-			
DB00220	Nelfinavir	0	0.101	+	+	47,9	
DB00178	Ramipril	1	0.087				
DB01122	Ambenonium	0	0.076	+	-		
DB00616	Candoxatril	0	0.066	-	+		*[113]*
DB01136	Carvedilol	0	0.060	-	+		*[114]*
DB06267	Udenafil	0	0.057	-			
DB00637	Astemizole	0	0.055	-			
DB00698	Nitrofurantoin	0	0.049	+	-		
DB01344	Polystyrene sulfonate	0	0.048	-			
DB00766	Clavulanate	0	0.036	+	-		
DB04861	Nebivolol	0	0.029	-	+		*[115, 116]*
DB01180	Rescinnamine	1	0.021				

Table 8. Results of the ACE inhibitory effect classification function and the *in vitro* tests. Class 1 is formed by the drugs originally registered as ACE inhibitors (14) while Class 0 entries are false positive hits (19) that were further examined. For three cases, the effect identified by Drug Profile Matching is verified by the literature. Four of the predictions were excluded from the experiments due to commercial unavailability or withdrawal from the market. Six of the twelve compounds suitable for the tests showed inhibitory effect on ACE. Inhibition% values measured at 500 μM are displayed. Each data is an average of two independent measurements. A star refers to the uncertainty of the data point, originated from solubility issues. Positive hits confirmed by literature or *in vitro* tests are highlighted with red.

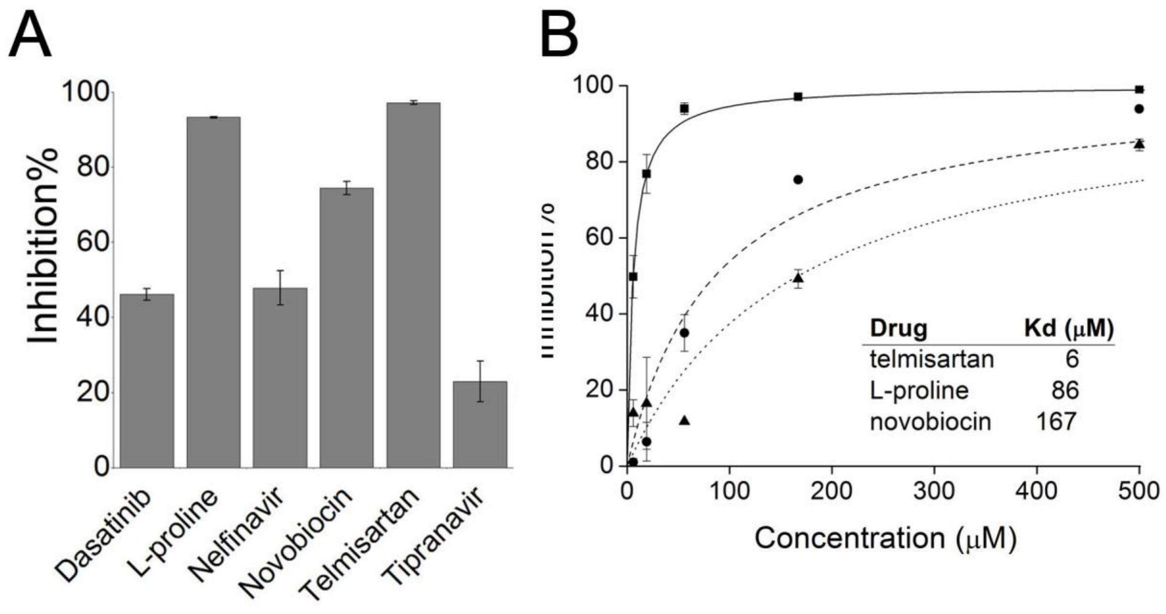


Figure 26. *In vitro* ACE inhibition test results. **a.** ACE inhibition values with standard deviation for the tested and active compounds. 500 μM drug concentrations were applied in each case. Altogether, six of the twelve compounds suitable for the tests showed inhibitory effect on ACE. P**anel b**. ACE inhibition curves for the three compounds possessing the highest inhibition activity: telmisartan (square; solid line), L-proline (circle; dashed line) and novobiocin (triangle; dotted line).

4.2 COX Inhibitors

COX inhibitors possess anti-inflammatory activity and are also used worldwide (*23*). This effect category yielded an AUC of 0.982 (Figure 22a) and a mean probability value of 0.673 (Appendix 3). These values are in the upper region therefore they make this effect category as a good example to evaluate the predictions. The enrichment value for COX inhibitors is 22.3 (the random dataset resulted in a mean probability value of 0.03).

In case of COX inhibitors, the prediction threshold was set to a level above which 90% of the registered COX inhibitors appeared as positives (33 out of 37, Table 9). Among them, 54 compounds were considered as “false” positives. The COX inhibitory properties for valproic acid, alpha-linolenic acid, oxybenzone and ciclopirox were confirmed in the literature [78, 118-121]. Valproic acid was described as a selective COX-2 inhibitor [78]. Two other compounds, ticlopidine and azathioprine are known as “tested but inactive” [122, 123]. Eleven drugs were excluded from tests due to lack of commercial availability or limited importance.

Totally, 39 compounds were tested for COX inhibition activity and 18 drugs yielded positive results. Since the studied effect category does not specify which of the two isoenzymes is affected by the compound, COX-1 and COX-2 isoforms were also tested for all cases. Overall, 47% of the predicted COX inhibitors showed the predicted activity (Figure 27b, d). Nitroxoline, alpha-linolenic acid, captopril, flutamide and nilutamide were found to be the strongest inhibitors in the test set.

COX inhibition curves of captopril (COX-1 and COX-2), nitroxoline (COX-2) and alpha-linolenic acid (COX-1) were determined. Captopril showed a reduced COX-1 and COX-2 inhibitory effect at high concentrations; its K_d values were 18 and 13μM for the two COX isoenzymes. This is comparable to the results of the classical COX inhibitor aspirin determined in our laboratory earlier ($K_{d,\,COX\text{-}1}$=62μM and $K_{d,\,COX\text{-}2}$=52μM; data not shown).

Linolenic acid showed a K_d of 4 μM for COX-1 and based on the inhibition results measured at 500 μM for COX-1, this compound possesses a strong, non-selective COX inhibition. This finding fits well to the results of Ren and Chung [120, 121]. Here, the authors proved that this compound has an anti-inflammatory effect through different mechanisms, including COX-2 inhibition while COX-1 was not mentioned. Thus, we extended the knowledge about the multi-target anti-inflammatory properties of alpha-linolenic acid.

Nitroxoline, a special antibiotic showed a K_d of 1 μM for COX-2 and also seems to be an extremely strong non-selective COX inhibitor.

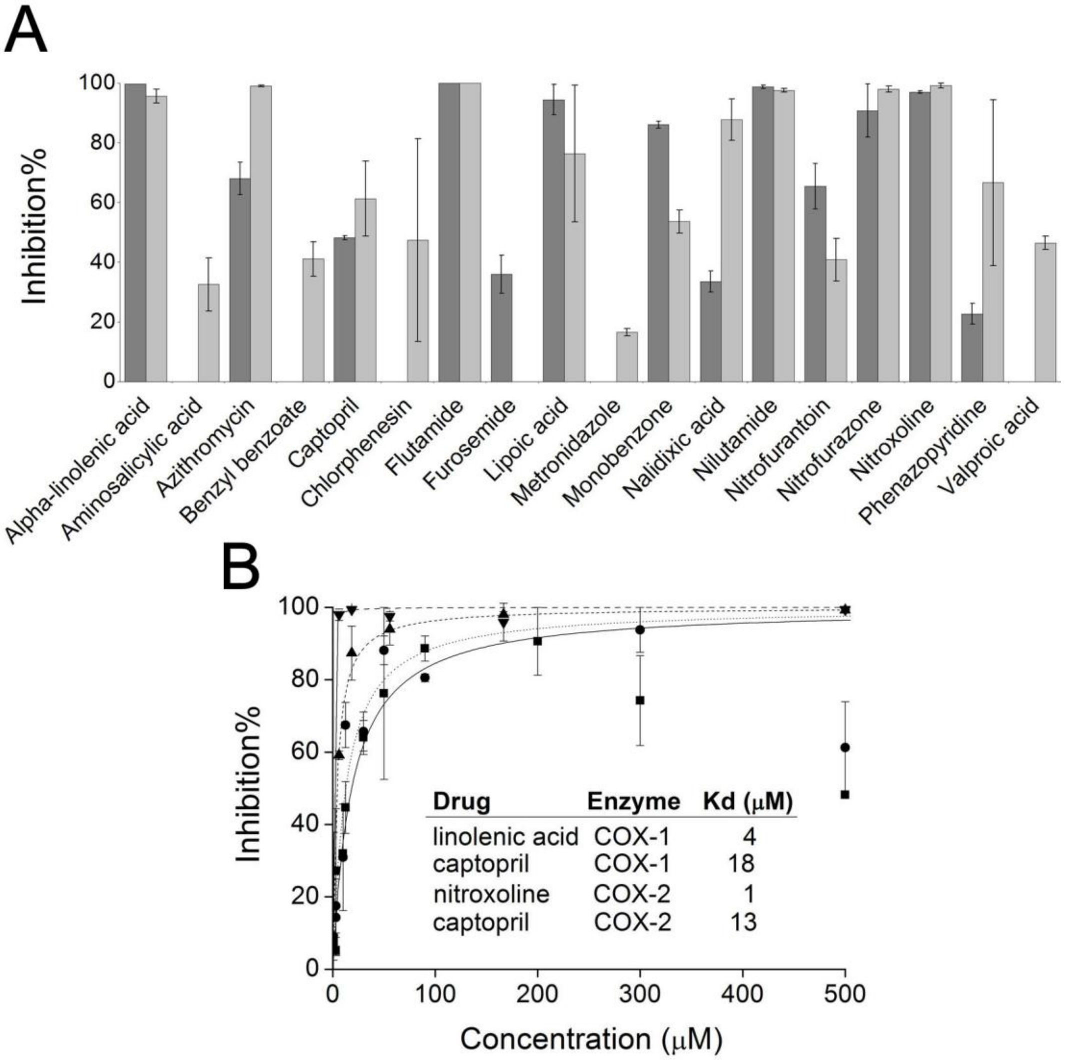
Valproic acid was also tested in order to reproduce the literature-based result [78]. It was confirmed that the compound possesses a moderate, selective COX-2 inhibition.

						Inhibition % at 500 μM		
DB code	**Drug name**	**Class**	**Probability**	**Tested**	**Active**	**COX1**	**COX2**	**Ref.**
DB00936	Salicyclic acid	1	1.000					
DB00784	Mefenamic acid	1	1.000					
DB00244	Mesalazine	1	1.000					
DB01600	Tiaprofenic acid	1	1.000					
DB00573	Fenoprofen	1	1.000					
DB01399	Salsalate	1	1.000					
DB00861	Diflunisal	1	1.000					
DB04552	Niflumic acid	1	1.000					
DB00939	Meclofenamic acid	1	1.000					
DB00586	Diclofenac	1	1.000					
DB01283	Lumiracoxib	1	1.000					
DB01009	Ketoprofen	1	1.000					
DB01250	Olsalazine	1	1.000					
DB00712	Flurbiprofen	1	1.000					
DB00465	Ketorolac	1	1.000					
DB00121	Biotin	0	1.000	+	-			
DB00788	Naproxen	1	1.000					

DB00945	Aspirin	1	1.000					
DB00821	Carprofen	1	1.000					
DB00233	Aminosalicylic acid	0	1.000	+	+	0	32,5	
DB00870	Suprofen	1	1.000					
DB01014	Balsalazide	1	0.999					
DB00499	Flutamide	0	0.999	+	+	100	100	
DB00336	Nitrofurazone	0	0.999	+	+	90,8	98,0	
DB00991	Oxaprozin	1	0.999					
DB00313	Valproic acid	0	0.998	+	+	0	46,5	*[78]*
DB00500	Tolmetin	1	0.998					
DB00166	Lipoic acid	0	0.998	+	+	94,5	76,4	
DB01050	Ibuprofen	1	0.998					
DB00963	Bromfenac	1	0.997					
DB00600	Monobenzone	0	0.997	+	+	86,1	53,6	
DB01241	Gemfibrozil	0	0.996	+	-			
DB00323	Tolcapone	0	0.995	-				
DB00749	Etodolac	1	0.987					
DB00814	Meloxicam	1	0.982					
DB00676	Benzyl benzoate	0	0.981	+	+	0	41,1	
DB00461	Nabumetone	1	0.979					
DB00817	Rosoxacin	0	0.979	-				
DB00328	Indomethacin	1	0.975					
DB00695	Furosemide	0	0.967	+	+	36,0	0	
DB01099	Flucytosine	0	0.966	+	-			
DB01053	Penicillin G	0	0.965	+	-			
DB01423	Stepronin	0	0.962	-				
DB01178	Chlormezanone	0	0.947	+	-			
DB00614	Furazolidone	0	0.940	+	-			
DB01607	Ticarcillin	0	0.932	+	-			
DB01422	Nitroxoline	0	0.922	+	+	97,0	99,2	
DB00911	Tinidazole	0	0.900	+	-			
DB00345	Aminohippurate	0	0.877	-				
DB00482	Celecoxib	1	0.860					
DB01206	Lomustine	0	0.837	+	-			
DB00208	Ticlopidine	0	0.820	-	-			*[123]*
DB00946	Phenprocoumon	0	0.819	-				
DB01168	Procarbazine	0	0.809	-				
DB01428	Oxybenzone	0	0.792	-	+			*[119]*
DB00406	Gentian violet	0	0.791	-				
DB00554	Piroxicam	1	0.775					
DB00665	Nilutamide	0	0.771	+	+	98,8	97,6	
DB00235	Milrinone	0	0.744	+	-			
DB01188	Ciclopirox	0	0.742	-	+			*[118]*
DB00132	Alpha-linolenic acid	0	0.738	+	+	99,8	95,7	[120, 121]
DB00291	Chlorambucil	0	0.694	+	-			

DB01424	Aminophenazone	1	0.679					
DB01438	Phenazopyridine	0	0.646	+	+	22,7	66,7	
DB00417	Penicillin V	0	0.644	+	-			
DB00815	Sodium lauryl sulfate	0	0.617	-				
DB01025	Amlexanox	0	0.600	-				
DB00207	Azithromycin	0	0.582	+	+	68,1	99,1	
DB00286	Estrone sulfate	0	0.577	+	-			
DB00385	Valrubicin	0	0.569	-				
DB00903	Ethacrynic acid	0	0.567	+	-			
DB00578	Carbenicillin	0	0.565	+	-			
DB00916	Metronidazole	0	0.558	+	+	0	16,5	
DB00731	Nateglinide	0	0.541	+	-			
DB00583	L-carnitine	0	0.510	+	-			
DB00459	Acitretin	0	0.508	+	-			
DB00307	Bexarotene	0	0.505	-				
DB00779	Nalidixic acid	0	0.504	+	+	33,5	87,8	
DB00605	Sulindac	1	0.499					
DB00172	L-proline	0	0.476	+	-			
DB00316	Acetaminophen	1	0.472					
DB00993	Azathioprine	0	0.441	-	-			*[122]*
DB00114	Pyridoxal phosphate	0	0.436	+	-			
DB00698	Nitrofurantoin	0	0.424	+	+	65,5	40,8	
DB01197	Captopril	0	0.422	+	+	48,2	61,3	
DB00856	Chlorphenesin	0	0.418	+	+	0	47,4	
DB00168	Aspartame	0	0.402	+	-			

Table 9. Results of the COX inhibitory effect classification function and the *in vitro* tests. Drugs belonging to Class 1 are originally registered as COX inhibitors (33) while Class 0 entries are false positive hits (54) that were further examined. Percent inhibition values at 500 µM on both isoforms are listed, each data is an average of two independent measurements. Positive hits confirmed by literature and/or *in vitro* tests are highlighted with red.



Drug	Enzyme	Kd (μM)
linolenic acid	COX-1	4
captopril	COX-1	18
nitroxoline	COX-2	1
captopril	COX-2	13

Figure 27. *In vitro* COX inhibition test results. **Panel a**. COX inhibition values with standard deviation for the tested and active compounds. 500 μM drug concentrations were applied in each case. Inhibitory effects on COX-1 and COX-2 enzymes are marked with dark and light gray bars, respectively. A missing bar signifies no observed activity. After excluding the known COX agents and the commercially unavailable drugs, 39 compounds were tested for COX inhibition activity. The 18 presented drugs yielded positive *in vitro* results. **Panel b.** COX inhibition curves for alpha-linolenic acid (COX-1) (up-triangle mark with short-dashed line), nitroxoline (COX-2) (down-triangle mark with dashed line) and captopril (COX-1, COX-2) (square mark with solid line and circle mark with dotted line, respectively).

4.3 Cell culture tests

Drugs developing adrenergic and dopaminergic effects are in the special focus of interest of our collaborators. We therefore selected amiloride and minoxidil for *in vivo* screens on the basis that they were the only adrenergic and dopaminergic predictions made by the one-dimensional *effect-focused* prediction method. They resulted in high prediction confidence values (C_P > 95% and 90% for each activity of amiloride and minoxidil, respectively) using the *neighbor-focused* prediction method as well. D_1, D_2, α_{1B}, α_{2A} and β_1 receptor agonist and antagonist activities were investigated in an independent study performed by EuroScreen, Ltd. (Figure 28).

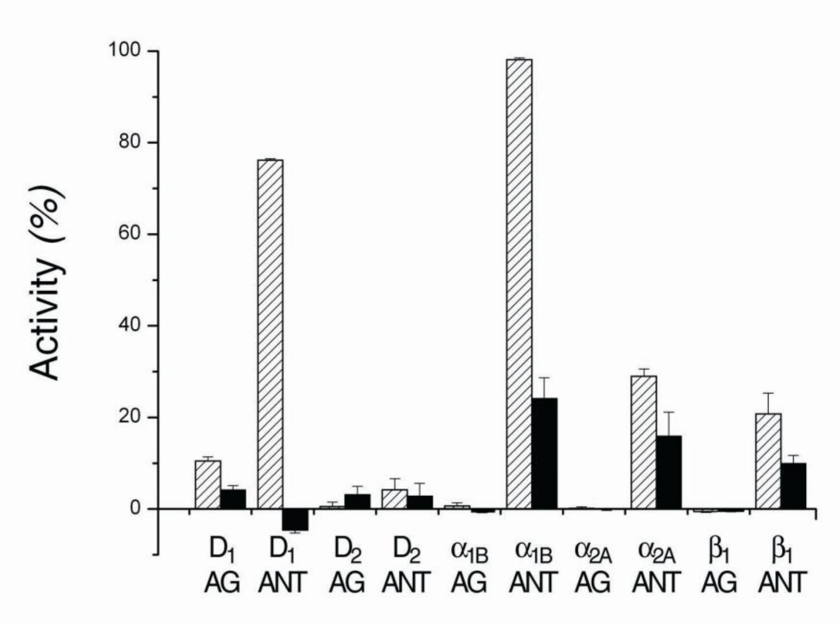


Figure 28. *In vivo* test results of amiloride and minoxidil. Different dopaminergic and adrenergic activities determined for amiloride (stripped bars) and minoxidil (filled bars). Abbreviations: AG – agonism, ANT – antagonism.

Amiloride, a sodium channel blocker diuretic agent produced strong α_{1B} antagonism (98 %, applying 50 μM of drug; K_d = 13.7 μM) and smaller D_1, α_{2A} and β_1 antagonism (76, 29 and 21 % respectively; applying 50 μM of amiloride except the D_1 antagonism test where 100 μM was used; $K_{d\ D1}$ = 51.2 μM). On the other hand, the peripheral vasodilator minoxidil showed some level of α_{1B} and α_{2A} antagonist effect (24 and 16 %, respectively; applying 25 μM of drug).

It is noteworthy that none of these effect categories were overrepresented in our database; specifically, adrenergic and dopaminergic effect related categories were registered in the unrefined effect list 64 (2.77 %) and 30 (1.30 %) times, respectively.

5. Summary

The presented examples highlight several important features of our method:

1. A clear association was revealed between IP and EP datasets, irrespective of the complexity of the applied evaluation method, i.e., one- or multidimensional data analyses.

2. The method allows a highly efficient identification of similarities in effects and mechanisms of action despite structural diversity.

3. The diversity of the applied protein set was large enough to obtain significant correlation between the IP and EP datasets.

4. Different scoring functions provided similar binding affinity patterns in terms of the level of canonical correlation between the IPs based on the different scoring functions and the binding site geometry descriptor data). A high level of correlation was obtained for the AutoDock4 and the X-SCORE-based IP set.

5. Target proteins are not necessarily for effect prediction as it was proved using one-dimensional (validation results based on an alternate, reduced protein set) and multidimensional analyses (canonical loading analysis of target proteins in the classification functions).

6. Several case studies, *in vitro* and cell culture tests were carried out to evaluate our predictions. Our approach, even applying the one-dimensional evaluation, was able to predict pharmacokinetic properties of ziprasidone, pointing to the possible future expansion of the method from PD towards PK event prediction. More than 50% of the ACE and COX inhibitory predictions were confirmed *in vitro*. The predicted adrenergic and dopaminergic profile of two compounds were also confirmed.

As a conclusion, our analyses revealed a strong predictive power of the IP-based effect prediction.

Part II: Internal viscosity:
the role of hinge residues in trypsin activation

Introduction

In Part II, the complex problem of protein flexibility was studied in the model system of the activation of human trypsin 4. During activation, four distinct regions of the protein undergo conformational change. This conformational rearrangement is a single-step, irreversible reaction in the applied conditions and it is coupled with an intrinsic fluorescence signal change, making the trypsin activation a suitable model system for assessing the question of flexibility. The protein parts involved in the conformational change, i.e., the activation domain is bordered by hinge glycines that secure the large conformational freedom needed for the rearrangement. In this study, the kinetic and thermodynamic parameters of the activation were modified by introducing side chains with different sizes at a hinge position. We examined the temperature and external viscosity dependence of the rate of the conformational change and evaluated the results applying Kramer’s theory. Based on this, a measure reflecting protein flexibility, i.e., the internal viscosity parameter of the activation of the wild type trypsin and two hinge mutants were determined. Our results suggest that the flexibility of the studied protein can be modulated by introducing point mutations in the hinge region.

Trypsin, a prototype of the S1 family of the serine proteases, is synthesized in an inactive zymogen form and is activated by proteolytic cleavage of the activation peptide. The α-amino group of the newly formed N terminus, Ile16 (chymotrypsinogen numbering) forms a stabilizing salt bridge with the Asp194 side chain carboxylate group which triggers a conformational change leading to the active enzyme [124-126]. This structural change affects a distinct region of the protein, i.e., 15% of the molecule, while 85% of the structures of the active and inactive forms are identical (Figure 1). Four peptide segments, collectively referred to as the activation domain, undergo large conformational rearrangement: 16-19, 142-152 (the autolysis loop), 184-194 and 216-223 (these latter two form the substrate binding pocket and the oxyanion hole). The conformational change completes the formation of the oxyanion hole.

This structural rearrangement can also be triggered by a pH-jump from pH 11.0 to pH 8.0 [127-129] and monitored by measuring the intrinsic fluorescence of the enzyme [130].

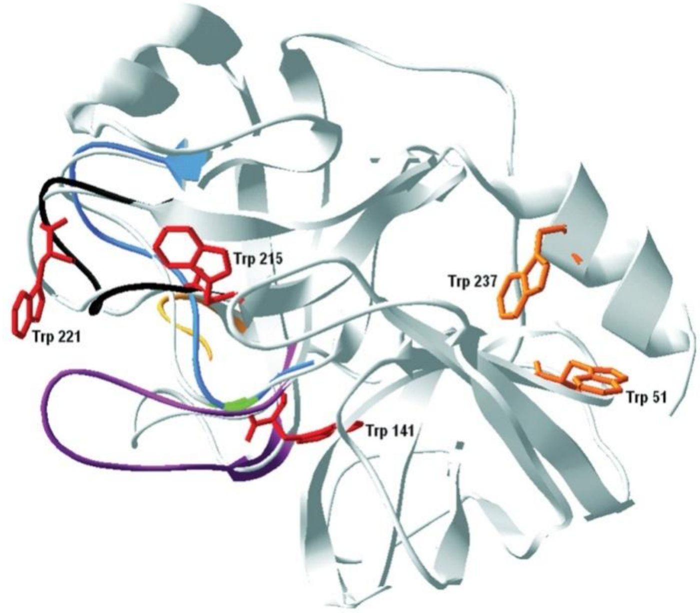


Figure 1. Superimposed structure of bovine trypsinogen (PDB ID: 1tgn) and human trypsin 4 (PDB ID: 1h4w) visualized using software DeepView-spdv 3.7 [131]. The peptide backbone segment of human trypsin 4 whose conformation differs from the conformation of bovine trypsinogen (gray) corresponding to the activation domain is shown as a colored ribbon. The 16-19 peptide segment is shown in yellow, the 142-152 segment is colored purple and the 184-194 and 216-223 segments are represented by a blue and a black ribbon, respectively. The backbone of residue 193, which goes through large dihedral angle transition in the course of activation is shown in green. Tryptophan residues are also highlighted, Trp141, 215 and 221 that might account for the fluorescence intensity change during the conformational change are colored red, while Trp51 and Trp237 is shown in orange (chymotrypsin numbering system is used to identify the residues).

Targeted molecular dynamics simulations of trypsinogen to trypsin transition showed that the largest changes in main chain dihedral angles occur at certain glycine residues among which Gly19, Gly142, Gly184, Gly193 and Gly216 border the activation domain peptides [132]. These glycine residues exhibit larger Φ and/or Ψ angle changes than the surrounding residues, as due to the absence of a side chain the rotation around the Cα-C and C-N bonds becomes energetically most favorable. Based on this finding, it is presumable that these glycines play an important role in the activation process, acting as hinges for the conformational change and the 4 peptide segments move as more rigid units. The presence of glycines at conserved positions of the activation domain seems to promote the conformational transition. As a consequence, replacement a hinge residue, e.g. Gly193 by an amino acid possessing a bulky side chain is supposed to have significant effect on the rate and thermodynamics of the conformational change upon activation. Position 193 is well conserved among serine proteases: it is occupied by a glycine residue except for rare examples. One of these exceptional enzymes is human trypsin 4 possessing an arginine at this position.

The sequence of the main substeps upon activation of bovine and rat chymotrypsinogen was deduced based on molecular dynamics simulations [133]. The conformational changes in the backbone of residue 192 trigger the reorientation of Gly193 towards the substrate binding site and a rotation around the C_α-C bond of Asp194 of ~180°. Consequently, a cavity is left behind that allows the penetration of Ile16 into the core of the molecule and the formation of the salt bridge between Ile16 and Asp194.

In order to study protein flexilibity in an experimental system with reduced dimensionality, we sought for a system where the change in the protein conformation is a single-step first-order transition, accompanied by an intrinsic signal change in the protein. The rearrangement of the activation domain of human trypsin 4 upon activation meets all these criteria and this enzyme is biochemically well characterized [134]. Our aim was to study the effect of point mutations R193G/A/Y/F in the hinge on the rate of the conformational change and characterize the thermodynamics of this structural rearrangement. We expressed wild type human trypsinogen 4 and its R193G/A/Y/F mutants and monitored their conformational change upon activation in pH-jump stopped flow experiments by detecting the intrinsic fluorescence change. We found that this conformational transition is highly affected by the mutations at position 193, and that its rate constant decreases with the size of the sidechain. We also studied the temperature dependence of the rate constant of the transition in the 5-38°C range with a conventional stopped flow apparatus. We developed a new heat-jump/stopped-flow setup which allowed the extension of the Arrhenius plots up to 60°C. Due to the relatively narrow temperature tolerance of enzymes, Arrhenius and van't Hoff plots can be determined only in narrow temperature ranges. The consequence is that the accuracy and confidence level of the calculated thermodynamic parameters are very low. In contrast, by using our newly developed technique [135], Arrhenius plots can be determined in a wider temperature range. This apparatus was applied to perform transient kinetic measurements between 34 and 60°C. [136]

Thermodynamic analysis revealed that the R193G/A/Y/F mutants differ only in the pre-exponential term of the Arrhenius equation, while the activation energy is unaffected by the mutations. Solvent viscosity dependence of the rate of the conformational transition of the R193G and R193A mutants was also determined. It was revealed that the rate of is inversely proportional to the solvent viscosity. This phenomenon is interpreted in terms of the Kramers' theory. Based on our results, we conclude that the rate of conformational change during activation of trypsinogen site 193 mutants is determined by the internal molecular friction around this hinge site. [136]

Aims

The main aim of this work was to characterize the thermodynamics of a special monomolecular structural rearrangement and to unravel the role of the internal viscosity by introducing residues with different sizes at a specific hinge point [136].

The development of a novel heat-jump/stopped flow was our secondary aim since using this equipment, the available temperature range for the studied enzyme reaction can be extended and the accuracy of Arrhenius plots can be improved [135].

Materials and methods

1. Development of heat-jump/stopped-flow

1.1 Heat-jump/stopped-flow apparatus

The modified stopped-flow apparatus is based on a KinTek 2004 Stopped Flow apparatus. The schematic representation of the heat-jump/stopped-flow can be seen in Figure 2. The apparatus is equipped with two thermo-controllers (Supertech STC05A). One of them is responsible for adjusting the temperature of the cuvette while the other adjusts the temperature of the so-called heating loop. The heating loop is a 55 μL Teflon loop built in a heating element (resistor based heating element, max 12V, 3A, 36W) including a high-tech thermo-sensor (Dallas Semiconsuctor DS1820) controller. The heating loop is connected to the mixing chamber with a 50 μL Teflon tube. 100 μL shot volume is sufficient to wash out the hot buffer from the heating loop and get into the 25 μL cuvette. The cuvette house is heated by a heating element (resistor based heating element, maximum 24V, 4A, 96W) including a similar thermo-sensor which was tightly fixed to the bottom surface of the cuvette house. The temperature of cuvette is detected directly by a thermo-sensor attached to the wall of the cuvette. Asymmetric mixing of the reactants is applied in order to reach the appropriate temperature of the reaction: 1 and 5 ml syringes were used and both were incubated at 20°C, controlled by a water circulator. The 5 ml syringe contained the non-heat-sensitive reactant while the 1 ml syringe was filled with the heat-sensitive reactant (regularly, the enzyme). The non-heat-sensitive reactant flows to the cuvette through an inserted heating loop. The temperature of the heating loop is adjusted to a higher temperature than the reaction temperature. The cuvette chamber is heated to the reaction temperature. Temperature calibration of the novel setup is carried out using NATA to ensure that the thermo controllers are adjusted correctly and thus the temperature of the reaction mixture is identical to the temperature of the cuvette.

1.2 Calibration of the adjusted temperature pairs of the heating loop and the cuvette house

NATA was used to establish correct temperature pairs for both heating elements, utilizing the temperature dependence of the intensity of tryptophan fluorescence. Both syringes contained 50 μM NATA and different experimental temperatures were generated with the heat-jump/stopped flow apparatus. The constant fluorescence intensity of NATA indicated that correct temperature pair values were adjusted by the thermo-controllers. Temperature-pairs to be adjusted were determined for the experimental temperatures in the range of 20-70 °C.

1.3 Dead Time Determination

Dead time determination was based on the NBS-NATA reaction as described in [137, 138]. 280 nm excitation wavelength was applied with 5 nm slit width. On the emission side a 340 nm interference filter was used. In these experiments phosphate buffer (140 mM NaCl, 2.7 mM KCl, 10.1 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.3) was applied. Measurements were carried out at 23°C applying 18 ml/s flow rate and the traces were recorded with a photomultiplier set to 694 V. The applied concentrations were as follows: 42 μM NATA was mixed with 33, 83, 167, 417, 833 and 1667 μM NBS. Single exponentials were fitted to the transients. The delay between the intercept of the fitted exponentials and the first data point that joins the fitted exponentials provides an estimate of the instrumental dead time at the applied flow rate.

2. Mutagenesis and expression of human trypsinogen 4 variants

Wild type human trypsinogen 4 was cloned as described previously [139]. The R193G mutant clone was generated as reported by Tóth *et al* [134]. The R193F and R193Y mutant clones were generous gifts from Dr. László Szilágyi (Department of Biochemistry, Eötvös Loránd University). The amino acid substitutions at position 193 were generated by the megaprimer mutagenesis method. Trypsinogens were expressed, renatured and activated, purified and purity was assessed as described previously [134]. The enzymes were dialyzed against 2.5 mM HCl and stored at -20°C. The concentration of the prepared enzymes were determined by active site titration with 4-methylumbelliferyl 4-guanidinobenzoate or based on their absorbance at 280 nm using the theoretical extinction coefficient ε_{280} = 40570 $M^{-1}cm^{-1}$.

Mutagenesis and expression were carried out by Júlia Tóth.

3. Steady state kinetic measurements

3.1 Determination of k_{cat} and K_m

Measurements were carried out with 0.5-2 nM enzymes on Z-Gly-Pro-Arg-pNA substrate in 50 mM Tricine, 10 mM $CaCl_2$ pH 8.0 buffer at 20.0 °C. Substrate stock solutions were prepared in dimethylformamide, and the final concentration of DMF in the assays was less than 1%. Hydrolysis of the substrate was monitored by measuring the generation of the paranitroanilin product at 405 nm using a Shimadzu UV-2101PC spectrophotometer. Initial velocities were measured at 6 different substrate concentrations in the range of 5-250 μM. Three parallel measurements were carried out for each data points. The values of k_{cat}, K_m and k_{cat}/K_m were determined from the parameters of the hyperbolas fitted to the initial velocities plotted against substrate concentration.

Steady state measurements were carried out by Péter Medveczky.

4. Transient kinetic measurements

4.1 Stopped flow measurements

Transients were recorded on a SF-2004 instrument (KinTek Corp.) equipped with a 450-watt Hg-Xe super-quiet lamp (Hamamatsu Corp.). Tryptophans were excited at 297 nm with a bandwidth of 2 nm and fluorescent emission was detected with a photomultiplier set to 700 V voltage through a 340 nm interference filter (Comar Instruments). The dead time of the stopped flow apparatus is 1 ms. The applied flow rate was 12 ml/sec and 40 μl shot volumes were mixed at 1:1 ratio. 1-4 μM enzymes in 20 mM CABS, 10 mM $CaCl_2$ pH 11.0 were mixed with 100 mM Tricine, 10 mM $CaCl_2$ pH 8.0 and the fluorescence emission intensity increase was monitored. The rate of the conformational change was measured with this setup in the 5-38 °C temperature range in 3 °C increments. The pHs of the buffers were adjusted at room temperature to different pH values to give the final value pH 8.0 ± 0.1 pH (and that pH ≥ 11.0) at the different experimental temperatures.

4.2. Heat-jump/stopped-flow experiments

Measurements were carried out on the novel heat-jump/stopped-flow instrument developed in our laboratory [135]. The enzyme was kept at 20 °C until mixing with the pH 8.0 buffer which flows through a heated loop. 20 μl enzyme in 20 mM CABS, 10 mM $CaCl_2$ pH 11.0 was mixed with 100 μl 100 mM Tricine, 10 mM $CaCl_2$ pH 8.0 and the fluorescent emission change was detected. The pHs of the buffers were adjusted at room temperature to yield the final value at the different experimental temperatures as described above. The 1:5 mixing ratio enables a greater temperature-jump and at the same time allows keeping the enzyme at non-denaturing temperatures until the reaction.

4.3. Determination of dependence of the rate constants on the relative external viscosity

Stopped flow measurements were carried out at 20.0 °C using buffers 10 mM CABS, 5 mM $CaCl_2$, pH 11.0 and 50 mM Tricine, 5 mM $CaCl_2$, pH 8.0 supplemented with viscogen to yield different relative viscosities. Maltose was applied as a viscogen in the concentration range of 0-1.46 M resulting in relative viscosities of 1-8.18 [140]. Maltose increases the relative viscosity to the greatest extent while reducing the dielectric constant of the solvent most slightly as compared to other frequently used viscogens, e.g. fructose and ethylene glycol. The concentration of the buffers was reduced compared to the other measurements as ionic strength influences the viscosity. Other experimental settings were the same as described in the “Stopped flow measurements” section.

4.4 Kinetic and thermodynamic analysis

5-8 recorded transient traces were averaged and analyzed by fitting to single or double exponential functions using the KinTek software (KinTek Corp.) and OriginLab v7.5 (MicroCal Software). Thermodynamic profiles were analyzed by fitting exponential functions following $y = a\exp(b/x)$ to the plots of the observed rate constants (k) *versus* temperature (T). The parameters of the fitted function can be directly corresponded to the parameters of the Arrhenius equation without linearization (y = k, a = A, b = -E_a/R, x = T).

Temperature dependence of a reaction rate constant is described by the Arrhenius equation [141]:

$$k = A\exp(-E_a / RT) \qquad \text{(Eq. 1)}$$

where k is the rate constant, A is the preexponential term, E_a stands for the activation energy of the process, R is the gas constant and T is the absolute temperature. Kramers' theory is a conventional theoretical approach to describe the effect of friction on the rate constants of unimolecular reactions in the condensed phase [142]. In this model the chemical reaction is modelled by a particle with a diffusive one-dimensional motion from a potential well over a barrier. Based on Kramers' rate theory, the preexponential term of the Arrhenius equation contains a friction parameter which is determined dominantly by viscosity and the rate constant is inversely proportional to this friction.

Ansari and co-workers modified this approach for proteins by separating the friction into two sources of friction because only a part of the protein interacts with the solvent molecules [143]. One of these terms is the friction of the solvent (external friction) restraining the motion of the atoms on the surface of the protein and the other term is the internal friction of the protein hindering the motion of the protein atoms relative to each other. The following equation was stated to describe this model:

$$k = \frac{C}{\sigma + \eta}\exp(-E_a / RT) \qquad \text{(Eq. 2)}$$

where η is the external (solvent) viscosity and σ is a parameter with units of viscosity that determines the internal friction of the protein (henceforth, internal viscosity), C includes the viscosity independent parameters. The solvent friction, according to Stokes' law, is proportional to the solvents viscosity. Based on this analogy, the internal molecular friction can also be referred to as the internal viscosity of the protein, and this viscosity-like parameter has units of viscosity. However, in contrast with (solvent) viscosity, this internal viscosity belongs to a specific structural change.

Assuming that the activation energy does not depend on the viscosity at constant temperature, modification of equation 2 results in:

$$k = \frac{C'}{\sigma + \eta} \qquad \text{(Eq. 3)}$$

where C' includes $C\exp(-E_a / RT)$. The linearized form of equation 3 will then be the following:

$$\frac{1}{k} = \frac{\eta}{C'} + \frac{\sigma}{C'} \qquad \text{(Eq. 4)}$$

As a consequence, plotting 1/k against the external viscosity gives a linear function, and the local internal friction of the protein can be deduced from its intercept multiplied by the reciprocal of the slope. In other words, at constant temperature (Equation 4), the internal friction can be calculated by the extrapolation of the rate constant to zero external viscosity (η = 0 cP):

$$\sigma = \frac{C'}{k} - \eta \qquad \text{(Eq. 5)}$$

Thus by measuring the rate of the conformational change as a function of the relative viscosity of the reaction buffer, the local internal friction of a protein can be determined.

Results

1. Principle and construction of the heat-jump/stopped-flow

In a conventional stopped-flow apparatus, reactants are rapidly pushed from two syringes (A and B) through a small mixing chamber into the cuvette where they combine. Then the reaction mixture reaches the cuvette through a short tube. The progress of the chemical reaction can be detected in the cuvette by an optical signal. Detection is started after the quick stop of the pistons but the reaction starts immediately as the reactants are mixed. Aging of the reaction mixture on the way from the mixing chamber to the cuvette causes the so-called dead time of the apparatus. Typically, the dead time of a stopped flow instrument is on the ms time scale which is short enough to investigate most of the enzymatic reactions. The syringes and the cuvette house are adjusted to the same temperature by a water circulator. The applied temperature is limited by the temperature sensitivity of the reactants because keeping at high temperature for longer times would denature them already in the syringe before the reaction starts.

The schematic view of the heat-jump/stopped-flow instrument is presented in Figure 2. In this equipment, the enzyme and its substrate are stored in syringes A and B at native temperature by using water bath temperature control and the cuvette is kept at the experimental temperature by means of an inserted heating element and temperature controllers. A heating loop is inserted between the substrate syringe (syringe B) and the mixing chamber in which the substrate (the non-heat-sensitive reactant) can be heated to higher temperature than the reaction temperature. The temperature of the heating loop is controlled by another thermo-controller. The temperature of the enzyme syringe (syringe A) and the heating loop are adjusted so that the mixture of the enzyme solution and the high-temperature substrate solution gives the reaction temperature. The temperature of the cuvette house is adjusted to the temperature of the mixture of the cold (heat-sensitive) and the hot reactant to keep the cuvette (reaction chamber) at a constant temperature during the reaction. Mixing of the reactants yields the new, high temperature of the reaction mixture, consequently the dead time of the measurement and the heat-jump are the same and is determined by the mixing time of the stopped-flow, which is 1 ms (Figure 3b). In order to achieve high temperature-jumps asymmetric mixing is applied: one volume of heat-sensitive cold reactant (e.g. enzyme in syringe A) is mixed with five volumes of non-heat-sensitive hot reactant (e.g. substrate in syringe B).

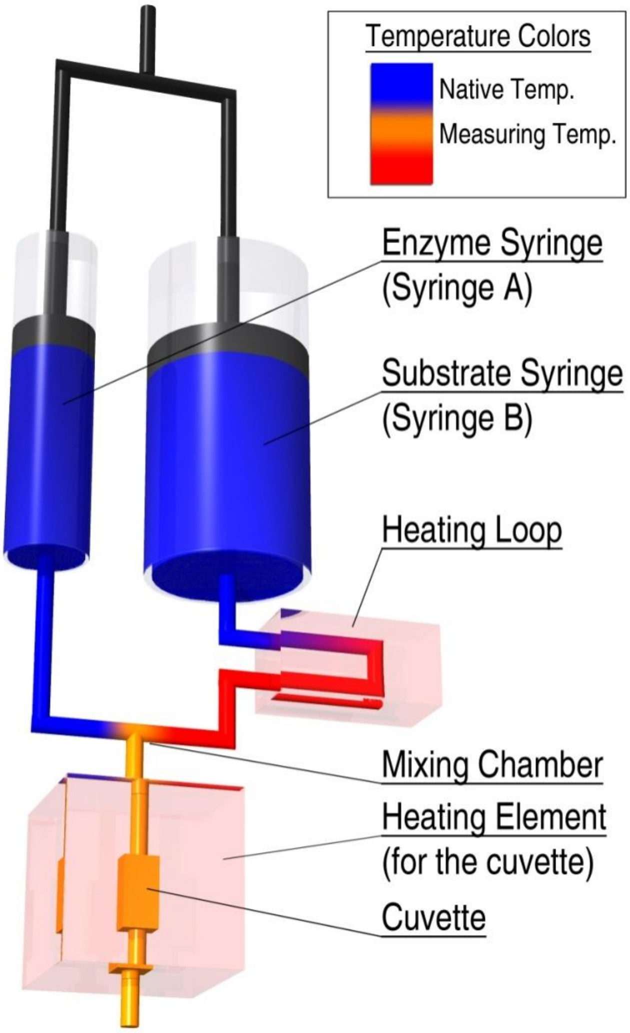


Figure 2. Schematic set-up of the heat-jump/stopped-flow apparatus. The heat-sensitive reactant (enzyme) is loaded into syringe A and the non-heat-sensitive reactant is loaded into syringe B both thermostated to 20 °C (indicated by blue). Upon the stopped flow push the non-heat-senitive reactant flows through a heating loop, in which it is heated above the reaction temperature (indicated by red). The enzyme and the hot reactant combine in the mixing chamber and they contribute to the new reaction temperature (indicated by orange) according to their temperatures and volumes. The process of the reaction at an elevated experimental temperature can be monitored in the cuvette using e.g. spectroscopic signals.

In a typical experiment after the first push of the drive syringes, the solution from the substrate syringe (syringe B) reaches the heating loop inserted before the mixing chamber, where the substrate solution can be warmed up even to 80-90 ºC. Due to the second push, the hot substrate solution from the heating loop and the colder enzyme solution (from syringe A) combine in the mixing chamber. The solutions of reactants contribute to the developing temperature according to their volumes thus even 40-50 ºC temperature-jump of the enzyme solution can be achieved immediately. While the reaction proceeds, the next portion of the substrate solution warms up in the heating loop, so the next shot is also appropriate for the forthcoming measurement.

Goldmann and Geeves suggested an arrangement called slow temperature-jump in which the syringes are kept at low temperature and the mixed cold reactants are shot into the hot cuvette [144]. Using this technique the temperature equilibration of the reaction mixture takes 150 ms. Verkman et al. constructed another equipment called stopped-flow temperature-jump which has 60 ms dead time [145]. The long dead times of these instruments severely limit their applicability. In comparison the dead time of the novel heat-jump/stopped-flow is much shorter because the reaction temperature is set upon mixing instead of warming up.

1.1 Calibration of heat jump stopped flow

To take advantage of the heat dependence of fluorescence, our system was calibrated to ensure that the temperature of the mixed solution and the cuvette are the same. Since the rate of reactions and the intensity of fluorescence are temperature sensitive, temperature equilibration of the reaction mixture in the cuvette during the courses of the reaction would cause artifacts. Fluorophores like NATA are useful to calibrate the temperatures of the reaction mixture and the cuvette house to the same value. During calibration, NATA was loaded into the cold syringe A and rapidly mixed with hot buffer pushed through the heating loop inserted between syringe B and the mixing chamber. Change in fluorescence intensity of NATA in the cuvette indicated that the temperature of the reaction mixture and the cuvette were not identical right after mixing and temperature re-equilibration occurred in the cuvette. Fluorescence intensity decrease indicates that the reaction mixture was colder than the cuvette and the solution warmed up in the cuvette and *vice versa* (Figure 3a). Constant fluorescence intensity indicated that the temperature of the cuvette and the reaction mixture was the same. The time constant of temperature equilibration in our system was determined to be 1.2 second. Since the temperature dependence of fluorescence intensity of NATA is sensitive enough to calibrate our system with 0.1°C accuracy, temperature pairs to be set by the temperature controllers can be easily determined for each experimental temperature.

The main advantage of this setup is that the fast heat-jump occurs simultaneously with the rapid mixing of the reactants so dead time of mixing and heat-jump are identical if the temperatures of reaction mixture and the cuvette are identical. We found that modification of the conventional stopped-flow apparatus did not increase the dead time which was determined with the reaction of NATA and NBS to be 0.9-1 ms (Figure 3b).

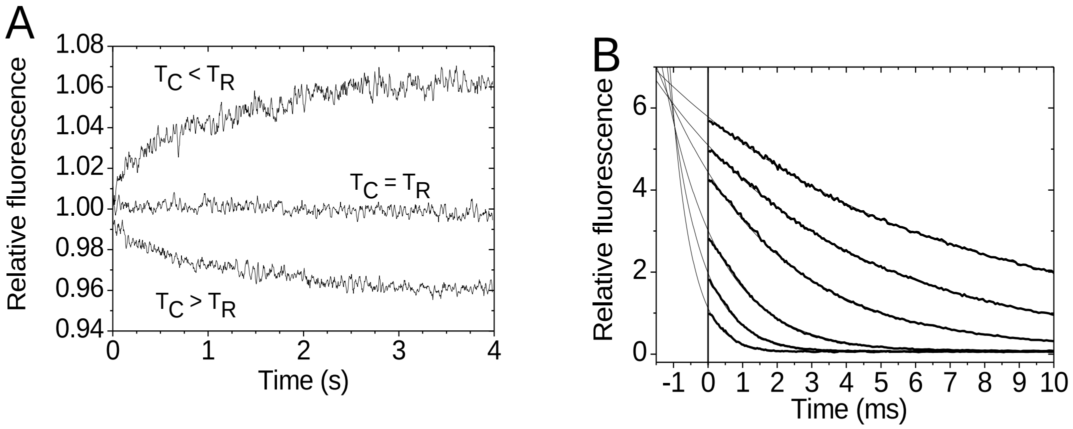


Figure 3. Temperature calibration of the heating elements (panel A) and determination of the dead time of the heat-jump/stopped-flow apparatus (panel B). **A:** Heat-jump/stopped-flow fluorescence records on mixing 50 μM NATA thermostated to 25 °C with hot buffer thermostated to 58 °C, 60 and 62 °C then pushed into a 52 °C cuvette house. Fluorescent intensity increase indicates that the temperature of the reaction mixture was higher than the temperature of the cuvette and *vice versa.* Constant fluorescence intensity suggests that the temperature of the reaction mixture was identical to the temperature of the cuvette. **B:** Heat-jump/stopped-flow records on the reaction of NATA with increasing concentrations of NBS at 25 °C to determine mixing dead time. Instrumental dead time was determined to be 0.9 ms.

2. Internal viscosity: the role of hinge residues in trypsin activation

2.1 Enzymatic activity of human trypsin 4 and its mutants at position 193

In order to test whether the enzymatic activity of the R193G/A/Y/F trypsin variants is affected by the mutations, their catalytic activity of amide bond hydrolysis was measured on Z-Gly-Pro-Arg-pNA substrate. The determined values for the Michaelis-Menten parameters are summarized in Table 1. Our data show that the mutations caused only slight changes in the k_{cat} and K_m values. The greatest change in the K_m is within 50% and for the k_{cat} value it does not exceed 30%. The k_{cat} value increases upon replacement of the glycine with a more bulky residue, and the degree of this increase correlates with the size of the sidechain. The same observation holds also for the K_m value. The greatest change in the k_{cat}/K_m value upon these amino acid substitutions is 30%, thus these site 193 variants can all be considered as active enzymes. These results indicate that both substrate binding and catalysis of hydrolysis of Z-Gly-Pro-Arg-pNA are only moderately affected by these mutations.

	k_{cat} *(s^{-1})*	K_m *(μM)*	k_{cat}/K_m *($s^{-1}\mu M^{-1}$)*
R193G	110 ± 10	13.5 ± 2.5	8.17 ± 1.68
R193A	124 ± 4	19.8 ± 3.2	6.23 ±1.02
WT	158 ± 8	25.2 ± 0.6	6.27 ±0.36
R193F	162 ± 5	27.4 ± 6.7	5.90 ±1.45
R193Y	133 ± 6	21.3 ± 6.1	6.21 ± 1.80

Table 1. Hydrolysis of Z-Gly-Pro-Arg-pNA amide substrate by wild type human trypsin 4 and its 193-variants. Assays were performed in 50 mM Tricine, 10 mM $CaCl_2$ pH 8.0 at 20.0 °C as described under Materials and methods. The values of the Michaelis-Menten parameters represent the mean ± SEM of 3 measurements. Errors of k_{cat}/K_ms were calculated taking the propagation of error into account.

2.2 Kinetic analysis of the conformational change during pH-jump activation

The rate of the conformational rearrangement in the course of activation was measured by following the intrinsic tryptophan fluorescent intensity change of the proteins in pH-jump stopped-flow experiments. One of the syringes of the stopped flow contained the enzyme in a buffer with a relatively low buffer capacity at pH 11 and the other syringe contained a buffer with high buffer capacity at pH 8.0. The pH of the solution after mixing was 8.0 ± 0.1. Our data show that the rates of the conformational change are affected by the mutations at position 193. The rate constants at 20.0 °C for the site 193 variants are as follows: $k_{R193G} = 1.66\ s^{-1}$, $k_{R193A} = 0.20\ s^{-1}$, $k_{WT} = 0.077\ s^{-1}$, $k_{R193Y} = 0.13\ s^{-1}$, $k_{R193F} = 0.090\ s^{-1}$. We note that burst phases were detected for wild type human trypsin 4, the R193Y and the R193F mutants with amplitudes between 11-16% of the total fluorescence change and rate constants 7-17 times larger than that of the analyzed dominant phases.

2.3 Thermodynamic analysis of the conformational change during pH-jump activation

2.3.1 Temperature dependence of the conformational transition

Temperature dependence of the rate constants was also measured in order to study the thermodynamics of the conformational change during activation. Thermodynamic parameters for the conformational change were derived from the non-linearized Arrhenius plots (k plotted against T) and are summarized in Table 2. The Arrhenius plots (ln k plotted against 1/T) were linear in the temperature range of 5-45 °C (Figure 4). No significant differences were found in the reliability of fitting when the Kramers' equation corrected with the temperature dependence of the viscosity of water was fitted to the data. Our most important finding is that the Arrhenius plots are parallel for the site 193 variants of human trypsin 4. Consequently, the activation energy (E_a) of the conformational transition is not affected by the amino acid substitutions. Thus exponents 'b' of the fitted exponential functions, i.e., the slope of the linearized Arrhenius plots are practically the same (Figure 4, Table 2). The largest difference in the activation energies was found between wild type and the R193F mutant human trypsin 4 amounting to 4%. The activation energies of the rest of the mutants differed only by 0.8-2%. In contrast, the intercepts of the linearized Arrhenius plots for these enzyme variants differ significantly, suggesting that these mutations selectively alter the preexponential term of the Arrhenius equation. The greatest difference in the preexponential terms was observed in case of the R193G and R193F mutants (26-fold; Table 2).

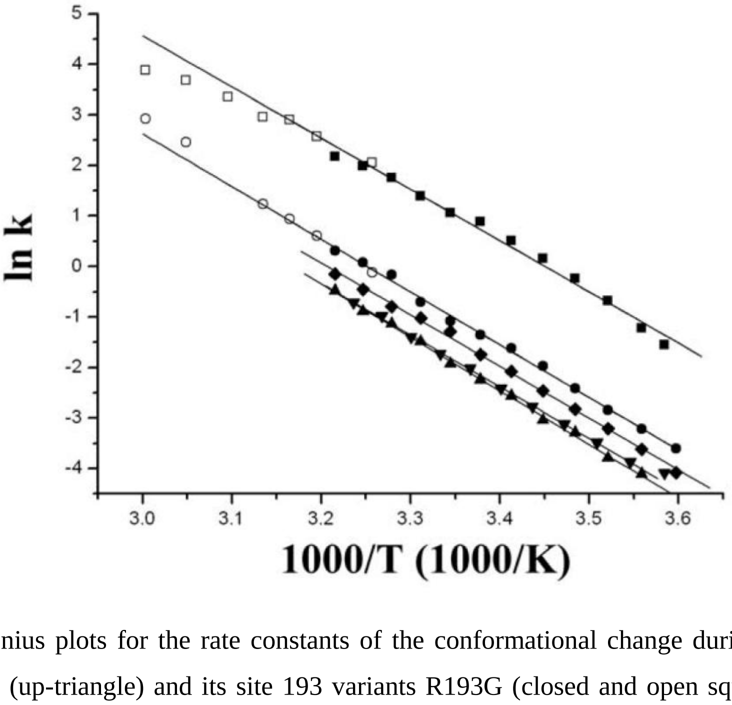


Figure 4. Arrhenius plots for the rate constants of the conformational change during activation for wild type human trypsin 4 (up-triangle) and its site 193 variants R193G (closed and open squares), R193A (closed and open circles), R193F (down-triangle) and R193Y (diamond), determined by stopped-flow and heat-jump/stopped-flow experiments. The rate of the conformational change was measured with the conventional stopped-flow setup in the 5-38 °C temperature range in 3 °C increments (closed marks). In the 34-60 °C temperature range the novel heat-jump/stopped-flow equipment was applied to extend the Arrhenius plots of R193G and R193A variants (open marks). Thermodynamic parameters were determined from the parameters of the fitted exponential functions following $y = a\exp(b/x)$ as described in Materials and methods.

	R193G	R193A	WT	R193F	R193Y
b*	-(10.4 ± 0.4) ×10^3	-(10.7 ± 0.2) ×10^3	-(10.8 ± 0.3) ×10^3	-(10.3 ± 0.2) ×10^3	-(10.5 ± 0.3) ×10^3
E_a *(kJmol^{-1})*	86.5 ± 3.4	88.7 ± 1.6	89.5 ± 2.4	85.7 ± 1.9	87.2 ± 2.7
a* = A	3.95 × 10^{15}	1.12 × 10^{15}	6.50 × 10^{14}	1.49 × 10^{14}	4.40 × 10^{14}

Table 2. Thermodynamic parameters derived from the non-linearized Arrhenius plots of wild type human trypsin 4 and its 193-variants. The rate of the conformational change during pH-jump activation was measured with a stopped flow apparatus monitoring the intrinsic fluorescent emission change of the proteins. Experimental conditions were as described under Materials and methods. Reaction rates were measured in the temperature range of 6-38 °C (or 6-60 °C in the case of R193G and R193A mutants) in 3 °C increments. Rate constants were plotted against temperature, and the exponential function following $* y = a \exp(b/x)$ was fitted to the data. Thermodynamic parameters were derived from the parameters ‘a’ and ‘b’ of the fitted functions as described under Materials and methods. Values represent the mean and ± SEM of the parameters for the fitted linear functions.

By developing a novel heat-jump/stopped-flow setup, we were able to extend the Arrhenius plots up to 60 °C. Deviations from the linear function could be observed above 45 °C which indicate another reaction step. It has to be noted, however, that denaturation happens on the ten seconds time scale at 45-60 °C (which could be detected by the decline of the fluorescence intensity) while the activation process is more than an order of magnitude faster at these temperatures.

2.3.2 External viscosity dependence of the conformational transition

To investigate the effects of these mutations on the preexponential term in more detail, we also performed experiments in which the relative viscosity of the buffers was varied. The question was if the rate constant of a conformational change decreases as the solvent viscosity is increased as predicted by the Kramers’ theory. To answer this question, pH-jump stopped-flow measurements were carried out on the wild type trypsin 4 and its R193G and R193A mutants in buffers of different relative viscosity from 1 to 8.18 at 20°C. The increased solvent viscosity caused dramatic effect on the rate constants of the pH-jump induced conformational change of both mutants (Figure 5); e.g., at relative external viscosity of 2 the rate constants decreased by 45% and 36% for the R193G and R193A mutants, respectively.

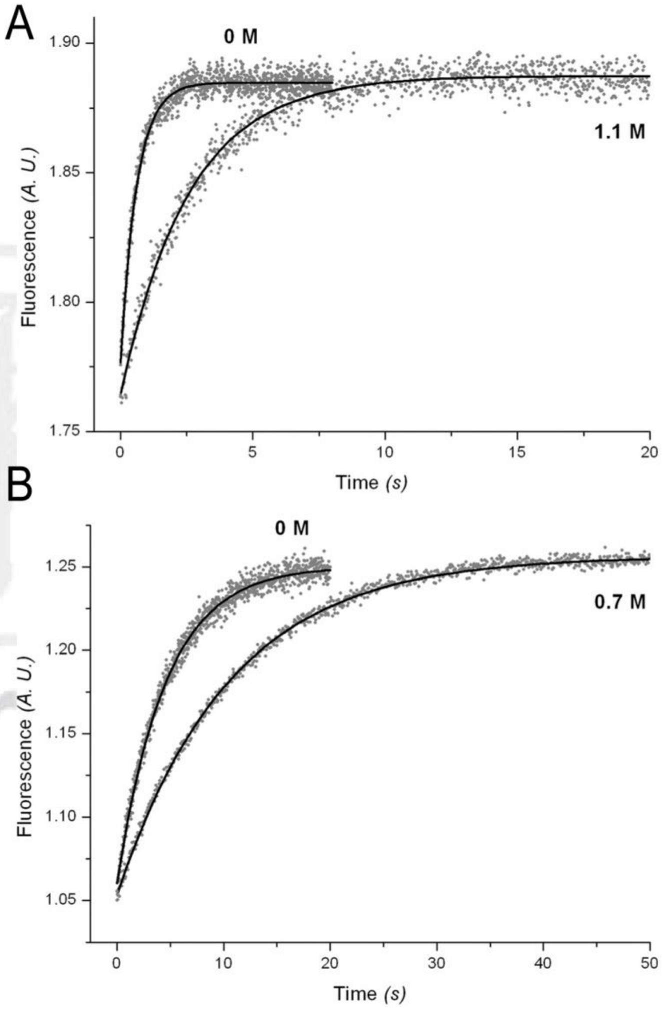


Figure 5. Stopped flow records on exposing human trypsin 4 R193G (**a**) and R193A mutant (**b**) to a pH-jump from pH 11.0 to pH 8.0 in buffers of different relative viscosity. Single exponentials were fit to the recorded traces. The rate constant of the conformational transition decreased as the solvent viscosity was increased by the viscogen. The observed rate constants for the presented traces are the following: $k_{R193G\ relative\ viscosity\ 1} = 1.65\ s^{-1}$, $k_{R193G\ relative\ viscosity\ 4.4} = 0.39\ s^{-1}$, $k_{R193A\ relative\ viscosity\ 1} = 0.22\ s^{-1}$ and $k_{R193A\ relative\ viscosity\ 2.2} = 0.095\ s^{-1}$.

2.3.3 Relative internal viscosity values of R193A/G mutants and wild type trypsin 4

We found hyperbolic relations between the observed rate constants and the external viscosity for the wild type enzyme and both mutants which confirm the validity of Kramers’ theory applied to this conformational change between two structurally definite conformers. Reciprocal values of the rate constants were plotted against the relative external solvent viscosity and linear functions were fitted to the data (Figure 6).

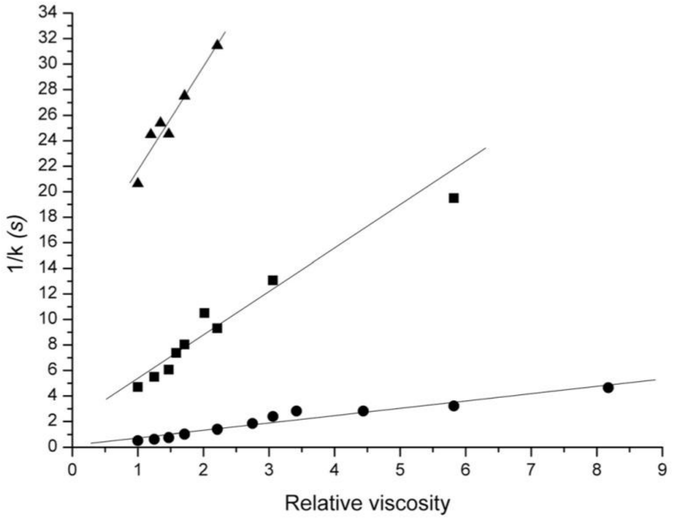


Figure 6. Dependence of the rate constants of the conformational change during activation on the external viscosity of the wild type human trypsin 4 (up-triangle) and its R193A (square and R193G (circle) mutants. Stopped flow measurements were carried out at 20.0 °C using buffers 10 mM CABS, 5 mM $CaCl_2$, pH 11.0 and 50 mM Tricine, 5 mM $CaCl_2$, pH 8.0, supplemented with 0-1.46 M maltose as viscogen to yield different relative viscosities from 1 to 8.18. 1/k was plotted against the relative external viscosity and linear functions were fitted to the measured data. Relative internal molecular frictions were calculated from the intercept and the slope of the linear fits according to Equation 4.

The parameters of the fitted linear functions and internal viscosities, calculated according to Eq. 5, are presented in Table 3. Our data show that the relative internal viscosity of the R193A mutant is 3-fold greater compared to the R193G mutant, the relative internal viscosity of the arginine possessing counterpart is 6-fold greater relative to the glycine mutant and 2-fold greater than that of the alanine mutant ($\sigma_{\text{R193G}} = 0.27$, $\sigma_{\text{R193A}} = 0.81$, $\sigma_{\text{WT}} = 1.67$). These data clearly suggest that a bulkier amino acid at the hinge region locally increases the molecular friction in the protein resulting in an increase of the steric hindrance for a specific conformational change.

	R193G	**R193A**	**WT**
Slope	0.575 ± 0.05	3.07 ± 0.25	8.13 ± 1.09
Intercept	0.156 ± 0.187	2.48 ± 0.67	13.56 ± 1.67
Relative internal viscosity	0.27	0.81	1.67

Table 3- Relative internal viscosity of R193G and R193A mutant and wild-type human trypsin 4 derived from the viscosity dependence of the reaction rate of the conformational change during activation. The internal viscosity of the proteins was calculated according to Equation 4. Values for the intercepts and slopes represent the mean and standard deviation of the fitted linear functions.

Discussion

The goal of this work was to determine if the rate and thermodynamics of the conformational change during pH-jump induced activation of trypsinogen is affected by the character of the residue at one of the hinge positions. We therefore investigated the temperature and viscosity dependence of the conformational change during activation in human trypsin 4 and its R193G/A/F/Y mutants in pH-jump stopped flow experiments. These amino acids at position 193 allowed different degree of freedom for the hinge around which the conformational rearrangement occurs.

1. Advantages of tryptophan signal detection

Human trypsin 4 possesses five tryptophan residues: Trp221 is located inside of segment 216-223, while Trp141 and Trp215 are located in the close vicinity of peptide segments 142-152 and 216-223 that are involved in the conformational transition of activation. These native tryptophans are appropriate probes to monitor the structural rearrangement induced by the pH-jump activation, as their fluorescence intensity increases significantly during this process. The tryptophan-based detection method has several advances compared to ligand-binding assays: this reaction follows first order kinetics, intrinsic fluorescence detection gives a direct read-out, and it even has greater sensitivity thus requires less protein.

2. Evaluation of the observed rate constants

Trypsin undergoes a reversible conformational change during pH-jump from an inactive zymogen-like structure at pH 11.0 to the active conformation at pH 8.0. Therefore, the observed rate constant measured by monitoring the intrinsic fluorescence change of the protein is the sum of the forward and reverse rate constant for the reaction. However, the equilibria at pH 8.0 are highly shifted towards the forward direction as it was shown for rat trypsin [146] thus the contribution of the reverse rate constant to the observed rate constant is negligible. Therefore, the transition is practically irreversible and the observed rate constant can be considered as the forward rate constant.

3. Structural consequences of the replacement of R193G with a bulkier residue

Activation domain peptides are bordered by glycine residues acting as hinges during activation. The values for the peptide backbone dihedral angles at Gly193 are Φ = -148.7° and Ψ = 18.1° in bovine trypsinogen (PDB ID: 1tgb), while in bovine trypsin (PDB ID: 2ptn) these values change to Φ = 105.4° and Ψ = -19.1°. The atomic structure of human trypsin 4 complexed with benzamidine has been resolved and it was shown that in spite of the G193R mutation, the overall fold of the molecule is highly similar to that of human trypsin 1 having a glycine at this position [139]. Earlier studies suggest that substitution of Gly193 with an amino acid having a bulkier side chain does not perturb significantly the overall fold of the molecule [147, 148]. Although there is only a slight structural change in the peptide backbone conformation, we supposed that some other dynamics-related physical parameter to be characterized might affect the rate of conformational change during activation in trypsin site 193 variants.

In the presented study we mutated R193 of human trypsin 4 to glycine, alanine, phenylalanine and tyrosine. These mutations did not cause large perturbation in the steady state enzyme kinetic values on Z-Gly-Pro-Arg-pNA substrate (Table 1). Besides from being one of the hinge residues during the activation, residue 193 has an important role in substrate binding and enzyme catalysis. The amido group of residue at position 193 is part of the oxyanion hole which stabilizes the developing tetrahedral intermediates during catalysis, thus the substitution of Gly193 slightly influences the k_{cat} value. These results suggest that the substrate binding pocket and the geometry of the oxyanion hole are not perturbed significantly by these mutations.

4. The effects of a bulkier residue at position 193 on the kinetic and thermodynamic parameters of the conformational transition

The rate of the conformational transition during pH-jump induced activation of wild type and R193G/A/F/Y mutant human trypsin 4 was measured by monitoring the intrinsic fluorescent intensity change of the proteins in stopped-flow experiments. The rates of the conformational change are influenced by the mutations at position 193, and the values correlate with the size of the sidechain. The temperature dependences of the rate of pH induced activation were also determined. Strikingly, we found that the Arrhenius plots of the reactions were parallel for all of the mutants. This phenomenon was investigated in a wide temperature range (between 5 °C and 60 °C) applying a new stopped flow equipment called heat-jump/stopped-flow developed in our lab. The wide temperature range allowed us to improve the reliability that the Arrhenius plots are parallels. These results suggest that activation energies are identical and the thermodynamic parameters for these trypsin mutants differ only in the preexponential term. Further important observation is that the larger the size of the substituted amino acid side chains at position 193 the smaller the value of the preexponential term. This aspect of the results indicates that restricting the conformational freedom of the hinge affects the preexponential term and not the activation energy.

Eyring-Polanyi transition state theory is a simplified model and it is adequate only for the description of temperature dependence of reactions in gas phases. As biomolecular reactions take place in the condensed phase and have complex multidimensional potential energy surfaces, and in a generalised transition state theory a transmission coefficient is required in the preexponential term [149], Kramers' theory is an appropriate and relatively simple approach for the description of reactions of complex molecules in condensed phase [150]. As stated in the work of Frauenfelder and co-workers, if the reaction rate depends strongly on solvent viscosity, data can be assessed using Kramers' theory. In the work of Beece *et al* the ligand binding of protoheme and myoglobin was studied in solvents in which the viscosity was varied over a wide range postulating that the solvent affects the protein reaction predominantly through the solvent viscosity [151]. This experimental approach is in agreement with the idea that protein barriers have dynamic origins. The transition state theory is valid only in a limited region of solvent viscosity, below 1 mP, thus the application of this theory to reactions even in aqueous solutions is uncertain, and rather Kramers' theory is appropriate. The phenomenon that the rate constant of a reaction is inversely proportional to the solvent viscosity is consistent with Kramers' theory.

To analyze the effect of the preexponential term on the rate of a structural rearrangement in detail, we examined whether the rate of the conformational transition is affected by the viscosity of the solvent. Kramers' equation predicts a hyperbolic dependence of the rate constants on solvent viscosity (Eq. 4). We measured the rate of the conformational change in buffers of different relative viscosity using maltose as a viscogen. Our data clearly show that the rate constant depends on the viscosity of the solvent (Figure 5), even at relatively low viscogen concentrations where the perturbation of charged-charged interactions were insignificant by the slightly decreased dielectric constant of the solutions. Furthermore, we found that the relation is hyperbolic, thus 1/k plotted against the relative external viscosity yields a linear function (Figure 6), which strongly indicates that Kramers' theory is an appropriate framework for the description of rates of enzyme conformational transitions under native conditions.

5. Internal viscosity of the trypsin hinge mutants

On the basis of the work of Ansari *et al* (Eq. 2) [143], the internal friction of a protein can be calculated by determining the dependence of the rate constant on the external viscosity and extrapolating to zero external viscosity (Eq. 5). We determined the viscosity dependence of the rate constant using the R193G and R193A mutant and linear functions were fitted to the plot of 1/k *versus* relative external viscosity. We found that both the slope and intercept of these linear functions are different in these mutants (Figure 6). The calculated internal viscosity of the alanine mutant is increased threefold as compared to the glycine mutant (Table 3). We conclude that the bulkier sidechain of alanine allows less conformational freedom for the peptide backbone in the conformational transition as compared to glycine which can be revealed as an increase in a viscosity-like parameter defined as internal viscosity. Also interesting, that this viscosity-like parameter is slightly lower than water viscosity which indicates relatively low restriction by the hinge region. Ansari *et al* determined σ=4.1 cP in myoglobin related to conformational change after ligand dissociation. This parameter is larger than we found in trypsin during its conformational change of activation. In myoglobin, relatively ordered water in the cavity of the protein may play an important role in the studied reaction which may increase the value of internal viscosity. Nevertheless, it has to be emphasized that internal viscosity is not a general parameter of the protein but it is associated with a specific conformational rearrangement.

In summary, our results illustrate that a specific conformational rearrangement of an enzyme is a Kramers' type reaction under native conditions. Furthermore, our data show that internal friction, therefore protein flexibility, can be modified specifically by mutations, in this way modulating the mobility of the hinge around which the structural change occurs.

Conclusions

Two levels of complexity were addressed in my thesis. In the first part, I presented an approach to predict the effect profiles of drug molecules. Then, I assessed protein flexibility by determining the internal viscosity of an interdomain conformational rearrangement. The methodology applied to process the two problems was different: from one hand, a number of statistical evaluation methods were used, combined with a small amount of experimental tests carried out to confirm the predicted bioactivity properties for certain drugs. On the other hand, an extensive experimental methodology was applied to study the effect of temperature and solvent viscosity on the rate of a specific conformational change, including the development of a novel combined heat-jump/stopped-flow equipment. Nevertheless, we were facing the same theoretical problem in both cases, i.e., the handling of complexity and the challenge of the extraction of the relevant features of a system in order to synchronize the found connections with a solid scientific model. In the case of protein flexibility, a model system was needed in which the problem of flexibility could be studied and modified specifically. In the case of bioactivity prediction, the issue of complexity was overcome by statistical analyses that enabled to reduce the dimensionality of the data without information loss that would lead to impaired prediction power.

Polypharmacology is a newly emerging approach which reflects the high complexity of the mechanism of actions of drugs. This aspect of pharmacology has not been fully exploited in drug development. Consequently, the entire effect profiles of drugs and drug candidates have remained unrevealed. We hypothesized that complex molecular feature sets of drugs correlate with the known part of effect profiles and may therefore provide predictive power to reveal the entire effect profiles of drugs.

We collected the structural data and registered effect profiles of all small-molecule drugs. Interactions to a series of non-target protein sites of each drug were calculated and an interaction pattern matrix was constructed. One-dimensional and multidimensional analyses unveiled a strong correlation between the effect profiles and IPs and this relationship was confirmed by independent validation. These findings allowed us to develop a robust and systematic effect prediction method, named Drug Profile Matching. *In vitro* analyses of tested effect categories further supported the accuracy and the robustness of the prediction.

To our knowledge this is the first method which directly relates distant levels of information, i.e., the information from the atomic interactions with the information from physiological effects. Moreover, unlike other similarity-based approaches [14, 84], no direct topological similarity information on drug molecules is involved; therefore, our approach is able to detect effect profile similarities even in the case of limited structural similarity between compounds. IPs, binding affinities of drugs to the same series of surfaces, mostly represent non-target interactions that cannot be measured experimentally because of the possible weak bindings. Nevertheless, these types of interactions might play an important role in the mechanism of actions in the organisms and could be considered as a key factor in polypharmacology.

The Drug Profile Matching method can be improved via different ways. After producing outstanding performance in effect re-prediction and prediction, the most interesting development would be the integration of side effect information into the interaction profile based bioactivity classification. The preliminary results presented before point to the possible application of Drug Profile Matching in this field. A number of publicly available sources for side effect information exist, e.g. SIDER at sideeffects.embl.de, maintained by the European Molecular Biology Laboratory [83]. Beside effect and adverse event data, target protein information could also been involved in the predictions. By the combination of these data sources, i.e. the effect, side effect and target data, a combined bioactivity profile predicting system can be set up. Our group has recently begun construction on this system, called Multicorrelated Drug Profile (MCDP), in parallel with a direction suggested by Fliri *et al* [152]. Prediction of drug metabolism, i.e. biotransformation by CYP isoforms, is also a topic of high interest. The Drug Profile Matching therefore offers an opportunity for systematic and rapid screening of approved drugs in order to discover new therapeutic indications and safety risks.

Moving away from approved drugs, the prediction system can be extended to druglike molecules as well. In this case, a new set of classification functions must be developed since the presented ones are trained and tested on the set of the existing drugs. Up to 2010, one hundred thousand druglike molecules were docked to the whole protein set and their respective IPs were generated. Processing of this huge set of data and getting more generally applicable effect classification rules are the next quests of our research group. After producing the generalized classification functions, Drug Profile Matching can be a valuable aid in the prediction of the pharmacological effect profiles of drug candidate molecules with high probability, thereby offering a novel approach for lead molecule design and optimization as well. As shown above, the good predictive power of the method holds out the promise for its use with marketed drugs or as a preclinical screen, bringing substantial improvement in the efficacy of future drug development and expediting the development process from drug discovery to marketing.

The examination of MAFs and binding site shape descriptors revealed that, except for few specific cases, the shapes of the binding pockets have relatively low weights in the determination of the affinity profiles of proteins. Since the MAF profile is closely related to the target specificity of ligand binding sites we can conclude that the shape of the binding site is not a pivotal factor in selecting drug targets. Nonetheless, based on strong specific associations between certain MAF profiles and specific geometric descriptors we identified, the shapes of the binding sites do have a crucial role in virtual drug design for certain drug categories, including morphine derivatives, benzodiazepines, barbiturates and antihistamines. Therefore we conclude that the application of shape-based drug design methodologies might prove better performance on this drug set than that for others.

In Part II, the role of internal friction, thus, the role of protein flexibility was studied in the model system of human trypsinogen 4 activation. Upon activation, distinct regions of the protein, bordered by hinge glycine residues, undergo conformational change. Since we presumed that rigidification of the hinge regions affect the rate of activation, we introduced side chains with different characters at a hinge position and studied their effects on the rate constant of conformational change. To analyze the thermodynamics of the reaction, temperature dependence of the reaction rate constants were examined in a wide temperature range using a novel heat-jump/stopped-flow apparatus developed in our laboratory. We found that an increase in the size of the side chain is associated with the decrease of the reaction rate constant. Our data show that the mutations do not affect the activation energy (the exponential term) of the reaction, but they significantly alter the preexponential term of the Arrhenius expression. The effect of solvent viscosity on the rate constants of the conformational change during activation of the 193G and 193A mutants were determined and evaluated by Kramers' theory. Based on this, we determined the internal viscosity parameter of the activation of the wild type trypsin and its R193A and R193G mutants experimentally. Therefore, we propose that the reaction rate of the studied conformational transition is regulated by the internal molecular friction which can be specifically modulated by mutagenesis in the hinge region.

Appendices

Appendix 1

List of the names and DrugBank codes of the applied drugs.

DB code	Drug name	DB code	Drug name
DB00114	Pyridoxal phosphate	DB00780	Phenelzine
DB00116	Tetrahydrofolic acid	DB00782	Propantheline
DB00117	L-histidine	DB00783	Estradiol
DB00118	S-adenosylmethionine	DB00784	Mefenamic acid
DB00119	Pyruvic acid	DB00786	Marimastat
DB00120	L-phenylalanine	DB00787	Aciclovir
DB00121	Biotin	DB00788	Naproxen
DB00122	Choline	DB00790	Perindopril
DB00123	L-lysine	DB00791	Uracil mustard
DB00125	L-arginine	DB00792	Tripelennamine
DB00126	Vitamin C	DB00793	Haloprogin
DB00127	Spermine	DB00794	Primidone
DB00128	L-aspartic acid	DB00795	Sulfasalazine
DB00129	L-ornithine	DB00796	Candesartan
DB00130	L-glutamine	DB00797	Tolazoline
DB00131	Adenosine monophosphate	DB00798	Gentamicin
DB00132	Alpha-linolenic acid	DB00799	Tazarotene
DB00133	L-serine	DB00800	Fenoldopam
DB00134	L-methionine	DB00801	Halazepam
DB00135	L-tyrosine	DB00802	Alfentanil
DB00136	Calcitriol	DB00804	Dicyclomine
DB00137	Xanthophyll	DB00805	Minaprine
DB00138	L-cystine	DB00806	Pentoxifylline
DB00139	Succinic acid	DB00807	Proparacaine
DB00140	Riboflavin	DB00808	Indapamide
DB00141	N-acetyl-D-glucosamine	DB00809	Tropicamide
DB00142	L-glutamic acid	DB00810	Biperiden
DB00143	Glutathione	DB00811	Ribavirin
DB00144	Phosphatidylserine	DB00812	Phenylbutazone
DB00145	Glycine	DB00813	Fentanyl
DB00146	Calcidiol	DB00814	Meloxicam
DB00147	Pyridoxal	DB00815	Sodium lauryl sulfate
DB00148	Creatine	DB00816	Orciprenaline
DB00149	L-leucine	DB00817	Rosoxacin
DB00150	L-tryptophan	DB00818	Propofol
DB00151	L-cysteine	DB00819	Acetazolamide
DB00152	Thiamine	DB00820	Tadalafil
DB00153	Ergocalciferol	DB00821	Carprofen
DB00154	Gamma-homolinolenic acid	DB00822	Disulfiram
DB00155	L-citrulline	DB00823	Ethynodiol diacetate
DB00156	L-threonine	DB00824	Enprofylline

DB00158	Folic Acid
DB00159	Icosapent
DB00160	L-alanine
DB00161	L-valine
DB00162	Vitamin A
DB00163	Vitamin E
DB00165	Pyridoxine
DB00166	Lipoic acid
DB00167	L-isoleucine
DB00168	Aspartame
DB00169	Cholecalciferol
DB00170	Menadione
DB00171	Adenosine triphosphate
DB00172	L-proline
DB00173	Adenine
DB00174	L-asparagine
DB00175	Pravastatin
DB00176	Fluvoxamine
DB00177	Valsartan
DB00178	Ramipril
DB00179	Masoprocol
DB00180	Flunisolide
DB00181	Baclofen
DB00183	Pentagastrin
DB00184	Nicotine
DB00185	Cevimeline
DB00186	Lorazepam
DB00187	Esmolol
DB00188	Bortezomib
DB00189	Ethchlorvynol
DB00190	Carbidopa
DB00191	Phentermine
DB00192	Indecainide
DB00193	Tramadol
DB00194	Vidarabine
DB00195	Betaxolol
DB00196	Fluconazole
DB00198	Oseltamivir
DB00199	Erythromycin
DB00201	Caffeine
DB00202	Succinylcholine
DB00203	Sildenafil
DB00204	Dofetilide
DB00205	Pyrimethamine
DB00206	Reserpine
DB00207	Azithromycin
DB00208	Ticlopidine
DB00209	Trospium
DB00210	Adapalene
DB00211	Midodrine
DB00212	Remikiren
DB00213	Pantoprazole
DB00825	Menthol
DB00826	Natamycin
DB00827	Cinoxacin
DB00828	Fosfomycin
DB00829	Diazepam
DB00830	Phenmetrazine
DB00831	Trifluoperazine
DB00832	Phensuximide
DB00833	Cefaclor
DB00834	Mifepristone
DB00835	Brompheniramine
DB00836	Loperamide
DB00837	Progabide
DB00838	Clocortolone
DB00839	Tolazamide
DB00841	Dobutamine
DB00842	Oxazepam
DB00843	Donepezil
DB00844	Nalbuphine
DB00845	Clofazimine
DB00846	Flurandrenolide
DB00847	Cysteamine
DB00848	Levamisole
DB00849	Methylphenobarbital
DB00850	Perphenazine
DB00851	Dacarbazine
DB00852	Pseudoephedrine
DB00853	Temozolomide
DB00854	Levorphanol
DB00855	Aminolevulinic acid
DB00856	Chlorphenesin
DB00857	Terbinafine
DB00858	Dromostanolone
DB00859	Penicillamine
DB00860	Prednisolone
DB00861	Diflunisal
DB00862	Vardenafil
DB00863	Ranitidine
DB00864	Tacrolimus
DB00865	Benzphetamine
DB00866	Alprenolol
DB00867	Ritodrine
DB00869	Dorzolamide
DB00870	Suprofen
DB00871	Terbutaline
DB00872	Conivaptan
DB00873	Loteprednol etabonate
DB00874	Guaifenesin
DB00875	Flupenthixol
DB00876	Eprosartan
DB00878	Chlorhexidine
DB00879	Emtricitabine

DB00214	Torasemide
DB00215	Citalopram
DB00216	Eletriptan
DB00217	Bethanidine
DB00218	Moxifloxacin
DB00219	Oxyphenonium
DB00220	Nelfinavir
DB00221	Isoetharine
DB00222	Glimepiride
DB00223	Diflorasone
DB00224	Indinavir
DB00225	Gadodiamide
DB00226	Guanadrel sulfate
DB00227	Lovastatin
DB00228	Enflurane
DB00229	Cefotiam
DB00230	Pregabalin
DB00231	Temazepam
DB00232	Methyclothiazide
DB00233	Aminosalicylic acid
DB00234	Reboxetine
DB00235	Milrinone
DB00236	Pipobroman
DB00237	Butabarbital
DB00238	Nevirapine
DB00239	Oxiconazole
DB00240	Alclometasone
DB00241	Butalbital
DB00242	Cladribine
DB00243	Ranolazine
DB00244	Mesalazine
DB00245	Benztropine
DB00246	Ziprasidone
DB00247	Methysergide
DB00248	Cabergoline
DB00249	Idoxuridine
DB00250	Dapsone
DB00251	Terconazole
DB00252	Phenytoin
DB00253	Medrysone
DB00254	Doxycycline
DB00255	Diethylstilbestrol
DB00256	Lymecycline
DB00257	Clotrimazole
DB00258	Calcium acetate
DB00259	Sulfanilamide
DB00260	Cycloserine
DB00261	Anagrelide
DB00262	Carmustine
DB00263	Sulfisoxazole
DB00264	Metoprolol
DB00265	Crotamiton
DB00880	Chlorothiazide
DB00881	Quinapril
DB00882	Clomifene
DB00883	Isosorbide dinitrate
DB00884	Risedronate
DB00885	Pemirolast
DB00887	Bumetanide
DB00888	Mechlorethamine
DB00889	Granisetron
DB00890	Dienestrol
DB00891	Sulfapyridine
DB00892	Oxybuprocaine
DB00894	Testolactone
DB00895	Benzylpenicilloyl polylysine
DB00896	Rimexolone
DB00897	Triazolam
DB00898	Ethanol
DB00899	Remifentanil
DB00900	Didanosine
DB00902	Methdilazine
DB00903	Ethacrynic acid
DB00904	Ondansetron
DB00905	Bimatoprost
DB00906	Tiagabine
DB00907	Cocaine
DB00908	Quinidine
DB00909	Zonisamide
DB00910	Paricalcitol
DB00911	Tinidazole
DB00912	Repaglinide
DB00913	Anileridine
DB00914	Phenformin
DB00915	Amantadine
DB00916	Metronidazole
DB00917	Dinoprostone
DB00918	Almotriptan
DB00919	Spectinomycin
DB00920	Ketotifen
DB00921	Buprenorphine
DB00922	Levosimendan
DB00923	Ceforanide
DB00924	Cyclobenzaprine
DB00925	Phenoxybenzamine
DB00927	Famotidine
DB00928	Azacitidine
DB00929	Misoprostol
DB00931	Methacycline
DB00932	Tipranavir
DB00933	Mesoridazine
DB00934	Maprotiline
DB00935	Oxymetazoline
DB00936	Salicyclic acid

DB00266	Dicumarol
DB00267	Cefmenoxime
DB00268	Ropinirole
DB00269	Chlorotrianisene
DB00270	Isradipine
DB00271	Diatrizoate
DB00272	Betazole
DB00273	Topiramate
DB00274	Cefmetazole
DB00275	Olmesartan
DB00276	Amsacrine
DB00277	Theophylline
DB00278	Argatroban
DB00279	Liothyronine
DB00280	Disopyramide
DB00281	Lidocaine
DB00282	Pamidronate
DB00283	Clemastine
DB00284	Acarbose
DB00285	Venlafaxine
DB00286	Estrone sulfate
DB00287	Travoprost
DB00288	Amcinonide
DB00289	Atomoxetine
DB00291	Chlorambucil
DB00292	Etomidate
DB00293	Raltitrexed
DB00294	Etonogestrel
DB00295	Morphine
DB00296	Ropivacaine
DB00298	Dapiprazole
DB00299	Penciclovir
DB00300	Tenofovir
DB00301	Flucloxacillin
DB00302	Tranexamic acid
DB00303	Ertapenem
DB00304	Desogestrel
DB00305	Mitomycin
DB00306	Talbutal
DB00307	Bexarotene
DB00308	Ibutilide
DB00309	Vindesine
DB00310	Chlorthalidone
DB00311	Ethoxzolamide
DB00312	Pentobarbital
DB00313	Valproic acid
DB00315	Zolmitriptan
DB00316	Acetaminophen
DB00317	Gefitinib
DB00318	Codeine
DB00319	Piperacillin
DB00320	Dihydroergotamine
DB00937	Diethylpropion
DB00938	Salmeterol
DB00939	Meclofenamic acid
DB00940	Methantheline
DB00941	Hexafluronium bromide
DB00942	Cycrimine
DB00943	Zalcitabine
DB00944	Demecarium bromide
DB00945	Aspirin
DB00946	Phenprocoumon
DB00947	Fulvestrant
DB00948	Mezlocillin
DB00949	Felbamate
DB00950	Fexofenadine
DB00951	Isoniazid
DB00952	Naratriptan
DB00953	Rizatriptan
DB00954	Dirithromycin
DB00955	Netilmicin
DB00956	Hydrocodone
DB00957	Norgestimate
DB00958	Carboplatin
DB00959	Methylprednisolone
DB00960	Pindolol
DB00961	Mepivacaine
DB00962	Zaleplon
DB00963	Bromfenac
DB00964	Apraclonidine
DB00966	Telmisartan
DB00967	Desloratadine
DB00968	Methyldopa
DB00969	Alosetron
DB00972	Azelastine
DB00973	Ezetimibe
DB00974	Edetic acid
DB00975	Dipyridamole
DB00976	Telithromycin
DB00977	Ethinyl estradiol
DB00978	Lomefloxacin
DB00979	Cyclopentolate
DB00980	Ramelteon
DB00981	Physostigmine
DB00983	Formoterol
DB00984	Nandrolone
DB00986	Glycopyrrolate
DB00987	Cytarabine
DB00988	Dopamine
DB00989	Rivastigmine
DB00990	Exemestane
DB00991	Oxaprozin
DB00992	Methyl aminolevulinate
DB00993	Azathioprine

DB00321	Amitriptyline
DB00322	Floxuridine
DB00323	Tolcapone
DB00324	Fluorometholone
DB00326	Calcium gluceptate
DB00327	Hydromorphone
DB00328	Indomethacin
DB00330	Ethambutol
DB00331	Metformin
DB00332	Ipratropium
DB00333	Methadone
DB00334	Olanzapine
DB00335	Atenolol
DB00336	Nitrofurazone
DB00337	Pimecrolimus
DB00339	Pyrazinamide
DB00340	Metixene
DB00341	Cetirizine
DB00342	Terfenadine
DB00343	Diltiazem
DB00344	Protriptyline
DB00345	Aminohippurate
DB00346	Alfuzosin
DB00347	Trimethadione
DB00348	Nitisinone
DB00349	Clobazam
DB00350	Minoxidil
DB00351	Megestrol
DB00352	Thioguanine
DB00353	Methylergonovine
DB00354	Buclizine
DB00355	Aztreonam
DB00356	Chlorzoxazone
DB00357	Aminoglutethimide
DB00358	Mefloquine
DB00359	Sulfadiazine
DB00360	Tetrahydrobiopterin
DB00361	Vinorelbine
DB00363	Clozapine
DB00365	Grepafloxacin
DB00366	Doxylamine
DB00367	Levonorgestrel
DB00368	Norepinephrine
DB00369	Cidofovir
DB00370	Mirtazapine
DB00371	Meprobamate
DB00372	Thiethylperazine
DB00373	Timolol
DB00374	Treprostinil
DB00375	Colestipol
DB00376	Trihexyphenidyl
DB00377	Palonosetron
DB00996	Gabapentin
DB00997	Doxorubicin
DB00998	Frovatriptan
DB00999	Hydrochlorothiazide
DB01000	Cyclacillin
DB01001	Salbutamol
DB01002	Levobupivacaine
DB01003	Cromoglicate
DB01004	Ganciclovir
DB01005	Hydroxyurea
DB01006	Letrozole
DB01007	Tioconazole
DB01008	Busulfan
DB01009	Ketoprofen
DB01010	Edrophonium
DB01011	Metyrapone
DB01012	Cinacalcet
DB01013	Clobetasol
DB01014	Balsalazide
DB01015	Sulfamethoxazole
DB01016	Glibenclamide
DB01017	Minocycline
DB01018	Guanfacine
DB01019	Bethanechol
DB01020	Isosorbide mononitrate
DB01021	Trichlormethiazide
DB01022	Phytonadione
DB01023	Felodipine
DB01024	Mycophenolic acid
DB01025	Amlexanox
DB01026	Ketoconazole
DB01028	Methoxyflurane
DB01029	Irbesartan
DB01030	Topotecan
DB01031	Ethinamate
DB01032	Probenecid
DB01033	Mercaptopurine
DB01034	Cerulenin
DB01035	Procainamide
DB01036	Tolterodine
DB01037	Selegiline
DB01038	Carphenazine
DB01039	Fenofibrate
DB01040	Hydroxystilbamidine isethionate
DB01041	Thalidomide
DB01042	Melphalan
DB01043	Memantine
DB01044	Gatifloxacin
DB01046	Lubiprostone
DB01047	Fluocinonide
DB01048	Abacavir
DB01049	Ergoloid mesylate

DB00378	Dydrogesterone
DB00379	Mexiletine
DB00380	Dexrazoxane
DB00381	Amlodipine
DB00382	Tacrine
DB00383	Oxyphencyclimine
DB00384	Triamterene
DB00385	Valrubicin
DB00387	Procyclidine
DB00388	Phenylephrine
DB00389	Carbimazole
DB00390	Digoxin
DB00391	Sulpiride
DB00392	Ethopropazine
DB00393	Nimodipine
DB00394	Beclomethasone
DB00395	Carisoprodol
DB00396	Progesterone
DB00397	Phenylpropanolamine
DB00398	Sorafenib
DB00399	Zoledronate
DB00400	Griseofulvin
DB00401	Nisoldipine
DB00402	Eszopiclone
DB00404	Alprazolam
DB00405	Dexbrompheniramine
DB00406	Gentian violet
DB00408	Loxapine
DB00409	Remoxipride
DB00410	Mupirocin
DB00411	Carbachol
DB00412	Rosiglitazone
DB00413	Pramipexole
DB00414	Acetohexamide
DB00415	Ampicillin
DB00417	Penicillin V
DB00418	Secobarbital
DB00419	Miglustat
DB00420	Promazine
DB00421	Spironolactone
DB00422	Methylphenidate
DB00423	Methocarbamol
DB00424	Hyoscyamine
DB00425	Zolpidem
DB00426	Famciclovir
DB00427	Triprolidine
DB00428	Streptozocin
DB00429	Carboprost tromethamine
DB00430	Cefpiramide
DB00431	Lindane
DB00432	Trifluridine
DB00433	Prochlorperazine
DB01050	Ibuprofen
DB01051	Novobiocin
DB01053	Penicillin G
DB01054	Nitrendipine
DB01055	Mimosine
DB01056	Tocainide
DB01057	Echothiophate iodide
DB01058	Praziquantel
DB01059	Norfloxacin
DB01060	Amoxicillin
DB01061	Azlocillin
DB01062	Oxybutynin
DB01063	Acetophenazine
DB01064	Isoproterenol
DB01065	Melatonin
DB01066	Cefditoren
DB01067	Glipizide
DB01068	Clonazepam
DB01069	Promethazine
DB01070	Dihydrotachysterol
DB01071	Mequitazine
DB01072	Atazanavir
DB01073	Fludarabine
DB01074	Perhexiline
DB01075	Diphenhydramine
DB01076	Atorvastatin
DB01077	Etidronic acid
DB01080	Vigabatrin
DB01081	Diphenoxylate
DB01082	Streptomycin
DB01083	Orlistat
DB01084	Emedastine
DB01085	Pilocarpine
DB01086	Benzocaine
DB01087	Primaquine
DB01088	Iloprost
DB01089	Deserpidine
DB01090	Pentolinium
DB01091	Butenafine
DB01092	Ouabain
DB01093	Dimethyl sulfoxide
DB01094	Hesperetin
DB01095	Fluvastatin
DB01096	Oxamniquine
DB01097	Leflunomide
DB01098	Rosuvastatin
DB01099	Flucytosine
DB01100	Pimozide
DB01101	Capecitabine
DB01102	Arbutamine
DB01103	Quinacrine
DB01104	Sertraline

DB00434	Cyproheptadine
DB00436	Bendroflumethiazide
DB00437	Allopurinol
DB00438	Ceftazidime
DB00440	Trimethoprim
DB00441	Gemcitabine
DB00442	Entecavir
DB00443	Betamethasone
DB00444	Teniposide
DB00445	Epirubicin
DB00446	Chloramphenicol
DB00447	Loracarbef
DB00448	Lansoprazole
DB00449	Dipivefrin
DB00450	Droperidol
DB00451	Levothyroxine
DB00452	Framycetin
DB00453	Clomocycline
DB00454	Meperidine
DB00455	Loratadine
DB00456	Cefalotin
DB00457	Prazosin
DB00458	Imipramine
DB00459	Acitretin
DB00460	Verteporfin
DB00461	Nabumetone
DB00462	Methylscopolamine
DB00463	Metharbital
DB00464	Sodium tetradecyl sulfate
DB00465	Ketorolac
DB00466	Picrotoxin
DB00467	Enoxacin
DB00468	Quinine
DB00469	Tenoxicam
DB00470	Marinol
DB00471	Montelukast
DB00472	Fluoxetine
DB00473	Hexylcaine
DB00474	Methohexital
DB00475	Chlordiazepoxide
DB00476	Duloxetine
DB00477	Chlorpromazine
DB00478	Rimantadine
DB00479	Amikacin
DB00480	Lenalidomide
DB00481	Raloxifene
DB00482	Celecoxib
DB00483	Gallamine triethiodide
DB00484	Brimonidine
DB00485	Dicloxacillin
DB00486	Nabilone
DB00487	Pefloxacin
DB01105	Sibutramine
DB01106	Levocabastine
DB01107	Methyprylon
DB01108	Trilostane
DB01110	Miconazole
DB01112	Cefuroxime
DB01113	Papaverine
DB01114	Chlorpheniramine
DB01115	Nifedipine
DB01116	Trimethaphan
DB01117	Atovaquone
DB01118	Amiodarone
DB01119	Diazoxide
DB01120	Gliclazide
DB01121	Phenacemide
DB01122	Ambenonium
DB01123	Proflavine
DB01124	Tolbutamide
DB01125	Anisindione
DB01126	Dutasteride
DB01127	Econazole
DB01128	Bicalutamide
DB01129	Rabeprazole
DB01130	Prednicarbate
DB01131	Proguanil
DB01132	Pioglitazone
DB01133	Tiludronate
DB01134	Desoxycorticosterone pivalate
DB01136	Carvedilol
DB01137	Levofloxacin
DB01138	Sulfinpyrazone
DB01139	Cephapirin
DB01140	Cefadroxil
DB01142	Doxepin
DB01143	Amifostine
DB01144	Dichlorphenamide
DB01145	Sulfoxone
DB01146	Diphenylpyraline
DB01147	Cloxacillin
DB01148	Flavoxate
DB01149	Nefazodone
DB01150	Cefprozil
DB01151	Desipramine
DB01153	Sertaconazole
DB01154	Thiamylal
DB01155	Gemifloxacin
DB01156	Bupropion
DB01157	Trimetrexate
DB01158	Bretylium
DB01159	Halothane
DB01160	Dinoprost tromethamine
DB01161	Chloroprocaine

DB00488	Altretamine
DB00489	Sotalol
DB00490	Buspirone
DB00491	Miglitol
DB00492	Fosinopril
DB00493	Cefotaxime
DB00494	Entacapone
DB00495	Zidovudine
DB00496	Darifenacin
DB00497	Oxycodone
DB00498	Phenindione
DB00499	Flutamide
DB00500	Tolmetin
DB00501	Cimetidine
DB00502	Haloperidol
DB00503	Ritonavir
DB00504	Levallorphan
DB00505	Tridihexethyl
DB00507	Nitazoxanide
DB00508	Triflupromazine
DB00513	Aminocaproic acid
DB00514	Dextromethorphan
DB00517	Anisotropine methylbromide
DB00518	Albendazole
DB00519	Trandolapril
DB00521	Carteolol
DB00522	Bentiromide
DB00524	Metolazone
DB00525	Tolnaftate
DB00526	Oxaliplatin
DB00527	Dibucaine
DB00528	Lercanidipine
DB00529	Foscarnet
DB00530	Erlotinib
DB00531	Cyclophosphamide
DB00532	Mephenytoin
DB00535	Cefdinir
DB00536	Guanidine
DB00537	Ciprofloxacin
DB00539	Toremifene
DB00540	Nortriptyline
DB00541	Vincristine
DB00542	Benazepril
DB00543	Amoxapine
DB00544	Fluorouracil
DB00545	Pyridostigmine
DB00546	Adinazolam
DB00547	Desoximetasone
DB00548	Azelaic acid
DB00549	Zafirlukast
DB00550	Propylthiouracil
DB00551	Acetohydroxamic acid
DB01162	Terazosin
DB01165	Ofloxacin
DB01166	Cilostazol
DB01167	Itraconazole
DB01168	Procarbazine
DB01170	Guanethidine
DB01171	Moclobemide
DB01172	Kanamycin
DB01173	Orphenadrine
DB01174	Phenobarbital
DB01175	Escitalopram
DB01176	Cyclizine
DB01177	Idarubicin
DB01178	Chlormezanone
DB01179	Podofilox
DB01180	Rescinnamine
DB01181	Ifosfamide
DB01182	Propafenone
DB01183	Naloxone
DB01184	Domperidone
DB01185	Fluoxymesterone
DB01186	Pergolide
DB01187	Iophendylate
DB01188	Ciclopirox
DB01189	Desflurane
DB01190	Clindamycin
DB01191	Dexfenfluramine
DB01192	Oxymorphone
DB01193	Acebutolol
DB01194	Brinzolamide
DB01195	Flecainide
DB01196	Estramustine
DB01197	Captopril
DB01199	Tubocurarine
DB01200	Bromocriptine
DB01202	Levetiracetam
DB01203	Nadolol
DB01204	Mitoxantrone
DB01205	Flumazenil
DB01206	Lomustine
DB01207	Ridogrel
DB01208	Sparfloxacin
DB01209	Dezocine
DB01210	Levobunolol
DB01212	Ceftriaxone
DB01213	Fomepizole
DB01214	Metipranolol
DB01215	Estazolam
DB01216	Finasteride
DB01217	Anastrozole
DB01218	Halofantrine
DB01219	Dantrolene

DB00552	Pentostatin
DB00553	Methoxsalen
DB00554	Piroxicam
DB00555	Lamotrigine
DB00556	Perflutren
DB00557	Hydroxyzine
DB00558	Zanamivir
DB00559	Bosentan
DB00560	Tigecycline
DB00561	Doxapram
DB00562	Benzthiazide
DB00563	Methotrexate
DB00564	Carbamazepine
DB00566	Succimer
DB00567	Cephalexin
DB00568	Cinnarizine
DB00570	Vinblastine
DB00571	Propranolol
DB00572	Atropine
DB00573	Fenoprofen
DB00575	Clonidine
DB00576	Sulfamethizole
DB00577	Valaciclovir
DB00578	Carbenicillin
DB00579	Mazindol
DB00581	Lactulose
DB00582	Voriconazole
DB00583	L-carnitine
DB00584	Enalapril
DB00585	Nizatidine
DB00586	Diclofenac
DB00587	Cinalukast
DB00588	Fluticasone propionate
DB00589	Lisuride
DB00590	Doxazosin
DB00591	Fluocinolone acetonide
DB00592	Piperazine
DB00593	Ethosuximide
DB00594	Amiloride
DB00595	Oxytetracycline
DB00596	Halobetasol propionate
DB00597	Gadoteridol
DB00598	Labetalol
DB00599	Thiopental
DB00600	Monobenzone
DB00601	Linezolid
DB00603	Medroxyprogesterone
DB00604	Cisapride
DB00605	Sulindac
DB00606	Cyclothiazide
DB00607	Nafcillin
DB00608	Chloroquine
DB01220	Rifaximin
DB01221	Ketamine
DB01222	Budesonide
DB01224	Quetiapine
DB01227	Levomethadyl acetate
DB01228	Encainide
DB01229	Paclitaxel
DB01231	Diphenidol
DB01232	Saquinavir
DB01233	Metoclopramide
DB01234	Dexamethasone
DB01235	Levodopa
DB01236	Sevoflurane
DB01237	Bromodiphenhydramine
DB01238	Aripiprazole
DB01239	Chlorprothixene
DB01240	Epoprostenol
DB01241	Gemfibrozil
DB01242	Clomipramine
DB01243	Chloroxine
DB01244	Bepridil
DB01245	Decamethonium
DB01246	Trimeprazine
DB01247	Isocarboxazid
DB01248	Docetaxel
DB01250	Olsalazine
DB01251	Gliquidone
DB01252	Mitiglinide
DB01253	Ergonovine
DB01254	Dasatinib
DB01255	Lisdexamfetamine
DB01256	Retapamulin
DB01258	Aliskiren
DB01259	Lapatinib
DB01260	Desonide
DB01261	Sitagliptin
DB01262	Decitabine
DB01264	Darunavir
DB01265	Telbivudine
DB01267	Paliperidone
DB01268	Sunitinib
DB01273	Varenicline
DB01274	Arformoterol
DB01275	Hydralazine
DB01280	Nelarabine
DB01283	Lumiracoxib
DB01288	Fenoterol
DB01289	Glisoxepide
DB01291	Pirbuterol
DB01295	Bevantolol
DB01296	Glucosamine
DB01297	Practolol

DB00609	Ethionamide
DB00610	Metaraminol
DB00611	Butorphanol
DB00612	Bisoprolol
DB00613	Amodiaquine
DB00614	Furazolidone
DB00616	Candoxatril
DB00617	Paramethadione
DB00618	Demeclocycline
DB00619	Imatinib
DB00620	Triamcinolone
DB00621	Oxandrolone
DB00622	Nicardipine
DB00623	Fluphenazine
DB00624	Testosterone
DB00625	Efavirenz
DB00627	Niacin
DB00628	Clorazepate
DB00629	Guanabenz
DB00630	Alendronate
DB00631	Clofarabine
DB00632	Docosanol
DB00633	Dexmedetomidine
DB00634	Sulfacetamide
DB00635	Prednisone
DB00636	Clofibrate
DB00637	Astemizole
DB00639	Butoconazole
DB00640	Adenosine
DB00641	Simvastatin
DB00642	Pemetrexed
DB00643	Mebendazole
DB00645	Dyclonine
DB00647	Propoxyphene
DB00648	Mitotane
DB00649	Stavudine
DB00650	Leucovorin
DB00651	Dyphylline
DB00652	Pentazocine
DB00654	Latanoprost
DB00655	Estrone
DB00656	Trazodone
DB00657	Mecamylamine
DB00659	Acamprosate
DB00660	Metaxalone
DB00661	Verapamil
DB00662	Trimethobenzamide
DB00663	Flumethasone pivalate
DB00664	Sulfametopyrazine
DB00665	Nilutamide
DB00667	Histamine Phosphate
DB00668	Epinephrine
DB01298	Sulfacytine
DB01299	Sulfadoxine
DB01301	Rolitetracycline
DB01319	Fosamprenavir
DB01320	Fosphenytoin
DB01324	Polythiazide
DB01325	Quinethazone
DB01326	Cefamandole
DB01327	Cefazolin
DB01328	Cefonicid
DB01330	Cefotetan
DB01331	Cefoxitin
DB01332	Ceftizoxime
DB01333	Cefradine
DB01340	Cilazapril
DB01342	Forasartan
DB01344	Polystyrene sulfonate
DB01348	Spirapril
DB01349	Tasosartan
DB01351	Amobarbital
DB01352	Aprobarbital
DB01353	Butethal
DB01354	Heptabarbital
DB01355	Hexobarbital
DB01357	Mestranol
DB01359	Penbutolol
DB01364	Ephedrine
DB01365	Mephentermine
DB01366	Procaterol
DB01367	Rasagiline
DB01380	Cortisone acetate
DB01382	Glycodiazine
DB01384	Paramethasone
DB01392	Yohimbine
DB01393	Bezafibrate
DB01394	Colchicine
DB01395	Drospirenone
DB01399	Salsalate
DB01400	Neostigmine
DB01403	Methotrimeprazine
DB01406	Danazol
DB01407	Clenbuterol
DB01408	Bambuterol
DB01409	Tiotropium
DB01410	Ciclesonide
DB01411	Pranlukast
DB01412	Theobromine
DB01413	Cefepime
DB01414	Cefacetrile
DB01415	Ceftibuten
DB01416	Cefpodoxime
DB01418	Acenocoumarol

DB00669	Sumatriptan
DB00670	Pirenzepine
DB00671	Cefixime
DB00672	Chlorpropamide
DB00673	Aprepitant
DB00674	Galantamine
DB00675	Tamoxifen
DB00676	Benzyl benzoate
DB00677	Isoflurophate
DB00678	Losartan
DB00679	Thioridazine
DB00680	Moricizine
DB00682	Warfarin
DB00683	Midazolam
DB00684	Tobramycin
DB00685	Trovafloxacin
DB00686	Pentosan polysulfate
DB00687	Fludrocortisone
DB00688	Mycophenolate mofetil
DB00689	Cephaloglycin
DB00690	Flurazepam
DB00691	Moexipril
DB00692	Phentolamine
DB00693	Fluorescein
DB00694	Daunorubicin
DB00695	Furosemide
DB00696	Ergotamine
DB00697	Tizanidine
DB00698	Nitrofurantoin
DB00699	Nicergoline
DB00700	Eplerenone
DB00701	Amprenavir
DB00703	Methazolamide
DB00704	Naltrexone
DB00705	Delavirdine
DB00706	Tamsulosin
DB00708	Sufentanil
DB00709	Lamivudine
DB00710	Ibandronate
DB00711	Diethylcarbamazine
DB00712	Flurbiprofen
DB00714	Apomorphine
DB00715	Paroxetine
DB00716	Nedocromil
DB00717	Norethindrone
DB00718	Adefovir dipivoxil
DB00719	Azatadine
DB00720	Clodronate
DB00721	Procaine
DB00722	Lisinopril
DB00723	Methoxamine
DB00724	Imiquimod
DB01419	Antrafenine
DB01420	Testosterone propionate
DB01422	Nitroxoline
DB01423	Stepronin
DB01424	Aminophenazone
DB01425	Alizapride
DB01426	Ajmaline
DB01427	Amrinone
DB01428	Oxybenzone
DB01429	Aprindine
DB01430	Almitrine
DB01431	Allylestrenol
DB01435	Antipyrine
DB01436	Alfacalcidol
DB01437	Glutethimide
DB01438	Phenazopyridine
DB01440	Gamma hydroxybutyric acid
DB01463	Fencamfamine
DB01544	Flunitrazepam
DB01558	Bromazepam
DB01559	Clotiazepam
DB01567	Fludiazepam
DB01576	Dextroamphetamine
DB01577	Methamphetamine
DB01579	Phendimetrazine
DB01580	Oxprenolol
DB01581	Sulfamerazine
DB01582	Sulfamethazine
DB01586	Ursodeoxycholic acid
DB01587	Ketazolam
DB01588	Prazepam
DB01589	Quazepam
DB01591	Solifenacin
DB01594	Cinolazepam
DB01595	Nitrazepam
DB01597	Cilastatin
DB01598	Imipenem
DB01599	Probucol
DB01600	Tiaprofenic acid
DB01602	Bacampicillin
DB01603	Meticillin
DB01604	Pivampicillin
DB01605	Pivmecillinam
DB01606	Tazobactam
DB01607	Ticarcillin
DB01608	Propericiazine
DB01609	Deferasirox
DB01610	Valganciclovir
DB01611	Hydroxychloroquine
DB01612	Amyl nitrite
DB01613	Erythrityl tetranitrate
DB01614	Acepromazine

DB00725	Homatropine methylbromide
DB00726	Trimipramine
DB00727	Nitroglycerin
DB00728	Rocuronium
DB00729	Diphemanil methylsulfate
DB00730	Thiabendazole
DB00731	Nateglinide
DB00733	Pralidoxime
DB00734	Risperidone
DB00735	Naftifine
DB00736	Esomeprazole
DB00737	Meclizine
DB00738	Pentamidine
DB00739	Hetacillin
DB00740	Riluzole
DB00741	Hydrocortisone
DB00742	Mannitol
DB00744	Zileuton
DB00745	Modafinil
DB00746	Deferoxamine
DB00747	Scopolamine
DB00748	Carbinoxamine
DB00749	Etodolac
DB00750	Prilocaine
DB00751	Epinastine
DB00752	Tranylcypromine
DB00753	Isoflurane
DB00754	Ethotoin
DB00755	Tretinoin
DB00756	Hexachlorophene
DB00757	Dolasetron
DB00758	Clopidogrel
DB00759	Tetracycline
DB00760	Meropenem
DB00762	Irinotecan
DB00763	Methimazole
DB00764	Mometasone
DB00765	Metyrosine
DB00766	Clavulanate
DB00767	Benzquinamide
DB00768	Olopatadine
DB00769	Hydrocortamate
DB00770	Alprostadil
DB00771	Clidinium
DB00772	Malathion
DB00773	Etoposide
DB00774	Hydroflumethiazide
DB00775	Tirofiban
DB00776	Oxcarbazepine
DB00777	Propiomazine
DB00778	Roxithromycin
DB00779	Nalidixic acid
DB01615	Aceprometazine
DB01616	Alverine
DB01618	Molindone
DB01619	Phenindamine
DB01620	Pheniramine
DB01621	Pipotiazine
DB01622	Thioproperazine
DB01623	Thiothixene
DB01624	Zuclopenthixol
DB01625	Isopropamide
DB01626	Pargyline
DB01627	Lincomycin
DB01628	Etoricoxib
DB02300	Calcipotriol
DB02546	Vorinostat
DB02703	Fusidic acid
DB04552	Niflumic acid
DB04570	Latamoxef
DB04573	Estriol
DB04575	Quinestrol
DB04794	Bifonazole
DB04835	Maraviroc
DB04837	Chlophedianol
DB04838	Cyclandelate
DB04839	Cyproterone
DB04840	Debrisoquin
DB04841	Flunarizine
DB04842	Fluspirilene
DB04843	Mepenzolate
DB04844	Tetrabenazine
DB04861	Nebivolol
DB04878	Voglibose
DB04880	Enoximone
DB04890	Bepotastine
DB04896	Milnacipran
DB04898	Ximelagatran
DB04930	Permethrin
DB04942	Tamibarotene
DB04948	Lofexidine
DB04967	Lucanthone
DB05245	Silver sulfadiazine
DB05246	Methsuximide
DB06144	Sertindole
DB06148	Mianserin
DB06151	Acetylcysteine
DB06155	Rimonabant
DB06262	Droxidopa
DB06267	Udenafil
DB06274	Alvimopan
DB06288	Amisulpride
DB06439	Tyloxapol
DB06689	Ethanolamine oleate

Appendix 2

List of the names and the Protein Data Bank entries of the used proteins.

PDB code	Protein name
13gs	Glutathione S-transferase pi
1a3b	Prothrombin
1aj0	Dihydropteroate synthase
1aj6	DNA topoisomerase II
1apy	Human aspartylglucosaminidase
1aq1	Human cyclin dependent kinase 2
1auk	Human arylsulfatase A
1b2y	Pancreatic alpha-amylase precursor
1b3d	Stromelysin-1
1bj4	Serine hydroxymethyltransferase, cytosolic
1bj5	Human serum albumin
1blc	Beta-lactamase
1bwc	Glutathione reductase (mitochondrial)
1bzm	Carbonic anhydrase I
1c5o	Urokinase type plasminogen activator
1cjf	Profilin
1cjy	Enoyl-[acyl-carrier-protein] reductase
1d3g	Human dihydroorotate dehidrogenase
1dfv	Human neutrophil gelatinase
1dkf	Retinoic acid receptor alpha
1dug	Glutathione S-transferase from *Schistosoma japonicum*
1e51	Delta-aminolevulinic acid dehydratase
1ewf	Bactericidal permeability-increasing protein
1exa	Retinoic acid receptor gamma-2
1ezf	Human squalene synthase
1f0x	D-lactate dehydrogenase
1f5f	Sex hormone-binding globulin precursor
1fcy	Retinoic acid receptor gamma-1
1fj4	3-oxoacyl-[acyl-carrier-protein] synthase I
1fkd	FK506-binding protein 1A
1g3m	Human estrogen sulfotransferase
1g9v	Hemoglobin
1gkc	Matrix metalloprotease 9
1hck	Human cyclin-dependent kinase 2
1hcn	Human chorionic gondadotropin
1hrn	Renin
1hso	Alcohol dehydrogenase alpha chain
1hsz	Alcohol dehydrogenase beta chain
1ht0	Alcohol dehydrogenase gamma chain
1hur	Human ADP-ribosylation factor 1
1hvr	HIV-1 protease
1ig3	Thiamine pyrophosphokinase
1j3j	Bifunctional dihydrofolate reductase-thymidylate synthase
1j8u	Phenylalanine-4-hydroxylase
1jmo	Prothrombin

1k0e	Para-aminobenzoate synthase component I
1kfy	Fumarate reductase flavoprotein subunit
1ki0	Human angiostatin
1kpg	Cyclopropane-fatty-acyl-phospholipid synthase 1
1ksp	DNA polymerase I
1kvo	Human phospholipase A2
1l7z	Calmodulin
1lo6	Human kallikrein 6
1lpb	Pancreatic triacylglycerol lipase precursor
1lpg	Coagulation factor X
1lxi	Bone morphogenic protein 7
1mf8	Calcineurin B subunit isoform 1
1mp8	Focal adhesion kinase
1mzs	3-oxoacyl-[acyl-carrier-protein] synthase III
1n52	Cap binding complex
1n5u	Serum albumin precursor
1nhz	Glucocorticoid receptor
1nrg	Pyridoxine-5'-phosphate oxidase
1of1	Thymidine kinase
1okc	ADP/ATP carrier protein heart isoform T1
1opb	Retinol-binding protein II
1oq5	Carbonic anhydrase II
1oth	Human ornithine transcarbamoylase
1p0p	Cholinesterase
1p60	Deoxycytidine kinase
1ph0	Tyrosine-protein phosphatase, non-receptor type 1
1qh5	Human glyoxalase II
1qkm	Human oestrogen receptor beta
1qon	Acetylcholinesterase precursor
1r1h	Enkephalinase
1r5l	Human alpha-tocopherol transfer protein
1r9o	Cytochrome P4502C9
1rbp	Plasma retinol-binding protein
1ro9	cAMP phosphodiesterase
1rsz	Purine nucleoside phosphorylase
1rwx	Caspase-1 precursor
1s1d	Human apyrase
1s2c	Prostaglandin D2 11-ketoreductase
1s3v	Dihydrofolate reductase
1sr7	Progesterone receptor hormone
1sz7	Human BET3
1t46	Mast/stem cell growth factor receptor precursor
1t65	Androgen receptor
1uae	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
1uhl	Retinoic acid receptor RXR-beta
1uze	Angiotensin-converting enzyme
1v97	Xanthine oxidase
1w6k	Human oxidosqualene cyclase
1x9d	Human class I alpha-1,2-mannosidase
1x9n	DNA ligase I
1xap	Retinoic acid receptor beta
1xkk	Epidermal growth factor receptor precursor

1xpc	Estrogen receptor
1xzx	Thyroid hormone receptor beta-1
1y6a	Vascular endothelial growth factor receptor 2 precursor
1yb5	Human zeta-crystallin
1ytv	Vasopressin V1a receptor
1z57	Human CLK1
1zcm	Human calpain protease
1zd3	Human soluble epoxide hydrolase
1zid	Enoyl-[acyl-carrier-protein] reductase
1zx0	Human guanidinoacetate N-methyltransferase
1zxm	Human topo IIa ATPase/AMP-PNP
1zy7	Human adenosine deaminase that acts on RNA
1zsq	Miotubularin-related protein 2
1zsx	Human potassium channel Kv beta subunit
2a1h	Human mitochondrial branched chain aminotransferase
2a3i	Mineralocorticoid receptor
2a5d	Human ADP-ribosylation factor 6
2aax	Mineralocorticoid receptor
2aeb	Human arginase I
2afw	Human glutaminyl cyclase
2ag4	GM2-activator protein
2aid	HIV-1 reverse transcriptase
2avd	Human catechol-O-methyltransferase
2axm	Heparin-binding growth factor 1 precursor
2axn	Human inducible form 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase
2az5	Tumor necrosis factor alpha
2b2u	Human CDH1
2bat	Neuraminidase
2bka	CC3
2bm2	Human beta-II tryptase
2bxs	Monoamine oxidase A
2c67	Monoamine oxidase B
2cbz	Human multidrug resistance protein 1
2cca	Peroxidase/catalase T
2cjz	Human protein tyrosine phosphatase PTPN5
2cmd	*E. coli* malate dehydrogenase
2cmw	Human casein kinase 1 gamma-1
2d0t	Human indoleamine 2,3-dioxygenase
2f4j	Proto-oncogene tyrosine-protein kinase ABL1
2f6q	Human peroxisomal delta3, delta2 enoyl CoA isomerase
2fbr	Transthyretin precursor
2fvv	Human diphosphoinositol polyphosphate phosphohydrolase 1
2fy3	Human choline acetyltransferase
2g5r	Siglec-7
2g72	Human phenylethanolamine N-methyltransferase
2gwh	Human sulfotranferase SULT1C2
2h7j	Cathepsin S
2ipx	Human fibrillarin
2iwz	Human mitochondrial beta-ketoacyl ACP synthase
2jis	Human cysteine sulfinic acid decarboxylase
2oaz	Human methionine aminopeptidase-2
2ozu	Human MYST histone acetyltransferase 3

2p0a	Human synapsin III
2p54	PPAR alpha
2pk4	Human plasminogen kringle
3fap	FKBP12-rapamycin complex-associated protein
3nos	Nitric-oxide synthase, endothelial

Appendix 3

Prediction and validation properties of the studied 181 effect categories. Mean value for a random classification was calculated by dividing the number of registrered drugs to a given effect by 1,226, the total number of drugs.

Effect	Number of drugs	Accuracy	Leave-one-out validation probability value				
		AUC	Mean	Mean of the upper 75%	Mean for a random classification	Mean/ random mean	Mean/random mean for the upper 75%
2-Hydroxy-3-aminopropoxy derivative	21	0.9942	0.6488	0.8714	0.0171	37.9	50.5
Adrenergic agent	133	0.9037	0.6165	0.7812	0.1085	5.7	7.6
Adrenergic agonist	38	0.9657	0.6181	0.8300	0.0310	19.9	26.6
Adrenergic alpha-agonist	20	0.9908	0.5755	0.7672	0.0163	35.3	47.0
Adrenergic alpha-antagonist	28	0.9766	0.4035	0.5325	0.0228	17.7	23.6
Adrenergic antagonist	62	0.9468	0.5920	0.7507	0.0506	11.7	15.6
Adrenergic beta-agonist	17	0.9937	0.6037	0.8237	0.0139	43.5	58.0
Adrenergic beta-antagonist	23	0.9957	0.6214	0.8173	0.0188	33.1	44.2
Adrenergic uptake inhibitor	19	0.9880	0.3700	0.5008	0.0155	23.9	31.8
Alkylating agent	17	0.9946	0.3077	0.4359	0.0139	22.2	29.6
Amphetamine	17	0.9730	0.5698	0.8072	0.0139	41.1	54.8
Analgesic agent	94	0.8908	0.5210	0.6825	0.0767	6.8	9.1
Analgesic agent, non-narcotic	13	0.9715	0.1113	0.1608	0.0106	10.5	14.0
Analgesic agent, opioid	23	0.9891	0.5881	0.7954	0.0188	31.3	41.8
Anesthetic agent	45	0.9502	0.4641	0.6159	0.0367	12.6	16.9
Anesthetic agent, intravenous	12	0.9942	0.2465	0.3287	0.0098	25.2	33.6
Anesthetic agent, local	25	0.9822	0.4394	0.6010	0.0204	21.5	28.7
Angiotensin-converting enzyme inhibitor	14	0.9984	0.4401	0.5986	0.0122	36.0	48.0
Anthelmintic agent	10	0.9989	0.1784	0.2549	0.0082	21.9	29.2
Anti-allergic agent	63	0.9195	0.5683	0.7436	0.0514	11.1	14.7
Antianginal agent	21	0.9684	0.2166	0.7500	0.0171	12.6	16.9
Anti-anxiety agent	50	0.9348	0.5564	0.4322	0.0408	13.6	18.2
Antiarrhythmic agent	51	0.9256	0.4833	0.7198	0.0416	11.6	15.5
Antiasthmatic agent	31	0.9718	0.3838	0.6573	0.0253	15.2	20.2
Antibacterial agent	129	0.9367	0.6647	0.7330	0.1052	6.3	8.4
Antibiotic	135	0.9313	0.6503	0.0543	0.1101	5.9	7.9
Anticholesteremic agent	13	0.9892	0.4454	0.4994	0.0106	42.0	56.0
Anticoagulant	12	0.9966	0.0259	0.7417	0.0098	2.6	3.5
Anticonvulsant	60	0.9390	0.5990	0.4471	0.0489	12.2	16.3

Antidepressant	40	0.9593	0.5725	0.4192	0.0326	17.5	23.4
Antidepressant, second-generation	14	0.9991	0.2219	0.3727	0.0114	19.4	25.9
Anti-dyskinesia agent	26	0.9704	0.3178	0.3033	0.0212	15.0	20.0
Anti-emetic agent	49	0.9537	0.5519	0.6391	0.0400	13.8	18.4
Antifungal agent	31	0.9777	0.2789	0.5104	0.0253	11.0	14.7
Antiglaucoma agent	23	0.9872	0.3344	0.8541	0.0188	17.8	23.8
Anti-HIV agent	25	0.9864	0.4758	0.8382	0.0204	23.3	31.1
Antihypertensive agent	113	0.8956	0.5203	0.6415	0.0922	5.6	7.5
Antihypocalcemic agent	14	0.9747	0.4187	0.0346	0.0114	36.7	48.9
Anti-infective agent	219	0.8686	0.5782	0.7907	0.1786	3.2	4.3
Anti-infective agent, local	13	0.9907	0.0376	0.7494	0.0106	3.5	4.7
Anti-infective agent, urinary	14	0.9875	0.3567	0.3095	0.0114	31.2	41.6
Anti-inflammatory agent	104	0.9036	0.5679	0.3728	0.0848	6.7	8.9
Antimalarial agent	19	0.9802	0.3247	0.4513	0.0155	21.0	27.9
Antimanic agent	12	0.9944	0.0191	0.6638	0.0098	2.0	2.6
Antimetabolite	31	0.9512	0.5168	0.5862	0.0253	20.4	27.3
Anti-migraine agent	19	0.9546	0.3295	0.4405	0.0155	21.3	28.3
Antimitotic agent	10	0.9977	0.1387	0.0253	0.0082	17.0	22.7
Antimuscarinic agent	36	0.9825	0.6366	0.6965	0.0294	21.7	28.9
Antineoplastic agent	120	0.8585	0.4635	0.1982	0.0979	4.7	6.3
Antineoplastic agent, alkylating	15	0.9990	0.3530	0.8135	0.0122	28.9	38.5
Antineoplastic agent, antimetabolite	15	0.9894	0.2150	0.5950	0.0122	17.6	23.4
Antineoplastic agent, hormonal	20	0.9869	0.4051	0.4813	0.0163	24.8	33.1
Anti-obesity agent	12	0.9908	0.3144	0.2932	0.0098	32.1	42.8
Antioxidant	10	0.9874	0.0139	0.5371	0.0082	1.7	2.3
Antiparkinson agent	30	0.9711	0.3847	0.0199	0.0245	15.7	21.0
Antiprotozoal agent	19	0.9848	0.1861	0.5221	0.0155	12.0	16.0
Antipruritic agent	41	0.9414	0.492	0.2525	0.0334	14.7	19.6
Antipsychotic	45	0.9418	0.5869	0.6616	0.0367	16.0	21.3
Antipyretic	25	0.9793	0.6511	0.7885	0.0204	31.9	42.6
Antirheumatic agent	18	0.9558	0.1723	0.9024	0.0147	11.7	15.6
Antispasmodic agent	26	0.9854	0.6094	0.2386	0.0212	28.7	38.3
Antitussive	10	0.9951	0.3521	0.8224	0.0082	43.2	57.6
Anti-ulcer agent	26	0.9809	0.2754	0.5029	0.0212	13.0	17.3
Antiviral agent	48	0.9770	0.5038	0.6572	0.0392	12.9	17.2
Barbiturate	17	1.0000	0.9953	1.0000	0.0139	71.8	95.7
Benzimidazole	12	0.9967	0.2378	0.3171	0.0098	24.3	32.4
Benzodiazepine	25	0.9995	0.8950	1.0000	0.0204	43.9	58.5
Beta-lactame antibiotic	56	0.9939	0.7990	0.9905	0.0457	17.5	23.3
Bone density conservation agent	18	0.9715	0.4389	0.6077	0.0147	29.9	39.9

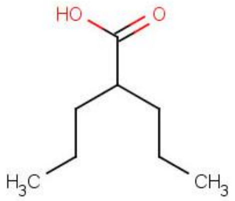
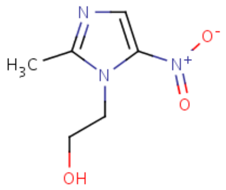
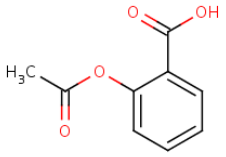
Bronchodilator agent	31	0.9587	0.3911	0.5220	0.0253	15.5	20.6
Calcium channel agent	31	0.9511	0.3786	0.5098	0.0253	15.0	20.0
Calcium channel blocker	29	0.9662	0.3866	0.5324	0.0237	16.3	21.8
Carbohydrate derivative	21	0.9992	0.4824	0.6752	0.0171	28.2	37.6
Cardiotonic agent	14	0.9922	0.2916	0.4083	0.0114	25.5	34.0
Cardiovascular agent	19	0.9881	0.3433	0.4657	0.0155	22.2	29.5
Catecholamine	11	0.9942	0.6160	0.8469	0.0090	68.7	91.5
Cell wall synthesis inhibitor	58	0.9821	0.7535	0.9707	0.0473	15.9	21.2
Central nervous system agent	26	0.9707	0.3362	0.4598	0.0212	15.9	21.1
Central nervous system stimulant	13	0.9846	0.3845	0.5554	0.0106	36.3	48.3
Cephalosporin	32	0.9936	0.7018	0.9256	0.0261	26.9	35.9
Cholinergic agent	43	0.9452	0.5340	0.7085	0.0351	15.2	20.3
Cholinergic antagonist	38	0.9551	0.5898	0.7749	0.0310	19.0	25.4
Cholinesterase inhibitor	14	0.9955	0.1605	0.2247	0.0114	14.1	18.7
Contraceptive agent	13	0.9997	0.7201	0.9641	0.0106	67.9	90.5
Corticosteroid	31	0.9978	0.8994	0.9995	0.0253	35.6	47.4
Corticosteroid, topical	12	0.9966	0.6550	0.8723	0.0098	66.9	89.2
Cyclooxygenase inhibitor	37	0.9817	0.6730	0.9087	0.0302	22.3	29.7
Depressant	37	0.9198	0.4986	0.6821	0.0302	16.5	22.0
Dermatologic agent	18	0.9891	0.2639	0.3650	0.0147	18.0	24.0
Dihydropyridine	10	0.9995	0.6420	0.9171	0.0082	78.7	104.9
Diuretic	29	0.9646	0.4986	0.6881	0.0237	21.1	28.1
Dopamine agent	76	0.9259	0.5696	0.7362	0.0620	9.2	12.3
Dopamine agonist	11	0.9969	0.1801	0.2475	0.0090	20.1	26.8
Dopamine antagonist	45	0.9652	0.6045	0.8057	0.0367	16.5	22.0
Dopamine uptake inhibitor	14	0.9921	0.1331	0.1860	0.0114	11.7	15.5
Ergoline derivative	11	0.9999	0.4524	0.6221	0.0090	50.4	67.2
Ergosterol synthesis inhibitor	12	0.9982	0.1465	0.1947	0.0098	15.0	20.0
Estrogen	11	0.9993	0.4901	0.6732	0.0090	54.6	72.8
Ethanolamine derivative	34	0.9211	0.3046	0.4110	0.0277	11.0	14.6
Fluoroquinolone	13	0.9999	0.9230	1.0000	0.0106	87.0	116.1
Folic acid antagonist	19	0.9796	0.5754	0.7805	0.0155	37.1	49.5
GABA agent	69	0.9580	0.7098	0.9230	0.0563	12.6	16.8
Gastrointestinal agent	12	0.9896	0.1099	0.1466	0.0098	11.2	15.0
Glucocorticoid	31	0.9983	0.9040	0.9996	0.0253	35.8	47.7
Glutamate receptor antagonist	20	0.9769	0.2999	0.3999	0.0163	18.4	24.5
Guanidine derivative	23	0.9863	0.5197	0.7030	0.0188	27.7	36.9
Histamine agent	74	0.9301	0.6606	0.8505	0.0604	10.9	14.6
Histamine antagonist	71	0.9283	0.6619	0.8511	0.0579	11.4	15.2
Histamine H1 antagonist	49	0.9737	0.6566	0.8516	0.0400	16.4	21.9
Histamine H1 antagonist,	10	0.9989	0.2074	0.2961	0.0082	25.4	33.9

non-sedating							
Hormone replacement agent	11	0.9990	0.2916	0.4008	0.0090	32.5	43.3
Hypnotic and/or sedative	63	0.9602	0.6423	0.8518	0.0514	12.5	16.7
Hypoglycemic agent	22	0.9858	0.4364	0.5956	0.0179	24.3	32.4
Imidazole derivative	36	0.9729	0.3502	0.4593	0.0294	11.9	15.9
Immunosuppressive agent	29	0.9734	0.3778	0.5210	0.0237	16.0	21.3
Indole derivative	20	0.9794	0.3432	0.4566	0.0163	21.0	28.1
Muscarinic agent	39	0.9818	0.6065	0.7818	0.0318	19.1	25.4
Muscle relaxant	62	0.9232	0.5066	0.6741	0.0506	10.0	13.4
Muscle relaxant, central	13	0.9945	0.1238	0.1789	0.0106	11.7	15.6
Muscle relaxant, skeletal	36	0.9612	0.5731	0.7640	0.0294	19.5	26.0
Narcotic	22	0.9892	0.6088	0.8369	0.0179	33.9	45.2
Neuroprotective agent	13	0.9785	0.0123	0.0177	0.0106	1.2	1.5
Neurotransmitter uptake inhibitor	43	0.9522	0.5795	0.7552	0.0351	16.5	22.0
Nitro compound	26	0.9929	0.6883	0.9352	0.0212	32.5	43.3
Nonsteroidal anti-inflammatory agent	71	0.9267	0.4938	0.6438	0.0579	8.5	11.4
Norepinephrine reuptake inhibitor	16	0.9949	0.4894	0.6520	0.0131	37.5	50.0
Nucleic acid synthesis inhibitor	86	0.9280	0.5633	0.7456	0.0701	8.0	10.7
Nucleoside or nucleotide	23	0.9979	0.7272	0.9599	0.0188	38.8	51.7
Nucleoside or nucleotide analogue	15	0.9990	0.4733	0.6454	0.0122	38.7	51.6
Opiate agent	31	0.9912	0.5960	0.8024	0.0253	23.6	31.4
Opiate agonist	27	0.9884	0.6181	0.8322	0.0220	28.1	37.4
Opioid	22	0.9929	0.6529	0.8977	0.0179	36.4	48.5
Parasympatholytic	18	0.9963	0.4366	0.5913	0.0147	29.7	39.6
Parasympathomimetic	12	0.9972	0.2974	0.3965	0.0098	30.4	40.5
Penicillin	20	0.9995	0.7437	0.8963	0.0163	45.6	60.8
Phenothiazine	25	0.9974	0.9101	0.9990	0.0204	44.6	59.5
Phosphodiesterase inhibitor	16	0.9939	0.1666	0.2220	0.0131	12.8	17.0
Piperazine derivative	58	0.9710	0.6615	0.8577	0.0473	14.0	18.6
Piperidine derivative	68	0.9423	0.5752	0.7286	0.0555	10.4	13.8
Platelet aggregation inhibitor	17	0.9882	0.0351	0.0494	0.0139	2.5	3.4
Potassium channel agent	18	0.9901	0.3758	0.5199	0.0147	25.6	34.1
Potassium channel blocker	16	0.9935	0.4107	0.5454	0.0131	31.5	42.0
Progestin	12	1.0000	0.8355	0.9576	0.0098	85.4	113.8
Prostaglandin	8	0.9923	0.0417	0.0556	0.0065	6.4	8.5
Prostaglandin derivative	12	0.9967	0.3578	0.4766	0.0098	36.6	48.7
Protein synthesis inhibitor	32	0.9861	0.5126	0.6832	0.0261	19.6	26.2
Purine derivative	12	0.9918	0.5476	0.7302	0.0098	55.9	74.6
Pyridine derivative	50	0.9234	0.3532	0.4650	0.0408	8.7	11.5

Pyrimidine derivative	18	0.9718	0.0767	0.1062	0.0147	5.2	7.0
Quaternary amine	39	0.9722	0.4599	0.6131	0.0318	14.5	19.3
Quinoline derivative	14	0.9877	0.2551	0.3570	0.0114	22.3	29.8
Quinolone	16	0.9990	0.8699	1.0000	0.0131	66.7	88.9
Respiratory depressant	10	0.9968	0.5089	0.7270	0.0082	62.4	83.2
Respiratory smooth muscle relaxant	14	0.9937	0.4645	0.6499	0.0114	40.7	54.2
Respiratory system agent	43	0.9356	0.4629	0.6142	0.0351	13.2	17.6
Reverse transcriptase inhibitor	14	0.9900	0.3775	0.5281	0.0114	33.1	44.1
Serotonin agent	63	0.9568	0.5719	0.7265	0.0514	11.1	14.8
Serotonin agonist	13	0.9978	0.6344	0.9028	0.0106	59.8	79.8
Serotonin antagonist	31	0.9813	0.4992	0.6625	0.0253	19.7	26.3
Serotonin reuptake inhibitor	21	0.9972	0.4714	0.6320	0.0171	27.5	36.7
Sodium channel blocker	39	0.9298	0.3872	0.5176	0.0318	12.2	16.2
Sodium chloride symporter inhibitor	13	0.9975	0.6236	0.8791	0.0106	58.8	78.4
Steroidal	74	0.9956	0.8656	0.9989	0.0604	14.3	19.1
Steroidal anti-inflammatory agent	33	0.9995	0.9611	0.9999	0.0269	35.7	47.6
Stimulant	17	0.9782	0.3476	0.4924	0.0139	25.1	33.4
Sulfonamide	80	0.9548	0.6260	0.7928	0.0653	9.6	12.8
Sulfone	17	0.9863	0.1137	0.1609	0.0139	8.2	10.9
Sulfonylurea	11	1.0000	0.4405	0.5910	0.0090	49.1	65.5
Sympatholytic	24	0.9726	0.4370	0.5823	0.0196	22.3	29.8
Sympathomimetic	33	0.9794	0.6028	0.8280	0.0269	22.4	29.9
Tetracycline	10	1.0000	0.7011	0.9993	0.0082	86.0	114.6
Tetrazole derivative	20	0.9641	0.5519	0.7352	0.0163	33.8	45.1
Thiazide	12	0.9978	0.6055	0.8004	0.0098	61.9	82.5
Thiazole	22	0.9959	0.5631	0.7736	0.0179	31.4	41.8
Tocolytic agent	11	0.9947	0.3098	0.4260	0.0090	34.5	46.0
Triazole derivative	16	0.9659	0.4133	0.5510	0.0131	31.7	42.2
Tricyclic antidepressant	14	0.9960	0.7065	0.9875	0.0114	61.9	82.5
Trifluormethyl derivative	36	0.9346	0.3159	0.4181	0.0294	10.8	14.3
Tropane derivative	11	0.9998	0.494	0.6780	0.0090	55.1	73.4
Vasoconstrictor	42	0.9628	0.5535	0.7387	0.0343	16.2	21.5
Vasodilator	78	0.9056	0.4399	0.5723	0.0636	6.9	9.2

Appendix 4

Pharmacologic comparison of valproic acid, metronidazole and acetylsalicylic acid.

	Effect	Drug		
		Valproic Acid 	**Metronidazole** 	**Acetylsalicylic Acid** 
EFFECTS /SIDE EFFECTS	Hyperammonemia with lethargy, vomiting and changes in mental status	Hyperammonemic encephalopathy [153]		Reye's syndrome: Heavy vomiting, Generalized lethargy, Hyperammonemia [154]
	Convulsions	Known side effect	Known side effect	Symptom of Aspirin overdose
	COX inhibition	Yes [78]	Yes (Figure 27a)	Yes [110]
	Strong psychiatric effects	Confusion Incoordination Abnormal dreams Personality disorder Abnormal thinking Emotional lability Aggression Hyperactivity Psychosis Depression	Confusion Incoordination Irritability Depression One report about relief of previously developed	Confusion Abnormal dreams Irrational behaviour Irritability Aggression

			psychiatric symptoms after metronidazole treatment [155]	
	Taste perversion	Known side effect	Known side effect	
	Hearing loss, tinnitus	Known side effect		Known side effect
	Renal effects	Dysuria Polyuria Urinary incontinence Urinary frequency	Dysuria Polyuria Urinary incontinence	

	Mitochondrial damage and hepatotoxicity mechanisms	CoA sequestration;
		Inhibition of β–oxidation enzymes;
		Opening of mitochondrial permeability transition pores which leads to apoptosis [156]
			Described in overdose	CoA sequestration;
				Opening of mitochondrial permeability transition pores which leads to apoptosis [156]

	Cytochrome P450 isoenzyme inhibition	CYP 2C9 [157]	CYP 2C9 [157]	CYP 2C9 (?) [158]
	Elongated bleeding time	Known side effect		Known side effect
APPLICATIONS	Suitability in migraine treatment	Yes [159]		Yes
	Anticonvulsive effect	Yes		Effective in large doses (mice);

				potentiates the effects of valproic acid [160]
	Antimicrobial effect	Yes [161] Note that diarrhea is a common side effect (13-23%)	Strong	Yes; *Heliobacter pylori* infections can be treated with Aspirin [162]
	Co-administration	With acetylsalicylic acid: co-administration should be avoided	With acetylsalicylic acid: synergistic effects	

Blank cells mean no data.

Summary

The primary aim of my PhD work was to develop an *in silico* system for the prediction of bioactivity properties of small-molecule compounds. The approach of Virtual Affinity Profiling was selected, adopting the recent paradigm of polypharmacology, i.e., the observation that drugs generally act on multiple proteins. Our starting hypothesis was that the *in silico* generated interaction profile of a drug, i.e., a series of calculated binding free energy values for a set of proteins, correlates with the bioactivity properties of the drug. We also assumed that no target proteins are needed to obtain a high level of correlation between the interaction profiles and the effect profiles since the interactions of a drug with a structurally diverse protein set mimics the possible interaction pattern with the human proteome.

To test our hypothesis, structural and effect information on 1,255 FDA-approves small-molecule drugs were collected and *in silico* interaction profiles were generated for the whole drug set against 154 proteins. The correlation between the resulting interaction profile database and the effect profile database of the drugs were statistically examined and a clear association was revealed, giving an opportunity to validate our system and to perform effect predictions. In order to check the diversity of the applied protein set, we also investigated the relationship between the virtual drug screening results and the shape of the protein binding sites and revealed that binding site geometry has a minor role in the description of affinity profiles in general. We also proved that relevant effect predictions can be performed regardless of the use of known target proteins.

Based on the quantitative correlations between the interaction profiles and effect profiles, hidden effects can be revealed for the existing drugs and their entire effect profiles can be predicted as presented in my thesis. The accuracy and the robustness of the effect prediction method, called Drug Profile Matching, were evaluated by successful *in vitro* analyses. The good predictive power of our approach gives an opportunity to its use with marketed drugs or as a preclinical screen, increasing the efficacy of drug development.

I also present another problem with a high level of complexity, i.e., the role of protein flexibility in a specific conformational rearrangement. Based on our findings, we deduce that flexibility can be quantitatively modified by introducing point mutations in a single dedicated site of the protein.

Összefoglalás (Summary in Hungarian)

Doktori munkám fő célkitűzése egy olyan *in silico* rendszer felépítése volt, mely képes kismolekulás vegyületek bioaktivitási profiljának előrejelzésére a polifarmakológiai paradigma alkalmazásával, tehát arra a megfigyelésre támaszkodva, hogy a gyógyszerek többsége nem egy, hanem több célfehérjére hat. Kiindulási hipotézisünk az volt, hogy egy *in silico* előállított interakciós mintázat szoros összefüggést mutat a gyógyszer bioaktivitási mintázatával. (Az interakciós mintázat egy kötési energiaértékekből álló egydimenziós vektor, mely a gyógyszer kölcsönhatási erősségét jelzi egy sor fehérjéhez.) Feltételeztük továbbá, hogy a célfehérjéknek nem kell szerepelniük a mintázat előállításához használt fehérjekészletben, mert egy szerkezetileg diverz készlet képes mimikálni azt a kölcsönhatási profilt, amit a gyógyszer az emberi proteommal alakít ki.

Feltevéseink igazolására 1255 jelenleg használt gyógyszer szerkezeti és bioaktivitási adatait összegeztük, és elkészítettük az *in silico* interakciós mintázatukat 154 fehérjével szemben. Szignifikáns korrelációt találtunk az előállított adatbázisok, tehát az interakciós és a hatásadatbázis között, amely lehetőséget adott a módszer validálására és új hatások predikciójára. Ellenőriztük a használt fehérjekészlet diverzitását, és megvizsgáltuk a zsebgeometria hatását a kötési mintázatokra. Kimutattuk, hogy az ismert célfehérjék hiánya nem befolyásolja a módszer predikciós erősségét.

A feltárt kvantitatív összefüggések felhasználásával az ismert gyógyszerek eddig rejtett hatásai azonosíthatók. Ezáltal lehetőség nyílik a szerek teljes hatásprofiljának előrejelzésére, amint azt a dolgozatomban részletesen bemutatom. A predikciós módszer pontosságát és robosztusságát számos sikeres *in vitro* teszttel igazoltuk. Mindezek figyelembevételével megállapítottuk, hogy megközelítésünk a gyógyszerfejlesztés számos pontján hozhat jelentős előrelépést: nem csak a forgalomban lévő szerek új hatásainak és mellékhatásainak felderítésében, de gyógyszerjelölt molekulák nagy áteresztőképességű preklinikai vizsgálatában is.

Dolgozatomban bemutatok egy másik komplex problémát is, a fehérje flexibilitásának szerepét egy specifikus konformációs átmenetben. Tripszin modellrendszerünkben végzett vizsgálataink alapján arra a következtetésre jutottunk, hogy a protein flexibilitása kvantitatívan módosítható specifikusan elhelyezett pontmutációk segítségével.

References

1. Gráf L: **The structural basis of serine protease action: the fourth dimension**. In: *Natural Sciences and Human Thought.* Edited by Zwilling R. Heidelberg: Springer-Verlag; 1995: 139-148.
2. Frauenfelder H, Chen G, Berendzen J, Fenimore PW, Jansson H, McMahon BH, Stroe IR, Swenson J, Young RD: **A unified model of protein dynamics**. *Proc Natl Acad Sci U S A* 2009, **106**(13):5129-5134.
3. Testa B, Kramer SD: **The biochemistry of drug metabolism--an introduction: part 1. Principles and overview**. *Chem Biodivers* 2006, **3**(10):1053-1101.
4. Ashburn TT, Thor KB: **Drug repositioning: identifying and developing new uses for existing drugs**. *Nat Rev Drug Discov* 2004, **3**(8):673-683.
5. Lindsay MA: **Finding new drug targets in the 21st century**. *Drug Discov Today* 2005, **10**(23-24):1683-1687.
6. Temple R: **Current definitions of phases of investigation and the role of the FDA in the conduct of clinical trials**. *Am Heart J* 2000, **139**(4):S133-135.
7. Pujol A, Mosca R, Farres J, Aloy P: **Unveiling the role of network and systems biology in drug discovery**. *Trends Pharmacol Sci* 2010, **31**(3):115-123.
8. Hopkins AL: **Network pharmacology**. *Nat Biotechnol* 2007, **25**(10):1110-1111.
9. Hopkins AL: **Network pharmacology: the next paradigm in drug discovery**. *Nat Chem Biol* 2008, **4**(11):682-690.
10. Hopkins AL, Mason JS, Overington JP: **Can we rationally design promiscuous drugs?** *Curr Opin Struct Biol* 2006, **16**(1):127-136.
11. Roth BL, Sheffler DJ, Kroeze WK: **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia**. *Nat Rev Drug Discov* 2004, **3**(4):353-359.
12. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity**. *Science* 2008, **321**(5886):263-266.
13. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: **Biospectra analysis: model proteome characterizations for linking molecular structure and biological response**. *J Med Chem* 2005, **48**(22):6918-6925.

14. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry**. *Nat Biotechnol* 2007, **25**(2):197-206.

15. Kola I, Landis J: **Can the pharmaceutical industry reduce attrition rates?** *Nat Rev Drug Discov* 2004, **3**(8):711-715.

16. Kapetanovic IM: **Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach**. *Chem Biol Interact* 2008, **171**(2):165-176.

17. **Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products**. In.: Food and Drug Administration; 2004.

18. Ekman P: **Finasteride in the treatment of benign prostatic hypertrophy: an update. New indications for finasteride therapy**. *Scand J Urol Nephrol Suppl* 1999, **203**:15-20.

19. Cleach LL, Bocquet H, Roujeau JC: **Reactions and interactions of some commonly used systemic drugs in dermatology**. *Dermatol Clin* 1998, **16**(2):421-429.

20. **Painful lessons**. *Nat Rev Drug Discov* 2005, **4**(10):800-803.

21. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling**. *Br J Pharmacol* 2007, **152**(1):9-20.

22. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: applications to targets and beyond**. *Br J Pharmacol* 2007, **152**(1):21-37.

23. Clark DE: **In silico prediction of blood-brain barrier permeation**. *Drug Discov Today* 2003, **8**(20):927-933.

24. Fu XC, Wang GP, Shan HL, Liang WQ, Gao JQ: **Predicting blood-brain barrier penetration from molecular weight and number of polar atoms**. *Eur J Pharm Biopharm* 2008, **70**(2):462-466.

25. Langowski J, Long A: **Computer systems for the prediction of xenobiotic metabolism**. *Adv Drug Deliv Rev* 2002, **54**(3):407-415.

26. Lahana R: **How many leads from HTS?** *Drug Discov Today* 1999, **4**(10):447-448.

27. Johnson M, Maggiora, GM: **Concepts and Applications Molecular Similarity**. New York: John Wiley & Sons; 2006.

28. Csizmadia F: **JChem: Java applets and modules supporting chemical database handling from web browsers**. *J Chem Inf Comput Sci* 2000, **40**(2):323-324.

29. Wermuth C, Ganellin CR, Lindberg P, Mitscher L: **Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998)**. *Pure Appl Chem* 1998, **70**:1129-1143.

30. Kurogi Y, Guner OF: **Pharmacophore modeling and three-dimensional database searching for drug design using catalyst**. *Curr Med Chem* 2001, **8**(9):1035-1055.

31. Nicklaus MC, Neamati N, Hong H, Mazumder A, Sunder S, Chen J, Milne GW, Pommier Y: **HIV-1 integrase pharmacophore: discovery of inhibitors through three-dimensional database searching**. *J Med Chem* 1997, **40**(6):920-929.

32. Cavasotto CN, Phatak SS: **Homology modeling in drug discovery: current trends and applications**. *Drug Discov Today* 2009, **14**(13-14):676-683.

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

34. Mestres J: **Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery**. *Drug Discov Today* 2005, **10**(23-24):1629-1637.

35. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC *et al*: **GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function**. *Science* 2007, **318**(5854):1266-1273.

36. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC: **The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist**. *Science* 2008, **322**(5905):1211-1217.

37. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF: **Structure of a beta1-adrenergic G-protein-coupled receptor**. *Nature* 2008, **454**(7203):486-491.

38. Huey R, Morris GM, Olson AJ, Goodsell DS: **A semiempirical free energy force field with charge-based desolvation**. *J Comput Chem* 2007, **28**(6):1145-1152.

39. Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP: **eHiTS: an innovative approach to the docking and scoring function problems**. *Curr Protein Pept Sci* 2006, **7**(5):421-435.

40. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP: **eHiTS: a new fast, exhaustive flexible ligand docking system**. *J Mol Graph Model* 2007, **26**(1):198-212.

41. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK *et al*: **Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy**. *J Med Chem* 2004, **47**(7):1739-1749.

42. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL: **Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening**. *J Med Chem* 2004, **47**(7):1750-1759.

43. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R: **Novel procedure for modeling ligand/receptor induced fit effects**. *J Med Chem* 2006, **49**(2):534-553.

44. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JW, Taylor RD, Taylor R: **Modeling water molecules in protein-ligand docking using GOLD**. *J Med Chem* 2005, **48**(20):6504-6515.

45. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD: **Improved protein-ligand docking using GOLD**. *Proteins* 2003, **52**(4):609-623.

46. Dominguez C, Boelens R, Bonvin AM: **HADDOCK: a protein-protein docking approach based on biochemical or biophysical information**. *J Am Chem Soc* 2003, **125**(7):1731-1737.

47. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ: **PatchDock and SymmDock: servers for rigid and symmetric docking**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W363-367.

48. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR: **Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go**. *Br J Pharmacol* 2008, **153 Suppl 1**:S7-26.

49. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP: **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* 1997, **11**(5):425-445.

50. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT: **Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes**. *J Med Chem* 2006, **49**(21):6177-6196.

51. Wang R, Lai L, Wang S: **Further development and validation of empirical scoring functions for structure-based binding affinity prediction**. *J Comput Aided Mol Des* 2002, **16**(1):11-26.

52. G.M. Morris DSG, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson: **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. *J Comp Chem* 1999, **19**(14):1639-1662.

53. Jones G, Willett P, Glen RC, Leach AR, Taylor R: **Development and validation of a genetic algorithm for flexible docking**. *J Mol Biol* 1997, **267**(3):727-748.

54. Velec HF, Gohlke H, Klebe G: **DrugScore(CSD)-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* 2005, **48**(20):6296-6303.

55. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R: **Comparing protein-ligand docking programs is difficult**. *Proteins* 2005, **60**(3):325-332.

56. Congreve M, Murray CW, Blundell TL: **Structural biology and drug discovery**. *Drug Discov Today* 2005, **10**(13):895-907.

57. Kovacs M, Toth J, Hetenyi C, Malnasi-Csizmadia A, Sellers JR: **Mechanism of blebbistatin inhibition of myosin II**. *J Biol Chem* 2004, **279**(34):35557-35563.

58. Hetenyi C, van der Spoel D: **Efficient docking of peptides to proteins without prior knowledge of the binding site**. *Protein Sci* 2002, **11**(7):1729-1737.

59. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V: **PASS: prediction of activity spectra for biologically active substances**. *Bioinformatics* 2000, **16**(8):747-748.

60. Poroikov VV, Filimonov DA, Borodina YV, Lagunin AA, Kos A: **Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds**. *J Chem Inf Comput Sci* 2000, **40**(6):1349-1355.

61. Poroikov VV, Filimonov DA, Ihlenfeldt WD, Gloriozova TA, Lagunin AA, Borodina YV, Stepanchikova AV, Nicklaus MC: **PASS biological activity spectrum predictions in the enhanced open NCI database browser**. *J Chem Inf Comput Sci* 2003, **43**(1):228-236.

62. Poulain R, Horvath D, Bonnet B, Eckhoff C, Chapelain B, Bodinier MC, Deprez B: **From hit to lead. Combining two complementary methods for focused library design. Application to mu opiate ligands**. *J Med Chem* 2001, **44**(21):3378-3390.

63. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: **Biological spectra analysis: Linking biological activity profiles to molecular structure**. *Proc Natl Acad Sci U S A* 2005, **102**(2):261-266.

64. Krejsa CM, Horvath D, Rogalski SL, Penzotti JE, Mao B, Barbosa F, Migeon JC: **Predicting ADME properties and side effects: the BioPrint approach**. *Curr Opin Drug Discov Devel* 2003, **6**(4):470-480.

65. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: **Analysis of drug-induced effect patterns to link structure and side effects of medicines**. *Nat Chem Biol* 2005, **1**(7):389-397.

66. Hetenyi C, Maran U, Karelson M: **A comprehensive docking study on the selectivity of binding of aromatic compounds to proteins**. *J Chem Inf Comput Sci* 2003, **43**(5):1576-1583.

67. Brewerton SC: **The use of protein-ligand interaction fingerprints in docking**. *Curr Opin Drug Discov Devel* 2008, **11**(3):356-364.

68. Deng Z, Chuaqui C, Singh J: **Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions**. *J Med Chem* 2004, **47**(2):337-344.

69. Toledo-Sherman LM, Chen D: **High-throughput virtual screening for drug discovery in parallel**. *Curr Opin Drug Discov Devel* 2002, **5**(3):414-421.

70. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J *et al*: **TarFisDock: a web server for identifying drug targets with docking approach**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W219-224.

71. Ehrlich P: **The theory and practice of chemotherapy**. *Folia Serologica* 1911, **7**:697-714.

72. Krueger KE: **Peripheral-type benzodiazepine receptors: a second site of action for benzodiazepines**. *Neuropsychopharmacology* 1991, **4**(4):237-244.

73. Ebert B, Andersen S, Krogsgaard-Larsen P: **Ketobemidone, methadone and pethidine are non-competitive N-methyl-D-aspartate (NMDA) antagonists in the rat cortex and spinal cord**. *Neurosci Lett* 1995, **187**(3):165-168.

74. Kroeze WK, Kristiansen K, Roth BL: **Molecular biology of serotonin receptors structure and function at the molecular level**. *Curr Top Med Chem* 2002, **2**(6):507-528.

75. Bristow LJ, Kramer MS, Kulagowski J, Patel S, Ragan CI, Seabrook GR: **Schizophrenia and L-745,870, a novel dopamine D4 receptor antagonist**. *Trends Pharmacol Sci* 1997, **18**(6):186-188.

76. Truffinet P, Tamminga CA, Fabre LF, Meltzer HY, Riviere ME, Papillon-Downey C: **Placebo-controlled study of the D4/5-HT2A antagonist fananserin in the treatment of schizophrenia**. *Am J Psychiatry* 1999, **156**(3):419-425.

77. Terbach N, Williams RS: **Structure-function studies for the panacea, valproic acid**. *Biochem Soc Trans* 2009, **37**(Pt 5):1126-1132.

78. Bosetti F, Weerasinghe GR, Rosenberger TA, Rapoport SI: **Valproic acid down-regulates the conversion of arachidonic acid to eicosanoids via cyclooxygenase-1 and -2 in rat brain**. *J Neurochem* 2003, **85**(3):690-696.

79. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV: **The topology of drug-target interaction networks: implicit dependence on drug properties and target families**. *Mol Biosyst* 2009, **5**(9):1051-1057.

80. Bolognesi ML, Cavalli A, Melchiorre C: **Memoquin: a multi-target-directed ligand as an innovative therapeutic opportunity for Alzheimer's disease**. *Neurotherapeutics* 2009, **6**(1):152-162.

81. Bolognesi ML, Rosini M, Andrisano V, Bartolini M, Minarini A, Tumiatti V, Melchiorre C: **MTDL design strategy in the context of Alzheimer's disease: from lipocrine to memoquin and beyond**. *Curr Pharm Des* 2009, **15**(6):601-613.

82. Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA: **Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases**. *Nat Chem Biol* 2008, **4**(11):691-699.

83. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P: **A side effect resource to capture phenotypic effects of drugs**. *Mol Syst Biol* 2010, **6**:343.

84. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB *et al*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**(7270):175-181.

85. Morphy R, Rankovic Z: **Designed multiple ligands. An emerging drug discovery paradigm**. *J Med Chem* 2005, **48**(21):6523-6543.

86. Zimmermann GR, Lehar J, Keith CT: **Multi-target therapeutics: when the whole is greater than the sum of the parts**. *Drug Discov Today* 2007, **12**(1-2):34-42.

87. Milletti F, Vulpetti A: **Predicting polypharmacology by binding site similarity: from kinases to the protein universe**. *J Chem Inf Model* 2010, **50**(8):1418-1431.

88. Schmitt S, Kuhn D, Klebe G: **A new method to detect related function among proteins independent of sequence and fold homology**. *J Mol Biol* 2002, **323**(2):387-406.

89. Weisel M, Proschak E, Schneider G: **PocketPicker: analysis of ligand binding-sites with shape descriptors**. *Chem Cent J* 2007, **1**:7.

90. Kortagere S, Krasowski MD, Ekins S: **The importance of discerning shape in molecular pharmacology**. *Trends Pharmacol Sci* 2009, **30**(3):138-147.

91. Joseph-McCarthy D: **Computational approaches to structure-based ligand design**. *Pharmacol Ther* 1999, **84**(2):179-191.

92. Zauhar RJ, Moyna G, Tian L, Li Z, Welsh WJ: **Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design**. *J Med Chem* 2003, **46**(26):5674-5690.

93. Venkatraman V, Chakravarthy PR, Kihara D: **Application of 3D Zernike descriptors to shape-based ligand similarity searching**. *J Cheminform* 2009, **1**:19.

94. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Res* 2008, **36**(Database issue):D901-906.

95. Morris GM, Huey R, Olson AJ: **Using AutoDock for ligand-receptor docking**. *Curr Protoc Bioinformatics* 2008, **Chapter 8**:Unit 8 14.

96. Jiang X, Kumar K, Hu X, Wallqvist A, Reifman J: **DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0**. *Chem Cent J* 2008, **2**:18.

97. Morris G, Goodsell D, Halliday R, Huey R, Hart W, Belew R, Olson A: **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. *J Comp Chem* 1999, **19**(14):1639-1662.

98. Simon Z, Vigh-Smeller M, Peragovics A, Csukly G, Zahoranszky-Kohalmi G, Rauscher AA, Jelinek B, Hari P, Bitter I, Malnasi-Csizmadia A, Czobor P: **Relating the shape of protein binding sites to binding affinity profiles: is there an association?** *BMC Struct Biol* 2010, **10**:32.

99. Qian G, Sural, S, Gu, Y, Pramanik, S: **Similarity between Euclidean and cosine angle distance for nearest neighbor queries**. *Journal of the American Society for Information Science* 1999, **60**(9):772-778.

100. Guttman L: **Some necessary conditions for common factor analysis**. *Psychometrika* 1954, **19**:149-161.

101. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility**. *J Comput Chem* 2009, **30**(16):2785-2791.

102. Dessailly BH, Lensink MF, Orengo CA, Wodak SJ: **LigASite--a database of biologically relevant binding sites in proteins with known apo-structures**. *Nucleic Acids Res* 2008, **36**(Database issue):D667-673.

103. Favia AD, Nobeli I, Glaser F, Thornton JM: **Molecular docking for substrate identification: the short-chain dehydrogenases/reductases**. *J Mol Biol* 2008, **375**(3):855-874.

104. Schalon C, Surgand JS, Kellenberger E, Rognan D: **A simple and fuzzy method to align and compare druggable ligand-binding sites**. *Proteins* 2008, **71**(4):1755-1778.

105. Krueger BA, Weil T, Schneider G: **Comparative virtual screening and novelty detection for NMDA-GlycineB antagonists**. *J Comput Aided Mol Des* 2009, **23**(12):869-881.

106. McInnes C: **Virtual screening strategies in drug discovery**. *Curr Opin Chem Biol* 2007, **11**(5):494-502.

107. Glassman AH: **Schizophrenia, antipsychotic drugs, and cardiovascular disease**. *J Clin Psychiatry* 2005, **66 Suppl 6**:5-10.

108. Prakash C, Kamel A, Cui D, Whalen RD, Miceli JJ, Tweedie D: **Identification of the major human liver cytochrome P450 isoform(s) responsible for the formation of the primary metabolites of ziprasidone and prediction of possible drug interactions**. *Br J Clin Pharmacol* 2000, **49 Suppl 1**:35S-42S.

109. Wong DT, Bymaster FP, Engleman EA: **Prozac (fluoxetine, Lilly 110140), the first selective serotonin uptake inhibitor and an antidepressant drug: twenty years since its first publication**. *Life Sci* 1995, **57**(5):411-441.

110. Vane JR: **Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs**. *Nat New Biol* 1971, **231**(25):232-235.

111. Ferreira SH: **Angiotensin converting enzyme: history and relevance**. *Semin Perinatol* 2000, **24**(1):7-10.

112. Hackam DG, Khan NA, Hemmelgarn BR, Rabkin SW, Touyz RM, Campbell NR, Padwal R, Campbell TS, Lindsay MP, Hill MD *et al*: **The 2010 Canadian Hypertension Education Program recommendations for the management of hypertension: part 2 - therapy**. *Can J Cardiol* 2010, **26**(5):249-258.

113. **DrugBank sheet of Candoxatril**. In.

114. Cohen Solal A, Jondeau G, Beauvais F, Berdeaux A: **Beneficial effects of carvedilol on angiotensin-converting enzyme activity and renin plasma levels in patients with chronic heart failure**. *Eur J Heart Fail* 2004, **6**(4):463-466.

115. Flordellis CS, Goumenos D, Kourounis G, Tsementzis SA, Paris H, Vlachojiannis J: **The shift in the "paradigm" of the pharmacology of hypertension**. *Curr Top Med Chem* 2004, **4**(4):487-498.

116. Margulies KB, Perrella MA, McKinley LJ, Burnett JC, Jr.: **Angiotensin inhibition potentiates the renal responses to neutral endopeptidase inhibition in dogs with congestive heart failure**. *J Clin Invest* 1991, **88**(5):1636-1642.

117. Borvendég J, Polák G, Váradi A: **Hatóanyagok, készítmények, terápia. Fókuszban a keringési rendszer**. Budapest: Melinda; 2004.

118. **DrugBank sheet of Ciclopirox**. In.

119. Jannesson L, Birkhed D, Scherl D, Gaffar A, Renvert S: **Effect of oxybenzone on PGE2-production in vitro and on plaque and gingivitis in vivo**. *J Clin Periodontol* 2004, **31**(2):91-94.

120. Ren J, Chung SH: **Anti-inflammatory effect of alpha-linolenic acid and its mode of action through the inhibition of nitric oxide production and inducible nitric oxide synthase gene expression via NF-kappaB and mitogen-activated protein kinase pathways**. *J Agric Food Chem* 2007, **55**(13):5073-5080.

121. Ren J, Han EJ, Chung SH: **In vivo and in vitro anti-inflammatory activities of alpha-linolenic acid isolated from Actinidia polygama fruits**. *Arch Pharm Res* 2007, **30**(6):708-714.

122. Naveh N, Weissman C, Dottan SA: **Azathioprine's inhibitory effect on prostaglandin E2 production is not via cyclooxygenase inhibition**. *Biochem Biophys Res Commun* 1988, **157**(2):727-732.

123. Sharis PJ, Cannon CP, Loscalzo J: **The antiplatelet effects of ticlopidine and clopidogrel**. *Ann Intern Med* 1998, **129**(5):394-405.

124. Birktoft JJ, Kraut J, Freer ST: **A detailed structural comparison between the charge relay system in chymotrypsinogen and in alpha-chymotrypsin**. *Biochemistry* 1976, **15**(20):4481-4485.

125. Huber R: **Structural basis of the activation and action of trypsin 14**. *Acc Chem Res* 1978(11):114-122.

126. Kossiakoff AA, Chambers JL, Kay LM, Stroud RM: **Structure of bovine trypsinogen at 1.9 A resolution**. *Biochemistry* 1977, **16**(4):654-664.

127. Fersht AR, Renard M: **pH dependence of chymotrypsin catalysis. Appendix: substrate binding to dimeric alpha-chymotrypsin studied by x-ray diffraction and the equilibrium method**. *Biochemistry* 1974, **13**(7):1416-1426.

128. Heremans L, Heremans K: **Raman spectroscopic study of the changes in secondary structure of chymotrypsin: effect of pH and pressure on the salt bridge**. *Biochim Biophys Acta* 1989, **999**(2):192-197.

129. Stoesz JD, Lumry RW: **Refolding transition of alpha-chymotrypsin: pH and salt dependence**. *Biochemistry* 1978, **17**(18):3693-3699.

130. Verheyden G, Matrai J, Volckaert G, Engelborghs Y: **A fluorescence stopped-flow kinetic study of the conformational activation of alpha-chymotrypsin and several mutants**. *Protein Sci* 2004, **13**(9):2533-2540.

131. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling**. *Electrophoresis* 1997, **18**(15):2714-2723.

132. Brunger AT, Huber R, Karplus M: **Trypsinogen-trypsin transition: a molecular dynamics study of induced conformational change in the activation domain**. *Biochemistry* 1987, **26**(16):5153-5162.

133. Matrai J, Verheyden G, Kruger P, Engelborghs Y: **Simulation of the activation of alpha-chymotrypsin: analysis of the pathway and role of the propeptide**. *Protein Sci* 2004, **13**(12):3139-3150.

134. Toth J, Gombos L, Simon Z, Medveczky P, Szilagyi L, Graf L, Malnasi-Csizmadia A: **Thermodynamic analysis reveals structural rearrangement during the acylation step in human trypsin 4 on 4-methylumbelliferyl 4-guanidinobenzoate substrate analogue**. *J Biol Chem* 2006, **281**(18):12596-12602.

135. Kintses B, Simon Z, Gyimesi M, Toth J, Jelinek B, Niedetzky C, Kovacs M, Malnasi-Csizmadia A: **Enzyme kinetics above denaturation temperature: a temperature-jump/stopped-flow apparatus**. *Biophys J* 2006, **91**(12):4605-4610.

136. Toth J, Simon Z, Medveczky P, Gombos L, Jelinek B, Szilagyi L, Graf L, Malnasi-Csizmadia A: **Site directed mutagenesis at position 193 of human trypsin 4 alters the rate of conformational change during activation: role of local internal viscosity in protein dynamics**. *Proteins* 2007, **67**(4):1119-1127.

137. Kovacs M, Malnasi-Csizmadia A, Woolley RJ, Bagshaw CR: **Analysis of nucleotide binding to Dictyostelium myosin II motor domains containing a single tryptophan near the active site**. *J Biol Chem* 2002, **277**(32):28459-28467.

138. Peterman BF: **Measurement of the dead time of a fluorescence stopped-flow instrument**. *Anal Biochem* 1979, **93**(2):442-444.

139. Katona G, Berglund GI, Hajdu J, Graf L, Szilagyi L: **Crystal structure reveals basis for the inhibitor resistance of human brain trypsin**. *J Mol Biol* 2002, **315**(5):1209-1218.

140. Weast R: **CRC Handbook of Chemistry and Physics**, 69 edn. Boca Raton: CRC; 1988.

141. Arrhenius S: **Über die Reaktiongeschwindigkeit bei der Inversion von Rohzucker durch Sauren**. *Z Phys Chem* 1889(4):226-248.

142. Kramers H: **Brownian motion in a field of force and the diffusion model of chemical reactions**. *Physica* 1940(7):284-304.

143. Ansari A, Jones CM, Henry ER, Hofrichter J, Eaton WA: **The role of solvent viscosity in the dynamics of protein conformational changes**. *Science* 1992, **256**(5065):1796-1798.

144. Goldmann WH, Geeves MA: **A "slow" temperature jump apparatus built from a stopped-flow machine**. *Anal Biochem* 1991, **192**(1):55-58.

145. Verkman AS, Dix JA, Pandiscio AA: **A simple stopped-flow temperature-jump apparatus**. *Anal Biochem* 1981, **117**(1):164-169.

146. Hedstrom L, Lin TY, Fast W: **Hydrophobic interactions control zymogen activation in the trypsin family of serine proteases**. *Biochemistry* 1996, **35**(14):4515-4523.

147. Bobofchak KM, Pineda AO, Mathews FS, Di Cera E: **Energetic and structural consequences of perturbing Gly-193 in the oxyanion hole of serine proteases**. *J Biol Chem* 2005, **280**(27):25644-25650.

148. Zivelin A, Ogawa T, Bulvik S, Landau M, Toomey JR, Lane J, Seligsohn U, Gailani D: **Severe factor XI deficiency caused by a Gly555 to Glu mutation (factor XI-Glu555): a cross-reactive material positive variant defective in factor IX activation**. *J Thromb Haemost* 2004, **2**(10):1782-1789.

149. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG: **How enzymes work: analysis by modern rate theory and computer simulations**. *Science* 2004, **303**(5655):186-195.

150. Frauenfelder H, Wolynes PG: **Rate theories and puzzles of hemeprotein kinetics**. *Science* 1985, **229**(4711):337-345.

151. Beece D, Eisenstein L, Frauenfelder H, Good D, Marden MC, Reinisch L, Reynolds AH, Sorensen LB, Yue KT: **Solvent viscosity and protein dynamics**. *Biochemistry* 1980, **19**(23):5147-5157.

152. Fliri AF, Loging WT, Volkmann RA: **Analysis of System Structure-Function Relationships**. *ChemMedChem* 2007, **2**(12):1774-1782.

153. Alqahtani S, Federico P, Myers RP: **A case of valproate-induced hyperammonemic encephalopathy: look beyond the liver**. *Cmaj* 2007, **177**(6):568-569.

154. Belay ED, Bresee JS, Holman RC, Khan AS, Shahriari A, Schonberger LB: **Reye's syndrome in the United States from 1981 through 1997**. *N Engl J Med* 1999, **340**(18):1377-1382.

155. Sandler RH, Bolte ER, Chez MG, Schrift MJ: **Relief of psychiatric symptoms in a patient with Crohn's disease after metronidazole therapy**. *Clin Infect Dis* 2000, **30**(1):213-214.

156. Pessayre D, Mansouri A, Haouzi D, Fromenty B: **Hepatotoxicity due to mitochondrial dysfunction**. *Cell Biol Toxicol* 1999, **15**(6):367-373.

157. Lee SY, Lee ST, Kim JW: **Contributions of CYP2C9/CYP2C19 genotypes and drug interaction to the phenytoin treatment in the Korean epileptic patients in the clinical setting**. *J Biochem Mol Biol* 2007, **40**(3):448-452.

158. Miners JO, Birkett DJ: **Cytochrome P4502C9: an enzyme of major importance in human drug metabolism**. *Br J Clin Pharmacol* 1998, **45**(6):525-538.

159. Leniger T, Pageler L, Stude P, Diener HC, Limmroth V: **Comparison of intravenous valproate with intravenous lysine-acetylsalicylic acid in acute migraine attacks**. *Headache* 2005, **45**(1):42-46.

160. Srivastava AK, Gupta YK: **Aspirin modulates the anticonvulsant effect of diazepam and sodium valproate in pentylenetetrazole and maximal electroshock induced seizures in mice**. *Indian J Physiol Pharmacol* 2001, **45**(4):475-480.

161. Esiobu N, Hoosein N: **An assessment of the in vitro antimicrobial effects of two antiepileptic drugs--sodium valproate and phenytoin**. *Antonie Van Leeuwenhoek* 2003, **83**(1):63-68.

162. Wang WH, Wong WM, Dailidiene D, Berg DE, Gu Q, Lai KC, Lam SK, Wong BC: **Aspirin inhibits the growth of Helicobacter pylori and enhances its susceptibility to antimicrobial agents**. *Gut* 2003, **52**(4):490-495.