Optimization of Clustering and Database Screening Procedures for Cavbase and Virtual Screening for Novel Antimalarial and Antibacterial Molecules

Dissertation
zur
Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem
Fachbereich Pharmazie
der Philipps-Universität Marburg
vorgelegt von

Serghei Glinca
aus Chișinău/Republik Moldau

Marburg/Lahn 2012

Vom Fachbereich Pharmazie der Philipps-Universität Marburg als Dissertation am 30.10.2012 angenommen.

Erstgutachter	Prof. Dr. Gerhard Klebe, Institut für Pharmazeutische Chemie, Philipps-Universität Marburg
Zweitgutachter	Prof. Dr. Martin Schlitzer, Institut für Pharmazeutische Chemie, Philipps-Universität Marburg

Tag der mündlichen Prüfung: 31.10.2012

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. Gerhard Klebe am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von Juli 2009 bis Oktober 2012 durchgeführt.

Für Sema Nur und Hilal

Contents

Abbreviations xi

List of Figures xv

List of Tables xvii

List of Publications xviii

Kurzfassung xxi

Summary xxv

I. Optimization of Clustering and Database Screening Procedures for Cavbase 1

1. Exploring Functional Relationships of Proteases via Binding Pockets 3

- 1.1. Introduction 3
- 1.2. Materials and methods 8
  - 1.2.1. Data sets selection and benchmark 8
  - 1.2.2. Data sets 9
  - 1.2.3. Similarity matrices 13
  - 1.2.4. Clustering workflow 13
- 1.3. Results and discussion 16
  - 1.3.1. Clustering workflow validation 16

1.3.2. Detailed analysis of the proteases clustering 19
1.3.3. Ligand, cavity, and sequence data: Cluster analysis of serine proteases 26
1.4. Conclusions 32
1.5. Supporting information 33
1.5.1. Cluster validation statistics - Internal criteria . . . 33
1.5.2. Cluster validation statistics - External criteria . . . 36

2. Automated Decomposition of Cavities Based on Bound Ligand Information 37
2.1. Preliminary remarks 37
2.2. Introduction 37
2.3. Automated binding site decomposition procedure 39
2.3.1. Ligand fragmentation methods 39
2.3.2. Cavity decomposition workflow 40
2.4. Case studies 42
2.4.1. Pairwise comparison 42
2.4.2. Database screening 47
2.5. Discussion 50

II. Virtual Screening for Novel Molecules with Antimalarial and Antibacterial Activity 53

3. Novel FAS II Inhibitors as Multistage Antimalarials 55
3.1. Preliminary remarks 55
3.2. Introduction 56
3.3. Results and discussion 60
3.3.1. Virtual screening 60
3.3.2. Biological evaluation and structure–activity relationship 64
3.3.3. Combinatorial library 67

3.3.4. Evaluation of the effect on pre-erythrocytic parasites 68
3.4. Conclusion 75
3.5. Materials and methods 76
3.5.1. Virtual screening and docking 76

4. Virtual Screening for Pyruvoyltetrahydropterin Synthase Scaffolds 79
4.1. Introduction 79
4.1.1. Folate metabolism in *P. falciparum* 79
4.1.2. Antifolates 80
4.1.3. Pyruvoyltetrahydropterin synthase orthologue of *P. falciparum* 82
4.1.4. Binding site of *Pf*PTPS 84
4.2. Targeted Library Design 85
4.3. Docking Setup 87
4.4. Results and Discussion 88

5. Virtual Screening for IspD Scaffolds 93
5.1. Inroductory remarks 93
5.2. Motivation 93
5.3. IspD as drug target 95
5.4. Active site of IspD 96
5.5. Docking setup and results 98
5.6. Experimental results 101

III. Appendix 103

Glossary 105

Bibliography 107

Acknowledgements 123

Declaration **127**

Curriculum vitae **129**

Abbreviations

α-CAs	α-carbonic anhydrases.
α-CA II	α-carbonic anhydrase II.
α-CA V	α-carbonic anhydrase V.
α-CA XIII	α-carbonic anhydrase XIII.
Å	Ångström (1 Å= 10^{-10} m= 100 pm).
"RO3"	"Rule of Three".
ACP	acyl carrier protein.
ARI	Adjusted Rand Index.
AS	Average Silhouettes.
ASP	Astex Statistical Potential.
BH_4	tetrahydrobiopterin.
CDP-ME	4-diphosphocytidyl-2-C-methylerythritol.
clog$P$	Calculated log of the octanol/water partition coefficient.
COX-2	cyclooxygenase-2.
CTP	cytidine triphosphate.
DAIM	**D**ecomposition **A**nd **I**dentification of **M**olecules.
DHF	7,8-dihydrofolate.
DHFR	dihydrofolate reductase.

DHN-PPP	7,8-dihydroneopterin triphosphate.
DHNA	dihydroneopterin aldolase.
DHPS	dihydropteroate synthase.
DHPT	7,8-dihydro-6-hydroxymethylpterin.
DMAPP	dimethylallyl diphosphate.
DUBs	deubiquitinating enzymes.
DXP/MEP	1-deoxy-D-xylulose 5-phosphate/2-C-methyl-D-erythritol 4-phosphate.
EC	Enzyme Commission.
ELO	fatty acid elongases.
EnCR	enoyl-CoA reductase.
ENR	enoyl ACP reductase.
FA	fatty acids.
FAS	Fatty acid biosynthesis.
FAS I	type I fatty acid biosynthesis.
FAS II	type II fatty acid biosynthesis.
fVIIa	Coagulation factor VIIa.
fXa	Coagulation factor Xa.
GA	Genetic Algorithm.
HBA	hydrogen-bond acceptor.
HBD	hydrogen-bond donor.
hBH	human bleomycin hydrolase.
HIV	human immunodeficiency virus.
HMG-CoA reductase	3-hydroxy-3-methyl-glutaryl-CoA reductase.
HSP70	70 kDa heat shock protein.
HSP90	90 kDa heat shock protein.
HT29	human colon carcinoma cell line 29.

IC_{50}	half-maximal inhibitory concentration.
IPP	isopentenyl diphosphate.
IspC	1-desoxy-D-xylulose-5-phosphate reductoisomerase.
IspD	4-diphosphocytidyl-2-C-methylerythritol synthetase.
LFA-1	lymphocyte function-associated antigen-1.
MCS	maximum common subgraph.
ME	methylerythritol.
MEP	2-C-methyl-D-erythritol 4-phosphate.
MSS	Median Split Silhouettes.
MW	Molecular weight (Dalton).
NMR	nuclear magnetic resonance.
PDB	Protein Data Bank.
PDK-1	3-phosphoinositide-dependent protein kinase 1.
PK	plasma kallikrein.
PPARγ	peroxisome proliferator-activated receptorγ.
PTP	6-pyruvoyltetrahydropterin.
PTPS	pyruvoyltetrahydropterin synthase.
RECAP	**Re**trosynthetic **C**ombinatorial **A**nalysis **P**rocedure.
RMSD	root-mean-square-deviation.
SBDD	structure-based drug design.
SI	selectivity index.

SMILES	**S**implified **M**olecular-**I**nput **L**ine-**E**ntry **S**ystem.
STD	saturation transfer difference.
THF	Tetrahydrofolate.
TNFα-IP3	tumor necrosis factor-α-induced protein-3.
tPA	tissue-type plasminogen activator.
TPSA	Total polar surface area (Å^2).
UCH-L1	ubiquitin carboxy-terminal hydrolase L1.
uPA	urokinase-type plasminogen activator.
VS	virtual screening.
WaterLOGSY	water ligand observation by gradient spectroscopy.
ZBG	zinc-binding group.

List of Figures

1.1. Clustering workflow 7
1.2. Correlation plot for cluster entropies 18
1.3. Cavity vs. sequence space of proteases 20
1.4. Superimposition of UCH-L1, TNFα-IP3, and otubain-2 cavities 21
1.5. Cross-reactivity between calpain-1 and cysteine cathepsins . . 25
1.6. Cluster entropy of the serine proteases data sets 27
1.7. Heatmaps of the clustered serine proteases data sets 30

2.1. Cavity decomposition workflow 41
2.2. Fragments of SC-558 43
2.3. Fragments of ritonavir 44
2.4. Fragments of lovastatin 45
2.5. PDK-1 inhibitor OSU-03012 and celecoxib 51

3.1. Type II fatty acid biosynthesis 57
3.2. *Pf*ENR binding sites 60
3.3. Docking poses of eight VS hits 62
3.4. VS hits and chemical modifications 63
3.5. *Pf*ENR and *Tg*ENR pockets 64
3.6. Docking pose for **1a** 65
3.7. Inhibition of *Tg*ENR 66
3.8. Inhibition of *Tg*ENR 69
3.9. Amide derivatives of salicylic acid 70
3.10. SAR of various aryl amide derivatives 71
3.11. Effect of **6g** and **5j** on invasion of HuH7 cells 73

3.12. Effect of **6g** and **5j** on the development of the exoerythrocytic form 74
3.13. Confocal pictures of malarial exoerythrocytic stages 75
3.14. Combinatorial library of salicylamides 77

4.1. Structure of dihydrofolate 80
4.2. Folate and biopterin biosynthetic pathways 81
4.3. DHNA and PTPS catalyzed reactions 83
4.4. *Pf*PTPS binding site with biopterin 85
4.5. Zinc-binding groups 87
4.6. Biopterin and virtual screening hits 89
4.7. Three screening hits in the *Pf*PTPS binding site 91

5.1. Inhibitors of the IspC. 95
5.2. IspD catalyzed reaction 96
5.3. IspD structure 97
5.4. IspD virtual screening hits 100
5.5. Top ranked IspD VS hit 102

List of Tables

1.1. 16 EC classes are represented in the EC data set composed by 502 binding sites. 9
1.2. Proteases data set: part 1 10
1.3. Proteases data set: part 2 11
1.4. Ligand data set of serine proteases 12
1.5. Estimated k for the EC and proteases data sets 17
1.6. Pearson correlation cluster statistics and ARI 18
1.7. Pre-filtering of the proteases data set 19
1.8. k estimation for serine proteases data sets 27
1.9. ARIs of the serine proteases data sets 28

2.1. 6COX vs. 1OQ5 43
2.2. 3PRS vs. 1RL8 45
2.3. 1CQP vs. 1HW9 46
2.4. Screening the pocket database: setup I 48
2.5. Screening the pocket database: setup II 49
2.6. Screening the pocket database: setup III 49

3.1. Fragment-like library for *Pf*ENR screening 61

4.1. Antifolates 82
4.2. *Pf*PTPS screening using GOLD 88

5.1. IspD screening using GOLD 98
5.2. Experimental binding of IspD hits 101

List of Publications

Articles

- Fober, T., Glinca, S., Klebe, G., Huellermeier, E. (2011) *Superposition and Alignment of Labeled Point Clouds*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(6): 1653-66
- Schrader F.C., Glinca S., Sattler J.M., Dahse H.-M., Afanador G.A., Prigge S.T., Lanzer M., Mueller A.-K., Klebe G., Schlitzer M. (2012) *Novel FAS II inhibitors as multistage antimalarials*, ChemMedChem, manuscript accepted
- Samsonova O., Biela A., Glinca S., Pfeiffer C., Klebe G., Kissel T. (2012) *Polymer Conformation in Aqueous Solution is Critical for DNA-Vector Formation: Isothermal Titration Calorimetry and Molecular Dynamics Disclose Causes for Variability in Transfection Performance*, manuscript submitted to Acta Biomaterialia
- Schmidt I., Tidten N., Immekus F., Glinca S., Nguyen P., Gerber H.-D., Heine A., Klebe G., Reuter K. (2012) *Investigation of Substrate Base Specificity Determinants in Bacterial tRNA-guanine Transglycosylase Reveals Queuine, the Natural Substrate of its Eucaryotic Counterpart, as Inhibitor*, manuscript submitted to PloS ONE
- Glinca S., Klebe G. (2012) *Cavities tell more than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets*, manuscript submitted to Journal of Chemical Information and Modeling

- Stock M., Fober T., Hüllermeier E., Glinca S., Klebe G., Heitzer C.B., Pahikkala T., Airola A., De Baets B., Waegeman W. *Supervised ranking for enhanced retrieval of enzyme functions*, manuscript in preparation

Posters

- Glinca S., Schrader F., Schlitzer M., Klebe G. (2010) *Characterization of P. falciparum Enoyl Acyl Carrier Protein Reductase (PfENR) for Structure Based Design and Initial Virtual Screening for Novel Inhibitors*, GDCh-Meeting “Frontiers in Medicinal Chemistry”, Münster, Germany

- Oppermann S., Elsaesser, K., Schrader, F., Glinca S., Klebe, G., Schlitzer, M., Culmsee C. (2011) *Development of Novel Bid Inhibitors for the Treatment of Neurodegenerative Diseases*, Joint Meeting of the Austrian and German Pharmaceutical Societies, Innsbruck, Austria; poster contribution

- Glinca S., Klebe G. *Bindetaschen-basierte Proteinvergleiche: Vorhersagen von unbekannten Funktionen und Arzneimittelnebenwirkungen*, CIB Partnering Konferenz, Industriepark Höchst, Frankfurt am Main, Germany

Talks

- 242nd ACS National Meeting, Emerging Technologies in Computational Chemistry, (2011) *Binding Site-Based Classification of Proteins using Clustering Techniques: Is a Similarity Score Enough?*, Denver, Colorado, USA
- 242nd ACS National Meeting, Mining Protein-Ligand Interaction Space, (2011) *Clustering the Pocket-Space of Protein Families*, Denver, Colorado, USA

Kurzfassung

Teil I: Optimierung von Cluster-Verfahren und Datenbank-Screening Methoden in Cavbase

Im Zyklus der rationellen Arzneimittelentwicklung werden Affinität und Selektivität von potentiellen Wirkstoffen intensiv erforscht. Da diese beiden Eigenschaften keine lineare Abhängigkeit zueinander aufweisen, führt höhere Affinität nicht gezwungenermaßen auch zu einer höheren Selektivität. Diese Informationen über potentielle Bindung an unerwünschte Zielproteine (Off-Targets) sind essenziell für eine erfolgreiche Arzneistoffentwicklung.

Computer-basierte Verfahren spielen eine immer größere Rolle für die Analyse und Vorhersage von Selektivitätsprofilen. Da die meisten erfolgreich eingesetzten niedermolekularen Arzneistoffe in Vertiefungen auf Proteinoberflächen binden, spielen physiko-chemische Eigenschaften von Bindetaschen eine zentrale Rolle in der Erkennung und damit auch der Bindung von Liganden. Cavbase ist eine Methode, die es ermöglicht Bindetaschen anhand der physiko-chemischen Eigenschaften dort exponierter Aminosäuren zu beschreiben und unabhängig von ihrer Proteinsequenz und Faltungsgeometrie zu vergleichen. Die Bindetaschen-basierte Klassifizierung von Proteinen ist ein effektiver Ansatz, um relevante Informationen für Selektivitätsanalysen zu extrahieren, die durch Anwendung von Clustermethoden erreicht werden kann. In der vorliegenden Arbeit wurde ein neuartiger Arbeitsablauf zur Untersuchung von wichtigen Parametern einer Clusterung entwickelt. Für einen Datensatz von Proteinen wird eine Ähnlichkeitsmatrix berechnet und anschließend dem entwickelten Arbeitsablauf übergeben. Dieser Ansatz wurde erfolgreich an zwei unterschiedlichen und anspruchsvollen Datensätzen getestet. Die vorhergesagte Anzahl der Cluster, die am besten

geeignete Clustermethode und die anschließende Clusterstruktur waren in Übereinstimmung mit den Referenzklassifikation der Proteine. Im Falle der Protease-Proteinfamilie führte die Bindetaschen-basierte Klassifizierung zur einer signifikanten Gruppierung von Proteineinträgen, die unabhängig von Sequenzinformation entstanden. Damit konnte auf struktureller Ebene die Kreuzreaktivität zwischen dem Protein Calpain-1 und Cysteincathepsinen detektiert werden, die bis jetzt nur auf Basis von Liganddaten beschrieben wurde. Im weiteren Verlauf wurden elf unterschiedliche Serinproteasen untersucht, indem die Topologie der Liganden, Bindetaschen- und Sequenzinformationen miteinander verglichen wurden. Die entstandenen Cluster zeigen einen Korrelationstrend zwischen der Ähnlichkeit im Liganden- und Bindetaschenraum. Diese Ergebnisse deuten darauf hin, dass Bindetaschenklassifizierungen wichtige Vorhersagen in Bezug auf unerwünschte Zielstrukturen geben können, die in Optimierungszyklen einer Leitstruktur berücksichtigt werden sollten.

Eine automatisierte Zerlegung von Bindetaschen in Subtaschen auf Basis von gebundenen Liganden wurde etabliert. Drei Fallbeispiele von Bindetaschen nicht verwandter Proteine, die jedoch den gleichen Liganden binden, wurden in einem Screening gegen eine Datenbank von etwa 275.000 Bindetaschen getestet. Die Subtaschen-basierte Methode führte dazu, dass die gesuchten Proteine auf höheren Rängen anzutreffen waren. Für die Subtaschen-basierte Suche der Cyclooxygenase-2 wurde das Celecoxib Analogon SC-558 verwendet. Das Ergebnis zeigt eine präzisere Ähnlichkeitsbewertung zwischen der Cyclooxygenase-2 und der Carboanhydrase-II sowie der 3-Phosphoinositid-abhägigen Proteinkinase 1. Interessanterweise wurden weitere mögliche Bindungspartner von Celecoxib vorgeschlagen. Diese Hypothese wird in Zukunft weiter untersucht.

Teil II: Virtuelles Screening nach neuen Molekülen mit antimalaria und antibakterieller Wirkung

Eine steigende Anzahl von Resistenzen auf derzeitig angewandte antiparasitäre und antibakterielle Arzneistoffe erfordert die Entwicklung neuartiger Antiinfektiva. Ein potentieller Wirkstoffkandidat sollte einen möglichst unterschiedlichen Wirkungsmechanismus aufweisen als die bereits in der Therapie verwendeten Arzneistoffe, für die vielseitige Resistenzen beschrieben wurden. Viele Bemühungen in der Forschung sind bestrebt, Stoffwechselwege aufzufinden, die mögliche Zielstrukturen für die Inhibitorentwicklung beinhalten.

Für den Parasiten *Plasmodium falciparum*, den Erreger der Malaria, wurde das Schlüsselenzym der Fettsäuresynthese Typ-2, Enoyl ACP Reduktase (ENR), als potentielle Zielstruktur vorgeschlagen. In einem virtuellen Screening einer hauseigenen virtuellen Datenbank von fragmentartigen Kleinmolekülen konnten acht vielversprechende Strukturen ausfindig gemacht werden, die von unserem Projektpartner synthetisiert und anschließend auf ihre biologische Wirkung getestet wurden. Ein Salicylsäureamidderivat zeigte in einem zellulären Assay inhibitorische Wirkung im erythrozytären Stadium der *Plasmodium* Parasiten. Diese Verbindung wurde in weiteren Schritten optimiert, in dem Struktur-Aktivitäts-Beziehungen und kombinatorisches Docking für Salicylamide analysiert wurden. Aus dieser Studie konnten zwei potente Verbindungen hervorgehen, die eine niedrige Zytotoxizität aufweisen und in einstellig mikromolarer Konzentration sowohl im erythrozytären als auch im prä-erythrozytären Stadium ihre hemmende Wirkung entfalten. Die Wirkung im prä-erythrozytären Stadium zeigte sich der Wirkung von Primaquin überlegen.

Die Biosynthese der Tetrahydrofolsäure ist ein essenzieller Stoffwechselweg für fast alle Organismen. Das Enzym Pyruvoyltetrahydropterin Synthase im *Plasmodium falciparum* (*Pf*PTPS) übernimmt in diesem Stoffwechselweg die Katalyse einer Reaktion, die gewöhnlich von Dihydroneopterin Aldolase katalysiert wird, das jedoch im *Plasmodium* Genom fehlt. Die Einbettung des Enzyms *Pf*PTPS in den Folatstoffwechsel qualifiziert es als eine potentielle Zielstruktur zur Entwicklung neuartiger Antifolate. Eine spezielle auf dieses Zielprotein hin aufgearbeitete Bibliothek weist Kleinmoleküle mit zink-bindenden funktionellen Gruppen auf. Die Durchführung eines virtuellen Screenings führte zur Auswahl von neun

Molekülen für die Synthese, die anschließend auf ihre biologische Wirkung evaluiert werden sollen.

Eine Vielzahl pathogener Mikroorganismen sind auf die Synthese der Isoprenoide aus dem Methylerithritolphosphatweg (MEP-Weg) angewiesen, daher eignet sich die Inhibition dieses Stoffwechselweges als eine sinnvolle Strategie für die Wirkstoffentwicklung, wie es für den IspC Inhibitor Fosmidomycin gezeigt wurde. IspD ist eines der Enzyme des MEP-Weges und wurde als Modellprotein zur Untersuchung der bestimmenden Faktoren für eine strukturbasierte Wirkstoffentwickung ausgewählt. Ein Datensatz von leitstrukturartigen Kleinmolekülen aus der ZINC Datenbank wurde für ein virtuelles Screening benutzt, das zur Auswahl von sieben Kandidaten führte. Sechs Verbindungen konnten kommerziell erworben und getestet werden. Für drei Verbindungen konnte eine Proteinbindung gemessen werden. Diese Ergebnisse liefern einen erfolgversprechenden Ausgangspunkt für weitere Experimente, wie Bestimmung der Bindungskonstanten und Proteinkokristallisation.

Summary

Part I: Optimization of Clustering and Database Screening Procedures for Cavbase

In rational drug design approaches two major properties of a drug candidate are exploited during the optimization cycles: affinity and selectivity. Since both properties do not correlate in a linear manner, i.e. high affinity does not necessarily lead to high selectivity, knowledge about the potential off-targets is essential for drug development.

Computational approaches play an increasingly important role for the analysis and prediction of selectivity profiles. As most of the successfully administered small molecule drugs bind in depressions on the surface of proteins, physicochemical properties of the pocket-exposed aminoacids play a central role in ligand recognition during the binding event. Cavbase is a methodology to describe binding sites in terms of the exposed physicochemical properties and to compare them independent of the sequence and fold homology. Classification of proteins by means of their binding site properties is a promising approach to achieve relevant information for selectivity modeling. For this purpose, a novel workflow has been developed to explore the important parameters of a clustering procedure, which will allow an accurate classification of proteins. For a given data set a similarity matrix can be generated and subsequently utilized as input for a clustering procedure. It has been successfully applied on two diverse and challenging data sets. The predicted number of clusters, suggested by the clustering methods, and the subsequent clustering of proteins are in agreement with considered expert classifications. In case of the human proteases data set, the binding site-based classification leads to significant

groups of proteins independent from sequence information. As a consequence, the cross-reactivity between calpain-1 and cysteine cathepsins on the structural level could be detected, which so far has only been described for the ligand data. In a benchmark study using ligand topology, binding site, and sequence information of eleven serine proteases the emerged clusters indicate a pronounced correlation between the cavity and ligand data. These results emphasize the importance of the binding site information which should be considered for ligand design during lead optimization cycles.

An automated procedure for binding site decomposition was established taking the information of the bound ligand into consideration. In the subsequent screening of three test cases against a database of about 275,000 pocket entries a more significant ranking of remotely related proteins was achieved. Using the subsites of cyclooxygenase-2 defined by fragments of the celecoxib analog SC-558 for screening, resulted in an improved ranking for known targets of celecoxib, such as carbonic anhydrase-II and 3-phosphoinositide-dependent protein kinase 1. Additional binding partners for celecoxib have been suggested and these predictions are planned to be evaluated in future.

Part II: Virtual Screening for Novel Antimalarial and Antibacterial Molecules

The increasing number of resistances to currently applied antimalarial and antibacterial drugs give rise to an urgent need for the development of new and affordable antiinfectives. A promising candidate should exhibit a mode of action differing from the available drugs for which pronounced resistance has been described. Many efforts are invested in the investigation of parasitic metabolism in order to find pathways that would provide putative targets for inhibitor design.

For the *Plasmodium* parasites, the pathogen of malaria, the key enzyme of the type II fatty acid biosynthesis, enoyl ACP reductase (ENR), has been suggested as a target. We performed a virtual screening for novel scaffolds of *Plasmodium falciparum* ENR using an in-house fragment-like virtual library and selected eight promising hits for synthesis and subsequent biological evaluation as multistage inhibitors by our project partners. A salicylamide derivative inhibited erythrocytic parasite growth in a cell-based assay and was considered for further optimization. A comprehensive analysis of the structure-activity relationships and the docking results of a combinatorial library of salicylamides resulted in two highly active structures. Both compounds comprise low cell-toxicity and display at one-digit micromolar concentrations potent inhibition of the parasitic growth in erythrocytic stage as well as superior inhibition profile compared to the gold-standard primaquine in pre-erythrocytic stages.

Biosynthesis of tetrahydrofolate is an essential pathway in almost all living organisms. Pyruvoyltetrahydropterin synthase of *Plasmodiun falciparum* (*Pf*PTPS) has been found to fill the gap in the folate biosynthetic pathway in *Plasmodium* parasites, since dihydroneopterin aldolase could not be identified in the genome. Integration of *Pf*PTPS in the folate metabolism qualifies the protein for inhibitor design, as antifolates are well-established and effective agents for prophylaxis and treatment of malaria. A focused library of compounds comprising zinc binding groups has been created and docked into the binding site of the protein. Nine virtual screening hits have been selected for synthesis and will be subjected further biological testing by our project partners.

Many pathogenic organisms rely on the synthesis of isoprenoids via the non-mevalonate pathway (DXP/MEP). Inhibition of this pathway is a promising strategy for the development of potent antiinfective agents, as it has been shown for the IspC inhibitor fosmidomycin. IspD catalyzes the third reaction step in the DXP/MEP pathway and has been selected for our study as model protein to elucidate the structural determinants for structure-based drug design. Virtual screening of the lead-like subset retrieved from the ZINC database resulted in the selection of seven promising hits. Six molecules were purchased and subsequently tested in an experimental enzyme binding assay. Two compounds showed weak and

one compound moderate binding affinity. These results deliver an adequate starting point for further experiment, such as measurement of the binding constants and co-crystallization trials.

Part I.

Optimization of Clustering and Database Screening Procedures for Cavbase

1

Chapter 1.

Exploring Functional Relationships of Proteases via Binding Pockets

1.1. Introduction

Greatly desired properties of a drug are either high affinity and selectivity toward one particular target or dependent on the mode-of-action also in special cases a promiscuous binding to a set of multiple targets might be important (Kawasaki and Freire, 2011). The latter situation can be given e.g. for kinases where a set of proteins of the kinome has to be downregulated in a disease situation. During lead optimization it is difficult to assign clear-cut criteria to the optimization strategy. In particular the information about the target and the target family have to be analyzed and considered in the optimization. General rules to follow in structure-based selectivity optimization such as shape and electrostatic complementary, flexibility, role of water, and allosteric or noncompetetive binding have been suggested (Huggins et al., 2012). Methods that make use of pocket information for selectivity analysis and prediction have gained an increasing importance over last years, yet their potential has to be utilized for routine applications (Pérot et al., 2010).

Various studies have used structural pocket similarity considerations leading to an accurate functional classification of kinases (Kuhn et al., 2007; Kinnings and Jackson, 2009; Spitzer et al., 2011). They found that on low sequence identity level binding sites can be highly conserved, on the contrary, kinases related by high sequence similarity can still expose significant differences in their pockets. Thereby experimentally observed cross-reactivities of known kinase inhibitors could be rationalized. Apart from cross-reactivity considerations regarding members of the same proteins family the prediction of potent binding to structurally and sequentially remote proteins is of utmost challenge. Therefore, the binding pocket of a given query protein can be screened against a database of pockets classified in the same way. A different picture emerges when a query binding site is screened against a pocket database. Local binding site similarities of remote proteins contributing to ligand binding can be detected (Weber et al., 2004; Milletti and Vulpetti, 2010). Bearing in mind the work of Mestres et al. (Mestres et al., 2009) that a drug interacts on average with six targets in a cell, both approaches, either the functional classification of protein families and the broad database screening for similarities in sequentially remote proteins provide indispensable information to be considered for the ligand selectivity profile.

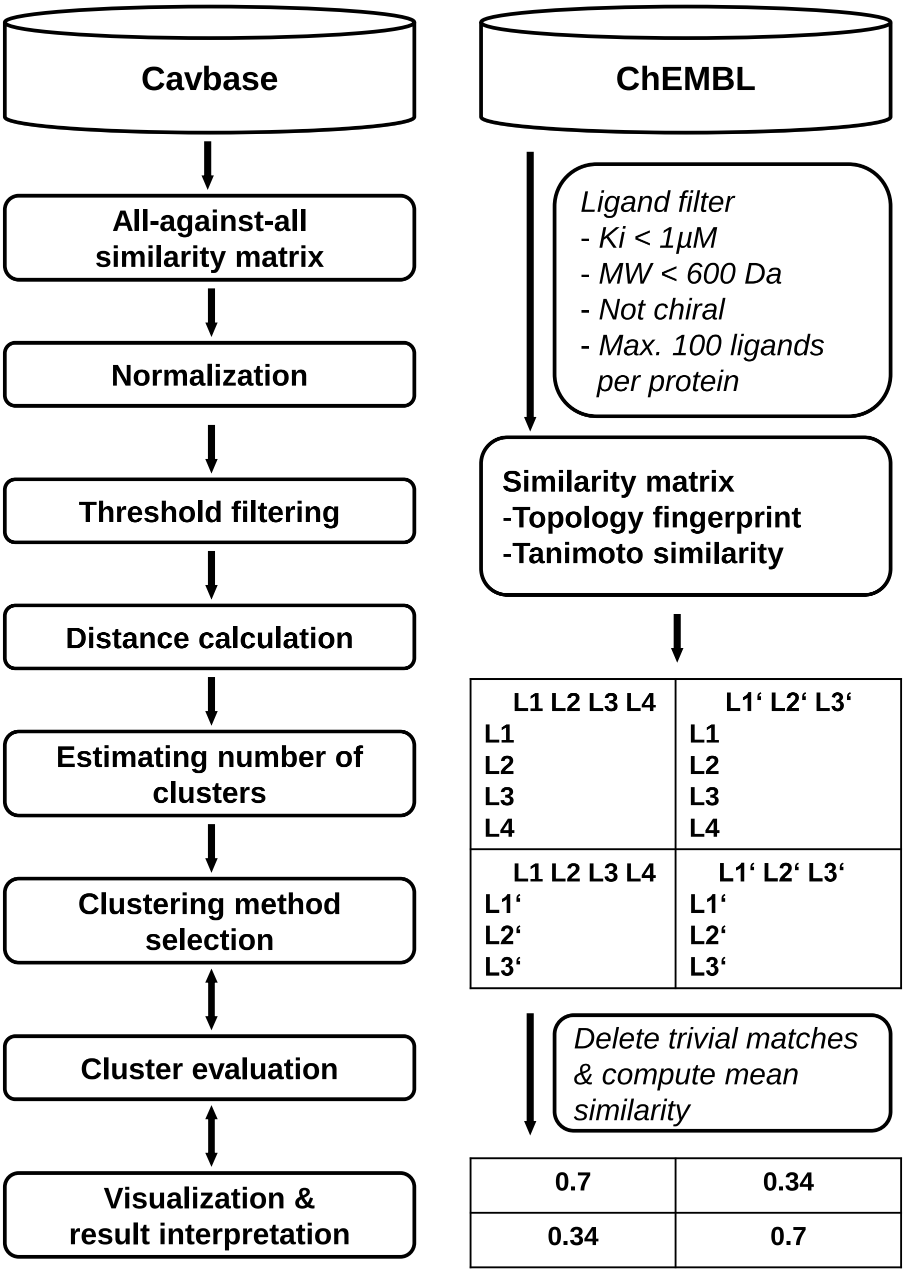
A wide range of different approaches for pocket detection and comparison have been developed up to now, which have been comprehensively reviewed elsewhere (Pérot et al., 2010). Since, our present study is based on Cavbase, a brief introduction of this methodology follows. Cavbase is able to detect, describe and compare protein binding pockets independent of their sequence and fold geometry (Schmitt et al., 2002). The pocket detection is performed using the Ligsite algorithm (Hendlich, 1997), which exploits geometric data about the protein structure only, and during this step any information about a possibly bound ligand is neglected. After pocket detection, pseudocenters are assigned to the cavity-flanking residues according to predefined rules (Schmitt et al., 2002; Kuhn et al., 2006).

The pseudocenters encode physicochemical properties that are exposed on the surface of the detected pocket. Currently, seven different types of pseudocenters are implemented in Cavbase, covering the following properties: metal, H-bond donor, acceptor, mixed donor-acceptor, hydrophobic, π (ability to form pi-pi interactions) and aromatic. The pocket comparison is computed by a clique-detection algorithm which relies only on the information stored by means of the pseudocenters. After the maximum common subgraph (MCS) is found, cavities are superimposed and a scoring function evaluates the overlap of the surfaces of the aligned pockets.

For a given data set of protein binding pockets, an all-against-all comparison can be performed, and based on the resulting similarity matrix a clustering procedure can be applied. Functional classifications based on Cavbase similarity scores have been presented for α-carbonic anhydrases (α-CAs) and kinases (Kuhn et al., 2006, 2007). In case of the α-CAs (Kuhn et al., 2006) a separation on subfamily level was achieved and conformational or mutational differences were easily detectable. Kinases could be clustered on superfamily level, different activation states in the subfamilies could be distinguished. In both studies the Cluto tool-kit (Zhao and Karypis, 2005) was applied for the clustering procedure. However, several limitations can occur when using Cluto. First, cavities that share only a marginal similarity are included and might end-up in the same cluster, which will bias the clustering. Second, Cluto provides a limited number of methods to evaluate the obtained clustering structure in order to choose the most suitable clustering strategy for the given problem. Third, Cluto requires as a prerequisite a predefined number of expected clusters, an assignment which usually appears quite arbitrary as the number of expected clusters is *a priori* not known.

In the present study we introduce a new clustering workflow which was designed and validated for clustering of data sets in terms of the Cavbase similarity metric, but the implemented routines can be applied to any similarity or distance matrix. The proposed procedure estimates the number of expected clusters, filters cavities using a user-defined threshold and compares different clustering strategies. In case, the user is unfamiliar with the clustering methods, application of cluster validation statistics can assist detecting the most appropriate clustering algorithm.

Structural data of binding sites can provide relevant information with respect to classification and prediction of ligand promiscuity and selectivity. Based on the developed clustering procedure we perform a comparative analysis of the cavity space of proteases. Evaluating the architecture in terms of similarity of the cavity, sequence, and ligand spaces for a subset of human serine proteases provides some important insights into the question whether a ligand-based classification correlates better with a cavity- or a sequence-based classification, as this issue is of utmost importance for the prediction of cross-reactivity among targets in computer-assisted drug design.



(a) (b)

Figure 1.1.: (a) Clustering workflow exemplified for Cavbase. (b) Computation of the similarity matrix for serine proteases using the inhibition data accessible on the ChEMBL database.

1.2. Materials and methods

1.2.1. Data sets selection and benchmark

The first data set is used for the validation of the clustering workflow and contains 502 cavities from 16 different proteins covering all six principal classes of enzymes according to the Enzyme Commission (Bairoch, 2000). This data set will be referred to in the following as *EC data set* (Table 1.1). It is worth mentioning that the data is challenging for classification issues due to two aspects. First, the number of individual entries accounted in the classes deviates strongly, ranging from 5 up to 70. Second, four classes are represented by proteins originating from multiple organisms and one group consists of four different enzyme isoforms, therefore common binding site motifs must be detected independent from any given sequence identity.

The second data set comprises human proteases only except bovine trypsin. These data are termed the *proteases data set* (Table 1.2, Table 1.3). In order to retrieve a reliable and methodologically orthogonal reference classification the Merops database (Rawlings et al., 2010), release 8.4, has been consulted. Merops is a manually curated database that classifies proteases in a hierarchical manner and assigns proteins to families and clans. A Merops family contains proteins for which relationships to a representative protease or another family members can be shown in terms of sequence comparison using a subset of residues only that are responsible for the catalyzed reaction. If possible, families are grouped into clans. A clan contains proteins for which relationships can be established and considers the three-dimensional arrangement of catalytic and non-catalytic residues. Hence, Merops clans include proteins for which relationships cannot be established merely based on sequence comparison, Cavbase should be able to detect sequence-independent relationships, which would then be reflected by the emerged clustering structure. The proteases data set is also used for the workflow validation, but in addition we investigate the differences between the computed clustering and the original Merops classification. The proteases data set considers 90 individual proteases from 12 Merops clans.

The last part of the present study is focused on the serine proteases, a subset of proteases. We generated and compared ligand-, cavity-, and sequence-based clustering. For this purpose we selected 11 proteins for which sufficient public data on ligand inhibition are available (Table 1.4). Ligand data for the regarded proteins have been retrieved from the ChEMBL database (ChEMBL Accessed July 2011). Only ligands have been included in the data set that fulfilled following criteria: molecular weight should be below 600 Da, inhibition constant K_i better than 1μM, achiral, and a maximum of 100 compounds per protein were considered.

1.2.2. Data sets

Table 1.1.: 16 EC classes are represented in the EC data set composed by 502 binding sites.

EC number	Name	Remarks (e.g. organism)	Number
1.1.1.21	Aldose/xylose reductase	Human, pig, C. tenius[1]	62
1.1.1.42	Isocitrate dehydrogenase	E.coli	21
1.1.1.62	Estradiol 17β-dehydrogenase	Human	16
1.14.13.2	Hydroxybenzoate-monooxygenase	P. fluorescens	30
2.7.1.37	Cyclin-dependent kinase 2	Human	46
2.7.1.112	C-Src tyrosine kinase	Human, mouse[1]	20
2.7.4.9	Thymidilate kinase	Human, M.tuberculosis, S.cerevisiase[1]	35
3.4.21.5	Thrombin	Human	41
3.4.23.16	HIV-1 protease	HIV	48
3.4.24.86	TNF-α converting enzyme	Human	16
4.1.1.23	COMP-decarboxylase	Human, S. cerevisiae[1]	36
4.2.1.1	α-Carbonic anhydrase I,II,III,IV[2]	Human	70
5.3.1.5	Xylose isomerase	A. missouriensis	13
5.4.2.1	Phosphoglycerate mutase	S. cerevisiae	5
6.3.2.1	Pantoate-β-alanine ligase	M.tuberculosis	27
6.3.4.4	Adenylosuccinate synthase	E. coli	16

1. Four proteins are regarded that originate from more than one organism.
2. The α-carbonic anhydrase group comprises four different human isoforms.

Table 1.2.: Proteases data set. Aspartate, cysteine, and metallo proteases.

Catalytic mechanism	Merops clan	Protease	pdb id
AP	AA	Pepsin A	1qrp
	AA	Memapsin-2	2vij
	AA	Renin	2g24
	AA	Cathepsin D	1lya
	AA	Cathepsin E	1tzsa
CS	CA	Cathepsin B	2ipp
	CA	Cathepsin C	2djg
	CA	Cathepsin F	1m6d
	CA	Cathepsin K	1mem
	CA	Cathepsin L	1mhw
	CA	Cathepsin S	1nqc
	CA	Cathepsin V	1fh0
	CA	Bleomycin hydrolase	1cb5
	CA	Calpain-1	1zcm
	CA	Calpain-2	1kfu
	CA	Calpain-9	1ziv
	CA	Ubitiquin carboxy-terminal hydrolase L1	2etl
	CA	TNFα-induced protein-3	3dkb
	CA	Otubain-2	1tff
	CD	Caspase-1	1rwn
	CD	Caspase-2	1pyo
	CD	Caspase-3	2dko
	CD	Caspase-7	2qlb
	CD	Caspase-8	2c2z
MP	MA	Angiotensin-converting enzyme peptidase unit-2	1o8a
	MA	Angiotensin-converting enzyme-2	1r42
	MA	Matrix metallo protease-1	2tcl
	MA	Matrix metallo protease-2	1qib
	MA	Matrix metallo protease-3	1ciz
	MA	Matrix metallo protease-7	1mmq
	MA	Matrix metallo protease-8	1zs0
	MA	Matrix metallo protease-9	1gkc
	MA	Matrix metallo protease-10	1q3a
	MA	Matrix metallo protease-12	3f19
	MA	Matrix metallo protease-13	1you
	MA	Matrix metallo protease-16	1rm8
	MA	TNFα-converting enzyme	3b92
	MA	ADAMTS1 protease	2v4b
	MA	ADAMTS4 protease	2rjp
	MA	ADAMTS5 protease	3b8z
	MA	Neprilysin	1r1h
	MA	Endothelin-converting enzyme-1	3dwb
	MC	Carboxypeptidase A2	1aye
	MC	Carboxypeptidase A4	2pcu
	MC	Carboxypeptidase M	1uwy
	MC	Carboxypeptidase U	3d68
	MS	Membrane dipeptidase	1itu
	MG	Methionyl aminopeptidase-2	1b6a
	MG	Xaa-Pro dipeptidase	2okn
	MP	AMSH-like protease	2znr

Table 1.3.: Proteases data set. Serine and threonine proteases.

Catalytic mechanism	Merops clan	Protease	pdb id
SP	PA	Trypsin (bovine)	2zft
	PA	Trypsin IVa	1h4w
	PA	Granzyme A	1orf
	PA	Granzyme B	1iau
	PA	α-Tryptase	2f9n
	PA	β-Tryptase	2bm2
	PA	Kallikrein 1	1spj
	PA	Kallikrein 3	2zch
	PA	Kallikrein 5	2psx
	PA	Kallikrein 6	1lo6
	PA	Kallikrein 7	2qxj
	PA	DECS1	2oq5
	PA	Cathepsin G	1cgh
	PA	Chymase	1klt
	PA	Prostasin	3e0p
	PA	Complement factor B	2ok5
	PA	Complement factor D	1bio
	PA	Complement component C1r	1md8
	PA	Complement component C1s	1elv
	PA	Complement component C2a	2odp
	PA	Plasma kallikrein	2any
	PA	Coagulation factor VIIa	1kli
	PA	Coagulation factor Xa	2jkh
	PA	Coagulation factor XIa	1zsk
	PA	Thrombin	1vzq
	PA	Activated Protein C	3f6u
	PA	Hepsin	1o5e
	PA	Mannan-binding lectin-associated serine peptidase 2	1q3x
	PA	u-plasminogen activator	1gj7
	PA	t-plasminogen activator	1a5h
	PA	Matriptase	1eax
	SC	dipeptidyl-peptidase-4	2g63
	SC	Fibroblast activation protein α-subunit	1z68
	SC	Cholinesterase	1p0i
	SC	Bile salt dependent lipase	1f6w
	SC	Serine carboxypeptidase A	1ivy
	SC	Valacylovir hydrolase	2ocg
	SC	Phosphatase methylesterase-1	3c5v
	SB	Tripeptidyl-peptidase I	3edy
TP	PB	Taspase-1	2a8j

'Catalytic mechanism' column: SP = serine protease, TP = threonine protease

Table 1.4.: Ligands of 11 serine proteases retrieved from the ChEMBL database.

Serine protease	Number of retrieved ligands
Chymase	52
Factor VIIa	100
Factor Xa	100
Kallikrein 1	23
Matriptase	19
Plasma kallikrein	27
Thrombin	100
Tissue-type plasminogen activator	36
Trypsin (bovine)	62
β2-Tryptase	72
Urokinase-type plasminogen activator	100

1.2.3. Similarity matrices

The pocket similarity matrix has been generated by Cavbase. The comparison of sequences was carried out by the fasta35 program (Pearson and Lipman, 1988). The sequence identity values were examined the same way as the Cavbase similarity scores. A sequence identity matrix was constructed and used as input for the clustering procedure.

The computation of the ligand-based similarity matrix for serine proteases has been performed as follows. 691 ligands were mutually compared using the RDKit topology fingerprint (Greg Landrum, RDKit) and a Tanimoto similarity measure (Greg Landrum, RDKit). Since it is known which ligand is associated with which target, the computed similarity matrix can be divided in 121 groups (11x11 proteins). For each group the average similarity is calculated disregarding the trivial matches of identical ligands. This step leads to a more compact matrix that is ready to compare with the other matrices (Figure 1.1b).

1.2.4. Clustering workflow

The basic process of a cluster analysis implies steps like feature selection or extraction, clustering algorithm design or selection, cluster validation, and result interpretation (Xu and Wunsch, 2009). These essential steps have been adopted to create a clustering workflow, which has been validated for the Cavbase similarity metric (Figure 1.1a). Each step is described in detail in the following sections.

1.2.4.1. Distance calculation

The generated Cavbase similarity matrix is symmetrized and normalized. Following the strategy of normalization implicates that all variables are given an equal weight and they can be converted to distances. A vector of normalization factors is determined by $F_N = \frac{1}{\sqrt{S_D}}$, where S_D are the similarity values from the main diagonal. Rows and columns are subsequently multiplied by this vector. The resulting normalized similarity matrix contains 1 on the main diagonal and other values are between 0 and 1. As clustering algorithms depend on the input given in the distance matrix, different clusterings will emerge applying different distance measures. In order to cover the search space as efficient as possible, four different distance measures ($D_1 - D_4$) were computed from the normalized input scores S_N.

$$D_1 = 1 - S_N$$
$$D_2 = \frac{1}{S_N} - 1$$
$$D_3 = (S_N - 1)^2$$
$$D_4 = \sqrt{1 - S_N}$$

1.2.4.2. Threshold filtering

In general, variables with no information content in a data set will make the clustering less clear-cut, therefore, they should be assigned a zero weight, which virtually discards them from the analysis (Kaufman and Rousseeuw, 1990). In order to avoid any unreasonable bias in our clustering we check for data points that fall below a predefined threshold. E.g. such a data point could correspond to a cavity that shares marginal similarity to any other members of the data set except itself. In the following, a threshold of 20 % was set a minimal mutual similarity for all data sets.

1.2.4.3. Estimating number of clusters

A crucial parameter in a clustering procedure is the number of expected clusters, therefore most algorithms require a predefined value given by the user during data set compilation. In case, the number of clusters is not predefined or the data set is not appropriately evaluated, silhouettes can be applied. We have selected two rather complementary approaches:

Average Silhouettes (AS) (Rousseeuw, 1986) and Median Split Silhouettes (MSS) (Pollard and van der Laan, 2002) (section 1.5). AS is a reliable global measure of the relevance of clustering results, whereas MSS analyzes the local structure within a cluster by calculating the average homogeneity of the clusters in the clustering result. Both methods were implemented for the partitioning around medoids clustering (Kaufman and Rousseeuw, 1990). We will stress the predictive power of these two methods when applied to Cavbase similarity matrices.

1.2.4.4. Clustering algorithms and cluster validity assessment

There is a wide range of clustering algorithms, which makes selection of the most appropriate algorithm difficult. We considered the most commonly used hierarchical agglomerative methods (Ward's method, single, complete, group average, median, and centroid linkage), the hierarchical divisive analysis and the partitioning around medoids method (Kaufman and Rousseeuw, 1990). For a given k different clustering methods can lead to different results. Therefore, internal and external criteria can be applied for cluster validity assessment. Detailed description of these approaches can be found elsewhere (Halkidi et al., 2001). In general, external criteria evaluate the results of a clustering algorithm using a predefined structure and internal criteria validate the results in terms of quantities using the proximity matrix itself. We made use of the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) as an external measure to compare the clustering results to an independent reference classification or for the comparison between each other (section 1.5). Furthermore, we tested the ability of nine internal cluster validity criteria (section 1.5) to discriminate between meaningful and less significant clustering structures with respect to the reference classifications.

1.3. Results and discussion

1.3.1. Clustering workflow validation

1.3.1.1. Threshold filter

The application of a threshold of 20 % for the normalized similarity score to discard data points, leads only in case of the proteases data set to an elimination of 16 cavities, which will be discussed in detail below. Subsequently the resulting proteases set contains 74 entries.

1.3.1.2. Number of clusters

In our approach, the estimation of number of clusters depends on one user-specified parameter, namely the maximum number of clusters. This means, the program subsequently computes AS and MSS for the possible number of clusters, from 2 to k_{max}. In case of the EC data set AS is able to find the number of 16 Enzyme Commission (EC) classes taking D_2 as distance. Interestingly, MSS performs better on the proteases data and suggest for three distance measures 7 clusters, which is the number of clans for the data points remaining after the filtering step (Table 1.5). For further analysis the number of clusters was set to 16 for the EC data set and for the proteases to 7, which is in accordance with the reference classifications.

Table 1.5.: Estimated number of clusters for the EC and proteases data set with k_{max}=25

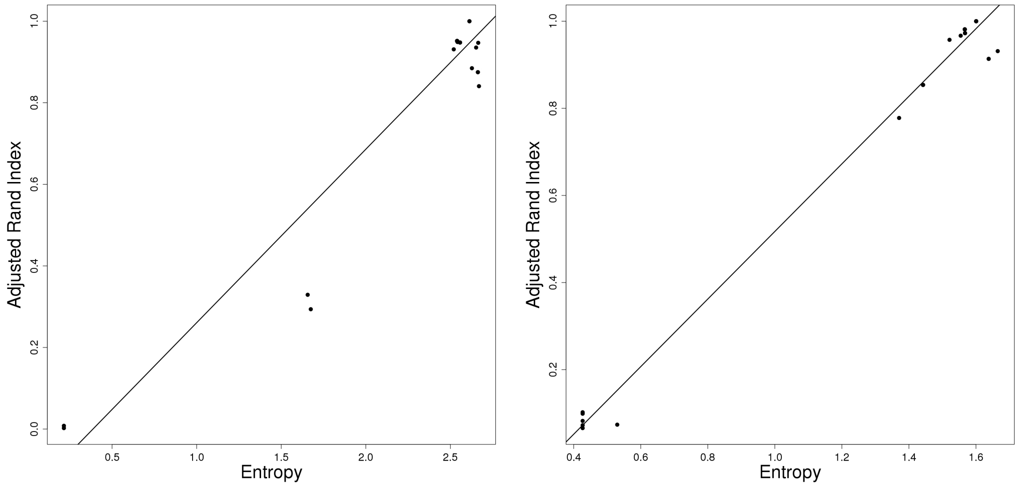
Data Set	Distance	Average Silhouette	Median Split Silhouette
EC	D_1	17	20
	D_2	**16**	21
	D_3	22	18
	D_4	14	23
Proteases	D_1	8	**7**
	D_2	8	24
	D_3	10	**7**
	D_4	9	**7**

1.3.1.3. Cluster validity

As mentioned above, we were interested whether for a given k any cluster validity criterium is able to discriminate between significant and less significant clustering structures. For this purpose the internal cluster validity criteria were checked for correlation with the ARI. An ARI of 1 means that the generated clustering matches a predefined splitting. The lower the ARI the less the agreement with the external classification. A correlation could be found for the entropy of the clustering (Meilă, 2007), which was derived from the information theory (Table 1.6). The correlation plot in Figure 1.2 shows that clustering structures with high ARI values have also a high entropy values. This finding can guide a user to consider only clustering methods best-ranked according to the entropy measure instead of considering all possible clusterings, and investigate them in more detail.

Table 1.6.: Pearson correlation coefficients between cluster statistics and the Adjusted Rand Index.
AB: **A**verage Distance **B**etween Clusters, **AW**: **A**verage Distance **W**ithin Clusters, **WB**: Ratio of Average Distance **W**ithin and **B**etween Clusters, **NB**: **N**umber of Distances **B**etween Clusters, **NW**: **N**umber of Distances **W**ithin Clusters, **DI**: **D**unn **I**ndex, **E**: *E*ntropy, **CH**: **C**alinski and **H**arabasz index, **AS**: **A**verage **S**ilhouette. For details see section 1.5.

cluster validity method	Pearson correlation coefficient	
	EC data set $k = 16$	Proteases data set $k = 7$
AB	0.23	-0.11
AW	-0.28	-0.46
WB	-0.72	-0.38
NB	0.93	0.99
NW	-0.93	-0.99
DI	0.19	-0.02
E	0.96	0.99
CH	0.26	0.31
AS	0.77	0.60



(a) EC data set **(b)** Proteases data set

Figure 1.2.: Pearson correlation of the clustering entropy to the ARI.

1.3.2. Detailed analysis of the proteases clustering

Cavbase is able to cluster successfully the proteases data on the Merops clan level, whereas sequence approaches fail rather miserably (Figure 1.3). This result demonstrates the advantage of a binding site-based classification particularly of remote or unrelated proteins. An all-against-all sequence comparison of proteins in such a data set reveals only a low signal-to-noise ratio, and the obtained clustering disagrees with available knowledge. Although binding site-based clustering of proteases can be matched to the Merops clan classification, there are differences. Due to the initial threshold filtering, 16 cavities are removed from the data set (Table 1.7). In order to illustrate the impact of this step, besides clustering bias prevention, the relationships of the discarded proteins to one another within their group and the relationships to other clustered proteins will be discussed.

Table 1.7.: Protease entries that comprise a similarity values below 0.2 (or 20 %) with any other members of the data set are discarded. 'Catalytic mechanism' column: CP = cysteine protease, MP = metallo protease, SP = serine protease, TP = threonine protease

Catalytic mechanism	Merops clan	Protein name
CP	CA	Human bleomycin hydrolase
CP	CA	Calpain-2
CP	CA	Calpain-9
CP	CA	Ubiquitin carboxy-terminal hydrolase L1
CP	CA	TNFα-induced protein-3
CP	CA	Otubain-2
MP	MC	Carboxypeptidase U
MP	MS	Membrane dipeptidase
MP	MG	Methionyl aminopeptidase-2
MP	MG	Xaa-Pro dipeptidase
MP	MP	AMSH-like protease
SP	SC	Serine carboxypeptidase A
SP	SC	Valacylovir hydrolase
SP	SC	Protein phosphatase methylesterase-1
SP	SB	Tripeptidyl-peptidase-1
TP	PB	Taspase-1

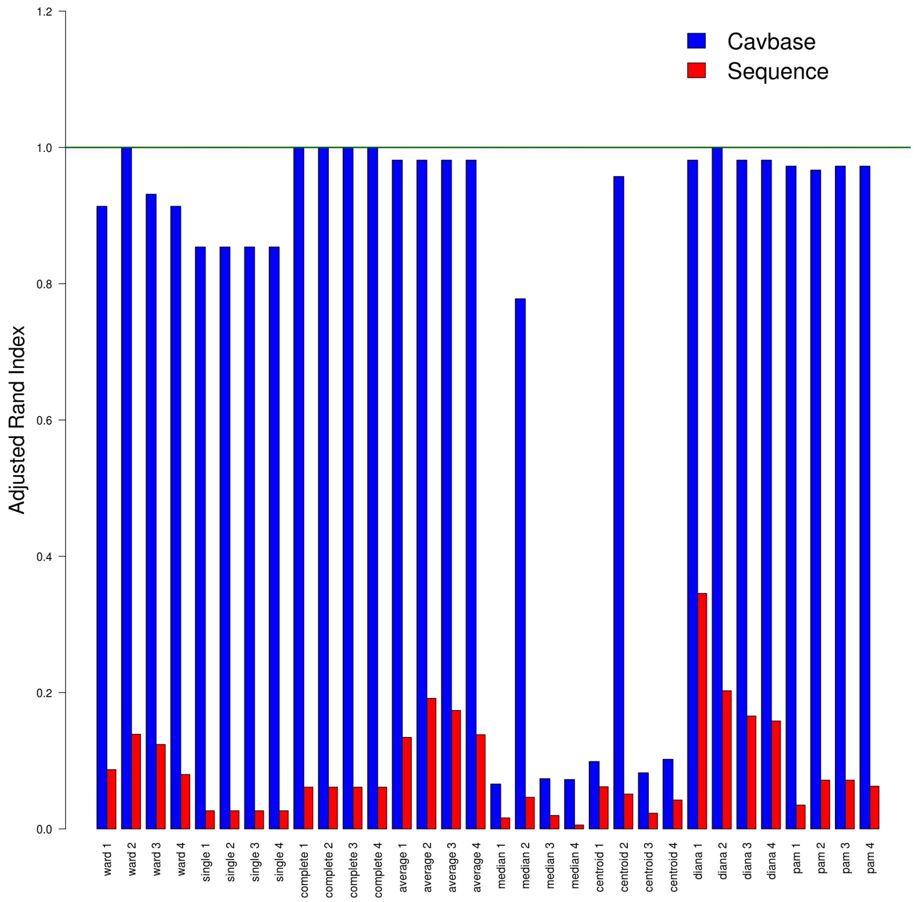
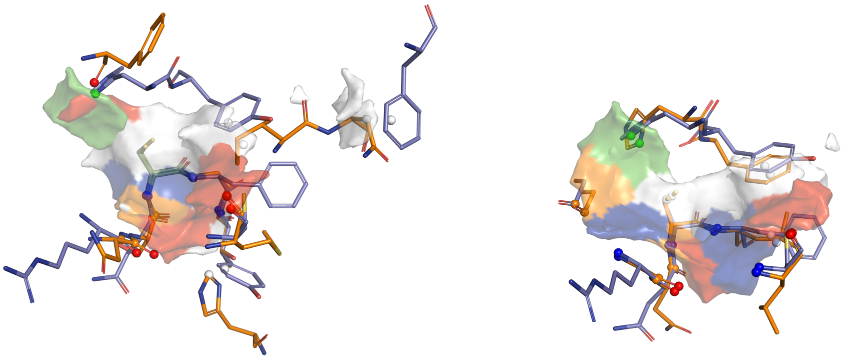


Figure 1.3.: Cavbase and sequence clustering of proteases. On the x-axis all implemented clustering methods and used distances are shown. The y-axis represents the ARI. As the Merops clan classification was defined as reference, a clustering that matches the reference classification, has an ARI of 1, which is depicted as a solid line intersecting the y-axis. ARIs of the Cavbase approach are shown in blue and the results from the sequence analysis are shown in red. Cavbase is able to reproduce the Merops clan classification in case of complete-linkage for all calculated distances, in case of Ward's method and divisive analysis for the D_2 distance measure. In contrast, sequence analysis of the proteases data set delivers rather poor performance.
Used clustering methods: ward = Ward's method, single = single linkage, complete = complete linkage, average = group average, median = median linkage, centroid = centroid linkage, diana = hierarchical divisive analysis, pam = partitioning around medoids.



(a) UCH-L1 vs. TNFα-IP3 **(b)** UCH-L1 vs. otubain-2

Figure 1.4.: Overlapping active site surface regions of ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) is superimposed onto the cavities of (a) tumor necrosis factor-α-induced protein-3 (TNFα-IP3) and (b) otubain-2.

Four proteins that represent an entire Merops clan on themselves were discarded from the data set. These proteins are: membrane-bound dipeptidase, AMSH-like peptidase, tripeptidyl-peptidase-1, and taspase-1.

Three other proteins which are cysteine proteases from the same clan belong to the group of deubiquitinating enzymes (DUBs). DUBs have different topologies and mechanisms of substrate recognition, but the spatial arrangement of the catalytic triad and the oxyanion hole are highly conserved (Nanao et al., 2004). The resulting selectivity and uniqueness of DUBs' binding sites is reflected by the filtering step of the Cavbase similarity matrix. UCH-L1, TNFα-IP3, and otubain-2 are from the same clan, but apparently the differences in their binding sites are significant. In Figure 1.4 the mutually matched binding site surface patches of both enzymes are superimposed and shown in a side-by-side view. The enzymes share only the conserved catalytic core in common. Although the arrangement of the proposed catalytically active histidine (His161) of UCH-L1 is too remote to create a catalytically active His-Cys diad (Das et al., 2006), Cavbase is able to match the conserved environment around the catalytically active cysteines.

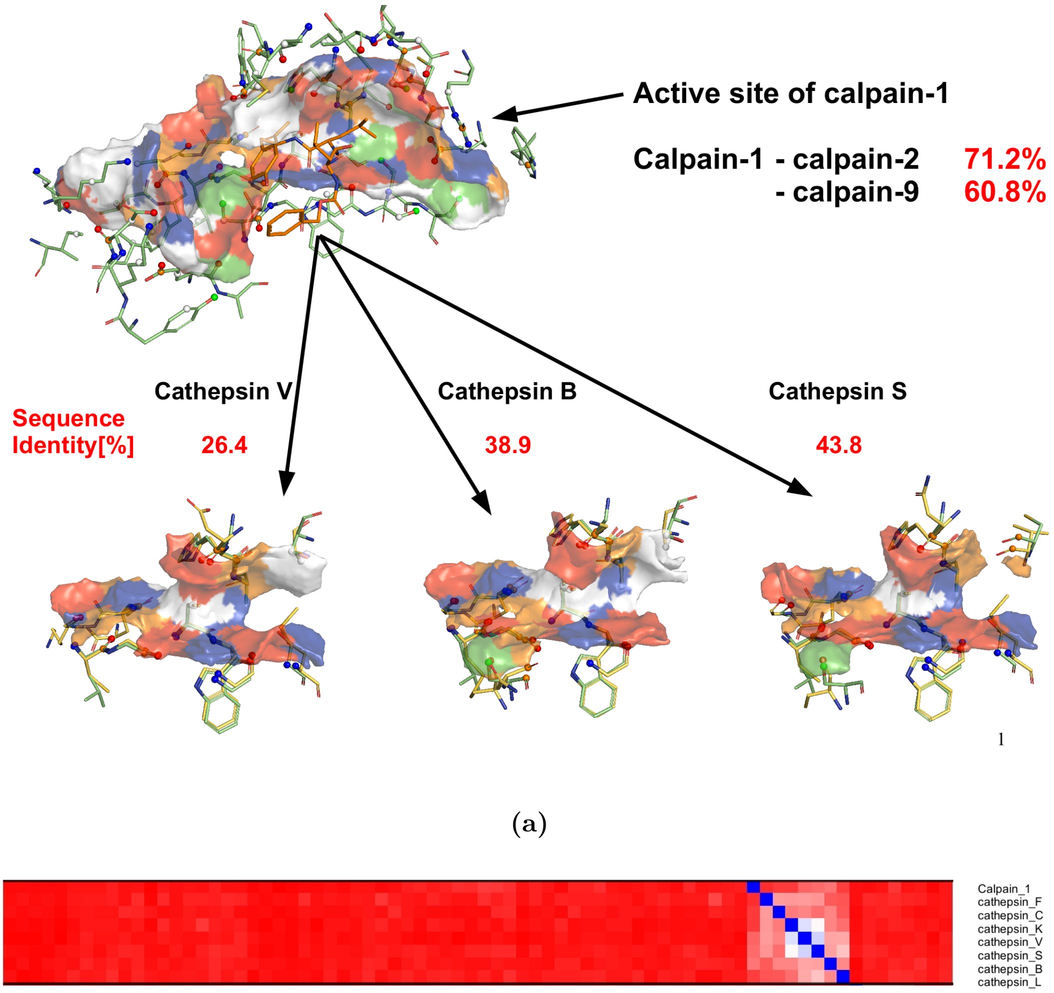
UCH-L1 enzyme is associated with Parkinson's disease and lung cancer (Das et al., 2006) and active-site inhibitors of this enzyme show antiproliferative effects in the H1299 lung cancer cell line (Liu et al., 2003). Even though the three DUBs are only sparsely populating the cysteine protease family, the application of the filtering step before clustering the matrix can provide valuable information about the uniqueness and singularity of particular binding sites which can help to resolve the selectivity issues of a specific target protein. A cavity screening of the UCH-L1 binding site against the entire Cavbase containing about 275 000 cavity entries revealed that most similar binding sites, apart from UCH-L1 itself, show a similarity of only 20 % or lower.

Interestingly, also proteins from the same Merops family and clan are discarded applying the 20 % threshold filter. This indicates high specificity of these enzymes towards their substrates, owing to differently exposed active site properties, which is not reflected in the sequence space. An example is valacylovir hydrolase and protein phosphatase methylesterase-1 which both fall into the same Merops family. Valacyclovir hydrolase displays high specificity for cleavage of amino acid esters (Lai et al., 2008), whereas the protein phosphatase methylesterase-1 binds selectively the carboxy-terminal residues of the catalytic subunit of protein phosphatase-2A (Xing et al., 2008).

The cysteine protease human bleomycin hydrolase (hBH) is also removed in cavity space although it shares high sequence identity of about 50 % to the cathepsins B, K, and S. hBH is a representative of a self-compartmentalizing protease. The flexibility of its C-terminus contributes to the active site and controls the activity of the enzyme (O'Farrell et al., 1999). The C-terminus near the catalytic cysteine is involved in substrate binding and forms a specific cavity that barely shares any similarity with the regarded cysteine cathepsins.

1.3.2.1. Selectivity of calpain-1 inhibitors over cysteine cathepsins

Calpains are also cysteine proteases that participate in many calcium regulated functions, e.g. cell proliferation, differentiation, and apoptosis. Their activity depends on the presence of calcium ions. All considered calpains are assigned to the same Merops family C2 and clan CA. In our cavity cluster analysis calpain-2 and calpain-9 are removed from the data set whereas calpain-1 is assigned to the cluster comprising the cysteine cathepsins from the Merops C1 family. From this finding several conclusions can be drawn. First, separation of individual calpains reflects the structural flexibility and diversity of active sites in calpains (Davis et al., 2007). Second, detecting calpain-1 in the same cluster with cysteine cathepsins leads to the assumption that their binding site properties are similar, despite the low sequence identity of calpain-1 with the other cathepsins, which varies from 25 to 44 %. It is interesting to see whether this assumption is also reflected by independent ligand data published in literature. Indeed, the majority of calpain-1 inhibitors lack selectivity over corresponding cathepsins (Donkor, 2011). The overlapping binding site surfaces are visualized for calpain-1 and three cysteine cathepsins V, B, S in Figure 1.5. The Ligsite algorithm detects as geometrically most distinct subsites of calpains S1, S2, S1', S2' which were used for the analysis. The S1 and S1' pockets of calpain-1 are highly similar to the corresponding subsites of cathepsins, but the S2 and S2' pockets differ in terms of their exposed properties. This observation suggests for the design of putatively selective ligands a stronger focus on specific interactions with the residues in the S2/S2' subpockets. Commonly used P2 residues to be considered in ligands are valine and leucine, which provide affinity towards calpains, however do not facilitate selectivity (Choe et al., 2006). For instance, Cuerrier et al. reported that placement of groups capable of hydrogen-bond formation at the P2 position improves ligand selectivity of calpain-1 over the cysteine cathepsins (Cuerrier et al., 2007). Similar effects to achieve selectivity by exchanging hydrophobic for hydrogen bonding groups have been also described for aspartyl proteases (Kawasaki and Freire, 2011).



(a)

(b)

Figure 1.5.: Cross-reactivity between calpain-1 and cysteine cathepsins.
(a) Sequence independent similarity of binding site surfaces is detected. Although calpain-2 and calpain-9 exhibit a high sequence identity to calpain-1, the binding site of calpain-1 shares a higher similarity to the cathepsins.
(b) Hence, calpain-1 is found in the same cluster with the cathepsins.

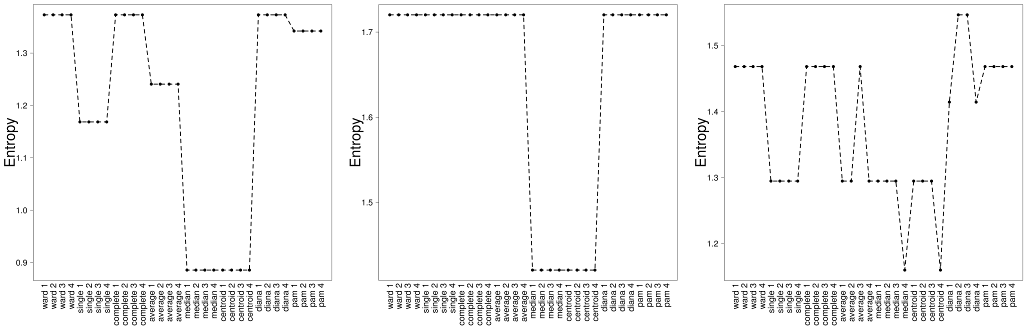
1.3.3. Ligand, cavity, and sequence data: Cluster analysis of serine proteases

Weskamp et al. have shown that the cavity space correlates well with ligand binding data and fold space (Weskamp et al., 2009). A more detailed analysis of Stegemann et al. concentrated on a data set of proteins with mutual sequence identity below 25 % that bind cofactors as ligands (Stegemann and Klebe, 2011). Comprehensive studies on kinases showed significant correlation between similarity of binding sites and the respective ligands they bind (Kinnings and Jackson, 2009; Spitzer et al., 2011). In the present work, we were interested whether the previously observed trends can be extended to serine proteases considering a broad range of bioactive ligands. Therefore we faced the ligand similarity matrix with those obtained from cavity and sequence space and performed a similar clustering.

In order to extract information by comparable means from the different matrices, k and the applied clustering method must be identical. The determination of an optimal k was performed for the three spaces. Our routine suggested for the ligand data a marked value of five and for the sequence data the number of clusters was estimated to four and six, respectively (Table 1.8). The values suggested for the cavity data are less clear-cut, as $k = 2$ is too small to come up with a meaningful clustering and $k = 8$ leads to a clustering comprising a large number of singletons. Hence, selection of the most appropriate clustering was performed for each k on sequence and ligand data using the above introduced entropy measure as evaluation criterion (Figure 1.6). The results have been mutually compared by the ARI. Sequence and cavity data reveal clusterings that deviate significantly, as indicated by the ARI. However, the emerging clustering based on cavity data correlates well with the clustering based on ligand topology and the obtained similarity is significantly more pronounced than on the corresponding sequence data level (Table 1.9).

Table 1.8.: Serine proteases. Estimated number of clusters based on the ligand similarity, Cavbase score, and sequence identity matrices of serine proteases with k_{max}=11

Data set	Distance	Average Silhouette	Median Split Silhouette
Ligands	D_{1-4}	5	5
Cavities	D_{1-4}	2	8
Sequences	D_{1-4}	4	6



(a) Sequences: $k = 4$ **(b)** Sequences: $k = 6$ **(c)** Ligands: $k = 5$

Figure 1.6.: Entropy measure for the (a,b) sequence and (c) ligand clusterings of serine proteases. The highest entropy values for the sequence data using $k = 4$ are found for three and using $k = 6$ for six of the applied clustering methods. Ligand data suggest for $k = 5$ the divisive analysis clustering for D_2 and D_3.

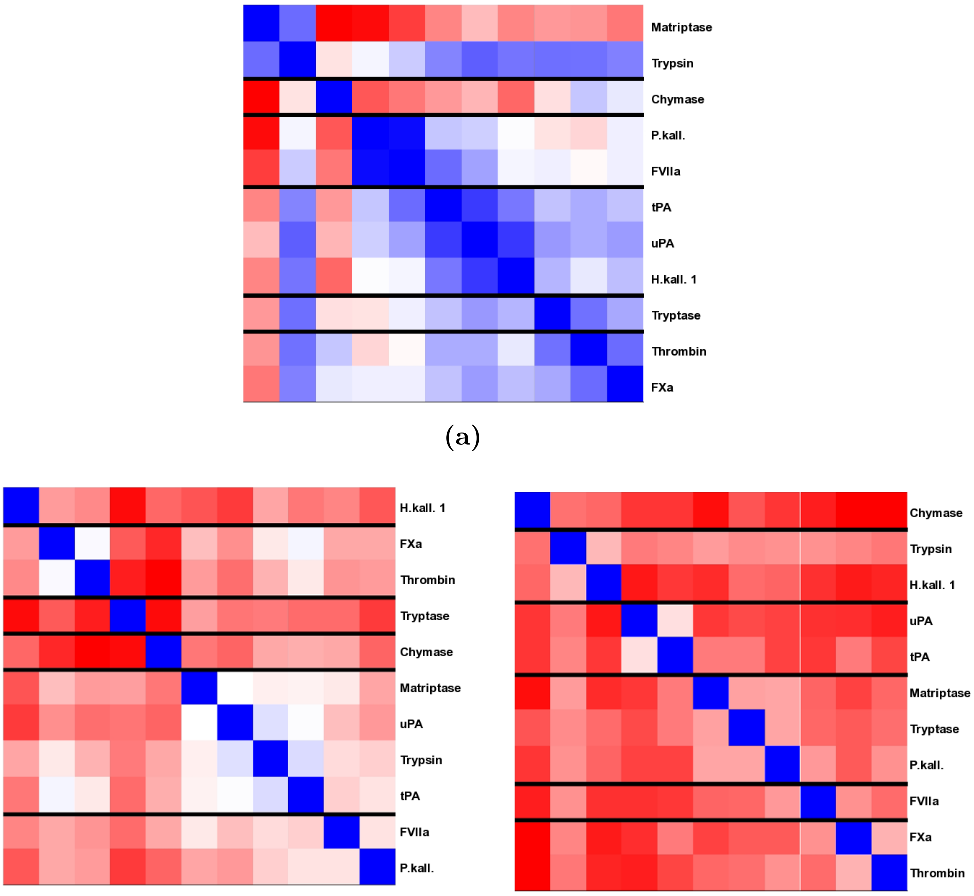
Table 1.9.: ARIs for the clustering structures of the ligand, cavity, and sequence space.

k	Clustering Method	Distance	Sequence-Cavity	Sequence-Ligand	Cavity-Ligand
4	ward	D_1	0.09	0	0.06
		D_2	-0.03	0	0.04
		D_3	-0.03	0	0.04
		D_4	0.09	0	0.06
	complete	D_1	-0.03	-0.1	0.04
		D_2	-0.03	-0.1	0.04
		D_3	-0.03	-0.1	0.04
		D_4	-0.03	-0.1	0.03
	diana	D_1	-0.01	-0.1	-0.05
		D_2	-0.01	-0.1	-0.05
		D_3	-0.01	-0.1	-0.05
		D_4	-0.01	-0.06	-0.06
5	diana	D_2	0.01	0.12	0.13
		D_3	0.01	0.12	0.13
6	ward	D_1	0.39	0.25	0.39
		D_2	0.11	0.25	0.25
		D_3	0.39	0.25	0.39
		D_4	0.11	0.25	0.39
	single	D_1	0.04	0.11	0.3
		D_2	0.04	0.11	0.3
		D_3	0.04	0.11	0.3
		D_4	0.04	0.11	0.3
	complete	D_1	0.18	0.25	0.51
		D_2	0.18	0.25	0.51
		D_3	0.18	0.25	0.51
		D_4	0.18	0.25	0.51
	average	D_1	0.18	0.25	0.51
		D_2	0.18	0.25	0.51
		D_3	0.18	0.25	0.51
		D_4	0.18	0.25	0.51
	diana	D_1	0.04	0.25	0.39
		D_2	0.04	0.25	0.39
		D_3	0.04	0.25	0.39
		D_4	0.04	0.25	0.39
	pam	D_1	0.04	0.35	0.2
		D_2	0.18	0.35	0.56
		D_3	0.04	0.35	0.2
		D_4	0.04	0.35	0.2

A more detailed analysis of the investigated spaces has been carried out using the following clustering settings: $k = 6$ with the complete-linkage clustering method and the previously introduced distance measure D_3. The results are depicted in terms of heatmaps in Figure 1.7. The more bluish the color the more similar are the data points, whereas, red color indicates increasing dissimilarity.

The overall structuring of three heatmaps suggests much higher discriminative power for the ligand and cavity data compared to the sequence data. The latter shows hardly any discriminative power apart from urokinase-type plasminogen activator (uPA) and tissue-type plasminogen activator (tPA) or trypsin and and kallikrein-1 which end up in joint clusters. On the ligand heatmap trypsin is of special evidence as strikingly an extended blue bar demonstrates the unspecific character of this particular enzyme. With respect to substrate cleavage trypsin is one of the most promiscuous enzymes in this family (Hilpert et al., 1994). Two clusters are identically indicated in the three input spaces based on ligand, cavity and sequence data. Coagulation factor Xa (fXa) and thrombin share a common cluster whereas chymase ends up in all cases as a singleton. Thrombin and fXa are closely related members of the blood coagulation cascade (Sanderson, 1999) and even the successful development of dual inhibitors acting equally potent against both enzymes has been accomplished (Nar et al., 2001). The human chymase, which is the only representative chymotrypsin-like protease in the data set, is found as a singleton, reflecting its distinct properties in the data set and its high selectivity towards its biological substrates (McGrath et al., 1997).

Focusing on the more discriminating spaces based on ligand and cavity information, related cluster patterns are found for the two-fold clusters formed by Coagulation factor VIIa (fVIIa) and plasma kallikrein (PK) and the singleton created by β-tryptase. The latter β-tryptase is a mast cell serine protease that has been directly linked to the pathology of asthma (Molinari et al., 1996). Its clustering as a singleton is supported by the fact that successful design of selective β-tryptase inhibitors could be achieved



(a)

(b) **(c)**

Figure 1.7.: Heatmaps obtained for clustering (a) ligand, (b) cavity, and (c) sequence data applying the complete-linkage method and $k = 6$. Deep blue color represents maximum possible similarity and deep red color maximum dissimilarity in the corresponding data set. The range between the maxima is mixed with white color. For reasons of clarity the six individual clusters are separated only by horizontal black lines and the entries are labeled to the right of the heatmap.

in several studies (Combrink et al., 1998; Hopkins et al., 2005; Lee et al., 2006a). On the contrary, only a few attempts have been described to address the selectivity problem depicted in the heatmaps of fVIIa and PK, which is particularly indicated for the ligand clustering (Olivero et al., 2005; Young et al., 2006). Interestingly, the latter selectivity issue reflected by the properties of known ligands is already suggested by our comparative analysis in cavity space as fVIIa and PK show the highest mutual similarity in this space. As a consequence and reflecting the current state of inhibitor development a similar clustering is therefore proposed for ligand topology information and exposed physicochemical properties of the binding pockets.

1.4. Conclusions

In the present study, we describe the development, validation, and application of a novel clustering workflow with particular focus on the Cavbase similarity metric. Owing to the implemented routines any proximity matrix can be provided as input. The program is able to predict the correct number of clusters for two data sets of binding sites and clusters them automatically in accordance with expert classifications, based on orthogonal information such as EC numbers and Merops clans.

As a case study, human proteases were analyzed in more detail. Clustering based on cavity information indicates a cross-reactivity between the cysteine protease calpain-1 and cysteine cathepsins, which has been reported upon calpain-1 inhibitor development in literature (Donkor, 2011). Unlike binding site information, the usage of a sequence identity matrix as input for clustering fails to produce any meaningful results, thereby making the detection of the described cross-reactivity virtually impossible.

Finally, we utilize our workflow in an attempt to investigate the relationships between ligand, cavity, and sequence spaces of serine proteases. Clustering of ligands, using solely similarities based on their topologies, leads to a pattern that shows higher correlation to the clustering of binding sites than to that of sequences. On the one hand, this result has to be treated with caution, as only eleven serine proteases were considered in the analysis. This fact results mainly from the limited access to the sparse ligand data stored in public databases. On the other hand, the evaluation of binding site information along with protein classification from orthogonal sources can deliver in a data mining approach valuable data to discriminate proteins with respect to selectivity criteria for the development of putative ligands that standard sequence comparison methods can hardly achieve.

1.5. Supporting information

1.5.1. Cluster validation statistics - Internal criteria

Let $Z = \{z_1, z_2, ...z_l\}$ be a set of numbers then $avg(Z) := \frac{1}{l}\sum_{i=1}^{l} z_i$. Let $X = \{x_1, x_2, ...x_n\}$ be the data set and $C_1, ...C_k$ the clusters with $n_j = |C_j|$ the number of elements in C_j $\forall j = 1...k$. As every element is in one and only one cluster, there is $\sum_{j=1}^{k} n_j = n$.

Let us define the function $\mathcal{C} : X \longrightarrow \{1, ..., k\}$ as $\mathcal{C}(x_i) = j \Leftrightarrow x_i \in C_j$

1.5.1.1. ***A***verage Distance ***B***etween Clusters (***AB***)

$$AB = avg(\{d(x_i, x_j)|\mathcal{C}(x_i) \neq \mathcal{C}(x_j)\})$$

1.5.1.2. ***A***verage Distance ***W***ithin Clusters (***AW***)

$$AW = avg(\{d(x_i, x_j)|\mathcal{C}(x_i) = \mathcal{C}(x_j) \wedge i \neq j\})$$

1.5.1.3. Ratio of Average Distance ***W***ithin and ***B***etween Clusters (***WB***)

$$WB = \frac{AW}{AB}$$

1.5.1.4. ***N***umber of Distances ***B***etween Clusters (***NB***)

$$NB = \sum_{i<j\leq k} n_i n_j$$

1.5.1.5. ***N***umber of Distances ***W***ithin Clusters (***NW***)

$$NW = \sum_{j=1}^{k} \frac{n_j(n_j - 1)}{2}$$

1.5.1.6. ***D***unn ***I***ndex (***DI***, Dunn (1973))

$$DI = \frac{\min(\{d(x_i, x_j)|\mathcal{C}(x_i) \neq \mathcal{C}(x_j)\}}{\max(\{d(x_i, x_j)|\mathcal{C}(x_i) = \mathcal{C}(x_j)\}}$$

1.5.1.7. ***E***ntropy (***E***, Meilă (2007))

$$E = -\sum_{j=1}^{k} \frac{n_j}{n} log \frac{n_j}{n}$$

k is the number of non-empty clusters. n is the number of data points in the data set and n_j is number of data points in cluster C_k.

1.5.1.8. ***C***alinski and ***H***arabasz index (***CH***, Calinski and Harabasz (1974))

$$CH = \frac{(n-k) \sum_{\mathcal{C}(x_i) \neq \mathcal{C}(x_j)} d(x_i, x_j)^2}{(k-1) \sum_{\mathcal{C}(x_i) = \mathcal{C}(x_j)} d(x_i, x_j)^2}$$

1.5.1.9. ***A***verage ***S***ilhouette (***AS***, Rousseeuw (1986))

Calculate average distance of object x_i to other elements of its own cluster.

$$a(x_i) = avg\{d(x_i, x_j)|\mathcal{C}(x_i) = \mathcal{C}(x_j) \wedge i \neq j\} = \frac{1}{n_{\mathcal{C}(x_i)} - 1} \sum_{\mathcal{C}(x_i) = \mathcal{C}(x_j)} d(x_i, x_j)$$

Calculate average distance of object x_i to members of clusters j.

$$b_j(x_i) = avg(\{d(x_i, x_l)|\mathcal{C}(x_l) = j\}) = \frac{1}{n_j} \sum_{\mathcal{C}(x_l) = j} d(x_i, x_l)$$

$$b_{j_0}(x_i) = \min_{j \neq \mathcal{C}(x_i)} b_j(x_i)$$

The silhouette of element x_i is defined as follows.

$$S_i = \frac{b_{j_0}(x_i) - a(x_i)}{max(b_{j_0}(x_i), a(x_i))}$$

The average silhouette over all elements of the clustering is calculated and should be minimized.

$$AS = avg(\{S_i\}_{i=1}^{n})$$

1.5.1.10. ***M*edian *S*plit *S*ilhouette (*MSS*, Pollard and van der Laan (2002))**

For a given clustering result with k clusters each cluster is splitted into two or more clusters and a new silhouette is computed for each element relative to other elements of the same parent cluster. The average for each parent cluster is the split silouette $SS_i, i = 1, 2, \ldots, k$. *MSS* is the median of split silhouettes over k clusters:

$$MSS(k) = \frac{1}{k} \sum_{i=1}^{k} SS_i$$

1.5.2. Cluster validation statistics - External criteria

1.5.2.1. *A*djusted *R*and *I*ndex (*ARI*, Hubert and Arabie (1985))

Given a clustering with partitions U and V, for all possible pairs of data points i and j the quantities of a, b, c and d and their cluster assignments $C_{U(i)}, C_{U(j)}, C_{V(i)}$ and $C_{V(j)}$ are computed.

$$a = |\{i, j | C_{U(i)} = C_{U(j)} \wedge C_{V(i)} = C_{V(j)}\}|$$

$$b = |\{i, j | C_{U(i)} = C_{U(j)} \wedge C_{V(i)} \neq C_{V(j)}\}|$$

$$c = |\{i, j | C_{U(i)} \neq C_{U(j)} \wedge C_{V(i)} = C_{V(j)}\}|$$

$$c = |\{i, j | C_{U(i)} \neq C_{U(j)} \wedge C_{V(i)} \neq C_{V(j)}\}|$$

a and d count the correspondeces, b and c count the deviations of two partitionings. The Rand Index (RI)is defined as follows

$$RI(U, V) = \frac{a + d}{a + b + c + d}$$

Using a different pepresentation based on the contingency table defined by U and V, the Adjusted Rand Index is defined as

$$ARI(U, V) = \frac{\sum_{lk}\binom{n_{lk}}{2} - [\sum_l \binom{n_l}{2} \cdot \sum_k \binom{n_k}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_l \binom{n_l}{2} + \sum_k \binom{n_k}{2}] - [\sum_l \binom{n_l}{2} \cdot \sum_k \binom{n_k}{2}]/\binom{n}{2}}$$

n_{lk} = the number of data points which where assigned to the same cluster.

Chapter 2.

2 Automated Decomposition of Cavities Based on Bound Ligand Information

2.1. Preliminary remarks

The present study was accomplished in collaboration with Thomas Rickmeyer, who stayed for an internship of six months in our research group.

2.2. Introduction

A general introduction of the Cavbase methodology has been presented in section 1.1. The current implementation of the binding site comparison and subsequent similarity calculation is able to find a *global* pairwise similarity (Schmitt et al., 2002), therefore, *local* similarities of subpockets are more difficult to detect. The following reasons can be put forward to explain this observation. First, the similarity calculation algorithm based on a clique detection as implemented in Cavbase is not exhaustive, since the maximal number of scored cliques is restricted to the 100 most largest ones. This fact becomes in particular a limitation when applied to large binding pockets. Second, the detection of the MCS does not guarantee that all local similarities are regarded and in consequence even scored on high ranks. Therefore, an approach is needed to detect local binding site similarities that sum-up and produce a more accurate representation of regions which are shared in common by two cavities. Such a procedure is of particular interest for the screening of a pocket database, in order to find pocket similarities of proteins remote in sequence or obviously not belonging to the same protein family.

One well-known example of more accurate similarity detection using subpockets in Cavbase has been described for a particular cross-reactivity of the drug celecoxib. Celecoxib was originally designed as a specific inhibitor of cyclooxygenase-2 (COX-2). However, celecoxib also shows nanomolar inhibition of α-carbonic anhydrase II (α-CA II). Reasons for this phenomenon could be successfully elucidated on structural level by comparing the crystal structures complexed by celecoxib and using subpockets defined by the functional groups of the SC-558 inhibitor, a bromo analog of celecoxib bound to COX-2 (Weber et al., 2004).

The primary goal of this study targeted the improvement of the binding site similarity detection between remote proteins using more appropriately the local information about subcavities. As in the previously mentioned case of celecoxib it was achieved using the subpockets defined by functional groups (fragments) of the bound ligand. Therefore a subcavity was defined as that part of the total cavity extracted by Ligsite that accommodated manually predefined ligand fragments. We attempted to develop an automatic procedure to perform such screenings. Since, different ligand fragmentation methods lead not only to different splitted fragments but also to a deviating number of fragments, the method will take directly impact on the size and shape as well as the number of generated subcavities for a given pocket. Therefore, we tested two different ligand fragmentation strategies in our approach, which will be described together with the cavity decomposition procedure in section 2.3.

2.3. Automated binding site decomposition procedure

2.3.1. Ligand fragmentation methods

Most common cheminformatic methods which are used to split bonds and/or match substructures of a molecule are encoded in line notation, as realized in the widely used **S**implified **M**olecular-**I**nput **L**ine-**E**ntry **S**ystem (SMILES). Since, 3-D information of the ligand must be retained in order to enable distance calculation to the adjacent pocket environment, a software that accepts and preserves ligand information in 3-D format is required. The program **D**ecomposition **A**nd **I**dentification of **M**olecules (DAIM) (Kolb and Caflisch, 2006) fulfills this requirement. Originally, DAIM was developed to compute a set of unique fragments using a virtual compound library as input. The generated fragments can be used further for fragment-based docking. An internal definition of fragments is implemented in DAIM (for details see Kolb and Caflisch (2006)). However, individual parameter definitions for uncleavable bonds can be provided by the user. This flexibility of the program allowed us to implement rules compatible with the well-known **Re**trosynthetic **C**ombinatorial **A**nalysis **P**rocedure (RECAP) (Lewell et al., 1998) for compound decomposition. The idea of RECAP is the generation of virtual but chemically more reasonable compound libraries. Primarily, eleven bond types, which in turn can be made by common chemical reactions, were considered in RECAP to generate a set of building blocks that can be reassembled in a subsequent combinatorial procedure.

The presented approaches, DAIM and RECAP, apply rather different ligand fragmentation rules as they were originally designed to accomplish different tasks. We adopted fragmentation rules from both strategies and investigated their suitability for subsequent cavity decomposition.

2.3.2. Cavity decomposition workflow

The developed workflow starts with the ligand extraction in Protein Data Bank (PDB) format. The ligand file is converted to MOL2 format using MOE whereby hydrogens and partial charges are added simultaneously. Next, the ligand is decomposed according to the two presented strategies. For each generated fragment the Cavbase assigned pseudocenters are selected, which reside on the protein residues within the radius of 4.5 Å to any atom of the fragment. This information along with the corresponding surface patches was considered as subcavity. In the next step, the program checks whether pseudocenters are assigned to more than one subcavity and re-assigns them to the closest one only. Subsequently, the subcavites are stored and ready to use as queries. If in a comparison several subcavities originating from the same Ligsite cavity are matched with a candidate cavity a total similarity score is computed as the sum of the individual scores which have been calculated separately for each individual subcavity match.

Figure 2.1 shows the established workflow exemplified for the COX-2 inhibitor SC-558. Worth mentioning, the default DAIM rules lead in case of SC-558 to rather small fragments (Figure 2.2). Therefore, we modified the initial fragment definition in DAIM in order achieve the most reasonable subcavity representation for COX-2 based on visual inspection.

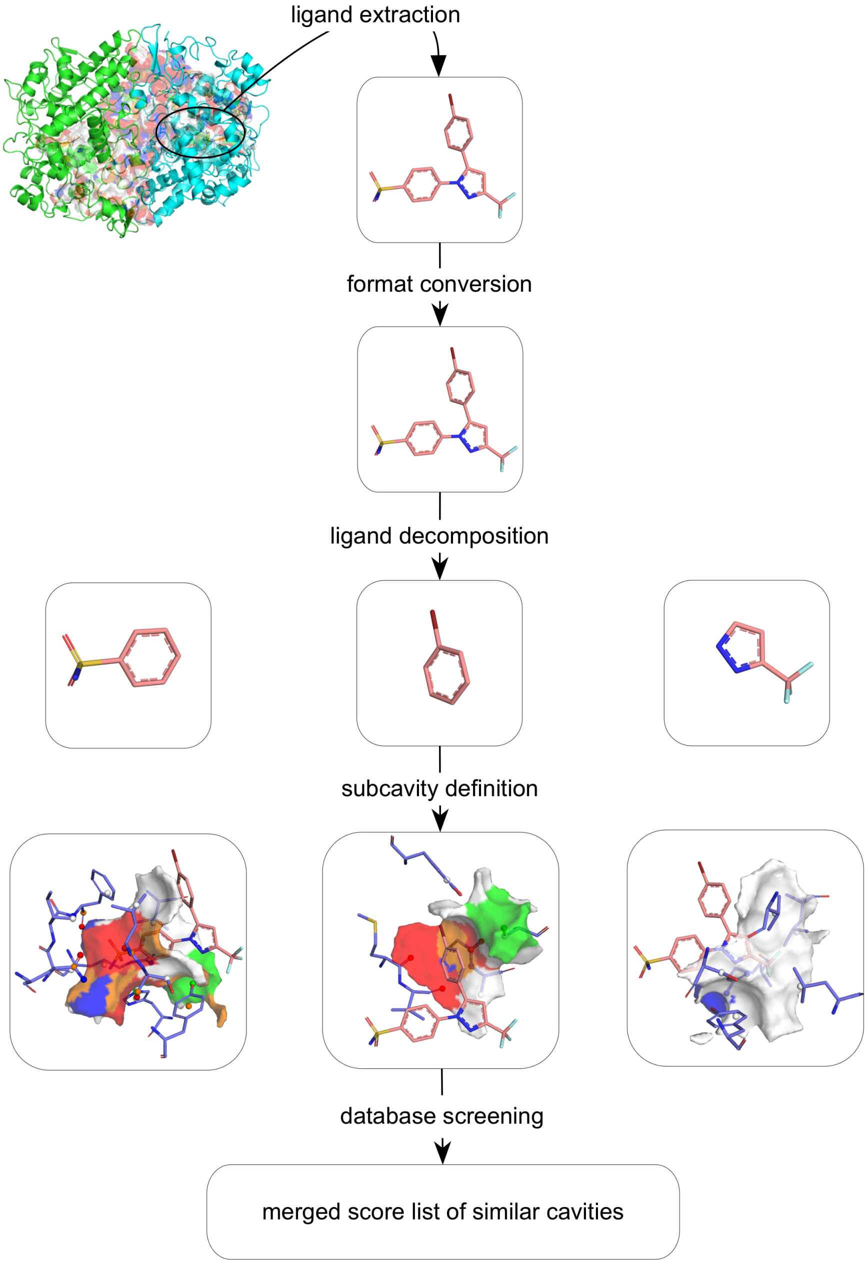


Figure 2.1.: Cavity decomposition workflow exemplified for COX-2 inhibitor SC-558 (PDB 6COX) applying modified DAIM rules for ligand fragmentation. The inhibitor is a celecoxib analog, which contains bromine in place of a methyl group.

2.4. Case studies

For test purposes ligands were required which are known to bind to remotely related proteins. Three cases could be identified, where crystal structures are available:

1. SC-558 bound to COX-2 (6COX) and celecoxib bound to α-CA II (1OQ5),
2. ritonavir bound to endothiapepsin (3PRS) and human immunodeficiency virus (HIV) protease (1RL8),
3. lovastatin bound to lymphocyte function-associated antigen-1 (LFA-1) (1CQP) and simvastatin bound to 3-hydroxy-3-methyl-glutaryl-CoA reductase (HMG-CoA reductase) (1HW9).

The results will be discussed in section 2.5.

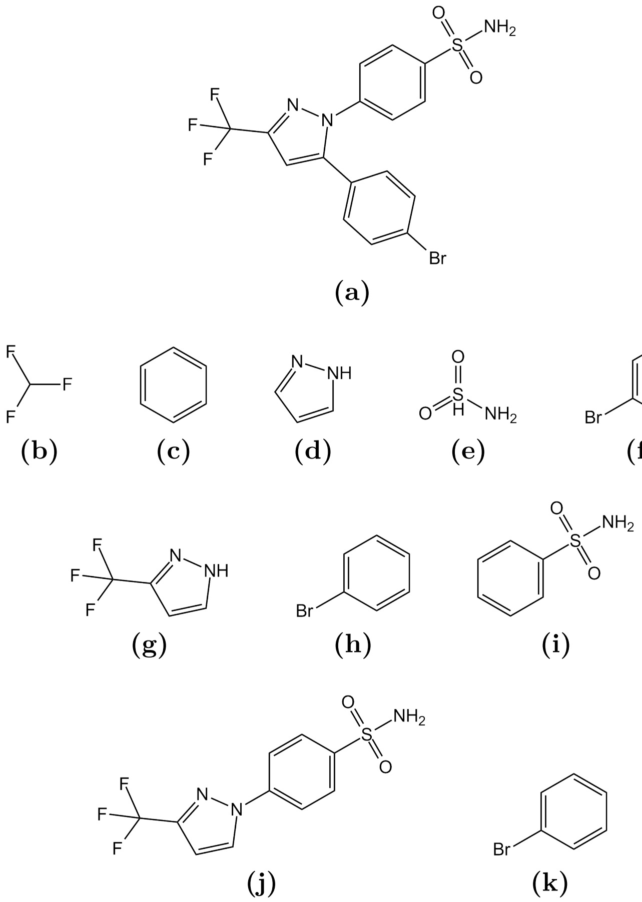
2.4.1. Pairwise comparison

Three different scenarios for cavity comparison with respect to the query have been constructed using a radius of 4.5 Å as default for cavity restriction:

1. complete Ligsite cavity,
2. ligand defined, and
3. fragment defined subcavity.

2.4.1.1. SC-558

The fragments of SC-558 after decomposition are shown in Figure 2.2. The similarity scores are presented in Table 2.1.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

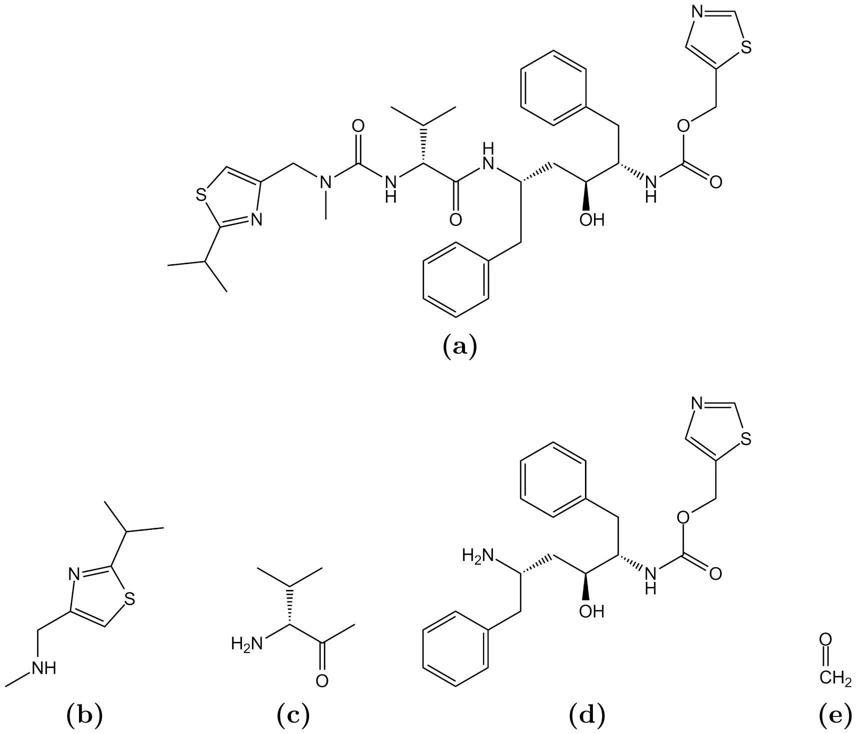
Figure 2.2.: Fragments of the inhibitor SC-558 (**a**) are shown. **b–f** are fragments of the default DAIM rules. **g–i** are fragments generated by modified DAIM rules. **j, k** are computed by RECAP.

Table 2.1.: Similarity scores of the Ligsite, SC-558, and respecitve fragment defined cavities of COX-2 vs. α-CA II (6COX vs. 1OQ5). Using the complete Ligsite cavity as query does not lead to correct match of active sites.

Ligsite	Complete	RECAP			DAIM			
cavity	ligand	frag 1	frag 2	$\sum$	frag 1	frag 2	frag 3	$\sum$
(4.696)	3.296	3.296	3.333	6.629	1.676	3.931	3.333	8.94

2.4.1.2. Ritonavir

The fragments of ritonavir after decomposition are shown in Figure 2.3. Here, the application of DAIM resulted in 17 fragments. Among them were also fragments such as methyl, phenyl, thiazole groups, which were unsuitable for reasonable subcavity definition, and therefore discarded. The similarity scores are presented in Table 2.2.



(a)

(b) (c) (d) (e)

Figure 2.3.: Fragments of the inhibitor ritonavir (**a**) are shown. **b–e** are computed by RECAP.

Table 2.2.: Similarity scores of the Ligsite, ritonavir, and the respecitve fragment defined cavities of endothiapepsin vs. HIV protease (3PRS vs. 1RL8). The formyl fragment Figure 2.3e was discarded for the cavity comparison, due to its small size.

Ligsite	Complete	RECAP			
cavity	ligand	frag 1	frag 2	frag 3	$\sum$
9.606	7.415	6.405	2.304	1.641	10.35

2.4.1.3. Lovastatin

The fragments of lovastatin after decomposition are shown in Figure 2.4. The similarity scores are presented in Table 2.3.

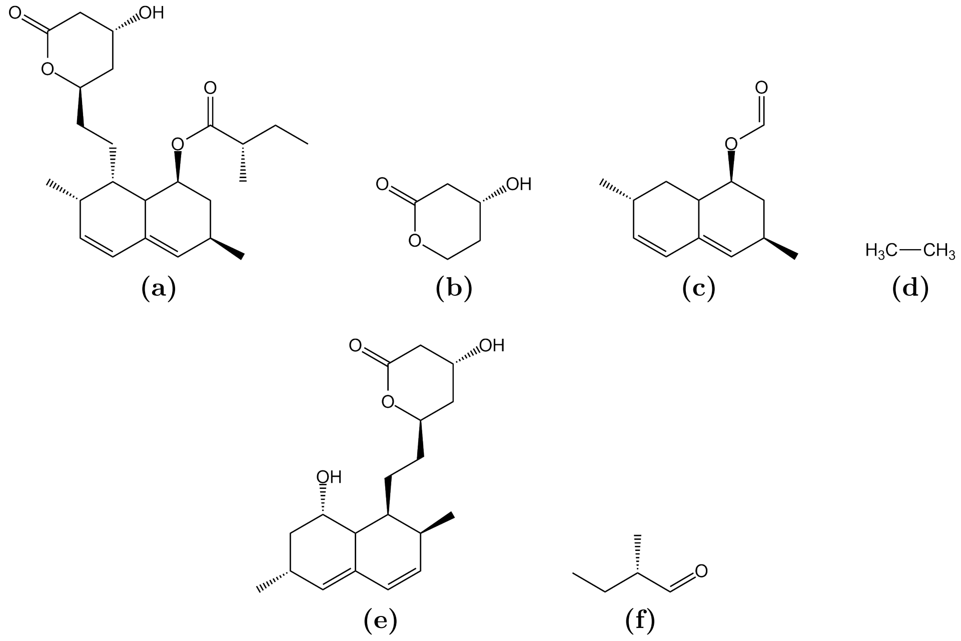


Figure 2.4.: Fragments of the inhibitor lovastatin (**a**) are shown. **b–d** are fragments generated by DAIM rules. **e, f** are computed by RECAP.

Table 2.3.: Similarity scores of the Ligsite, lovastatin, and the respecitve fragment defined cavities of LFA-1 vs. HMG-CoA reductase (1CQP vs. 1HW9). The ethyl moiety (Figure 2.4d) was discarded for the cavity comparison, due to is small size.

Ligsite cavity	Complete ligand	RECAP			DAIM		
		frag 1	frag 2	$\sum$	frag 1	frag 2	$\sum$
3.517	2.531	2.226	2.361	4.587	0.914	2.396	3.31

2.4.2. Database screening

The introduced cavity decomposition strategy was applied on a database containing about 275,000 cavities. COX-2 was used as query. The respective EC number is 1.14.99.1.

Proteins from diverse EC classes were selected for the comparison of the emerging rankings when different screening setups were tested. The respective EC numbers are listed in brackets, if available.

- α-CA II (4.2.1.1),
- α-carbonic anhydrase V (α-CA V) (4.2.1.1),
- α-carbonic anhydrase XIII (α-CA XIII) (4.2.1.1),
- 70 kDa heat shock protein (HSP70) (3.6.1.3),
- 90 kDa heat shock protein (HSP90) (3.6.1.3),
- peroxisome proliferator-activated receptorγ (PPARγ), and
- 3-phosphoinositide-dependent protein kinase 1 (PDK-1) (2.7.11.1).

In the following result tables only the top ranked protein belonging to identical EC class are shown, which is the case for classes 4.2.1.1 and 3.6.1.3.

Results from the pairwise comparison studies suggested that matched subpockets of the target protein can be found rather distantly from each other in the candidate proteins, hence the ligand present in the query pocket will hardly bind simultaneously to all the matched subpockets in the candidate proteins. Therefore, we included a routine that checks whether binding of the original query ligand would be geometrically feasible in the hit candidate pockets. The centers of each subpocket found in the candidate proteins were determined based on the coordinates of the matched pseudocenters. If the distances between these centers fall below a given threshold, then scores of the subpockets are added, otherwise the scores are discarded. The applied threshold limits depend on the geometry of the original ligand. Therefore, the distance between the farthest ligand atoms is considered. The resulting scores and the subsequently generated ranking only includes those proteins which were able to match all defined subcavities with appropriate mutual distances.

Table 2.4.: Pocket database screening using Ligsite cavity and complete ligand as queries.

Rank	%	Score	Protein name	PDB id
Ligsite cavity. Top 40 are COX entries.				
1023	0.379	7.933	α-CA XIII	3D0H
106	0.039	9.996	PPARγ	2G0H
233	0.086	9.913	HSP70	3FZF
45387	16.810	6.047	PDK-1	2PE0
Complete ligand. Top 37 are COX entries				
625	0.213	6.622	murine α-CA V	1DMX
1397	0.517	6.160	PPARγ	3IA6
227	0.084	7.485	bovine HSP70	1BA0
85589	3.181	5.218	PDK-1	2PE0

Table 2.5.: Pocket database screening using modified DAIM rules.

Rank	%	Score	Protein name	PDB id
DAIM setup, no distance check. Top 35 are COX entries.				
599	0.222	11.542	α-CA II	1ZFK
59	0.022	13.188	PPARγ	2VV3
30	0.014	13.863	bovine HSP70	1BA0
1358	0.503	11.055	PDK-1	2PE0
DAIM setup, with distance check. Top 32 are COX entries.				
252	0.093	11.542	α-CA II	1ZFK
38	0.014	13.190	PPARγ	2VV3
82	0.030	12.312	bovine HSP70	1KAY
563	0.209	11.055	PDK-1	2PE0

Table 2.6.: Pocket database screening using RECAP rules.

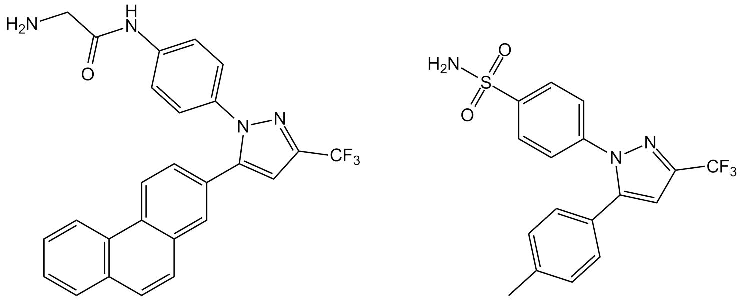
Rank	%	Score	Protein name	PDB id
RECAP setup, no distance check. Top 37 are COX entries.				
2054	0.761	9.331	α-CA V	1DMX
167	0.061	10.864	PPARγ	2VV3
66	0.024	11.678	bovine HSP70	1BA0
5538	2.051	8.797	PDK-1	2PE0
RECAP setup, with distance check. Top 35 are COX entries.				
1207	0.447	9.331	α-CA V	1DMX
96	0.035	10.864	PPARγ	2VV3
119	0.044	10.731	HSP90	3K99
3423	1.268	8.797	PDK-1	2PE0

2.5. Discussion

One of the first aspects we can learn from the pairwise comparisons of subcavities is the impact that the ligand fragmentation strategy takes on the achieved results. First, fragmentation according to the used strategies does not lead in all cases to a reasonable complementarity of the protein binding site with respect to the structurally determined subcavities. Second, too small fragments are inadequate for a subcavity definition, since they result in the selection of subcavities comprising rather a small subset of the original pseudocenters, and therefore no relevant distribution of the exposed physicochemical properties is any longer given. These two observations indicate that an automated cavity decomposition procedure requires a compilation and the subsequent evaluation of rules with particular focus on fragmentation of ligands bound in protein pockets.

Concerning the screening of COX-2 subpockets defined by SC-558, a significant improvement could be observed for α-CA II ranking. Using the original Ligsite cavity from COX-2 as query the matched α-CA II cavity was found on rank 10,036. Restriction of the Ligsite cavity using contiguous spheres of 4.5 Å around all ligand atoms lead to rank 5,107. Modifying the DAIM rules, the composite approach based on subcavity ranking moved the α-CA II as candidate hit on rank 599 and after the distance compatibility check on position 252. In case of RECAP α-CA II is found at rank 3302 and 1977 respectively. The following conclusions can be drawn. The definition of fragments plays a crucial role for the scoring and ranking, which explains the discrepancy between the performance of the modified DAIM and RECAP rules. Furthermore, the strategy of checking the distance compatibility of the matched subcavities with respect to each other improves the ranking of α-CA II as a hit in both cases. Unfortunately, we were unable to solve the case, where different subpockets overlap partly the same region in the target pocket, therefore, leading to a bias toward a higher score. This drawback could be potentially solved by computing a penalty term for these cases or by screening the queries against a database of subpockets as targets.

Most notable and unexpected result is the ranking of the PDK-1 as a hitted candidate pocket for celecoxib on rank 563 using modified DAIM rules. In the contrary, Ligsite but also the ligand defined cavity definition have only scored this protein at position 45,387 and 85,387 respectively. According to the Drug Bank database (Wishart et al., 2006) PDK-1 is listed as a second target for celecoxib apart from COX-2. Arico et al. used anti-PDK1 immunoprecipitates derived from human colon carcinoma cell line 29 (HT29) to measure the inhibition of the Ser/Thr kinase activity. They found that PDK-1 is inhibited by celecoxib at the half-maximal inhibitory concentration (IC_{50}) of 3.5 μM (Arico et al., 2002). Zhu et al. reported an IC_{50} of 48 μM measured in an enzyme assay using the recombinant PDK-1 protein. Subsequent structure-based optimization cycles lead to the celecoxib analog OSU-03012 that inhibits PDK-1 kinase activity in low micromolar range (IC_{50}=5 μM, (Zhu et al., 2004)).



(a) **(b)**

Figure 2.5.: Structure of the PDK-1 inhibitor OSU-03012 (a) and celecoxib (b).

Interestingly, PPARγ and HSP70 are also found on high ranks suggesting that these proteins are potential binding partners for celecoxib as well. This hypothesis should be validated experimentally. The low solubility of celecoxib in water and most likely its weak affinity to the suggested proteins has to be considered in the experimental setup to allow for sufficient sensitivity in the assay.

Part II.

Virtual Screening for Novel Molecules with Antimalarial and Antibacterial Activity

3

Chapter 3.

Novel FAS II Inhibitors as Multistage Antimalarials

3.1. Preliminary remarks

This chapter has been prepared as part of a contribution for a scientific journal. Meanwhile, it has been accepted for publication in ChemMedChem. Dr. Florian Schrader performed the synthesis of the compounds. In a collaboration with the Department for Infectious Diseases and Parasitology Unit at the Heidelberg University Hospital the compounds were tested for their growth inhibition of the *Plasmodium* parasites in cell-based assays. The cytotoxicities to human cell lines were measured at the Department for Infection Biology of Hans-Knöll-Institute in Jena. *Tg*ENR enzyme inhibition values were assayed at the Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA. My contribution to this project was to carry out the initial virtual screening and hit identification. Furthermore, I supported the optimization of the lead compound using combinatorial and docking techniques.

3.2. Introduction

Over two billion people in about 100 countries run the risk of falling ill with malaria, a disease, which often turns out to be fatal for young children (Murray et al., 2012). Their chances to survive depend not least on the development of new, affordable drugs, as increasing resistance against the currently used drugs is observed. Malaria is caused by the protozoan *Plasmodium* and is transmitted by the *Anopheles* mosquito. When this insect feeds on an infected individual, it may ingest *Plasmodium* gametocytes along with the blood meal. These represent the only stages in the life cycle of *Plasmodium*, which are infectious to the mosquito. Inside of the mosquito host *Plasmodium* sporozoites are subsequently formed. These motile stages travel to the salivary gland of the mosquito and may be injected along with its saliva while feeding on another human host. There, *Plasmodium* sporozoites migrate actively into the circulation and finally end up in the liver, where they infect hepatocytes. It takes up to 7-10 days for *P. falciparum* sporozoites to subsequently develop into several thousand first-generation merozoites (Sturm et al., 2006). As a consequence, the liver-stage is therefore characterized by huge metabolic demands. Merozoites are then released into the blood to begin their pathological blood-stage development by infecting erythrocytes. Subsequently, the patient begins to suffer from symptoms like fever, pain and nausea (WHO).

Replicating in the liver as well as in erythrocytes, *Plasmodium* parasites require vast amounts of fatty acids (FA). Metabolism in these two stages of the life cycle, however, is fundamentally different (Yu et al., 2008; Tarun et al., 2009). In the blood-stage, the vast majority of the FAs are acquired from the host (Vial et al., 1982). In earlier studies it was assumed that *Plasmodium* acquires FAs merely by scavenging, however, it was recently found to be capable of type II fatty acid biosynthesis (FAS II). FAS enzymes are targeted to the apicoplast, a relict, non-photosynthetic plastid of algal origin. In most plants as well as in bacteria, discrete enzymes catalyze the distinct steps in plasmodial FAS II (Figure 3.1). In contrast to this,

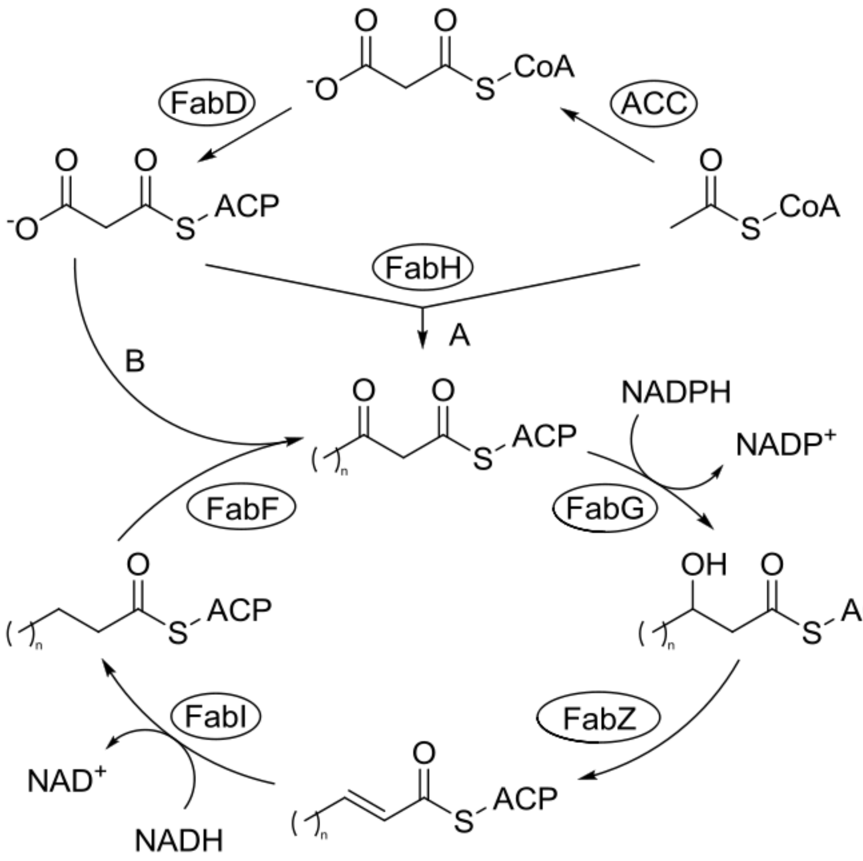


Figure 3.1.: Type II fatty acid biosynthesis. The most vital precursor to fatty acids in *Plasmodium* is acetyl-CoA, which is provided by acetyl-CoA-Synthase or pyruvate dehydrogenase. Fatty acid biosynthesis (FAS) begins with the carboxylation of acetyl-CoA by acetyl-CoA carboxylase (ACC). The resulting product malonyl-CoA is then converted to malonyl-ACP by the malonyl-CoA ACP transacylase (FabD). The acyl carrier protein (ACP) is a small, acidic protein which binds acyl intermediates as thioesters during fatty acid synthesis and carries them from one to the other enzymes. The first reaction is a condensation catalyzed by β-ketoacyl-ACP synthase III (FabH), which uses acetyl-CoA and malonyl-ACP as substrates. Next is a NADPH-dependent reduction of β-ketoacyl-ACP to β-hydroxyacyl-ACP catalyzed by β-ketoacyl-ACP reductase (FabG). β-hydroxy-acyl-ACP dehydratase (FabZ) then forms trans-2-enoyl-ACP by removing a molecule of water from the acyl chain. Trans-enoyl-ACP is finally NADH-dependently reduced by enoyl-ACP reductase (FabI), which presents the rate limiting step in fatty acid chain elongation. Another molecule of malonyl-ACP may subsequently be used to add two more carbon atoms to the nascent acyl chain by β-ketoacyl-ACP synthase II. This cycle continues with the length of the acyl chain increasing by two carbons until the desired fatty acid is produced.

in mammals FAS is performed by a multi enzyme complex, handling all four of the enzymatic steps of the elongation of fatty acids (type I fatty acid biosynthesis (FAS I)). Although there is no difference in mechanism in the elongation of fatty acid chains, this fundamentally distinct setup of enzymes makes FAS I insensitive to a number of FAS II inhibitors and qualifies fatty acid biosynthesis in *Plasmodium* as a potential drug target.

Triclosan is an antibacterial and antifungal agent, commonly used in a large variety of consumer products such as toothpastes and pillowcases. In 1998 it was first shown to be an *E. coli* FAS inhibitor and to specifically inhibit *E. coli* enoyl ACP reductase (ENR) (McMurry et al., 1998). The contemporary discovery of the plastidial origin of the apicoplast and its suggestion as a drug target (McFadden et al., 1996; Kohler, 1997) prompted efforts to assay Triclosan as an antiplasmodial agent. In the following Triclosan proved to inhibit *Pf*ENR with a K_i of 0.4 nM (Kapoor et al., 2001) and, in addition to that, the growth of blood-stage *P. falciparum* at a low micromolar concentration (Surolia and Surolia, 2001; McLeod et al., 2001). Subsequent work by Yu et al. showed that Triclosan inhibits another essential target in blood stage parasites: Disruption of the gene encoding *Pf*ENR did not affect parasite growth or Triclosan susceptibility.

Similarly, other inhibitors of plasmodial FAS, and *Pf*ENR inhibitors in particular, frequently do show an inhibitory effect on blood-stage cultured parasites as well (Tasdemir et al., 2006, 2007). Hence it is tempting to speculate on a common off-target (Vaughan et al., 2009; Spalding and Prigge, 2008). The genome of *Plasmodium* appears to encode for three different fatty acid elongases (ELO). In contrast to FAS I and FAS II, ELO pathways use CoA rather than ACP as an acyl carrier. Importantly, ELO pathways contain an enoyl-CoA reductase (EnCR) which catalyzes a similar reaction to that of *Pf*ENR. Although ELO pathways typically elongate long-chain FAs such as palmitate (Lee et al., 2006b), trypanosomes were recently shown to synthesize most of their FAs from butyryl-CoA precursors (Kohlwein et al., 2001).

Type II fatty acid synthesis, and ENR in particular, has been shown to play a key role in the development of liver-stage malaria parasites (Yu et al., 2008; Vaughan et al., 2009). ENR-deficient *P. berghei* sporozoites are markedly less infective to mice and typically fail to complete liver-stage development *in vitro.* This defect is characterized by an inability to form intrahepatic merosomes, which normally initiate blood-stage infections. Even though it is not clear how Triclosan and other FAS II inhibitors act upon blood-stage parasites, present data suggest that FAS II inhibitors may provide true causal chemoprophylaxis and could simultaneously cure blood-stage Malaria (Yu et al., 2008; Vaughan et al., 2009).

3.3. Results and discussion

3.3.1. Virtual screening

In an effort to find structurally novel, potent inhibitors, we performed a virtual screening based on two different *Pf*ENR crystal structures (PDB codes: 2O2Y, 2OOS), as the structures indicate that the active-site residue Phe368 can adopt two alternative conformations (Figure 3.2).

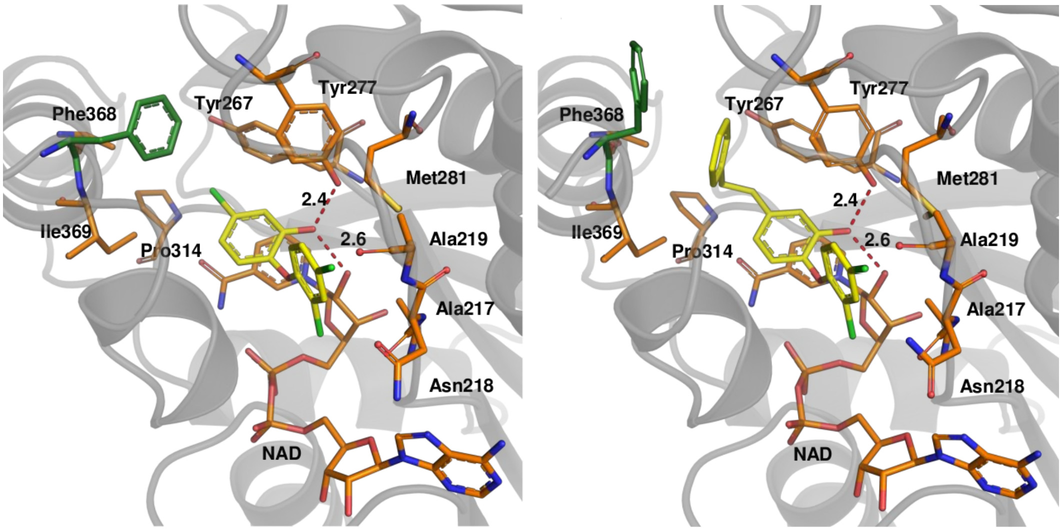


Figure 3.2.: Crystal structures of *Pf*ENR and NAD$^+$ with bound inhibitor Triclosan (PDB 2O2Y) (a), and a 5-substituted Triclosan derivative (PDB 2OOS) (b). Introduction of a large hydrophobic substituent at the 5-position of Triclosan induces a conformational transition of Phe368. Hydrogen bond interactions of the inhibitor's phenolic OH group to Tyr277 and 2'-hydroxyl group of the nicotinamide ribose are depicted as red dashed lines. For reasons of clarity Ile323 and Val222 are not shown.

The *Pf*ENR binding site hosts an NAD$^+$ molecule, which was retained during docking as integral part of the pocket. A total of 13,200 compounds retrieved from an in-house fragment-like library (Table 3.1) were docked into the respective pocket using GOLD (Jones et al., 1997). Results were re-ranked applying the DSX scoring function (Neudert and Klebe, 2011a).

Table 3.1.: Properties of the fragment-like library used for screening.

Property	Min	Max
No. of heavy atoms	8	20
Molecular weight (Dalton) (MW)	122	360
Lipinski donor	0	4
Lipinski acceptor	1	8
Calculated log of the octanol/water partition coefficient (clog$P$)	-1.2	7.6
Free rotatable bonds	0	7
Total polar surface area (Å^2) (TPSA)	12	126

Eight chemically diverse hits were selected, which satisfied the pattern of interactions supposed to be essential for inhibitor binding as indicated by the reference crystal structures (Perozzo et al., 2002). Docked compounds were requested to exhibit stacking interactions with the nicotinamide moiety of NAD$^+$ and to form hydrogen bonds to Tyr277 and the 2'-hydroxyl group of the nicotinamide ribose. The achieved respective docking poses are depicted in Figure 3.3 a-h.

Out of the set of compounds derived from the virtual screening (VS), eight molecules were selected with respect to sufficient drug-likeliness, chemical diversity and synthetic accessibility and easy scope of variation. In a first step, minor modifications were performed to the chemical structure of these eight promising screening hits in order to overcome obvious metabolic instability or to facilitate convenient synthesis (Figure 3.4). The intended modifications were validated by subsequent docking whether binding to the target protein and consistency with the derived pharmacophore were still fulfilled.

With the outlined modifications the eight target compounds (Figure 3.4) were synthesized by Dr. Florian Schrader (Schrader, 2012) and subsequently tested for biological activity (subsection 3.3.2).

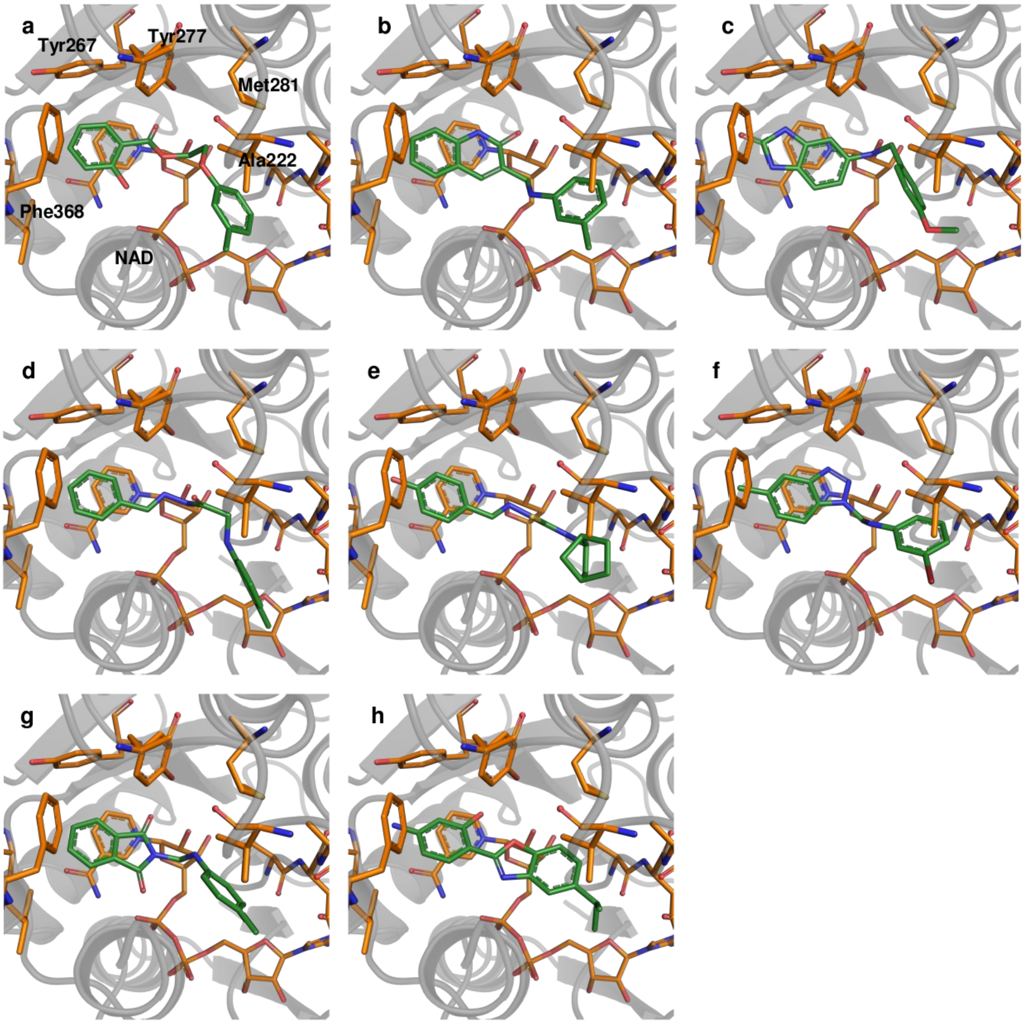
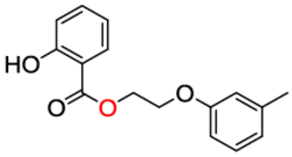
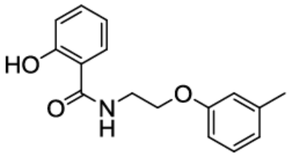
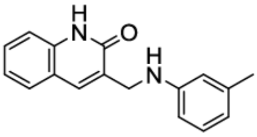
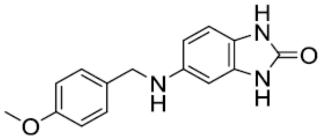
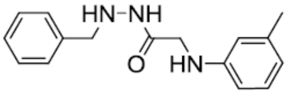
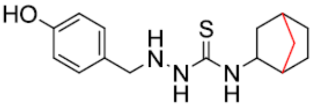
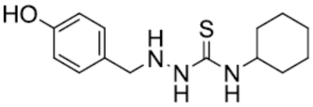
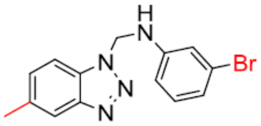
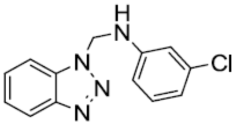
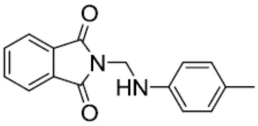
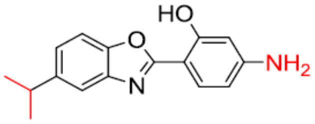
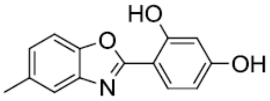


Figure 3.3.: Docking poses of the eight selected most promising hits from virtual screening (a-h). All compounds comprise an aromatic moiety, which is able to establish stacking interactions with the nicotinamide portion of NAD$^+$. Hydrogen bonding to Tyr277 and/or the 2'-hydroxyl group of the ribose can be formed by a carbonyl oxygen (a, b, g, h) or by nitrogen atoms (c-f).

	Suggested compound		Synthesized compound	IC 50 *P. falciparum* (μM)	CC_{50} (μM)^a
0a		1a		7.9 ± 0.3	59.3
1b			as suggested	> 50	< 189.2
1c			as suggested	> 50	< 185.7
1d			as suggested	> 50	< 185.6
0e		1e		> 50	136.4
0f		1f		> 50	< 185.7
1g			as suggested	> 50	159.2
0h		1h		> 50	26.11

^{a} Cytotoxicity (HeLa cells)

Figure 3.4.: The eight selected VS hits and the respective chemical modifications.

3.3.2. Biological evaluation and structure–activity relationship

The inhibitor-binding pocket of both *Pf*ENR and *Toxoplasma gondii* ENR (*Tg*ENR) are highly conserved with only one amino acid difference (Figure 3.5, Muench et al. (2007)). We therefore used the *Toxoplasma* homolog for the enzyme inhibition assay.

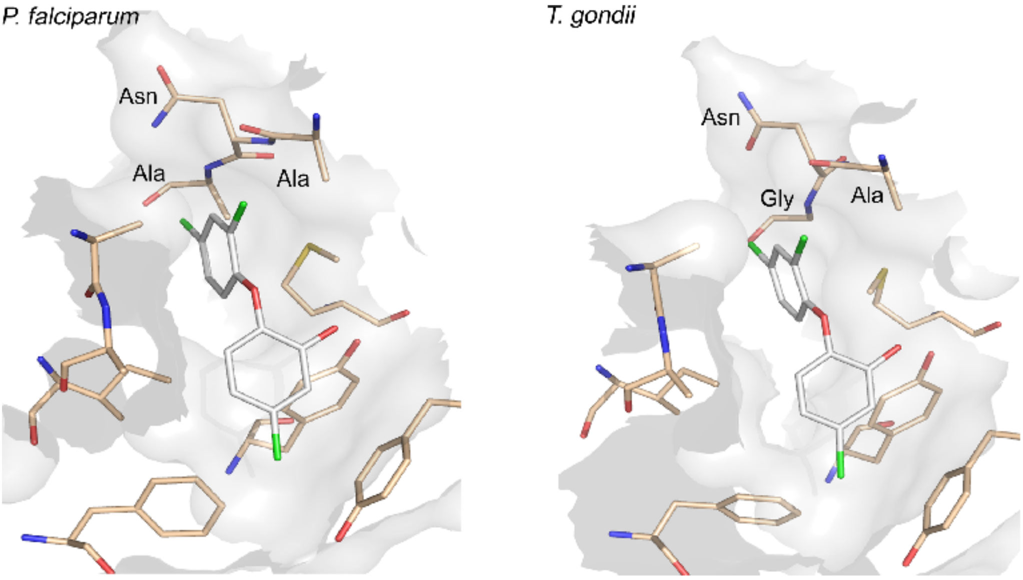
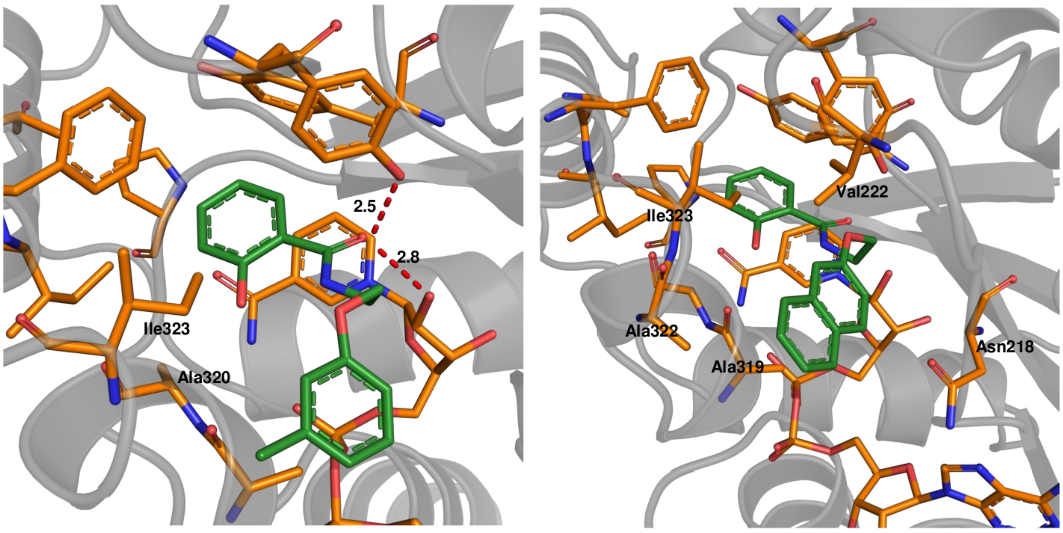


Figure 3.5.: Inhibitor binding sites of the *Pf*ENR and *Tg*ENR Enzymes

Activity against the asexual blood-stage is considered to be a prerequisite for antimalarial drug activity. Therefore, all synthesized compounds were first tested for their inhibitory activity against cultured blood-stage *P. falciparum* (multidrug-resistant Dd2 isolate). This way, the compounds could also prove themselves to be sufficiently cell-permeable, which poses a critical hurdle to overcome by appropriate drug design of compounds targeting living parasites intra-cellularly.

Out of our eight promising hits from virtual screening the aryloxyalkylbenzamide derivative inhibited the growth of bloodstage Dd2 malaria parasites in the cell-based assay with an IC_{50} of 7.9 μM (Figure 3.4). The inhibition of the *Tg*ENR enzyme was rather poor at a 1 μM concentration (10 %). Nevertheless, in the following experiments we tried to evaluate whether this scaffold could be chemically optimized to display an inhibitor of blood-stage *Plasmodium* and ENR at the same time.

The respective docking pose of **1a** suggests that its polar 2-OH-substituent does not specifically interact with the inhibitor binding site of ENR (Figure 3.6a). At this position, non-polar substituents might lead to an enhanced affinity toward the enzyme as they could potentially interact with the hydrophobic amino acids Ile323 and Ala320. We therefore synthesized derivatives **6a** and **5e**, which either bear such a non-polar 2-substituent or lack this group. Interestingly, both derivatives inhibit the *Tg*ENR enzyme as depicted in Figure 3.7.



(a) (b)

Figure 3.6.: Docking pose for (a) **1a** and (b) **5j**.

In the cell-based assay, removal of the 2-substituent (**5e**) results in a three-fold decline of activity (24.3 μM). However, the 2-chloro-derivative (**6a**) shows no reduction in activity against blood-stage parasites (7.5 μM) and displays an even lower cytotoxicity (>172.6 μM compared to 59.3 μM for the 2-OH substituted derivative **1a**). As the 2-chlorobenzoic acid

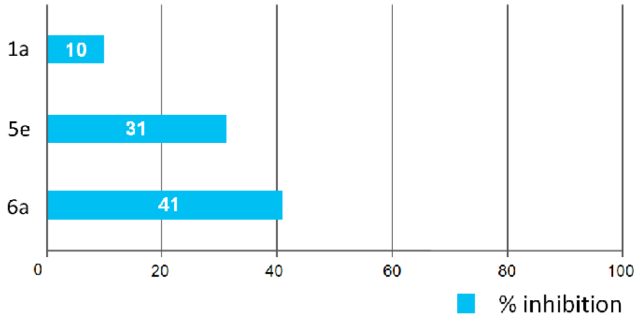


Figure 3.7.: Inhibition of *Tg*ENR at a concentration of 1 μM of derivatives of **1a**

derivative inhibits blood-stage malaria parasites and the *Tg*ENR as well, it qualifies as a lead structure for subsequent optimization.

Initially, we evaluated whether selected heteroatoms are mandatory in this compound. We subsequently varied the apparently important 2-substituent introducing a fluoro- and trifluoromethyl group. For the 2-fluoro (**6i**, 10.5 μM) and 2-trifluoromethyl derivates (**6j**, 43.1 μM) there was no considerable improvement in the cell-based assay. We also prepared a derivative with a 3-chloro-substituted benzoic acid moiety (**6l**) instead of attachment at the 2-position. Compared to the 2-chloro substituted derivative (**6a**) this change led to a more than three-fold reduction of activity (26.6 μM) in the cell-based assay. These data stimulated us to the assumption that the 2-position of the benzoic acid moiety is optimal for substitution. Apparently, the substituent does not necessarily act as an H-bond donor or acceptor as either the OH or Cl substituted derivatives show equipotent inhibition. We prepared the sulphur homolog (**6d**) by using thiocresolate as a nucleophile for ether synthesis. As there was about a four-fold reduction in activity (32.0 μM) in the cell-based assay we kept the original composition.

Based on the 2-chloro benzoic acid moiety a series of derivatives have been prepared to optimize the lead structure. Docking of **6a** suggested a hydrogen bonding interaction to the hydroxy group of Tyr267 within the binding site. Therefore, a series incorporating hydrogen bond acceptor and/or donor functionalities in 3- and 4-position of the benzoic moiety have been synthesized and tested. None of these compounds showed improved activity in the cell-based assay compared to the original compound. Strikingly, introduction of a polar group in almost all cases led to a complete loss of activity in the cell-based assay. The only derivatives which at least partially maintained activity was the 4-amino (**6q**, 10.5 μM) and the 4-hydroxy derivatives (**6t**, 17.8 μM). We therefore abandoned the idea of introducing polar groups at the 4-position. In an attempt to increase activity through an entropic gain in binding, we conformationally rigidified the molecule by incorporating a fixed linker between the aromatic rings as a 3-substituted azetidine (**6c**, 32.6 μM) and a 3-subsitituted piperidine (**6b**, >50 μM) heterocycle. As these modifications were clearly detrimental to activity in the cell-based assay, it seems that the most favorable conformations required for binding of the inhibitors are no longer easily accessible by incorporation of the two tested rigid linkers. This prompted us to remain with the original oxyethylamide moiety as a linker.

3.3.3. Combinatorial library

To elaborate the chemical space and the potential of aryloxyalkylbenzamide derivatives, a combinatorial virtual library was generated considering commercially available agents. This library contained about 430,000 molecules with a MW 600 Da. 10,000 compounds were pre-selected from the computed library using a coarse-grained docking procedure, followed by a second, more extensive docking run. The results were re-ranked by DSX and visually inspected. The most promising derivatives were selected for synthesis and subsequent biological testing. The top-ranked compound comprised a naphthaleneoxy instead of the m-tolyloxy moiety. This derivative displayed the same interaction patterns as the initial compound (**1a**) but the increased molecular surface of the naphthyl moiety provided more van der Waals interactions to hydrophobic residues Val222, Ala319, Ala322, Ile323, and additionally Asn218 (Figure 3.6b).

Docking into the crystal structure suggested the oxyaryl moiety of a 4-isopropylbenzoxy- (**6e**) a 5,6,7,8-tetrahydronaphthaleneoxy- (**6f**) and a naphthaleneoxy derivative (**6g**) to interact with lipophilic amino acid residues. We synthesized these compounds and observed a decline in activity in the cell-based assay for the isopropylbenzeneoxy derivate (22.0 μM). In contrast, for the tetrahydronaphthyloxy derivate (2.5 μM) and the naphthyloxy (1.7 μM) we observed a three-fold, respectively more than four-fold improvement in the cell-based assay. At a 1 μM concentration, the naphthaleneoxy derivative showed slightly higher inhibition (58 %) in the *Tg*ENR assay than the original m-tolyloxy derivative (**6a**, 41 %). In addition, it displayed a selectivity index (SI) of more than 50 (Figure 3.10).

Encouraged by this fact, we also prepared a derivative that included both the naphthoxy- and the salicylic acid moiety connected by the ethylamine linker (**5j**). This compound is the most active derivate in our series of salicylic acid amides in the cell-based assay (3.0 μM). It shows 58 % inhibition of *Tg*ENR at 1 μM concentration and displays a SI of more than 50. In order to confirm that this improvement can be attributed to a directional interaction with the target protein, we also prepared the 1-naphthoxy derivative (**5k**). This compound showed an eight-fold drop in activity (25.0 μM) compared with the 2-naphthoxy derivative. Therefore, **6g** and **5j** represent the most promising compounds in this series so far and were tested for plasmodial sporozoite- and liver-stage inhibition.

3.3.4. Evaluation of the effect on pre-erythrocytic parasites

In order to address the question of whether compound **6g** and **5j** have an effect on the motor machinery of the parasite - e.g. the infectious sporozoite - crucial for the invasion of liver cells, we first performed two-color host-cell invasion assays. Therefore, salivary gland sporozoites of *P. berghei* were applied to immortalized human hepatoma cell lines (HuH7) and allowed to

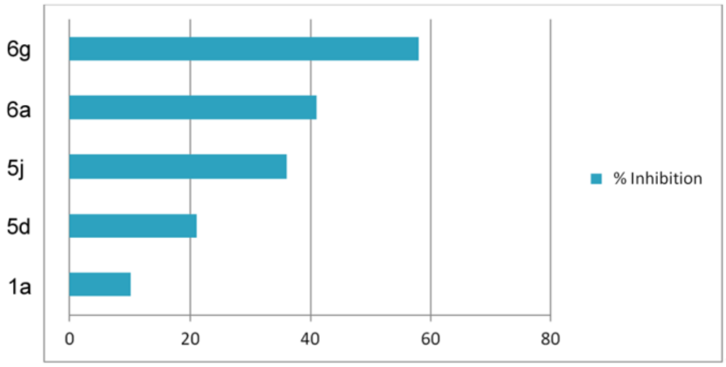
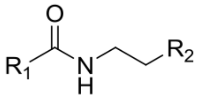
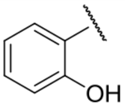
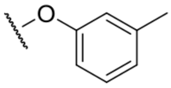
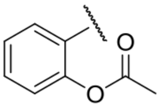
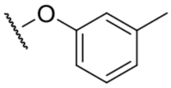
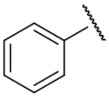
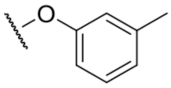
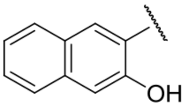
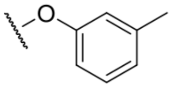
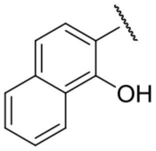
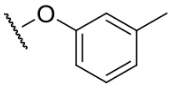
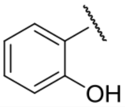
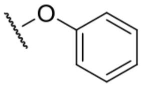
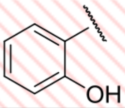
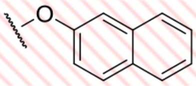
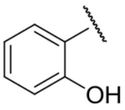
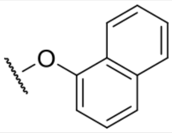


Figure 3.8.: Inhibition of *Tg*ENR at a concentration of 1 μM of structurally optimized derivatives of **1a**.

invade for 90 minutes in the presence of different concentrations of either compound **6g** or **5j**. Quantification of the numbers of extracellular versus intracellular parasites revealed that we could not observe any effect on the invasion of liver cells compared to our DMSO-treated control (Figure 3.11) (Aikawa et al., 1984; Tsuji et al., 1994).

Next we tested the inhibitory effect of the compounds on the development of the exoerythrocytic form (EEF) of the parasite (the clinically silent liver-stage) by applying **6g** and **5j** to the culture medium after invasion of infectious *P. berghei* sporozoites into HuH7 cells. After early- (24 h), mid- (40 h) and late- (60 h) intrahepatic development, cells were fixed and liver-stages were visualized by immunostaining of intracellular malarial HSP70 (Tsuji et al., 1994; Pinzon-Ortiz et al., 2001). Interestingly, the number of liver-stages by immunofluorescence microscopy was not significantly different in the various conditions tested (Figure 3.12a-c) except for a decrease in liver-stage numbers when compound **6g** was applied in a concentration of 40 μM (Figure 3.12a). However we have to take into consideration that due to the overall inhibitory growth effect in consequence of compound incubation, i.e. resulting in reasonable small liver-stages by size, we most likely failed to incorporate every single fluorescent liver-stage. Strikingly, when we measured the diameter, i.e. developmental status of the maturing



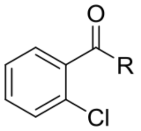
Cmpd.	R_1	R_2	IC_{50} *P. falciparum* (µM)	CC_{50} (µM)[a]	SI[b]
1a			7.9 ± 0.3	59.3	12.1
5c			7.7 ± 0.4	66.7	14.2
5e			24.3 ± 2.2	>195.8	8.0
5f			8.4 ± 0.4	133.5	15.9
5g			>50	73.1	ND
5d			13.4 ± 0.6	137.2	10.2
5j			3.0 ± 0.2	>162.7	>54.2
5k			25.0 ± 1.8	19.5	0.8

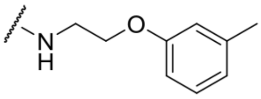
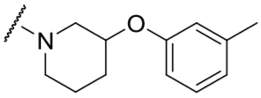
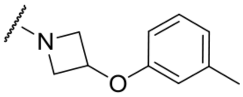
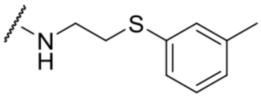
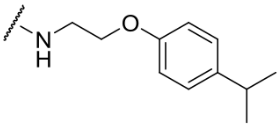
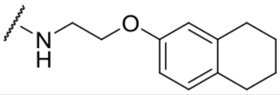
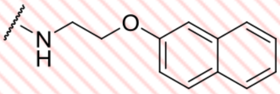
[a] **C**ytotoxicity (HeLa cells)

[b] SI = selectivity index = CC_{50} (HeLa)/IC_{50} (*P. falciparum*)

ND = not determined

Figure 3.9.: Structure and antimalarial activity of amide derivatives of salicylic acid.



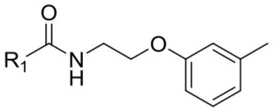
Cmpd.	R¹	IC $_{50}$ *P. falciparum* (µM)	CC$_{50}$ (µM)[a]	SI[b]
6a		7.5 ± 0.5	>172.6	>20.3
6b		>50	113.7	ND
6c		32.6 ± 1.7	127.9	3.9
6d		32.0 ± 1.9	60.8	1.9
6e		22.0 ± 0.5	105.4	4.8
6f		2.5 ± 0.1	60.1	24.4
6g		1.7 ± 0.1	89.6	52.7

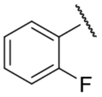
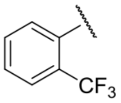
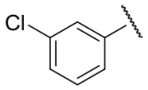
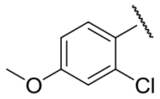
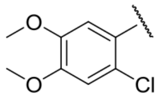
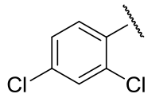
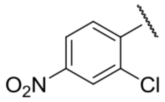
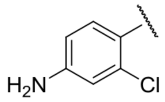
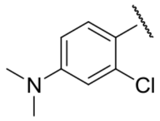
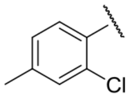
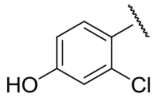
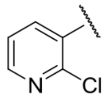
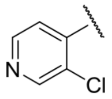
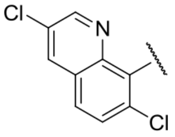
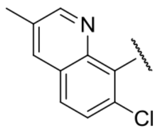
[a] Cytotoxicity (HeLa cells)

[b] SI = selectivity index = CC$_{50}$ (HeLa)/IC$_{50}$ (*P. falciparum*)

ND = not determined

Figure 3.10.: Structure and antimalarial activity of amide derivatives of 2-chloro-benzoic acid and various aryl amide derivatives (next page).



Cmpd.	R^1	IC$_{50}$ *P. falciparum* (μM)	CC$_{50}$ (μM)[a]	SI[b]
6i		10.5 ± 0.9	170.0	16.2
6j		43.1 ± 1.3	>154.7	>3.6
6l		26.6 ± 0.8	132.9	5.0
6m		>50	>156.5	ND
6n		>50	>142.9	ND
6o		>50	124.6	ND
6p		50.0 ± 0.8	>149.4	>3.0
6q		10.5 ± 0.8	>164.1	>15.6
6r		>50	117.5	ND
6s		>50	154.1	ND
6t		17.8 ± 1.0	119.4	6.7
6u		>50	>172.0	ND
6v		>50	>172.0	ND
6w		>50	>133.3	ND
6x		>50	>140.9	ND

[a] Cytotoxicity (HeLa cells)

[b] SI = selectivity index = CC$_{50}$ (HeLa)/IC$_{50}$ (*P. falciparum*)

ND = not determined

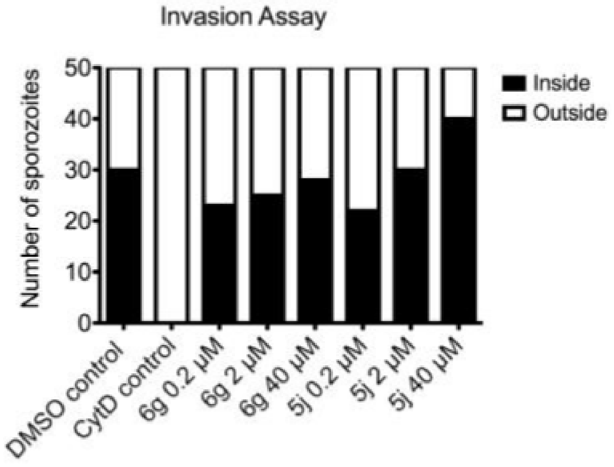


Figure 3.11.: **6g** and **5j** have no effect on the invasion of HuH7 cells *in vitro*. Infectious *P. berghei* sporozoites were pre-incubated with either 0.5 % DMSO, 10 μM CytochalasinD, or compound of interest at different concentrations (0.2 μM, 2 μM or 40 μM) and allowed to invade for 90 min under drug cover. After sporozoite double-staining 50 sporozoites were counted and classified as invaded or non-invaded.

intrahepatic liver-stages by confocal microscopy and subsequently image-processed with Zeiss Image Examiner we could clearly find that already at reasonable early time points after sporozoite inoculation at 24 h and **6g** treatment 40 μM caused a significant developmental delay compared to the control infection (Figure 3.12d; $p<0.0001$). More importantly, at later time points during liver-stage development we observed a significant attenuation of liver-stage growth even at lower concentrations of 2 μM for both compounds **6g** and **5j** tested (Figure 3.12e and f; $p<0.0001$). Interestingly, when compared to Primaquine (a member of the 8-aminoquinoline group of antimalarials exclusively active against the intrahepatic stages) at a standard inhibitory concentration of 10 μM, compounds **6g** and **5j** exert a more potent inhibition of malarial liver-stage growth with a calculated IC_{90} at 60 h of 2.79 μM and 3.14 μM, respectively.

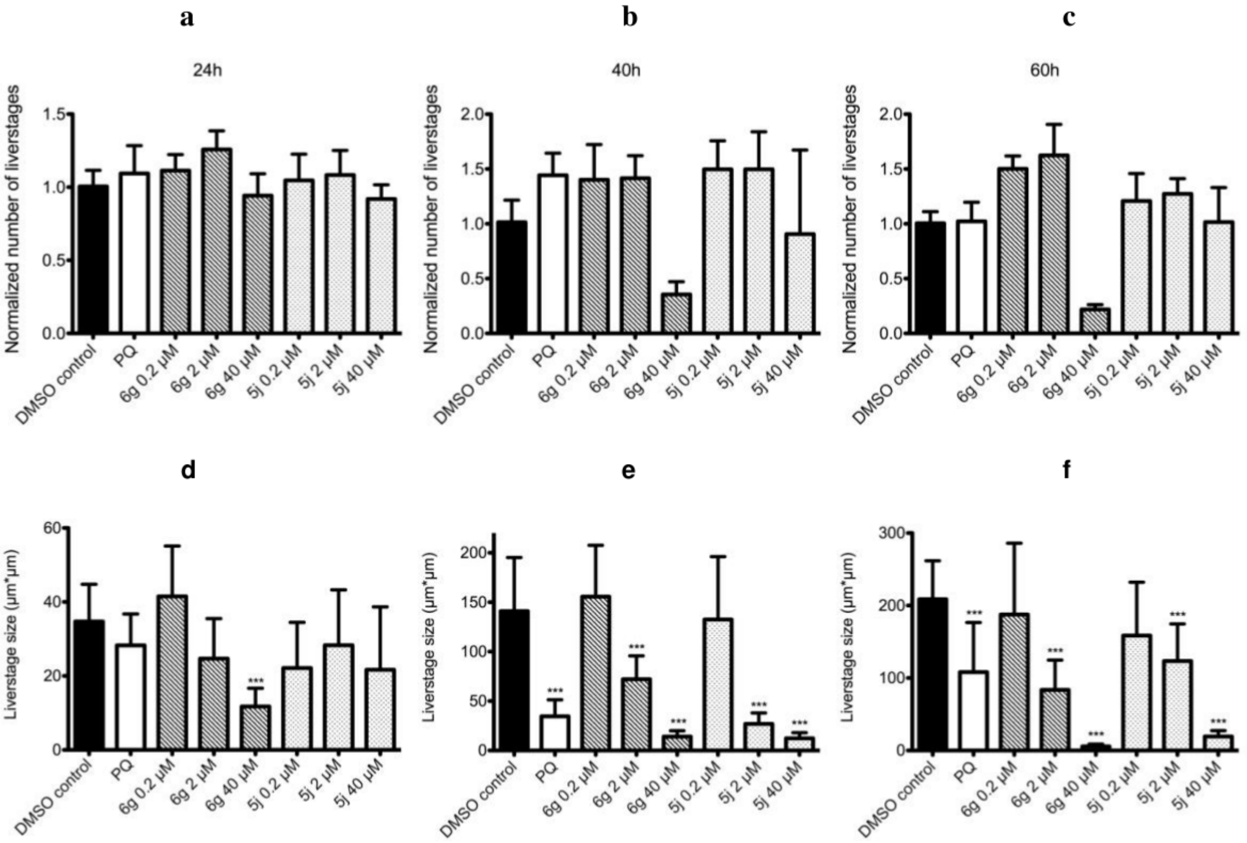


Figure 3.12.: Establishment of exoerythrocytic stages is not impaired but the development is significantly compromised *in vitro*. 24 h, 40 h or 60 h after infection of HuH7 with infectious *P. berghei* sporozoites cells were fixed, stained with anti-PbHSP70 and analyzed by immunofluorescence microscopy. We could not observe any difference in numbers of exoerythrocytic forms (a-c) only **6g** applied at a concentration of 40 μM caused a prominent reduction of liver-stage numbers at 40 h and 60 h after infection (b and c). Measurements of liver-stage growth over time revealed that already at 24 h p.i. 40 μM of **6g** caused a significant developmental delay compared to the DMSO control (d). At later time points even 2 μM of **6g** or **5j** could arrest the growth of exoerythrocytic stages significantly (e and f). PQ, Primaquine 10 μM; $p<0.0001$.

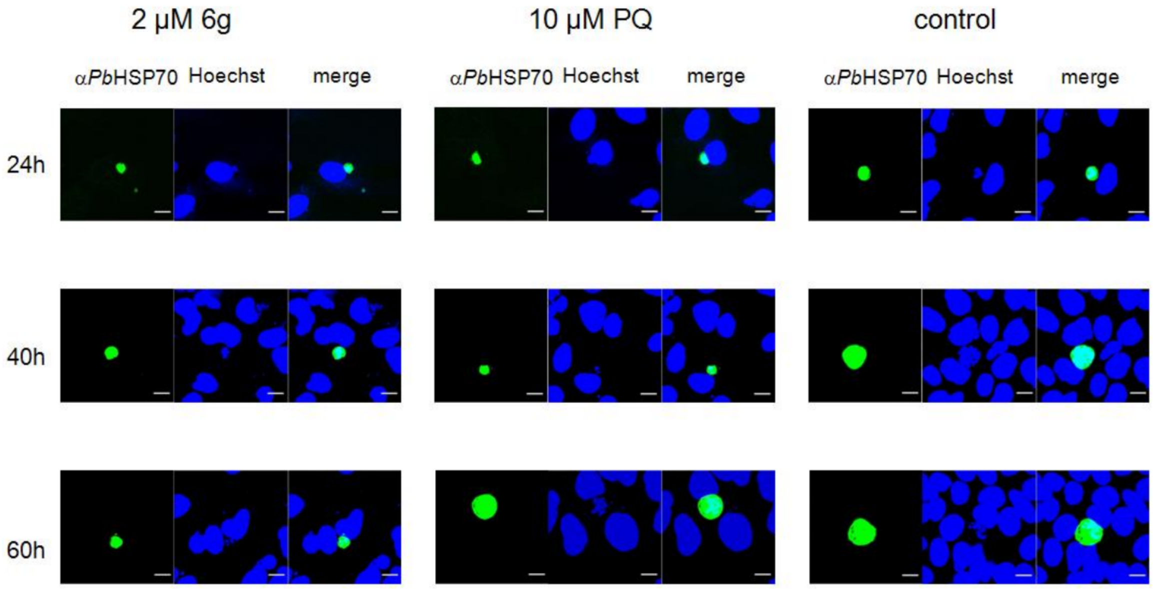


Figure 3.13.: Representative confocal pictures of malarial exoerythrocytic stages at 24 h, 40 h and 60 h after infection with infectious *Pb*ANKA sporozoites (scale bar 10 μM).

3.4. Conclusion

In summary, we showed that FAS II inhibitors qualify as antimalarial agents against pre-erythrocytic malarial parasites. In addition, compared to the gold-standard Primaquine, compounds **6g** and **5j** show at least a five-fold enhanced inhibitory effect on both, the development of clinically-silent liver-stages as well as disease-inducing erythrocytic blood-stage *P. falciparum* parasites. Due to their low cell-toxicity, these substances can be considered as most promising candidates to further evaluate their potency in *in vivo* experimental models. Altogether, this work provides evidence for novel concepts in chemical treatment of pre-erythrocytic malarial parasites and the pharmacological management of *Plasmodium* infections.

3.5. Materials and methods

3.5.1. Virtual screening and docking

The fragment-like library was designed according to the methodology reported by Köster et al. (Köster et al., 2011). Lipinski acceptor, clog$P$, TPSA were slightly modified in order to expand the chemical space of the library, in particular the number of lipophilic compounds was increased. All properties were calculated within the MOE software. A total of 13,200 ligands were minimized and protonated using the MMFF94x implemented in MOE. Ligands were docked into the binding sites of two PfENR X-ray structures (2O2Y, 2OOS) using the GOLD program. For the docking procedure solvent molecules were removed and the binding site was defined by a radius of 7 Åaround the respective inhibitor, considering NAD$^+$ as a part of the protein. Docking runs were performed applying 50 Genetic Algorithm (GA) runs, a search efficiency of 50 %, and the Astex Statistical Potential (ASP) scoring function (Mooij and Verdonk, 2005). For each compound 5 best-scored solutions were further evaluated. After clustering the poses according to root-mean-square-deviation (RMSD) < 2 Å, the results were locally minimized, re-scored and re-ranked by DSX using the per-atom-score as indicator. Top ranked 200 solutions were visually inspected. Graphical representations of protein-ligand interactions were prepared using PyMOL (Schrödinger, LLC, 2010).

For the creation of a combinatorial library of aryloxyalkylbenzamides the ZINC database was screened for commercially available derivatives of salicylic acid and bromoethylethers applying the *fconv* (Neudert and Klebe, 2011b) substructure search. We were able to retrieve 3068 salicylic acid and 190 bromoethylether non-redundant derivatives. The hydroxyl group of the carboxy moiety was defined as linker L1 and the bromine atom as linker L2, respectively. All fragments were connected *via* a nitrogen atom (Figure 3.14) using the combinatorial routine of CoLibri (BioSolveIT). The computed library of about 580,000 molecules was filtered using a

MW threshold < 450 Da resulting in nearly 110,000 compounds ready for screening. The parameters for the initial docking were set to 30 GA runs, search efficiency 30 %, ASP scoring function, and only the best-scored pose per compound was considered for further evaluation. The DSX per-atom-score was considered in the re-ranking procedure. Top 10,000 compounds were used for a next docking run. All next settings for the docking and the analysis of results were set to those described for the fragment-like library screening.

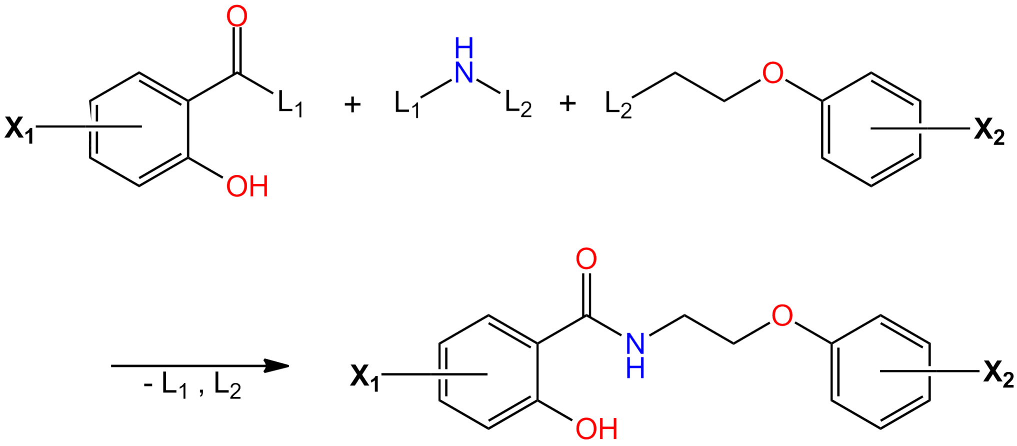


Figure 3.14.: Schematic illustration of the combinatorial library generation for salicylamides.

4

Chapter 4.

Virtual Screening for Pyruvoyltetrahydropterin Synthase Scaffolds

4.1. Introduction

4.1.1. Folate metabolism in *P. falciparum*

Tetrahydrofolate (THF) plays an important role in the metabolism of almost all living organisms. It is the major one-carbon carrier in cells and serves as a cofactor in methyltransferase reactions. Cellular processes, such as synthesis of nucleotides for DNA replication, synthesis of amino acids glycine and methionine, and metabolism of histidine, glutamic acid and serine, rely on the availability of THF (Ragsdale, 2008). THF is a crucial cofactor in metabolism, which in turn serves as a booster for growth and rapid cell division like those in tumors, bacteria, and malarial parasites. Hence, the folate biosynthetic pathway has been successfully exploited as a target in anti-cancer and anti-infective drug design (Nzila et al., 2005).

De novo synthesis of 7,8-dihydrofolate (DHF) (Figure 4.1), the precursor of THF, is absent in mammalian cells. On the contrary, plants, most bacteria as well as unicellular eukaryotes are able to synthesize THF. For unclear reasons, some parasitic protozoa such as *Plasmodium* and *Toxoplasma* have both, the folate biosynthetic and the salvage pathways (Hyde et al., 2008). The folate pathway is depicted in Figure 4.2 a.

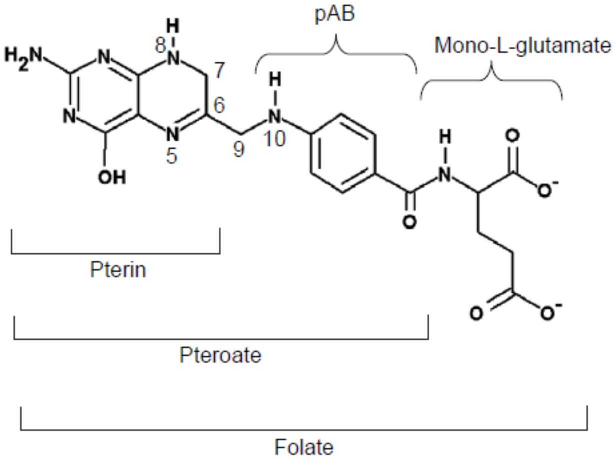


Figure 4.1.: Structure of DHF and its components. Pterin: 2-amino-4-hydroxy-7,8-dihydropteridine; pAB: p-aminobenzoic acid. Adapted from (Nzila et al., 2005)

4.1.2. Antifolates

Inhibitors of the folate biosynthetic pathway are effective agents for prophylaxis and treatment of malaria. They are classified into two classes: dihydropteroate synthase (DHPS) inhibitors (class I antifolates) and dihydrofolate reductase (DHFR) inhibitors (class II antifolates). DHPS inhibitors, such as sulfadoxine, sulfalene, and dapsone, are sulfur-based drugs and show as single drug administration only a weak antiparasitic effect, but they display in combination with DHFR inhibitors synergistic effects. Well known DHFR inhibitors are pyrimethamine, and the prodrugs proguanil and chlorproguanil that are used as monotherapy and in combination with other drugs. The brand names of antifolates and antifolate combinations are summarized in Table 4.1. More detailed information

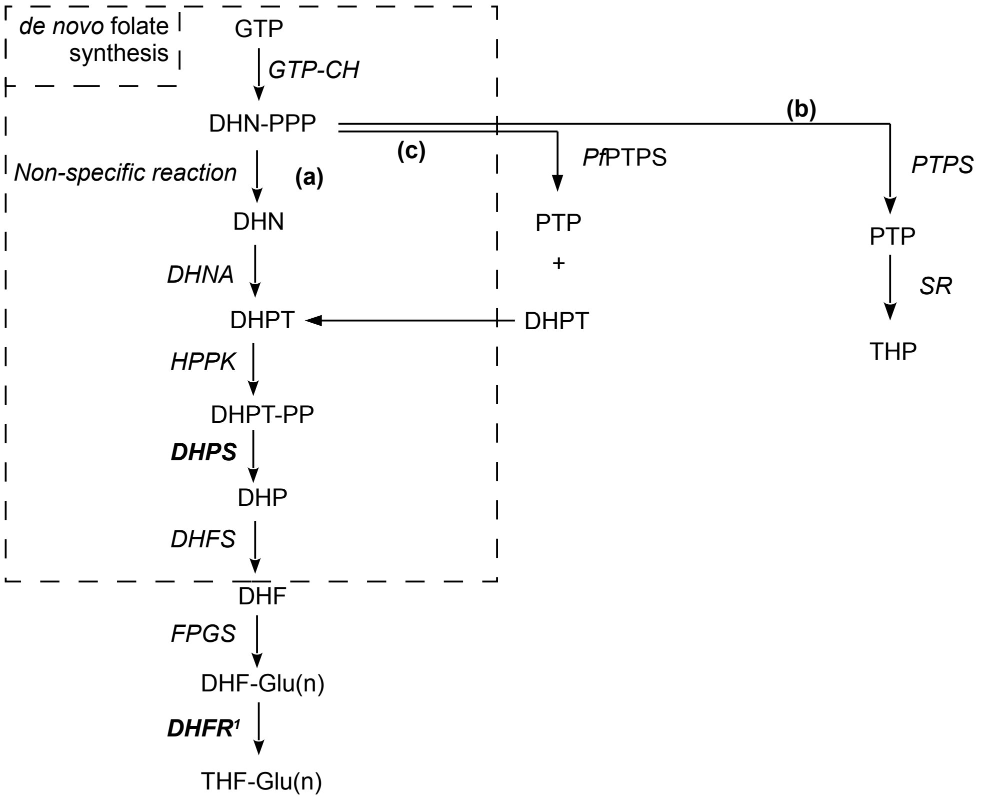


Figure 4.2.: Conventional folate (a) and biopterin (b) biosynthetic pathways. Pathway (c) shows the *Pf*PTPS catalyzed reaction. Abbreviations: GTP, guanosine triphosphate; GTP-CH, GTP-cyclohydrolase I; DHN-PPP, 7,8-dihydroneopterin triphosphate; DHN, 7,8-dihydroneopterin; DHNA, dihydroneopterin aldolase; DHPT, 7,8-dihydro-6-hydroxymethylpterin; HPPK, 6-hydroxymethyl-dihydropterin pyrophosphokinase; DHPT-PP, 7,8-dihydro-6-hydroxymethylpterin pyrophosphate; DHPS, dihydropteroate synthase; DHP, 7,8 dihydropteroate; DHFS, dihydrofolate synthase; DHF, 7,8-dihydrofolate; FPGS, folylpoly-gamma-glutamate synthetase; DHFR, dihydrofolate reductase; THF, tetrahydrofolate; Glu, glutamate; PTPS, pyruvoyl-tetrahydropterin synthase; PTP, 6-pyruvoyltetrahydropterin; SR, sepiapterin reductase; THP, tetrahydrobiopterin.

DHPS, ***DHFR*** are key enzymes that are inhibited by commonly used antifolates (section 4.1.2).

[1] DHFR plays a central role in the folate pathway and has three functions: control of the *de novo* synthesis, salvage of exogenous folate derivatives, and recycling of DHF (Nzila, 2006a).

about medicinal chemistry, mechanism of action and spread of resistances of antifolates can be found elsewhere (Schlitzer, 2007; Nzila, 2006a,b).

Table 4.1.: Commonly used antimalarials with antifolate activity.

brand name	DHPS inhibitor	DHFR inhibitor
Paludrine®	-	proguanil
Daraprim®	-	pyrimethamine
Fansidar®	sulfadoxine	pyrimethamine
Metakelfin®	sulfalene	pyrimethamine
Maloprim®	dapsone	pyrimethamine
LapDap®	dapsone	proguanil

4.1.3. Pyruvoyltetrahydropterin synthase orthologue of *P. falciparum*

Following statement regarding the folate biosynthesis can be found in the *P. falciparum* genome paper: "All but one of the enzymes dihydroneopterin aldolase (DHNA) required for *de novo* synthesis of folate from GTP were identified" (Gardner et al., 2002). The reasons remain unclear, whether the *dhna* gene was missed by the BLAST (Altschul, 1997) search or the respective gene is missing in the genome of *P. falciparum*. The lack of knowledge how *Plasmodium* parasites are able to fill the gap in the folate biosynthetic pathway lead the group of John Hyde to the discovery of a pyruvoyltetrahydropterin synthase (PTPS) orthologue using both bioinformatic methods and biochemical assay techniques (Dittrich et al., 2008). It is worth mentioning that the success of their work was supported by bioinformatic tools which make use of secondary and tertiary structure elements, since sequence-based approaches failed to deliver any statistically significant hits.

Originally, PTPS is embedded in the *de novo* biosynthesis of tetrahydrobiopterin (BH_4). BH_4 is required as a cofactor for aromatic amino acid hydrolases, for all NO synthases and glyceryl-ether monooxygenases (Thöny et al., 2000). PTPS catalyzes the reaction of 7,8-dihydroneopterin triphosphate (DHN-PPP) to 6-pyruvoyltetrahydropterin (PTP) (Figure 4.2 b). The difference between the conventional PTPS enzyme and the PTPS orthologue found in *P. falciparum* is an active site mutation of a cysteine residue to glutamate leading to different catalytic properties of the enzyme. The respective glutamate is labeled as Glu161 in the *Pf*PTPS binding site (Figure 4.4). Interestingly, PTPS substrate DHN-PPP is closely related to that of DHNA (Figure 4.3). Most likely the mutation and the substrate similarity give rise to the ability of *Pf*PTPS to produce both PTP, an intermediate of the BH_4 pathway, and 7,8-dihydro-6-hydroxymethylpterin (DHPT), an intermediate of the folate pathway, from the precursor DHN-PPP (Figure 4.2 c). Regarding the equilibrium it has been shown that the balance is shifted to the side of DHPT formation due to the higher efficiency of *Pf*PTPS(Dittrich et al., 2008).

Since, it has been shown that PTPS is an integral part of the *Plasmodium* folate pathway and its inhibition is likely to block parasite growth, a VS was performed in order to identify scaffolds as putatitive inhibitors of this enzyme.

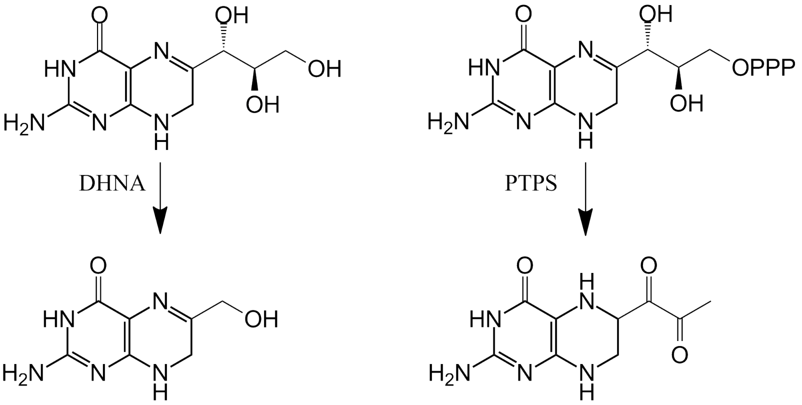


Figure 4.3.: Substrates and products of conventional DHNA and PTPS catalyzed reactions.

4.1.4. Binding site of *Pf*PTPS

The crystal structure of *Pf*PTPS reveals a rather small, zinc containing active site. The volume of the binding site occupied by the native ligand is about 300 Å^3 (Relibase+). Biopterin is bound in the interface between two protein chains. The zinc ion is chelated by oxygens of the 1,2-dihydroxypropyl moiety of biopterin, hence qualifying it as a zinc-binding group (ZBG). The interactions of biopterin with the protein chains are not only rich in hydrogen bonds but perpendicular the heterocyclic ring system also forms $\pi - \pi$-stacking interactions to adjacent phenylalanine and tyrosine side chains (Figure 4.4).

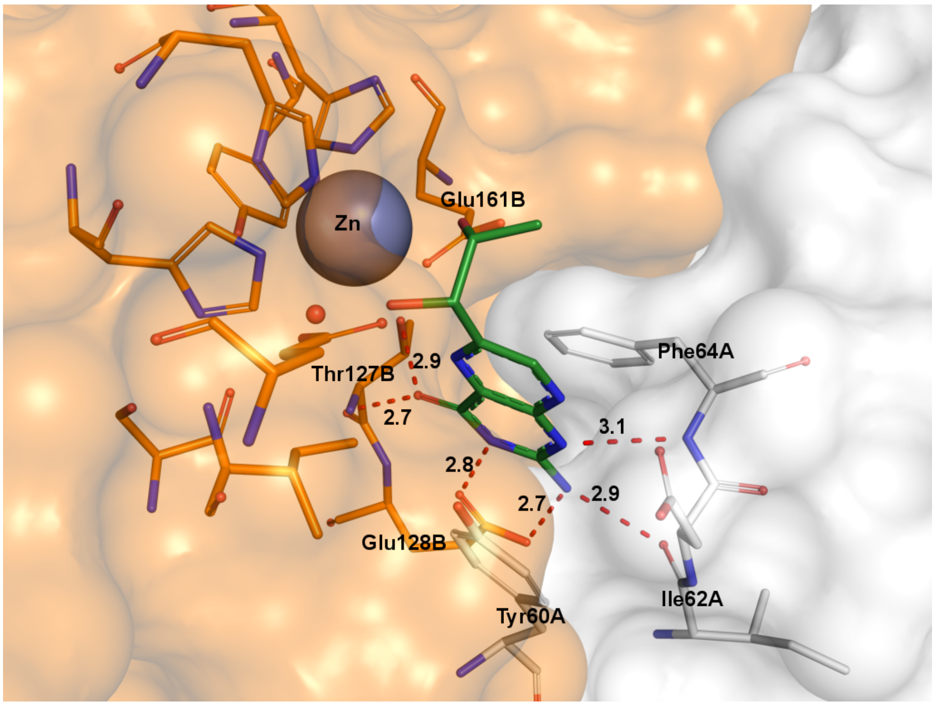


Figure 4.4.: (a) *Pf*PTPS binding site with biopterin. Biopterin binds in the interface between two protein chains (white: chain A, orange: chain B.) The 1*H*-pteridin-4-one moiety forms two hydrogen bonds to the carboxyl group of Glu128B and one to the hydroxyl group of Thr127B. Three hydrogen bonds are established to the backbone atoms of Ile62A, Phe64A, and Thr127B. In addition, face-to-face aromatic interactions to the side chains of Tyr60A and Phe64A are formed. A water molecule occupies a small subpocket; up to now any function has been attributed to this water molecule in literature.

4.2. Targeted Library Design

A ligand library has been created prior to VS considering both properties of the *Pf*PTPS binding site which are: The small volume and the presence of a zinc ion. Regarding the pocket size of *Pf*PTPS, the fragment-like subset of the ZINC database (Irwin and Shoichet, 2006) was selected as a matter of choice for the ligand library, although a general definition for fragment-like compounds is absent in literature. The so-called "Rule of Three" ("RO3") (Congreve et al., 2003) provides filter criteria for the construction of fragment libraries: MW is < 300, number of hydrogen-bond donor (HBD)s is ≤ 3, number of hydrogen-bond acceptor (HBA)s is ≤ 3, and clog$P$ is ≤ 3. A more recent study demonstrates that expansions of the "RO3" thresholds lead to an increase of hit rates as well as coverage of a large variety of chemotypes (Köster et al., 2011). However, the criteria of the ZINC database for fragment-like compounds deviate from the previously described rules and are defined as follows: MW ≤ 250, clog$P$ ≤ 3.5, and number of rotatable bonds ≤ 5. This library has been screened for known ZBGs (Figure 4.5) using the substructure search implemented in *fconv* (Neudert and Klebe, 2011b). Compounds which contained at least one ZBG were allowed to pass the filter. Starting initially with 413,796 entries, the library size was reduced to 12,068 molecules. Redundant entries were subsequently discarded, resulting in a ligand set of 4,140 structurally unique compounds. After all, the ligands were protonated and minimized by MOE using the MMFF94x.

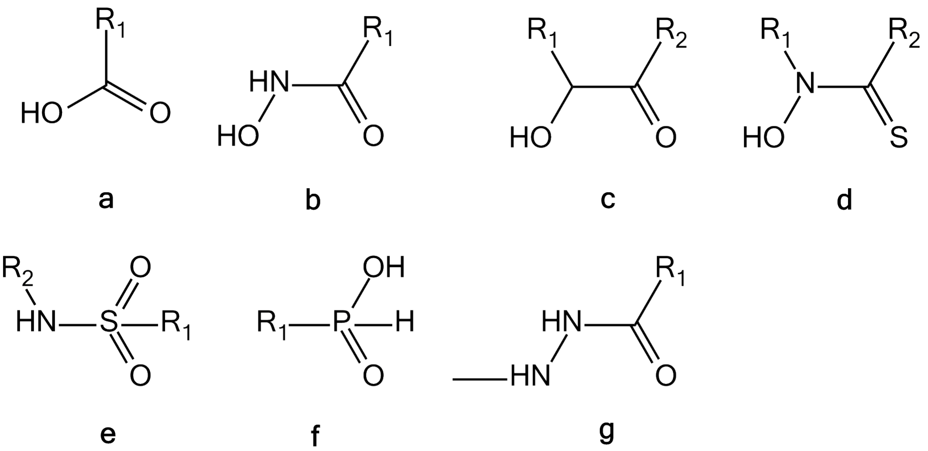


Figure 4.5.: Seven ZBGs were used to extract molecules from the ZINC fragment-like subset for the docking experiments. Cyclic as well as non-cyclic groups were considered from literature (Puerta et al., 2006; Klebe, 2009). Worth mentioning, cyclic ZBGs have been designed in order to overcome the limitations, such as reactivity or metabolic instability of non-cyclic zinc-chelating groups (Puerta et al., 2006).

4.3. Docking Setup

For the docking studies a *Pf*PTPS crystal structure (PDB 1Y13) comprising a resolution of 2.2 Å was used. The receptor site was defined by applying a radius of 8.0 Å around the native ligand biopterin in the active site. All ligands (section 4.2) were docked into the pocket using the GOLD program (Jones et al., 1997). Important parameters are summarized in Table 4.2. Five top ranked docking solution for each compound were generated and subsequently clustered taking a RMSD of 2.0 Å of all ligand atoms as cluster criterion. All solution were re-scored and re-ranked using the DSX per-atom-score (Neudert and Klebe, 2011a). Out of the generated solutions 200 top ranked docking poses were visually inspected and nine compounds were selected for further investigation (Figure 4.6).

Table 4.2.: Parameters for GOLD used in the *Pf*PTPS screening (PDB 1Y13).

Parameter	Value
Binding site	8.0 Å around biopterin
GA runs	30
Scoring function	ASP (Mooij and Verdonk, 2005)
Top N solutions	5
Search efficiency	100 %
Early termination	off

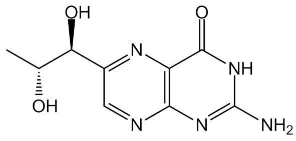
4.4. Results and Discussion

All nine selected structures provide apart from the ZBG a HBD group most likely required for the interaction with the acidic side chain of Glu128B at the bottom of the binding pocket, as it has been also observed for the native ligand. For the design of high affinity binders, the interaction of a strong HBD functionality, e.g. a positively charged basic group, with Glu128B is most likely as much important as the key interaction to the zinc ion.

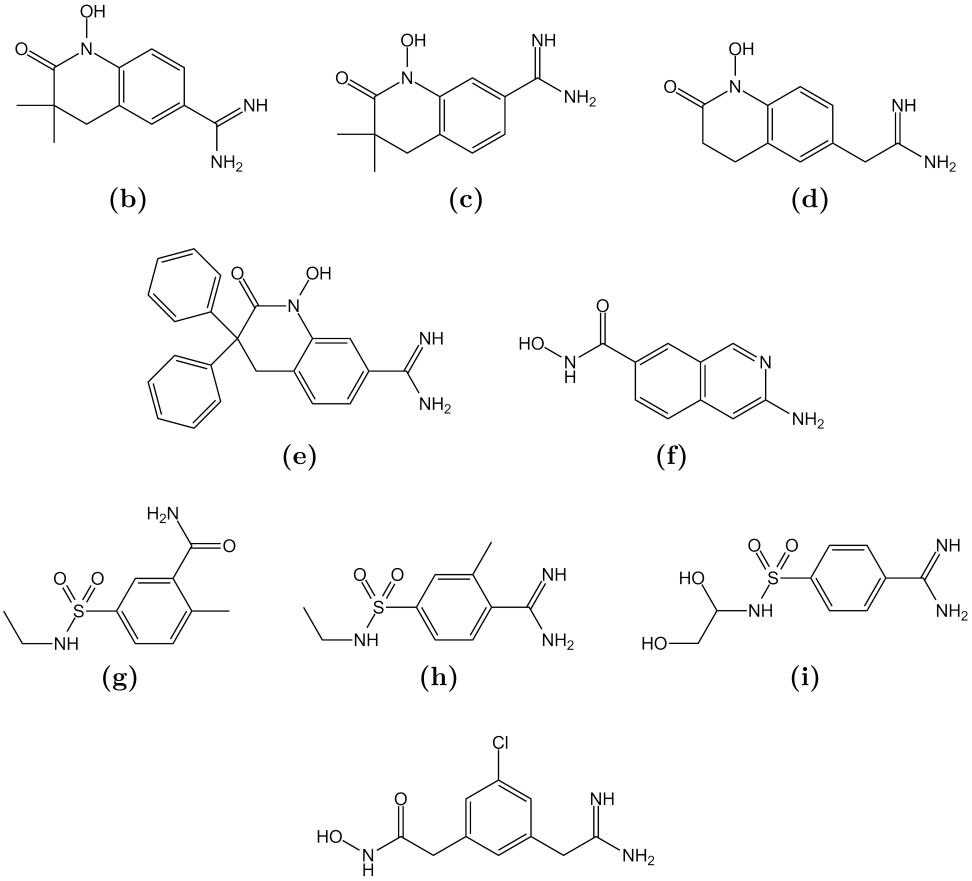
Four selected compounds (Figure 4.6 b–e) incorporating an endo cyclic ZBG share the 1-hydroxy-2-oxo-3,4-dihydroquinoline core structure in common. The docking pose of one derivative (Figure 4.7 a) indicates hydrogen bonding of the amidine moiety to the side chain of Glu128B and in addition to the backbone carbonyl of Ile62A.

One screening hit (Figure 4.6 d), a 3-aminoisoquinoline derivative, provides an exocyclic ZBG at position 7, it therefore shows a different substitution pattern to the latter compounds. The hit list also comprises three molecules with a sulfonamide moiety as ZBG (Figure 4.6 g–i). In Figure 4.7 b the docking solution of a representative compound is depicted. The nitrogen atom of the 3-carboxamide moiety forms a hydrogen bond to Glu128B and the oxygen atom accepts hydrogens from both, the backbone nitrogen and the hydroxy group, of Thr127B.

The last compound to mention (Figure 4.6j) provides an interesting three-armed substitution pattern of benzene. This particular 1,3,5-substituted



(a) biopterin

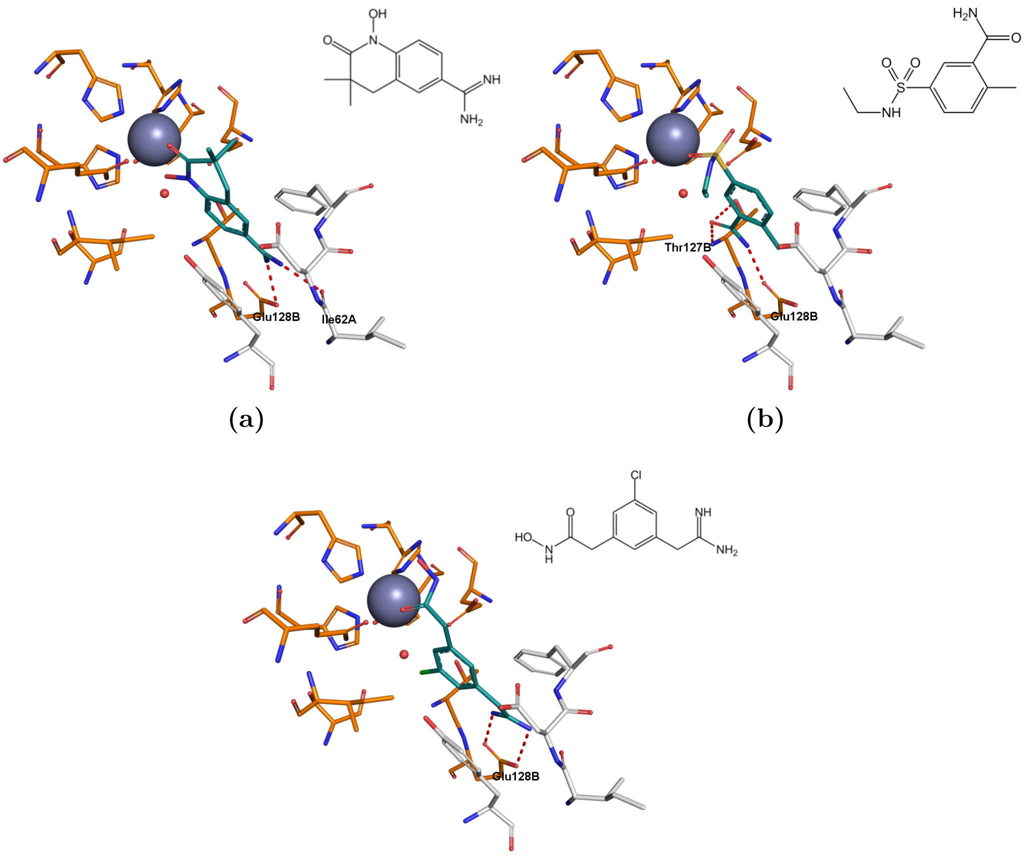


(b) **(c)** **(d)** **(e)** **(f)** **(g)** **(h)** **(i)** **(j)**

Figure 4.6.: Nine fragment-like compounds (b-j) selected from the *Pf*PTPS virtual screening.

benzene derivative has a ZBG at position 1, a methylene carboxamidine group as a HBD at position 3, and a chlorine atom at position 5, which might be capable to displace the water molecule from the binding site (Figure 4.7 c).

Summing up, inhibition of the folate pathway up-stream the key enzyme DHFR, e.g. the DHPS, brings about synergistic effects, thereby benefits for malaria treatment. Following this strategy, inhibition of *Pf*PTPS should be associated with a similar result. Up to now no inhibitors for *Pf*PTPS have been reported in literature. Using a targeted fragment-like library of commercially available compounds for the *Pf*PTPS screening resulted in selection of nine promising hits. Their predicted binding has to be validated experimentally.



(a) **(b)**

(c)

Figure 4.7.: Docking poses for three (Figure 4.6 b, g, j) selected virtual screening hits are shown. Hydrogen bonds which contribute to the inhibitor binding mode are depicted as red dashed lines and the respective amino acids are labeled. A water molecule is shown in the binding site for the sake of completeness, but it has been discarded for the docking experiments.
(c) The distance between the chlorine atom and the water molecule is 2.0 Å, indicating a clash between the atoms. The predicted ligand binding mode would lead to a displacement of the water molecule by the chlorine atom.

5

Chapter 5.

Virtual Screening for IspD Scaffolds

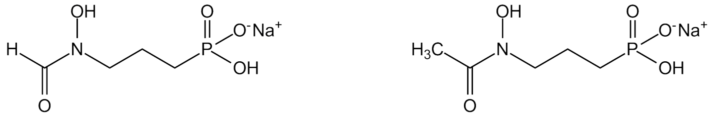
5.1. Inroductory remarks

The present study was carried out in cooperation with Thomas Rickmeyer during his internship and Kan Fu who performed the experimental measurements.

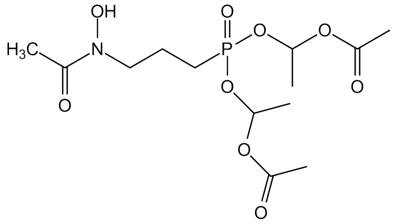
5.2. Motivation

Isoprenoids play essential roles in primary and secondary metabolism in all living organisms, including thousands of mono-, sesqui-, di-, and triterpenes, sterols, and carotenoids (Sacchettini, 1997). The fundamental precursors of isoprenoids are isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). IPP can be produced *via* the mevalonate (Beytía and Porter, 1976) or the 1-deoxy-D-xylulose 5-phosphate/2-C-methyl-D-erythritol 4-phosphate (DXP/MEP) (Rohmer, 1999) pathways. There are several reasons which make the DXP/MEP pathway attractive for the development of new antibacterial and antimalarial drugs. First, a wide range of pathogenic organisms such as *M. tuberculosis, P. falciparum* synthesize IPP via this pathway. Second, the DXP/MEP pathway is absent in mammalian cells. Third, the DXP/MEP pathway comprises eight distinct enzymes, therefore providing a large amount of targets which can be exploited for synergistic effects by inhibiting more than one enzyme.

An example for the DXP/MEP pathway inhibition is the well-studied inhibitor fosmidomycin (Figure 5.1a). Fosmidomycin inhibits the second reaction catalyzed by 1-desoxy-D-xylulose-5-phosphate reductoisomerase (IspC). The inhibitor shows antibacterial, including multidrug-resistant strains (Davey et al., 2011), as well as antimalarials effects. It displays an IC_{50} for IspC from *E. coli* at 30 nM, from *M. tuberculosis* at 310 nM, and from *P. falciparum* at 35 nM (Jawaid et al., 2009; Schlitzer, 2007). The parasite growth in case of *P. falciparum* is inhibited at nanomolar IC_{50} values (Schlitzer, 2007). In an effort to improve the potency and bioavailability of fosmidomycin for the treatment of malaria structural variations have been performed, leading to the compound FR-900098 which displays higher activity against the *Plasmodium* parasites (Figure 5.1b, Jomaa (1999)). Prodrug strategies were successfully applied to increase the efficacy of FR-900098 in a murine malaria model (Figure 5.1c, Ortmann et al. (2003)). These examples demonstrate that inhibition of the DXP/MEP pathway turns out to be an effective strategy for the development of effective antibacterials as well as antimalarials.



(a) Fosmidomycin **(b)** FR900098



(c) SCHL-7166

Figure 5.1.: Inhibitors of the IspC.

5.3. IspD as drug target

For a better understanding of the structural determinants for structure-based drug design (SBDD) 4-diphosphocytidyl-2-C-methylerythritol synthetase (IspD) from *E. coli* has been selected as model protein. The term IspD will refer in the following to the enzyme of *E. coli.* IspD catalyzes the third reaction step in the DXP/MEP pathway converting 2-C-methyl-D-erythritol 4-phosphate (MEP) to 4-diphosphocytidyl-2-C-methylerythritol (CDP-ME) (Figure 5.2). Up to now seven protein crystal structures of IspD have been published. Four structures are in the apo form, two are complexed with the native substrate or the product in the active site respectively. One crystal structure contains the 1,2-propanediol molecule in the active site of IspD mimicking the binding mode of methylerythritol (ME) (Behnen et al., 2012). Thus, none of the published crystal structures of IspD from *E. coli* accommodate any known inhibitor in the active site.

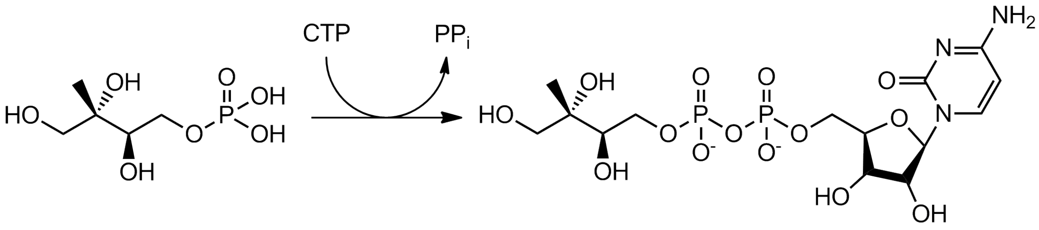
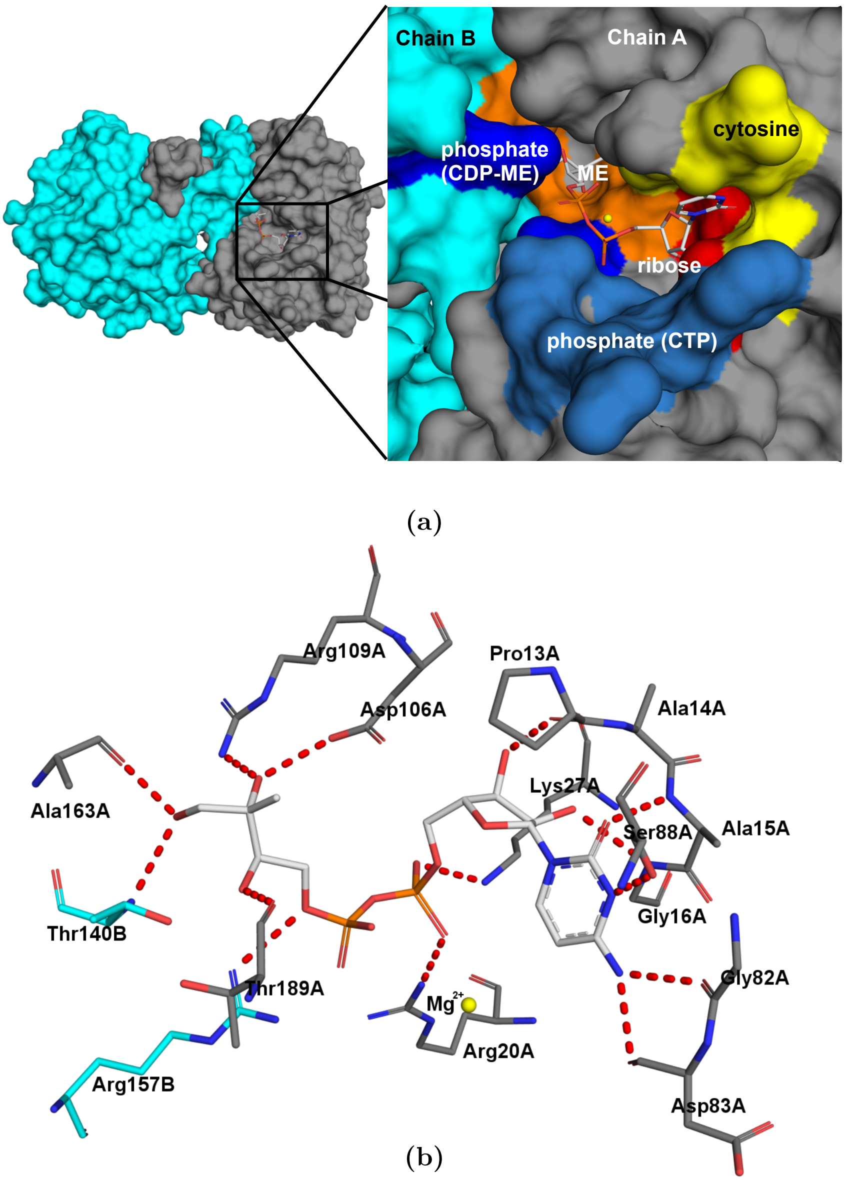


Figure 5.2.: IspD catalyzed reaction step.

5.4. Active site of IspD

In this section the exemplified active site originates from the IspD crystal structure containing CDP-ME and a magnesium ion, which is essentially required for the cytidyltransferase activity (PDB 1INI, resolution 1.8 Å, Richard et al. (2001)). This structure was also utilized for the VS.

IspD is organized as a homodimer. The monomeric units are clung together by a subdomain which resembles a curved arm. The active site is between the interface and MEP as well as phosphate pockets are partly formed by the interlocking arms (Figure 5.3a). The substrate specificity for the pyrimidine base cytosine is achieved *via* the hydrogen bond interactions and the steric constrictions. The cytosine moiety shows hydrogen bonds to the backbone atoms of Ala14A, Ala15A, Gly82A, and Asp83A, and to the hydroxyl group of Ser88A. The two hydroxyl groups of the ribose display hydrogen bonds to the backbone carbonyl of Pro13A and backbone amide of Gly16A. The phosphate groups are bound in an arginine and lycine rich region. They participate in hydrogen bonds to the side chains of Arg20A, Lys27A, and Arg157B. Last substructure of the CDP-ME molecule to mention is ME, which interacts with backbone atoms of Thr140B, Arg157B, Thr189 and with the side chains of Asp106A and Arg109A respectively (Figure 5.3b).



(a)

(b)

Figure 5.3.: IspD homodimer and its active site. (a) The active site is located adjacent to the interface. For reasons of clarity the surface of the binding site is colored according to the moieties it interacts with: cytosine in yellow, ribose in red, ME in orange, phosphate groups of the CDP-ME molecule in dark blue. The surface area which interacts with the phosphates of the substrate cytidine triphosphate (CTP) is depicted in light blue. The small yellow sphere represents the magnesium ion complexed by the phosphate groups. (b) Hydrogen bond interactions between IspD and bound CDP-ME are depicted as red dashed lines.

5.5. Docking setup and results

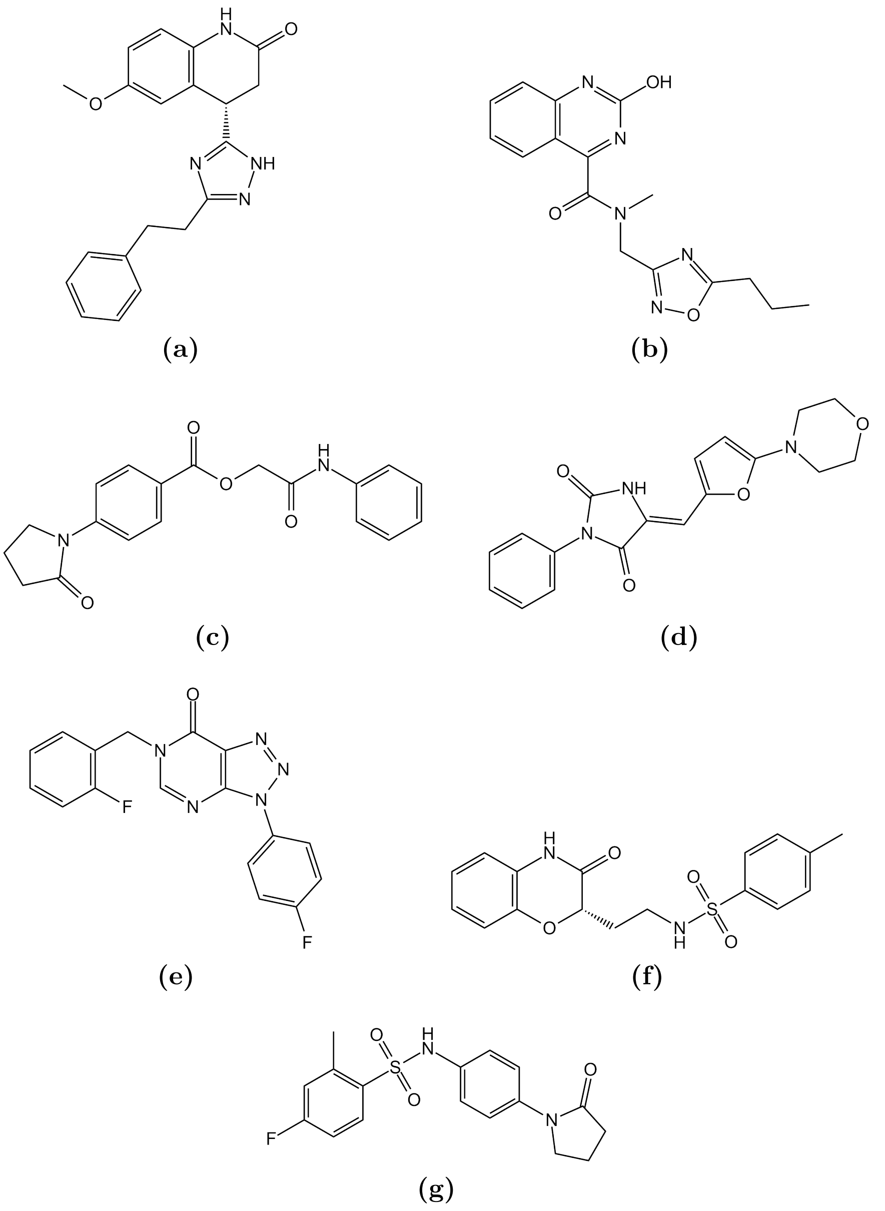
Initial studies indicated that lead-like compounds are suitable for docking into the pocket of IspD, mostly due to their size. Drug-like (Lipinski, 2000) molecules were too large and bulky to fit into the binding pocket accordingly. Therefore, nearly four million compounds were retrieved from the ZINC (Irwin and Shoichet, 2006) considering the lead-like subset. Worth mentioning, the used ligand data included also fragment-like compounds (section 4.2) and the ZINC content is very dynamic. The rules for subset definition have been also modified since the retrieval of the compounds used for the docking. Protonation states were assigned and minimization of the ligands was performed by MOE using the MMFF94x.

The architecture of the used protein structure and its respective active site has been described in the previous section 5.4. The native ligand CDP-ME, ions and water molecules were removed from the binding site prior to docking. Protein residues coinciding with a radius of 7 Å around the native reaction product CDP-ME were included in the receptor site definition, and contributing residues originated from both chains of the homodimer. Key parameters for the docking studies using GOLD (Jones et al., 1997) are summarized in Table 5.1.

Table 5.1.: Parameters for GOLD used in the IspD screening (PDB 1INI).

Parameter	Value
Binding site	7.0 Å around CDP-ME
GA runs	10
Scoring function	ASP (Mooij and Verdonk, 2005)
Top N solutions	1
Search efficiency	30 %
Early termination	on (top 5 solutions & rmsd tolerance 1.5 Å)

Nearly 40,000 best ASP scored compounds were re-scored and re-ranked by DSX (Neudert and Klebe, 2011a). A visual inspection of the top 200 compounds lead to the selection of six instantly available compounds (Figure 5.4a-e, g) for experimental testing (section 5.6). One molecule (Figure 5.4f) comprised a promising docking pose, in which the oxygen atoms of the sulfonamide group are positioned in a similar fashion as the oxygens of diphosphate moiety of CDP-ME. Owing to the fact that this compound is not directly available for purchase, a synthesis is planned in collaboration with the group of Professor Schlitzer.



(a) (b) (c) (d) (e) (f) (g)

Figure 5.4.: IspD virtual screening hits ordered according to the DSX ranking.

5.6. Experimental results

The experimental enzyme binding assay of six purchased ligands was performed by Kan Fu at the Institute of Organic Chemistry, University of Frankfurt, in collaboration with Dr. Marcus Maurer and Dr. Krishna Saxena from the group of Prof. Harald Schwalbe. Three distinct nuclear magnetic resonance (NMR) techniques were applied:

- saturation transfer difference (STD) (Mayer and Meyer, 1999),
- water ligand observation by gradient spectroscopy (WaterLOGSY) (Dalvit et al., 2000), and
- T2 method (Hajduk et al., 1997).

These preliminary experiments allowed the estimation of the binding strength of the selected ligands (Table 5.2). Two compounds could not be measured due to insufficient solubility and one compound showed no binding. Interestingly, the top ranked compound displayed the highest binding affinity amongst the VS hits. The docking pose is visualized in Figure 5.5. Two ligands showed only weak binding. The explicit K_i and/or IC_{50} values of the active ligands are planned to be characterized in near future.

Table 5.2.: Experimental binding of IspD hits. Compounds are labeled according to their appearance in Figure 5.4.

Compound	ZINC id	DSX rank	Binding
a	65381879	1	**moderate**
b	19148434	32	**weak**
c	4891045	40	(poor solubility)
d	4872921	53	no binding
e	4139891	64	(poor solubility)
f	49739899	144	(not available)
g	5329261	175	**weak**

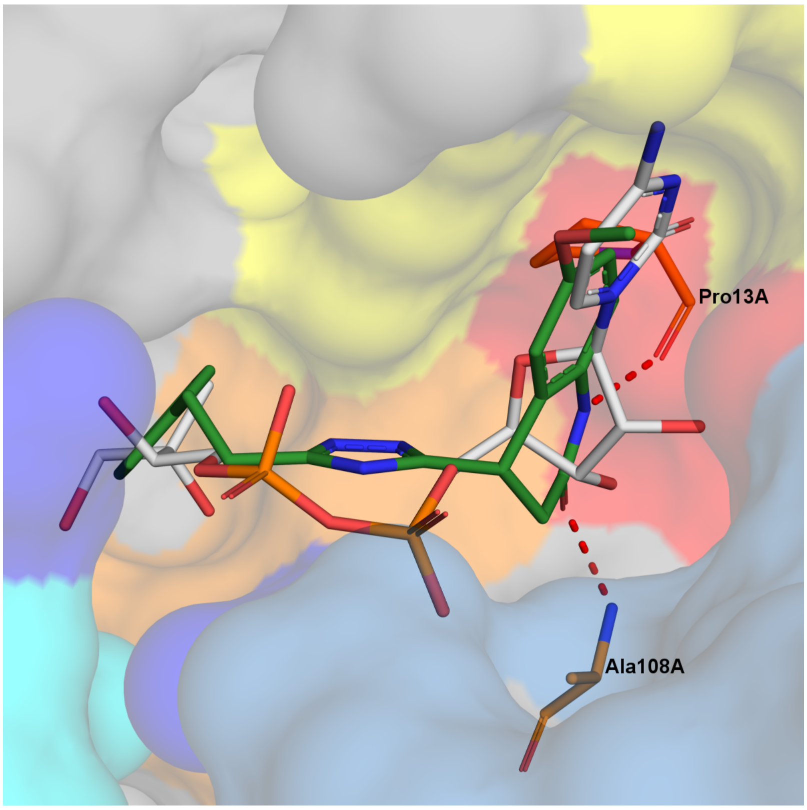


Figure 5.5.: Top ranked docking pose emerged from the IspD VS is depicted. The putative binding mode exhibits two hydrogen bonds to the backbone amide of Ala180A and to the backbone carbonyl of Pro13A. The dihydroquinoline-2-one moiety occupies the cystosine and ribose pockets. The triazol and the phenylethyl moieties extend from the region of phosphate pocket to the ME pocket. The color code of CDP-ME subpockets has been adapted from Figure 5.3a

Part III.

Appendix

Glossary

Active site Active sites are in most cases depressions on protein surface. The active site is that region where substrates bind which leads in case of enzymes to a catalytic reaction leading to one or several products. Binding of a substrate to the active site of a receptor protein leads to signal transduction. 22, 43, 105

Binding pocket see binding site. 4, 38

Binding site The region of a protein where a ligand binds, but it is not necessarily an active site. 4, 37, 105

Cavity A depression on the protein surface. It can be an active site or a binding site. 6, 38

Merck molecular force field 94x This force field was parameterized for gas phase small organic molecules in medicinal chemistry. 76, 86, 98

Virtual screening A computational method that screens huge libraries of small molecules for a putative binder with respect to a particular target and/or model. 60, 89, 100

Bibliography

Aikawa, M., Schwartz, A., Uni, S., Nussenzweig, R., and Hollingdale, M. (1984). Ultrastructure of *in vitro* cultured exoerythrocytic stage of *Plasmodium berghei* in a hepatoma cell line. *Am. J. Trop. Med. Hyg.*, 33(5):792–9.

Altschul, S. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389.

Arico, S., Pattingre, S., Bauvy, C., Gane, P., Barbat, A., Codogno, P., and Ogier-Denis, E. (2002). Celecoxib induces apoptosis by inhibiting 3-phosphoinositide-dependent protein kinase-1 activity in the human colon cancer ht-29 cell line. *J. Biol. Chem.*, 277(31):27613–21.

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic Acids Research*, 28(1):304.

Behnen, J., Köster, H., Neudert, G., Craan, T., Heine, A., and Klebe, G. (2012). Experimental and computational active site mapping as a starting point to fragment-based lead discovery. *ChemMedChem*, 7(2):248–61.

Beytía, E. D. and Porter, J. W. (1976). Biochemistry of polyisoprenoid biosynthesis. *Annu. Rev. Biochem.*, 45:113–42.

BioSolveIT. BioSolveIT GmbH "`http://www.biosolveit.de/`".

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1.

ChEMBL Accessed July 2011. `http://www.ebi.ac.uk/chembl/`.

Choe, Y., Leonetti, F., Greenbaum, D. C., Lecaille, F., Bogyo, M., Brömme, D., Ellman, J. A., and Craik, C. S. (2006). Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.*, 281(18):12824–32.

Combrink, K. D., Gülgeze, H. B., Meanwell, N. A., Pearce, B. C., Zulan, P., Bisacchi, G. S., Roberts, D. G., Stanley, P., and Seiler, S. M. (1998). 1,2-benzisothiazol-3-one 1,1-dioxide inhibitors of human mast cell tryptase. *J. Med. Chem.*, 41(24):4854–60.

Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today*, 8(19):876 – 877.

Cuerrier, D., Moldoveanu, T., Campbell, R. L., Kelly, J., Yoruk, B., Verhelst, S. H. L., Greenbaum, D., Bogyo, M., and Davies, P. L. (2007). Development of calpain-specific inactivators by screening of positional scanning epoxide libraries. *J. Biol. Chem.*, 282(13):9600–11.

Dalvit, C., Pevarello, P., Tatò, M., Veronesi, M., Vulpetti, A., and Sundström, M. (2000). Identification of compounds with binding affinity to proteins *via* magnetization transfer from bulk water. *Journal of Biomolecular NMR*, 18(1):65.

Das, C., Hoang, Q. Q., Kreinbring, C. A., Luchansky, S. J., Meray, R. K., Ray, S. S., Lansbury, P. T., Ringe, D., and Petsko, G. A. (2006). Structural basis for conformational plasticity of the parkinson's disease-associated ubiquitin hydrolase uch-l1. *Proc. Natl. Acad. Sci. U.S.A.*, 103(12):4675–80.

Davey, M. S., Tyrrell, J. M., Howe, R. A., Walsh, T. R., Moser, B., Toleman, M. A., and Eberl, M. (2011). A promising target for treatment of multidrug-resistant bacterial infections. *Antimicrob. Agents Chemother.*, 55(7):3635–6.

Davis, T. L., Walker, J. R., Finerty, P. J., Mackenzie, F., Newman, E. M., and Dhe-Paganon, S. (2007). The crystal structures of human calpains 1 and 9 imply diverse mechanisms of action and auto-inhibition. *J. Mol. Biol.*, 366(1):216–29.

Dittrich, S., Mitchell, S. L., Blagborough, A. M., Wang, Q., Wang, P., Sims, P. F. G., and Hyde, J. E. (2008). An atypical orthologue of 6-pyruvoyltetrahydropterin synthase can provide the missing link in the folate biosynthesis pathway of malaria parasites. *Mol. Microbiol.*, 67(3):609–18.

Donkor, I. O. (2011). Calpain inhibitors: a survey of compounds reported in the patent and scientific literature. *Expert Opin Ther Pat*, 21(5):601–36.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511.

Greg Landrum, RDKit. Rdkit: Open-source cheminformatics. `http://www.rdkit.org`.

Hajduk, P. J., Olejniczak, E. T., and Fesik, S. W. (1997). One-dimensional relaxation- and diffusion-edited nmr methods for screening compounds that bind to macromolecules. *Journal of the American Chemical Society*, 119(50):12257.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107.

Hendlich, M. (1997). Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359.

Hilpert, K., Ackermann, J., Banner, D. W., Gast, A., Gubernator, K., Hadvary, P., Labler, L., Mueller, K., and Schmid, G. (1994). Design and synthesis of potent and highly selective thrombin inhibitors. *Journal of Medicinal Chemistry*, 37(23):3889.

Hopkins, C. R., Czekaj, M., Kaye, S. S., Gao, Z., Pribish, J., Pauls, H., Liang, G., Sides, K., Cramer, D., Cairns, J., Luo, Y., Lim, H.-K., Vaz, R., Rebello, S., Maignan, S., Dupuy, A., Mathieu, M., and Levell, J. (2005). Design, synthesis, and biological activity of potent and selective inhibitors of mast cell tryptase. *Bioorg. Med. Chem. Lett.*, 15(11):2734–7.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Huggins, D. J., Sherman, W., and Tidor, B. (2012). Rational approaches to improving selectivity in drug design. *J. Med. Chem.*, 55(4):1424–44.

Hyde, J. E., Dittrich, S., Wang, P., Sims, P. F. G., de Crécy-Lagard, V., and Hanson, A. D. (2008). *Plasmodium falciparum*: a paradigm for alternative folate biosynthesis in diverse microorganisms? *Trends Parasitol.*, 24(11):502–8.

Irwin, J. J. and Shoichet, B. K. (2006). ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1):177–82.

Jawaid, S., Seidle, H., Zhou, W., Abdirahman, H., Abadeer, M., Hix, J. H., van Hoek, M. L., and Couch, R. D. (2009). Kinetic characterization and phosphoregulation of the *Francisella tularensis* 1-deoxy-d-xylulose 5-phosphate reductoisomerase (mep synthase). *PLoS ONE*, 4(12):e8288.

Jomaa, H. (1999). Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, 285(5433):1573.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–48.

Kapoor, M., Dar, M. J., Surolia, A., and Surolia, N. (2001). Kinetic determinants of the interaction of enoyl-acp reductase from *Plasmodium falciparum* with its substrates and inhibitors. *Biochem. Biophys. Res. Commun.*, 289(4):832–7.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data An Introduction to Cluster Analysis.* Wiley Interscience.

Kawasaki, Y. and Freire, E. (2011). Finding a better path to drug selectivity. *Drug Discov. Today*, 16(21-22):985–90.

Kinnings, S. L. and Jackson, R. M. (2009). Binding site similarity analysis for the functional classification of the protein kinase family. *J Chem Inf Model*, 49(2):318–29.

Klebe, G. (2009). *Wirkstoffdesign.* Spektrum Akademischer Verlag Heidelberg.

Kohler, S. (1997). A plastid of probable green algal origin in apicomplexan parasites. *Science*, 275(5305):1485.

Kohlwein, S. D., Eder, S., Oh, C. S., Martin, C. E., Gable, K., Bacikova, D., and Dunn, T. (2001). Tsc13p is required for fatty acid elongation and localizes to a novel structure at the nuclear-vacuolar interface in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, 21(1):109–25.

Kolb, P. and Caflisch, A. (2006). Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.*, 49(25):7384–92.

Köster, H., Craan, T., Brass, S., Herhaus, C., Zentgraf, M., Neumann, L., Heine, A., and Klebe, G. (2011). A small nonrule of 3 compatible fragment library provides high hit rate of endothiapepsin crystal structures with various fragment chemotypes. *J. Med. Chem.*, 54(22):7784–96.

Kuhn, D., Weskamp, N., Hüllermeier, E., and Klebe, G. (2007). Functional classification of protein kinase binding sites using cavbase. *ChemMedChem*, 2(10):1432–47.

Kuhn, D., Weskamp, N., Schmitt, S., Hüllermeier, E., and Klebe, G. (2006). From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.*, 359(4):1023–44.

Lai, L., Xu, Z., Zhou, J., Lee, K.-D., and Amidon, G. L. (2008). Molecular basis of prodrug activation by human valacyclovirase, an alpha-amino acid ester hydrolase. *J. Biol. Chem.*, 283(14):9318–27.

Lee, C.-S., Liu, W., Sprengeler, P. A., Somoza, J. R., Janc, J. W., Sperandio, D., Spencer, J. R., Green, M. J., and McGrath, M. E. (2006a). Design of novel, potent, and selective human beta-tryptase inhibitors based on alpha-keto-[1,2,4]-oxadiazoles. *Bioorg. Med. Chem. Lett.*, 16(15):4036–40.

Lee, S. H., Stephens, J. L., Paul, K. S., and Englund, P. T. (2006b). Fatty acid synthesis by elongases in trypanosomes. *Cell*, 126(4):691–9.

Lewell, X., Judd, D., Watson, S., and Hann, M. (1998). Recap-retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Modeling*, 38(3):511.

Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, 44(1):235.

Liu, Y., Lashuel, H., Choi, S., Xing, X., Case, A., Ni, J., Yeh, L. A., Cuny, G. D., Stein, R. L., and Lansbury, P. J. (2003). Discovery of inhibitors that elucidate the role of uch-l1 activity in the h1299 lung cancer cell line. *Chemistry & Biology*, 10(9):837–846.

Mayer, M. and Meyer, B. (1999). Characterization of ligand binding by saturation transfer difference nmr spectroscopy. *Angewandte Chemie International Edition*, 38(12):1784.

McFadden, G. I., Reith, M. E., Munholland, J., and Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature*, 381(6582):482.

McGrath, M. E., Mirzadegan, T., and Schmidt, B. F. (1997). Crystal structure of phenylmethanesulfonyl fluoride-treated human chymase at 1.9 a. *Biochemistry*, 36(47):14318–24.

McLeod, R., Muench, S. P., Rafferty, J. B., Kyle, D. E., Mui, E. J., Kirisits, M. J., Mack, D. G., Roberts, C. W., Samuel, B. U., Lyons, R. E., Dorris, M., Milhous, W. K., and Rice, D. W. (2001). Triclosan inhibits the growth of *Plasmodium falciparum* and *Toxoplasma gondii* by inhibition of apicomplexan Fab I. *International Journal for Parasitology*, 31(2):109.

McMurry, L. M., Oethinger, M., and Levy, S. B. (1998). Triclosan targets lipid synthesis. *Nature*, 394(6693):531–2.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873.

Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. (2009). The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol Biosyst*, 5(9):1051–7.

Milletti, F. and Vulpetti, A. (2010). Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J Chem Inf Model*, 50(8):1418–31.

MOE. Molecular Operating Environment 10.2010 "`http://www.chemcomp.com/`".

Molinari, J. F., Scuri, M., Moore, W. R., Clark, J., Tanaka, R., and Abraham, W. M. (1996). Inhaled tryptase causes bronchoconstriction in sheep via histamine release. *American Journal of Respiratory and Critical Care Medicine*, 154(3):649–53.

Mooij, W. T. M. and Verdonk, M. L. (2005). General and targeted statistical potentials for protein-ligand interactions. *Proteins*, 61(2):272–87.

Muench, S. P., Prigge, S. T., McLeod, R., Rafferty, J. B., Kirisits, M. J., Roberts, C. W., Mui, E. J., and Rice, D. W. (2007). Studies of *Toxoplasma gondii* and *Plasmodium falciparum* enoyl acyl carrier protein reductase and implications for the development of antiparasitic agents. *Acta Crystallogr. D Biol. Crystallogr.*, 63(Pt 3):328–38.

Murray, C. J., Rosenfeld, L. C., Lim, S. S., Andrews, K. G., Foreman, K. J., Haring, D., Fullman, N., Naghavi, M., Lozano, R., and Lopez, A. D. (2012). Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet*, 379(9814):413.

Nanao, M. H., Tcherniuk, S. O., Chroboczek, J., Dideberg, O., Dessen, A., and Balakirev, M. Y. (2004). Crystal structure of human otubain 2. *EMBO Rep.*, 5(8):783–8.

Nar, H., Bauer, M., Schmid, A., Stassen, J.-M., Wienen, W., Priepke, H. W., Kauffmann, I. K., Ries, U. J., and Hauel, N. H. (2001). Structural basis for inhibition promiscuity of dual specific thrombin and factor xa blood coagulation inhibitors. *Structure*, 9(1):29.

Neudert, G. and Klebe, G. (2011a). DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. *J. Chem. Inf. Model.*, 51(10):2731–45.

Neudert, G. and Klebe, G. (2011b). fconv: format conversion, manipulation, and feature computation of molecular data. *Bioinformatics (Oxford, England)*.

Nzila, A. (2006a). Inhibitors of *de novo* folate enzymes in *Plasmodium falciparum*. *Drug Discov. Today*, 11(19-20):939–44.

Nzila, A. (2006b). The past, present and future of antifolates in the treatment of *Plasmodium falciparum* infection. *J. Antimicrob. Chemother.*, 57(6):1043–54.

Nzila, A., Ward, S. A., Marsh, K., Sims, P. F. G., and Hyde, J. E. (2005). Comparative folate metabolism in humans and malaria parasites (part i): pointers for malaria treatment from cancer chemotherapy. *Trends Parasitol.*, 21(6):292–8.

O'Farrell, P. A., Gonzalez, F., Zheng, W., Johnston, S. A., and Joshua-Tor, L. (1999). Crystal structure of human bleomycin hydrolase, a self-compartmentalizing cysteine protease. *Structure*, 7(6):619–627.

Olivero, A. G., Eigenbrot, C., Goldsmith, R., Robarge, K., Artis, D. R., Flygare, J., Rawson, T., Sutherlin, D. P., Kadkhodayan, S., Beresini, M., Elliott, L. O., DeGuzman, G. G., Banner, D. W., Ultsch, M., Marzec, U., Hanson, S. R., Refino, C., Bunting, S., and Kirchhofer, D. (2005). A selective, slow binding inhibitor of factor viia binds to a nonstandard active site conformation and attenuates thrombus formation in vivo. *J. Biol. Chem.*, 280(10):9160–9.

Ortmann, R., Wiesner, J., Reichenberg, A., Henschker, D., Beck, E., Jomaa, H., and Schlitzer, M. (2003). Acyloxyalkyl ester prodrugs of FR900098 with improved *in vivo* anti-malarial activity. *Bioorganic & Medicinal Chemistry Letters*, 13(13):2163.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci*, 85(8):2444–2448.

Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A.-C., and Villoutreix, B. O. (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today*, 15(15-16):656–67.

Perozzo, R., Kuo, M., Sidhu, A. b. S., Valiyaveettil, J. T., Bittman, R., Jacobs, W. R., Fidock, D. A., and Sacchettini, J. C. (2002). Structural elucidation of the specificity of the antibacterial agent triclosan for malarial enoyl acyl carrier protein reductase. *J. Biol. Chem.*, 277(15):13106–14.

Pinzon-Ortiz, C., Friedman, J., Esko, J., and Sinnis, P. (2001). The binding of the circumsporozoite protein to cell surface heparan sulfate proteoglycans is required for plasmodium sporozoite attachment to target cells. *J. Biol. Chem.*, 276(29):26784–91.

Pollard, K. S. and van der Laan, M. J. (2002). New methods for identifying significant clusters in gene expression data. *Proceedings of the American Statistical Association, Biometrics Section.*

Puerta, D. T., Griffin, M. O., Lewis, J. A., Romero-Perez, D., Garcia, R., Villarreal, F. J., and Cohen, S. M. (2006). Heterocyclic zinc-binding groups for use in next-generation matrix metalloproteinase inhibitors: potency, toxicity, and reactivity. *J. Biol. Inorg. Chem.*, 11(2):131–8.

Ragsdale, S. W. (2008). *Catalysis of Methyl Group Transfers Involving Tetrahydrofolate and B12*, volume 79 of *Vitamins and Hormones*. Academic Press.

Rawlings, N. D., Barrett, A. J., and Bateman, A. (2010). Merops: the peptidase database. *Nucleic Acids Res.*, 38(Database issue):D227–33.

Richard, S. B., Bowman, M. E., Kwiatkowski, W., Kang, I., Chow, C., Lillo, A. M., Cane, D. E., and Noel, J. P. (2001). Structure of 4-diphosphocytidyl-2-c- methylerythritol synthetase involved in mevalonate- independent isoprenoid biosynthesis. *Nat. Struct. Biol.*, 8(7):641–8.

Rohmer, M. (1999). The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Natural Product Reports*, 16(5):565.

Rousseeuw, P. J. (1986). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*.

Sacchettini, J. C. (1997). Creating isoprenoid diversity. *Science*, 277(5333):1788.

Sanderson, P. E. J. (1999). Small, noncovalent serine protease inhibitors. *Medicinal Research Reviews*, 19(2):179.

Schlitzer, M. (2007). Malaria chemotherapeutics part I: History of antimalarial drug development, currently used therapeutics, and drugs in clinical development. *ChemMedChem*, 2(7):944–86.

Schmitt, S., Kuhn, D., and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *Journal Of Molecular Biology*, 323(2):387–406.

Schrader, F. C. (2012). *Development of novel type II FAS-Inhibitors targeting apicomplexan-borne diseases and development of potential Bid-Inhibitors as neuroprotectants.* PhD thesis, Fachbereich Pharmazie, Philipps-Universität Marburg. In German.

Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1.

Spalding, M. D. and Prigge, S. T. (2008). Malaria pulls a FASt one. *Cell Host Microbe*, 4(6):509–11.

Spitzer, R., Cleves, A. E., and Jain, A. N. (2011). Surface-based protein binding pocket similarity. *Proteins*, 79(9):2746–63.

Stegemann, B. and Klebe, G. (2011). Cofactor-binding sites in proteins of deviating sequence: Comparative analysis and clustering in torsion angle, cavity, and fold space. *Proteins*.

Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., Pollok, J.-M., Menard, R., and Heussler, V. T. (2006). Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science*, 313(5791):1287–90.

Surolia, N. and Surolia, A. (2001). Triclosan offers protection against blood stages of malaria by inhibiting enoyl-acp reductase of *Plasmodium falciparum. Nat. Med.*, 7(2):167–73.

Tarun, A. S., Vaughan, A. M., and Kappe, S. H. I. (2009). Redefining the role of *de novo* fatty acid synthesis in *Plasmodium parasites. Trends Parasitol.*, 25(12):545–50.

Tasdemir, D., Lack, G., Brun, R., Rüedi, P., Scapozza, L., and Perozzo, R. (2006). Inhibition of *Plasmodium falciparum* fatty acid biosynthesis: Evaluation of FabG, FabZ, and FabI as drug targets for flavonoids. *J. Med. Chem.*, 49(11):3345–53.

Tasdemir, D., Topaloglu, B., Perozzo, R., Brun, R., O'Neill, R., Carballeira, N. M., Zhang, X., Tonge, P. J., Linden, A., and Rüedi, P. (2007). Marine natural products from the turkish sponge agelas oroides that inhibit the enoyl reductases from *Plasmodium falciparum*, *Mycobacterium tuberculosis* and *Escherichia coli*. *Bioorg. Med. Chem.*, 15(21):6834–45.

Thöny, B., Auerbach, G., and Blau, N. (2000). Tetrahydrobiopterin biosynthesis, regeneration and functions. *Biochem. J.*, 347(1):1–16.

Tsuji, M., Mattei, D., Nussenzweig, R. S., Eichinger, D., and Zavala, F. (1994). Demonstration of heat-shock protein 70 in the sporozoite stage of malaria parasites. *Parasitol Res*, 80(1):16–21.

Vaughan, A. M., O'Neill, M. T., Tarun, A. S., Camargo, N., Phuong, T. M., Aly, A. S. I., Cowman, A. F., and Kappe, S. H. I. (2009). Type II fatty acid synthesis is essential only for malaria parasite late liver stage development. *Cell. Microbiol.*, 11(3):506–20.

Vial, H. J., Thuet, M. J., and Philippot, J. R. (1982). Phospholipid biosynthesis in synchronous *Plasmodium falciparum* cultures. *J. Protozool.*, 29(2):258–63.

Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C. T., Scozzafava, A., and Klebe, G. (2004). Unexpected nanomolar inhibition of carbonic anhydrase by cox-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.*, 47(3):550–7.

Weskamp, N., Hüllermeier, E., and Klebe, G. (2009). Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins*, 76(2):317–30.

WHO. World Malaria Report 2010 "`http://www.who.int/malaria/world_malaria_report_2010/en/index.html`".

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–72.

Xing, Y., Li, Z., Chen, Y., Stock, J. B., Jeffrey, P. D., and Shi, Y. (2008). Structural mechanism of demethylation and inactivation of protein phosphatase 2a. *Cell*, 133(1):154–63.

Xu, R. and Wunsch, D. C. (2009). *Clustering*. Wiley-IEEE Press.

Young, W. B., Mordenti, J., Torkelson, S., Shrader, W. D., Kolesnikov, A., Rai, R., Liu, L., Hu, H., Leahy, E. M., Green, M. J., Sprengeler, P. A., Katz, B. A., Yu, C., Janc, J. W., Elrod, K. C., Marzec, U. M., and Hanson, S. R. (2006). Factor viia inhibitors: chemical optimization, preclinical pharmacokinetics, pharmacodynamics, and efficacy in an arterial baboon thrombosis model. *Bioorg. Med. Chem. Lett.*, 16(7):2037–41.

Yu, M., Kumar, T. R. S., Nkrumah, L. J., Coppi, A., Retzlaff, S., Li, C. D., Kelly, B. J., Moura, P. A., Lakshmanan, V., Freundlich, J. S., Valderramos, J.-C., Vilcheze, C., Siedner, M., Tsai, J. H.-C., Falkard, B., Sidhu, A. B. S., Purcell, L. A., Gratraud, P., Kremer, L., Waters, A. P., Schiehser, G., Jacobus, D. P., Janse, C. J., Ager, A., Jacobs, W. R., Sacchettini, J. C., Heussler, V., Sinnis, P., and Fidock, D. A. (2008). The fatty acid biosynthesis enzyme FabI plays a key role in the development of liver-stage malarial parasites. *Cell Host Microbe*, 4(6):567–78.

Zhao, Y. and Karypis, G. (2005). Data clustering in life sciences. *Molecular Biotechnology*, 31(1):055.

Zhu, J., Huang, J.-W., Tseng, P.-H., Yang, Y.-T., Fowble, J., Shiau, C.-W., Shaw, Y.-J., Kulp, S. K., and Chen, C.-S. (2004). From the cyclooxygenase-2 inhibitor celecoxib to a novel class of 3-phosphoinositide-dependent protein kinase-1 inhibitors. *Cancer Res.*, 64(12):4309–18.

Danksagung

Prof. Dr. Gerhard Klebe danke ich für die sehr interessante Aufgabenstellung der Doktorarbeit und die herausragende, aber vor allem motivierende Betreuung während dieser Zeit. Ich danke ihm ebenfalls für das mir entgegengebrachte Vertrauen und die damit verbundene Möglichkeit auf einer internationalen Konferenz die Arbeit in einem Vortrag vorzustellen. Sehr dankbar bin ich ihm für seine ständige Bereitschaft für ein fachliches aber auch persönliches Gespräch und die vielen ausführlichen Korrekturen der Manuskripte. Letztendlich konnte diese Dissertation deshalb entstehen, da *Prof. Klebe* mir großen Freiraum für eine selbständige Arbeitsweise geboten hat.

Prof. Dr. Martin Schlitzer danke ich für die Bereitschaft diese Arbeit als Zweitgutachter zu beurteilen und für die Zusammenarbeit bei der Entwicklung von Inhibitoren gegen *Plasmodien*.

Dr. Florian Schrader danke ich für die hervorragende Zusammenarbeit bei *ENR*- und *BID*-Projekten.

Bei *Florian Meyer* möchte ich mich für seine Expertise auf dem Gebiet der Statistik und Clusteranalyse bedanken. Die zahlreichen Diskussionen haben viel zur Entwicklung des neuen Workflows zur Clusterung von Cavbase-Scores beigetragen.

Ich danke *Prof. Hüllermeier*, *Thomas Fober* und *Dr. Marco Mernberger* für die zahlreichen und kritischen Diskussionen zu *Cavbase*. An dieser

Stelle möchte ich ebenfalls *Dr. Willem Waegeman*, *Michiel Stock*, *Christina Heitzer*, *Matthias Leinweber*, *Lars Baumgärtner*, und *Timo Krotzky* für die verschiedenen Projekte auf dem Gebiet der Bindetaschenvergleiche danken.

Thomas Rickmeyer danke ich für die hervorragende Zusammenarbeit sowohl im *Cavbase*- als auch im *IspD*-Projekt und wünsche ihm viel Erfolg bei der Fortführung der Projekte.

Dr. Simon Cottrell danke ich für die technische und *Dr. Tjelvar Olsson* für die fachliche Unterstützung im *Cavbase*-Projekt und die angenehme Zeit beim *CCDC*.

Dr. Finton Sirockin danke ich für die anfängliche Initiation des *Cavbase*-Projektes und den regen fachlichen Austausch während des Aufenthalts bei *Novartis* in Basel und im Weiteren per E-Mail.

Prof. Dr. Klaus Reuter und *Inna Schmidt* danke ich für die interessante Fragestellung und die vielen Diskussionen im *Tgt*-Projekt.

Für die Zusammenarbeit zum Verhalten von Polymeren in Wasser danke ich *Dr. Olga Samsonova*.

Ich danke *Dr. Olayinka Oyewale Ajani* für das Korrekturlesen des *Pf*PTPS-Kapitels und die zukünftige Umsetzung der Vorschläge aus dem *Virtual Screening*.

Bei *Dr. Peter Kolb* und *Dr. Juan-Carlos Mobarec* bedanke ich mich für das Korrekturlesen der Manuskripte. Auch bei *Denis Schmidt* bedanke ich mich für die vielen Diskussionen und seine Hilfsbereitschaft Texte zum *Molecular Modeling* Korrektur zu lesen.

Den Administratoren der AG Klebe danke ich für ihre Mühe, die Soft- und

Hardware am Laufen zu halten. Ohne euch wäre die Arbeit viel schwieriger gewesen! Danke an *Sven Siebler*, *Michael Betz*, *Dr. Gerd Neudert*, *Timo Krotzky*, *Denis Schmidt* und *Felix Gut*.

Frau Lydia Hartleben möchte ich für ihre Geduld und Hilfsbereitschaft bei all meinen Anliegen danken.

Bei *Dr. Gerd Neudert* möchte ich mich bedanken fur die LaTeXVorlage der Dissertation.

Ein besonderer Dank geht an alle Mitglieder von *SAGD e.V, Marburg* und vor allem *Dr. Li Lee*. Für die vielen spontanen philosophischen Intermezzi zum chinesischen Kulturgut im Arbeitsalltag danke ich *Stefan Merkl*.

Ein großer Dank geht an *Kan Fu* für die vielen kulinarischen Verkostungen und fachlichen sowie persönlichen Diskussionen während der Promotionszeit.

Allen Mitgliedern der AG Klebe danke für die nette und freundliche Arbeitsatmosphäre.

Ich danke *meiner ganzen Familie* für die großartige Unterstützung während der Promotionszeit. Ohne euch wäre diese Arbeit nicht möglich gewesen!

Ein ganz besonderer Dank gilt meinem Großvater *Konstantin Glinka*, der mein Interesse an naturwissenschaftlichen Fragestellungen, vor allem der Chemie, sehr stark geprägt hat.

Erklärung

Ich versichere, dass ich meine Dissertation

Optimization of Clustering and Database Screening Procedures for Cavbase and Virtual Screening for Novel Antimalarial and Antibacterial Molecules

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

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

Marburg, den

....
(Serghei Glinca)

Curriculum vitae

Persönliche Angaben

Geburtsdatum	24.05.1983
Geburtsort	Chișinău/Republik Moldau
Staatsangehörigkeit	deutsch
Familienstand	verheiratet, 1 Kind

Ausbildung

1995 – 1998	**Georg-August-Zinn-Schule**, *Gesamtschule - Europaschule Kassel*.
1998 – 2001	**Goetheschule**, *Gymnasium Kassel*.
2001 – 2003	**Herderschule**, *Gymnasium Kassel*.
11/06/2003	**Abiturabschluss**.
04/2004 – 03/2008	**Pharmaziestudium**, *Philipps Universität Marburg*.
05/2008 – 10/2008	**Engel-Apotheke am Rathaus Kassel**.
12/2008 – 05/2009	**Wissenschaftlicher Mitarbeiter**, *Institut für pharmazeutische Chemie, Philipps Universität Marburg*, Arbeitskreis Prof. Dr. G. Klebe.
17/07/2009	**Erteilung der Approbation als Apotheker**.
07/2009 – 10/2012	**Promotion**, *Institut für pharmazeutische Chemie, Philipps Universität Marburg*, Arbeitskreis Prof. Dr. G. Klebe.