Computer Aided Drug Discovery

Descriptor Improvement and Application to Obesity-related Therapeutics

Von der Fakultät für Biowissenschaften, Pharmazie und Psychologie

der Universität Leipzig

genehmigte

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium

Dr. rer. nat.

vorgelegt

von

Master in Pharmacology Gregory Richard Sliwoski

geboren am 30.07.1981 in Rochester, New Hampshire, USA

Dekan: Prof. Dr. Erich Schröger

Gutachter: Prof. Dr. Annette Beck-Sickinger

Prof. Dr. Vsevolod Gurevich

Tag der Verteidigung 04.12.2015

For my father and my mother

Table of Contents

Table of Contents iv

Abbreviations viii

Bibliographische Darstellung xiii

Summary xiv

Zusammenfassung xx

Background and Hypothesis xxvii

Chapter 1 29

1.1 Abstract 30

1.2 Introduction 31

1.2.1 Position of CADD in the drug discovery pipeline 33

1.2.2 Ligand databases for CADD 36

1.2.3 Preparation of Ligand Libraries for CADD 37

1.2.4 Representation of small molecules as “SMILES” 38

1.2.5 Small Molecule Representations for Modern Search Engines: InChIKey 39

1.2.6 Target databases for CADD 40

1.2.7 Benchmarking Techniques of CADD 41

1.3 Structure-Based Computer-Aided Drug Discovery (SB-CADD) 42

1.3.1 Preparation of a Target Structure 43

1.3.2 Binding site detection and characterization 48

1.3.3 Representing Small Molecules and Target Protein for Docking Simulations 51

1.3.4 Sampling Algorithms for Protein-Ligand Docking 52

1.3.5 Scoring Functions for Evaluation Protein-Ligand Complexes 58

1.3.6 Structure-Based virtual High-Throughput Screening 60

1.3.7 Atomic-detail / High Resolution Docking 61

1.3.8 Binding Site Characterization 62

1.3.9 Pharmacophore Model 64

1.3.10 Automated de novo Design of Ligands 66

1.4 Ligand-Based Computer-Aided Drug Design 70

1.4.1 Molecular Descriptors / Features 71

1.4.2 Molecular fingerprint and similarity searches 81

1.4.3 Quantitative Structure Activity Relationship models 84

1.4.4 Selection of optimal descriptors/features 93

1.4.5 Pharmacophore mapping 95

1.5 Prediction and Optimization of Drug Metabolism and Pharmacokinetics Properties Including Absorption, Distribution, Metabolism, Excretion, and the Potential for Toxicity Properties 104

1.5.1 Compound Library Filters 104

1.5.2 Lead improvement: metabolism and distribution 106

1.5.3 Prediction of human Ether-a-go-go related gene binding 109

1.5.4 Drug Metabolism and Pharmacokinetics/Absorption, Distribution, Metabolism, and Excretion and the Potential for Toxicity Prediction Software Packages and Algorithms 112

1.5.5 Drug Metabolism and Pharmacokinetics/Absorption, Distribution, Metabolism, and Excretion and the Potential for Toxicity: Clinical Trial Prediction and Dosing 113

1.6 Conclusions 114

1.7 References 118

Chapter 2 145

2.1 Abstract 146

2.2 Introduction 147

2.3 Results and Discussion 150

2.4 Conclusions 161

2.5 Experimental Methods 162

2.6 References ... 166

2.7 Supplementary information ... 169

Chapter 3 ... 171

3.1 Abstract ... 172

3.2 Introduction ... 173

3.3 Results ... 177

3.4 Discussion ... 183

3.5 Methods ... 185

3.6 References ... 186

Chapter 4 ... 189

4.1 Abstract ... 190

4.2 Introduction ... 191

4.3 Results ... 192

4.4 Discussion ... 199

4.5 Materials and Methods ... 201

4.6 References ... 203

4.7 Supplementary Information ... 206

Chapter 5 ... 209

5.1 Abstract ... 210

5.2 Introduction ... 211

5.3 Results ... 215

5.4 Discussion ... 221

5.5 Methods ... 222

5.6 References ... 224

5.7 Supplementary Information ... 229

Chapter 6 ... 241

6.1 Abstract 242
6.2 Introduction 243
6.3 Results 246
6.4 Discussion 255
6.5 Methods 258
6.6 References 265
6.7 Supplementary Information 269
Chapter 7 275
7.1 Future Directions 276
7.2 Closing Remarks 281
7.3 References 283
Appendix 285
Introduction 286
Results and Discussion 288
Methods 291
References 292
Publications and Presentations 296
Curriculum Vitae 298
SELBSTSTÄNDIGKEITSERKLÄRUNG 299
Acknowledgements 300

Abbreviations

2DA	2D auto-correlation
3DA	3D auto-correlation
ACD	Available Chemical Directory
ACE	angiotensin-converting enzyme
ACTH	adrenocorticotrophic hormone
ADMET	adsorption, distribution, metabolism, excretion, and toxicity
ADR	adverse drug reaction
AgRP	agouti related hormone
ANN	artificial neural network
AUC	area under the curve
BCL	BioChemical Library
CADD	computer aided drug discovery
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CCD	cyclic coordinate descent
CGB	corticosteroid binding globulins
CIP	Cahn-Ingold Prelog
CoMFA	comparative molecular field analysis
CoMSIA	comparative molecular similarity indices
CPE	chemical penetration enhancers
DAS	directional asymmetry score

DMPK	drug metabolism and pharmacokinetics
DUD	Directory of Useful Decoys
DVC	dorsal vagal comlex
ECL	extracellular loops
EMAS	enantioselective molecular asymmetry descriptor
FEFF	free energy force field
FFS	forward-feature selection
FPP	fraction positive predictions
GCOD	grid cell occupancy descriptor
GDB	Generating a DataBase
GDT-TS	global distance test-total scores
GLP-1	glucagon-like peptide-1
GOLD	Genetic Optimization for Ligand Docking
GPCR	G-protein coupled receptor
GRIND	grid-independent descriptor
HCV	hepatitis C virus
hERG	human ether-a-go-go related gene
HIV	human immunodeficiency virus
HTS	high-throughput screen
ICL	intracellular loops
ICM	internal coordinate mechanics

IN	integrase
InChI	International Chemical Identifier
ISIS	Integrated Scientific Information System
IUPAC	International Union of Pure and Applied Chemistry
KIC	kinematic loop closure
LB-CADD	ligand-based computer aided drug discovery
LiBERO	ligand-guided backbone ensemble receptor optimization algorithm
logAUC	AUC with logarithmic x-axis
LOO	leave-one-out cross-validation
MACCS	Molecular Access System
mAChR	muscarinic acetylcholine receptor
MAPK	mitogen-activated protein kinase
MC	Monte Carlo
MC4R	melanocortin 4 receptor
MCM	Monte Carlo with Metropolis criterion
MD	molecular dynamics
MDL	Molecular Design Limited
mGluR	metabotropic glutamate receptor
MIF	molecular-interaction field
MLR	multivariate linear regression
mQSAR	multidimensional QSAR

MSH	melanocyte stimulating hormone
NAM	negative allosteric modulation
NDP-MSH	[Nle4,dPhe7]-MSH
NPY	neuropeptide Y
NST	nucleus of the solitary tract
p38 MAPK	p38 mitogen-activated protein kinase
P450	cytochrome P450
PAM	positive allosteric modulator
PCA	principal component analysis
PDB	Protein DataBank
PEOE	partial equalization of orbital electronegativity
PLA_2	phospholipase A_2
PLS	partial least squares
POMC	proopiomelanocortin
PP	pancreatic polypeptide
PPV	positive predictive value
PPAR	peroxisome proliferator-activated receptor
PYY	peptide tyrosine tyrosine
QSAR	quantitative structure-activity relationship
RDF	radial distribution function
REU	Rosetta energy units

RMSD	root-mean-square deviation
ROC	receiver operating characteristic
ROCS	rapid overlay of chemical structures
SB-CADD	structure-based computer aided drug discovery
SEA	similarity ensemble approach
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line System
SVM	support vector machine
SYNOPSIS	synthesize and optimize systems in silico
TGF-β1	transforming growth factor-β1
TM	transmembrane
VDW	van der Waals
VEGF	vascular endothelial growth factor
VEGFR2	vascular endothelial growth factor receptor 2
vHTS	virtual-HTS
WOMBAT	World of Molecular Bioactivity
Y_1R	Y_1 receptor
Y_2R	Y_2 receptor
Y_4R	Y_4 receptor
Y_5R	Y_5 receptor

Bibliographische Darstellung

Gregory Richard Sliwoski

Computer Aided Drug Discovery – Descriptor Improvement and Applications to Obesity-related Therapeutics

Universitat Leipzig, Fakultat für Biowissenschaften, Pharmazie und Psychologie

Dissertation

310 Seiten, 825 Literaturangaben, 39 Abbildungen, 17 Tabellen

---

When applied to drug discovery, modern computational systems can provide insight into the highly complex systems underlying drug activity and predict compounds or targets of interest. Many tools have been developed for computer aided drug discovery (CADD), focusing on small molecule ligands, protein targets, or both. The aim of this thesis is the improvement of CADD tools for describing small molecule properties and application of CADD to several stages of drug discovery regarding two targets for the treatment of obesity and related diseases: the neuropeptide Y_4 receptor (Y_4R) and the melanocortin-4 receptor (MC4R).

In the first chapter, the major categories of CADD are outlined, including descriptions for many of the popular tools and examples where these tools have directly contributed to the discovery of new drugs. Following the introduction, several improvements for encoding stereochemistry and signed property distribution are introduced and tested in scenarios meant to simulate applications in virtual high-throughput screening. Y_4R and MC4R are both class A G-protein coupled receptors (GPCRs) with endogenous peptide ligands that play critical roles in the signaling of satiety and energy metabolism. So far, no structures from either receptor family have been experimentally elucidated. CADD was combined with high-throughput screening (HTS) to discover the first small molecule positive allosteric modulators (PAMs) of Y_4R. Secondly, CADD techniques were used to model the interaction of Y_4R and pancreatic polypeptide based on experimental results that elucidate specific binding contacts. Similar SB-CADD approaches were used to model the interaction of MC4R with its high affinity peptide agonist α-MSH. Due to its role in monogenic forms of obesity, these models were used to predict which residues directly participate in binding and correlate mutated residues with their potential role in the binding site.

Summary

Drug discovery is a cornerstone of medical research that draws from many disciplines. In recent decades, these disciplines have extended into fields such as computer science, drawing from modern technology's ability to simulate complex processes that demand billions of calculations per second. Computer aided drug discovery (CADD) is a collective term for the many *in silico* methods being developed and applied to the discovery and design of new therapeutics. The goal of CADD is not to replace traditional *in vitro* and *in vivo* experimental techniques, but to supplement them. With CADD, it becomes possible to model complex processes and narrow the seemingly endless list of possible experiments to a manageable strategy designed for efficiency and cost effectiveness. **Chapter 1** introduces the two categories of CADD: ligand-based CADD (LB-CADD) and structure-based CADD (SB-CADD).

The overall focus of this dissertation is to improve three dimensional descriptors for use in quantitative structure-activity relationship (QSAR) models and apply LB-CADD and SB-CADD techniques to the modeling and drug discovery of two peptide binding class A G-protein coupled receptors (GPCRs) that represent promising therapeutic targets for obesity and related diseases.

Capturing stereochemistry in 3D-QSAR

QSAR descriptors encode physicochemical properties used to train models for predicting biological activity. The Radial distribution function (RDF) and 3D auto-correlation (3DA) are two commonly used 3D-QSAR descriptors that encode the geometry and distribution of properties within a molecule. The major difference between 3DA and RDF is the smoothing function applied by RDF to compensate for positional uncertainty caused by bond vibration and minor conformational changes. One of the major disadvantages of the RDF and 3DA descriptors is their failure to differentiate certain stereoisomers. This can hinder QSAR model performance when enantiomer pairs have different biological activities or toxicities.

Chapter 2 presents the enantioselective molecular asymmetry descriptor (EMAS), a 3D-QSAR descriptor that implicitly distinguishes between enantiomers. Traditionally, stereoisomers are distinguished with the Cahn-Ingold-Prelog (CIP) ruleset. However, CIP is not sufficient to cover all cases of stereochemistry and suffers from limited application in QSAR. EMAS avoids the limitations of CIP by encoding the overall stereochemistry of a molecule and implicitly distinguishing between enantiomer pairs using geometric

properties rather than rulesets. EMAS takes advantage of the transformation-invariant and smoothing properties of RDF by applying a similar iterative framework and calculates an asymmetry score for each atom triplet. This asymmetry score captures the direction and extent of asymmetry by combining triplets with the molecule's geometric center to create tetrahedrons that vary in shape and volume.

The utility of EMAS was evaluated against a small dataset of 31 compounds commonly used to evaluate novel stereochemistry descriptors. Artificial neural network (ANN) models trained with EMAS performed as well as or better than half of the previously published stereochemistry descriptors. Although EMAS did not outperform all published methods, the broad applicability of EMAS makes it an attractive descriptor since it is the only stereochemistry descriptor that does not require molecule superimposition or ruleset-based identification of stereocenters.

To evaluate the utility of EMAS with large datasets, ANN models were trained over a high throughput screening (HTS) dataset for inhibitors or substrates of cytochrome P450 2D6. Models trained with feature sets including EMAS were able to predict active compounds with success rate increase of approximately 11.7% compared to models trained without EMAS descriptors.

Improving 3D descriptors to avoid information loss: 3DA_Sign and other modifications

Chapter 3 presents modifications to 3DA designed to avoid several sources of information loss. As mentioned, the fundamental difference between RDF and 3DA is the application of Gaussian smoothing. This smoothing has the potential to increase descriptor performance but in its traditional implementation, RDF leads to underrepresentation of atom pair distances falling between distance centers. A 3DA/RDF hybrid descriptor called 3DA_Smooth was designed to apply Gaussian smoothing to 3DA to avoid this problem with RDF. ANN model's trained with 3DA, RDF, or 3DA_Smooth showed comparable prediction success across nine HTS datasets with varying target protein classes. Because the application of smoothing increases computational demand, these results suggest that the extra cost of RDF and 3DA_Smooth does not increase model performance and 3DA may be used in place of RDF.

Secondly, **chapter 3** presents a variation of 3DA called 3DA_Sign. 3DA_Sign is designed to avoid the loss of information that can occur when weighting a 3DA with signed atom properties. Traditionally, property weighting coefficients are calculated as the product of two atom properties. When atom properties are signed, this can lead to information loss as the product of two negative properties is equal to the product of two equivalent positive properties. 3DA_Sign splits all atom property pairs into one of three curves: negative-negative, positive-positive, and opposite signs. ANN models trained using 3DA_Sign for signed atom properties outperformed models trained with standard 3DA. This increase in performance was seen for all HTS datasets with an average prediction success rate increase of approximately 4.4%.

Lastly, a variation of 3DA that limits maximum atom pair distance encoding to six angstrom was tested. The commonly applied 3DA cutoff of twelve angstroms captures the maximum width of most small molecules. However, conformational flexibility leads to higher chances of variability in distant atom pairs. Because 3DA encodes a single conformation of each molecule, atom pair distances that do not reflect active conformations can hinder model performance. The reduced distance cutoff was designed to focus on molecule fragments less susceptible to this problem. Models trained with a distance cutoff of six angstroms outperformed models trained with a distance cutoff of twelve angstroms across all HTS datasets with an average increase in prediction success of approximately 6.4%.

Applications of computer aided drug design to the discovery of obesity therapeutics

Chapters 4 through **6** present different applications of CADD to the discovery of novel obesity therapeutics. Obesity is a medical problem that has doubled in worldwide prevalence over the past several decades and is a major risk factor for diabetes, heart disease, cancer, and mortality. Currently, the most effective treatment for obesity is bariatric surgery and less invasive pharmacological approaches have seen moderate to little success. Two potential therapeutic targets are explored with methods from LB-CADD and SB-CADD: the neuropeptide Y_4 receptor and the melanocortin 4 receptor.

Discovery of the first positive allosteric modulators of the human Y_4 receptor

Hormonal changes following bariatric surgery have become promising pharmacological targets due to their contribution to the long term effect of this surgery. The neuropeptide Y_4 receptor (Y_4R) is one such target with its endogenous agonist pancreatic polypeptide (PP) acting as a satiety factor released in response and in proportion to food intake. To date, no small molecule potentiators of Y_4R have been published nor has the three dimensional structure of any neuropeptide Y receptors been elucidated. At this stage in the drug discovery process CADD may be applied in several beneficial ways.

Chapter 4 presents the first small molecule positive allosteric modulators (PAMs) of Y_4R. High throughput screening was coupled with fingerprint-based cheminformatics to discover five Y_4R PAMs sharing a common scaffold. Further verification and Y receptor selectivity analyses were performed with an orthogonal IP accumulation assay.

The first screen of two thousand compounds yielded niclosamide as the hit of greatest interest. Fingerprint-based similarity was used to enrich the second set of screened compounds for those that were structurally similar to niclosamide. Compounds in the Vanderbilt Institute for Chemical Biology library were compared with niclosamide using the Tanimoto coefficient measure. This measure compares the average occupancy similarity of two compounds' molecular fingerprints for shared functional groups and overall geometric features.

The second screen of 33,288 compounds was enriched with 1,288 niclosamide-similar compounds, yielding four verified hits structurally similar to niclosamide. These compounds showed varying selectivity profiles across different neuropeptide Y receptor subtypes, allowing for development of preliminary structure-activity relationships around this scaffold that suggest an electron-rich substituent on the benzoyl ring important for Y_4R potency and a nitro-benzoyl substitution that decreases potency at Y_1R.

Structure-based computational modeling of Y_4R and PP

Chapter 5 presents the application of structure-based computational methods to model the interaction of Y_4R and PP. This project involves a collaboration combining several rounds of complimentary *in vitro* cellular assays and *in silico* modeling. The primary role of the presenting author was the application of computational modeling. Mutagenesis and cell-based assays were performed by Xavier Pedragosa-Badia and Diana Lindner of the Beck-Sickinger laboratory and have been published along with the computational models in a paper titled "Pancreatic polypeptide is recognized by two hydrophobic domains of the Y4 receptor binding pocket". Chapter 5 focuses and expands on the computational strategies applied and resulting models.

Comparative modeling with the Rosetta Molecular Modeling Suite was used to take advantage of the shared topology of class A GPCRs for modeling Y_4R. Highly disordered regions with low sequence conservation were extensively remodeled with cyclic coordinate descent (CCD) and refined with kinematic loop closure (KIC).

The NMR structure ensemble of PP reveals a structured α-helix and highly flexible C-terminal region. Therefore, the rigid helix portion of PP was docked to Y_4R first using standard protein-protein docking.

Because this protocol is unable to capture the extensive flexibility of the loop regions, the PP helix was docked in the absence of extracellular loops to avoid interference from rigid loops. An experimentally derived potential contact between Tyr$^{2.64}$ of Y_4R and Tyr^{27} of PP was used to guide helix docking.

Unlike the helix region, the five C-terminal residues of PP exhibit substantial conformational flexibility. Therefore, Rosetta's *de novo* folding was used to comprehensively model these five residues in the presence of the Y_4R helices. *In vitro* results provided three contacts between PP and Y_4R that guided modeling. Finally, CCD and KIC were used to reconstruct the extracellular loops of Y_4R in the presence of PP. This approach is designed to capture changes in loop conformations that may occur when PP binds to Y_4R.

An ensemble of nine energetically comparable high-resolution models of PP and Y_4R was generated that captured experimentally determined interactions including a salt bridge between Asp$^{6.59}$ and Arg^{35}, a hydrogen bond between Asn$^{7.32}$ and Arg^{33}, and cation-pi interactions between Phe$^{7.35}$ and Arg^{33}. Residue contacts were examined across all conformations to propose potential interactions beyond those previously explored. One such putative contact between Ser$^{5.28}$ of Y_4R and Thr^{32} of PP presents a target for future mutagenesis studies.

Modeling the interaction of the melanocortin 4 receptor and α-MSH

The melanocortin 4 receptor (MC4R) is a promising target for the treatment of obesity due to its contributions to monogenic forms of obesity. Approximately 150 naturally occurring MC4R gene mutations have been identified among obese patients and MC4R deficiency is characterized by hyperphagia, increased adiposity, and severe hyperinsulinemia. Anorexigenic signaling from α-MSH activation of MC4R appears to be critical for the regulation of feeding and metabolism. Binding studies with α-MSH have revealed several critical interactions between MC4R and α-MSH. However, the flexibility of α-MSH, a 13 amino acid linear peptide only restrained by a single reverse β-turn around the central residues, makes it difficult to elucidate a precise binding pose.

Chapter 6 presents a comparative modeling and docking approach that is tailored to the flexibility of α-MSH. A comparative modeling application recently added to Rosetta called RosettaCM was used for its hybrid multi-template approach. After modeling the MC4R with RosettaCM, experimental evidence guided a two–phase docking approach where the central region of the peptide was docked followed by the remodeling of flexible terminals in parallel with the receptor's extracellular loops. In α-MSH, a core tetrapeptide spanning residues 6-9 was shown to be critical and sufficient for activation of MC4R. Mutant binding assays consistently reveal two binding sites: an acidic pocket facing Arg^{8} of α-MSH including MC4R residues Glu$^{2.60}$, Asp$^{3.25}$, and Asp$^{3.29}$ and a hydrophobic interaction between Phe$^{6.51}$ of MC4R and Phe^{7} of α-MSH.

Rosetta FlexPepDock was used to dock this tetrapeptide. Extracellular loop regions of MC4R and flexible terminals of α-MSH were modeled and refined simultaneously using the same combination of CCD and KIC as in chapter 5. An additional restraint within α-MSH was used during the loop building phase to enforce the active conformation β-turn.

Models reveal convergence to a single binding pose of the central tetrapeptide region of α-MSH and significant conformational flexibility in the terminal regions. Examination of an ensemble of energetically-comparable conformations shows a binding interface on MC4R that encompasses three receptor regions: residues from transmembrane helices two and three that contact Arg^{8}, transmembrane helices six and seven and extracellular loop three that contact Glu^{5} and His^{6}, and transmembrane helices four and five and extracellular loop two that contact Trp^{9}. Phe^{7} of α-MSH, considered to contain the most important pharmacophore for activation of MC4R, points downwards into the transmembrane pore in all models, engaging residues from transmembrane helices three, six, and seven.

This ensemble approach identified twelve binding interactions in addition to the four used to guide docking. These interactions were compared with previously published binding assay results, eight of which are supported with published experimental results. Additionally, these models were used to propose a previously unidentified contact between Met$^{7.32}$ of MC4R and Ser^{4} or Glu^{5} of α-MSH.

In summary, this thesis presents improvements and applications for both categories of CADD with two therapeutic targets for obesity: Y$_4$R and MC4R. **Chapter 7** serves as a closing chapter that presents future projects to integrate results from different chapters and methods in this dissertation. An **appendix** following chapter 7 describes preliminary results that integrate chapters 4 and 5 to dock niclosamide to Y$_4$R-PP models. This will help to lay important groundwork for future studies aimed at elucidating an allosteric binding site on Y$_4$R and improving future drug discovery endeavors.

Zusammenfassung

Arzneimittelforschung ist ein essentieller Bestandteil der Medizinforschung und kombiniert verschiedene Disziplinen. In den vergangenen Jahrzehnten wurde sie um Bereiche wie Informatik erweitert, um von deren modernen Methoden zur Simulierung komplexer Prozesse, die Milliarden von Berechnungen pro Sekunde erfordern, zu profitieren. Computer Aided Drug Discovery (CADD) ist ein Sammelbegriff für verschiedene in-silico-Methoden, welche für die Entdeckung und Entwicklung neuer Therapeutika angewendet werden. Das Ziel von CADD ist nicht herkömmliche in vitro und in vivo Techniken zu ersetzen, sondern deren Ergänzung. CADD ermöglicht die Eingrenzung einer scheinbar endlosen auf eine überschaubare Zahl an möglichen Experimenten, und eine auf Effizienz und Wirtschaftlichkeit ausgelegte Strategie. **Kapitel 1** liefert eine Einführung in die zwei CADD-Kategorien: Ligand-basierte CADD (LB-CADD) und struktur-basierte CADD (SB-CADD).

Der Fokus der vorliegenden Arbeit ist die Verbesserung dreidimensionaler Deskriptoren für den Einsatz in Modellen für quantitative Struktur-Wirkungs-Beziehungen (QSAR), sowie die Anwendung von LB-CADD und SB-CADD zur Modellierung und Wirkstoffforschung von zwei peptidbindenden G-Protein-gekoppelten Rezeptoren (GPCR), welche vielversprechende therapeutische Ziele für Adipositas und verwandte Erkrankungen sind.

Beschreibung von Stereochemie in 3D-QSAR

QSAR-Deskriptoren kodieren physikochemische Eigenschaften, welche verwendet werden um Modelle für die Vorhersage biologischer Aktivität zu trainieren. Die radiale Verteilungsfunktion (RDF) und 3D-Autokorrelation (3DA) sind zwei häufig verwendete 3D-QSAR-Deskriptoren, welche die Geometrie und die Verteilung von Eigenschaften innerhalb eines Moleküls kodieren. Der Hauptunterschied zwischen 3DA und RDF ist die durch RDF angewandte Glättungsfunktion um die durch Bindungsschwingungen und kleinere Konformationsänderungen verursachte Positionsunsicherheit zu kompensieren. Einer der Hauptnachteile der RDF- und 3DA-Deskriptoren ist ihre Unfähigkeit, bestimmte Stereoisomere zu unterscheiden. Dies kann die Leistung von QSAR-Modellen beeinträchtigen, wenn Enantiomerenpaare unterschiedliche biologische Aktivitäten oder Toxizitäten aufweisen.

Kapitel 2 stellt den enantioselektiven, molekularen Asymmetrie Deskriptor (EMAS) vor, ein 3D-QSAR-Deskriptor, welcher implizit zwischen Enantiomeren unterscheidet. Traditionell werden Stereoisomere durch den Cahn-Ingold-Prelog (CIP) Regelsatz unterschieden. Jedoch deckt CIP nicht alle Fälle der

Stereochemie ab und leidet unter seiner begrenzten Anwendung in QSAR. EMAS umgeht die Einschränkungen des CIP-Regelsatzes durch Kodierung der allgemeinen Konfiguration eines Moleküls und implizite Unterscheidung zwischen Enantiomerenpaaren anhand geometrischer Eigenschaften und nicht anhand von Regelsätzen. EMAS nutzt die Transformationsinvarianz und Glättungseigenschaften der RDF durch Verwendung eines iterativen Rahmenwerks und berechnet die Asymmetrie jedes Atom-Tripletts. Dieser Asymmetrie-Score erfasst die Richtung und das Ausmaß der Asymmetrie durch Kombination der Tripletts mit der geometrischen Mitte des Moleküls zu Tetraedern, welche in Form und Volumen variieren.

Die Nützlichkeit des EMAS wurde auf Basis eines kleinen Datensatzes bestehend aus 31 Verbindungen evaluiert, welcher üblicherweise verwendet wird um neue Stereochemie-Deskriptoren zu evaluieren. Künstliche neuronale Netz (ANN) Modelle, welche EMAS verwenden, erzielten Ergebnisse, die vergleichbar zu zuvor publizierten Deskriptoren sind. Obwohl EMAS zuvor veröffentlichte Methoden nicht übertrifft, macht ihn seine breite Anwendbarkeit attraktiv, da EMAS für die stereochemische Beschreibung eines Moleküls weder eine strukturelle Überdeckung noch eine Regelsatz-basierte Identifizierung von Sterozentren erfordert. Um die Nützlichkeit des EMAS für großen Datensätze zu bewerten, wurden ANN-Modelle auf einem Hochdurchsatz-Screening (HTS) Datensatz für Inhibitoren oder Substrate des Cytochrom P450 2D6 trainiert. Modelle, welche mit Feature-Sets einschließlich EMAS trainiert wurden, konnten Wirkstoffe mit einer um ca. 11,7% höheren Erfolgsquote gegenüber Modellen ohne EMAS-Deskriptoren voraussagen.

Verbesserung von 3D-Deskriptoren um Informationsverlust zu vermeiden: 3DA_Sign und andere Modifikationen

In **Kapitel 3** werden Änderungen an 3DAs vorgestellt, um verschiedene Ursachen für Informationsverlust zu vermeiden. Wie erwähnt ist der grundlegende Unterschied zwischen RDF und 3DA die Anwendung der Gaußglättung. Diese Glättung hat das Potenzial die Leistung der Deskriptoren zu verbessern, führt in ihrer traditionellen Implementierung jedoch zur Unterrepräsentierung von Atompaaren, deren Abstand zwischen den Zentren der Abstanden liegt. 3DA_Smooth, ein 3DA/RDF-Hybrid-Deskriptor, wurde entwickelt, um eine Gaußglättung auf 3DAs anzuwenden um dieses Problem mit RDFs zu umgehen. ANN-Modelle, welche mit 3DA, RDF oder 3DA_Smooth trainiert wurden, weisen in neun HTS-Datensätzen mit unterschiedlichen Zielprotein-Klassen vergleichbare Vorhersageergebnisse auf. Da die Anwendung der Glättung erhöhten Rechenbedarf mit sich zieht, legen diese Ergebnisse nahe, dass die zusätzlichen Rechenkosten für RDF und 3DA_Smooth die Modellleistung nicht erhöhen und 3DA anstelle von RDF verwendet werden kann.

Als zweites stellt **Kapitel 3** eine Variante der 3DA, genannt 3DA_Sign, vor. 3DA_Sign wurde entwickelt, um Informationsverlust zu vermeiden, welcher bei der Gewichtung eines 3DA mit vorzeichenbehafteten Atomeigenschaften auftreten kann. Traditionell werden Gewichtungskoeffizienten für Eigenschaften als das Produkt von zwei Atomeigenschaften berechnet. Wenn Atomeigenschaften vorzeichenbehaftet sind, kann dies zu Datenverlust führen, da das Produkt von zwei negativen Eigenschaften gleich dem Produkt von zwei gleichartigen, positiven Eigenschaften ist. 3DA_Sign unterteilt alle Atomeigenschaftspaare in eine von drei Kurven: negativ-negativ, positiv-positiv und entgegengesetzte Vorzeichen. ANN-Modelle, welche mit 3DA_Sign für vorzeichenbehaftete Atomeigenschaften trainiert wurden, übertrafen Modelle, welche mit Standard-3DA trainiert wurden. Die Leistungssteigerung konnte für alle HTS-Datensätze beobachtet werden und betrug rund 4,4%.

Abschließend wurde eine Variante des 3DA, welche den kodierten Maximal-Abstand zwischen Atompaaren auf sechs Angström begrenzt, getestet. Die am häufigsten angewendeten 3DA Cutoffs von zwölf Angström decken zwar die maximale Breite der meisten kleinen Molekülen ab. Allerdings führt Konformationsflexibilität zu höheren Wahrscheinlichkeiten für Variabilität in entfernten Atompaaren. Da 3DA einzelne Konformationen eines Moleküls kodiert, können Entfernungen zwischen Atompaaren, welche die aktiven Konformationen nicht widerspiegeln, die Modellleistung beeinträchtigen. Der verringerte Abstands-Cutoff wurde für Molekülfragmente entwickelt, welche weniger anfällig für dieses Problem sind. Modelle die mit einem Abstands-Cutoff von sechs Angström trainiert wurden, übertrafen Modelle mit einem Abstand von zwölf Angström in allen HTS-Datensätzen mit einer durchschnittlichen Verbesserung des Vorhersageerfolges von rund 6,4%.

Anwendungen des computergestützten Wirkstoffdesigns zur Entdeckung von Therapeutika gegen Fettleibigkeit.

Kapitel 4 bis **6** zeigen verschiedene Anwendungen des CADD zur Entdeckung neuartiger Therapeutika gegen Fettleibigkeit. Übergewicht ist ein medizinisches Problem, das sich in weltweiter Prävalenz in den letzten Jahrzehnten verdoppelt hat und ist ein wichtiger Risikofaktor für Diabetes, Herzerkrankungen, Krebs und Mortalität. Derzeit sind die effektivsten Behandlungen für Fettleibigkeit Chirurgie. Weniger invasive, pharmakologische Ansätze, haben bisher nur zu schwachem bis mäßigem Erfolg geführt. Der Neuropeptid-Y_4-Rezeptor und der Melanocortin 4-Rezeptor sind in diesem Zusammenhang zwei potenzielle therapeutische Ziele. Beide Rezeptoren wurden in dieser Arbeitmit Methoden aus der LB-CADD und SB-CADD untersuchen.

Entdeckung der ersten positiven allosterischen Modulatoren des menschlichen Y_4-Rezeptors

Hormonelle Veränderungen folgend im Anschluß an eineder Adipositaschirurgie haben durch ihren Beitrag zum langfristigen Effekt dieser Operation Hinweise auf sich in vielversprechende pharmakologische Targets entwickeltgeliefert durch ihren Beitrag zum langfristigen Effekt dieser Operation. Der Neuropeptid-Y_4-Rezeptor (Y_4R) ist ein solches Ziel mit seinem endogenen Agonisten Pankreatischess-Polypeptid (PP), welcher welches als Sättigungsfaktor agiert und im Verhältnis zur Nahrungsaufnahme freigesetzt wird. Bisher wurden keine kleinen Molekülniedermolekularen Potentiatoren von des Y_4R veröffentlicht. Auch die, noch wurde die dreidimensionale Struktur konnte bisher von keinem dervon jedem Neuropeptid Y-Rezeptoren bestimmt werden. Zu diesem Zeitpunkt in der Wirkstoffentwicklung kann CADD kann auf mehrere vorteilhafte Weisen eingesetzt werden.

Kapitel 4 präsentiert die ersten niedermolekularen positiv allosterischen Modulatoren (PAMs) von Y_4R. In einem Hochdurchsatz-Screening wurden fünf Y_4R PAMs mit gemeinsamer Grundstruktur identifiziert. Studien zur Affinität, Potenz und Rezeptorelektivität wurden mit einem orthogonalen IP Akkumulationsassay durchgeführt.

Der erste Screen von 2000 Verbindungen ergab Niclosamid als interessantesten Treffer. Eine Suche nach Niclosamid-ähnlichen Verbindungen in der gesamten Substanzbibliothek des Vanderbilt-Instituts für chemische Biologiesollte einen zweiten Screen mit Verbindungen anreichern, welche eine dem Niclosamid ähnliche Grundstruktur aufweisen. Der Grad der Ähnlichkeit wurde hierbei mit dem Tanimoto-Score bewertet. Diese Messung vergleicht die durchschnittliche Ähnlichkeit der beiden Verbindungen durch den molekularen Fingerabdruck für gemeinsame funktionellen Gruppen und geometrische Merkmale.

Der zweite Screen von 33.288 Verbindungen wurde so mit 1.288 Niclosamid-ähnlichen Verbindungen angereichert, was in vier überprüften Treffern resultierte, welche strukturelle Ähnlichkeit zu Niclosamid aufwiesen. Diese Verbindungen zeigten variierende Selektivitätsprofile an den verschiedenen Neuropeptid-Y-Rezeptor-Subtypen, was die Entwicklung von vorläufigen Struktur-Wirkungsbeziehungen um dieses Gerüst ermöglichte, welches die Wichtigkeit eines elektronenreichen Substituenten am

Benzoylrings fuer eine hohe Y_4R Potenz vermuten laesst, wohingegen eine Nitrobenzoylsubstitution potenzverringernd am Y_1R zu wirken scheint.

Kapitel 5 stellt die Anwendung struktur-basierter Rechenmethoden zur Modellierung von Y_4R und PP vor. Dieses Projekt besteht aus einer Kombination von multiplen Iterationen aus zellulären in vitro Assays und in silico Modellierung. Die vornehmliche Rolle des Autors war die computergestützte Modellierung. Mutagenese und zellbasierte Assays wurden von Xavier Pedragosa-Badia und Diana Lindner im Labor von Frau Prof. Beck-Sickinger durchgeführt und gemeinsam mit den Rechenmethoden als “Pancreatic polypeptide is recognized by two hydrophobic domains of the Y4 receptor binding pocket” publiziert. Somit beschäftigt sich Kapitel 5 vornehmlich mit angewandten Rechenmethoden und deren Resultaten.

Um davon zu profitieren, dass Klasse A GPCRs dieselbe Topologie haben, wurde Y_4R durch Comparative Modelling mit der „Rosetta Molecular Modeling Suite“ modelliert. Stark unstruktierte Regionen mit geringer Sequenzkonservierung wurden ausführlich durch cyclic coordinate descent (CCD) remodelliert und durch kinematic loop closure (KIC) optimiert.

Das NMR-Struktur-Ensemble von PP weist eine strukturierte α-Helix und eine hochflexible C-terminale Region auf. Daher wurde der starre helikale Abschnitt von PP zuerst unter Verwendung von Standard-Protein-Protein-Docking an Y_4R angedockt. Da dieses Protokoll nicht die umfassende Flexibilität der Loop-Regionen erfasst, wurde die PP-Helix in der Abwesenheit von extrazellulären Loops angedockt, um Interferenzen durch starre Loops zu vermeiden. Ein experimentell abgeleiteter potentieller Kontakt zwischen $Tyr^{2.64}$ von Y_4R und Tyr^{27} des PP wurde verwendet, um das Helix-Docking zu lenken.

Im Gegensatz zu der Helix-Region, weisen die fünf C-terminalen Reste von PP erhebliche Konformationsflexibilität auf. Daher wurde die Rosetta de novo Faltung verwendet, um diese Reste umfassend in Gegenwart der Y_4R Helices zu modellieren. In-vitro-Ergebnisse lieferten drei Kontakte zwischen PP und Y_4R, welche die Modellierung lenkten. Abschließend wurden CCD und KIC verwendet, um die extrazellulären Loops des Y_4R in Gegenwart von PP zu rekonstruieren. Dieser Ansatz wurde entwickelt, um Änderungen in den Loop-Konformationen, die auftreten können, wenn PP an Y_4R bindet, zu erfassen.

Ein Ensemble von neun energetisch vergleichbaren, hochauflösenden Modellen von PP und Y_4R, welches experimentell bestimmte Interaktionen mit einer Salzbrücke zwischen $Asp^{6.59}$ und Arg^{35}, einer Wasserstoffbrücke zwischen $Asn^{7.32}$ und Arg^{33} und Kation-Pi Wechselwirkungen zwischen $Phe^{7.\ 35}$ und Arg^{33} erfasst, wurde erzeugt. Weitere Kontakte wurden in allen Konformationen untersucht, um mögliche, zuvor nicht erkannte, Wechselwirkungen nahezulegen. Ein solcher vermeintlicher Kontakt zwischen $Ser^{5.28}$ von Y_4R und Thr^{32} von PP stellt ein Ziel für künftige Mutagenesestudien dar.

Modellieren der Interaktion zwischen Melanocortin-4-Rezeptor und α-MSH

Der Melanocortin-4-Rezeptor (MC4R) ist ein vielversprechendes Ziel für die Behandlung von Fettleibigkeit aufgrund seiner Beiträge zu monogenen Formen von Adipositas. Rund 150 natürlich vorkommende MC4R-Gen-Mutationen wurden bei adipösen Patienten identifiziert und MC4R-Mangel wird durch Hyperphagie, erhöhte Adipositas und schwere Hyperinsulinämie gekennzeichnet. Anorexigenes Signaling von α-MSH durch Aktivierung des MC4R scheint kritisch für die Regulierung der Nahrungsaufnahme und des Metabolismus sein. Bindungsstudien mit α-MSH haben mehrere kritische Wechselwirkungen zwischen MC4R und α-MSH enthüllt. Jedoch macht es die Flexibilität von α-MSH, einem 13 Aminosäure langem, linearen Peptid, welches nur durch einen einzigen Rückwärts β-turn um die zentralen Reste eingeschränkt wird, schwierig einen genauen Bindungsmodus zu bestimmen.

Kapitel 6 enthält einen Comparative Modelling- und Docking-Ansatz, der auf die Flexibilität von α-MSH zugeschnitten ist. Eine Modellierungsanwendung, welche kürzlich zu Rosetta hinzugefügt und RosettaCM genannt wurde, wurde für einen Hybrid-Multi-Template-Ansatz verwendet. RosettaCM mischt Templatefragmente um ein Modell zu erstellen, den energetisch günstige Abschnitte über verschiedene Templates ausnutzt. Da die Anwesenheit oder Abwesenheit von Resten wie Prolin und Glycine die Topologie und das Verhalten der einzelnen Helices signifikant verändern kann, ist dieser Hybridisierungsansatz besonders geeignet für GPCR Vergleichsmodellierung.

Nach der Modellierung des MC4R mit RosettaCM, lenkten experimentelle Daten einen zweiphasigen Docking Ansatz, bei dem die zentrale Region des Peptids gedockt wurde, gefolgt von einer Remodellierung der flexiblen Terminalregion parallel zu den extrazellulären Loops. In α-MSH, einem Kern-Tetrapeptid, wurde gezeigt, dass die Reste 6-9 kritisch und ausreichend für die Aktivierung des MC4R ist. Bindungsassays mit Mutanten offenbarten zwei Bindungsstellen: eine saure Tasche gerichtet auf Arg^8 von α-MSH, einschließlich den MC4R-Resten $Glu^{2.60}$, $Asp^{3.25}$ und $Asp^{3.29}$ sowie eine hydrophobe Wechselwirkung zwischen $Phe^{6.51}$ des MC4R und Phe^7 von α-MSH.

Rosetta FlexPepDock wurde verwendet, um dieses Tetrapeptid andocken. Extrazelluläre Loop-Regionen des MC4R und flexible Terminalregionen der α-MSH wurden modelliert und gleichzeitig verfeinert unter Verwendung der gleichen Kombination aus CCD und KIC wie in Kapitel 5. Eine zusätzliche Beschränkung in α-MSH während der Loopkonstruktion wurde verwendet, um die aktive Konformation des β-turns zu erzwingen.

Die Modelle zeigen eine Konvergenz auf eine einzige Bindungs-Pose der zentralen Tetrapeptid-Region des α-MSH und signifikante konformative Flexibilität in den Terminalbereichen. Die Untersuchung eines Ensembles von energetisch vergleichbaren Konformationen zeigt ein Bindungsepitop auf, das drei MC4R Rezeptorbereiche umfasst: Reste aus Transmembranhelices zwei und drei, die mit Arg^8 interagieren, Reste aus Transmembranhelices sechs und sieben und drei extrazellulären Loops, die mit Glu^5 und His^6 interagieren, und Reste aus Transmembranhelices vier und fünf und sowie extrazelluläre Loops, die mit Trp^9 interagieren. Phe^7 von α-MSH, von welchem vermutet wird, dass es die wichtigste Pharmakophor-Aktivierung von MC4R enthält, ist in allen Modellen nach unten in die Transmembranpore gerichtet und greift Reste aus den Transmembranhelices drei, sechs und sieben an.

Zusätzlich zu den vier Bindungsinteraktionen die für das Docking verwendet wurden, identifizierte der Ensembleansatz zwölf Bindungsinteraktionen. Diese Interaktionen wurden mit zuvor veröffentlichten Bindungsassayergebnissen verglichen: acht Interaktionen werden durch publizierte experimentellen Ergebnisse gestützt. Zusätzlich wurden diese Modelle verwendet, um eine bisher nicht identifizierte Interaktion zwischen $Met^{7.32}$ des MC4R und Ser^4 oder Glu^5 in α-MSH vorzuschlagen.

Zusammenfassend zeigt diese Arbeit Verbesserungen und Anwendungen für beide CADD-Kategorien auf mit zwei therapeutischen Zielen für Übergewicht: Y_4R und MC4R. **Kapitel 7** dient als Schlusskapitel, das zukünftige Projekte für die Integration der verschiedenen Kapitel und Methoden in dieser Arbeit vorstellt. Ein Anhang, welcher auf Kapitel 7 folgt, beschreibt die vorläufigen Ergebnisse für die Kombination von Ergebnissen aus Kapitel 4 und 5, um Niclosamid an Y_4R-PP-Modelle andocken. Dies wird helfen, wichtige Grundlagen für zukünftige Studien, welche auf die Aufklärung einer allosterischen Bindungsstelle in Y_4R und die Verbesserung der künftigen Wirkstoffforschung gerichtet sind, zu legen.

Background and Hypothesis

The aim of this thesis is the application of computer aided drug discovery (CADD) to the study of two potential targets for the treatment of obesity and related disease: the neuropeptide Y_4 receptor (Y_4R) and the melanocortin 4 receptor (MC4R). Because CADD applications are continuously evolving, a sub-aim of this thesis is the improvement of specific CADD applications with novel descriptors for reducing information loss.

CADD may be applied to many different stages of the drug discovery pipeline, aiding in the analysis of large datasets, prioritizing experiments, and proposing studies aimed at elucidating specific interactions or activities. One commonly used ligand-based CADD technique is called quantitative structure activity relationship (QSAR). In this technique, quantitative descriptors are generated for known active and inactive compounds designed to capture the physicochemical properties that give rise to their activity at a specific protein target. These descriptors are used to train models capable of predicting activity for previously untested compounds.

Three dimensional descriptors such as 3D autocorrelation (3DA) and radial distribution function (RDF) encode the spatial distribution of physicochemical properties within a molecule by iterating over all interatomic distances. 3DA and RDF are attractive descriptors because they are transformation independent and do not rely on molecule superimposition to compare 3D structure. However, their current implementation contains several potential sources of information loss. Specifically, enantiomers are indistinguishable with 3DA and RDF because the interatomic distances of enantiomer pairs are identical. This is problematic in drug discovery projects where opposite enantiomers display different activities. Another source of information loss comes with the inclusion of signed atom properties. To describe the spatial distribution of properties, the interatomic distances of 3DA and RDF are often weighted with the product of the two atom properties. When these properties are signed such as with partial charge, their products become incapable of distinguishing between negative and positive pairs. Novel descriptors based on the framework of 3DA and RDF can specifically address these shortcomings.

The Y_4R is a class A G-protein coupled receptor (GPCR) with strong anorexigenic potential. Its endogenous agonist pancreatic polypeptide (PP) is released from pancreatic islets in response and proportion to food ingestion to inhibit gastrointestinal peristalsis and relay anorexigenic signals. Specific CADD techniques can be applied to different aspects of Y_4R signaling to model the interaction of Y_4R and PP and accelerate discovery of small molecule modulators of Y_4R.

Despite its potential as a pharmacological target for the treatment of obesity, no small molecule agonists or potentiators of Y_4R have been described. High throughput screening (HTS) can be used to rapidly assess thousands of compounds for Y_4R activity. To complement HTS, CADD may be applied to enrich hit rates or prioritize compounds for screening based on similarity to known active compounds. Binary molecular fingerprints encode the presence or absence of specific geometric properties and functional groups as predefined bit strings. Once one or more active compounds have been identified, large compound libraries may be queried for structurally similar compounds and prioritized for screening.

Inherent challenges in the experimental elucidation of membrane protein structures have so far prevented the elucidation of many GPCR structures, including Y_4R. Therefore, structure-based CADD techniques such as comparative modeling and protein docking can help characterize the structure of Y_4R and interactions with PP. Comparative modeling uses known protein structures to guide the modeling of similar protein structures. Different GPCR types share common topology despite sometimes low sequence identity and are well suited for a multi-template comparative modeling approach that incorporates multiple GPCR structures into the prediction of a target structure. With a modeled structure of Y_4R, it becomes possible to model the interactions of Y_4R and PP based on experimentally determined residue contacts.

The MC4R is another class A GPCR that relays anorexigenic signals in response to its endogenous peptide agonist α-MSH. As with Y_4R, no three dimensional structure of MC4R is available. However, several contacts between MC4R and α-MSH have been experimentally elucidated. Therefore, a similar comparative modeling and protein docking approach can be used to model the structure of MC4R and interactions with α-MSH.

Chapter 1

Computational Methods in Drug Discovery

---

published in Pharmacological Reviews (2013) **66**, 334-395

Gregory Sliwoski, Sandeep Kothiwale, Jens Meiler, Edward Will Lowe: **Computational Methods in Drug Discovery**

1.1 Abstract

Computer-aided drug discovery/design methods have played a major role in the development of therapeutically important small molecules for over three decades. These methods are broadly classified as either structure-based or ligand-based methods. Structure-based methods are in principle analogous to high-throughput screening in that both target and ligand structure information is imperative. Structure-based approaches include ligand docking, pharmacophore, and ligand design methods. The article discusses theory behind the most important methods and recent successful applications. Ligand-based methods use only ligand information for predicting activity depending on its similarity/dissimilarity to previously known active ligands. We review widely used ligand-based methods such as ligand-based pharmacophores, molecular descriptors, and quantitative structure-activity relationships. In addition, important tools such as target/ligand databases, homology modeling, ligand fingerprint methods, etc., necessary for successful implementation of various computer-aided drug discovery/design methods in a drug discovery campaign are discussed. Finally, computational methods for toxicity prediction and optimization for favorable physiologic properties are discussed with successful examples from literature.

1.2 Introduction

On October 5, 1981, *Fortune* magazine published a cover article entitled the “Next Industrial Revolution: Designing Drugs by Computer at Merck” [1]. Some have credited this as the beginning of intense interest in the potential for Computer Aided Drug Design (CADD). While progress was being made in CADD, high-throughput screening (HTS) was taking priority as a means for finding novel therapeutics. This brute force approach relies on automation to screen high numbers of molecules in search of those which elicit the desired biological response. HTS has the advantage of requiring minimal compound design or prior knowledge and newer technologies make screening these large libraries efficient and relatively fast. However, while traditional HTS can result in multiple hit compounds, some of which are capable of being modified into a lead and then a novel therapeutic, the hit rate for HTS is often extremely low. This low hit rate limits the application of HTS to research programs capable of screening large compound libraries. In the past decade, CADD has reemerged as a way to significantly decrease the number of compounds necessary to screen while retaining the same level of lead compound discovery. Many compounds predicted to be inactive can be skipped and those predicted to be active can be prioritized. This reduces the cost and workload of a full HTS screen without compromising lead discovery. Additionally, traditional HTS assays often require extensive development and validation before they can be employed. Since CADD requires significantly less preparation time, experimenters can perform CADD studies while the traditional HTS assay is being prepared. The fact that both of these tools can be used in parallel provides an additional benefit for CADD in a drug discovery project.

For example, researchers at Pharmacia (now part of Pfizer) used CADD tools to screen for inhibitors of tyrosine phosphatase-1B, an enzyme implicated in diabetes. Their virtual screen yielded 365 compounds, 127 of which showed effective inhibition, a hit rate of nearly 35%. Simultaneously, this group performed a traditional HTS against the same target. Of the 400,000 compounds tested, 81 showed inhibition, producing a hit rate of only 0.021%. This comparative case effectively displays the power of CADD [2]. CADD has already been used in the discovery of compounds which have passed clinical trials and become novel therapeutics in the treatment of a variety of diseases. Some of the earliest examples of approved drugs that owe their discovery in large part to the tools of CADD include the carbonic anhydrase inhibitor dorzolamide, approved in 1995 [3], the angiotensin-converting enzyme (ACE) inhibitor captopril, approved in 1981 as an antihypertensive drug [4], three therapeutics for the treatment of HIV: saquinavir (approved in 1995), ritonavir, and indinavir (both approved in 1996) [1] and tirofiban, a fibrinogen antagonist approved in 1998 [5].

One of the most striking examples of the possibilities presented from CADD occurred in 2003 with the search for novel Transforming Growth Factor-β1 (TGF-β1) receptor kinase inhibitors. One group at Eli Lilly used a traditional HTS to identify a lead compound that was subsequently improved through structure activity relationship (SAR) studies using *in vitro* assays [6], whereas a group at Biogen Idec used a CADD approach involving virtual HTS based on the structural interactions between a weak inhibitor and TGF-β1 receptor kinase [7]. Through virtual screening, the group at Biogen Idec identified 87 hits, the best being identical in structure to the lead compound discovered through the traditional HTS approach at Eli Lilly [8]. In this example CADD, a method involving reduced cost and workload, was capable of producing the same lead as a full-scale HTS.

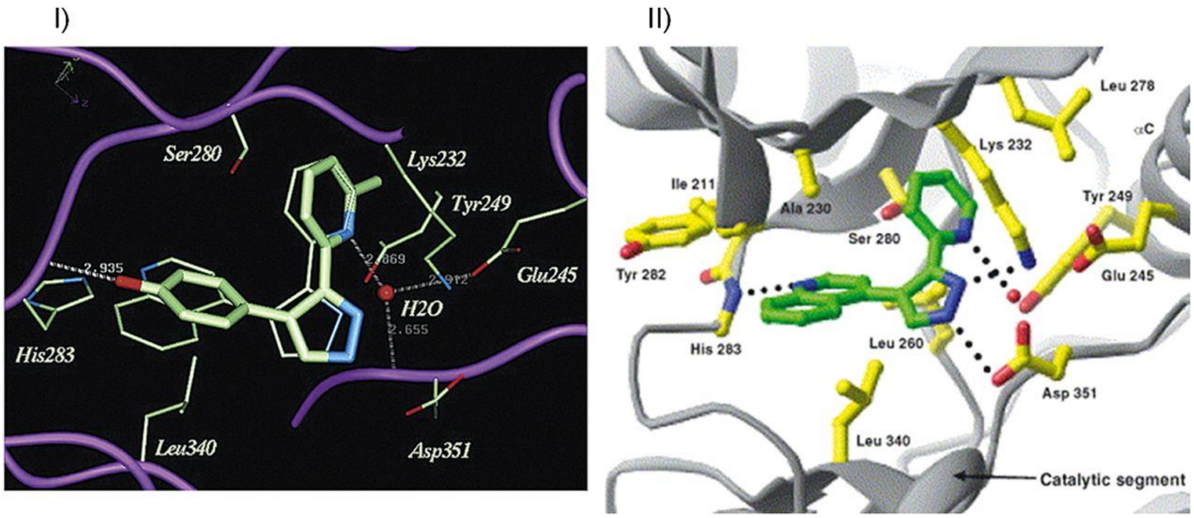


Figure 1.1: Identical lead compounds are discovered in a traditional high-throughput screen and structure-based virtual high-throughput screen. I) X-ray crystal structures of 1 and 18 bound to the ATP-binding site of the TβR-I kinase domain discovered using traditional high-throughput screening. Compound 1, shown as the thinner wire-frame is the original hit from the HTS and is identical to that which was discovered using virtual screening. Compound 18 is a higher affinity compound after lead optimization. II) X-ray crystal structure of compound HTS466284 bound to the TβRI active site. This compound is identical to compound 1 in I but was discovered using structure-based virtual high-throughput screening. Source: [6, 7]

1.2.1 Position of CADD in the drug discovery pipeline

CADD is capable of increasing the hit rate of novel drug compounds because it uses a much more targeted search than traditional HTS and combinatorial chemistry. It not only aims to explain the molecular basis of therapeutic activity, but also to predict possible derivatives that would improve activity. In a drug discovery campaign, CADD is usually used for three major purposes: a) filter large compound libraries into smaller sets of predicted active compounds that can be tested experimentally, b) guide the optimization of lead compounds, whether to increase its affinity or optimize drug metabolism and pharmacokinetics (DMPK) properties including absorption, distribution, metabolism, excretion, and the potential for toxicity (ADMET), c) design novel compounds, either by “growing” starting molecules one functional group at a time or by piecing together fragments into novel chemotypes. Figure 1.2 illustrates the position of CADD in drug discovery pipeline.

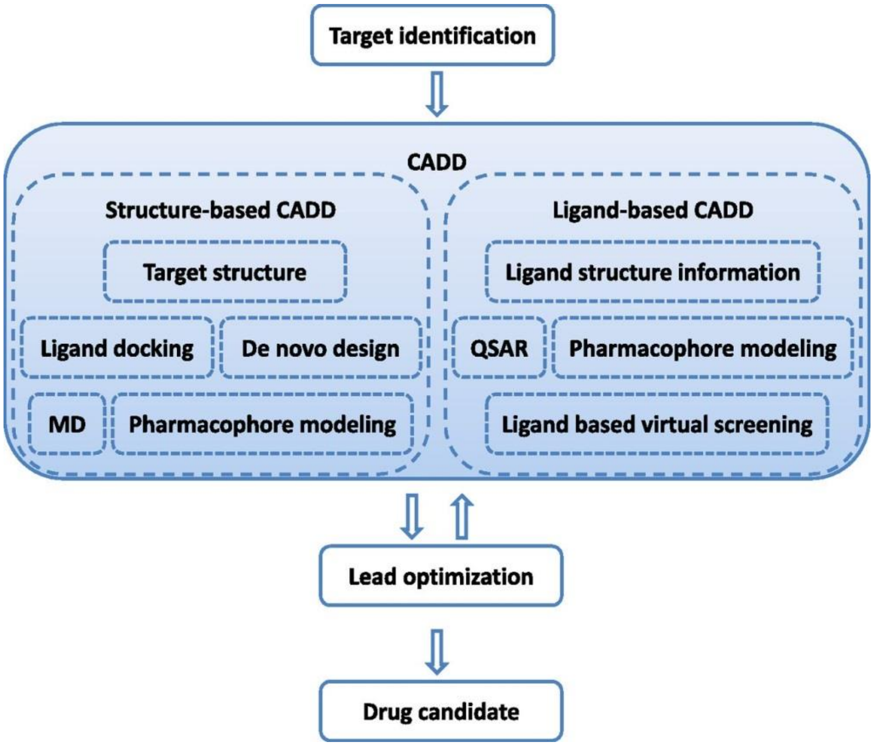


Figure 1.2 CADD in drug discovery/design pipeline. A therapeutic target is identified against which a drug has to be developed. Depending on the availability of structure information, a structure-based approach or a ligand-based approach is used. A successful CADD campaign will allow identification of multiple lead compounds. Lead identification is often followed by several cycles of lead optimization and subsequent lead identification using CADD. Lead compounds are tested in vivo to identify drug candidates.

CADD can be classified into two general categories: structure-based and ligand-based. Structure-based CADD relies on the knowledge of the target protein structure to calculate interaction energies for all compounds tested, while ligand-based CADD exploits the knowledge of known active and inactive molecules through chemical similarity searches or construction of predictive Quantitative Structure-Activity Relation (QSAR) models [9]. Structure-based CADD is generally preferred when high resolution structural data of the target protein is available, i.e. for soluble proteins that can readily be crystallized. Ligand-based CADD is generally preferred when no or little structural information is available, often for membrane protein targets. The central goal of structure-based CADD is to design compounds that bind tightly to the target, i.e. with large reduction in free energy, improved DMPK/ADMET properties, and are target specific, i.e. have reduced off-target effects [10]. A successful application of these methods will result in a compound that has been validated *in vitro* and *in vivo,* and its binding location has been confirmed, ideally through a co-crystal structure.

One of the most common uses in CADD is the screening of virtual compound libraries, also known as virtual high-throughput screening (vHTS). This allows experimentalists to focus resources on testing compounds likely to have any activity of interest. In this way, a researcher can identify an equal number of hits while screening significantly less compounds, because compounds predicted to be inactive may be skipped. Avoiding a large population of inactive compounds saves money and time, because the size of the experimental HTS is significantly reduced without sacrificing a large degree of hits. Ripphausen *et al* note that the first mention of vHTS was in 1997 [11] and chart an increasing rate of publication for the application of vHTS between 1997 and 2010. They also found that the largest fraction of hits has been obtained for G-protein coupled receptors (GPCR's), followed by kinases [12].

Virtual HTS comes in many forms, including chemical similarity searches by fingerprints or topology, selecting compounds by predicted biological activity or pharmacophore mapping, and virtual docking of compounds into a target of interest, known as structure-based docking [13]. These methods allow the ranking of "hits" from the virtual compound library for acquisition. The ranking can reflect a property of interest such as percent similarity to a query compound or predicted biological activity, or in the case of docking, the lowest energy scoring poses for each ligand bound to the target of interest [14]. Often initial hits are rescored and ranked using higher level computational techniques that are too time consuming to be applied to full-scale vHTS. It is important to note that vHTS does not aim to identify a drug-compound that is ready for clinical testing, but rather to find leads with chemotypes that have not previously been associated with a target. This is not unlike a traditional HTS where a compound is generally considered a hit if its activity is close to 10 µM. Through iterative rounds of chemical synthesis and *in vitro* testing, a compound is first developed into a "lead" with higher affinity, some understanding of its structure-activity-relation, and initial tests for DMPK/ADMET properties. Only after further iterative rounds of lead-to-drug optimization and *in vivo* testing does a compound reach clinically appropriate potency and acceptable DMPK/ADMET properties [15]. For example, the literature survey performed by Ripphausen *et al* revealed that a majority of successful vHTS applications identified a small number of hits that are usually active in the micromolar range, and hits with low nanomolar potency are only rarely identified [12].

The cost benefit of using computational tools in the lead optimization phase of drug development is substantial. Development of new drugs can cost anywhere in the range of 400 million to 2 billion dollars with synthesis and testing of lead analogues being a large contributor to that sum [16]. Therefore, it is beneficial to apply computational tools in hit-to-lead optimization in order to cover a wider chemical space while reducing the number of compounds that must be synthesized and tested *in vitro*. The computational optimization of a hit compound can involve a structure-based analysis of docking poses and energy profiles for hit analogs, ligand-based screening for compounds with similar chemical structure or improved predicted biological activity, or prediction of favorable DMPK/ADMET properties. The comparably low-cost of CADD compared to chemical synthesis and biological characterization of compounds make these methods attractive to focus, reduce, and diversify the chemical space that is explored [13].

De novo drug design is another tool in CADD methods, but rather than screening libraries of previously synthesized compounds it involves the design of novel compounds. A structure generator is needed to sample the space of chemicals. Given the size of the search space (more than 10^{60} molecules) [17] heuristics are used to focus these algorithms on molecules that are predicted to be highly active, readily synthesizable, devoid of undesirable properties, often derived from a starting scaffold with demonstrated activity, etc. Additionally, effective sampling strategies are utilized while dealing with large search spaces such as evolutionary algorithms, metropolis search, or simulated annealing [18]. The construction algorithms are generally defined as either linking or growing techniques. Linking algorithms involve docking of small fragments or functional groups such as rings, acetyl groups, esters, etc., to particular binding sites followed by linking fragments from adjacent sites. Growing algorithms, on the other hand, begin from a single fragment placed in the binding site to which fragments are added, removed, and changed to improve activity. Similar to vHTS, the role of *de novo* drug design is not to design the single compound with nanomolar activity and acceptable DMPK/ADMET properties, but to design a lead compound that can be subsequently improved.

1.2.2 Ligand databases for CADD

Virtual HTS uses high-performance computing to screen large chemical databases and prioritize compounds for synthesis. Current hardware and algorithms allow structure-based screening of up to 100,000 molecules per day using parallel processing clusters [19]. To perform a virtual screen, however, a virtual library must be available for screening. Virtual libraries can be acquired in a variety of sizes and designs including general libraries that can be used to screen against any target, focused libraries that are designed for a family of related targets, and targeted libraries that are specifically designed for a single target.

General libraries can be constructed using a variety of computational and combinatorial tools. Early systems used molecular formula as the only constraint for structure generation, resulting in all possible structures for a predetermined limit in the number of atoms. As comprehensive computational enumeration of all chemical space is and will remain infeasible, additional restrictions are applied. Typically, chemical entities difficult to synthesize or known/expected to cause unfavorable DMPK/ADMET properties are excluded. Fink *et al.* proposed a generation method for the construction of virtual libraries that involved the use of connected graphs populated with C, N, O, and F atoms and pruned based on molecular structure constraints and removal of unstable structures. The final database proposed with this method is called the GDB (Generated a DataBase) and contains 26.4 million chemical structures that have been used for vHTS [20, 21]. A more recent variation of this database called GDB-13 includes atoms C, N, O, S, and Cl (F is not included in this variation to accelerate computation) and contains 970 million compounds [22].

Most frequently, vHTS focuses on drug-like molecules that have been synthesized or can be easily derived from already available starting material. For this purpose several small molecule databases are available that provide a variety of information including known/available chemical compounds, drugs, carbohydrates, enzymes, reactants, and natural products [23, 24]. Some widely used databases are listed in table 1-1.

Table 1-1 Widely used chemical compound repositories along with content information about class of compounds they host and size of repositories

DataBase	Type	Size
PubChem [25, 26]	Biological activities of small molecules	~68,000,000
Accelrys Available Chemicals Directory (ACD) [27]	Consolidated catalog from major chemical suppliers	~7,000,000
PDBeChem [28]	Ligands and small molecules referred in PDB	19,838
Zinc [29]	Annotated commercially available compounds	~90,000,000
DrugBank [30]	Detailed drug data with comprehensive drug target information	7469
ChemDB [31, 32]	Annotated commercially available molecules	~5,000,000
WOMBAT Database (World Of Molecular BioAcTivity) [33, 34]	Bioactivity data for compounds reported in medicinal chemistry journals	331,872
MDDR (MDL Drug Data Report) [34]	Drugs under development or released; descriptions of therapeutic	180,000
3D MIND [35]	Molecules with target interaction and tumor cell line screen data	100,000

1.2.3 Preparation of Ligand Libraries for CADD

Ligand libraries are often constructed by enriching ligands for drug likeness or certain desirable physiochemical properties suitable for the target of interest. Even with rapid docking algorithms, docking millions of compounds requires considerable resources, and time can be saved through the elimination of non-drug like, unstable, or unfavorable compounds. Drug likeness is commonly evaluated using Lipinski's rule of five [36] which states that in general, an orally active drug should have no more than one violation of the following criteria a) maximum of five hydrogen bond donors b) no more than 10 oxygen and nitrogen atoms c) molecular mass less than 500 daltons d) an octanol-water partition coefficient of not greater than five. If two or more of the conditions are violated, poor adsorption can be expected. Similarly, polar molecular surface is also used to predict oral absorption and brain penetration [37]. It is a common practice to filter molecules based on predicted DMPK/ADMET properties before initializing a vHTS campaign. Ligand-based methods to predict DMPK/ADMET properties use statistical and learning approaches, molecular descriptors, and experimental data to model biological processes like oral bioavailability, intestinal absorption/permeability, half-life time, and distribution in human blood plasma etc.

Compound libraries are often enriched for a particular target or family of targets. Physiochemical filters derived from observed ligand-target complexes are used for enriching a library with compounds that satisfy specific geometric or physicochemical constraints. Such libraries are prepared by searching for ligands that are similar to known active ligands [38, 39]. Several target-specific libraries exist in Cambridge Structure Database (CSD) including kinase-biased, GPCR-biased, and ion channel-biased sets. In addition, a small molecule library requires preparations such as conformational sampling, and assigning proper stereo isometric and protonation state [40, 41]. Molecules are flexible in solvent environment and hence representation of conformational flexibility is an important aspect of molecular recognition. Often conformations of protein and ligand are precomputed using simulation or knowledge-based methods [42, 43].

1.2.4 Representation of small molecules as “SMILES”

Development and efficient use of ligand databases require universally applicable methods for the virtual representation of small molecules. SMILES (Simplified Molecular Input Line System) [44] was developed as an unambiguous and reproducible method for computationally representing molecules. It was developed as an improvement over the Wiswesser Line Notation [45] which had a cumbersome set of rules, but was a preferred method due to the representation of molecular structure as a linear string of symbols that could be efficiently read and stored by computer systems.

Commonly, SMILES does not explicitly encode hydrogen atoms (hydrogen-suppressed graph) and conventionally assumes that hydrogens make up the remainder of an atom’s lowest normal valence. All non-hydrogen atoms are represented by their atomic symbols enclosed in square brackets. Atoms may also be listed without square brackets, implying the presence of hydrogens. Formal charges are specifically assigned as + or – followed by an optional digit inside the appropriate brackets. Aromatic atoms are specified using the lowercase atomic symbols. Single bonds, double bonds, triple bonds, and aromatic bonds are denoted by “-“, “=”, “#”, and “:”, respectively. Branched systems are specified by enclosing them in parentheses. Cyclic structures are represented by breaking a ring at a single or aromatic bond and numbering the atoms on either side of the break with a number. For example, cyclohexane is represented with the SMILES string C1CCCCC1. Disconnected compounds are separated by a period, and ionic bonds are considered disconnected structures with complimentary formal charges [46].

SMILES algorithms are capable of detecting most aromatic compounds with an extended version of Huckel’s rule (all atoms in the ring must be sp^2 hybridized and the number of available π electrons must satisfy 4N + 2) [47]. Therefore, aromaticity does not necessarily need to be defined beforehand. However, tautomeric structures must be explicitly specified as separate SMILES strings. There are no SMILES definitions for tautomeric bonds or mobile hydrogens. SMILES was designed to have good human readability as a molecular file format. However, there are usually many different but equally valid SMILES descriptions for the same structure. It is most commonly used for storage and retrieval of compounds across multiple computer platforms.

SMARTS (SMILES ARbitrary Target Specification) is an extension of SMILES that allows for variability within the represented molecular structures. This provides substructure search functionality to SMILES. In addition to the SMILES naming conventions, SMARTS includes logical operators, such as “AND” (&), “OR” (,), and “NOT” (!) and special atomic and bond symbols that provide a level of flexibility to chemical names. For example, in SMARTS notation, [C,N] represents an atom that can be either an aliphatic carbon or an aliphatic nitrogen and the symbol “~” will match any bond type [48].

1.2.5 Small Molecule Representations for Modern Search Engines: InChIKey

InChI (International Chemical Identifier) was released in 2005 as an open source structure representation algorithm that is meant to unify searches across multiple chemical databases using modern internet search engines. It is maintained by the InChI Trust and currently supports chemical elements up to 112 [49]. The purpose of InChI and the hash-key version InChIKey is to provide a nonproprietary machine-readable code unique for all chemical structures that can be indexed by major search engines such as Google without any alteration. By use of this protocol, researchers can search for chemicals in a routine and straightforward manner. Prior to INChI, chemical searches spanning multiple databases using typical search engines were unreliable. Different systems have their own proprietary identification method for indexing chemicals; SMILES-based searches are insufficient as different databases have adopted their own unique SMILES.

InChI is made up of several layers that represent different classes of structural information. The first two layers contain only general information, including the chemical formula and connections. More specific conformational information such as stereochemistry, tautomerism, and isotopic information is represented in additional optional layers. Bonds between atoms can be partitioned into up to three sublayers depending on the level of specification desired. These layers represent all bonds to nonbridging hydrogen atoms, immobile hydrogen atoms, and mobile hydrogen atoms, respectively. The InChI algorithm includes six normalization rules that apply qualities such as variable protonation and identification of tautomeric patterns and resonances to achieve a unique and consistent chemical representation [49].

InChIKey is a hash-key version of InChI that generates two blocks using a truncated SHA-256 cryptographic hash function. This allows the keys to contain a fixed length of 27 characters with high collision resistance (minimal chance of two different molecules having the same hash key). Use of InChIKeys to search multiple database with typical search engines was tested and the incidence of false-positive hits was low [50]. Publically available web applets are available that allow chemists to draw molecules and automatically search the web using an automatically calculated InChIKey.

1.2.6 Target databases for CADD

The knowledge of the structure of the target protein is required for structure-based CADD. The Protein Data Bank (PDB) [51], established in 1971 at the Brookhaven National Laboratory, and the Cambridge Crystallographic Data Center, are among the most commonly used databases for protein structure. PDB currently houses more than 100,000 protein structures, the majority of which (90%) have been determined using X-ray crystallography and a smaller set determined using NMR spectroscopy. When an experimentally determined structure of a protein is not available, it is often possible to create a comparative model based on the experimental structure of a related protein. When this protein is evolutionarily related, the term “homology modeling” is used in place of comparative modeling. The Swiss-Model server is one of the most widely used web-based tools for homology modeling [52]. Initially, static protein structures were used for all structure-based design methods. However, proteins are not static structures but exist as ensembles of different conformational states. The protein fluctuates through this ensemble depending on the relative free energies of each of these states, spending more time in conformations of lower free energy. Ligands are thought to interact with some conformations but not others, thus stabilizing conformational populations in the ensemble. Therefore, docking compounds into a static protein structure can be misleading, as the chosen conformation may not be representative of the conformation capable of binding the ligand. Recently, it has become state of the art to use additional computational tools such as molecular dynamics and molecular mechanics to simulate and evaluate a protein's conformational space. Conformational sampling provides a collection of snapshots that can be used in place of a single structure that reflect the breadth of fluctuations the ligand may encounter *in vivo*. This approach was proven to be invaluable in CADD by Schames *et al* [53] in the 2004 identification of novel HIV Integrase inhibitors [54]. Some methods, such as ROSETTA-LIGAND [55], are capable of incorporating protein flexibility during the actual docking procedure, alleviating the need for snapshot ensembles.

The collection of events that occurs when a ligand binds a receptor extends far beyond the noncovalent interactions between ligand and protein. Desolvation of ligand and binding pocket, shifts in the ligand and protein conformational ensembles, and reordering of water molecules in the binding site all contribute to binding free energies. Consideration of water molecules as an integral part of binding sites is necessary for key mechanistic steps and binding [56, 57]. These water molecules shift the free energy change of ligand binding by either facilitating certain noncovalent interactions between the ligand and protein, or by being displaced into a more favorable direct interactions between the ligand and protein, causing an overall change in free energy upon binding [58, 59]. Improvements in computational resources allows inclusion of better representations of physiochemical interactions in computational methods to increase their accuracies [60].

1.2.7 Benchmarking Techniques of CADD

Effective benchmarks are essential for assessment of performance and accuracy of CADD algorithms. Design of the benchmark in terms of number and type of target proteins, size and composition of active and inactive chemicals, and selection of quality measures play a key role when comparing new CADD methods with existing ones. Scientific benchmarks usually involve screening a library of compounds that include a subset of known actives combined with known inactive compounds and then evaluating the number of known actives that were identified by the CADD technique used [61].

Performance is commonly reported by correlating predicted activities with experimentally observed activities through the use of receiver operating characteristic (ROC) curves. These curves plot the number of true positive predictions on the *y*-axis versus the false positive predictions on the *x*-axis. A random predictor would result in a plot of a line with a slope of 1, whereas curves with high initial slopes above this line represent increasing performance scores for the method tested [34, 62]. ROC curves are therefore analyzed by determining the area under the curve (AUC), positive predictive value (PPV) – the ratio of true positives in a subset selected in a vHTS screen, or enrichment – a benchmark that normalizes PPV by the background ratio of positives in the dataset.

For structure-based CADD it is now common also to include decoy molecules that further test a technique's ability to discern actives from inactives at high resolution. Irwin *et al* created the Directory of Useful Decoys (DUD) dataset designed for high-resolution benchmarking. It includes experimental data for approximately 3000 ligands covering up to 40 different targets and a set of carefully chosen decoys [63]. These decoys were designed to resemble positive ligands physically but not topologically [64]. These decoys, however, are not experimentally validated and are only postulated to be "inactive" against the targets. Good and Oprea developed clustered versions of DUD with added data sets from sources such as WOMBAT to avoid challenges in enrichment comparisons between methods due to different parameters and limited diversity [65].

1.3 Structure-Based Computer-Aided Drug Discovery (SB-CADD)

Structure-based computer-aided drug discovery (SB-CADD) relies on the ability to determine and analyze 3D structures of biological molecules. The core hypothesis of this approach is that a molecule's ability to interact with a specific protein and exert a desired biological effect depends on its ability to favorably interact with a particular binding site on that protein. Molecules that share those favorable interactions will exert similar biological effects. Therefore, novel compounds can be elucidated through the careful analysis of a protein's binding site. Structural information about the target is a prerequisite for any SB-CADD project. Scientists have been using a target protein's structure to aid in drug discovery since the early 1980s [66]. Since then, SB-CADD has become a commonly used drug discovery technique thanks to advances in genomics and proteomics that have led to the discovery of a large number of candidate drug targets [67, 68]. Extensive use of biophysical techniques such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy has led to the elucidation of a number of 3D structures of human and pathogenic proteins. Drug discovery campaigns leveraging target structure information have sped up the discovery process and led to the development of several clinical drugs. A prerequisite for the drug discovery process is the ability to rapidly determine potential binders to the target of interest. Computational methods in drug discovery allow rapid screening of a large compound library and determination of potential binders through modeling/simulation and visualization techniques.

1.3.1 Preparation of a Target Structure

A target structure experimentally determined through x-ray crystallography or NMR techniques and deposited in the PDB is the ideal starting point for docking. Structural genomics has accelerated the rate at which target structures are being determined. In the absence of experimentally determined structures, several successful virtual screening campaigns have been reported based on comparative models of target proteins [69-71]. Efforts have also been made to incorporate information about binding properties of known ligands back into comparative modeling process [72, 73].

Success of virtual screening is dependent upon the amount and quality of structural information known about both the target and the small molecules being docked. The first step is to evaluate the target for the presence of an appropriate binding pocket [74, 75]. This is usually done through the analysis of known target-ligand cocrystal structures or using *in silico* methods to identify novel binding sites [76].

Comparative modeling

Advances in biophysical techniques such as X-ray crystallography and NMR spectroscopy have increased the availability of protein structures. This provides structural information to guide drug discovery. In the absence of experimental structures, computational methods are used to predict the 3D structure of target proteins. Comparative modeling is used to predict target structure based on a template with a similar sequence, taking advantage of the fact that protein structure is better conserved than sequence, i.e. proteins with similar sequences have similar structures. Homology modeling is a specific type of comparative modeling in which the template and target proteins share the same evolutionary origin. Comparative modeling involves the following steps: a) Identification of related proteins to serve as template structures, b) sequence alignment of the target and template proteins, c) copying coordinates for confidently aligned regions, d) constructing missing atom coordinates of target structure, and e) model refinement and evaluation. Figure 1.3 illustrates the steps involved in comparative modeling. Several computer programs and web servers exist which automate the comparative modeling process e.g. PSIPRED [77] , MODELLER [78].

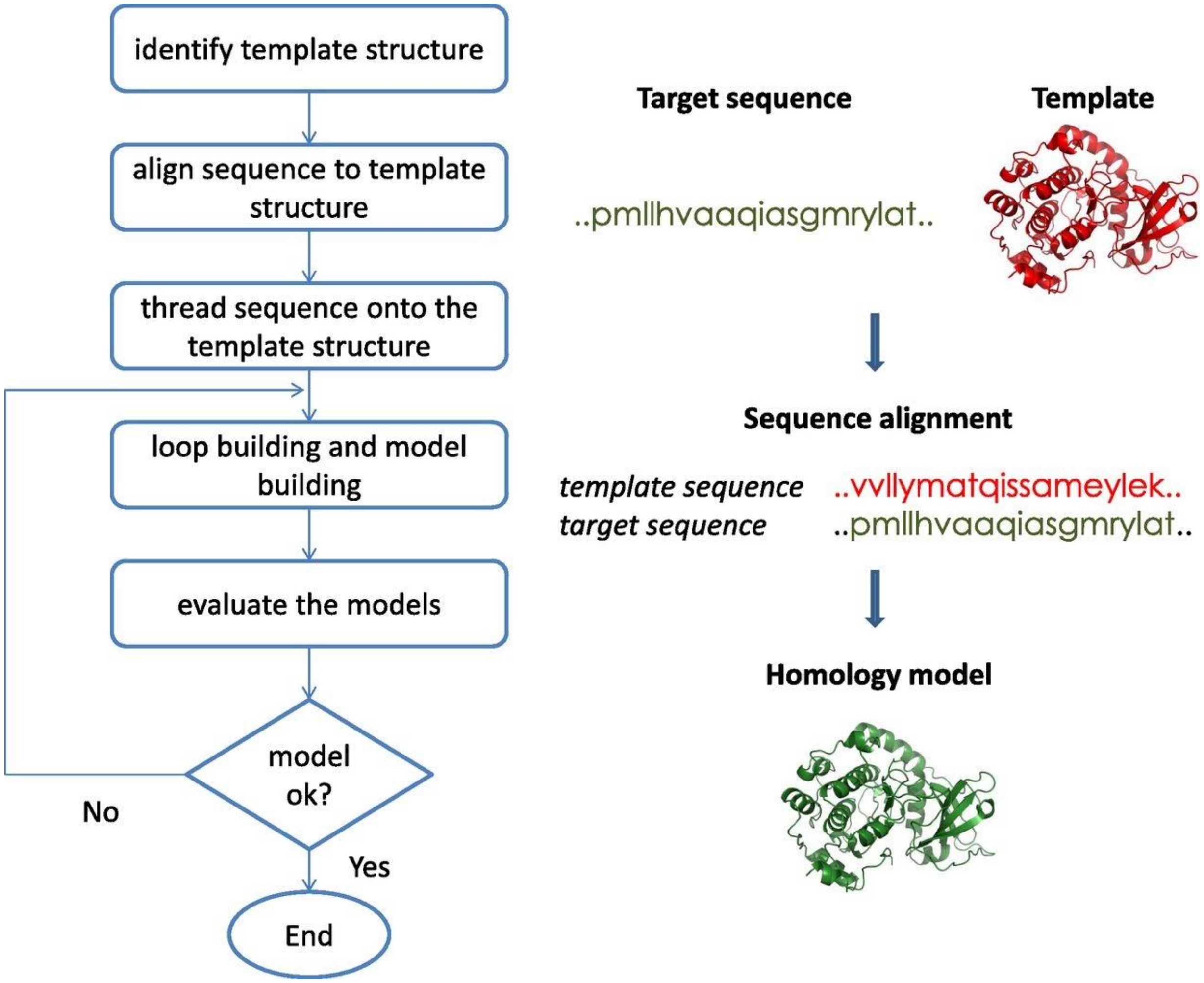

identify template structure
align sequence to template structure
thread sequence onto the template structure
loop building and model building
evaluate the models
model ok?
No
Yes
End
Target sequence
Template
..pmllhvaaqiasgmrylat..
Sequence alignment
template sequence ..vvllymatqissameylek..
target sequence ..pmllhvaaqiasgmrylat..
Homology model

Figure 1.3 Steps in homology model building process

Template identification and alignment

In the first step the target sequence is used as a query for the identification of template structures in the PDB. Templates with high sequence similarity can be determined by a straight-forward PDB-BLAST search [79]. More sophisticated fold recognition methods are available if PDB-BLAST does not yield any hits [80, 81]. Search for template structure is followed by sequence alignment using methods like CLUSTALW [82] which is a multiple sequence alignment tool. For closely related protein structures, structurally conserved regions are identified and used to build the comparative model. Construction and evaluation of multiple comparative models from multiple good-scoring sequence alignments improves the quality of the comparative model [83, 84]. It has been demonstrated that combination of multiple templates can improve comparative models by leveraging well-determined regions that are mutually exclusive [85]. Appropriate template selection is critical for successful comparative modeling. Careful consideration should be given to alignment length, sequence identity, resolution of template structure and consistency of secondary structure between target and templates.

Model building

Gaps or insertions in the original sequence alignment occur most frequently outside secondary structure elements and lead to chain breaks (gaps and insertions) and missing residues (gaps) in the initial target protein model. Modeling these missing regions involves connecting the anchor residues, which are the N- or C-terminal residues of protein segments on either side of the missing region. Two broad classes of loop-modeling methods exist: knowledge-based methods and *de novo* methods. Knowledge-based methods use loops from protein structures that have approximately the same anchors as found in target models. Loops from such structures are applied to the target structure. *De novo* methods generate a large number of loop conformations and use energy functions to judge the quality of predicted loops [86]. Both methods, however, solve the “loop closure” problem, i.e. identifying low-energy loop conformations from a large conformational sample space that justify the structural restraint of connecting the two anchor points. Cyclic coordinate descent (CCD) [87] and kinematic closure (KIC) [88] algorithms optimally search for conformations that satisfy constraints for loop closure in a target structure. CCD iteratively changes dihedral angles one at a time such that a distance constraint between anchor residues is satisfied [87]. The KIC algorithm derives from kinematic methods which allow geometric analysis of possible conformations of a system of rigid objects connected by flexible joints. The KIC algorithm generates a Fourier polynomial in N variables for a system of N rotatable bonds by analyzing bond lengths and bond angles constraints [89]. Atom coordinates of the loop are then determined using the polynomial equation.

Loop modeling can be affected by two classes of errors: scoring function errors and insufficient sampling. The former arises when nonnative conformations are assigned better scores. Confidence in scoring can be improved by scoring with different functions, assuming that true native conformation will likely be best ranked across multiple scoring methods. Insufficient sampling arises when near native conformations are not sampled. Sufficient sampling can be achieved by running multiple independent simulations to establish convergence.

The next step in comparative modeling is prediction of side-chain conformations. A statistical clustering of observed side-chain conformations in PDB, called a rotamer library is used in most side-chain construction methods [90]. Methods like dead-end elimination [91] implemented in SCRWL [92-94] and Monte Carlo searches [95] are used for side-chain conformation sampling. Dead-end elimination imposes conditions to identify rotamers that cannot be members of global minimum energy conformation. For example, the algorithm prunes a rotamer *a* if a second rotamer *b* exists, such that lowest energy conformation containing *a* is greater than highest energy conformations containing *b.* The SCRWL algorithm evaluates steric interactions between side chains through the use of a backbone dependent rotamer library which expresses frequency of rotamers as a function of dihedral angles ϕ and ψ. Monte Carlo algorithms search the side chain conformational space stochastically using the Metropolis criterion to guide the search into energetic minima.

Binding pockets in homology models or even crystal structures are often not amenable for ligand docking because of insufficient accuracy. Ligand information has been used to improve comparative models. Tanrikulu et al used a pseudoreceptor modeling method to improve a homology model of human histamine H_4 receptor. Pseudoreceptor methods map binding pockets around one or more reference ligands by capturing their shape and interactions with the target. Conformation snapshots of the homology model were obtained by MD simulation, and pocket-forming coordinates were extracted. Binding pockets of MD frames that matched pseudoreceptor were prioritized for virtual screening. Hits from virtual screening were tested experimentally and two compounds with diverse chemotypes exhibited $pK_i > 4$ [96, 97]. *Katritch et al*. used a combined homology modeling and ligand-guided backbone ensemble receptor optimization algorithm (LiBERO) for prediction of a protein-ligand complex in CASP experiments. The approach was identified as the best in that it identified 40% of the 70 contacts that the antagonist ZM241385 makes with adenosine A2a receptor (PDB:3EML). In LiBERO framework multiple models are generated and normal mode analysis is used to generate backbone conformation ensembles. Conformers are selected according to docking performance through an iterative process of model building and docking [98]. Ligand information assisted homology modeling is contingent on the availability of high-affinity ligands and structurally similar homologs to ensure high quality homology models.

Model refinement and evaluation

Atomic models are refined by introducing ideal bond geometries and by removing unfavorable contacts introduced by the initial modeling process. Refinement involves minimizing models using techniques such as molecular dynamics [99], Monte Carlo Metropolis minimization [100] or genetic algorithms [101]. For example, the ROSETTA refinement protocol fixes bond lengths and angles at ideal values and removes steric clashes in an initial low-resolution step. ROSETTA then minimizes energy as a function of backbone torsional angles ϕ, ψ, and ω using a Monte Carlo minimization strategy [100]. Molecular dynamics-based refinement techniques have been used widely as refinement strategy in drug-design oriented homology models [102, 103].

Model evaluation involves comparison of observed structural features with experimentally determined protein structures. Melo and Sali [104] applied a genetic algorithm that used 21 input model features like sequence alignment scores, measures of protein packing, and geometric descriptors to assess folds of models. Critical Assessment of Techniques for Protein Structure Prediction (CASP) [105] is a worldwide competition in which many groups participate for an objective assessment of methods in the area of protein structure prediction. Models are numerically assessed and ranked by estimating similarity between a model and corresponding experimental structure. Some evaluation methods used in CASP are full model root mean square deviation (RMSD), global distance test-total scores (GDT-TS) and alignment accuracy (AL0 score). GDT-TS is the average maximum number of residues in predicted model that deviate from corresponding residues in the target by no more than a specified distance while AL0 represents the percentage of correctly aligned residues [105].

Model databases

SWISS-MODEL [106] and MODBASE [107] databases store annotated comparative protein structure models. SWISS-MODEL repository contains annotated 3D protein models generated by homology modeling of all sequences in SWISS-PROT [106]. As of March 2015, SWISS-MODEL contained 3.18 million entries for 2.3 million unique sequences in UNIPROT database. MODBASE is organized into datasets of models for specific projects which include datasets of 9 archaeal genomes, 13 bacterial genomes and 18 eukaryotic genomes. Together with other datasets, MODBASE currently houses 34 million models across 5.7 million unique protein sequences [107].

Park *et al* [108] used a homology model of Cdc25A phosphatase, a drug target for cancer therapy, to identify novel inhibitors. The crystal structure of protein Cdc25B served as a template to generate structural models of Cdc25A. Docking of a library of 85,000 compounds led to the discovery of structurally diverse compounds with IC_{50} values ranging from 0.8 to 15 µM.

1.3.2 Binding site detection and characterization

Protein-ligand interaction is a prerequisite for drug activity. Often possible binding sites for small molecules are known from cocrystal structures of the target or a closely related protein with a natural or nonnatural ligand. In the absence of a cocrystal structure, mutational studies can pinpoint ligand binding sites. However, the ability to identify putative high-affinity binding sites on proteins is important if the binding site is unknown or if new binding sites are to be identified, e.g. for allosteric molecules. Computational methods like POCKET, SURFNET, Q-SITEFINDER, etc. [76, 109] are often used for binding site identification. Computational methods for identifying and characterizing binding sites can be divided into three general classes: a) geometric algorithms to find shape concave invaginations in the target, b) methods based on energetic consideration, and c) methods considering dynamics of protein structures.

Geometric method

Geometric algorithms identify binding sites through the detection of cavities on a protein's surface. These algorithms frequently use grids to describe molecular surface or 3D structure of protein. The boundary of a pocket is determined by rolling a "spherical probe" over the grid surface. A pocket is identified if there is a period of noninteraction i.e. probe doesn't touch any target atoms, between periods of contact with protein. This technique is employed by POCKET [110] and LIGSITE [111]. SURFNET [112] places spheres between all pairs of target atoms and then reduces the radius of spheres until each sphere contains only a pair of atoms. The program thus accumulates spheres in pockets, both inside the target and on the surface. The SPHGEN program [113] generates overlapping spheres to describe the 3D shape of binding pocket. The algorithm creates a negative image of invaginations for target surface. Spheres are calculated all over the entire surface such that each sphere touches the molecular surface at two points. The overlapping dense representation of spheres is then filtered to include only largest sphere associated with each target surface atom. The main disadvantages of geometric-based methods include that geometric descriptors are method dependent and subjective, the target protein is typically rigid, and the methods are often tied to a generalized concept of a binding pocket and may miss unorthodox binding sites within channels or on protein-protein interaction interfaces [76].

Trypanosoma brucei is the causative agent of human trypanosomiasis in Africa [114]. A binding pocket identified by LIGSITE was used for identifying inhibitors of ornithine decarboxylase which is a molecular target for treatment of African trypanosomiasis. SPHGEN was used to identify putative binding sites in BCL6 [115], a therapeutic target for B cell lymphomas. Docking of a library of 1,000,000 commercially available compounds into the identified sites led to successful identification of inhibitors of BCL6 [115].

Energy-based approaches

Energy-based approaches calculate van der Waals, electrostatic, hydrogen-binding, hydrophobic, and solvent interactions of probes that could result in energetically favored binding. Simple energy-based methods tend to be as fast as geometric methods, but are more sensitive and specific. The Q-SITEFINDER [116] algorithm calculates the VDW interaction energy for aliphatic carbon probes on a grid, and retains pockets with favorable interactions. The GRID [117, 118] algorithm samples the potential on a 3D grid to determine favorable binding positions for different probes. GRID determines interaction energy as a sum of Lennard-Jones, Coulombic and hydrogen-bond terms. Other algorithms like POCKETPICKER [119] and FLAPSITE [109] use similar approaches but different metrics to evaluate the quality of a putative binding site. For example, POCKETPICKER defines "buriedness" indices in its binding site elucidation. A serious limitation of these methods is that they result in many different energy minima on the surface of the protein, including many false-positives [76]. These shortcomings can be addressed in part by including the solvation term in the scoring potential as is done in CS-Map algorithm [120]. More complex tools distinguish solvent accessible from solvent inaccessible surfaces. Kim *et al* present a method for defining the topology of the protein as a Voronoi diagram of spheres and its use to elucidate binding pocket locations [121].

Segers *et al* [122] applied Q-SITEFINDER and POCKETFINDER to identify the binding site for the C2 domain of coagulation factor V whose interaction with platelet membrane is necessary for coagulation. Excessive coagulation caused by high thrombin production could be controlled by small molecule inhibitors of factor V. Docking of 300,000 compounds into the predicted sites identified four inhibitors with $IC_{50} < 10$ µM. Novel putative drug binding regions were identified in Avian Influenza Neuraminidase H5N1 using computational solvent mapping [123]. Virtual screening of the binding site with a library of compounds led to the discovery of novel small-molecule inhibitor of H5N1 [124].

Pocket matching

Methods like CATALYTIC SITE ATLAS (CSA) [125], AFT [126], SURFACE [127], POCKET-SURFER [128] and PATCH-SURFER [129] detect similar pockets based on reference ligand binding sites. CSA contains annotated descriptors of enzyme active site residues as well as equivalent sites in related proteins found by sequence alignment. Query made by PDB code returns annotated catalytic residues highlighted on amino acid sequence and on the structure via RasMol [130]. SURFACE is a repository of annotated protein functional sites with sequence and structure-derived information about function or interactions. The comparison algorithm explores all combinations of similar/identical residues in a sequence-independent way between query protein and database structures. Pocket-surfer and patch-surfer describe property of binding pockets. Pocket-surfer captures global similarity of pockets, whereas patch-surfer evaluates and compares binding pockets in small circular patches. These methods describe patches using four properties, surface shape, visibility, hydrophobicity, and electrostatic potential.

Molecular dynamics-based detection

The dynamic nature of biomolecules sometimes makes it insufficient to use a single static structure to predict putative binding sites. Multiple conformations of target are often used to account for structural dynamics of target. Classical molecular dynamics (MD) simulations can be used for obtaining an ensemble of target conformations beginning with a single structure. The MD method uses principles of Newtonian mechanics to calculate a trajectory of conformations of a protein as a function of time. The trajectory is calculated for a specific number of atoms in small time steps, typically 1-10 fs [131]. Classical MD methods tend to get trapped in local energy minima. Several advanced MD algorithms like targeted-MD [132], SWARM-MD [133], conformational flooding simulations [134], temperature accelerated MD simulations [135], and replica exchange MD [136] have been implemented for traversing multiple-minima energy surface of proteins.

MD simulations elucidated a novel binding trench in HIV integrase (IN), which led to development of raltegravir, a drug used to treat HIV infection. MD simulations of 5CITEP, a known inhibitor of IN, showed that the inhibitor underwent various movements including entry into a novel binding trench (shown in figure 1.4) that went undetected with a static crystal structure [53]. The discovery of this trench led to the development of raltegravir, by Merck [137]. Frembgen-Kesner and Elcock [138] reproduced a cryptic drug binding site in an explicit-solvent MD simulation of ligand-free p38 MAP kinase protein, a target in the treatment of inflammatory diseases.

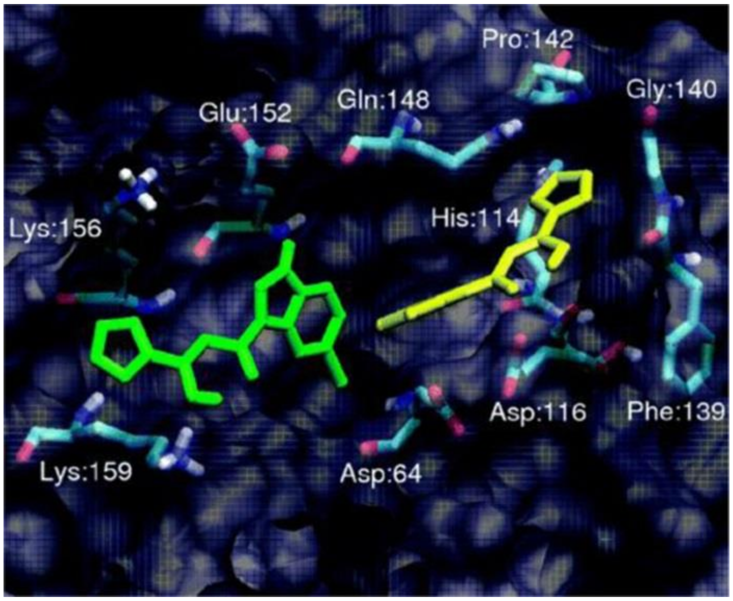


Figure 1.4 Discovery of novel binding trench in HIV-1 IN. Ligand in green is similar to the crystal structure binding pose while the one in yellow is in the novel trench. Source: [53].

1.3.3 Representing Small Molecules and Target Protein for Docking Simulations

There are three basic methods to represent target and ligand structures *in silico*: atomic, surface, and grid representations [139, 140]. Atomic representation of the surface of the target is typically used when scoring and ranking is based on potential energy functions. An example is DARWIN which uses CHARMM force field to calculate energy [141]. Surface methods represent the topography of molecules using geometric features. The surface is represented as a network of smooth convex, concave, and saddle shape surfaces. These features are generated by mapping part of van der Waals surface of atoms that is accessible to probe a sphere [142]. Docking is then guided by a complementary alignment of ligand and binding site surfaces. Earliest implementation of DOCK [143] used a set of nonoverlapping spheres to represent invaginations of target surface and the surface of the ligand (method described earlier in detail for SPHGEN). Geometric matching begins by systematically pairing one ligand sphere a_1 with one receptor sphere b_1. This is followed by pairing a second set of spheres, a_2 and b_2. The move is accepted if the change in atomic distances is less than an empirically determined cut-off value. The cut-off value specifies the maximum allowed deviation between ligand and receptor internal distance. The pairing step is repeated for a third pair of atoms with the same internal distance checks as above. A minimum of four assignable pairs is essential for determining orientation, otherwise the match is rejected. For the grid representation, the target is encoded as physicochemical features of its surface. A grid method described by Katchalskikatzir *et al* [144] digitizes molecules using a 3D discrete function that distinguishes the surface from the interior of the target molecule. Molecules are scanned in relative orientation in three dimensions, and the extent of overlap between molecules is determined using a correlation function calculated from a Fourier Transform. Best overlap is determined from a list of overlap functions [144]. Physiochemical properties may be represented on the grid by storing energy potentials on surface grid points.

1.3.4 Sampling Algorithms for Protein-Ligand Docking

Docking methods can be classified as rigid-body docking and flexible docking applications depending on the degree to which they consider ligand and protein flexibility during the docking process [139, 145]. Rigid body docking methods consider only static geometric/physiochemical complementarities between ligand and target and ignore flexibility and induced-fit [139] binding models. More advanced algorithms consider several possible conformations of ligand or receptor or both at the same time according to the conformational selection paradigm [146]. Rigid docking simulations are generally preferred when time is critical, i.e., when a large number of compounds are to be docked during an initial vHTS. However, flexible docking methods are still needed for refinement and optimization of poses obtained from an initial rigid docking procedure. With the evolution of computational resources and efficiency, flexible docking methods are becoming more commonplace. Some of the most popular approaches include systematic enumeration of conformations, molecular dynamic simulations, Monte Carlo search algorithms with Metropolis criterion (MCM), and genetic algorithms.

Systematic methods

Systematic algorithms incorporate ligand flexibility through a comprehensive exploration of a molecule’s degrees of freedom. In systematic algorithms, the current state of the system determines the next state. Starting from the same exact state and same set of parameters, systematic methods will yield exactly the same final state. Systematic methods can be categorized into exhaustive search algorithms or fragmentation algorithms.

Exhaustive searches elucidate ligand conformations by systematically rotating all possible rotatable bonds at a given interval. Large conformational space often prohibits an exhaustive systematic search. Algorithms such as GLIDE [147] use heuristics to focus on regions of conformational space that are likely to contain good scoring ligand poses. GLIDE precomputes a grid representation of target's shape and properties. Next, an initial set of low-energy ligand conformations in ligand torsion-angle space is created. Initial favorable ligand poses are identified by approximate positioning and scoring methods (shape and geometric complementarities). This initial screen reduces the conformational space over which the high-resolution docking search is applied. High-resolution search involves the minimization of the ligand using standard molecular mechanics energy function followed by a Monte Carlo procedure for examining nearby torsional minima.

Fragmentation methods sample ligand conformation by incremental construction of ligand conformations from fragments obtained by dividing the ligand of interest. Ligand conformations are obtained by docking fragments in the binding site one at a time and incrementally growing them, or by docking all fragments into the binding site and linking them covalently. DesJarlais *et al* modified the DOCK algorithm to allow for ligand flexibility by separately docking fragments into the binding site and subsequently joining them [148]. FLEXX [149] uses the "anchor and grow method" for ligand conformational sampling. A base fragment has to be interactively selected by the user, which is followed by automatic determination of placements for the fragment that maximize favorable interactions with the target protein. The base fragment is grown incrementally by adding new fragments in all possible conformations, and the extended fragment is selected if no significant steric clashes (overlap volume $\leq$ 4.5 Å^3) are observed between ligand and target atoms. Extended ligands are optimized if new interactions are found or if minor steric interactions exist [149]. Fully automated "anchor and grow" methods have been implemented in several methods like FLOG [150], SURFLEX [151] and SEED [152]. In a benchmark study in which performance of eight docking algorithm was compared on 100 protein-ligand complex, GLIDE and SURFLEX were among the methods that showed best accuracy [153]. GLIDE and SURFLEX generated poses close to X-ray conformation for 68 protein-ligand complexes in the Directory of Useful Decoys [154].

Human Pim-1 kinase, responsible for cell survival/apoptosis, differentiation and proliferation, is a valuable anticancer target as it is over expressed in a variety of leukemia. Pierce *et al* [155] used GLIDE to dock approximately 700,000 commercially available compounds and identified four compounds with K_i values less than 5 µM. Chiu *et al* [156] used SURFLEX to identify novel inhibitors of anthrax toxin lethal factor, responsible for anthrax-related cytotoxicity. Docking study of a compound library derived from seven databases including DrugBank [30], ZINC [29], National Cancer Institute (NCI) database [157] identified lead compounds which eventually led to the development of nanomolar inhibitors upon optimization. Table 1-2 illustrates some examples of drug discovery campaigns where systematic docking algorithms have been used.

Table 1-2 Successful docking applications of some widely used docking software.

Algorithm	Target
SEED	Plasmepsin [158], target for malaria Flavivirus Proteases [159, 160], target for WNV and dengue virus Tyrosine Kinase Erythropoietin Producing Human Hepatocellular Carcinoma Receptor B4(EphB4) [161]
FlexX	Plasmepsin II and IV Inhibitors [162], malaria Anthrax edema factor [163] Pneumococcal peptidoglycan deacetylase inhibitors [164]
Glide	Aurora kinases inhibitors [69] Falcipain inhibitors [165] Cytochrome450 inhibitors [166]
Surflex	Topoisomerase I , anti-cancer (optimization)
DOCK	FK506 Immunophilin [167] BCL6, oncogene in B cell lymphomas [115]

Molecular dynamics simulations

Molecular dynamics (MD) simulation calculates the trajectory of a system by the application of Newtonian mechanics. However, standard MD methods depend heavily on the starting conformation and are not readily appropriate for simulation of ligand-target interactions. Because of its nature, MD is not able to cross high-energy barriers within the simulation's lifetime and is not efficient for traversing the rugged hyper surface of protein-ligand interactions. Strategies like simulated annealing have been applied for more efficient use of MD in docking. Mangoni *et al* described a MD protocol for docking small flexible ligands to flexible targets in water [168]. They separated the center of mass movement of ligand from its internal and rotational motions. The center of mass motion and internal motions were coupled to different temperature baths, allowing independent control to the different motions. Appropriate values of temperature and coupling constants allowed for flexible or rigid ligand and/or receptor.

The McCammon group developed a "relaxed-complex" approach that explores binding conformations that may occur only rarely in the unbound target protein. A 2-ns MD simulation of ligand free target is carried out to extensively sample its conformations. Docking of ligands is then performed in target conformation snapshots taken at different time points of the MD run. This relaxed complex method was used to discover novel modes of inhibition for HIV integrase and led to the discovery of the first clinically approved HIV integrase inhibitor, Raltegravir. This MD method was also used in several other campaigns to identify inhibitors of target of interest [169, 170].

Metadynamics is a MD-based technique for predicting and scoring ligand binding. The method maps the entire free energy landscape in an accelerated way as it keeps track of history of already sampled regions. During the MD simulation of a protein-ligand complex, a Gaussian repulsive potential are added on explored regions, steering the simulation toward new free energy regions [54, 171, 172].

Millisecond timescale MD simulations are now possible with special purpose machines like Anton [173]. Such long simulations have allowed study of drug binding events to their protein target [174]. Anton has been used successfully for full atomic resolution protein folding [175]. Advances in computer hardware capabilities means protein flexibility can be accessed more routinely on longer timescales. This would allow better descriptions of conformational flexibility in future.

Monte Carlo search with metropolis criterion

Stochastic algorithms make random changes to either ligand being docked or to its target binding site. These random changes could be translational or rotational in the case of ligand or random conformational sampling of residue side-chains in the target binding site. Whether a step is accepted or rejected in such a stochastic search is decided based on the Metropolis criterion, which generally accepts steps that lower the overall energy and occasionally accepts steps that increase energy to enable departure from a local energy minimum. The probability of acceptance of an uphill step decreases with increasing energy gap and depends on the “temperature” of the MCM simulation [176]. MCM simulations have been adopted for flexible docking applications such as in MCDOCK [177], Internal Coordinate Mechanics (ICM) [178], and ROSETTALIGAND [55, 179]. MCM samples conformational space faster than molecular dynamics because it only requires energy function evaluation and not the derivative of the energy functions. Although traditional MD drives a system towards a local energy minimum, the randomness introduced with Monte Carlo allows hopping over the energy barriers, preventing the system from getting stuck in local energy minima. A disadvantage is that any information about the timescale of the motions is lost.

ROSETTALIGAND [180, 181] uses a knowledge-based scoring procedure with a Monte Carlo-based energy minimization scheme that reduces the number of conformations that must be sampled while providing a more rapid scoring system than offered through molecular mechanics force fields. ROSETTALIGAND incorporates side-chain and ligand flexibility during a high-resolution refinement step through a Monte Carlo-based sampling of torsional angles. All torsion angles of protein and ligand are optimized through gradient-based minimization mimicking an induced fit scenario [179]. MCDOCK uses two stages of docking and a final energy minimization step for generating target-ligand structure. In the first docking stage, the ligand and docking site are held rigid while the ligand is placed randomly into the binding site. Scoring is done entirely on the basis of short contacts. This allows identification of nonclashing binding poses. In the next stage, energy-based Metropolis sampling is done to sample the binding pocket [177]. QXP [182] optimizes grid map energy and internal ligand energy for searching ligand-target structure. The algorithm performs a rigid body alignment of ligand-target complex followed by MCM translation and rotation of ligand. This step is followed by another rigid body alignment and scoring using energy grid map. ICM [183] relies on a stochastic algorithm for global optimization of entire flexible ligand in receptor potential grid. The relative positions of ligand and target molecule make up the internal variables of the method. Internal variables are subject to random change followed by local energy minimization and selection by Metropolis criterion. ICM performed satisfactorily in generating protein-ligand complexes for 68 diverse, high-resolution X-ray complexes found in DUD [154].

ROSETTALIGAND was used by Kaufmann *et al* [184] to predict the binding mode of serotonin with serotonin transporters. The binding site predicted to be deep within the binding pocket was consistent with mutagenesis studies. QXP has been used to optimize inhibitors of Human β-Secretase (BACE1) [185-187], an important therapeutic target for treating Alzheimer's disease by diminishing β-amyloid deposit formation. ICM was used successfully to identify inhibitors for a number of targets, including Tumor necrosis factor-α [188], dysregulation of which is implicated in tumorigenesis and autoinflammatory diseases like rheumatoid arthritis and psoriatic arthritis. Computational screening of 230,000 compounds from the NCI database against neuraminidase using ICM identified 4-[4-[(3-(2-amino-4-hydroxy-6-methyl-5-pyrimidinyl)propyl)amino]phenyl]-1-chloro-3-buten-2-one which inhibited influenza virus replication at a level comparable to known neuraminidase inhibitor oseltamivir [124].

Genetic Algorithms

Genetic algorithms introduce molecular flexibility through recombination of parent conformations to child conformations. In this simulated evolutionary process, the “fittest” or best scoring conformations are kept for another round of recombination. In this way, the best possible set of solutions evolves by retaining favorable features from one generation to the next. In docking, a set of values that describe the ligand pose in the protein are state variable, i.e., the genotype. State variables may include sets of values describing translation, orientation, conformation, number of hydrogen bonds, etc. The state corresponds to the genotype; the resulting structural model of the ligand in the protein corresponds to the phenotype, and binding energy corresponds to the fitness of the individual. Genetic operators may swap large regions of parent’s genes or randomly change (mutate) the value of certain ligand states to give rise to new individuals.

Genetic Optimization for Ligand Docking (GOLD) [189] explores full ligand flexibility with partial target flexibility using a genetic algorithm. The GOLD algorithm optimizes rotatable dihedrals and ligand-target hydrogen bonds. The fitness of a generation is evaluated based on a maximization of intermolecular hydrogen bonds. The fitness function is the sum of a hydrogen bonding term, a term for steric energy interaction between the protein and the ligand and a Lennard-Jones potential for internal energy of ligand. AutoDock [190] uses the Lamarckian genetic algorithm, which allows favorable phenotypic characteristics to become inheritable. GOLD has demonstrated better accuracy than most docking algorithms, except GLIDE, in various benchmark studies [153, 191, 192].

Inhibition of α-glucosidase has shown to retard glucose absorption and decrease postprandial blood glucose level, making it an attractive target for treating diabetes and obesity. Park *et al* [193] used AUTODOCK to identify four novel inhibitors of α-glucosidase by screening a library of 85,000 compounds obtained from INTERBIOSCREEN chemical database . AUTODOCK was also used to identify inhibitors of RNA Editing Ligase-1 enzyme of *Trypanosoma brucei*, causative agent of Human African trypanosomniasis [194].

Incorporating target flexibility in docking

Conformational variability is seen in unbound form and different apo structures [195, 196]. It is widely believed that the ligand-bound state is selected from an ensemble of protein conformations by the ligand [197]. Accounting for receptor flexibility in the form of protein side-chain and backbone movement is essential for predicting correct binding pose. An ensemble of nonredundant low energy target structures covers a larger conformational space than a single conformation. Methods for inducing receptor flexibility include induced-fit docking and MD simulation snapshot ensembles. Induced-fit algorithms allow small overlap between the ligand and the target along with side-chain movements, resulting in elasticity. GLIDE uses an induced fit model in which all side-chain residues are changed to alanine before initial docking. Side-chain sampling is followed by energy minimization of the binding site and ligand. ROSETTALIGAND allows for full protein backbone and side-chain flexibility in the active site. Multiple fixed receptor conformations are used in docking protocols, known as ensemble-based screening, to incorporate receptor flexibility [198]. Receptor conformations may either be experimentally determined by crystallography or NMR or computationally generated from MD simulations, normal mode analysis and MC sampling [199]. Schames *et al*. used the relaxed complex scheme (RCS) to describe a novel trench in HIV integrase which led to the discovery of the integrase inhibitor raltegravir [53]. In RCS, multiple conformations are determined from MD simulations to perform docking studies against. Other sampling methods include umbrella-sampling, metadynamics, accelerated MD etc [196].

1.3.5 Scoring Functions for Evaluation Protein-Ligand Complexes

Docking applications need to rapidly and accurately assess protein-ligand complexes, i.e., approximate the energy of the interaction. A ligand docking experiment may generate hundreds of thousands of target-ligand complex conformations, and an efficient scoring function is necessary to rank these complexes and differentiate valid binding mode predictions from invalid predictions. More complex scoring functions attempt to predict target-ligand binding affinities for hit-to-lead and lead-to-drug optimization. Scoring functions can be grouped into four types: a) force-field or molecular mechanics-based scoring functions b) empirical scoring functions c) knowledge-based scoring functions d) consensus scoring functions.

Force-field or molecular mechanics based scoring functions

Force-field scoring functions use classic molecular mechanics for energy calculations. These functions use parameters derived from experimental data and *ab initio* quantum mechanical calculations. The parameters for various force terms including prefactor variables are obtained by fitting to high-quality *ab initio* data on intermolecular interactions [200]. The binding free energy of protein-ligand complexes are estimated by the sum of van der Waals and electrostatic interactions. DOCK uses the AMBER force fields in which van der Waals energy terms are represented by the Lennard-Jones potential function while electrostatic terms are accounted for by coulomb interaction with a distance-dependent dielectric function. Standard force fields are however biased to select highly charged ligands. This can be corrected by handling ligand solvation during calculations [201, 202]. Terms from empirical scoring functions (discussed below) are often added to force-field functions to treat solvation and electronic polarizability. A semi-empirical force field has been implemented in AUTODOCK to evaluate the contribution of water surrounding the receptor-ligand complex in the form of empirical enthalpic and entropic terms, for example [203].

Empirical Scoring Functions

Empirical scoring functions fit parameters to experimental data. An example is binding energy, which is expressed as a weighted sum of explicit hydrogen bond interactions, hydrophobic contact terms, desolvation effects, and entropy. Empirical function terms are simple to evaluate and are based on approximations. The weights for different parameters are obtained from regression analysis using experimental data obtained from molecular data. Empirical functions have been used in several commercially available docking suits like LUDI [204] , FLEXX [149] and SURFLEX.

Knowledge-Based Scoring Function

Knowledge-Based scoring functions employ the information contained in experimentally determined complex structures. They are formulated under the assumption that interatomic distances occurring more often than average distances represent favorable contacts. On the other hand, interactions that are found to occur with lower frequencies are likely to decrease affinity. Several knowledge based potentials have been developed to predict binding affinity like potential of mean force [205], DRUGSCORE [206], SMOG [207] and BLEEP [208].

Consensus-Scoring Functions

More recently, consensus-scoring functions have been demonstrated to achieve improved accuracies through a combination of basic scoring functions. Consensus approaches rescore predicted poses several times using different scoring functions. These results can then be combined in different ways to rank solutions [209]. Some strategies for combining scores include a) weighted combinations of scoring functions b) a voting strategy in which cutoffs established for each scoring method is followed by decision based on number of passes a molecule has c) a rank by number strategy that ranks each compound by its average normalized score values d) a rank by rank method that sorts compounds based on average rank determined by individual scoring functions. O'Boyle *et al* [210] evaluated consensus scoring strategies to investigate the parameters for the success of properly combined rescoring strategies. It turns out that combining scoring functions that have complementary strengths leads to better results over those with consensus in their predictions. For example, scoring functions whose strengths are distinguishing actives from inactive compounds are complemented by scoring functions that can distinguish correct from incorrect binding poses. One disadvantage of consensus scoring is that a single inappropriate scoring function can lead to false negatives.

Okamoto *et al* [211] used consensus scoring to identify inhibitors of death-associated protein kinases that may contribute to ischemic diseases in the brain, kidney, and other organs. They used DOCK4.0 and three scoring functions including an empirical scoring function implemented in FLEXX, a knowledge-based PMF scoring function [212], and the force-field function in DOCK4.0. Approximately 400,000 compounds from a corporate compound library were docked followed by simultaneous scoring with the three functions. The consensus score was defined as the highest among the three. In another successful application of consensus scoring scheme, Friedman and Caflisch [158] discovered plasmepsin inhibitors for use as antimalarial agents using a scoring based on median ranking of four field-based scoring functions.

1.3.6 Structure-Based virtual High-Throughput Screening

Structure-based virtual high-throughput screening (SB-vHTS), is an *in silico* screening method for identifying putative hits out of hundreds of thousands of compounds to targets of known structure that relies on a comparison of the 3D structure of a ligand with the putative binding pocket. SB-vHTS selects for ligands predicted to bind to a particular site as opposed to traditional HTS that evaluates the ligand's general ability bind and modulate a protein's function. To make screening of large compound libraries in finite time feasible, SB-vHTS often uses limited conformational sampling of protein and ligand and a simplified approximation of binding energy that can be rapidly computed. The inaccuracies introduced by these approximations lead to false-positive hits that can be subsequently removed during a refinement stage where all putative hits are rescored with more sophisticated and computationally expensive methods including iterative docking and clustering of ligand poses. The key steps in SB-vHTS are: 1) preparation of the target protein and compound library for docking 2) determining a favorable binding pose for each compound, and 3) ranking the docked structures. SB-vHTS has been used successfully in identifying novel and potent hits in several drug discovery campaigns [70, 71, 167, 213-219]. Notably, SB-vHTS played a pivotal role in discovery of lead compounds for the following studies.

Inhibitors of Hsp90

Hsp90 is a molecular chaperone that modulates the activity of multiple oncogenic processes, making it an important therapeutic target for oncology. Roughley *et al* [213] virtually screened 0.7 million compounds from rCat [220] against Hsp90 to identify potential inhibitors of Hsp90. Crystal structures of Hsp90 bound to previously known inhibitors were used in the docking-based virtual screen. From over 9000 non-redundant hits identified after the virtual screen, a set of 719 chemically diverse compounds were purchased. A total of 13 compounds with $IC_{50} < 100$ µM and seven with $IC_{50} < 10$ µM were identified. Following lead-optimization, compound AUY922 was carried forward and evaluated against multiple myeloma, breast, lung and gastric cancers.

Discovery of M_1 Acetylcholine Receptor Agonists

Selective agonism of M_1 mAChR, a class A G-protein coupled receptor (GPCR), has therapeutic potential for treating dementia including Alzheimer's disease and cognitive impairment associated with schizophrenia. Budzik *et al* [70] used a homology model of M_1 based on the crystal structure of bovine rhodopsin to virtually screen a corporate compound collection. The docking of compounds into a previously known allosteric binding site yielded approximately 1000 putative hits. *In vitro* testing and optimization for potency and selectivity led to the development of a series of novel 1-(N substituted piperidin-4-yl) benzimidazolones, which resulted in compounds that were potent, central nervous system penetrant, and orally active M1 mAChR agonists.

1.3.7 Atomic-detail / High Resolution Docking

As mentioned, scoring function and sampling algorithms are kept simple to evaluate large libraries of compounds in realistic time frames. The most promising hit compounds often are evaluated with more sophisticated scoring functions, for example, using an electrostatic solvation model for evaluating energetics of protein-ligand interaction. The implicit electrostatic solvation model is achieved by assuming the solvent as a continuum high-dielectric-constant medium through the use of numerical solutions of Poisson equation [221] or a generalized-Born approximation [222]. Realistic conformational sampling, for example, through the inclusion of protein conformational changes is often done for lead compounds. The objective of this atomic-detail refinement of initial docking poses is threefold: a) improved judgment if ligand will actually engage the target, b) accurate prediction of complex conformation, and c) accurate prediction of binding affinity. The following example illustrates this two-stage approach.

Inhibitors of casein kinase by hierarchical docking

Casein kinase 2 (CK2), a target for antineoplastic and anti-infectious drugs, is involved in a large variety of important cell functions and many viruses exploit CK2 as phosphorylating agent of proteins essential to their life cycle. Cozza *et al* [223, 224] used a hierarchical docking process to identify a potent inhibitor from an in-house molecular database containing approximately 2000 compounds including several families of polyphenolic compounds including catechins, coumarins, and others. A rigid body docking step using FRED was used to dock ligand conformations generated by OMEGA v.1.1. The top 50% of poses ranked by FRED score were selected, and one unique pose for each of the best-scored compounds used for subsequent steps. The selected poses were optimized via a flexible ligand-docking protocol with three different programs: MOE-DOCK, GLIDE and GOLD. A consensus scoring scheme was developed in which each docked set, i.e. FRED-DOCK, FRED-GLIDE and FRED-GOLD was scored by five different scoring functions MOE-Score, GlideScore, GoldScore, ChemScore and Xscore, leading to three docking/scoring sets. Common compounds among the top 5% of compounds ranked by consensus scores from each list were prioritized for *in vitro* testing. The hierarchical docking process allowed identification of nanomolar CK2 inhibitors such as ellagic acid (IC_{50} 40 nM) and quinalizarin (IC_{50} 50 nM).

1.3.8 Binding Site Characterization

The success of SB-CADD methods depends on the understanding of physiochemical interactions between molecules. Optimization of lead molecules into high-affinity compounds that can be tested *in vivo* requires both the optimization of binding affinity and pharmacological properties. This process requires a deep understanding of the molecular interactions between ligand and target. Structural studies aimed at elucidating binding modes are commonly done using experimental methods such as X-ray and NMR. However, the time necessary to generate samples and determine structures can prevent applicability to the repetitive cycles of lead optimization. Computational methods such as molecular docking, molecular dynamics simulation, and quantum-mechanical simulations can be used to accelerate this process.

Experimentally determined protein structures in complex with ligand often serve as starting point for SB-CADD campaigns. For example, the cocrystal structure (PDB code 2BEL) of 11β-hydroxysteroid dehydrogenase (11β-HSD1) and its inhibitor, a semisynthetic derivative of 18β-glycyrrhetnic acid was used to generate a model of the binding site. Increased expression of 11β-HSD1 in liver and adipose tissue has been linked to obesity, insulin resistance, diabetes, and cardiovascular diseases in humans. The crystal structure illustrates interaction of carbenoxolone and active site residues Ser170, Tyr183 and Lys187, as shown in figure 1.5. In addition, two hydrophobic pockets exist on either side of the catalytic site which is exploited by a number of adamantine containing 11β-HSD1 inhibitors. A proprietary structure-based drug design program, Contour, was used to develop binding models of inhibitors containing an N-(2-adamantyl) amide moiety. Structural insight of binding site allowed the investigators to apply ligands containing an N-(2-adamantyl) amide moiety in a drug design program. With the help of the model and modeling studies, the authors discovered an 11β-HSD1 inhibitor which is orally bioavailable in three species and is active in a primate pharmacodynamic model [225].

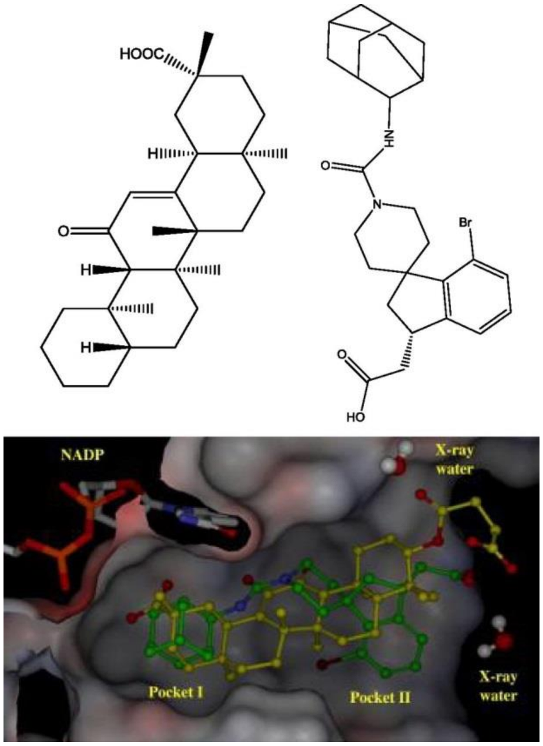


Figure 1.5 Carboxynoxolone and 10j2. Overlap of carenoxolone (yellow) and urea 10j2 (green) in binding site of 11*β*-HSD1. Source: [226].

1.3.9 Pharmacophore Model

A pharmacophore model of the target binding site summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Molecular properties most commonly used to define pharmacophores include hydrogen bonding potential (acceptors and donors), basic groups, acidic groups, partial charges, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening, *de novo* design, and lead optimization [227]. A pharmacophore model of the target binding site can be used to virtually screen a compound library for putative hits. Apart from querying a database for active compounds, pharmacophore models can also be used by *de novo* design algorithms to guide the design of new compounds.

Structure-based pharmacophore models are developed based on an analysis of the target binding site or on a target-ligand complex structure. LigandScout [228] uses protein-ligand complex data to map interactions between ligand and target. A knowledge based rule set obtained from the PDB is used to automatically detect and classify interactions into hydrogen bond interactions, charge transfers, and lipophilic regions [228]. The Pocket v.2 [229] algorithm is capable of automatically developing a pharmacophore model from a target-ligand complex. This algorithm creates regularly spaced grids around the ligand and the surrounding residues. Probe atoms that represent a hydrogen bond donor, a hydrogen bond acceptor, and a hydrophobic group are used to scan the grids. An empirical scoring function, SCORE, is used to describe the binding constant between probe atoms and the target. SCORE includes terms to account for van der Waals interactions, metal-ligand bonding, hydrogen bonding and desolvation effects upon binding [230]. A pharmacophore model is developed by rescoring the grids followed by clustering and sorting to extract features essential for protein-ligand interaction. During rescoring, hydrogen bond donor/acceptor scores lower than 0.2 and hydrophobic scores lower than 0.47 are reset to zero. Grids with three zero scores are filtered out, and the “neighbor number” for each grid is determined by counting the number of grids within 2 Å having non-zero score for a particular type. Grids with less than 50 donor neighbors, 30 acceptor neighbors, and 40 hydrophobic neighbors are reset to zero for their donor score, acceptor score, and hydrophobic scores, respectively. Grids are filtered by eliminating those with three zero scores, leaving only those grids that represent key interaction sites. The algorithm then superimposes the ligand on the grid, and a given grid is selected as a candidate if it is close to an atom type that can mediate the same interaction. Candidates with non-zero donor, acceptor, or hydrophobic scores are gathered into separate clusters, and the grid with highest score is defined as the center of donor, acceptor, or hydrophobic property.

Virtual screening using a pharmacophore model

17β-hydroxysteroid dehydrogenase type 1 (17β-HSD1) plays an important role in the synthesis of the most potent estrogen estradiol. Its inhibition could be important for breast cancer prevention and treatment. Schuster *et al* [231] used LigandScout2.0 to generate pharmacophore models of 17β-HSD1 from cocrystallization complexes with inhibitors (PDB codes 1EQU and 1I5R). These pharmacophore models represent the binding mode of a steroidal compound and small hybrid compounds (consisting of a steroidal part and an adenosine), respectively. The 1I5R-based pharmacophore model was used to screen the NCI and SPECS databases for new inhibitors using CATALYST. Best scoring hit compounds were docked into the binding pocket of 1EQU using GOLD, and final selection for *in vitro* testing was performed according to the best fit value, visual inspection of predicted docking pose and the ChemScore (GOLD scoring function) value. Four of 14 compounds tested *in vitro* showed an IC_{50} value of less than 50 µM with the most potent being 5.7 µM. Brvar *et al* [232] applied pharmacophore models to discover novel inhibitors of bacterial DNA gyrase B, a bacterial type II topoisomerase originating from gyrase and a target for antibacterial drugs. A pharmacophore model obtained using LigandScout was used to screen the ZINC database which yielded a novel class of thiazole-based inhibitors with IC_{50} value of 25 µM.

Multitarget inhibitors using common pharmacophore models

Wei *et al* [233] used Pocket v.2 to identify a common pharmacophore for two targets involved in inflammatory signaling, human leukotriene A4 hydrolase (LTA4H-h) and human nonpancreatic secretory phospholipase A2 (PLA2). The cocrystal structure (PDB code 1HS6) of LTA4H-h with 2-(3-amino-2-hydroxy- 4-phenylbutyrylamino)-4-methyl-pentanoic acid (bestatin) and the structure (PDB code 1DB4) of PLA2 with [3-(1-benzyl-3-carbamoylmethyl-2-methyl-1H-indol-5-yloxy)propyl]phosphonic acid (indole 8) were used to derive pharmacophores of the two targets. For LTA4H-h, six pharmacophore centers were identified that included four hydrophobic centers, one hydrogen bond acceptor, and one zinc metal coordination pharmacophore. In the binding pocket of PLA2, three hydrophobic centers, one hydrogen bond acceptor, and two calcium ion coordination centers were identified. The comparison of two sets of pharmacophore models revealed that two hydrophobic pharmacophores and a pharmacophore that coordinated with metal, shown in figure 1.6, was common to both proteins. The authors hypothesized that compounds satisfying the common pharmacophores would inhibit both the proteins. The MDL chemical database was screened virtually with LTA4H-h and PLA2 using Dock4.0 and binding conformation of top 150,000 compounds (60% of database) ranked by Dock score were extracted and checked for conformity to common pharmacophores. This identified 163 compounds whose binding conformations were reanalyzed using Autodock3.5 followed by comparison with common pharmacophores. Finally, nine compounds whose conformations matched the common pharmacophores were tested *in vitro* for binding with PLA2 and LTA4H-h. The best inhibitor, compound 10, inhibited LTA4H-h at submicromolar range and inhibited PLA2 with an IC_{50} value of 7.3 µM.

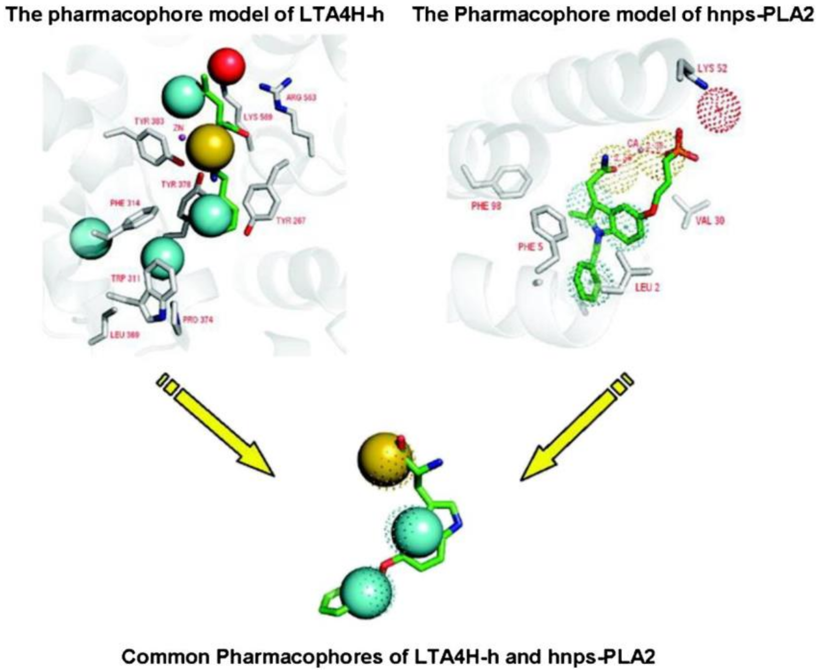

The pharmacophore model of LTA4H-h
The Pharmacophore model of hnps-PLA2
ARG 563
LYS 589
TYR 383
ZN
TYR 378
PHE 314
TYR 267
TRP 311
PRO 374
LEU 369
LYS 52
CA
PHE 98
VAL 30
PHE 5
LEU 2
Common Pharmacophores of LTA4H-h and hnps-PLA2

Figure 1.6 Extracting common pharmacophores of LTA4H-h and human-PLA$_2$. Cyan spheres depict hydrophobic centers, red spheres represent H-bond acceptor, and yellow spheres stand for feature that coordinates with a metal. Source: [233].

1.3.10 Automated de novo Design of Ligands

De novo structure-based ligand design can be accomplished by either a ligand-growing or ligand-linking approach. With the ligand-growing approach, a fragment is docked into the binding site and the ligand is extended by adding functional groups. The linking method, on the other hand, docks multiple small fragments into adjacent binding pockets and then links them to form a single compound. This approach is a computational version of the popular SAR by NMR technique introduced by Shuker *et al* [234].

Several methods have been developed implementing both ligand-growing and ligand-linking strategies for designing ligands that can bind to a given target. LigBuilder [235] builds ligands in a step by step fashion using a library of fragments. The design process can be carried out by various operations like ligand growing and linking and the construction process is guided by a genetic algorithm. The target-ligand complex binding affinity is evaluated with an empirical scoring function. The program first reads the target protein and analyzes the binding pocket. Depending on user preference, it can then either use a growing or a linking strategy. In the growing strategy, a seed structure is placed in a binding pocket and the program replaces user defined growing sites with candidate fragments. This gives rise to a new seed structure that can then be used in further rounds of growing. For the linking strategy, several fragments placed at different locations on the target protein serve as the seed structure. The growing scheme happens simultaneously on each fragment biased towards linking these fragments. The LUDI [204] algorithm, which precedes LigBuilder, uses a ligand linking strategy. It positions seed fragments into binding pockets of the target structure, optimizing their interactions individually before. This step is followed by linking the fragments into a single molecule. The synthetic accessibility of ligands can also be taken into account. For example, LigBuilder 2.0 analyzes designed using a chemical reaction database and a retrosynthesis analyzer [236].

The biggest challenge of *de novo* drug design is inseparable from its greatest advantage. By defining compounds that have never been seen before, one is invariably necessitating synthetic effort for acquisition prior to testing. This forces any *de novo* protocol to incorporate synthesizability metrics into its scoring. This increases the effort required in terms of cost, time, and expertise. Synthesizability is most important when designing a large number of different compounds and scaffolds. One tool that approaches the constraint of synthesizability is SYNOPSIS (SYNthesize and OPtimize System in Silico) [237], which enforces synthesizability throughout the design process by starting with available compounds and creating novel compounds by virtually using known chemical reactions. 70 different reaction types may be selected based on the presence of different functional groups in the evolving molecule. SYNOPSIS also provides additional restraints for desired properties such as solubility.

De novo design by linking fragments has been successfully applied in the design of inhibitors of p38 MAPK [238], a key regulator in signaling pathways that control the production of cytokines such as tumor necrosis factor-α and interleukin-1β. Inhibitors of MAPK can potentially be used for the treatment of various autoimmune diseases. Figure 1.7***A*** shows four classes of interactions of a clinical compound BIRB 796 with MAPK: (1) interaction with residues in ATP binding site (Met109), (2) interaction with the "Phe pocket" (dotted arc), (3) hydrophobic interaction with the kinase specificity pocket (solid arc), and (4) interaction of the urea with backbone NH-bond of Asp168 and carboxylate of Glu71. A design strategy for exploring structurally distinct scaffolds by leveraging the interactions of BIRB 796 [1-(5-tert-butyl-2-p-tolyl-2H-pyrazol-3-yl)-3-[4-(2-morpholin-4-yl-ethoxy)naphthalen-1-yl] was devised as follows: a) A tert-butyl group was used as "Phe pocket" seed structure in place of pyrazole ring of BIRB 796 b) An N-formyl group was appended to tert-butyl fragment to access the hydrogen bonds with Glu71 and Asp168 c) a carbonyl group was used as the second seed fragment to access the hydrogen bond with Met109 as shown in figure 1.7***B***. LigandBuilder software was used to link the two seed fragments, the tert-butyl linked to N-formyl group, and the carbonyl group. The program consistently introduced a 4-tolyl group in the kinase specificity pocket. However, LigandBuilder failed to predict favorable rigid linkers for connecting tolyl group to carbonyl group which would be essential for carbonyl display at the proper distance to interact with Met109. Modeling indicated N-linked azoles connected to tolyl group via an N-linkage as a suitable linker. Derivatives of this designed molecule were synthesized leading to the discovery of compound 28 as shown in figure 1.7***D***, with an IC_{50} of 83 nM.

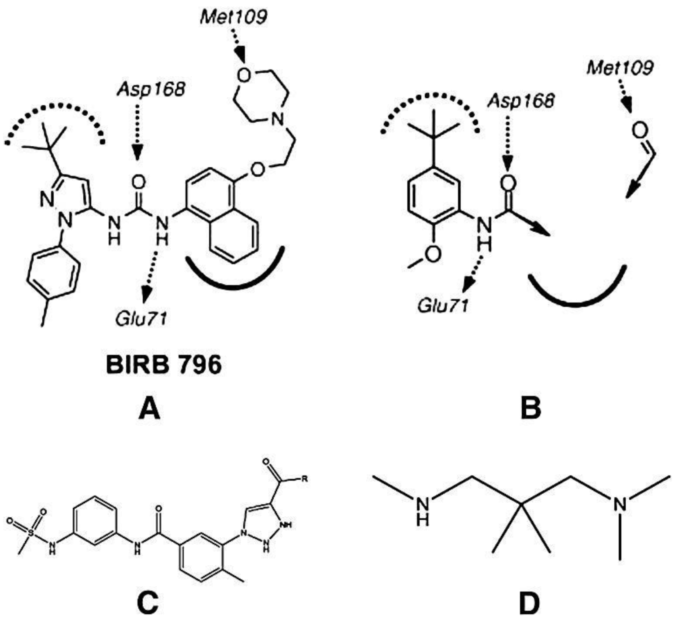


Figure 1.7 Design strategy for inhibitors of p38 MAPK. A) Key interactions of BIRB-796 inhibitor with MAPK. B) A fragment linking strategy to link two seed structures was applied using LigBuilder. A tert-butyl

phenyl fragment was used in the first pocket, whereas a carbonyl fragment was used to access the hydrogen bond with Met109 in the second site. An N-formyl group was attached to the first seed fragment to access hydrogen bonds with Glu71 and Asp168. C) General structure of optimized structures which showed potent activity. D) R group for compound 28, which showed IC_{50} value of 83 nM. Source: [238].

`The fragment extension approach was employed by Zhang *et al* [239] in the discovery of vascular endothelial growth factor receptor 2 (VEGFR2) inhibitors, a therapeutic target for tumor-induced angiogenesis. The authors used quinazoline as the seed fragment, since three of the nine clinically approved kinase inhibitor drugs are 4-anilinoquinazoline derivatives [240]. These inhibitors bind the active site of their respective targets such that the quinazoline ring is located at the front of ATP binding pocket. The ligand building process involved placing the quinazoline fragment in the binding pocket in the same orientation as found for known inhibitors. The design strategy sought to create a ligand that would extend to fit a specific hydrophobic pocket at the back of the ATP binding cleft. An NH_2, OH, or SH group was added in the C4 position of the quinazoline ring to allow for a turn owing to orientation of quinazoline and the spatial arrangement of the hydrophobic pocket. A fragment-growth-based de novo method was applied in which various fragments (approximately 1200 fragments) were allowed to grow on the turn fragment to extend into the hydrophobic pocket. Designed molecules were then rescored and ranked using GOLD. The design process led to the development of a potent and specific VEGFR2 inhibitor, SKLB1002, [2-((6,7-dimethoxyquinazolin-4-yl)thio)-5-methyl-1,3,4-thiadiazole], shown in figure 1.8, that inhibits angiogenic processes in zebra fish embryo and athymic mice with human tumor xenografts.

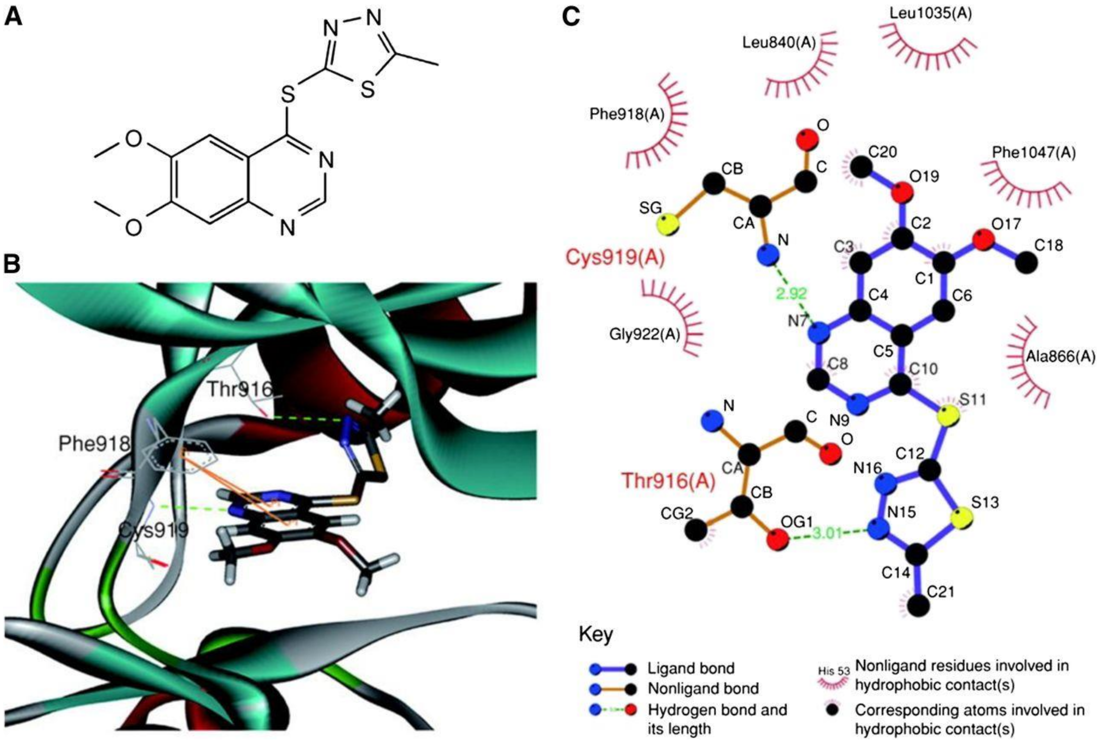


Figure 1.8 Computational design of novel VEGFR2 inhibitor SKLB1002. A) Chemical structure of SKLB1002. B) SKLB1002 is docked into the active site of VEGFR2, showing interactions between SKLB1002 and VEGFR2 by using the in silico model. C) A 2D interaction map of SKLB1002 and VEGFR2. Source: [239].

1.4 Ligand-Based Computer-Aided Drug Design

The ligand-based computer-aided drug discovery (LB-CADD) approach focuses on ligands known to interact with a target of interest. These methods analyze 2D or 3D structures of multiple ligands for the same target. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained, whereas extraneous information not relevant to the interactions is discarded. It is considered an indirect approach to drug discovery in that it does not necessitate knowledge of the target structure. The two fundamental approaches of LB-CADD are a) selection of compounds based on chemical similarity to known actives using some similarity measure or b) the construction of a QSAR model that predicts biological activity from chemical structure. Either approach can be applied to vHTS, hit-to-lead and lead-to-drug optimization, and the optimization of DMPK/ADMET properties. LB-CADD is based on the Similar Property Principle, published by Johnson *et al*, which states that structurally similar molecules are likely to have similar properties [241]. In contrast to SB-CADD, LB-CADD can also be used when the structure of the biological target is unknown. Additionally, active compounds identified by Ligand-Based virtual High-Throughput Screening (LB-vHTS) methods are often more potent than those identified in (SB-vHTS) [61].

1.4.1 Molecular Descriptors / Features

LB-CADD techniques use a variety of computational algorithms to describe small molecule features that balance efficiency and information content. The optimal descriptor set depends on the biological function predicted as well as on the LB-CADD technique used. Molecular descriptors can be structural as well as physicochemical and can be described on multiple levels of complexity. Chemical properties may include molecular weight, geometry, volume, surface area, ring content, rotatable bonds, interatomic distances, bond types, atom types, planar and non-planar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others [242-248]. These descriptors are generated through knowledge-based methods, graph-theoretical methods, molecular-mechanical, or quantum-mechanical tools [249, 250] and are classified according to the "dimensionality" of the chemical representation from which they are computed [33]: 1D, scalar physicochemical properties such as molecular weight; 2D, molecular constitution-derived descriptors, 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors. More complex descriptors often incorporate information from simpler ones. For example, many 2D and 3D descriptors use physicochemical properties to weight their functions and to describe the overall distribution of these properties.

Functional groups

Functional groups are defined by the International Union of Pure and Applied Chemistry (IUPAC) as atoms or groups of atoms that have similar chemical properties across different compounds. These groups are attached to a central backbone of the molecule, also called the scaffold or chemotype. The spatial positioning of the functional groups dictated by the backbone defines the physical and chemical properties of compounds. Therefore, the location and nature of functional groups for a given compound contain key information for most ligand-based CADD methods. There are many different kinds of functional groups composed primarily of hydrocarbons, halogens, oxygens, nitrogens, sulfur, and phosphorous including alcohols, esters, amides, carboxylates, ethers, nitro group, thiols, and many others [251].

Functional groups can either be explicitly described by their atomic composition and bonding or may be implicitly encoded with their general properties. For example, under physiological conditions carboxyl groups are often negatively charged, whereas amine groups are positively charged. This property is reflected both in the geometry of the functional group as well as its charge. Because it is the properties conferred by the functional groups that are most important to the biochemical activity of a given compound, many CADD applications treat functional groups containing different atoms but conferring the same properties as similar or even identical. For example, the capacity for hydrogen bonding can heavily influence a molecule's properties. These interactions frequently occur between a hydrogen atom and an electron donor such as oxygen or nitrogen. Hydrogen bonding interactions influence the electron distribution of neighboring atoms and the site's reactivity, making it an important functional property for therapeutic design. Commonly, hydrogen bonding groups are separated simply as hydrogen bond donors with strong electron-withdrawing substituents (OH, NH, SH, and CH) and hydrogen bond acceptors (PO, SO, CO, N, O, and S) [252, 253]. The applications Phase, Catalyst, DISCO, and GASP (Genetic Algorithm Superposition Program) as well as pharmacophore mapping algorithms discussed in greater detail below focus primarily on hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic regions, ionizable groups, and aromatic rings.

Prediction of physicochemical properties

Properties within the same dimensionality can show a range of complexity. The simplest properties, such as molecular weight and total hydrogen bond donors, may be rapidly and accurately computed. More complex properties such as solubility and partial charge, on the other hand, may be more difficult to compute but can provide higher information content [254]. Prediction of these complex physicochemical properties, though more computationally expensive, may be critical for an effective set of molecular descriptors. These trade-offs must be considered on a case-by-case basis when designing a LB-CADD project. Modern computational algorithms and approximations, however, allow for the incorporation of certain highly complex properties.

Electronegativity and partial charge

Electron distribution plays an important role in a molecule's properties and activities. Therefore, it was important to develop a descriptor capable of modeling the charge distribution over an entire molecule. One useful method is to assign a partial charge to all atoms in a molecule. Initially, electron distribution was assigned to individual atoms through quantum mechanical calculations. However, when screening thousands or millions of compounds, a much faster and more efficient method became necessary. Gasteiger and Marsili developed a method for assigning partial charges to individual atoms called the Partial Equalization of Orbital Electronegativity (PEOE) [255]. This method is based on a definition of electronegativity introduced by Mulliken that relates the electronegativity of an atom to its ionization potential I and electron affinity E with the equation electronegativity = ½(I+E) [256]. The values for E and I depend on the valence state of the atom and takes advantage of a concept of orbital electronegativity introduced by Hinze *et al* [257, 258] that describes the electronegativity of a specific orbital in a given valence state and depends on hybridization and occupation number of the orbital.

PEOE improves upon the concept of electronegativity equalization first proposed by Sanderson [259, 260] that states bonded atoms change electron density until total equalization of electronegativity is reached. Sanderson's simple model leads to chemically unacceptable calculations, necessitating a more complex model of electronegativity equalization. Gasteiger and Marsili first introduced an approximation function that joins the electronegativity values of an atom in its anionic, neutral, and cationic state with appropriate ionization potentials and electron affinities and relates orbital occupation with orbital electronegativity. They also added a damping function to account for the fact that charge transfer generates an electrostatic field that inhibits further electron transfer and prevents complete equalization. Finally, they introduced an iterative procedure to account for the changes in charge separations following a round of electronegativity modification. Progressive iterations include wider spheres of neighboring atoms until the total transfer drops below a cutoff. The total charge of an atom is then calculated as the sum of the individual charge transfers following the iteration.

For small-member rings, special bonds based on the valence bond model [261] were used as additional parameters in the PEOE method [262]. The valence bond model states that the bonds of three and four membered ring systems arise from orbitals with varying amounts of *s* and *p* character depending on the type and number of rings involved and whether exo- or endocyclic bonds are considered. The extra coefficients provided charge dependence for the different hybridization states interpolated from the values of electronegativities for sp^3, sp^2, *sp*, and *p* states [263].

Gasteiger and Saller later introduced a method for applying the PEOE method to molecules with multiple resonance structures [264]. Charge distribution in π-systems could be calculated on the basis of resonance structure weights. These weights were calculated by including a topological weight and electronic weight. The topological weight was based on whether resonance structures involved the loss of covalent bonds, decrease in aromatic systems, or charge separation. The electronic weight was based on the idea that resonance structures are more important when negative charge is localized on the more strongly electronegative atom. Therefore, it was a measure of how well the donor atom can donate its lone pair of electrons and how stable a negative charge on the acceptor atom is. To calculate this weight, the electronegativity concept is applied. Finally, by adding the changes in charge of the individual resonance structures to the scaling factor the charge distribution could be calculated.

Additionally, orbital electronegativity is often separated into σ and π bond systems. Standard connection tables describe connections between two atoms as twice the number of electrons per bond order (single bonds contain two electrons; double bonds contain four, etc). This valence bond structure, however, is insufficient to describe some compounds and may fail to distinguish between the different excited states of a molecule. Separating σ and π electrons has been shown to be advantageous to this representation scheme [265]. Bauershmidt and Gasteiger describe computational representation of chemical species using three electron systems: σ-electron systems, π-electron systems, and coordinative bonds [266].

σ-electron systems contain electrons localized in the σ part of a bond and single bond electrons. These systems may consist of more than two atoms when multicenter bonds are described, including overlapping orbitals that point into a central region between bonded atoms and open bridging α-electron systems where one atom is located between the other atoms part of the same system. π-electron systems encode free electrons. One π-electron system is generated for each electron pair. For example, the electrons of a triple bond are distributed into one σ-electron system and two π-electron systems, each with two electrons. Properties such as orbital electronegativity and partial charges are more accurately described using the σ- and π-electron systems. Therefore, it is common to implement descriptors separated as σ charges, π charges, σ electronegativity, and π electronegativity.

These methods provide a means to quantitatively calculate electronegativity and partial charge on a per-atom basis without the need for quantum mechanics. PEOE charges have been shown to be useful information for predicting chemical properties such as taste [267]. Additionally, these properties are often used to weight three-dimensional descriptors that would, on their own, only capture purely structural information. By weighting these descriptors with these properties, information regarding the three-dimensional distribution of electrons is available.

Polarizability

Effective polarizability or mean molecular polarizability is another widely used molecular descriptor. It quantifies the response of electron density to an external field leading to an induced dipole moment [268]. Polarizability contributes to dispersion forces and influences intermolecular interactions. Brauman and Blair described stabilization effects of substituent polarizability [269]. For example, induced dipole moments in unsubstituted alkyl groups are believed to stabilize charges in gaseous ions formed by protonation or deprotonation [270]. The magnitude of the induced dipole is calculated as the product of the electric field operator and the polarizability tensor of the molecule. The average polarizability of a molecule is calculated as the average of the three principal components of this tensor [271].

Miller and Savchik introduced a formula for calculating mean molecular polarizabilities using a polarizability contribution for each atom based on its atom type and hybridization state and the total number of electrons in the molecule [272]. Gasteiger and Hutchings improved this formula to account for the attenuation of substituent influence. This was accomplished through the introduction of a damping factor dependent on the distance in bonds between the atom and the charged reaction center [273].

Glen [271] defined a method for calculating static molecular polarizability using a modified calculation of atomic nuclear screening constants based on effective nuclear charge described by Slater [274]. This calculation divides electrons into different groups with different shielding constants. These shielding constants reflect the fact that inner-shell electrons modify the view of the nucleus for outer-shell electrons and adjust the field of nuclear charge for each group of electrons.

Octanol/water partition coefficient

LogP (logarithm of partition coefficient between *n*-octanol and water) is an important molecular descriptor that has been widely used in QSAR since the work of Leo *et al* [275]. Lipinksi's rule of five, a class set of rules describing the "druggability" of a compound, includes measurement of the compound's logP. Traditionally, logP is determined experimentally by measuring its partitioning behavior in the insoluble mixture of *n*-octanol and water and reflects the molecule's hydrophobicity. This molecular property has been shown to be important in solubility, oral availability, transport, penetration of the blood-brain-barrier, receptor binding, and toxicity [276, 277]. For virtual screening applications, several methods for calculating logP based on molecular constitution have been established.

LogP calculations largely rely on an additive method introduced by Rekker and Mannhold [278] where the contributions to logP by basic fragments of a molecule (atoms and functional groups) are summed. Additivity methods improved with the incorporation of additional molecular properties have also been used to calculate logP [279, 280].

Wang *et al* developed the very popular additivity method called XLOGP [281]. This method originally defined 80 basic atom types for carbon, nitrogen, oxygen, sulfur, phosphorous, and halogen atoms. Hydrogen atoms are implicitly included in the different atom types. This method was later improved to include 90 atom types and ten correction factors [282].

Additional corrections became necessary when many simple summation approaches resulted in incorrect logP calculations. These corrections account for specific intramolecular interactions affecting a molecule's logP beyond individual fragments. For example, simple summation underestimated compounds with long hydrocarbon chains due to their flexibility and aggregation behavior. Additional interactions that can obscure simple fragment summation include dipole shielding in compounds containing two or more halogen atoms, internal hydrogen bonding, the unusually strong internal hydrogen bonding with salicylic acids, and the existence of α-amino acids as zwitterions. Correction factors are often included for aromatic nitrogen pairs, ortho sp^3 oxygen pairs, para donor pairs, sp^2 oxygen pairs, and amino sulfonic acids.

Xing and Glen introduced an alternative logP calculation that was based on the evidence that molecular size and hydrogen-bonding account for a major part of logP [283]. They created a statistical model by combining molecular size and dispersion interactions using molecular polarizability and the sum of squared partial atomic charges on oxygen and nitrogen atoms. The final model showed that molecular polarizability is more significant than atomic charges and that an increase in polarizability is correlated with an increase in logP, whereas a decrease in charge densities on nitrogen and oxygen correlated with a decrease in logP. They theorized that the importance of molecular polarizability on logP was due in part to the relative energy required for a larger molecule to create a cavity in water or octanol.

Converting properties into descriptors

Molecule properties must be converted into numerical vectors known as descriptors for use in LB-CADD. For many applications, descriptors must have a constant length independent of molecule size. Each position in the vector of descriptors therefore encodes a well-defined property or feature which facilitates the direct comparison of two compounds via mathematical algorithms.

Binary molecular fingerprints

Fingerprints are bit string representations of molecular structure and/or properties [284-286] where a 1 indicates the presence of a particular functional group or property and 0 indicates its absence. This allows chemical identity to be unambiguously assigned entirely by the presence or absence of a specific set of features [287]. The features described in a molecular fingerprint can vary in number and complexity (from hundreds of bits for structural fragments to thousands for connectivity fingerprints, and millions for the complex pharmacophore-like fingerprints) [286], depending on the computational resources available and the intended application. Fingerprints which rely solely on interatomic connectivity, i.e., molecular constitution, are known as 2D fingerprints [287]. In the prototypic 2D keyed fingerprint design, each bit position is associated with the presence or absence of a specific substructure pattern – for example carbonyl group attached to sp^3 carbon, hydroxyl group attached to sp^3 carbon, etc. [288].

Molecular structure itself comprises several levels of organization between the atoms within a molecule and, therefore, fingerprints may differ in their own levels of organization. For example, the simplest fingerprint may state that a given compound contains six carbon atoms and six hydrogen atoms. However, up to 217 different isomers may be encoded by this fingerprint. 2D fingerprints containing connectivity may distinguish between some of these isomers, however, stereochemistry, which separates compounds with identical constitutions, is beyond the realm of most 2D fingerprints. One extension to fingerprints is the use of hash codes. These are bit strings of fixed length that contain information about connectivity, stereo centers, isotope labeling, and other properties. This information is compressed to avoid redundancies [289]. Unfortunately, it is not always obvious which of these characteristics are important in a given context and which are not [287].

Commonly used fingerprints include the ISIS (Integrated Scientific Information System) keys with 166 bits and the MDL (Molecular Design Limited) MACCS (Molecular ACCess System) keys [290] with 960 bits. The ISIS keys are small topological substructure fragments while the MACCS keys consist of the ISIS keys plus algorithmically generated more abstract atom-pair descriptors. MDL keys are commonly used when optimizing diversity [291]. For example, the PubChem database uses a fingerprint that is 881 bits long to rank substances against a query compound. This fingerprint is comprised of the number and type of elements, ring systems (saturated and unsaturated up to a size of 10), pair-wise atom combinations, sequences, and substructures [287].

2D Description of molecular constitution

2D descriptors can be computed solely from the constitution or topology of a molecule, whereas 3D descriptors are obtained from the 3D structure of the molecule [33]. Many 2D molecular descriptors are based on molecular topology derived from graph-theoretical methods. Topological indices treat all atoms in a molecule as vertices and index-specific information for all pairs of vertices. A simple topological index, for example, will contain only constitutional information such as which atoms are directly bound to each other. This is known as an adjacency matrix and an entry of 1 for vertices v_i and v_j indicates their corresponding atoms are bonded while an entry of 0 for v_i and v_i indicates that the corresponding atoms are not [292]. For an adjacency matrix, the sum of all entries is equal to twice the total number of bonds in the molecule.

Complex topological indices are created by performing specific operations to an adjacency matrix that allow for the encoding of more complex constitutional information. These indices are based on local graph invariants that can represent atoms independent of their initial vertex numbering [293]. For example, topological indices may contain entries for the number of bonds linking the vertices. Information gathered from such an index can include the number of bonds linking all pairs of atoms and the number of distinct ways a path can be superimposed on the molecular graph. A topological index that includes information such as heteroatoms and multiple bonds through the weighting of vertices and edges was introduced by Bertz [294]. Randic and Basak introduced an augmented adjacency matrix by replacing the zero diagonal entries (where $v_i = v_j$) with empirically obtained atomic properties. This adjacency matrix includes atom type information as well as connectivity [295]. Topological indices that describe the molecular charge distribution as evaluated by charge transfers between pairs of atoms and global charge transfers have also been developed [296, 297]. Additionally, topological indices known as geometrical indices have been derived to describe molecular shape. For example, the shape index E measures how elongated the molecular graph is [296, 298]. Statistical methods such as linear discriminant analysis are often applied to topological indices and biological properties to create predictive descriptors relating indices to molecular activity [296, 299].

Topological autocorrelation (2D autocorrelation) is designed to represent the structural information of a molecular diagram as a fixed-length vector that can be applied to molecules of any shape or size. It encodes the constitutional information as well as atom property distribution by analyzing the distances between all pairs of atoms. Topological autocorrelations are independent of conformational flexibility because all distances are measured as the shortest path of bonds between the two atoms. The autocorrelation vector is created by summing all products for atom pairs within increasing distance intervals in terms of number of bonds. In other words, it creates a frequency plot for a specific range of atom pair distances. By including atom property coefficients for all atom pairs, autocorrelations are capable of plotting the arrangement of specific atom properties. For example, information such as the frequency at which two negatively charged atoms are three bonds apart versus four bonds apart is stored in an autocorrelation plot weighted by partial atomic charge [300].

3D Description of molecular configuration and conformation

The physicochemical meaning of topological indices and autocorrelations is unclear and incapable of representing some qualities that are inherently three-dimensional (stereochemistry). 3D molecular descriptors were developed to address some of these issues [301].

The 3D Autocorrelation is similar to the 2D autocorrelation but measures distances between atoms as Euclidian distances between their 3D coordinates in space. This allows a continuous measure of distances and encodes the spatial distribution of physicochemical properties. Instead of summing all pairs within discrete shortest path differences, the pairs are summed into interval steps [302].

Radial distribution function (RDF) is another very popular 3D descriptor. It maps the probability distribution to find atoms in a spherical volume of radius r. In its simplest form, the RDF maps the interatomic distances within the entire molecule. Often it is combined with characteristic atom properties to fit the information requirements [242]. RDFs not only provide information regarding interatomic distances between atoms and properties, they contain information such as bond distances, ring types, and planar versus nonplanar molecules. These functions allow estimation of molecular flexibility through the use of a “fuzziness” coefficient that extends the width of all peaks to allow for small changes in interatomic distances.

GRIND (Grid-Independent Descriptor) is another 3D descriptor that does not require prior alignment [303]. GRIND was designed to retain characteristics that could be directly traced to the molecules themselves, rather than producing purely mathematical descriptors that are not obviously related to the molecular structures they describe. GRIND is comprised of three steps. The first step calculates a molecular-interaction field (MIF). Probes with different chemical properties to scan the molecule and identify regions showing favorable interaction energy [304].

Initial MIFs may contain up to 100,000 nodes. Therefore, the second step of GRIND reduces this set of nodes to focused regions of greatest favorable interaction energies. Initial implementation of GRIND used a Fedorov-like optimization algorithm [305] to reduce the number of nodes to several hundred by considering both the intensity of a field and the mutual node-node distances between the selected nodes. In the second iteration of GRIND (GRIND-2), this method was replaced with a new algorithm called AMANDA [306]. While the original GRIND requires users to define the number of nodes to extract per molecule, AMANDA allows GRIND-2 to automatically adjust the number of nodes per compound. After a prefiltering step that removes all nodes failing an energy cutoff, every atom in the molecule is assigned a set of nodes and the number of nodes to extract per atom is calculated using a weighting factor and function that automatically assigns additional nodes to larger regions. The node selection uses a recursive technique designed to assign initial selection weight based entirely on energy values. As the iterations continue through lower energy nodes, however, the internode distances become more important than the individual energy score of each node.

The final step of GRIND-2 (and GRIND) encodes this set of nodes into descriptors using auto- and cross-correlation methods. Pairs of interaction energies are multiplied and only the greatest product is retained for each inter-node distance. This is called maximum auto- and cross-correlation (MACC) and allows GRIND-2 (and GRIND) to contain information that directly correlates with the initial molecular structure.

GRIND-PP [307] improves GRIND-2 by removing much of the inherent repetition in the calculated descriptors. Structural features are often repeated across many GRIND-2 variables which can artificially weight certain features and reduce computational efficiency [308]. Principle Properties (PP) replace the original variables in GRIND and are calculated using principle component analysis. These variables are linear combinations of the original variables selected to explain as much of the variance in the original set of variables as possible.

Comparative field molecular analysis (CoMFA) [248] is a 3D-QSAR technique that aligns molecules and extracts aligned features that can be related to biological activity. This method focuses on the alignment of molecular interaction fields rather than the features of each individual atom. CoMFA was established over 20 years ago as a standard technique for constructing 3D models in the absence of direct structural data of the target. In this method, 3D molecules are aligned within a grid and the values of steric (Van der Waals interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. Comparative Molecular Similarity Indices (CoMSIA) is an extension to CoMFA where the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type functions are used to avoid extreme values [309]. These methods, however, are limited to static structures with similar scaffolds and neglect the dynamical nature of the ligands [249].

1.4.2 Molecular fingerprint and similarity searches

Molecular fingerprint-based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or to cluster collections based on structural similarity. These methods are less hypothesis driven and less computationally expensive than pharmacophore mapping or QSAR models. They rely entirely on chemical structure and omit known biological activity of the compound, making the approach more qualitative in nature than other LB-CADD approaches [286]. Additionally, fingerprint-based methods consider all parts of the molecule equally and avoid focusing only on parts of a molecule that are thought to be most important for activity. This is less error prone to overfitting and requires smaller datasets to begin with. However, model performance suffers from the influence of unnecessary features and the often narrow chemical space evaluated [286]. Despite this drawback, 2D fingerprints continue to be the representation of choice for similarity-based virtual screening [310]. Not only are these methods the computationally least expensive way to compare molecular structures [287], but their effectiveness has been demonstrated in many comparative studies [310].

Similarity searches in LB-CADD

Fingerprint methods may be employed to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be tested for improved activity over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for *in vitro* testing [4]. Bologa *et al* used 2D fingerprint and 3D shape-similarity searches to identify novel agonists of the estradiol receptor family receptor GPR30. Estrogen is an important hormone responsible for many aspects of tissue development and physiology [311, 312]. The GPCR GPR30 has recently been shown to bind estrogen with high affinity and its specific role in estrogen-regulated signaling is being studied [313]. This group used virtual screening to identify compounds selective for GPR30 that could be used to study this target. 10,000 molecules provided by Chemical Diversity Laboratories were enriched with GPCR binding ligands and screened for fingerprint-based similarity to the reference molecule 17β-estradiol. Fingerprints used were Daylight and MDL and similarities were scored using Tanimoto and Tversky scores. The top 100 ranked hits were selected for biological testing and a first-in-class selective agonist with a K_i of 11 nM for GPR30 was discovered. [314].

Stumpfe *et al* used SecinH3 and analogs as reference compounds for a combined fingerprint and fingerprint-based support vector machine modeling screen aimed at inhibitors targeting the multifunctional cytohesins. Cytohesins are small guanine nucleotide exchange factors that stimulate Ras-like GTPases, which control various regulatory networks implicated in a variety of diseases [315-320]. The group screened approximately 2.6 million compounds in the ZINC database [29] and the top 145 candidates were selected for biological testing. Of those tested, 40 compounds showed measurable activity and 26 were more potent than SecinH3 [321].

Ijjalli *et al* created 2D pharmacophoric fingerprints using a query data set of 19 published T-type calcium channel blockers. T-type calcium channels underlie the generation of rhythmical firing patters in the CNS and have been implicated in the pathologies of epilepsy and neuropathic pain [322-324]. Specifically, T-type calcium channel 3.2 has been identified as a promising target for novel analgesic drugs for pathological pain syndromes [324]. A database of two million compounds was collected from various commercial catalogues and filtered for drug-like qualities, uniqueness, and standardization. The group used ChemAxon's PF and CGC GpiDAPH3 [325] fingerprints and tested a subset of 38 unique hits biologically. 16 hits showed more than 50% blockade of CaV3.2-mediated T-type current. These compounds proved to be an interesting collection of T-type calcium channel blockers. Some showed reversible inhibition, whereas others resulted in irreversible inhibition, and one of the compounds caused alterations in depolarization/repolarization kinetics [326].

In addition to the enrichment of lead compound population, fingerprints are also used to increase molecular diversity of test compounds. Fingerprints can be used to cluster large libraries of hits to allow the sampling of a wide range of compounds without the need to sample the entire library. The Jarvis-Patrick method clusters compounds by calculating a list of nearest neighbors for each molecule. Two structures cluster together if they are in each-others list of nearest neighbors and they have at least K of their J nearest neighbors in common. The MDL keys are also used to eliminate compounds least likely to satisfy the drug-likeness criterion [291].

Polypharmacology: similarity networks and off-target predictions

Chemical similarity measures such as Tanimoto coefficients are being used to generate networks capable of clustering drugs that bind to multiple targets in an effort to predict novel off-target effects. Keiser et al [327] used a Similarity Ensemble Approach (SEA) [328] to compare drug targets based on the similarity of their ligands. SEA predicts whether a ligand and target will interact using a statistical model to control for chemical similarity due to chance. Sets of ligands that interact with each target are compared by calculating Tanimoto coefficients based on standard 2D Daylight fingerprints [329] for each pair of molecules between two sets. Raw similarity scores between all pairs of ligand sets are calculated as the sum of all Tanimoto coefficients between the sets greater than 0.57. Because the probability of achieving Tanimoto coefficients greater than 0.57 increases with set size, this is normalized by expected similarity due entirely to chance. This model for random chemical similarity is achieved by randomly generating 300,000 pairs of molecule sets spanning logarithmic size intervals from 10 to 1000 molecules. Expectation scores are calculated based on raw scores and the probability of achieving the raw score by random chance and used to sequentially link ligand sets into a clustered map. Keiser et al collected over 900,000 drug-target comparisons from 65,241 ligands and 246 targets in the MDL Drug Data Report database [330] to generate a target similarity network. Another drug database, WOMBAT [331] included interactions not listed in the MDDR database and the authors tested the predictability of their networks by searching their networks for interactions found in WOMBAT but not MDDR. They found that 19% of the off-target effects listed in WOMBAT but not in MDDR were captured in their network. In addition to those found in MDDR and WOMBAT, 257 additional drug-target predictions were captured in their network, 184 of which had not been documented. The authors tested 30 of these undocumented predictions using radioligand competition assays and verified 23 interactions with binding constants less than 15 μM. Some of these interactions may help to explain well-known side effects. For example, the authors discovered an interaction between β-adrenergic receptors and Selective Serotonin Reuptake

Inhibitors Prozac (fluoxetine) and Paxil (paroxetine). This may explain the selective serotonin reuptake inhibitors discontinuation syndrome seen with these drugs that are analogous to discontinuation syndrome seen with β-blockers.

Lounkine et al [332] used the SEA approach combined with adverse drug reaction (ADR) information to generate a drug-target-ADR network. This network was then used to predict off-target interactions that may explain specific ADRs. The authors experimentally tested 694 predictions and verified 151 interactions with IC_{50} values less than 30 µM. The clinical relevance of these off-target interactions was explored through the enrichment of target-ADR pairs within their network. For example, abdominal pain has been reported for 45 drugs that interact with COX-1, and based on their network, the ADR-target pair abdominal pain-COX-1 was enriched (represented in a greater degree within the network than average) 2.3-fold, reflecting a predicted correlation between abdominal pain and COX-1 interaction. Another target-ADR correlation is predicted for sedation and H1 interaction with an enrichment of 4.9.

1.4.3 Quantitative Structure Activity Relationship models

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals [333]. Classical QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules [334]. In 1964, Free and Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution), and the two methods were later combined to create the Hansch/Free-Wilson method [335, 336].

The general workflow of a QSAR-based drug discovery project is to first collect a group of active and inactive ligands and then create a set of mathematical descriptors that describe the physicochemical and structural properties of those compounds. A model is then generated to identify the relationship between those descriptors and their experimental activity, maximizing the predictive power. Finally, the model is applied to predict activity for a library of test compounds that were encoded with the same descriptors. Success of QSAR depends not only on the quality of the initial set of active/inactive compounds, but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this “training” set of compounds will not be represented in the final model, and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set. For a comprehensive guide on performing a QSAR-based virtual screen, please see the review by Zhang [333].

Multidimensional QSAR: 4D and 5D Descriptors

Multidimensional QSAR (mQSAR) goes beyond the self-contained properties of a compound and quantifies all energy contributions of ligand binding including desolvation, loss of conformational entropy, and binding pocket adaptation.

4D-QSAR is an extension of 3D-QSAR that treats each molecule as an ensemble of different conformations, orientations, tautomers, stereoisomers, and protonation states. The fourth dimension in 4D-QSAR refers to the ensemble sampling of spatial features of each molecule. A receptor-independent (RI) 4D-QSAR method was proposed by Hopfinger *et al* [337]. This method begins by placing all molecules into a grid and assigning interaction pharmacophore elements to each atom in the molecule (polar, nonpolar, hydrogen bond donor, etc.). Molecular dynamic simulations are used to generate a Boltzmann weighted conformational ensemble of each molecule within the grid. Trial alignments are performed within the grid across the different molecules, and descriptors are defined based on occupancy frequencies within each of these alignments. These descriptors are called grid cell occupancy descriptors (GCODs). A conformational ensemble of each compound is used to generate the GCODs rather than a single conformation.

5D-QSAR has been developed to account for local changes in the binding site that contribute to an induced fit model of ligand binding. In a method developed by Vedani and Dobler [338], induced fit is simulated by mapping a “mean envelope” for all ligands in a training set on to an “inner envelope” for each individual molecule. Their method involves several protocols for evaluating induced-fit models including a linear scale based on the adaptation of topology, adaptations based on property fields, energy minimization, and lipophilicity potential. Using this information, the energetic cost for adaptation of the ligand to the binding site geometry is calculated.

Receptor-Dependent 3D/4D-QSAR

Although QSAR methods are especially useful when structural information regarding target binding site is not available, more recent QSAR methods that specifically include such information may be used when possible. One method, known as free energy force field (FEFF) 3D-QSAR trains a ligand-receptor force field QSAR model that describes all thermodynamic contributions for binding [339]. A 4D-QSAR version of FEFF has also been developed to apply this method to the RI-4D-QSAR methods described above [339]. Structurally, the analysis is focused solely on the site of interaction between the ligand and target, and all atoms of interest are assigned partial charges. Molecular dynamic simulations are applied to these structures to generate a conformational ensemble following energy minimization. This approach avoids any alignment issues present in the RI-4D-QSAR method, since the binding site constrains the three-dimensional orientations of the ligands. The conformation ensembles of receptor-ligand complexes generated are placed in a similar grid-cell lattice as used in RI-4D-QSAR, and occupancy profiles are calculated to generate receptor-dependent RD-4D-QSAR models. When tested alongside RI-4D-QSAR against a set of glucose analogue inhibitors of glycogen phosphorylase, predictability of RD-4D-QSAR models outperformed those of RI-4D-QSAR [339].

Linear regression and related methods

Linear QSAR models may be generated using multivariable linear regression analysis (MLR), principal component analysis (PCA), or partial least square analysis (PLS) [249]. MLR computes biological activity as a weighted sum of descriptors or features. The method requires typically 4 or 5 data points for every descriptor used. PCA increases the efficiency of MLR by extracting information from multiple variables into a smaller number of uncorrelated variables. Analysis of results is however not always straightforward [340, 341]. It can be applied with smaller sets of compounds than MLR. PLS combines MLR and PCA and extracts the dependent variable (biological activity) into new components to optimize correlations [342]. PCA or PLS are commonly used for developing models for the molecular interaction field algorithm CoMFA and CoMSIA [249]. A major advantage to these models is that they can be rapidly trained with the tools of linear algebra. The major drawback, however, is that chemical structure often correlates with biological activity in a non-linear fashion.

Nonlinear models using machine learning algorithms

Artificial Neural Networks (ANNs) are one of the most popular nonlinear regression models applied to QSAR-based drug discovery [343]. These models belong to the class of self-organizing algorithms in which a neural network learns the relationship between descriptors and biological activity through iterative cycles of prediction and improvement [249]. A major concern with neural networks is their sensitivity to overtraining, resulting in excellent performance within the training set but reduced ability to assess novel compounds. During the iterative learning process, therefore, ANN performance is commonly measured against an “independent” set of compounds not used to train the model.

Support Vector Machine (SVM) is a kernel-based supervised learning method that was introduced by Vapnik and Lerner [344, 345]. It is based on statistical learning theory and the Vapnik-Chervonenkis dimension [346, 347] and seeks to divide sets of patterns (molecules described with descriptors) based on their classification (biological function). Once this separation is performed on a training dataset, novel patterns can be classified based on which side of the boundary they fall. The simplest form of separation can be imagined as a straight line down the center of a graph with the two classes clustered in opposite corners of the graph. Because two classes can be separated by many potential lines, SVM, a maximal margin classifier, defines the hyperplane with the widest margin between these two classes. The patterns (compounds) that line the closest border of each class are known as support vectors. They define the two hyperplanes separated by that margin and are used to predict classes for novel unclassified patterns. All patterns that lie further from these boundaries are not support vectors and have no influence on the classification of novel patterns. Hyperplanes defined by the lowest number of support vectors are preferred. The solution is a parallel decision boundary that lies equidistant from the two hyperplanes defined by their respective support vectors [348-350].

Ideally, the margin between hyperplanes contains no patterns (molecules). However, to account for noise within datasets and other issues that prevent a linear solution from being reached, a soft-margin classifier is used that allows for misclassification of some data and the existence of patterns within the margin between hyperplanes. In this approach, a penalization constant can be adjusted, with higher values stressing classification accuracy and lower values providing more flexibility.

SVM was initially designed for datasets that could be separated linearly. However, especially in CADD application, this is not always possible. Therefore, SVM incorporates a high-dimensional space in which linear classification becomes possible. This involves the preprocessing of input data using feature functions where the input variables are mapped into a Hilbert space of finite or infinite dimension [348]. This strategy, however, must be offset by the fact that higher dimensional space creates more computational burden and contributes to overfitting [351]. SVM utilizes kernel functions to ease the computational demand imposed by the existence of higher dimensional data. These special nonlinear functions combine the feature functions in a way that avoids explicit transformation and preprocessing using feature functions [348]. In other words, the higher dimensional space that allows for linear separation does not need to be dealt with directly.

Several methods of SVM optimization have been considered. SVM parameter optimization is accomplished by solving the quadratic programming problem with a termination condition called the Klarush-Kuhn-Tucker condition that defines when parameters are at their minima. This can be computationally demanding and difficult to implement. Therefore, decompositional methods have been used to discard all zero parameters [352]. The sequential minimization optimization algorithm is a commonly used alternative introduced by Platt [353]. This method breaks the overall quadratic programming problem into subproblems and solves the smallest possible optimization problem at every stop involving only two parameters. One problem with sequential minimization optimization, however, is that it can result in selection of support vectors that include more than those necessary for the optimal model. Researchers have found that identical solutions can be achieved even after several of these support vectors have been removed [354]. Because the time needed to predict a pattern classification with an SVM model is dependent on the number of support vectors, it is beneficial to eliminate unnecessary or redundant support vectors. Zhan and Shen describe a four step method for removing unnecessary support vectors [354]. Once the SVM has completed training, the support vectors that contribute to the most curvature along the hyper-surface are removed. The SVM model is then retrained and the hyper-surface is further approximated with a subset of support vectors.

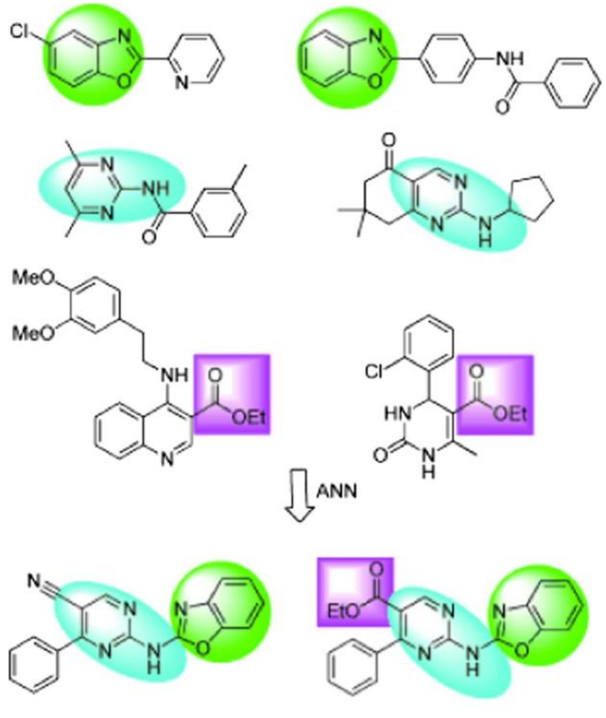
Decision tree learning is a supervised learning algorithm that works by iteratively grouping the training dataset into smaller and more specific groups. The resulting classification resembles a tree in which each feature is broken into different values and each of these values is subsequently divided based on values of a different feature. The order in which features are divided is usually based on an information gain (difference between information before and after the branching) parameter with the highest valued features appearing first [355, 356]. Various methods are used to sort the features, with the overall goal being the smallest possible decision tree providing the best performance. C4.5 is a widely used DT algorithm that calculates information gain based on information entropy [357, 358]. The information entropy of a given classification that can divide the dataset into two classes is calculated based on the number of compounds in either class. The information entropy of the system when dividing the dataset into two subsets using a specific feature is calculated based on the number of compounds from each class in either of the feature subsets. Finally, the information gain for that specific feature is calculated as the difference between the information entropy of the classification and the information entropy of the system.

Once the decision tree has been optimized for the training set, new compounds can be classified by applying their descriptors to the decision tree and activities can be predicted based on which subset they fall into and the activities of the training compounds that are contained in that subset.

Quantitative Structure-Activity Relationship applications in computer-aided drug design

QSAR has been used to screen for novel therapeutics in the same way both pharmacophore models and fingerprint similarity methods have been applied to virtual libraries. Casanola-Martin *et al* used Dragon (Talete S.R.L., Italy) software to define descriptors for tyrosinase inhibitors. Tyrosinase is a copper-containing enzyme that catalyzes two reactions in the melanin biosynthesis pathway [359, 360]. Altered melanin synthesis is found in multiple disease states including hyperpigmentation, melisma, and age spots. Additionally, this protein has been implicated in dopamine neurotoxicity in Parkinson's disease [361]. Descriptors were generated using a highly variable training set of 245 active tyrosinase inhibitors and 408 inactive molecules. These descriptors include constitutional, topological, BCUT, Galvez, topological charge, 2D autocorrelations, and empirical properties and descriptors. Seven models were created using linear discriminant analysis. *In vitro* testing revealed their most potent inhibitor with an IC_{50} of 1.72 µM. This presents a more potent inhibition of tyrosinase than the current reference drug L-mimosine (IC_{50} = 3.68 µM) [362].

Mueller *et al* used ANN QSAR models to identify novel positive and negative allosteric modulators of mGlu5. This receptor has been implicated in neurologic disorders including anxiety, Parkinson's disease, and schizophrenia [363, 364]. For the identification of positive allosteric modulators (PAMs), they first performed a traditional high throughput screen of approximately 144,000 compounds. This screen yielded a total of 1356 hits, a hit rate of 0.94%. The dataset from this HTS was then used to develop a QSAR model that could be used in a virtual screen. To generate the QSAR model, a set of 1252 different descriptors across 35 categories were calculated using the ADRIANA (Molecular Networks GmbH, Erlangen, Germany) software package. The descriptors included scalar, 2D, and 3D descriptor categories. The authors iteratively removed the least-sensitive descriptors to create the optimal set. This final set included 276 different descriptors, including scalar descriptors such as molecular weight up to 3D descriptors, including the radial distribution function weighted by lone-pair electronegativity and π electronegativity. A virtual screen was performed against approximately 450,000 commercially available compounds in the ChemBridge database. Eight hundred twenty-four compounds were tested experimentally for the potentiation of mGlu5 signaling. Of these compounds, 232 were confirmed as potentiators or partial agonists. This hit rate of 28.2% was approximately 30 times greater than that of the original HTS, and the virtual screen took approximately 1 hour to complete once the model had been optimized (figure 1.9) [365].



13 IC_{50} = 75 nM

14 IC_{50} = 124 nM

Figure 1.9 QSAR-based virtual screening of mGlu5 negative allosteric modulators yields lead compounds that contain substructure combinations taken across several known actives used for model generation. Source: [365]

In a separate study, Mueller *et al* [366] used a similar approach to identify negative allosteric modulators for mGlu5. Rodriguez *et al* previously performed a traditional HTS screen of 160,000 compounds for allosteric modulators of mGlu5 and found 624 antagonists [367]. The QSAR model was used to virtually screen over 700,000 commercially available compounds in the ChemDiv Discovery database. Hits were filtered for drug-like properties, and fingerprint techniques were used to remove hits that were highly similar to known actives to identify new chemotypes. Seven hundred forty-nine compounds were tested *in vitro*, and 27 compounds were found to modulate mGlu5 signaling. This hit rate of 3.6% was a significant increase over the 0.2% hit rate of the traditional HTS screen. The most potent of the compounds showed *in vitro* IC_{50} values of 75 and 124 nM, respectively, and contained a previously unidentified scaffold. After analog synthesis and stability optimization, the experimenters tested the effect of their best lead *in vivo* against two behaviors known to involve mGlu5: operant sensation seeking behavior [368] and the burying of foreign objects in deep bedding [369]. Both behaviors were found to be inhibited given intraperitoneal administration of their lead analogue.

QSAR has also been applied to *de novo* drug design techniques when structural information regarding the target is unknown. Descriptor and model generation is used to score the *de novo* generated molecules in place of other structure-based scoring techniques such as docking. Most commonly, compound generation involves iterative algorithms in which structures are repeatedly modified and their biological activities are estimated using QSAR models. In the simplest case, modifications can be achieved by randomly swapping parts of the structure such as functional groups. Ligand-based *de novo* drug design, however, is less practiced than structure-based *de novo* design because of the inherent challenges of accurately evaluating a new molecule in the absence of the receptor structure. To address the challenge of scoring the newly generated molecules, similarity based methods have been applied in addition to QSAR models [370].

Feher *et al* used five selective norepinephrine reuptake inhibitors as a training set to generate 2200 molecules using a combination of structural similarity, 2D pharmacophore similarity, and other properties to drive the evolution [371]. One of the top scoring compounds was found to be highly active and has been selected as a lead compound in a project at Neurocrine [371].

Golla *et al* applied QSAR-based methods to the design of novel chemical penetration enhancers (CPEs) to be used in transdermal drug delivery [372]. This group used a genetic algorithm to design novel CPEs. In this paradigm, new molecules are generated based on crossover and mutation operations randomly applied to candidates. All generated molecules are scored based on the QSAR model and predicted property values, and the highest scoring molecules are retained for new rounds of evolution. Two hundred seventy-two CPEs were used to generate the QSAR model and provide seed molecules for the genetic algorithm. The QSAR model was created using sequential regression analysis and heuristic analysis using CODESSA and contained a final set of 40 descriptors that optimally predicted properties, including skin penetration coefficient, logP, melting point, skin sensitization, and irritation. The top scoring molecules were validated experimentally for permeation and toxicity using Franz Cell with porcine skin and HPLC analysis as well as toxicity effects on human foreskin fibroblasts and porcine abdominal skin. The study resulted in the identification of 18 novel CPEs, four of which showed minimal or no toxic effects [372].

Hoeglund *et al* used QSAR modeling combined with synthetic optimization in a follow-up to their most potent hit from a 2008 *in silico* screen for inhibitors of autotaxin. Autotaxin is an autocrine motility factor and has been linked to cancer progression, multiple sclerosis, obesity, diabetes, Alzheimer's Disease, and chronic pain through the production of LPA [373-378]. Analogues of the lead compound were tested and 4 of the 30 exhibited IC_{50} less than or equal to the lead. The most potent compound showed threefold higher affinity for autotaxin than the lead, whereas another compound showed twofold higher affinity [379].

CoMFA and CoMSIA 3D-QSAR methods have also been used to predict novel therapeutic compounds for a variety of disease targets. Ke *et al* [380] generated CoMFA and CoMSIA models using 66 previously discovered pyrazole- and furanopyrimidine-based Aurora Kinase inhibitors [381-383]. Aurora kinase A is a serine/threonine kinase involved in mitosis [384] that has been shown to be involved in various forms of cancer [385, 386]. Using the model that showed the best predictive performance, the group synthesized a novel compound (compound 67). This compound was tested *in vitro* and displayed an IC_{50} of 25 nM against Aurora kinase A. Additionally, compound 67 displayed antiproliferative activity with an IC_{50} of 23 nM against the HCT-116 colon cancer cell line.

Chai *et al* [387] used 26 previously identified anti-Hepatitis B (HBV) compounds [388, 389] to generate CoMFA models based on steric and electrostatic fields and CoMSIA models based on steric, electrostatic, hydrophobic, and H-bond acceptor fields. Three compounds were designed using these models and subsequently tested against replication of HBV DNA in HBV-infected 2.2.15 cells. The most potent compound displayed an IC_{50} of 3.1 µM, whereas the other two showed IC_{50} values of 5.1 µM and 3.3 µM. These compounds were comparatively more potent than the control lamivudine which displays an IC_{50} of 994 µM.

Jiao *et al* [390] generated CoMFA models using 38 styrylquinoline derivatives in an effort to understand and design potential HIV integrase inhibitors. Their model suggested that a bulky group near the carboxyl group at C-7 in the quinolone ring may confer increased inhibition. Additionally, the presence of an H-bonding donor is favorable near the C-7 atom. Based on these predictions, they designed several compounds that were tested against purified HIV Integrase to determine inhibitory activity on the strand transfer reaction of integrase. Four of these compounds showed higher inhibitory activity than their positive control Baicalein (Sigma-Aldrich, St. Louis, MO).

Over the past several decades, over 18,000 QSAR models have been reported for a variety of targets with a variety of descriptors. C-QSAR was used to generate a comprehensive database of QSAR models [391]. This collection has provided not only access to models for novel applications, but allows the analysis of QSAR models to identify challenges in the field. Kim examined the C-QSAR database for outlier patterns, i.e., compounds that showed poor prediction when the average prediction for the model was good. They found that over the 47 QSAR models examined, the number of compounds scoring as outliers ranged from 3 to 36%. Twenty-six of the 47 datasets showed 20% or more compound outliers [392]. They presented several theories as to why QSAR models are so sensitive to the generation of outliers. One possibility came from analysis of the RCSB protein databank where they discovered examples where related analogs were shown to bind in very different poses. Another explanation offered was protein flexibility, leading to multiple binding modes and or binding sites on the same protein. These different binding modes/sites may reflect different structure-activity relationships for molecules within a given dataset. Analogous compounds that do not share the same binding mode, therefore, present unique challenges to the classification of ligands [392].

1.4.4 Selection of optimal descriptors/features

Hristozov *et al* analyzed the performance of different descriptors across a range of benchmarking datasets and found that the performance of a particular descriptor was often dependent on the activity class. It was found that topological autocorrelation usually offers the best dimensionality/performance ratio. The fusion of the ranked lists obtained with RDF codes and 2D descriptor improved results because RDF codes, while giving similar results, covered different parts of the activity spaces under investigation [34]. This suggests that it is not possible to select an optimal set of descriptors independent of the problem; a custom-optimized descriptor set is needed for optimal performance of LB-CADD.

Excessive numbers of descriptors or features can add noise to a model, reducing its predictive power. Feature selection techniques remove unnecessary features to minimize the number of degrees of freedom of the model. Thus, the ratio of data points versus degrees of freedom increases, leading to models of increased predictive power. Techniques that have proven successful in QSAR modeling include selecting features by measures such as information gain [393] and F–score, sequential feature forward selection or feature backward elimination [394], genetic algorithm [395, 396], swarm optimization [395], and input sensitivity analysis [366].

Information gain measures the change of information entropy from the data distribution of two classes (active and inactive compounds) of one feature compared with the entropy of the feature overall. Thus, discriminatory power of the individual feature increases with information gain. An *F*-score is calculated that considers the mean and standard deviation of each feature across data classes. The higher the *F*-score value, the greater discriminatory power of that feature. Selecting features by individual benchmarks has the disadvantage that correlation between features is ignored. For example, let us assume a feature has a high information gain. However, if a second feature highly correlated is already part of the model, no improved model will result from adding the feature. More complex feature selection schemes address this limitation.

Sequential feature forward selection is a deterministic, greedy search algorithm. In each round, the best feature set from the previous round *N* appends a single feature from the pool of M remaining features and trains the M models using the *N* + 1 features. The best performing feature set from this round then advances to the next round. This continues until all features are used in a final feature set. The best performing model over all iterations is then chosen as the best feature set. This process is time consuming and not guaranteed to yield the optimal feature set; the single best performing feature will always be part of the model. However, there is no guarantee that it is needed. Feature backward elimination inverts the process starting from a model trained from all features, eliminating one after the other. Although the process is more robust in terms of identifying the optimal model, it also requires substantial computer time. Therefore, alternative approaches have been explored to optimize feature sets.

Genetic algorithms mimic the process of evolution to create an efficient search heuristic. This method uses a population of individuals (distinct feature sets) to encode candidate solutions. The initial individuals can be generated randomly. In each iteration, or generation, the fitness of each individual is evaluated, i.e., the predictive power of the derived LB-CADD model. This fitness function is the performance metric of a model trained using that individual as the feature set. Individuals are then selected based on the fitness and undergo recombination and/or mutation to form the next generation. The algorithm continues until a desired fitness score is achieved or a set number of generations have been completed.

Swarm optimization algorithms, such as ant colony optimization [397], particle swarm optimization, and artificial bee colony optimization [398], are optimization techniques based on the organized behavior of social animals such as birds. The algorithm iteratively searches for a best solution by moving individuals around the search space guided by both the local best solution as well as the best solutions found so far in the entire population. The best overall solution is constantly updated, letting the swarm converge towards the optimal solutions.

Input sensitivity analysis seeks to combine speed of individual benchmark values with accuracy of methods that take correlation into account. First, a model is constructed using all features. Next, the influence of each feature on the model output is determined: Each feature x_i is perturbed, and the change in output y is computed. This procedure numerically estimates the partial derivative of the output with respect to each input, a measure that is effective in selecting optimal descriptor sets [366].

1.4.5 Pharmacophore mapping

In 1998, the IUPAC formally defined a pharmacophore as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” [399]. In terms of drug activity, it is the spatial arrangement of functional groups that a compound or drug must contain to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the target, as well as information regarding the type of noncovalent interactions and interatomic distances between these functional groups/interactions. This arrangement can be derived either in a structure-based manner by mapping the sites of contact between a ligand and binding site or in a ligand-based approach.

To generate a ligand-based pharmacophore, multiple active compounds are overlaid in such a way that a maximum number of chemical features overlap geometrically [400]. This can involve rigid 2D or 3D structural representations or, in more precise applications, incorporate molecular flexibility to determine overlapping sites. This conformational flexibility can be incorporated by precomputing the conformational space of each ligand and creating a general-purpose conformational model or conformations can be explored by changing molecule coordinates as needed by the alignment algorithm [400]. For example, one popular pharmacophore-generation software package, Catalyst (Accelrys, Inc., San Diego, CA), uses the “polling” algorithm [401] to generate approximately 250 conformers that it uses in its pharmacophore generation algorithm [249].

Superimposing active compounds to create a pharmacophore

Molecules are commonly aligned through either a point-based or property-based technique. The point-based technique involves superposing pairs of points (atoms or chemical features) by minimizing Euclidean distances. These alignment methods typically use a root-mean-square distance (RMSD) to maximize overlap [402]. Property-based alignment techniques, on the other hand, use molecular field descriptors to generate alignments [400]. These fields define 3D grids around compounds and calculate the interaction energy for a specific probe at each point. The distribution of interaction energies is represented by Gaussian functions, and the degree of overlap between Gaussian functions of two aligned compounds is used as the objective scoring function to maximize alignment [402]. One popular field generation method for property-based alignments is GRID [304].

Molecular flexibility is always an important consideration when aligning compounds of interest and several approaches are used to efficiently sample conformational space. These approaches include rigid, flexible, and semiflexible methods. Rigid methods require knowledge of the active conformation of known ligands and align only these active conformations. This is only applicable, however, when the active conformation is known with confidence. Semiflexible methods overlay pregenerated ensembles of static conformations and flexible methods, being the most computationally expensive, perform conformational search during the alignment process, often using molecular dynamics or random sampling of rotatable bonds. Because the conformational space can increase substantially with an increase in the number of rotatable bonds, strategies are often used to limit the exploration of conformational space through the use of reference geometry (often an active ligand with low flexibility). This method is known as the active analog approach [403].

Pharmacophore feature extraction

A pharmacophore feature map is carefully constructed so as to balance generalizability with specificity. A general definition might categorize all functional groups having similar physiochemical properties (i.e., similar hydrogen-bonding behavior, ionizability) into one group, whereas specific feature definitions may include specific atom types at specific locations. More general feature definitions increase the population of compounds that match the pharmacophore. They allow the identification of novel scaffolds but also increase the ratio of false-positives. The level of feature definition generalizability is usually determined by the algorithm used to extract feature maps and through user-specified parameters. The most common features used to define pharmacophore maps are hydrogen bond acceptors and donors, acidic and basic groups, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties [249]. Features are commonly implemented as spheres with a certain tolerance radius for pharmacophore matching [400].

Pharmacophore Algorithms and Software Packages

The most common software packages employed for ligand-based pharmacophore generation include Phase [404], MOE [325], Catalyst [405, 406], LigandScout [407], DISCO [408], and GASP [409]. These packages use different approaches to molecular alignment, flexibility, and feature extraction. Catalyst approaches alignment and feature extraction by identifying common chemical features arranged in certain positions in three-dimensional space. These chemical features focus on those expected to be important for interaction between ligand and protein and include hydrophobic regions, hydrogen-bond donors, hydrogen-bond acceptors, positive ionizable, and negative ionizable regions. Chemical groups that participate in the same type of interaction are treated as identical. Catalyst contains two algorithms that can be used for pharmacophore construction. HipHop is the simpler of the two algorithms and looks for common 3D arrangements of features only for compounds with a threshold activity against the target. It begins with best alignment of only two features (scored by RMS deviations) and continues expanding the model to include more features until no further improvements are possible. This method is only capable of producing a qualitative distinction between active and inactive predictions. HypoGen, on the other hand, employs biological assay data such as IC_{50} values for active compounds as well as a set of inactive compounds. Initial pharmacophore construction in HypoGen is identical to HipHop but includes additional algorithms that incorporate inactive compounds and experimental values. These algorithms compare the best pharmacophore from the “HipHop” stage with the inactive compounds and features common to the inactive set are removed. Finally, HypoGen performs an optimization routine that attempts to improve the predictive power of the pharmacophore by making adjustments and scoring the accuracy in predicting the specific experimental activities [405, 410]. This results in models that are capable of quantitative predictions that can predict specific levels of activity. Ten different models are created following a simulated annealing optimization [411]. Both

Catalyst methods incorporate molecular flexibility by storing compounds as multiple conformations per molecule. The Poling algorithm published by Smellie *et al* [401] is employed to increase the conformational variation within the set of conformations per molecule. This allows Catalyst to cover the greatest extent of conformational space while keeping the number of conformations at a minimum.

Phase approaches alignment and feature extraction using a tree-based partitioning algorithm and an RMS deviation-based scoring function that considers the volume of heavy atom overlap. It incorporates molecular flexibility through a preparation step where conformational space is sampled using a Monte Carlo or torsional search [402].

DISCO regards compounds as sets of interpoint distances between heavy atoms containing features of interest. Alignments are based on the spatial orientation of common points among all active compounds. DISCO considers multiple conformations that have been prespecified by the user during the alignments and uses a clique-detection algorithm for scoring alignments [410].

GASP uses a genetic algorithm with iterative generations of the best models for pharmacophore construction [409]. Flexibility is handled during the alignment process through random rotations and translations. Conformations are optimized by fitting them to similarity constraints and weighing the conformations that fit these constraints more than conformations that do not [411].

Different software packages can produce different results for the same datasets, and their strengths and weaknesses should be considered prior to any application. For example, Catalyst only permits a single bonding feature per heavy atom, whereas LigandScout allows a hydrogen-bond donor or acceptor to be involved in more than one hydrogen-bonding interaction [400]. MOE, on the other hand, allows a more customizable approach to hydrogen-bonding features. Lipophilic areas are generally represented as spheres located on hydrophobic atom chains, branches, or groups in a similar manner across software packages but with slight nuances. Although subtle, these differences have important consequences on prediction models. Additionally, software packages that do not attach a hydrophobic feature to an aromatic ring are unable to predict that an aromatic group may be positioned in a lipophilic binding pocket [400]. The level of customizability also differs across pharmacophore software packages and can influence predictions. Catalyst allows the specification of one or more chemical groups that satisfy a particular feature, whereas Phase allows not only matching chemical groups but also a list of exclusions for a given feature. MOE offers a level of customization that allows the user to implement entirely novel pharmacophore schemes as well as modification of existing schemes. However, this requires additional levels of expertise to program [400]. For a comprehensive analysis of the differences between commercial pharmacophore software packages, please see the 2007 review by Wolber *et al* [400] and a 2002 comparison of Catalyst, DISCO, and GASP by Patel *et al* [412].

Ligand-based pharmacophore methods have been used for the discovery of novel compounds across a variety of targets. New compounds can have activity in the micromolar and nanomolar range and reflect proof of concept with *in vivo* disease models. Al-Sha'er and Taha used a diverse set of 83 known Hsp90-α inhibitors and the HypoGen module of Catalyst to generate a pharmacophore model. Hsp90-α is a molecular chaperone that is involved in protein folding, stability, and function [413]. By interacting with many oncogenic proteins, it has been shown to be a valid anticancer drug target [414, 415]. The pharmacophore model was used to screen the NCI list of compounds (238,000) using the "Best Flexible" search option. The top 100 hits were evaluated *in vitro* and their most potent compound had an IC_{50} of 25 nM [416].

Schuster *et al* used three steroidal inhibitors and two non-steroidal inhibitors of 17β-HSD3 and Catalyst to create a pharmacophore model that was used to screen for novel 17β-HSD3 inhibitors. Hydroxysteroid dehydrogenases (HSD3) catalyze the oxidoreduction of alcohols or carbonyls and the final step in male and female sex hormone biosynthesis. Therefore, these enzymes are suggested therapeutic targets for control of estrogen- and androgen-dependent diseases such as breast and prostate cancer, acne, and hair loss [417]. Eight commercial databases were screened, and 15 top scoring hits were tested *in vitro* at 2 μM. Five were verified to be inhibitors of 17β-HSD3 with the most potent compound able to inhibit 17β-HSD3 by 67.1% at 2 μM [418].

Noha *et al* developed 5-point pharmacophore models using the HipHop algorithm of Catalyst based on a training set of compounds with IC_{50} < 100 nM against IKK-β as potential anti-inflammatory and chemosensitizing agents. The authors used 128 active and 44 inactive compounds to develop a pharmacophore model [419]. Their model was further refined with exclusion volume spheres and shape constraints to improve the scoring of compounds in their virtual high-throughput screen against the National Cancer Institute molecular database. Ten compounds were selected and the most potent compound (NSC719177, $C_{26}H_{31}NO_4$) showed inhibitory activity against IKK-β in a cell free *in vitro* assay with IC_{50} of 6.95 μM. Additionally, this compound inhibited NF-κB activation induced by TNF-α in HEK293 cells with an IC_{50} of 5.85 μM [419].

Chiang *et al* used the HypoGen module of Catalyst to generate four-feature pharmacophore models based on an indole series of 21 compounds that showed antiproliferative activity through the inhibition of tubulin polymerization/microtubule depolymerization. Disruption of microtubules during the mitotic phase of the cell cycle can induce cell-cycle arrest and apoptosis [420]. Therefore, inhibitors of tubulin polymerization are useful cancer treatments. One hundred thirty thousand compounds of the ChemDiv database and an in-house compound collection were screened, and the top 142 hits were tested *in vitro*. Four novel compounds were discovered with antiproliferative activity. The most potent compound displayed antiproliferative activity in human cancer KB cells with an IC_{50} of 187 nM. This compound also inhibited the proliferation of other cancer cell types, including MCF-7, NCI-H460, and SF-268 and demonstrated anticancer effects in a histoculture system. *In vitro* assays revealed that this compound inhibited tubulin polymerization with an IC_{50} of 4.4 µM [421].

Doddareddy *et al* generated a pharmacophore model containing three hydrophobic regions, one positive ionizable center, and two hydrogen bond acceptor groups for the identification of novel selective T-type calcium channel blockers. The most potent hit showed an IC_{50} of 100 nM [422, 423]. T-type calcium channels are involved in rhythmical firing patterns in the CNS and present therapeutic targets for the treatment of epilepsy and neuropathic pain [326].

Lanier generated pharmacophores containing five feature points using Catalyst and CombiCode (Deltagen Research Laboratories, San Diego CA) software and an exclusion sphere generated in MOE based on a training set of 100 active and 1000 inactive compounds. This model was used to guide and evaluate variations of a core molecule, leading them to a gonadotropin releasing hormone GnRH receptor antagonist with receptor affinity below 10 nM [424]. GnRH is involved in the regulatory pathways of follicle stimulating hormone and luteinizing hormone. It is a target for disease therapeutics including endometriosis, uterine fibroids, and prostate cancer [425, 426].

Roche *et al* used known H3 antagonists to generate a pharmacophore model with four features including a distal positive charge, an electron-rich position, a central aromatic ring, and either a second basic amine or another aromatic [427]. Histamine is a central modulator in the central and peripheral nervous systems through four receptors (H1-H4) [428]. H3 is a presynaptic autoreceptor that modulates production and release of histamine and other neurotransmitters [429]. H3 antagonists have been studied in Alzheimer's disease, attention deficit disorder, and schizophrenia [430]. Additionally, it has been suggested to be involved in appetite and obesity [431] .This model was used in a *de novo* approach with the Skelgen software [432] to generate novel compounds from fragment libraries that match the pharmacophoric restraints. They found a series of four compounds with high potency and selectivity for H3. Their most potent compound showed inverse agonist activity with an EC_{50} of 200 pM in a GTPγS functional assay and a binding affinity K_i towards H3 of 9.8 nM [427].

Chao *et al* used pharmacophore-based design to take advantage of the therapeutic benefits of Indole-3-carbinol (I3C) in the treatment of cancer. I3C is known to suppress proliferation and induce apoptosis of various cancer cells through the inhibition of Akt activation [433, 434]. I3C, however, has a poor metabolic profile and low potency, likely due to the fact that its therapeutic behavior comes from only four of its metabolites. By overlaying these low energy conformers of these four metabolites, Chao *et al* was able to identify similar N-N' distances and overlapping indole rings (figure 1.10) [435]. This led them to design SR13650, which showed an IC_{50} of 80 nM. Tumor xenograft studies using MCF-7 cells revealed antitumor effects at 10 mg/kg for 30 days. Computational analysis was also applied to increase the bioavailability, and three compounds showed 45-60% tumor growth inhibition *in vivo* compared to the 26% growth inhibition of SR13650. SR13668 was the most potent compound and also displayed antitumor effects in other xenograft models. *In vitro*, SR13668 was shown to inhibit Akt activation by blocking growth-factor stimulated phosphorylation and showed favorable toxicological profiles [435]. This drug is currently in phase 0 trials for the treatment of cancer (figure 1.10) [436].

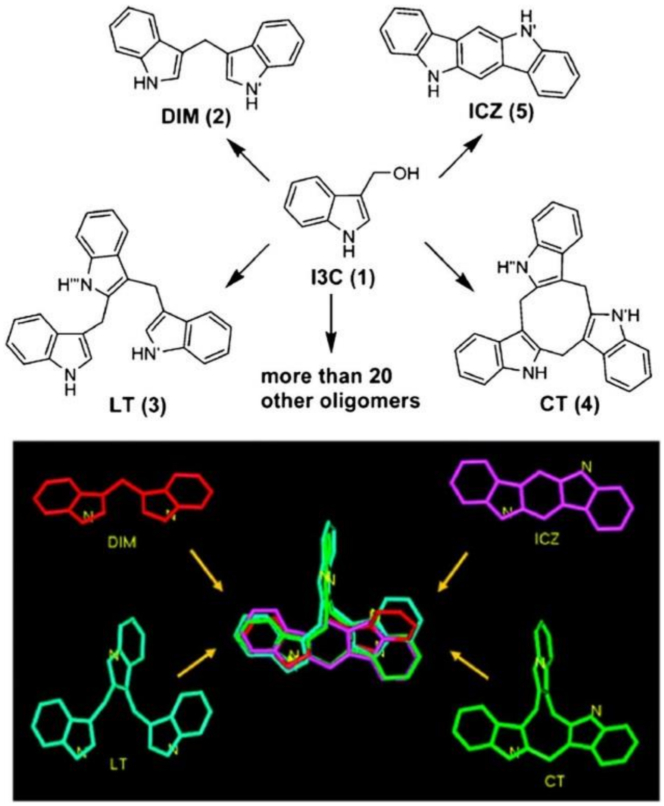


Figure 1.10 SR13668, an anticancer therapeutic was discovered using ligand-based pharmacophore screening based on active components of indole-3-carbinol. Source: [435].

Dayam *et al* [437] used pharmacophore modeling in an effort to identify novel HIV-1 integrase (IN) inhibitors. IN is the third viral enzyme in HIV and is responsible for integration of viral DNA into host cell chromosomes through 3’-processing and strand transfer [438, 439]. This model was created with the HipHop algorithm within Catalyst and was based on the Quinolone 3-carboxylic acid class of IN inhibitors that show IC_{50} values ranging from 43.5 to 7.2 nM and EC_{50} against HIV-1 replication of 805 to 0.9 nM [440]. The final pharmacophore hypothesis consisted of four features including a negatively ionizable feature, hydrogen-bond acceptor, and two hydrophobic aromatic features (figure 1.11). Three hundred sixty-two thousand two hundred sixty commercially available compounds were screened and 56 selected for *in vitro* evaluation. Eleven of those tested inhibited the IN catalytic activity with an IC_{50} value < 100 µM. Five compounds had an IC_{50} less than 20 µM, and the most potent compound inhibited both the 3’ processing (IC_{50} 14 µM) as well as strand transfer activities (IC_{50} 5 µM) of IN [437]. Mugnaini *et al* created a pharmacophore model using 30 known inhibitors of the 3’-processing step of HIV-1 IN and screened the ASINEX gold database of over 200,000 compounds for inhibitors of IN. Twelve hits were tested *in vitro* and discovered one compound with a novel scaffold and anti-integrase activity with IC_{50} of 164 µM. Further improvement of this compound yielded an analogue with IC_{50} of 12 µM [441].

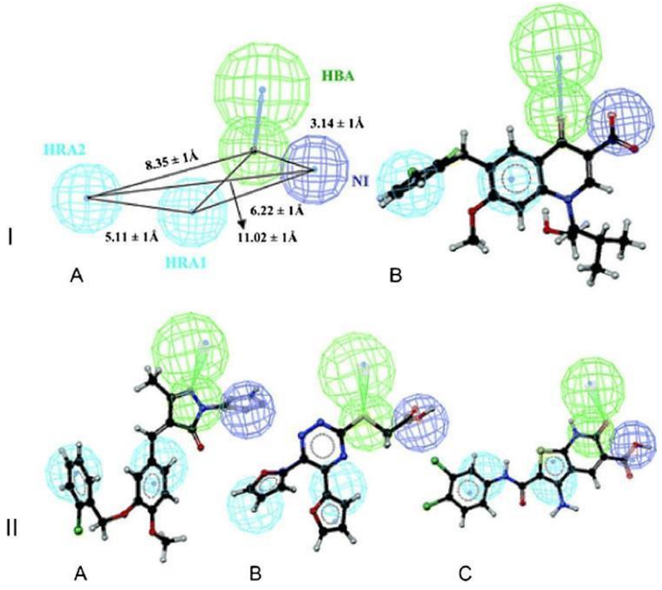


Figure 1.11 HIV-1 Integrase inhibitor pharmacophore I) A) Novel HIV-1 Integrase inhibitor using ligand-based virtual screening with a pharmacophore model of quinolone 3-carboxylic acid IN inhibitors. B) Pharmacophore query generated from the quinolone 3-carboxylic acid IN inhibitors accompanied with an overlay onto a known HIV-1 integrase inhibitor. Features are color-coded, and their 3D arrangement/distances are shown in angstroms. Green sphere represent H-bond acceptor regions, blue spheres represent negatively ionizable regions, and cyan spheres represent hydrophobic aromatic regions. II) Pharmacophore query overlayed with 3 potent hits from the ligand-based virtual screen: compounds 8 (A), 9 (B), and 17 (C). Source: [437].

Noeske *et al* [442] used 2D-pharmacophore-based virtual screening to identify novel mGlu1 antagonists. Antagonism of this receptor has been studied in regards to therapeutic potential in neurodegenerative diseases, anxiety, pain, and schizophrenia [443, 444]. Six reference mGlu1 antagonists were used to construct 2D-pharmacophores with the CATS software package [445]. This software assigns all atoms in a compound as either a hydrogen-bond donor, hydrogen-bond acceptor, positively charged, negatively charged, lipophilic, or non-interest atom type. Then, all compounds of a library are compared with the distances between these different atom types in the reference molecule and similarity scores are calculated to rank molecules that most closely fit this 2D-pharmacophore. Screening the Gold Collection of Asinex Ltd yielded six different hit lists (one for each reference molecule). The top hits were collected from all lists as well as hits that appeared in three or more different lists and 23 compounds were tested experimentally for mGlu1 antagonism. Their most potent compound yielded an IC_{50} of 360 nM and was further optimized to a compound with an IC_{50} of 123 nM.

1.5 Prediction and Optimization of Drug Metabolism and Pharmacokinetics Properties Including Absorption, Distribution, Metabolism, Excretion, and the Potential for Toxicity Properties

In addition to high biological activity and selectivity for the target of interest, drug metabolism and pharmacokinetics (DMPK) properties including absorption, distribution, metabolism, excretion, and the potential for toxicity (ADMET) in humans are critical to the success of any candidate therapeutic. After lead discovery or design, there is considerable attention given to improving the compound's *in vivo* DMPK/ADMET properties without losing its biological activity. It is common to apply some DMPK/ADMET-based restrictions early on in the discovery process to reduce the number of compounds necessary to evaluate, saving time and resources. Therefore, computational techniques extend to predicting this very important aspect of drug design and discovery. Methods used are structure-based to study the interaction of candidate compounds with key proteins involved in DMPK/ADMET and ligand-based to predict of key properties using quantitative structure property relation (QSPR) models.

1.5.1 Compound Library Filters

Computational tools are routinely used to filter large data bases so that compounds predicted to have poor DMPK/ADMET profiles may be avoided. One of the earliest and still the most popular filters to apply to any compound database when performing a vHTS is Lipinski's rule of 5. These rules are: a) molecular weight of **5**00 or less, b) logP coefficient less than **5**, c) **5** or fewer hydrogen-bond donor sites d) 2x**5** or fewer hydrogen-bond accepting sites [446]. The rule set is based on an analysis of 2245 compounds from the World Drug Index that had reached phase II trials or higher. The rules were based on distributions for molecular weight, logP, hydrogen bond donors, and hydrogen bond acceptors for the top percentile of these compounds [446]. This set of rules suggests the necessary properties for good oral bioavailability [447] and reflects the notion that pharmacokinetics, toxicity, and other adverse effects are directly linked to the chemical structure of a drug. Although this criteria is well established and offers a relatively fast and simple way to apply DMPK/ADMET filters before any sort of screening is performed, it is incapable of predicting with any certainty whether a compound will make an appropriate therapeutic. It has been estimated that almost 69% of available compounds in the Available Chemical Directory (ACD) Screening Database (2.4 million compounds) and 55% of the compounds in the ACD (240,000) do not violate this rule of 5 [448]. Accordingly, this rule set has always been intended to be a guide and not necessarily a hard-set filter. It is expected that such a simple rule of thumb will remove lead compounds; for example, many peptidomimetics, transporter substrates, and natural products will violate Lipinski's rule. Approximately 16% of oral drugs violate at least one criterion and 6% fail two or more criteria, and multiple examples exist of highly successful drugs that fail one or more of Lipinksi's criteria including Lipitor and Singulair [449]. At the same time the Lipinski's rule will not, for example, recognize and remove compounds with structural features that give rise to toxicity. It is limited to evaluating oral bioavailability through passive transport only. When used to train models with machine learning, Lipinski's rule failed to provide better than random classification of drugs and nondrugs [450]. Additionally, it is not designed to provide any discrimination beyond a binary pass or fail. Any compound that violates two or more criteria is treated as an equal fail, whereas any compound that does not is treated as an equal pass.

On the basis of its shortcomings, several improvements and replacements have been proposed for the rule of 5. For example, two additional criteria have been suggested that include the number of rotatable bonds being less than or equal to ten and the polar surface area being less than 140 $Å^2$ [451]. Bickerton *et al* [449] introduced the quantitative estimate of drug-likeness that is a score ranging from 0 (all properties unfavorable) to 1 (all properties favorable). This score is taken as a geometric mean of individual desirability functions, each of which corresponds to a different molecular descriptor. These descriptors include molecular weight, logP, hydrogen bond donors and acceptors, rotatable bonds, aromatic rings, and the number of structural alerts [452].

However, the simple application of filters such as these during a lead compound search can be problematic by nature of the limitation of these descriptors and the evolution of lead compound to drug. For example, Hann *et al* found that, on average, over a set of 470 lead-drug pairs, lead compounds had lower molecular weight, lower logP, fewer aromatic rings, and fewer hydrogen-bond acceptors compared with their eventual drugs [453]. Therefore, it can be problematic to apply filters designed around the average properties of drugs to libraries that are intended for the discovery of lead compounds.

Additionally, some of the properties used in these filters can depend on conformation and environment. Kulkarni *et al* [454] state that permeability and hydrophobicity can change depending on the free energy of solvation, interaction of the drug with a phospholipid monolayer, and the drug's flexibility. Vistoli *et al* [455] state that hydrophobicity and hydrogen bonding are both dependent on the dynamic nature of molecules and that chemical information is limited without the use of dynamic descriptors. For a comprehensive review on the concept of drug likeness please see the 2011 review by Ursu *et al* [456].

The same computational tools used to predict activity can be applied to predict a more detailed DMPK/ADMET profile, including solubility, membrane permeability, metabolism, interaction with influx/efflux transporter proteins, interaction with transcription proteins, and different aspects of toxicity. For example, QSAR-based techniques have been especially important in predicting the toxicology profiles for drugs very early on in their development. These tools collect information regarding known toxins such as carcinogens, neurotoxins, and skin irritating agents, and create statistical models that can predict the likelihood that a particular compound will reflect these undesirable properties [457].

1.5.2 Lead improvement: metabolism and distribution

Aside from general filters applied to compound libraries preceding a screen, computational tools can be used to guide hit-to-lead optimization where a compound's metabolic profile is fine tuned. This requires a precise balancing act as the changes necessary to improve a compound's metabolic profile may also significantly reduce its target affinity. During this stage of drug development, efforts are made in changing the compound's structure not only to improve affinity but also to improve its metabolism. Therefore, although computational tools are useful in predicting the effects on target affinity from any proposed changes to the lead structure, they can be used in parallel to predict the affinity and interactions the compound may have with metabolizing enzymes and their regulators [458]. The metabolism of a drug can have significant impacts not only on its bioavailability but also on its half-life and generation of harmful metabolites. When metabolic stability is lowered, a drug can lose its efficacy. Increasing stability can amplify harmful side effects owing to a long half-life. Physiologically, there are two important phases in drug metabolism that have been studied extensively. The phase I reactions include hydrolysis, reduction, and oxidation and are primarily performed by cytochrome p450 enzymes. Phase II reactions are more diverse and include glucuronidation, sulfation, acetylation, methylation, and glutathione conjugation [459]. These reactions accelerate the drug's elimination from the body but can result in toxic products like highly reactive electrophiles or free radicals [458].

Computational tools have been developed to address the phase I metabolism reactions performed by Cytrochrome P450 enzymes, mainly through docking and QSAR procedures to predict the likelihood that a particular compound will bind to a cytochrome P450. At least 57 P450 isoforms exist in the human body, but phase I metabolism is dominated by the isoforms 1A2, 2C9, 2C19, 2D6, and 3A4 [460] and computational methods are routinely directed against these particular P450 isoforms. In addition to the elimination of the drug and generation of metabolites, P450s can also be the source of drug-drug interactions in that one drug can reduce the elimination of another drug by blocking access to metabolizing enzymes or can increase elimination by upregulating expression of those enzymes. For example, in the early development of CCR5 antagonists, experimenters discovered hits that contained functional groups that are common among CYP2D6 inhibitors. By modeling the binding of these ligands to CYP2D6, imidazopyridines were replaced with benzimidazoles so that possible drug-drug interactions arising from inhibition of CYP2D6 were avoided early on [461].

Structure-based methods are the most popular computational tools for predicting the interaction between a compound and P450 enzymes. Binding poses predicted through docking studies may provide further insight into the specific sites of metabolism within the compound. For example, structure-based methods successfully predicted the metabolism of celecoxib and its 13 analogues through CYP2C9 [462, 463]. In addition to some P450 isoforms, x-ray structures of the ligand-binding domain of prenane X receptor (PXR) [464], the transcription regulator of CYP3A4 [465], glutathione-S-transferases [466], and drug transporters such as P-glycoprotein [467] have been determined. Structural information about PXR and drug transporters can be used to predict drug-drug interactions through the induction of CYP3A4 or transport channels.

One of the major challenges in modeling P450 binding is the dynamic nature of the binding site that accommodates a wide variety of ligands. Another challenge with docking studies involving P450 enzymes is the fact that the goal is often fundamentally opposite to that of most docking studies in that weaker binding is usually preferred over stronger binding. Monte Carlo and stochastic simulations of a wide variety of cocrystal structures have allowed development of several dynamic models of P450 binding sites exploring the different orientations amino acid side chains [458]. GOLD, FlexX, DOCK, AutoDock, and the scoring function C-Score are most commonly used for structure-based methods with

P450 predictions [468]. For modeling the catalytic reaction encountered when the ligand binds to the P450 enzyme, *ab initio* calculations using Hartree-Fock or density functional theory have been used [458].

For example, the formation of the hydroquinone metabolite and electrophilic quinonone from remoxipride was calculated using hybrid density functional theory. This information was then used to redesign remoxipride [469]. Density functional theory calculations were used to eliminate the formation of reactive metabolites from a series of tyrosine kinase-2 inhibitors. These calculations correctly predicted the necessary changes that avoided the formation of these harmful metabolites [470]. Park and Harris used DFT on CYP2E1 homology models along with docking and MD to predict the metabolism profiles for seven compounds [471]. Li *et al* used homology modeling and MD to dock ligands into CYP2J2 in an effort to describe the binding characteristics of this enzyme. CYP2J2 is involved in the creation of eicosatrienoic acids from arachidonic acid. They were able to identify key residues that were important for the substrate specificity of CYP2J2. Additionally, they discovered that different ligands, although sharing the same scaffold, show different binding modes [103]. Bazeley *et al* used structural information of CYP2D6 to identify invariant segments and performed conformational sampling with MD. Combining this data with neural-network based feature selection they found that only three out of 20 conformations are relevant for CYP2D6 binding. They also analyzed the docking of 82 compounds and showed that the most important attributes that conferred a compound’s affinity for CYP2D6 was the number of hydrogen-bonding sites, molecular weight, the number of rotatable bonds, AlogP, formal charge, number of aromatic rings, and the number of positive atoms. With these findings, they were able to achieve a prediction accuracy of 85% [472].

In addition to these structural methods, reactivity rules are also used to predict the metabolism of small molecules. Databases such as Accelrys Metabolite [473] contain curated metabolic transformations from the literature. This information can be used to predict the various metabolic transformations that will be produced from an input structure. META [474] is a model of mammalian xenobiotic metabolism that incorporates metabolic data from literature, textbooks, and monographs to define chemical transformation rules called transforms, which can identify and substitute functional groups. These focus on both phase 1 and phase 2 metabolism.

Another method uses electronics and intramolecular sterics to predict sits of CYP3A4 metabolism. This approach focuses on the rate-limiting step of the hydroxylation by CYP3A4, namely the removal of the hydrogen-atom [475]. The model assumes that the susceptibility for removal depends mainly on the electronic environment surrounding the hydrogen. Therefore, the method calculates a hydrogen abstraction energy for each hydrogen atom and this information is used to predict sites of metabolism [476].

SMARTCyp [477] is another rule-based method that determines the reactivity of molecular fragments based on activation energies calculated by quantum mechanical methods. It combines a reactivity descriptor and accessibility descriptor. The reactivity descriptor estimates energy required for P450 metabolism at a given site by looking up fragments in an energy table for each atom. The accessibility descriptor is a calculation that determines the 2D distance from the center of the molecule a given atom is and always ranges between 0.5 and 1.

The activation energy table used for the reactivity descriptor combines 11 previously defined rules for aliphatic, aromatic, and alkene carbon atoms for 50 carbon sites [478] with new data generated by the authors. This produced a collection of 139 transition states that can represent different types of P450 reactions.

Other aspects of a drug's DMPK/ADMET profile that are predicted with computational tools include membrane permeability, which is a large part of bioavailability as well as volume of distribution and penetration of the blood-brain barrier, and blood plasma protein binding, involved in a drug's volume of distribution and effective plasma concentrations. The evolution of predictive models for blood-brain barrier penetration is reviewed in detail by Norinder and Haeberlein [479]. Additionally, the structure of human serum albumin is used to predict plasma protein binding and volume of distribution changes [480].

1.5.3 Prediction of human Ether-a-go-go related gene binding

The human ether-a-go-go related gene (hERG) protein is a voltage-gated potassium channel expressed in the heart and nervous system. The tetramer has six transmembrane spanning regions per protamer and is important for repolarization during the cardiac action potential [481-483]. The delayed rectifier repolarizing current, an outward potassium current comprised of a rapid and slow component, is involved in plateau repolarization and the configuration of the action potential. Alterations in this channel's conductance, especially blockade of the channel, can lead to an altered refractory period and action potential duration [483], often resulting in what is known as drug-induced QT syndrome and a severe cardiac side effect called torsades de points [484]. The QT interval is the period of a cardiac cycle where ventricular repolarization occurs [482] and drug-induced QT syndrome can lead to sudden death [485]. Because of its importance in the proper regulation of cardiac action potential, off-target interactions with hERG have caused several drugs to be removed from the market and/or linked to arrhythmias and sudden death [481]. hERG has been termed an “antitarget” in the pharmaceutical industry [486]. It has been estimated that 2-3% of prescribed medications include some unintended QT elongation [483]. Though most drugs have been shown to inhibit the rapid component of the outward potassium current [487], interaction between drugs and hERG is not completely understood, and high-affinity ligands tend to interact with the inactivated channel with low voltage-dependency, whereas low-affinity ligands tend to interact with the activated state with high voltage-dependent kinetics [488]. However, key residues involved in the interaction between hERG and at least some ligands have been identified. For example, Phe656 and Tyr652 in the channel pore may engage in π-π and cation-π interactions with the ligand. Thr623 and Ser624 are thought to interact with the polar tails of some ligands and some evidence exists of a second binding site [482, 483, 486, 489]. *In vitro* and *in vivo* methods are commonly used to evaluate drug candidates for potential hERG blockade activity, especially patch clamp techniques and radioligand binding assays [490, 491]. However, these methods are difficult to scale to high-throughput candidate evaluation, making the computational approach attractive for this aspect of drug discovery.

SB-CADD and LB-CADD have both been used to develop models to discriminate hERG blockers and non-blockers [492, 493]. SB-CADD techniques have mainly relied on docking with homology models and this method has not been validated with large, highly diverse data sets [494]. LB-CADD-based hERG models have been created using tools including ligand-based pharmacophore [495, 496], CoMFA [497], Bayesian classification with QSAR [498], and 2D fragment based descriptors [499].

Wang *et al* developed discrimination models based on molecular property descriptors and fingerprints [500]. Descriptors were calculated using Discovery Studio molecular simulation package (Accelrys) and included several variations on logP, molecular weight, hydrogen-bonding, the number of rotatable bonds, rings, and aromatic rings, the sum of oxygen and nitrogen atoms, and fractional polar surface area. The fingerprints included SciTegic extended-connectivity fingerprints and Daylight-style path-based fingerprints using the Morgan algorithm [501]. Bayesian classifiers and decision tree methods were used to create models based on these descriptors.

Wang *et al* [500]analyzed the results of their models and found that increased hydrophobicity was correlated with increased hERG binding. Additionally, molecular weight showed a significant, although lesser impact on hERG binding, with molecules having a molecular weight under 250 being less likely to be a hERG blocker. Additionally, analysis of their fingerprints revealed that most hERG-binding fragments contained nitrogen atoms, with four of the top five containing positively charged nitrogen atoms. These top five fragments also contained at least one oxygen atom or a carboxylic acid. Despite these correlations, the authors stressed that no single molecular property can be used to discriminate between hERG blockers and nonblockers.

Obrezanova and Segall [502] used the Gaussian process to build models for hERG inhibition as well as other ADMET properties. The Gaussian process [503, 504] is a nonlinear regression technique that is resistant to overtraining. It uses Bayesian inference to link the descriptors of a molecule with the probability of the molecule falling into a specific class. Eventually, a posterior probability distribution is created that defined which functions best describe the observed data. The mean value over all functions can provide the prediction, whereas the full distribution can provide a measure of uncertainty for each prediction. The hERG inhibitor model was trained on 117 active and 51 inactive compounds evaluated through patch clamp in mammalian cells with descriptors generated in StarDrop's Auto-Modeler [505]. These 2D descriptors were based on SMARTS and included atom type counts, functionality, and molecular properties such as logP, molecular weight, and polar surface areas. Datasets were also clustered using 2D fingerprints and tanimoto similarity.

Nisius and Göller [506] used the Tripos Topomer Search technology [507] to design a modeling approach termed topoHERG. This method screens reference datasets for molecules similar to a query compound and returns pharmacophore and shape-based distances between a query molecule and its neighbors. The dataset contained 115 inactive compounds, 90 moderately active hERG blockers, and 70 highly active hERG blockers. The topomer is defined as a 3D representation of a molecular fragment that is based on 2D topology and a rule set that generates an absolute conformation [508] so that distances between topomers of different molecules in large databases can be calculated. To differentiate between hERG active and inactive neighbors, the inverse of the topomer search distance was multiplied by one if the topomor search neighbor was active and negative one if it was inactive. A molecule was predicted to be an active hERG blocker if its overall sum was greater than zero. A two-stage approach using two optimized models yielded a prediction accuracy of 76-81% [506].

Garg *et al* [487] used a genetic function approximation to generate quantitative structure-toxicity relationship (QSTR) models using 2D descriptors generated using the QSAR+ module of Cerius (Accelrys). These models were trained with 56 hERG blockers and descriptors included electrotopological descriptors that contained information regarding the topological environments for all atoms in the molecule as well as electronic interactions with other atoms in the molecule. To perform genetic function approximation, the authors generated a number of random equations that were randomly selected as pairs. These parent pairs underwent crossover operations to generate new equations, and those that showed improved fitness scores were kept [509]. In parallel, the authors generated a toxicophore (pharmacophore-based toxicity model) using Catalyst's HypoGen that included hydrogen-bonding, hydrophobic, aromatic, and positive ionizable features. Upon analysis of their models, the authors noted that both basic and neutral hERG blockers had highly flexible linkers and various molecular fragments.

1.5.4 Drug Metabolism and Pharmacokinetics/Absorption, Distribution, Metabolism, and Excretion and the Potential for Toxicity Prediction Software Packages and Algorithms

There are currently many models available for predicting absorption, bioavailability, transporter binding, metabolism, volume of distribution, and P450 interactions [510-516]. Comprehensive software packages have been developed such as QikProp which can be used to predict an array of ADMET-related properties such as solubility, membrane permeability, partition coefficients, blood-brain barrier penetration, plasma protein binding, and the formation of metabolites [517]. These predictions mainly come from statistical models such as regression and neural networks that are trained on known ADMET properties for many compounds. The OSIRIS Property Explorer allows scientists to draw chemical structures and predict ADMET profile [35]. The software package MetaSite (Molecular Discovery Ltd, Middlesex UK) is used to predict the site of metabolism using structural information from both the ligand and the enzyme. A probability function is created for the site(s) of metabolism using the free energy of P450-ligand binding and reactivity. This software uses structure-based techniques to identify the relevant amino acids and proposes compound modifications that can optimize its metabolism profile [518]. Ahlstrom *et al* proposed a three-step procedure using MetaSite to identify metabolic sites, *in silico* modification of these sites, and docking of new compounds [462]. These software packages aim at predicting overall ADMET properties with convenient and accessible tools and have shown great benefit in drug development. For example, computational modeling of ADMET properties prevented a potential blood pressure-lowering drug from being lost early in the development process. The proposed compound showed low EC_{50} values, indicating that it was less potent than another compound of consideration. However, pharmacokinetic modeling showed that this compound would actually have greater efficacy than the one that showed higher potency. This compound did indeed show superior efficacy in the clinic [519].

1.5.5 Drug Metabolism and Pharmacokinetics/Absorption, Distribution, Metabolism, and Excretion and the Potential for Toxicity: Clinical Trial Prediction and Dosing

Computational tools are also being developed to address the possibility of simulating early clinical trials to avoid the waste resources inherent in testing drugs with poor ADMET profiles. This is a prevalent problem in drug development because up to 90% of drugs fail during clinical development and the time between reaching clinical trials and approval is up to 8 years [520]. These simulations aim at modeling the pathophysiology of biological systems and the pharmacology of treatments and can often incorporate things such as disease progression, placebo response, and dropout rates.

For example, clinical trial simulation was used by Laer *et al* to propose appropriate doses for sotalel [CAS 959-24-0; N-[4-[1-hydroxy-2-[(1-methylethyl)amino] ethyl] phenyl] methanesulfonamide hydrochloride] in children [521] and the Food and Drug Administration approved dosing changes for etanercept (Immunex Corporation, Thousand Oaks CA) in juvenile rheumatoid arthritis due to clinical trial simulations performed by Yim *et al* [522]. Simcyp (Simcyp Ltd, Sheffield UK) is a software package that creates virtual populations of participants with specifiable genetic and physiological characteristics using literature data. *In vitro* metabolism data can be applied to the *in-vitro-in-vivo* extrapolation process to simulate whole-live and hepatic clearances for these virtual populations [523]. Kowalski *et al* used the NONMEM software package (ICON plc, Dublin, Ireland) and PK/PD modeling to suggest a dosing regimen for SC-75416, a selective COX-2 inhibitor that would be comparable to the pain relief afforded from 50 mg of rofecoxib. This simulation saved an estimated nine months of development [524].

1.6 Conclusions

The extensive variety of computational tools used in drug discovery campaigns suggests that there are no fundamentally superior techniques. The performance of methods varies greatly with target protein, available data, and available resources. For example, Kruger and Evers completed a performance benchmark between structure- and ligand-based vHTS tools across four different targets, including angiotensin-converting enzyme, cyclooxygenase-2, thrombin and HIV-1 protease [525]. Docking methods including Glide, GOLD, Surflex, and FlexX were used to dock ligands into rigid target crystal structures obtained from PDB. A single ligand was used as a reference for ligand-based similarity search strategies such as 2D (fingerprints and feature trees) and 3D (Rapid Overlay of Chemical Structures (ROCS, OpenEye Scientific Software, Santa Fe, NM)), a similarity algorithm that calculates maximum volume overlap of two 3D structures [243, 526]. In general the authors found that docking methods performed poorly for HIV-1 protease and thrombin because of the flexible nature of the targets and the fact that the known ligands for these proteins have large molecular weight and peptidomimetic character.

Enrichments based on 3D similarity searches were poor for HIV-1 protease and thrombin datasets compared with ACE, which is likely due to the higher level of diversity in the HIV-1 protease and thrombin ligand datasets. Similarity scoring algorithms like ShapeTanimoto, ColorScore, and ComboScore were compared with the performance of ROCS [525]. It was found that even within the scoring, algorithm performance varied across targets. For example, ColorScore performed best for ACE and HIV-1 protease, whereas ShapeTanimoto for COX-2 and ComboScore was the method of choice for thrombin. All vHTS tools performed comparatively well for ACE, but ligand-based 2D fingerprint approach generally outperformed docking methods. The authors also note an important observation in that, especially for HIV-1 protease, the structure-based and ligand-based approaches yielded complimentary hit lists. Therefore, performance metrics are not the only benchmark to consider when comparing CADD techniques. In some cases, discovery of novel chemotypes is more important than high hit rates or high activity. In the current study, Kruger and Evers found that ROCS and feature trees were more successful in retrieving compounds with novel scaffolds compared to other fingerprints [525].

Warren *et al* published an in-depth assessment of the capabilities and shortcomings for docking programs and their scoring techniques against eight proteins of seven evolutionarily diverse target types. They found that docking programs were well adept at generating poses that included ones similar to those found in complex crystal structures. In general, although the molecular conformation was less precise across docking programs, they were fairly accurate in terms of the ligand's overall positioning. With regards to scoring, their findings agree with others that docking programs lack reliable scoring algorithms. So while the tools were able to predict a set of poses that included those that were seen in the crystal structure, the preference for the crystal structure pose was not necessarily reflected in the scoring. For five of the seven targets that were evaluated, the success rate, however, was greater than 40%. It was found that the enrichment of hits could be increased by applying previous knowledge regarding the target. However, there was little statistically significant correlation between docking scores and ligand affinity across the targets. The study concluded that a docking program's ability to reproduce accurate binding poses did not necessarily mean that the program could accurately predict binding affinities. This analysis underscores the necessity not only to re-rank the top hits from a docking-based vHTS using computationally expensive tools but also to continue evaluating novel scoring functions that can efficiently and accurately predict binding affinities [527].

Improvements in scoring functions involve the use of consensus scoring methods and free energy scoring with docking techniques. Consensus scoring methods have been shown to improve enrichments and prediction of bound conformations and poses by balancing out errors of individual scoring functions. In 2008, Enyedy and Egan compared docking scores of ligands with known IC_{50} and found that docking scores were incapable of correctly ranking compounds and were sometimes unable to differentiate active from inactive compounds. They concluded that individual scoring methods can be used successfully to enrich a dataset with increased population of actives but are insufficient to identify actives against inactives [13]. Page *et al* concluded that although binding energy calculations such as MM-PBSA are one of the more successful methods of estimating free energy of complexes, these techniques are more applicable to providing insights into the nature of interactions rather than prediction or screening [528]. Consensus scoring functions where free energy scores of different algorithms have been combined or averaged have been shown to substantially improve performance [529-532].

In their literature survey, Ripphausen *et al* reported that structure-based virtual screening was used much more frequently than ligand-based virtual screening (322 to 107 studies). Despite a preference for structure-based methods, ligand-based methods on average yield hits with higher potency than structure-based methods. Most ligand-based hits had activities better than 1 µM while structure-based hits fall frequently in the range of 1-100 µM [12]. Scoring algorithms in docking functions have been found to be biased toward known protein ligand complexes; for example more potent hits against protein kinase targets are discovered when compared to other target classes (figure 1.12) [61].

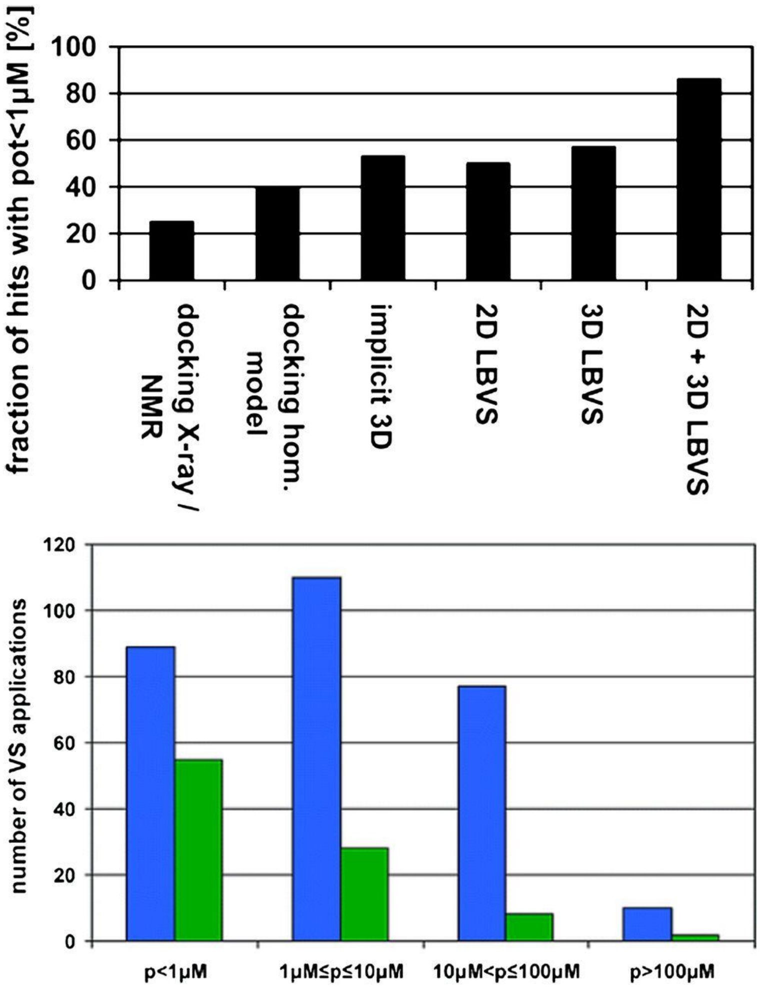


Figure 1.12 Ligand-based and structure-based lead compounds. Ripphausen, et al. report that ligand-based computationally approaches yield compounds with higher affinity than structure-based computationally approaches. Source: [533].

One CADD approach that has been gaining considerable momentum is the combination of structure-based and ligand-based computation techniques [534]. For example, the GRID-GOLPE method docks a set of ligands at a common binding site using GRID and then calculates descriptors for the binding interactions by probing these docking poses with GOLPE [535]. Multivariate regression is then used to create a statistical model that can explain the biological activity of these ligands. Structure-based interactions between a ligand and target can also be used in similarity-based searches to find compounds that are similar only in the regions that participate in binding rather than cross the entire ligand. LigandScout uses such a technique to define a pharmacophore based on hydrogen bonding and charge-transfer interactions between a ligand and its target. Another technique known as the pseudoreceptor technique [96] uses pharmacophore mapping-like overlaying techniques for a collection of ligands that bind to the same binding site to establish a virtual representation of the binding site's structure, which is then used as a template for docking and other structure-based vHTS. This approach has been utilized by VirtualToxLab [536] for the creation of nuclear receptors and cytochrome P450 binding site models in ADMET prediction tools and by Schneider *et al* in the modeling of the H4 receptor binding site subsequently used to identify novel active scaffolds [97]. In a recent review by Wilson and Lill [537], these methods are grouped into a major class of combined techniques called interaction based methods. A second major class involves the use of QSAR and similarity methods to enrich a library of virtual compounds prior to a molecular docking project. This can increase the efficiency of the project by reducing the number of compounds to be docked. This is similar to the application of CADD to enrich libraries prior to traditional HTS projects. This review also presents comprehensive descriptions of software packages using a combination of ligand- and structure-based techniques as well as several case studies testing the performance of these tools.

As discussed earlier, these methods are often used in serial where ligand-based methods are first used to enrich libraries that will subsequently be used in structure-based vHTS. The most common application is at the ligand library creation stage through the use of QSAR techniques to filter out compounds with low similarity to a query compound or no predicted activity based on a statistical model. QSAR has also been used as a means to refine the docking scores of a structure-based virtual screen. 2D and 3D QSAR can also be used to track docking errors. This method has been used by Novartis where a QSAR model is built from docking scores rather than observed activities, and this model is applied to that set to provide additional score weights for each compound [538].

Although CADD has been applied quite extensively in drug discovery campaigns, certain lucrative therapeutic targets like protein-protein interaction and protein-DNA interactions are still formidable, problems mainly because of the relatively massive size of interaction sites (in excess of 1500 $Å^2$) [1]. Lastly, accessibility has also been a problem with CADD as many tools are not designed with a friendly user interface in mind. In many cases, there can be an overwhelming number of variables that must be configured on a case-by-case basis and the interfaces are not always straightforward. A great deal of expertise is often required to use these tools to get desired measure of success. Increasingly, efforts are being made to develop user friendly interfaces especially in commercially available tools. For example, ICM-Pro (MolSoft L.L.C., San Diego, CA) is a software package designed to be a user friendly docking tool and replaces the front-end of current docking algorithms with an interface that is manageable to a wider audience [198]. More recently gamification of the ROSETTA folding program, known as Foldit [539], has allowed individuals outside of the scientific community to help solve the structure of M-PMV retroviral protease [540] and for predicting backbone remodeling of computationally designed biomolecular Diels-Alderase that increased its activity [541]. The successful application of crowd-sourced biomolecule design and prediction suggests further potential of CADD methods in drug discovery.

Acknowledgements

This work was supported by the National Science Foundation through the Office for Cyber Infrastructure Transformative Computational Sciences Fellowship [OCI-1122919] (E.W.L.); the National Institutes of Health National Institute of Mental Health [Grant R01 MH090192] (to the Meiler laboratory); the National Institutes of Health National Institute of General Medical Sciences [Grant R01 GM099842]; and the National Institutes of Health National Institute of Diabetes and Digestive and Kidney Diseases [Grant R01 DK097376].

1.7 References

1. Van Drie JH (2007) Computer-aided drug design: the next 20 years. *J Comput Aided Mol Des* 21(10-11):591-601.
2. Doman TN*, et al.* (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* 45(11):2213-2221.
3. Vijayakrishnan R (2009) Structure-based drug design and modern medicine. *J Postgrad Med* 55(4):301-304.
4. Talele TT, Khedkar SA, & Rigby AC (2010) Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current topics in medicinal chemistry* 10(1):127-141.
5. Hartman GD*, et al.* (1992) Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *Journal of medicinal chemistry* 35(24):4640-4642.
6. Sawyer JS*, et al.* (2003) Synthesis and activity of new aryl- and heteroaryl-substituted pyrazole inhibitors of the transforming growth factor-beta type I receptor kinase domain. *Journal of medicinal chemistry* 46(19):3953-3956.
7. Singh J*, et al.* (2003) Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFbeta receptor kinase (TbetaRI). *Bioorganic & medicinal chemistry letters* 13(24):4355-4359.
8. Shekhar C (2008) In silico pharmacology: computer-aided methods could transform drug development. *Chem Biol* 15(5):413-414.

9. Kalyaanamoorthy S & Chen YP (2011) Structure-based drug design to augment hit discovery. *Drug Discov Today* 16(17-18):831-839.
10. Jorgensen WL (2010) Drug discovery: Pulled from a protein's embrace. *Nature* 466(7302):42-43.
11. Horvath D (1997) A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J Med Chem* 40(15):2412-2423.
12. Ripphausen P, Nisius B, Peltason L, & Bajorath J (2010) Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* 53(24):8461-8467.
13. Enyedy IJ & Egan WJ (2008) Can we use docking and scoring for hit-to-lead optimization? *J Comput Aided Mol Des* 22(3-4):161-168.
14. Joffe E (1991) Complication during root canal therapy following accidental extrusion of sodium hypochlorite through the apical foramen. *Gen Dent* 39(6):460-461.
15. Jorgensen WL (2004) The many roles of computation in drug discovery. *Science* 303(5665):1813-1818.
16. Basak SC (2012) Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. *Curr Comput Aided Drug Des* 8(1):1-2.
17. Bohacek RS, McMartin C, & Guida WC (1996) The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal research reviews* 16(1):3-50.
18. Schneider G*, et al.* (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends in biotechnology* 27(1):18-26.
19. Agarwal AK & Fishwick CW (2010) Structure-based design of anti-infectives. *Ann N Y Acad Sci* 1213:20-45.
20. Fink T, Bruggesser H, & Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angewandte Chemie* 44(10):1504-1508.
21. Fink T & Reymond JL (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of chemical information and modeling* 47(2):342-353.
22. Blum LC & Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society* 131(25):8732-8733.
23. Song CM, Lim SJ, & Tong JC (2009) Recent advances in computer-aided drug design. *Brief Bioinform* 10(5):579-591.
24. Ortholand JY & Ganesan A (2004) Natural products and combinatorial chemistry: back to the future. *Curr Opin Chem Biol* 8(3):271-280.
25. Wheeler DL*, et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 34(Database issue):D173-180.
26. Information NCfB (2013) PubMed. (NCBI).
27. accelrys (2012) Accelrys Available Chemicals Directory (ACD). (accelrys).
28. Dimitropoulos D, Ionides, J. and Henrick K (2006) Using PDBeChem to Search the PDB Ligand Dictionary. *Current Protocols in Bioinformatics* (John Wiley & Sons), pp 14.13.11-14.13.13.
29. Irwin JJ & Shoichet BK (2005) ZINC - A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45(1):177-182.
30. Wishart DS*, et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 34:D668-D672.
31. Chen J, Swamidass SJ, Bruand J, & Baldi P (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* 21(22):4133-4139.

32. Chen JH, Linstead E, Swamidass SJ, Wang D, & Baldi P (2007) ChemDB update - full-text search and virtual chemical space. *Bioinformatics* 23(17):2348-2351.
33. Ekins S, Mestres J, & Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology* 152(1):9-20.
34. Hristozov DP, Oprea TI, & Gasteiger J (2007) Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J Comput Aided Mol Des* 21(10-11):617-640.
35. Mandal S, Moudgil M, & Mandal SK (2009) Rational drug design. *Eur J Pharmacol* 625(1-3):90-100.
36. Lipinski CA, Lombardo F, Dominy BW, & Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* 46(1-3):3-26.
37. Kelder J, Grootenhuis PDJ, Bayada DM, Delbressine LPC, & Ploemen JP (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharmaceut Res* 16(10):1514-1519.
38. Orry AJW, Abagyan RA, & Cavasotto CN (2006) Structure-based development of target-specific compound libraries. *Drug Discov Today* 11(5-6):261-266.
39. Harris CJ, Hill RD, Sheppard DW, Slater MJ, & Stouten PFW (2011) The Design and Application of Target-Focused Compound Libraries. *Comb Chem High T Scr* 14(6):521-531.
40. Anderson AC (2012) Structure-based functional design of drugs: from target to lead compound. *Methods in molecular biology* 823:359-366.
41. Cavasotto CN & Phatak SS (2011) Docking methods for structure-based library design. *Methods in molecular biology* 685:155-174.
42. Liwo A, Czaplewski C, Oldziej S, & Scheraga HA (2008) Computational techniques for efficient conformational sampling of proteins. *Current opinion in structural biology* 18(2):134-139.
43. Foloppe N & Chen IJ (2009) Conformational Sampling and Energetics of Drug-Like Molecules. *Current Medicinal Chemistry* 16(26):3381-3413.
44. Wiswesser WJ (1985) Historic development of chemical notations. *Journal of chemical information and computer sciences* 25(3):258-263.
45. Wiswesser WJ (1954) *A line-formula chemical notation* (Crowell, New York,) p 149 p.
46. Weininger D (1988) Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules. *Journal of chemical information and computer sciences* 28(1):31-36.
47. Weininger SJ & Stermitz FR (1984) *Organic chemistry* (Academic Press, Orlando) pp xviii, 1121 p.
48. Daylight Chemical Information Systems I (2008) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns.
49. InChITRUST (2013) InChI FAQ.
50. Southan C (2013) InChI in the wild: an assessment of InChIKey searching in Google. *J Cheminform* 5(1):10.
51. RCSB (2013) RCSB Protein Data Bank. (RCSB).
52. Arnold K, Bordoli L, Kopp J, & Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2):195-201.
53. Schames JR*, et al.* (2004) Discovery of a novel binding trench in HIV integrase. *J Med Chem* 47(8):1879-1881.
54. Durrant JD & McCammon JA (2010) Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr Opin Pharmacol* 10(6):770-774.

55. Meiler J & Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* 65(3):538-548.
56. Ball P (2008) Water as an active constituent in cell biology. *Chemical Reviews* 108(1):74-108.
57. Levitt M & Park BH (1993) Water - Now You See It, Now You Dont. *Structure* 1(4):223-226.
58. Ladbury JE (1996) Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol* 3(12):973-980.
59. Li Z & Lazaridis T (2007) Water at biomolecular binding interfaces. *Physical Chemistry Chemical Physics* 9(5):573-581.
60. de Beer SB, Vermeulen NP, & Oostenbrink C (2010) The role of water molecules in computational drug design. *Current topics in medicinal chemistry* 10(1):55-66.
61. Stumpfe D, Ripphausen P, & Bajorath J (2012) Virtual compound screening in drug discovery. *Future medicinal chemistry* 4(5):593-602.
62. Cleves AE & Jain AN (2006) Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem* 49(10):2921-2938.
63. Huang N, Shoichet BK, & Irwin JJ (2006) Benchmarking sets for molecular docking. *J Med Chem* 49(23):6789-6801.
64. Irwin JJ (2008) Community benchmarks for virtual screening. *J Comput Aided Mol Des* 22(3-4):193-199.
65. Good AC & Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J Comput Aided Mol Des* 22(3-4):169-178.
66. NIH-structure_based (
67. Lundstrom K (2011) Genomics and drug discovery. *Future Med Chem* 3(15):1855-1858.
68. Bambini S & Rappuoli R (2009) The use of genomics in microbial vaccine development. *Drug Discov Today* 14(5-6):252-260.
69. Warner SL*, et al.* (2006) Identification of a lead small-molecule inhibitor of the Aurora kinases using a structure-assisted, fragment-based approach. *Mol Cancer Ther* 5(7):1764-1773.
70. Budzik B*, et al.* (2010) Novel N-Substituted Benzimidazolones as Potent, Selective, CNS-Penetrant, and Orally Active M(1) mAChR Agonists. *Acs Medicinal Chemistry Letters* 1(6):244-248.
71. Becker OM*, et al.* (2006) An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J Med Chem* 49(11):3116-3135.
72. Evers A, Gohlke H, & Klebe G (2003) Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J Mol Biol* 334(2):327-345.
73. Evers A & Klebe G (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of Medicinal Chemistry* 47(22):5381-5392.
74. Fauman EB, Rai BK, & Huang ES (2011) Structure-based druggability assessment - identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol* 15(4):463-468.
75. Hajduk PJ, Huth JR, & Tse C (2005) Predicting protein druggability. *Drug Discov Today* 10(23-24):1675-1682.
76. Laurie ATR & Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for Structure-Based Drug Design and virtual ligand screening. *Curr Protein Pept Sc* 7(5):395-406.
77. Buchan DW*, et al.* (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 38(Web Server issue):W563-568.
78. Marti-Renom MA*, et al.* (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Bioph Biom* 29:291-325.

79. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of molecular biology* 215(3):403-410.
80. Soding J & Remmert M (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Current opinion in structural biology* 21(3):404-411.
81. Kelley LA & Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols* 4(3):363-371.
82. Thompson JD, Higgins DG, & Gibson TJ (1994) Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic acids research* 22(22):4673-4680.
83. Misura KMS, Chivian D, Rohl CA, Kim DE, & Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. *P Natl Acad Sci USA* 103(14):5361-5366.
84. Chivian D & Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic acids research* 34(17).
85. Rai BK & Fiser A (2006) Multiple mapping method: A novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. *Proteins-Structure Function and Bioinformatics* 63(3):644-661.
86. Hillisch A, Pineda LF, & Hilgenfeld R (2004) Utility of homology models in the drug discovery process. *Drug Discov Today* 9(15):659-669.
87. Canutescu AA & Dunbrack RL (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science* 12(5):963-972.
88. Mandell DJ, Coutsias EA, & Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods* 6(8):551-552.
89. Coutsias EA & Seok C (2004) Kinematic view of loop closure. *Abstr Pap Am Chem S* 228:U534-U534.
90. Krivov GG, Shapovalov MV, & Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins-Structure Function and Bioinformatics* 77(4):778-795.
91. Desmet J, Demaeyer M, Hazes B, & Lasters I (1992) The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning. *Nature* 356(6369):539-542.
92. Dunbrack RL & Karplus M (1993) Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction. *Journal of molecular biology* 230(2):543-574.
93. Dunbrack RL & Karplus M (1994) Conformational-Analysis of the Backbone-Dependent Rotamer Preferences of Protein Side-Chains. *Nat Struct Biol* 1(5):334-340.
94. Bower MJ, Cohen FE, & Dunbrack RL (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *Journal of molecular biology* 267(5):1268-1282.
95. Rohl CA, Strauss CEM, Misura KMS, & Baker D (2004) Protein structure prediction using rosetta. *Method Enzymol* 383:66-+.
96. Tanrikulu Y & Schneider G (2008) Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. *Nature reviews. Drug discovery* 7(8):667-677.
97. Tanrikulu Y*, et al.* (2009) Homology model adjustment and ligand screening with a pseudoreceptor of the human histamine H4 receptor. *ChemMedChem* 4(5):820-827.
98. Katritch V, Rueda M, Lam PC, Yeager M, & Abagyan R (2010) GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins-Structure Function and Genetics* 78(1):197-211.

99. Raval A, Piana S, Eastwood MP, Dror RO, & Shaw DE (2012) Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins-Structure Function and Genetics*.
100. Misura KMS & Baker D (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins-Structure Function and Bioinformatics* 59(1):15-29.
101. Xiang Z (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci* 7(3):217-227.
102. Serrano ML, Perez HA, & Medina JD (2006) Structure of C-terminal fragment of merozoite surface protein-1 from Plasmodium vivax determined by homology modeling and molecular dynamics refinement. *Bioorg Med Chem* 14(24):8359-8365.
103. Li W*, et al.* (2008) Probing ligand binding modes of human cytochrome P450 2J2 by homology modeling, molecular dynamics simulation, and flexible molecular docking. *Proteins-Structure Function and Genetics* 71(2):938-949.
104. Melo F & Sali A (2007) Fold assessment for comparative protein structure modeling. *Protein Science* 16(11):2412-2426.
105. Cozzetto D*, et al.* (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins-Structure Function and Bioinformatics* 77:18-28.
106. Kiefer F, Arnold K, Kunzli M, Bordoli L, & Schwede T (2009) The SWISS-MODEL Repository and associated resources. *Nucleic acids research* 37:D387-D392.
107. Pieper U*, et al.* (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research* 37:D347-D354.
108. Park H, Bahn YJ, & Ryu SE (2009) Structure-based de novo design and biochemical evaluation of novel Cdc25 phosphatase inhibitors. *Bioorganic & Medicinal Chemistry Letters* 19(15):4330-4334.
109. Henrich S*, et al.* (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit* 23(2):209-219.
110. Levitt DG & Banaszak LJ (1992) Pocket - a Computer-Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino-Acids. *J Mol Graphics* 10(4):229-234.
111. Hendlich M, Rippmann F, & Barnickel G (1997) LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15(6):359-+.
112. Laskowski RA (1995) Surfnet - a Program for Visualizing Molecular-Surfaces, Cavities, and Intermolecular Interactions. *J Mol Graphics* 13(5):323-&.
113. Desjarlais RL*, et al.* (1988) Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor-Binding Site of Known 3-Dimensional Structure. *J Med Chem* 31(4):722-729.
114. Smithson DC, Lee J, Shelat AA, Phillips MA, & Guy RK (2010) Discovery of Potent and Selective Inhibitors of Trypanosoma brucei Ornithine Decarboxylase. *Journal of Biological Chemistry* 285(22):16771-16781.
115. Cerchietti LC*, et al.* (2010) A Small-Molecule Inhibitor of BCL6 Kills DLBCL Cells In Vitro and In Vivo. *Cancer Cell* 17(4):400-411.
116. Laurie ATR & Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21(9):1908-1916.
117. Reynolds CA, Wade RC, & Goodford PJ (1989) Identifying Targets for Bioreductive Agents - Using Grid to Predict Selective Binding Regions of Proteins. *J Mol Graphics* 7(2):103-&.
118. Wade RC, Clark KJ, & Goodford PJ (1993) Further Development of Hydrogen-Bond Functions for Use in Determining Energetically Favorable Binding-Sites on Molecules of Known Structure .1. Ligand Probe Groups with the Ability to Form 2 Hydrogen-Bonds. *J Med Chem* 36(1):140-147.

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

120. Kortvelyesi T, Dennis S, Silberstein M, Brown L, 3rd, & Vajda S (2003) Algorithms for computational solvent mapping of proteins. *Proteins* 51(3):340-351.

121. Kim D*, et al.* (2008) Pocket extraction on proteins via the Voronoi diagram of spheres. *J Mol Graph Model* 26(7):1104-1112.

122. Segers K*, et al.* (2007) Design of protein-membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V. *P Natl Acad Sci USA* 104(31):12697-12702.

123. Landon MR*, et al.* (2008) Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem Biol Drug Des* 71(2):106-116.

124. An JH*, et al.* (2009) A Novel Small-Molecule Inhibitor of the Avian Influenza H5N1 Virus Determined through Computational Screening against the Neuraminidase. *J Med Chem* 52(9):2667-2672.

125. Porter CT, Bartlett GJ, & Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32(Database issue):D129-133.

126. Arakaki AK, Zhang Y, & Skolnick J (2004) Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* 20(7):1087-1096.

127. Ferre F, Ausiello G, Zanzoni A, & Helmer-Citterich M (2004) SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res* 32(Database issue):D240-244.

128. Chikhi R, Sael L, & Kihara D (2010) Real-time ligand binding pocket database search using local surface descriptors. *Proteins-Structure Function and Genetics* 78(9):2007-2028.

129. Sael L & Kihara D (2012) Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins-Structure Function and Genetics* 80(4):1177-1195.

130. Sayle RA & Milner-White EJ (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20(9):374.

131. Vangunsteren WF & Berendsen HJC (1990) Computer-Simulation of Molecular-Dynamics - Methodology, Applications, and Perspectives in Chemistry. *Angew Chem Int Edit* 29(9):992-1023.

132. Schlitter J, Engels M, & Kruger P (1994) Targeted Molecular-Dynamics - a New Approach for Searching Pathways of Conformational Transitions. *J Mol Graphics* 12(2):84-89.

133. Huber T & van Gunsteren WF (1998) SWARM-MD: Searching conformational space by cooperative molecular dynamics. *J Phys Chem A* 102(29):5937-5943.

134. Grubmuller H (1995) Predicting Slow Structural Transitions in Macromolecular Systems - Conformational Flooding. *Phys Rev E* 52(3):2893-2906.

135. Abrams CF & Vanden-Eijnden E (2010) Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *P Natl Acad Sci USA* 107(11):4961-4966.

136. Sugita Y & Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314(1-2):141-151.

137. Summa V*, et al.* (2008) Discovery of Raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *J Med Chem* 51(18):5843-5855.

138. Frembgen-Kesner T & Elcock AH (2006) Computational sampling of a cryptic drug binding site in a protein receptor: Explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase. *Journal of molecular biology* 359(1):202-214.

139. Halperin I, Ma B, Wolfson H, & Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47(4):409-443.

140. Kitchen DB, Decornez H, Furr JR, & Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* 3(11):935-949.

141. Taylor JS & Burnett RM (2000) DARWIN: a program for docking flexible molecules. *Proteins* 41(2):173-191.

142. Connolly ML (1983) Analytical Molecular-Surface Calculation. *J Appl Crystallogr* 16(Oct):548-558.

143. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, & Ferrin TE (1982) A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of molecular biology* 161(2):269-288.

144. Katchalskikatzir E*, et al.* (1992) Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *P Natl Acad Sci USA* 89(6):2195-2199.

145. Dias R & de Azevedo WF (2008) Molecular Docking Algorithms. *Curr Drug Targets* 9(12):1040-1047.

146. Changeux JP & Edelstein S (2011) Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep* 3:19.

147. Friesner RA*, et al.* (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 47(7):1739-1749.

148. Desjarlais RL, Sheridan RP, Dixon JS, Kuntz ID, & Venkataraghavan R (1986) Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape. *J Med Chem* 29(11):2149-2153.

149. Rarey M, Kramer B, Lengauer T, & Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* 261(3):470-489.

150. Miller MD, Kearsley SK, Underwood DJ, & Sheridan RP (1994) Flog - a System to Select Quasi-Flexible Ligands Complementary to a Receptor of Known 3-Dimensional Structure. *J Comput Aided Mol Des* 8(2):153-174.

151. Jain AN (2003) Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* 46(4):499-511.

152. Majeux N, Scarsi M, & Caflisch A (2001) Efficient electrostatic solvation model for protein-fragment docking. *Proteins-Structure Function and Genetics* 42(2):256-268.

153. Kellenberger E, Rodrigo J, Muller P, & Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins-Structure Function and Genetics* 57(2):225-242.

154. Cross JB*, et al.* (2009) Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model* 49(6):1455-1474.

155. Pierce AC, Jacobs M, & Stuver-Moody C (2008) Docking study yields four novel inhibitors of the protooncogene Pim-1 kinase. *J Med Chem* 51(6):1972-1975.

156. Chiu TL*, et al.* (2009) Identification of Novel Non-Hydroxamate Anthrax Toxin Lethal Factor Inhibitors by Topomeric Searching, Docking and Scoring, and in Vitro Screening. *J Chem Inf Model* 49(12):2726-2734.

157. Milne GWA, Nicklaus MC, Driscoll JS, Wang SM, & Zaharevitz D (1994) National-Cancer-Institute Drug Information-System 3d Database. *Journal of chemical information and computer sciences* 34(5):1219-1224.

158. Friedman R & Caflisch A (2009) Discovery of Plasmepsin Inhibitors by Fragment-Based Docking and Consensus Scoring. *ChemMedChem* 4(8):1317-1326.

159. Ekonomiuk D*, et al.* (2009) Flaviviral Protease Inhibitors Identified by Fragment-Based Library Docking into a Structure Generated by Molecular Dynamics. *Journal of medicinal chemistry* 52(15):4860-4868.

160. Ekonomiuk D*, et al.* (2009) Discovery of a Non-Peptidic Inhibitor of West Nile Virus NS3 Protease by High-Throughput Docking. *Plos Neglect Trop D* 3(1).

161. Lafleur K, Huang DZ, Zhou T, Caflisch A, & Nevado C (2009) Structure-Based Optimization of Potent and Selective Inhibitors of the Tyrosine Kinase Erythropoietin Producing Human Hepatocellular Carcinoma Receptor B4 (EphB4). *Journal of medicinal chemistry* 52(20):6433-6446.
162. Luksch T*, et al.* (2008) Computer-aided design and synthesis of nonpeptidic plasmepsin II and IV inhibitors. *ChemMedChem* 3(9):1323-1336.
163. Chen D*, et al.* (2008) Novel inhibitors of anthrax edema factor. *Bioorganic & medicinal chemistry* 16(15):7225-7233.
164. Bui NK*, et al.* (2011) Development of screening assays and discovery of initial inhibitors of pneumococcal peptidoglycan deacetylase PgdA. *Biochemical pharmacology* 82(1):43-52.
165. Shah F*, et al.* (2011) Design, synthesis and biological evaluation of novel benzothiazole and triazole analogs as falcipain inhibitors. *Medchemcomm* 2(12):1201-1207.
166. Caporuscio F, Rastelli G, Imbriano C, & Del Rio A (2011) Structure-Based Design of Potent Aromatase Inhibitors by High-Throughput Docking. *Journal of medicinal chemistry* 54(12):4006-4017.
167. Zhao L*, et al.* (2006) FK506-binding protein ligands: structure-based design, synthesis, and neurotrophic/neuroprotective properties of substituted 5,5-dimethyl-2-(4-thiazolidine) carboxylates. *Journal of medicinal chemistry* 49(14):4059-4071.
168. Mangoni R, Roccatano D, & Di Nola A (1999) Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins-Structure Function and Bioinformatics* 35(2):153-162.
169. Durrant JD, Urbaniak MD, Ferguson MAJ, & McCammon JA (2010) Computer-Aided Identification of Trypanosoma brucei Uridine Diphosphate Galactose 4 '-Epimerase Inhibitors: Toward the Development of Novel Therapies for African Sleeping Sickness. *J Med Chem* 53(13):5025-5032.
170. Amaro RE*, et al.* (2008) Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei. *P Natl Acad Sci USA* 105(45):17278-17283.
171. Leone V, Marinelli F, Carloni P, & Parrinello M (2010) Targeting biomolecular flexibility with metadynamics. *Curr Opin Struc Biol* 20(2):148-154.
172. Biarnes X, Bongarzone S, Vargiu AV, Carloni P, & Ruggerone P (2011) Molecular motions in drug design: the coming age of the metadynamics method. *J Comput Aid Mol Des* 25(5):395-402.
173. Shaw DE*, et al.* (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* 51(7):91-97.
174. Shan YB*, et al.* (2011) How Does a Drug Molecule Find Its Target Binding Site? *J Am Chem Soc* 133(24):9181-9183.
175. Lindorff-Larsen K, Piana S, Dror RO, & Shaw DE (2011) How Fast-Folding Proteins Fold. *Science* 334(6055):517-520.
176. Sousa SF, Fernandes PA, & Ramos MJ (2006) Protein-ligand docking: Current status and future challenges. *Proteins-Structure Function and Bioinformatics* 65(1):15-26.
177. Liu M & Wang SM (1999) MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* 13(5):435-451.
178. Abagyan R, Totrov M, & Kuznetsov D (1994) Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of computational chemistry* 15(5):488-506.
179. Davis IW & Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. *Journal of molecular biology* 385(2):381-392.

180. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, & Meiler J (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49(14):2987-2998.
181. RosettaCommons (2013) Rosetta - The premier software suite for macromolecular modeling. (RosettaCommons).
182. McMartin C & Bohacek RS (1997) QXP: Powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des* 11(4):333-344.
183. Totrov M & Abagyan R (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* Suppl 1:215-220.
184. Kaufmann KW*, et al.* (2009) Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins* 74(3):630-642.
185. Malamas MS*, et al.* (2009) Aminoimidazoles as Potent and Selective Human beta-Secretase (BACE1) Inhibitors. *J Med Chem* 52(20):6314-6323.
186. Nowak P*, et al.* (2010) Discovery and initial optimization of 5,5 '-disubstituted aminohydantoins as potent beta-secretase (BACE1) inhibitors. *Bioorganic & Medicinal Chemistry Letters* 20(2):632-635.
187. Malamas MS*, et al.* (2010) Novel pyrrolyl 2-aminopyridines as potent and selective human beta-secretase (BACE1) inhibitors. *Bioorganic & Medicinal Chemistry Letters* 20(7):2068-2073.
188. Chan DSH*, et al.* (2010) Structure-Based Discovery of Natural-Product-like TNF-alpha Inhibitors. *Angew Chem Int Edit* 49(16):2860-2864.
189. Jones G, Willett P, Glen RC, Leach AR, & Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology* 267(3):727-748.
190. Morris GM, Goodsell DS, Halliday RS, Huey R, & Olson A (1998) Automated Docking Using a Lamarckian Genetic Algotithm and an Empirical Binding Free energy Function. *J. Comp. Chem.* 19(14)):1639-1662.
191. Kontoyianni M, McClellan LM, & Sokol GS (2004) Evaluation of docking performance: Comparative data on docking algorithms. *J Med Chem* 47(3):558-565.
192. Narlawar R*, et al.* (2010) Hybrid Ortho/Allosteric Ligands for the Adenosine A1 Receptor. *Journal of medicinal chemistry* 53(8):3028-3037.
193. Park H*, et al.* (2008) Discovery and biological evaluation of novel alpha-glucosidase inhibitors with in vivo antidiabetic effect. *Bioorganic & Medicinal Chemistry Letters* 18(13):3711-3715.
194. Durrant JD*, et al.* (2010) Novel Naphthalene-Based Inhibitors of Trypanosoma brucei RNA Editing Ligase 1. *Plos Neglect Trop D* 4(8).
195. B-Rao C, Subramanian J, & Sharma SD (2009) Managing protein flexibility in docking and its applications. *Drug Discov Today* 14(7-8):394-400.
196. Sinko W, Lindert S, & McCammon JA (2013) Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem Biol Drug Des* 81(1):41-49.
197. Carlson HA (2002) Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* 6(4):447-452.
198. Abagyan R*, et al.* (2006) Disseminating structural genomics data to the public: from a data dump to an animated story. *Trends in biochemical sciences* 31(2):76-78.
199. Cozzini P*, et al.* (2008) Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J Med Chem* 51(20):6237-6255.
200. Halgren TA (1996) Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem* 17(5-6):490-519.
201. Shoichet BK, Leach AR, & Kuntz ID (1999) Ligand solvation in molecular docking. *Proteins-Structure Function and Genetics* 34(1):4-16.

202. Kukic P & Nielsen JE (2010) Electrostatics in proteins and protein-ligand complexes. *Future Med Chem* 2(4):647-666.

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

204. Bohm HJ (1992) The Computer-Program Ludi - a New Method for the Denovo Design of Enzyme-Inhibitors. *J Comput Aided Mol Des* 6(1):61-78.

205. Shimada J, Ishchenko AV, & Shakhnovich EI (2000) Analysis of knowledge-based protein-ligand potentials using a self-consistent method. *Protein Science* 9(4):765-775.

206. Velec HFG, Gohlke H, & Klebe G (2005) DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem* 48(20):6296-6303.

207. DeWitte RS & Shakhnovich E (1997) SMoG: De novo design method based on simple, fast and accurate free energy estimates. *Abstr Pap Am Chem S* 214:6-Comp.

208. Mitchell JBO, Laskowski RA, Alex A, Forster MJ, & Thornton JM (1999) BLEEP - Potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *Journal of computational chemistry* 20(11):1177-1185.

209. Feher M (2006) Consensus scoring for protein-ligand interactions. *Drug Discov Today* 11(9-10):421-428.

210. O'Boyle NM, Liebeschuetz JW, & Cole JC (2009) Testing Assumptions and Hypotheses for Rescoring Success in Protein-Ligand Docking. *J Chem Inf Model* 49(8):1871-1878.

211. Okamoto M*, et al.* (2009) Identification of Death-Associated Protein Kinases Inhibitors Using Structure-Based Virtual Screening. *J Med Chem* 52(22):7323-7327.

212. Muegge I & Martin YC (1999) A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J Med Chem* 42(5):791-804.

213. Roughley S, Wright L, Brough P, Massey A, & Hubbard RE (2012) Hsp90 inhibitors and drugs from fragment and virtual screening. *Topics in current chemistry* 317:61-82.

214. Lu IL*, et al.* (2006) Structure-based drug design of a novel family of PPAR gamma partial agonists: Virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. *J Med Chem* 49(9):2703-2712.

215. Li N*, et al.* (2009) Discovery of novel inhibitors of Streptococcus pneumoniae based on the virtual screening with the homology-modeled structure of histidine kinase (VicK). *Bmc Microbiol* 9.

216. Izuhara Y*, et al.* (2010) A novel inhibitor of plasminogen activator inhibitor-1 provides antithrombotic benefits devoid of bleeding effect in nonhuman primates. *J Cereb Blood Flow Metab* 30(5):904-912.

217. Triballeau N*, et al.* (2008) High-Potency Olfactory Receptor Agonists Discovered by Virtual High-Throughput Screening: Molecular Probes for Receptor Structure and Olfactory Function. *Neuron* 60(5):767-774.

218. Simmons KJ, Chopra I, & Fishwick CW (2010) Structure-based discovery of antibacterial drugs. *Nat Rev Microbiol* 8(7):501-510.

219. Ruiz FM*, et al.* (2008) Structure-based discovery of novel non-nucleosidic DNA alkyltransferase inhibitors: Virtual screening and in vitro and in vivo activities. *J Chem Inf Model* 48(4):844-854.

220. Baurin N*, et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *Journal of chemical information and computer sciences* 44(2):643-651.

221. Honig B & Nicholls A (1995) Classical Electrostatics in Biology and Chemistry. *Science* 268(5214):1144-1149.

222. Bashford D & Case DA (2000) Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry* 51:129-152.
223. Cozza G*, et al.* (2006) Identification of ellagic acid as potent inhibitor of protein kinase CK2: A successful example of a virtual screening application. *J Med Chem* 49(8):2363-2366.
224. Cozza G*, et al.* (2009) Quinalizarin as a potent, selective and cell-permeable inhibitor of protein kinase CK2. *Biochem J* 421:387-395.
225. Tice CM*, et al.* (2010) Spirocyclic ureas: Orally bioavailable 11 beta-HSD1 inhibitors identified by computer-aided drug design. *Bioorganic & Medicinal Chemistry Letters* 20(3):881-886.
226. Tice CM*, et al.* (2010) Spirocyclic ureas: Orally bioavailable 11 beta-HSD1 inhibitors identified by computer-aided drug design. *Bioorganic & medicinal chemistry letters* 20(3):881-886.
227. Yang SY (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today* 15(11-12):444-450.
228. Wolber G & Langer T (2005) LigandScout: 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters. *J Chem Inf Model* 45(1):160-169.
229. Chen J & Lai LH (2006) Pocket v.2: Further developments on receptor-based pharmacophore modeling. *J Chem Inf Model* 46(6):2684-2691.
230. Wang RX, Liu L, Lai LH, & Tang YQ (1998) SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J Mol Model* 4(12):379-394.
231. Schuster D*, et al.* (2008) Discovery of nonsteroidal 17 beta-hydroxysteroid dehydrogenase 1 inhibitors by pharmacophore-based screening of virtual compound libraries. *J Med Chem* 51(14):4188-4199.
232. Brvar M, Perdih A, Oblak M, Masic LP, & Solmajer T (2010) In silico discovery of 2-amino-4-(2,4-dihydroxyphenyl)thiazoles as novel inhibitors of DNA gyrase B. *Bioorganic & Medicinal Chemistry Letters* 20(3):958-962.
233. Wei DG*, et al.* (2008) Discovery of Multitarget Inhibitors by Combining Molecular Docking with Common Pharmacophore Matching. *Journal of medicinal chemistry* 51(24):7882-7888.
234. Shuker SB, Hajduk PJ, Meadows RP, & Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274(5292):1531-1534.
235. Wang RX, Gao Y, & Lai LH (2000) LigBuilder: A multi-purpose program for structure-based drug design. *J Mol Model* 6(7-8):498-516.
236. Yuan YX, Pei JF, & Lai LH (2011) LigBuilder 2: A Practical de Novo Drug Design Approach. *J Chem Inf Model* 51(5):1083-1091.
237. Vinkers HM*, et al.* (2003) SYNOPSIS: SYNthesize and OPtimize System in Silico. *J Med Chem* 46(13):2765-2773.
238. Cogan DA*, et al.* (2008) Structure-based design and subsequent optimization of 2-tolyl-(1,2,3-triazol-1-yl-4-carboxamide) inhibitors of p38 MAP kinase. *Bioorganic & medicinal chemistry letters* 18(11):3251-3255.
239. Zhang S*, et al.* (2011) SKLB1002, a Novel Potent Inhibitor of VEGF Receptor 2 Signaling, Inhibits Angiogenesis and Tumor Growth In Vivo. *Clin Cancer Res* 17(13):4439-4450.
240. Li WW*, et al.* (2010) Taking Quinazoline as a General Support-Nog to Design Potent and Selective Kinase Inhibitors: Application to FMS-like Tyrosine Kinase 3. *ChemMedChem* 5(4):513-516.
241. Johnson MA, Maggiora GM, & American Chemical Society. Meeting (1990) *Concepts and applications of molecular similarity* (Wiley, New York) pp xix, 393 p.
242. Hemmer MC, Steinhauer V, & Gasteiger J (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* 19(1):151-164.
243. Schuur JH, Selzer P, & Gasteiger J (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and

studies of biological activity. *Journal of chemical information and computer sciences* 36(2):334-344.
244. Pearlman RS & Smith KM (1999) Metric validation and the receptor-relevant subspace concept. *Journal of chemical information and computer sciences* 39(1):28-35.
245. Bravi G*, et al.* (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J Comput Aided Mol Des* 11(1):79-92.
246. Randic M (1995) Molecular Profiles - Novel Geometry-Dependent Molecular Descriptors. *New J Chem* 19(7):781-791.
247. Hong H*, et al.* (2008) Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* 48(7):1337-1344.
248. Cramer RD, Patterson DE, & Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110(18):5959-5967.
249. Acharya C, Coop A, Polli JE, & Mackerell AD, Jr. (2011) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr Comput Aided Drug Des* 7(1):10-22.
250. Marrero-Ponce Y, Santiago OM, Lopez YM, Barigye SJ, & Torrens F (2012) Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application. *J Comput Aided Mol Des* 26(11):1229-1246.
251. March J (1977) *Advanced organic chemistry : reactions, mechanisms, and structure* (McGraw-Hill, New York) 2d Ed pp xv, 1328 p.
252. Pimentel GC & McClellan AL (1960) *The hydrogen bond* (W.H. Freeman).
253. Vinogradov SN & Linnell RH (1971) *Hydrogen bonding* (Van Nostrand Reinhold, New York,) pp xi, 319 p.
254. Zhou T, Huang D, & Caflisch A (2010) Quantum mechanical methods for drug design. *Current topics in medicinal chemistry* 10(1):33-45.
255. Gasteiger J & Marsili M (1980) Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. *Tetrahedron* 36(22):3219-3228.
256. Mulliken RS (1934) A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *The Journal of Chemical Physics* 2(11):782-793.
257. Hinze J & Jaffe HH (1962) Electronegativity. I. Orbital Electronegativity of Neutral Atoms. *Journal of the American Chemical Society* 84(4):540-546.
258. Hinze J, Whitehead MA, & Jaffe HH (1963) Electronegativity. II. Bond and Orbital Electronegativities. *Journal of the American Chemical Society* 85(2):148-154.
259. Sanderson RT (1951) An Interpretation of Bond Lengths and a Classification of Bonds. *Science* 114(2973):670-672.
260. Sanderson RT (1960) *Chemical periodicity* (Reinhold Pub. Corp., New York,) p 330 p.
261. Coulson CA & Moffitt WE (1947) Strain in Non-Tetrahedral Carbon Atoms. *The Journal of Chemical Physics* 15(3):151.
262. Guillen MD & Gasteiger J (1983) Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring-Systems. *Tetrahedron* 39(8):1331-1335.
263. Hinze J (1985) Citation Classic - Electronegativity .1. Orbital Electronegativity of Neutral Atoms. *Cc/Phys Chem Earth* (28):20-20.
264. Gasteiger J & Saller H (1985) Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angewandte Chemie International Edition in English* 24(8):687-689.

265. Gasteiger J (1979) A Representation of π Systems for Efficient Computer Manipulation. *Journal of chemical information and computer sciences* 19(2):111-115.

266. Bauerschmidt S & Gasteiger J (1997) Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species. *Journal of Chemical Information and Computer Sciences* 37(4):705-714.

267. Belitz HD*, et al.* (1979) Sweet and Bitter Compounds: Structure and Taste Relationship. *Food Taste Chemistry,* ACS Symposium Series, (AMERICAN CHEMICAL SOCIETY), Vol 115, pp 93-131.

268. Le Fèvre RJW (1965) Molecular Refractivity and Polarizability. *Advances in Physical Organic Chemistry*, ed Gold V (Academic Press), Vol Volume 3, pp 1-90.

269. Brauman JI & Blair LK (1968) Gas-phase acidities of alcohols. Effects of alkyl groups. *Journal of the American Chemical Society* 90(23):6561-6562.

270. Gasteiger J & Hutchings MG (1984) Quantitative Models of Gas-Phase Proton-Transfer Reactions Involving Alcohols, Ethers, and Their Thio Analogs - Correlation Analyses Based on Residual Electronegativity and Effective Polarizability. *Journal of the American Chemical Society* 106(22):6489-6495.

271. Glen RC (1994) A Fast Empirical-Method for the Calculation of Molecular Polarizability. *J Comput Aided Mol Des* 8(4):457-466.

272. Miller KJ & Savchik J (1979) A new empirical method to calculate average molecular polarizabilities. *Journal of the American Chemical Society* 101(24):7206-7213.

273. Gasteiger J & Hutchings MG (1983) New Empirical-Models of Substituent Polarizability and Their Application to Stabilization Effects in Positively Charged Species. *Tetrahedron Letters* 24(25):2537-2540.

274. Slater JC (1930) Atomic Shielding Constants. *Physical Review* 36(1):57-64.

275. Leo A, Hansch C, & Elkins D (1971) Partition Coefficients and Their Uses. *Chem Rev* 71(6):525-+.

276. Hansch C, Bjorkroth JP, & Leo A (1987) Hydrophobicity and central nervous system agents: on the principle of minimal hydrophobicity in drug design. *Journal of pharmaceutical sciences* 76(9):663-687.

277. Hansch C, Maloney PP, Fujita T, & Muir RM (1962) Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* 194(4824):178-180.

278. Rekker RF & Mannhold R (1992) *Calculation of drug lipophilicity : the hydrophobic fragmental constant approach* (VCH, Weinheim ; New York) p 112 p.

279. Kellogg GE, Semus SF, & Abraham DJ (1991) Hint - a New Method of Empirical Hydrophobic Field Calculation for Comfa. *J Comput Aided Mol Des* 5(6):545-552.

280. Meng EC, Kuntz ID, Abraham DJ, & Kellogg GE (1994) Evaluating Docked Complexes with the Hint Exponential Function and Empirical Atomic Hydrophobicities. *J Comput Aided Mol Des* 8(3):299-306.

281. Wang RX, Fu Y, & Lai LH (1997) A new atom-additive method for calculating partition coefficients. *Journal of chemical information and computer sciences* 37(3):615-621.

282. Wang RX, Gao Y, & Lai LH (2000) Calculating partition coefficient by atom-additive method. *Perspect Drug Discov* 19(1):47-66.

283. Xing L & Glen RC (2002) Novel methods for the prediction of logP, pK(a), and logD. *Journal of chemical information and computer sciences* 42(4):796-805.

284. Bajorath J (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of chemical information and computer sciences* 41(2):233-245.

285. Bajorath J (2002) Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 1(11):882-894.
286. Auer J & Bajorath J (2008) Molecular similarity concepts and search calculations. *Methods in molecular biology* 453:327-347.
287. Hutter MC (2011) Graph-based similarity concepts in virtual screening. *Future Med Chem* 3(4):485-501.
288. Barnard JM & Downs GM (1997) Chemical fragment generation and clustering software. *Journal of chemical information and computer sciences* 37(1):141-142.
289. Ihlenfeldt WD & Gasteiger J (1994) Hash codes for the identification and classification of molecular structure elements. *Journal of computational chemistry* 15(8):793-813.
290. Durant JL, Leland BA, Henry DR, & Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 42(6):1273-1280.
291. McGregor MJ & Pallai PV (1997) Clustering of large databases of compounds: Using the MDL ''keys'' as structural descriptors. *Journal of chemical information and computer sciences* 37(3):443-448.
292. Trinajstić N (1992) *Chemical graph theory* (CRC Press, Boca Raton) 2nd Ed p 322 p.
293. Devillers J & Balaban AT (1999) *Topological indices and related descriptors in QSAR and QSPR* (Gordon and Breach, Amsterdam) pp x, 811 p.
294. Bertz SH (1983) On the Complexity of Graphs and Molecules. *B Math Biol* 45(5):849-855.
295. Randic M & Basak SC (2001) Characterization of DNA primary sequences based on the average distances between bases. *Journal of chemical information and computer sciences* 41(3):561-568.
296. Galvez J, Garciadomenech R, Dejulianortiz JV, & Soler R (1995) Topological Approach to Drug Design. *Journal of chemical information and computer sciences* 35(2):272-284.
297. Galvez J, Garcia R, Salabert MT, & Soler R (1994) Charge Indexes. New Topological Descriptors. *Journal of chemical information and computer sciences* 34(3):520-525.
298. Galvez J, Garcia-Domenech RY, & de Julian-Oritz JV (1998) Design of new antineoplastic lead drugs by molecular topology. *Expert Opinion on Therapeutic Targets* 2(2):265-268.
299. Galvez J, Garciadomenech R, Dejulianortiz V, & Soler R (1994) Topological Approach to Analgesia. *Journal of chemical information and computer sciences* 34(5):1198-1203.
300. Moreau G & Broto P (1980) The Auto-Correlation of a Topological-Structure - a New Molecular Descriptor. *Nouv J Chim* 4(6):359-360.
301. Kubinyi H, Folkers G, & Martin YC (1998) *3D QSAR in drug design* (Kluwer Academic, Dordrecht ; Boston, Mass) pp v. < 2- >.
302. Broto P, Moreau G, & Vandycke C (1984) Molecular-Structures - Perception, Auto-Correlation Descriptor and Sar Studies - Perception of Molecules - Topological-Structure and 3-Dimensional Structure. *Eur J Med Chem* 19(1):61-65.
303. Pastor M, Cruciani G, McLay I, Pickett S, & Clementi S (2000) GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 43(17):3233-3243.
304. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry* 28(7):849-857.
305. Fedorov VV (1972) *Theory of optimal experiments* (Academic Press, New York,) pp xi, 292 p.
306. Duran A, Martinez GC, & Pastor M (2008) Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in Molecular Interaction Fields. *J Chem Inf Model* 48(9):1813-1823.

307. Duran A, Zamora I, & Pastor M (2009) Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening. *J Chem Inf Model* 49(9):2129-2138.
308. Pastor M (2006) Alignment-independent Descriptors from Molecular Interaction Fields. *Molecular Interaction Fields*, (Wiley-VCH Verlag GmbH & Co. KGaA), pp 117-143.
309. Klebe G, Abraham U, & Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 37(24):4130-4146.
310. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* 11(23-24):1046-1053.
311. Osborne CK & Schiff R (2005) Estrogen-receptor biology: continuing progress and therapeutic implications. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 23(8):1616-1622.
312. Hall JM, Couse JF, & Korach KS (2001) The multifaceted mechanisms of estradiol and estrogen receptor signaling. *The Journal of biological chemistry* 276(40):36869-36872.
313. Revankar CM, Cimino DF, Sklar LA, Arterburn JB, & Prossnitz ER (2005) A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* 307(5715):1625-1630.
314. Bologa CG*, et al.* (2006) Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat Chem Biol* 2(4):207-212.
315. Klarlund JK*, et al.* (1997) Signaling by phosphoinositide-3,4,5-trisphosphate through proteins containing pleckstrin and Sec7 homology domains. *Science* 275(5308):1927-1930.
316. Fuss B, Becker T, Zinke I, & Hoch M (2006) The cytohesin Steppke is essential for insulin signalling in Drosophila. *Nature* 444(7121):945-948.
317. Ogasawara M*, et al.* (2000) Similarities in function and gene structure of cytohesin-4 and cytohesin-1, guanine nucleotide-exchange proteins for ADP-ribosylation factors. *The Journal of biological chemistry* 275(5):3221-3230.
318. Kliche S*, et al.* (2001) Signaling by human herpesvirus 8 kaposin A through direct membrane recruitment of cytohesin-1. *Molecular cell* 7(4):833-843.
319. Perez OD*, et al.* (2003) Leukocyte functional antigen 1 lowers T cell activation thresholds and signaling through cytohesin-1 and Jun-activating binding protein 1. *Nature immunology* 4(11):1083-1092.
320. Hafner M*, et al.* (2006) Inhibition of cytohesins by SecinH3 leads to hepatic insulin resistance. *Nature* 444(7121):941-944.
321. Stumpfe D*, et al.* (2010) Targeting Multifunctional Proteins by Virtual Screening: Structurally Diverse Cytohesin Inhibitors with Differentiated Biological Functions. *Acs Chem Biol* 5(9):839-849.
322. Huguenard JR & Prince DA (1992) A novel T-type current underlies prolonged Ca(2+)-dependent burst firing in GABAergic neurons of rat thalamic reticular nucleus. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 12(10):3804-3817.
323. Perez-Reyes E (2003) Molecular physiology of low-voltage-activated T-type calcium channels. *Physiol Rev* 83(1):117-161.
324. Bourinet E & Zamponi GW (2005) Voltage gated calcium channels as targets for analgesics. *Current topics in medicinal chemistry* 5(6):539-546.
325. MOE (2011) Molecular Operating Environment (MOE) (Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7), 2011.10.
326. Ijjaali I, Barrere C, Nargeot J, Petitet F, & Bourinet E (2007) Ligand-based virtual screening to identify new T-type calcium channel blockers. *Channels (Austin)* 1(4):300-304.

327. Keiser MJ*, et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462(7270):175-181.
328. Keiser MJ*, et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25(2):197-206.
329. Daylight Chemical Information Systems I (2013) Daylight Theory Manual.
330. Schuffenhauer A*, et al.* (2002) An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* 42(4):947-955.
331. Olah M*, et al.* (2005) WOMBAT: World of Molecular Bioactivity. *Chemoinformatics in Drug Discovery*, (Wiley-VCH Verlag GmbH & Co. KGaA), pp 221-239.
332. Lounkine E*, et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486(7403):361-367.
333. Zhang S (2011) Computer-aided drug discovery and development. *Methods in molecular biology* 716:23-38.
334. Hansch C (1982) Citation Classic - Rho-Sigma-Pi-Analysis - a Method for the Correlation of Biological-Activity and Chemical-Structure. *Cc/Life Sci* (47):18-18.
335. Free SM, Jr. & Wilson JW (1964) A Mathematical Contribution to Structure-Activity Studies. *J Med Chem* 7:395-399.
336. Tmej C*, et al.* (1998) A combined Hansch/Free-Wilson approach as predictive tool in QSAR studies on propafenone-type modulators of multidrug resistance. *Archiv der Pharmazie* 331(7-8):233-240.
337. Hopfinger AJ*, et al.* (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J Am Chem Soc* 119(43):10509-10524.
338. Vedani A & Dobler M (2002) 5D-QSAR: the key for simulating induced fit? *J Med Chem* 45(11):2139-2149.
339. Pan D, Tseng Y, & Hopfinger AJ (2003) Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J Chem Inf Comput Sci* 43(5):1591-1607.
340. Wold S, Esbensen K, & Geladi P (1987) Principal Component Analysis. *Chemometr Intell Lab* 2(1-3):37-52.
341. Kubinyi H (1997) QSAR and 3D QSAR in drug design .1. methodology. *Drug Discov Today* 2(11):457-467.
342. Zheng W & Tropsha A (2000) Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *Journal of chemical information and computer sciences* 40(1):185-194.
343. Livingstone D (2008) *Artificial neural networks : methods and applications* (Humana Press, Totowa, NJ) pp ix, 254 p.
344. Vapnik V & Lerner A (1963) Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control* 24.
345. Boser BE, Guyon IM, & Vapnik VN (1992) A training algorithm for optimal margin classifiers. in *Proceedings of the fifth annual workshop on Computational learning theory* (ACM, Pittsburgh, Pennsylvania, United States), pp 144-152.
346. Blumer A, Ehrenfeucht A, Haussler D, & Warmuth MK (1989) Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* 36(4):929-965.
347. Vapnik VN (1999) An overview of statistical learning theory. *Neural Networks, IEEE Transactions on* 10(5):988-999.
348. Ivanciuc O (2007) Applications of Support Vector Machines in Chemistry. *Reviews in Computational Chemistry*, (John Wiley & Sons, Inc.), pp 291-400.

349. Liang Y (2011) *Support vector machines and their application in chemistry and biotechnology* (CRC Press, Boca Raton) pp x, 201 p.
350. Boyle BH (2011) *Support vector machines : data analysis, machine learning, and applications* (Nova Science Publishers, New York) pp x, 202 p.
351. Cristianini N & Shawe-Taylor J (2000) *An introduction to support vector machines : and other kernel-based learning methods* (Cambridge University Press, Cambridge ; New York) pp xiii, 189.
352. Vapnik VN (2006) *Estimation of dependences based on empirical data ; Empirical inference science : afterword of 2006* (Springer, New York, N.Y.) 2nd enl. Ed p 505 p.
353. Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, (MIT Press), pp 185-208.
354. Zhan Y & Shen D (2005) Design efficient support vector machine for fast classification. *Pattern Recognition* 38(1):157-161.
355. Han J & Kamber M (2006) *Data mining : concepts and techniques* (Elsevier; Morgan Kaufmann, Amsterdam ; Boston; San Francisco, CA) 2nd Ed pp xxviii, 770 p.
356. Mitchell TM (1997) *Machine Learning* (McGraw-Hill, New York) pp xvii, 414 p.
357. Fukunishi Y (2009) Structure-Based Drug Screening and Ligand-Based Drug Screening with Machine Learning. *Comb Chem High T Scr* 12(4):397-408.
358. Quinlan JR (1993) *C4.5 : programs for machine learning* (Morgan Kaufmann Publishers, San Mateo, Calif.) pp x, 302 p.
359. Briganti S, Camera E, & Picardo M (2003) Chemical and instrumental approaches to treat hyperpigmentation. *Pigm Cell Res* 16(2):101-110.
360. Sanchezferrer A, Rodriguezlopez JN, Garciacanovas F, & Garciacarmona F (1995) Tyrosinase - a Comprehensive Review of Its Mechanism. *Bba-Protein Struct M* 1247(1):1-11.
361. Xu Y*, et al.* (1997) Tyrosinase mRNA is expressed in human substantia nigra. *Brain research. Molecular brain research* 45(1):159-162.
362. Casanola-Martin GM*, et al.* (2007) Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *Eur J Med Chem* 42(11-12):1370-1381.
363. Gasparini F, Bilbe G, Gomez-Mancilla B, & Spooren W (2008) mGluR5 antagonists: discovery, characterization and drug development. *Current opinion in drug discovery & development* 11(5):655-665.
364. Conn PJ, Christopoulos A, & Lindsley CW (2009) Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature reviews. Drug discovery* 8(1):41-54.
365. Mueller R*, et al.* (2012) Discovery of 2-(2-benzoxazoyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu(5)): from an artificial neural network virtual screen to an in vivo tool compound. *ChemMedChem* 7(3):406-414.
366. Mueller R*, et al.* (2010) Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem Neurosci* 1(4):288-305.
367. Rodriguez AL*, et al.* (2010) Discovery of novel allosteric modulators of metabotropic glutamate receptor subtype 5 reveals chemical and functional diversity and in vivo activity in rat behavioral models of anxiolytic and antipsychotic activity. *Molecular pharmacology* 78(6):1105-1123.
368. Olsen CM, Childs DS, Stanwood GD, & Winder DG (2010) Operant sensation seeking requires metabotropic glutamate receptor 5 (mGluR5). *PloS one* 5(11):e15085.
369. Deacon RM (2006) Digging and marble burying in mice: simple methods for in vivo identification of biological impacts. *Nature protocols* 1(1):122-124.

370. Brown N, McKay B, Gilardoni F, & Gasteiger J (2004) A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *Journal of chemical information and computer sciences* 44(3):1079-1087.

371. Feher M, Gao Y, Baber JC, Shirley WA, & Saunders J (2008) The use of ligand-based de novo design for scaffold hopping and sidechain optimization: two case studies. *Bioorganic & medicinal chemistry* 16(1):422-427.

372. Golla S*, et al.* (2012) Virtual design of chemical penetration enhancers for transdermal drug delivery. *Chem Biol Drug Des* 79(4):478-487.

373. Baumforth KR*, et al.* (2005) Induction of autotaxin by the Epstein-Barr virus promotes the growth and survival of Hodgkin lymphoma cells. *Blood* 106(6):2138-2146.

374. Euer N*, et al.* (2002) Identification of genes associated with metastasis of mammary carcinoma in metastatic versus non-metastatic cell lines. *Anticancer research* 22(2A):733-740.

375. Kawagoe H, Stracke ML, Nakamura H, & Sano K (1997) Expression and transcriptional regulation of the PD-Ialpha/autotaxin gene in neuroblastoma. *Cancer research* 57(12):2516-2521.

376. Boucher J*, et al.* (2005) Potential involvement of adipocyte insulin resistance in obesity-associated up-regulation of adipocyte lysophospholipase D/autotaxin expression. *Diabetologia* 48(3):569-577.

377. Umemura K*, et al.* (2006) Autotaxin expression is enhanced in frontal cortex of Alzheimer-type dementia patients. *Neuroscience letters* 400(1-2):97-100.

378. Inoue M, Ma L, Aoki J, & Ueda H (2008) Simultaneous stimulation of spinal NK1 and NMDA receptors produces LPC which undergoes ATX-mediated conversion to LPA, an initiator of neuropathic pain. *Journal of neurochemistry* 107(6):1556-1565.

379. Hoeglund AB*, et al.* (2010) Optimization of a pipemidic acid autotaxin inhibitor. *J Med Chem* 53(3):1056-1066.

380. Ke YY*, et al.* (2013) 3D-QSAR-assisted drug design: identification of a potent quinazoline-based Aurora kinase inhibitor. *ChemMedChem* 8(1):136-148.

381. Coumar MS*, et al.* (2009) Structure-based drug design of novel Aurora kinase A inhibitors: structural basis for potency and specificity. *J Med Chem* 52(4):1050-1062.

382. Coumar MS*, et al.* (2010) Fast-forwarding hit to lead: aurora and epidermal growth factor receptor kinase inhibitor lead identification. *J Med Chem* 53(13):4980-4988.

383. Coumar MS*, et al.* (2010) Identification, SAR studies, and X-ray co-crystallographic analysis of a novel furanopyrimidine aurora kinase A inhibitor. *ChemMedChem* 5(2):255-267.

384. Li M*, et al.* (2010) Aurora kinase inhibitor ZM447439 induces apoptosis via mitochondrial pathways. *Biochem Pharmacol* 79(2):122-129.

385. Fu J, Bian M, Jiang Q, & Zhang C (2007) Roles of Aurora kinases in mitosis and tumorigenesis. *Mol Cancer Res* 5(1):1-10.

386. Agnese V*, et al.* (2007) The role of Aurora-A inhibitors in cancer therapy. *Ann Oncol* 18 Suppl 6:vi47-52.

387. Chai HF*, et al.* (2011) Identification of novel 5-hydroxy-1H-indole-3-carboxylates with anti-HBV activities based on 3D QSAR studies. *J Mol Model* 17(8):1831-1840.

388. Chai HF, Zhao YF, Zhao CS, & Gong P (2006) Synthesis and in vitro anti-hepatitis B virus activities of some ethyl 6-bromo-5-hydroxy-1H-indole-3-carboxylates. *Bioorganic & Medicinal Chemistry* 14(4):911-917.

389. Zhao CS, Zhao YF, Chai HF, & Gong P (2006) Synthesis and in vitro anti-hepatitis B virus activities of some ethyl 5-hydroxy-1H-indole-3-carboxylates. *Bioorganic & Medicinal Chemistry* 14(8):2552-2558.

390. Jiao ZG, *et al.* (2010) Design, synthesis and anti-HIV integrase evaluation of N-(5-chloro-8-hydroxy-2-styrylquinolin-7-yl)benzenesulfonamide derivatives. *Molecules* 15(3):1903-1917.
391. Kurup A (2003) C-QSAR: a database of 18,000 QSARs and associated biological and physical data. *J Comput Aided Mol Des* 17(2-4):187-196.
392. Kim KH (2007) Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? *J Comput Aided Mol Des* 21(8):421-435.
393. KENT JT (1983) Information gain and a general measure of correlation. *Biometrika* 70(1):163-173.
394. Mao KZ (2004) Orthogonal forward selection and backward elimination algorithms for feature subset selection. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 34(1):629-634.
395. Goodarzi M, Freitas MP, & Jensen R (2009) Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. *Journal of chemical information and modeling* 49(4):824-832.
396. Davis L (1991) *Handbook of genetic algorithms* (Van Nostrand Reinhold, New York) pp xii, 385 p.
397. Zhou Y, Lai X, Li Y, & Dong W (2012) Ant Colony Optimization With Combining Gaussian Eliminations for Matrix Multiplication. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*.
398. Lv J, Wang Y, Zhu L, & Ma Y (2012) Particle-swarm structure prediction on clusters. *The Journal of chemical physics* 137(8):084104.
399. Wermuth CG (2006) Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist. *Pharmacophores and Pharmacophore Searches*, (Wiley-VCH Verlag GmbH & Co. KGaA), pp 1-13.
400. Wolber G, Seidel T, Bendix F, & Langer T (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today* 13(1-2):23-29.
401. Smellie A, Teig SL, & Towbin P (1995) Poling - Promoting Conformational Variation. *Journal of computational chemistry* 16(2):171-187.
402. Poptodorov K, Luu T, & Hoffmann RD (2006) Pharmacophore Model Generation Software Tools. *Pharmacophores and Pharmacophore Searches*, (Wiley-VCH Verlag GmbH & Co. KGaA), pp 15-47.
403. Marshall GR, Barry CD, Bosshard HE, Dammkoehler RA, & Dunn DA (1979) Conformational Parameter in Drug Design - Active Analog Approach. *Abstr Pap Am Chem S* (Apr):29-29.
404. Dixon SL, *et al.* (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des* 20(10-11):647-671.
405. Kurogi Y & Guner OF (2001) Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Current medicinal chemistry* 8(9):1035-1055.
406. Anonymous (2002) Catalyst (Accelrys Inc.; San Diego, CA), 2002.
407. Anonymous (2012) LigandScout - advanced structure-based pharmacophore modeling.
408. Martin YC, *et al.* (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comput Aided Mol Des* 7(1):83-102.
409. Jones G, Willett P, & Glen RC (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comput Aided Mol Des* 9(6):532-549.
410. Güner OF (2000) *Pharmacophore perception, development, and use in drug design* (International University Line, LaJolla, CA) pp xiii, 537 p., xx p. of col. plates.

411. Chang C & Swaan PW (2006) Computational approaches to modeling drug transporters. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences* 27(5):411-424.
412. Patel Y, Gillet VJ, Bravi G, & Leach AR (2002) A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J Comput Aided Mol Des* 16(8-9):653-681.
413. Prodromou C & Pearl LH (2003) Structure and functional relationships of Hsp90. *Current cancer drug targets* 3(5):301-323.
414. Solit DB & Rosen N (2006) Hsp90: a novel target for cancer therapy. *Current topics in medicinal chemistry* 6(11):1205-1214.
415. Chiosis G, Rodina A, & Moulick K (2006) Emerging Hsp90 inhibitors: from discovery to clinic. *Anti-cancer agents in medicinal chemistry* 6(1):1-8.
416. Al-Sha'er MA & Taha MO (2010) Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90alpha inhibitors. *J Chem Inf Model* 50(9):1706-1723.
417. Poirier D (2009) Advances in development of inhibitors of 17beta hydroxysteroid dehydrogenases. *Anti-cancer agents in medicinal chemistry* 9(6):642-660.
418. Schuster D*, et al.* (2011) Identification of chemically diverse, novel inhibitors of 17beta-hydroxysteroid dehydrogenase type 3 and 5 by pharmacophore-based virtual screening. *J Steroid Biochem Mol Biol* 125(1-2):148-161.
419. Noha SM*, et al.* (2011) Discovery of a novel IKK-beta inhibitor by ligand-based virtual screening techniques. *Bioorganic & medicinal chemistry letters* 21(1):577-583.
420. Valiron O, Caudron N, & Job D (2001) Microtubule dynamics. *Cellular and molecular life sciences : CMLS* 58(14):2069-2084.
421. Chiang YK*, et al.* (2009) Generation of ligand-based pharmacophore model and virtual screening for identification of novel tubulin inhibitors with potent anticancer activity. *J Med Chem* 52(14):4221-4233.
422. Doddareddy MR*, et al.* (2007) 3D pharmacophore based virtual screening of T-type calcium channel blockers. *Bioorganic & medicinal chemistry* 15(2):1091-1105.
423. Annoura H*, et al.* (2002) Synthesis and biological evaluation of new 4-arylpiperidines and 4-aryl-4-piperidinols: dual Na(+) and Ca(2+) channel blockers with reduced affinity for dopamine D(2) receptors. *Bioorganic & medicinal chemistry* 10(2):371-383.
424. Lanier MC*, et al.* (2007) Selection, synthesis, and structure-activity relationship of tetrahydropyrido[4,3-d]pyrimidine-2,4-diones as human GnRH receptor antagonists. *Bioorganic & medicinal chemistry* 15(16):5590-5603.
425. Cheng KW & Leung PC (2000) The expression, regulation and signal transduction pathways of the mammalian gonadotropin-releasing hormone receptor. *Canadian journal of physiology and pharmacology* 78(12):1029-1052.
426. Huirne JA & Lambalk CB (2001) Gonadotropin-releasing-hormone-receptor antagonists. *Lancet* 358(9295):1793-1803.
427. Roche O & Rodriguez Sarmiento RM (2007) A new class of histamine H3 receptor antagonists derived from ligand based design. *Bioorganic & medicinal chemistry letters* 17(13):3670-3675.
428. Hough LB (2001) Genomics meets histamine receptors: New subtypes, new receptors. *Molecular pharmacology* 59(3):415-419.
429. Alguacil LF & Perez-Garcia C (2003) Histamine H3 receptor: a potential drug target for the treatment of central nervous system disorders. *Current drug targets. CNS and neurological disorders* 2(5):303-313.

430. Witkin JM & Nelson DL (2004) Selective histamine H3 receptor antagonists for treatment of cognitive deficiencies and other disorders of the central nervous system. *Pharmacology & therapeutics* 103(1):1-20.

431. Hancock AA & Brune ME (2005) Assessment of pharmacology and potential anti-obesity properties of H3 receptor antagonists/inverse agonists. *Expert opinion on investigational drugs* 14(3):223-241.

432. Stahl M*, et al.* (2002) A validation study on the practical use of automated de novo design. *J Comput Aided Mol Des* 16(7):459-478.

433. Howells LM, Gallacher-Horley B, Houghton CE, Manson MM, & Hudson EA (2002) Indole-3-carbinol inhibits protein kinase B/Akt and induces apoptosis in the human breast tumor cell line MDA MB468 but not in the nontumorigenic HBL100 line. *Molecular cancer therapeutics* 1(13):1161-1172.

434. Li Y, Chinni SR, & Sarkar FH (2005) Selective growth regulatory and pro-apoptotic effects of DIM is mediated by AKT and NF-kappaB pathways in prostate cancer cells. *Frontiers in bioscience : a journal and virtual library* 10:236-243.

435. Chao WR, Yean D, Amin K, Green C, & Jong L (2007) Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling. *Journal of medicinal chemistry* 50(15):3412-3415.

436. Reid JM*, et al.* (2011) Phase 0 clinical chemoprevention trial of the Akt inhibitor SR13668. *Cancer Prev Res (Phila)* 4(3):347-353.

437. Dayam R*, et al.* (2008) Quinolone 3-carboxylic acid pharmacophore: design of second generation HIV-1 integrase inhibitors. *Journal of medicinal chemistry* 51(5):1136-1144.

438. Palmisano L (2007) Role of integrase inhibitors in the treatment of HIV disease. *Expert review of anti-infective therapy* 5(1):67-75.

439. Gordon CP, Griffith R, & Keller PA (2007) Control of HIV through the inhibition of HIV-1 integrase: A medicinal chemistry perspective. *Medicinal chemistry* 3(2):199-220.

440. Sato M*, et al.* (2006) Novel HIV-1 integrase inhibitors derived from quinolone antibiotics. *J Med Chem* 49(5):1506-1508.

441. Mugnaini C*, et al.* (2007) Toward novel HIV-1 integrase binding inhibitors: molecular modeling, synthesis, and biological studies. *Bioorganic & medicinal chemistry letters* 17(19):5370-5373.

442. Noeske T*, et al.* (2007) Virtual screening for selective allosteric mGluR1 antagonists and structure-activity relationship investigations for coumarine derivatives. *ChemMedChem* 2(12):1763-1773.

443. Bordi F & Ugolini A (1999) Group I metabotropic glutamate receptors: implications for brain diseases. *Progress in neurobiology* 59(1):55-79.

444. Spooren W*, et al.* (2003) Insight into the function of Group I and Group II metabotropic glutamate (mGlu) receptors: behavioural characterization and implications for the treatment of CNS disorders. *Behavioural pharmacology* 14(4):257-277.

445. Schneider G, Neidhart W, Giller T, & Schmid G (1999) "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie* 38(19):2894-2896.

446. Lipinski CA, Lombardo F, Dominy BW, & Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* 23(1-3):3-25.

447. Lajiness MS, Vieth M, & Erickson J (2004) Molecular properties that influence oral drug-like behavior. *Current opinion in drug discovery & development* 7(4):470-477.

448. Hou T, Wang J, Zhang W, Wang W, & Xu X (2006) Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Current medicinal chemistry* 13(22):2653-2667.
449. Bickerton GR, Paolini GV, Besnard J, Muresan S, & Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nat Chem* 4(2):90-98.
450. Frimurer TM, Bywater R, Naerum L, Lauritsen LN, & Brunak S (2000) Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *Journal of chemical information and computer sciences* 40(6):1315-1324.
451. Veber DF*, et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry* 45(12):2615-2623.
452. Brenk R*, et al.* (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* 3(3):435-444.
453. Hann MM, Leach AR, & Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences* 41(3):856-864.
454. Kulkarni A, Han Y, & Hopfinger AJ (2002) Predicting caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *Journal of chemical information and computer sciences* 42(2):331-342.
455. Vistoli G, Pedretti A, & Testa B (2008) Assessing drug-likeness--what are we missing? *Drug discovery today* 13(7-8):285-294.
456. Ursu O, Rayan A, Goldblum A, & Oprea TI (2011) Understanding drug-likeness. *Wires Comput Mol Sci* 1(5):760-781.
457. Schnecke V & Bostrom J (2006) Computational chemistry-driven decision making in lead generation. *Drug Discov Today* 11(1-2):43-50.
458. Sun H & Scott DO (2010) Structure-based drug metabolism predictions for drug design. *Chem Biol Drug Des* 75(1):3-17.
459. Goldstein A, Aronow L, & Kalman SM (1973) *Principles of drug action; the basis of pharmacology* (Wiley, New York,) 2d Ed pp xv, 854 p.
460. Ortiz de Montellano PR (2005) *Cytochrome P450 : structure, mechanism, and biochemistry* (Kluwer Academic/Plenum Publishers, New York) 3rd Ed pp xx, 689 p.
461. Armour D*, et al.* (2006) The discovery of CCR5 receptor antagonists for the treatment of HIV infection: Hit-to-lead studies. *ChemMedChem* 1(7):706-+.
462. Ahlstrom MM, Ridderstrom M, & Zamora I (2007) CYP2C9 structure-metabolism relationships: substrates, inhibitors, and metabolites. *J Med Chem* 50(22):5382-5391.
463. Ahlstrom MM, Ridderstrom M, Zamora I, & Luthman K (2007) CYP2C9 structure-metabolism relationships: optimizing the metabolic stability of COX-2 inhibitors. *J Med Chem* 50(18):4444-4452.
464. Xue Y*, et al.* (2007) Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism. *Bioorganic & medicinal chemistry* 15(5):2156-2166.
465. Yano JK*, et al.* (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution. *The Journal of biological chemistry* 279(37):38091-38094.
466. Udomsinprasert R*, et al.* (2005) Identification, characterization and structure of a new Delta class glutathione transferase isoenzyme. *The Biochemical journal* 388(Pt 3):763-771.
467. Aller SG*, et al.* (2009) Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* 323(5922):1718-1722.

468. de Graaf C, Pospisil P, Pos W, Folkers G, & Vermeulen NP (2005) Binding mode prediction of cytochrome p450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J Med Chem* 48(7):2308-2318.
469. Erve JC, Svensson MA, von Euler-Chelpin H, & Klasson-Wehler E (2004) Characterization of glutathione conjugates of the remoxipride hydroquinone metabolite NCQ-344 formed in vitro and detection following oxidation by human neutrophils. *Chemical research in toxicology* 17(4):564-571.
470. Sun H*, et al.* (2009) Differences in CYP3A4 catalyzed bioactivation of 5-aminooxindole and 5-aminobenzsultam scaffolds in proline-rich tyrosine kinase 2 (PYK2) inhibitors: retrospective analysis by CYP3A4 molecular docking, quantum chemical calculations and glutathione adduct detection using linear ion trap/orbitrap mass spectrometry. *Bioorganic & medicinal chemistry letters* 19(12):3177-3182.
471. Park JY & Harris D (2003) Construction and assessment of models of CYP2E1: predictions of metabolism from docking, molecular dynamics, and density functional theoretical calculations. *J Med Chem* 46(9):1645-1660.
472. Bazeley PS, Prithivi S, Struble CA, Povinelli RJ, & Sem DS (2006) Synergistic use of compound properties and docking scores in neural network modeling of CYP2D6 binding: predicting affinity and conformational sampling. *J Chem Inf Model* 46(6):2698-2708.
473. accelrys (2013) Accelrys Metabolite.
474. Talafous J, Sayre LM, Mieyal JJ, & Klopman G (1994) META. 2. A dictionary model of mammalian xenobiotic metabolism. *J Chem Inf Comput Sci* 34(6):1326-1333.
475. Shaik S, de Visser SP, Ogliaro F, Schwarz H, & Schroder D (2002) Two-state reactivity mechanisms of hydroxylation and epoxidation by cytochrome P-450 revealed by theory. *Curr Opin Chem Biol* 6(5):556-567.
476. Singh SB, Shen LQ, Walker MJ, & Sheridan RP (2003) A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J Med Chem* 46(8):1330-1336.
477. Rydberg P, Gloriam DE, Zaretzki J, Breneman C, & Olsen L (2010) SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *Acs Medicinal Chemistry Letters* 1(3):96-100.
478. Rydberg P, Vasanthanathan P, Oostenbrink C, & Olsen L (2009) Fast prediction of cytochrome P450 mediated drug metabolism. *ChemMedChem* 4(12):2070-2079.
479. Norinder U & Haeberlein M (2002) Computational approaches to the prediction of the blood-brain distribution. *Advanced drug delivery reviews* 54(3):291-313.
480. Davis AM & Riley RJ (2004) Predictive ADMET studies, the challenges and the opportunities. *Curr Opin Chem Biol* 8(4):378-386.
481. Mitcheson JS & Perry MD (2003) Molecular determinants of high-affinity drug binding to HERG channels. *Current opinion in drug discovery & development* 6(5):667-674.
482. Sanguinetti MC & Tristani-Firouzi M (2006) hERG potassium channels and cardiac arrhythmia. *Nature* 440(7083):463-469.
483. Recanatini M, Poluzzi E, Masetti M, Cavalli A, & De Ponti F (2005) QT prolongation through hERG K+ channel blockade: Current knowledge and strategies for the early prediction during drug development. *Medicinal research reviews* 25(2):133-166.
484. Hancox JC & Mitcheson JS (2006) Combined hERG channel inhibition and disruption of trafficking in drug-induced long QT syndrome by fluoxetine: a case-study in cardiac safety pharmacology. *British Journal of Pharmacology* 149(5):457-459.
485. Keating MT & Sanguinetti MC (1996) Molecular genetic insights into cardiovascular disease. *Science* 272(5262):681-685.

486. Aronov AM (2005) Predictive in silico modeling for hERG channel blockers. *Drug discovery today* 10(2):149-155.
487. Garg D, Gandhi T, & Gopi Mohan C (2008) Exploring QSTR and toxicophore of hERG K+ channel blockers using GFA and HypoGen techniques. *Journal of molecular graphics & modelling* 26(6):966-976.
488. Ficker E, Obejero-Paz CA, Zhao S, & Brown AM (2002) The binding site for channel blockers that rescue misprocessed human long QT syndrome type 2 ether-a-gogo-related gene (HERG) mutations. *Journal of Biological Chemistry* 277(7):4989-4998.
489. Choe H*, et al.* (2006) A novel hypothesis for the binding mode of HERG channel blockers. *Biochem Bioph Res Co* 344(1):72-78.
490. Polak S, Wisniowska B, & Brandys J (2009) Collation, assessment and analysis of literature in vitro data on hERG receptor blocking potency for subsequent modeling of drugs' cardiotoxic properties. *Journal of applied toxicology : JAT* 29(3):183-206.
491. Wood C, Williams C, & Waldron GJ (2004) Patch clamping by numbers. *Drug discovery today* 9(10):434-441.
492. Bridgland-Taylor MH*, et al.* (2006) Optimisation and validation of a medium-throughput electrophysiology-based hERG assay using IonWorks HT. *Journal of pharmacological and toxicological methods* 54(2):189-199.
493. Thai KM & Ecker GF (2007) Predictive models for HERG channel blockers: ligand-based and structure-based approaches. *Current medicinal chemistry* 14(28):3003-3026.
494. Xi B, Chandak GR, Shen Y, Wang Q, & Zhou D (2012) Association between common polymorphism near the MC4R gene and obesity risk: a systematic review and meta-analysis. *PloS one* 7(9):e45731.
495. Ekins S, Crumb WJ, Sarazan RD, Wikel JH, & Wrighton SA (2002) Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *The Journal of pharmacology and experimental therapeutics* 301(2):427-434.
496. Cianchetta G*, et al.* (2005) Predictive models for hERG potassium channel blockers. *Bioorganic & medicinal chemistry letters* 15(15):3637-3642.
497. Cavalli A, Poluzzi E, De Ponti F, & Recanatini M (2002) Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers. *Journal of medicinal chemistry* 45(18):3844-3853.
498. Sun HM (2006) An accurate and interpretable Bayesian classification model for prediction of hERG liability. *ChemMedChem* 1(3):315-322.
499. Song MH & Clark M (2006) Development and evaluation of an in silico model for hERG binding. *Journal of chemical information and modeling* 46(1):392-400.
500. Wang S*, et al.* (2012) ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Molecular pharmaceutics* 9(4):996-1010.
501. Rogers D & Hahn M (2010) Extended-Connectivity Fingerprints. *Journal of chemical information and modeling* 50(5):742-754.
502. Obrezanova O & Segall MD (2010) Gaussian processes for classification: QSAR modeling of ADMET and target activity. *Journal of chemical information and modeling* 50(6):1053-1061.
503. Rasmussen CE & Williams CKI (2006) *Gaussian processes for machine learning* (MIT Press, Cambridge, Mass.) pp xviii, 248 p.
504. Gibbs MN & MacKay DJC (2000) Variational Gaussian process classifiers. *Ieee T Neural Networ* 11(6):1458-1464.

505. Obrezanova O, Gola JM, Champness EJ, & Segall MD (2008) Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J Comput Aided Mol Des* 22(6-7):431-440.

506. Nisius B & Goller AH (2009) Similarity-based classifier using topomers to provide a knowledge base for hERG channel inhibition. *Journal of chemical information and modeling* 49(2):247-256.

507. Cramer RD, Jilek RJ, & Andrews KM (2002) Dbtop: topomer similarity searching of conventional structure databases. *Journal of molecular graphics & modelling* 20(6):447-462.

508. Jilek RJ & Cramer RD (2004) Topomers: A validated protocol for their self-consistent generation. *Journal of chemical information and computer sciences* 44(4):1221-1227.

509. Rogers D & Hopfinger AJ (1994) Application of Genetic Function Approximation to Quantitative Structure-Activity-Relationships and Quantitative Structure-Property Relationships. *Journal of chemical information and computer sciences* 34(4):854-866.

510. Pintore M, de Waterbeemd HV, Piclin N, & Chretien JR (2003) Prediction of oral bioavailability by adaptive fuzzy partitioning. *Eur J Med Chem* 38(4):427-431.

511. Yoshida F & Topliss JG (2000) QSAR model for drug human oral bioavailability. *J Med Chem* 43(13):2575-2585.

512. Turner JV, Glass BD, & Agatonovic-Kustrin S (2003) Prediction of drug bioavailability based on molecular structure. *Anal Chim Acta* 485(1):89-102.

513. Ekins S*, et al.* (2002) Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol Pharmacol* 61(5):974-981.

514. Lombardo F, Obach RS, Shalaeva MY, & Gao F (2004) Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J Med Chem* 47(5):1242-1250.

515. de Groot MJ & Ekins S (2002) Pharmacophore modeling of cytochromes P450. *Advanced drug delivery reviews* 54(3):367-383.

516. Lewis DF (2003) Quantitative structure-activity relationships (QSARs) within the cytochrome P450 system: QSARs describing substrate binding, inhibition and induction of P450s. *Inflammopharmacology* 11(1):43-73.

517. Jorgensen WL & Duffy EM (2002) Prediction of drug solubility from structure. *Advanced drug delivery reviews* 54(3):355-366.

518. Cruciani G*, et al.* (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* 48(22):6970-6979.

519. Rajman I (2008) PK/PD modelling and simulations: utility in drug development. *Drug Discov Today* 13(7-8):341-346.

520. Holford N, Ma SC, & Ploeger BA (2010) Clinical trial simulation: a review. *Clin Pharmacol Ther* 88(2):166-182.

521. Laer S*, et al.* (2005) Development of a safe and effective pediatric dosing regimen for sotalol based on population pharmacokinetics and pharmacodynamics in children with supraventricular tachycardia. *Journal of the American College of Cardiology* 46(7):1322-1330.

522. Yim DS*, et al.* (2005) Population pharmacokinetic analysis and simulation of the time-concentration profile of etanercept in pediatric patients with juvenile rheumatoid arthritis. *Journal of clinical pharmacology* 45(3):246-256.

523. Jamei M*, et al.* (2009) The Simcyp population-based ADME simulator. *Expert opinion on drug metabolism & toxicology* 5(2):211-223.

524. Kowalski KG, Olson S, Remmers AE, & Hutmacher MM (2008) Modeling and simulation to support dose selection and clinical development of SC-75416, a selective COX-2 inhibitor for the treatment of acute and chronic pain. *Clin Pharmacol Ther* 83(6):857-866.

525. Kruger DM & Evers A (2010) Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* 5(1):148-158.

526. Rush TS, 3rd, Grant JA, Mosyak L, & Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of medicinal chemistry* 48(5):1489-1495.

527. Warren GL*, et al.* (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49(20):5912-5931.

528. Page CS & Bates PA (2006) Can MM-PBSA calculations predict the specificities of protein kinase inhibitors? *Journal of computational chemistry* 27(16):1990-2007.

529. Plewczynski D, Lazniewski M, von Grotthuss M, Rychlewski L, & Ginalski K (2011) VoteDock: consensus docking method for prediction of protein-ligand interactions. *Journal of computational chemistry* 32(4):568-581.

530. Bar-Haim S, Aharon A, Ben-Moshe T, Marantz Y, & Senderowitz H (2009) SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization. *J Chem Inf Model* 49(3):623-633.

531. Fukunishi H, Teramoto R, Takada T, & Shimada J (2008) Bootstrap-based consensus scoring method for protein-ligand docking. *Journal of chemical information and modeling* 48(5):988-996.

532. Teramoto R & Fukunishi H (2008) Consensus scoring with feature selection for structure-based virtual screening. *Journal of chemical information and modeling* 48(2):288-295.

533. Ripphausen P, Nisius B, Peltason L, & Bajorath J (2010) Quo vadis, virtual screening? A comprehensive survey of prospective applications. *Journal of medicinal chemistry* 53(24):8461-8467.

534. Nicolotti O, Miscioscia TF, Carotti A, Leonetti F, & Carotti A (2008) An integrated approach to ligand- and structure-based drug design: development and application to a series of serine protease inhibitors. *J Chem Inf Model* 48(6):1211-1226.

535. Baroni M*, et al.* (1993) Generating Optimal Linear Pls Estimations (Golpe) - an Advanced Chemometric Tool for Handling 3d-Qsar Problems. *Quantitative Structure-Activity Relationships* 12(1):9-20.

536. Vedani A, Dobler M, Spreafico M, Peristera O, & Smiesko M (2007) VirtualToxLab - in silico prediction of the toxic potential of drugs and environmental chemicals: evaluation status and internet access protocol. *Altex* 24(3):153-161.

537. Wilson GL & Lill MA (2011) Integrating structure-based and ligand-based approaches for computational drug design. *Future medicinal chemistry* 3(6):735-750.

538. Klon AE, Glick M, Thoma M, Acklin P, & Davies JW (2004) Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *Journal of medicinal chemistry* 47(11):2743-2749.

539. Khatib F*, et al.* (2011) Algorithm discovery by protein folding game players. *P Natl Acad Sci USA* 108(47):18949-18953.

540. Khatib F*, et al.* (2012) Crystal structure of a monomeric retroviral protease solved by protein folding game players (vol 18, pg 1175, 2011). *Nat Struct Mol Biol* 19(3):365-365.

541. Eiben CB*, et al.* (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol* 30(2):190-192.

Chapter 2

BCL::EMAS — Enantioselective Molecular Asymmetry Descriptor for 3D-QSAR

---

published in Molecules (2012) **17**, 9971-9989

Gregory Sliwoski, Edward Will Lowe, Mariusz Butkiewicz, Jens Meiler: **BCL::EMAS – enantioselective molecular asymmetry descriptor for 3D-QSAR**

2.1 Abstract

Stereochemistry is an important determinant of a molecule's biological activity. Stereoisomers can have different degrees of efficacy or even opposing effects when interacting with a target protein. Stereochemistry is a molecular property difficult to represent in 2D-QSAR as it is an inherently three-dimensional phenomenon. A major drawback of most proposed descriptors for 3D-QSAR that encode stereochemistry is that they require a heuristic for defining all stereocenters and rank-ordering its substituents. Here we propose a novel 3D-QSAR descriptor termed Enantioselective Molecular ASymmetry (EMAS) that is capable of distinguishing between enantiomers in the absence of such heuristics. The descriptor aims to measure the deviation from an overall symmetric shape of the molecule. A radial-distribution function (RDF) determines a signed volume of tetrahedrons of all triplets of atoms and the molecule center. The descriptor can be enriched with atom-centric properties such as partial charge. This descriptor showed good predictability when tested with a dataset of thirty-one steroids commonly used to benchmark stereochemistry descriptors (r2 = 0.89, q2 = 0.78). Additionally, EMAS improved enrichment of 4.38 versus 3.94 without EMAS in a simulated virtual high-throughput screening (vHTS) for inhibitors and substrates of cytochrome P450 (PUBCHEM AID891).

2.2 Introduction

Stereoisomers are defined as different molecular species of equal constitution which are separated by energy barriers [1]. For organic molecules stereochemistry is most frequently caused by carbon atoms with four different substituents. However, other stereocenters exist such as positively charged nitrogen atoms with four different substituents, double bonds with different substituents on each of the two carbon atoms, stereoisomeric allenes, atropisomeric biphenyls, *etc*. Enantiomers are a subset of stereoisomers that are defined as non-superimposable mirror images (*enantios* being Greek for opposite and *meros* for part). Despite their structural similarities, enantiomers can display very different pharmacological profiles. Stereoisomers that are not enantiomers are called diastereomers. Stereoselectivity is widely prevalent in nature as most proteins are formed from the genetically encoded L-amino acids making small molecule binding pockets enantioselective [2]. In drug discovery, there are examples in which different enantiomers show different efficacies, e.g., dexrabeprazole [3] and beta blockers [4], and different toxicities, e.g., levobupivacaine [5]. In 1992, the FDA issued a statement requiring that the development of any racemate (mixture of a compound's stereoisomers) carry a justification for the inclusion of both isomers [6] and in the year 2000, chiral drugs accounted for over $100 billion in sales [7]. Between 1985 and 2004, the number of single enantiomer drugs as a percentage of chiral molecules increased from 31.6% to 89.8% [8].

Given the importance of stereoselectivity in drug design, it is necessary that any computational approach to drug discovery distinguishes between stereoisomers. In Structure-Based Computer-Aided Drug Discovery (SB-CADD) stereochemistry is explicitly accounted for as the molecule is docked into a structural model of the protein binding site. The 3D structure of the molecule in complex with the protein is evaluated taking its stereochemistry into account. In complex with the target protein even enantiomers turn into diastereomers and can be distinguished. In Ligand-Based Computer-Aided Drug Discovery (LB-CADD) the chemical structures of active compounds are compared to derive common features that determine activity. The task of distinguishing stereoisomers and in particular enantiomers becomes more challenging as stereochemistry needs to be defined in the absence of the protein. This is impossible in 2D molecular descriptors where only the constitution of a molecule is taken into account. Therefore extensions to 2D molecular descriptors have been developed—sometimes described as 2.5D descriptors—that describe configuration and can therefore define stereochemistry. Lastly, 3D descriptors based on the molecular conformation can define stereochemistry, if appropriately designed.

The IUPAC convention for distinguishing stereoisomers is the Cahn-Ingold-Prelog (CIP) convention distinguishing *R* (*rectus*) and *S* (*sinister*) configuration of stereocenters. It requires a priority weighting system for the different substituents that is incapable of dealing with some complex scenarios. Extensions to the CIP system have been introduced to handle situations in which the chiral center did not rest on an atom (chirality plane or axis) and for stereoisomers which do not possess centers of chirality at all (stereisomeric allenes, atropisomeric biphenyls, and ansa-compounds) [1]. Further complications arise for pseudoasymmetric stereogenic units, defined as pairs of enantiomorphic ligands together with two ligands which are non-enantiomorphic. In cases such as these, the priorities of two substituents depend on their own chiral centers. One particular disadvantage is that the CIP nomenclature does not always follow chemical intuition. For example, take the two molecules $HC(CH_3)(OH)F$ and $HC(CH_3)(SH)F$. Naively we would align these close derivatives by superimposing H with H, CH_3 with CH_3, OH with SH and F with F. This assigns *R*-$HC(CH_3)(OH)F$ to *S*-$HC(CH_3)(SH)F$ and *vice versa*. In fact, closely related derivatives that place similar functional groups in the same regions of space and are likely to have similar activity can have opposite CIP assignment. Therefore, the CIP convention is not suitable to describe stereochemistry effectively for LB-CADD.

Extensions to 2D-QSAR have been proposed to distinguish enantiomers. Golbraikh and co-workers introduced a series of chirality descriptors that use an additional term called the chirality correction added to the vertex degrees of asymmetric atoms in a molecular graph [9]. This method is similar to one proposed by Yang and Zhong [10] where the chiral index was instead appended to the substituents attached to the chiral center. Multiple similar algorithms have also been proposed [11-14]. For example, Brown, *et al* [11] added chirality to their graph kernel method. The drawbacks of these methods include their reliance on the problematic R/S designations as well as the combination of spatial and atom property information such that their indices become a principally mathematical concept with little interpretation on physical terms.

Another approach proposed by Benigni and co-workers [15] describes a chirality measure based on the comparison of the 3D structure for a molecule with all others in a data set. Zabrodsky [16] proposed a similar continuous symmetry measure which quantifies the minimal distance movement for points of an object in order to transform it into a shape of desired symmetry. However, these molecular similarity indices are very sensitive to relative orientation and depend on pairwise molecular indices which can complicate QSAR-based high throughput screening.

Aires-de-Sousa, *et al* [17-19] introduced a 3D-QSAR method for handling enantiomers. Classical 3D-QSAR descriptors such as radial distribution functions are incapable of distinguishing between enantiomers based on their nature. This method employs an RDF-like function that utilizes a ranking system for each chiral center introduced by Zhang and Aires-de-Sousa that reinterpreted the CIP rules in terms of more meaningful physicochemical properties. Additionally, it has the benefit of being a vector rather than single value which is equal and opposite for enantiomer pairs. However, this method requires the identification and appropriate labeling of all stereogenic units and suffers from the fact that spatial information is combined with atom properties where some physical interpretability is lost. It is also worth mentioning that it is not clear if it is pharmacologically relevant to specify every stereogenic component of a molecule, but rather if different profiles between enantiomers depend on specific chiral centers and/or an overall chirality of the molecule as a whole.

CoMFA [20] is an appealing method for distinguishing between enantiomers as it avoids the necessity to identify stereogenic centers. Rather, it intrinsically takes chirality into account as the molecular fields of chiral isomers are inherently different. However, the method relies on superimposition of all molecules[9] which is difficult to achieve for large or diverse substance libraries.

Here we propose a novel enantioselective 3D descriptor for QSAR that is similar to the RDF-like function proposed by Aires-de-Sousa and co-workers but with important differences to address the concerns raised above. We call this new method Enantioselective Molecular ASymmetry (EMAS). Our method does not rely on any priority ranking or distinction of every stereogenic unit, thereby eliminating the need to combine spatial and atomic properties and bypassing the difficulties that arise in non-conventional chiral centers. Rather, the enantiomeric distinctions “emerge” from the spatial distribution of atoms within the molecule. Additionally, EMAS is designed to avoid a rigid distinction between enantiomers but rather to represent the overall asymmetry of a molecule as it compares to other similar molecules as well as its own enantiomorphs. Therefore, EMAS intends to describe overall molecular asymmetry while including a directionality component that can distinguish between enantiomers.

2.3 Results and Discussion

Enantiomorphism is Determined by Asymmetry in Shape or Property Distribution

Enantiomorphism in small molecules is impacted by two phenomena. The first factor is the shape of the molecule—*i.e.*, the distribution of its atom coordinates in space. If the mirror image of this shape cannot be superimposed with the original version, the two molecules are enantiomers. Beyond the overall shape the distribution of properties plays a role. We can envision molecules that have a (near) perfect symmetric shape. Image and mirror image will be identical shape wise. However, distribution of partial charge, polarizability, and electronegativity can be enantiomorphic. While both contributions are coupled they represent two dimensions of one phenomenon. For a specific molecule one of the other factors might be more pronounced. For example steroids can have enantiomorph shapes but have relatively uniform property distributions as they are dominated by apolar CH groups. On the other hand, the molecule CFClBrI is an almost perfect regular tetrahedron with a highly enantiomorph distribution of partial charge and polarizability. As both contributions can determine properties and activities of small molecules, stereochemical descriptors should capture and ideally distinguish both contributions.

Radial Distribution Functions Separate Shape Information and Property Distribution

Radial Distribution Functions (RDFs) are often applied in 3D-QSAR [21, 22]. As a means of comparison, the general form of the atomic radial distribution function is shown:

$$f(r) = \sum_{i}^{n} \sum_{j}^{n-1} P_i P_j e^{-\beta\left(r-r_{ij}\right)^2} \tag{1}$$

In this equation, β is a smoothing parameter, often called the 'temperature' while r_{ij} is the distance between atoms i and j, n is the total number of atoms in the molecule, and r is the running variable for the function $f(r)$. Often, such equations are 'weighted' with a property coefficient for both atoms P_iP_j. The function plots shape (i.e., distance between two atoms) on the x-axis, the respective property coefficient on the y-axis thereby separating geometry from property distribution. With $P_iP_j = 1$ this function is a representation of the overall shape of the molecule based on the frequencies of all atom pair distances within each radial distance step. As distances are invariant to mirroring, enantiomers share identical RDF functions. Note that diastereomers have distinct RDFs as not all atom pair distances are identical.

Expanding RDFs to 'Signed' Volumes that Are Sensitive to Shape Enantiomorphy

We first look for the simplest geometric form that would be sensitive to mirroring. This shape would be a tetrahedron. We choose tetrahedrons consisting of all combinations of three atoms i, j, k and the center of the molecule. Other approaches use all permutations of four atoms. The present approach reduces the computational demand. The geometric property plotted for the tetrahedron is volume. c_i, c_j, and c_k are the coordinates of the three atoms. The center of the molecule is defined by point o. Then, we compute the signed volume as:

$$signed\ volume = \frac{1}{6}\left(\overrightarrow{c_i c_j} \times \overrightarrow{c_i c_k}\right) \cdot \overrightarrow{o c_i} \tag{2}$$

While the absolute term always reflects volume, it is important to note that the result can have a positive or negative sign, depending on the order of points which is initially arbitrary. We note that the volume has an arbitrary sign that inverts when the molecule is converted into its mirror image. We note further that the volume becomes 0 if the plane defined by c_i, c_j, and c_k includes o. This property is beneficial as a planar arrangement of atoms cannot be enantiomorphic. However, for a tetrahedron to contribute to enantiomorphy, its edges $\left\|\overrightarrow{c_i c_j}\right\|$, $\|\overrightarrow{c_i c_k}\|$, and $\left\|\overrightarrow{c_j c_k}\right\|$ must be of different length. This property is captured by a stereochemistry score:

$$stereochemistry = \frac{\left(\left\|\overrightarrow{c_i c_j}\right\| - \|\overrightarrow{c_i c_k}\|\right) * \left(\|\overrightarrow{c_i c_k}\| - \left\|\overrightarrow{c_j c_k}\right\|\right) * \left(\left\|\overrightarrow{c_j c_k}\right\| - \left\|\overrightarrow{c_i c_j}\right\|\right)}{0.0962243 * \max\left(\left\|\overrightarrow{c_i c_j}\right\|, \|\overrightarrow{c_i c_k}\|, \left\|\overrightarrow{c_j c_k}\right\|\right)^{3}} \tag{3}$$

Two things emerge from the numerator: the asymmetry is evaluated based on the variation in distances between the three atoms. If any two distances are equal, the triangle formed from the three atom coordinates will contain perfect symmetry and the score will be 0. Additionally, the directional (enantiomorphic) information emerges based on the order of distances. For example, if $\left\|\overrightarrow{c_i c_j}\right\| > \|\overrightarrow{c_i c_k}\| > \left\|\overrightarrow{c_j c_k}\right\|$, then this product will have a negative sign $(+) * (+) * (-)$. However, if, from the vantage point of the molecular center, the order of distances has been shuffled (as would be seen in an enantiomer $\|\overrightarrow{c_i c_k}\| > \left\|\overrightarrow{c_i c_j}\right\| > \left\|\overrightarrow{c_j c_k}\right\|$), the sign changes as well $(-) * (+) * (-)$. Figure 2.1 demonstrates how opposite directions emerge depending on the ordering of instances. Recall that by allowing a signed volume, we ensure that the order of distances does not rely on the order of atoms coordinates encountered, but rather as the order of distances seen from the molecular center in terms of the cross product's direction. The score is normalized by a constant factor of 0.0962243 which is calculated as the maximum possible score when the largest of the three distances is 1. Details can be found in the supplementary information. Figure 2.2 compares atom triplets that give rise to high versus low scores as well as scores with opposite directions.

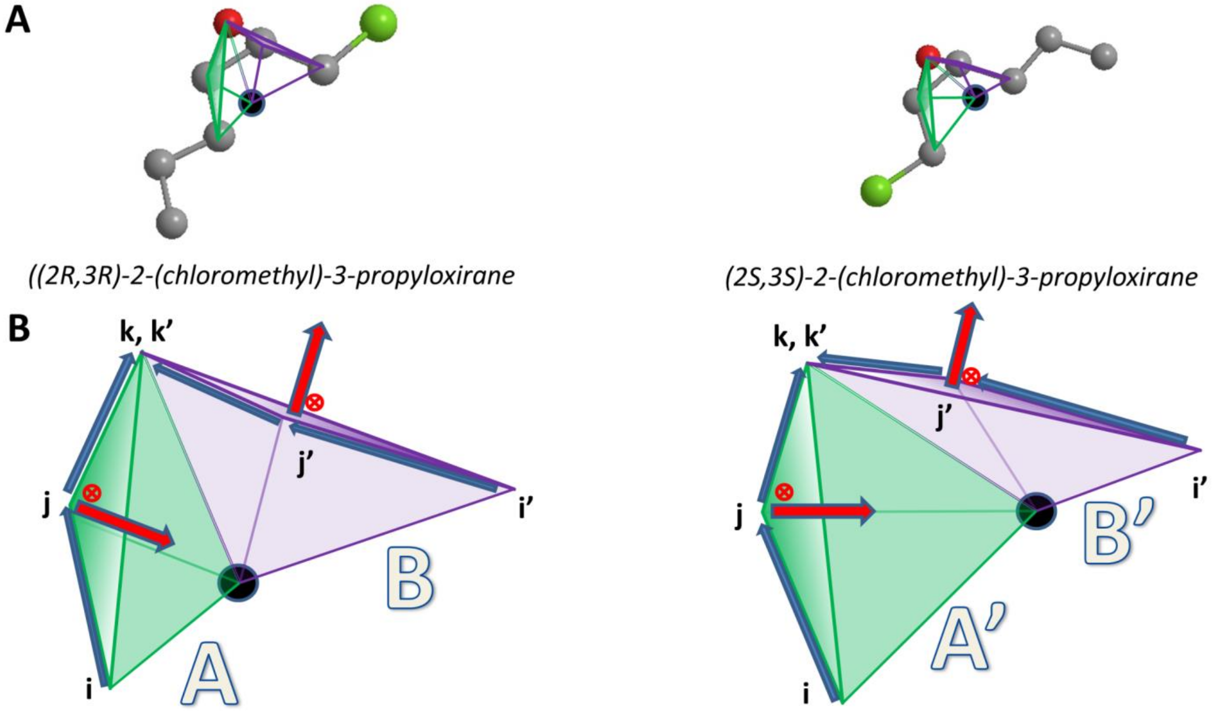
The final directional asymmetry score (DAS) of any given atom triplet becomes:

$$DAS = \sqrt[3]{signed\ volume_{ijk} * stereochemistry_{ijk}} \tag{4}$$

Note that the products cube-root has been taken to achieve a dimension of distance resembling a common RDF. This procedure preserves the sign and expands the range of frequently occurring low-scoring triplets at the cost of rare triplets with high scores. Substituting this directional asymmetry in place of atom distance, the EMAS function becomes:

$$EMAS(r) = \sum_{i}^{n} \sum_{j}^{n-1} \sum_{k}^{n-2} sign(DAS) \times e^{-\beta\left(r\ -\left|DAS_{ijk}\right|\right)^{2}} \tag{5}$$

where β is the smoothing parameter, n is the total number of non-hydrogen atoms, and r is the running variable of the function $EMAS(r)$. The alternate sign preceding the exponential function transfers the "directionality" of the score to the overall function so that at any given score, the intensity reflects the subtraction of negative (one direction) from positive (opposite direction). Figure 2.3 maps the EMAS plot for epothilone B and its mirror image.


A
((2R,3R)-2-(chloromethyl)-3-propyloxirane
(2S,3S)-2-(chloromethyl)-3-propyloxirane
B
k, k'
j'
i'
j
i
A
B
k, k'
j'
i'
j
i
A'
B'

C

	A	A'	B	B'
Signed Volume	+ 0.1931	- 0.1931	- 0.4168	+ 0.4168
Stereochemistry	- 0.0715	- 0.0715	- 0.0715	- 0.0715
DAS	- 0.0138	+ 0.0138	+ 0.0298	- 0.0298

Figure 2.1 Calculating DAS (A) Scores reflect opposing enantiomorphs based on cross-product direction and geometric center. Enantiomers [(2*R*,3*R*)-2-(chloromethyl)-3-propyloxirane and (2*S*,3*S*)-2-(chloromethyl)-3-propyloxirane] with two stereocenters are shown. (B) Two triangles are visualized in both enantiomers. These triangles encompass the same triplets of atoms between the two molecules. Four tetramers formed by the atom triplets and molecular center are visualized. i, j, k, and i', j', k' reflect the order of these atoms in either molecule. Importance of atom ordering is shown based on the direction of cross product (red arrow) and location of molecular center (black circle). (C) Volume and score calculations for the four tetrahedrons across both enantiomers are shown. Note the opposite signs and scores between the two enantiomers' tetrahedrons.

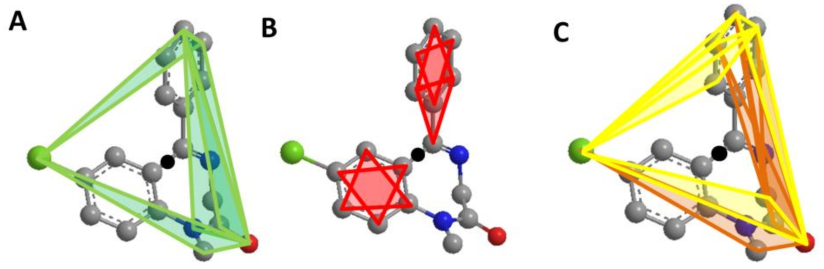


Figure 2.2 Encoding diazepam: atom triplets (A) Top five scoring atom triplets in diazepam are shown. The black circle in all figures represents the molecular center. (B) Lowest five scoring atom triplets in diazepam. All triplets shown here score 0 and do not contribute to the RDF-like code. (C) Top five positive and top five negative scoring triplets in diazepam. Yellow: positive; orange: negative.

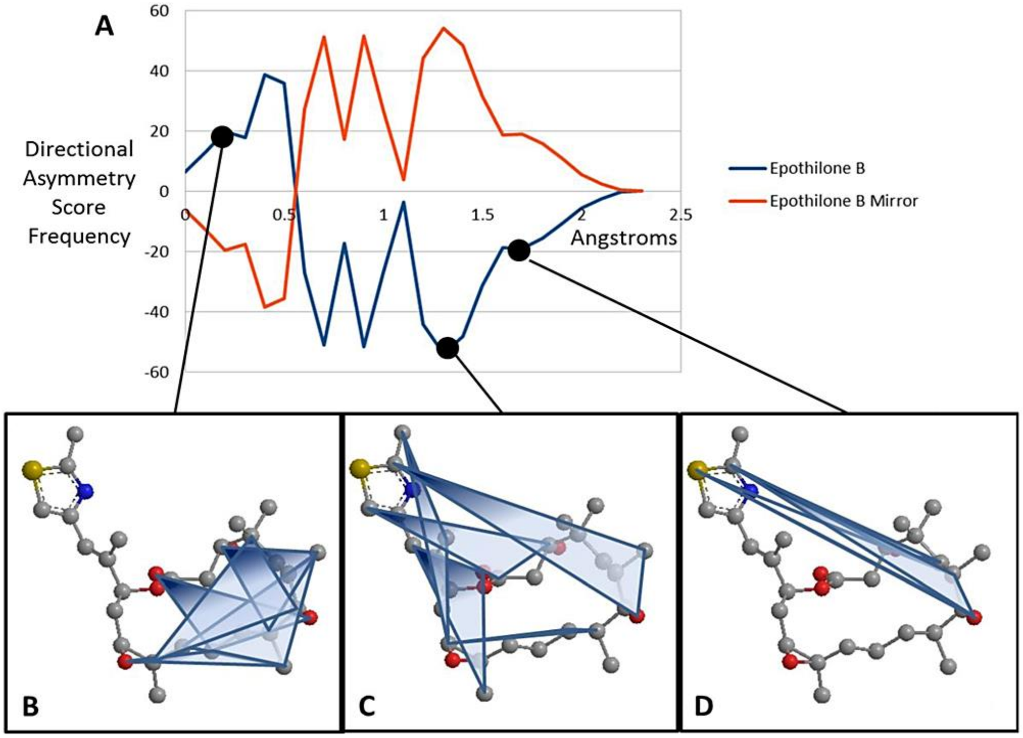


Figure 2.3 EMAS curves for epothilone B (A) EMAS curves for epothilone B (blue) compared with its mirror image (red). X-axis represents the Directional Asymmetry Score in angstroms while the y-axis indicates the frequency of these scores across the entire molecule. (B) Atom triplets with a directional asymmetry score of approximately 0.3 angstroms. (C) Atom triplets with a directional asymmetry score of approximately 1.3 angstroms. (D) Atom triplets with a directional asymmetry score of approximately 1.7 angstroms.

As with the basic radial distribution function, the absence of any weighting coefficient results in a descriptor that encodes only spatial information. While this is important information in and of itself, the addition of a property weighting coefficient increases the utility of this descriptor. Since we are iterating over all atom triplets, the possibility that one atom property can throw off two other atom properties in unintended ways made it problematic in some cases to simply multiply the three atom properties together. Adding the properties, on the other hand, can circumvent this issue but two atom properties of equal magnitude and opposite signs can cancel each other out. Therefore, we retained the functionality for both property coefficient methods and suggest that any use of this descriptor in larger datasets test either method since one may outperform the other depending on the dataset.

Evaluation of EMAS as a Novel Descriptor

Predictability Benchmarking: Cramer's Steroids

A commonly used dataset for evaluating the predictive capability of novel stereochemistry-based descriptors was introduced by Cramer *et al*. in 1988 [20] and several structures were corrected in a subsequent publication [23]. These thirty-one steroid structures are accompanied with their experimental binding affinities to human corticosteroid-binding globulins (CGB) and provide a small dataset containing many stereocenters. Additionally, the rigidity of these compounds makes them an ideal benchmark set for 3D-QSAR algorithms eliminating the factor of conformational flexibility. Since EMAS can be employed in three forms: spatial only, property weighting coefficient via summation, and property weighting coefficient via multiplication, we trained three separate artificial neural network (ANN) models using descriptors derived in each of these three methods. To predict binding affinities over the entire dataset, we used a cross-validated leave-one-out approach. To compare the predictive power of our model versus other descriptors that have been tested against the steroid set, we calculated the correlation coefficient r^2 of predicted versus experimental affinities and the "cross-validated r^2" q^2.

As expected, the ANN model generated using no property weighting (solely spatial information) performed the worst of the three, producing a r^2 of 0.78 and a q^2 of 0.60. By weighting with a multiplicative property coefficient, the performance increased considerably, resulting in a r^2 of 0.86 and a q^2 of 0.74. Weighting with the property summation coefficient yielded the best predictions with a r^2 of 0.89 and a q^2 of 0.78.

Since we began with an interest in generating a molecular asymmetry descriptor that could distinguish between enantiomers, we wanted to ensure that the inclusion of directionality increased the information contained in the descriptor. Therefore, we created a version of the descriptor that incorporates just the absolute value of all stereochemistry scores, thereby eliminating all directional information while retaining all other spatial information. We found that by training our model without directional information, the predictive capabilities for the steroid affinities decreased to a r^2 of 0.65 and a q^2 of 0.41, reinforcing our original design to capture stereochemistry. We also compared the model employing EMAS with one created with a traditional RDF. This model performed worse than any of our three methods giving a r^2 of 0.75 and a q^2 of 0.56. Weighting the RDF's with the same properties used to weight EMAS did not produce any significant improvement in the model (data not shown). Cross-validated predictions for all variations of EMAS as well as the experimental affinities can be found in Table 2-1.

Since this dataset is well-established across similar descriptors in the literature, we compared our predictive power to other methods and found that our best q^2 fell at the average q^2 of all of these methods ($0.63 < q^2 < 0.94$). This result is somewhat difficult to interpret for several reasons: (a) different statistical models are utilized; (b) different degrees of cross validation were employed, and (c) our descriptor solely describes stereochemistry and is meant to be complemented by other descriptors (read below). Most of the competing descriptors include more information on molecule size, shape, and property distribution. However, it is important to note that while EMAS does not require any molecular alignment or pre-annotated stereocenters, it is capable of performing well with a dataset that contains a great deal of stereochemistry. Additionally, the inclusion of directional information outperforms a similar implementation lacking directional information as well as the similar RDF descriptor weighted with or without atom properties. For a comparison of our q^2 with other documented tests against Cramer's steroids, see Table 2-2.

Table 2-1 Experimental and predicted binding affinities for the 31 Cramer's steroids using novel stereoselective descriptor to train ANN models. Spatial predictions utilize the novel descriptor without any atom property weighting. Multiply properties utilize the novel descriptor weighted by the product of atom properties. Sum properties utilize the novel descriptor weighted by the sum of atom properties.

Molecule	**Observed CBG affinity (pKa)**	**Predicted [spatial]**	**Predicted [multiply properties]**	**Predicted [sum properties]**	**Predicted [no stereochemistry]**
aldosterone	−6.28	−7.47	−7.31	−7.25	−7.22
androstanediol	−5.00	−5.47	−5.46	−5.33	−5.56
5-androstenediol	−5.00	−5.47	−5.43	−5.36	−5.75
4-androstenedione	−5.76	−5.64	−5.60	−5.79	−6.36
androsterone	−5.61	−5.78	−5.81	−5.55	−5.42
corticosterone	−7.88	−7.30	−7.37	−7.32	−7.34
cortisol	−7.88	−7.63	−7.58	−7.64	−7.33
cortisone	−6.89	−7.22	−6.83	−7.39	−7.07
dehydroepiandrosterone	−5.00	−5.39	−5.13	−5.46	−5.80
11-deoxycorticosterone	−7.65	−7.48	−7.47	−7.50	−6.85
11-deoxycortisol	−7.88	−7.66	−7.53	−7.59	−7.52
dihydrotestosterone	−5.92	−5.38	−5.70	−5.43	−5.96
estradiol	−5.00	−5.40	−5.36	−5.32	−5.21
estriol	−5.00	−5.25	−5.26	−5.43	−6.10
estrone	−5.00	−5.30	−5.21	−5.54	−5.42
etiocholanolone	−5.23	−6.42	−6.44	−6.22	−6.27
pregnenolone	−5.23	−5.30	−5.25	−5.37	−6.37
17a-hydroxypregnenolone	−5.00	−5.20	−5.28	−5.29	−6.65
progesterone	−7.38	−7.17	−7.27	−7.13	−6.46
17a-hydroxyprogesterone	−7.74	−7.42	−7.39	−6.97	−6.70
testosterone	−6.72	−6.08	−6.36	−6.19	−5.94
prednisolone	−7.51	−7.61	−7.36	−7.65	−7.03
cortisolacetat	−7.55	−6.74	−6.90	−7.63	−6.00
4-pregnene-3,11,20-trione	−6.78	−6.40	−6.83	−6.09	−6.46
epicorticosterone	−7.20	−5.98	−6.00	−7.03	−7.15
19-nortestosterone	−6.14	−5.58	−5.86	−5.54	−5.45
16a,17a-dihydroxy-progesterone	−6.25	−7.25	−7.04	−7.46	−7.36
16a-methylprogesterone	−7.12	−6.69	−6.39	−6.78	−6.60
19-norprogesterone	−6.82	−6.01	−6.30	−7.25	−6.19
2a-methylcortisol	−7.69	−6.62	−7.22	−7.68	−6.57
2a-methyl-9a-fluorocortisol	−5.80	−7.56	−6.97	−6.22	−6.74
	r^2	0.78	0.86	0.89	0.65
	q^2	0.60	0.74	0.78	0.42

Table 2-2 Comparison of novel stereoselective descriptor predictability with other published QSAR methods against the Cramer's steroid set. Calculation of q^2 can be found in the methods section. Statistical model generation method is indicated as well as QSAR method employed are indicated for each reference.

QSAR Method	Model Creation	q^2	Reference
Purely Spatial EMAS	Artificial Neural Network	**0.56**	
Property weighted EMAS (product)	Artificial Neural Network	**0.74**	
Property weighted EMAS (sum)	Artificial Neural Network	**0.78**	
Stochastic 3D-chiral linear indices	Multiple Linear Regression	0.87	[13]
Chiral Topological Indices	Stepwise Regression Analysis	0.85	[10]
Chiral Graph Kernels	Support Vector Machine	0.78	[11]
Chirality Correction and Topological Descriptors	K-nearest neighbor	0.83	[9]
Molecular Quantum Similarity Measures	Multilinear Regression	0.84	[24]
Shape and Electrostatic Similarity Matrixes	Non-linear Neural Network	0.94	[25]
Comparative Molecular Moment Analysis	Partial Least Squares (PLS)	0.83	[23]
Comparative Molecular Similarity Indices Analysis	PLS	0.67	[26]
Comparative Molecular Field Analysis	PLS	0.65	[20]
E-state Descriptors	PLS	0.62	[27]
Molecular Electronegativity Distance Vector	Genetic Algorithm PLS	0.78	[28]
Molecular Quantum Similarity Measures	Multilinear Regression and PLS	0.80	[29]

vHTS Utility and Enrichment Benchmarking: PUBMED AID891

We provide the above analysis for comparison. However, realistically the steroid dataset is too small to provide a good benchmark for EMAS as often the number of features (24 features) is in the same order of magnitude as the number of data points (31 molecules). Therefore we tested the descriptor in a virtual high-throughput screening (vHTS) endeavor. For the benchmark dataset, we used the publicly available results of a conformational screen for inhibitors and substrates of cytochrome P450 2D6 (AID 891). This dataset is of moderate size (approximately 10,000 molecules) and contains both active (18%) and inactive (82%) compounds. We employed a forward-feature selection (FFS) analysis that selects optimal descriptors from RDF's, 3D Autocorrelations (3DA), and 2D Autocorrelations (2DA) functions labeled with atom properties including charge, electronegativity, and effective polarizability (see Experimental section). For a complete list of features tested in forward-feature selections, please see supplementary Table S2-1. ANN 3D-QSAR models were trained with and without inclusion of the EMAS descriptors in the list of descriptors for FFS to choose from. Hence the utility of the EMAS descriptor can be evaluated in two ways: (a) are the EMAS descriptors chosen by the FFS procedure? and (b) has the final model that includes EMAS descriptors an increased predictive power? The FFS with the default set of initial features resulted in a best descriptor set of 9 features distributed evenly across RDF’s, 3D Autocorrelations (3DA), and 2D Autocorrelations (2DA). Cross-validated predictions from the ANN model constructed with this feature set produced an enrichment of 3.94 and a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.826.

An identical FFS analysis was performed by combining the default set of features with 34 EMAS features including all three variations of EMAS (spatial, property weighting via sum, and property weighting via product) weighted with the same list of properties used to test RDFs, 3DAs, and 2DAs. The best set of features contained 20 total features distributed across RDF’s, 3DA’s, 2DA’s, number of hydrogen bond donors, and several EMAS features. There were a total of seven EMAS features represented in the best feature set. Therefore, almost one third of the total features in the best feature set generated through this analysis were EMAS features. This set of seven features contained a spatial EMAS weighted by van der Waals surface areas, three EMAS features weighted via the product method and three EMAS features weighted via the sum method. This substantial representation of EMAS in the best feature set suggests that EMAS successfully provides useful information for the model development that may not be represented in any other feature in the original set. Cross-validated predictions from the ANN model constructed from this EMAS-inclusive feature set produced an enrichment of 4.38 and a ROC curve with an area under the curve of 0.837. Positive predictive value (PPV) is a related measure of a model’s predictive capability which tracks predictive precision as more and more positive predictions are made. By comparing the average PPV precision over a range of the fraction of total predictions made (fraction positive predictions, FPP) of interest, it is possible to compare predictive capabilities for two models. Over the FPP range of 0.005 to 0.05, we find that our model trained with the EMAS features performed significantly better than the model trained without EMAS features (0.727 PPV precision compared with 0.651). A paired t-test for the cross-validated models comparing precisions in this FPP range showed that this is a statistically significant improvement ($p < 0.005$) over the analysis completed without EMAS features. For a complete list of the best features determined from both forward feature analyses, please see the supplementary Table S2-2. Comparative ROC and PPV curves from the forward feature analyses for the control set of features and the control set combined with EMAS features are shown in Figure 2.4.

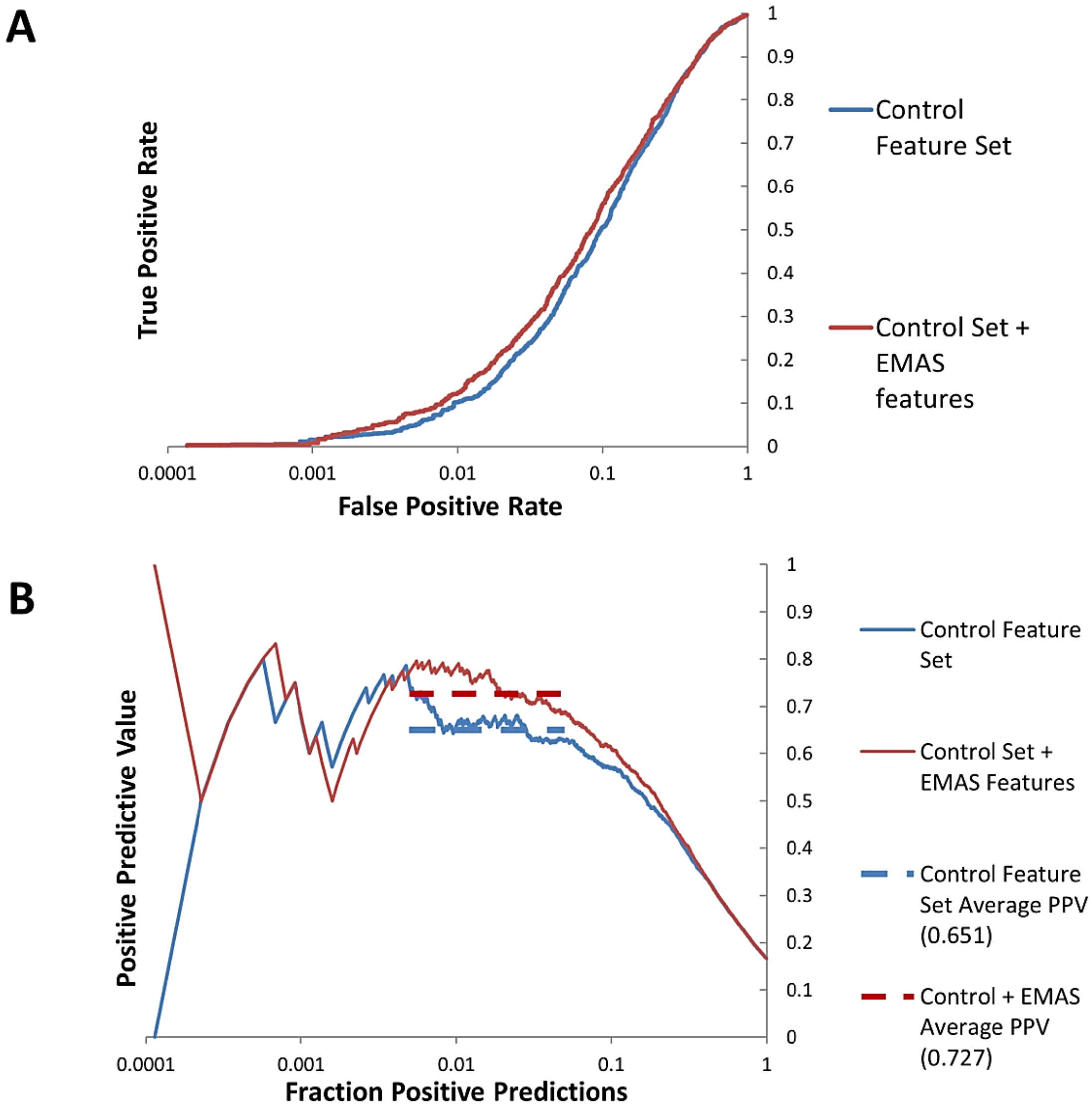


Figure 2.4 ROC and PPV results for the feature forward analysis with the control set of features compared with the control set combined with EMAS features (A) AID891 prediction ROC curves generated from the ANN models trained with the best descriptor set generated from the forward feature analysis beginning with the control set of features combined with the novel EMAS features (red) show improved performance when compared with ROC curves generated from the ANN models trained with the best descriptor set generated from the forward feature analysis beginning with the control set of features (blue) (B) PPV curves for models trained with the best descriptor set of control features combined with the EMAS features (red) shows improved performance over those models trained with the best descriptor set of control features only (blue). Dashed lines of corresponding colors show the average PPV values over the FPP region from which the models were optimized (0.005 to 0.05 fraction positive predicted values).

2.4 Conclusions

The goal of this project was to develop a 3D-QSAR descriptor that was capable of not only distinguishing between enantiomers, but also of describing the overall degree of asymmetry for a molecule. This was accomplished by developing an RDF-like curve that described the distribution of 'directional asymmetry scores (DAS)' rather than inter-atomic distances. The DAS is designed to incorporate information regarding the degree and direction of asymmetry between each atom triplet in the molecule. The degree of asymmetry is calculated as a product of how asymmetrically the three atoms are distributed and the distance they lie from the center of the molecule. This asymmetry is related to the differences between their interatomic distances and the distance from the center of the molecule is related to the volume of the tetrahedron created by the three atom coordinates and the geometric center of the molecule. The direction of asymmetry is related to the distribution of the interatomic distances between these three atom coordinates from the point of view of the center of the molecule. If the sides of the triangle created by these three atoms are different, then identical triangles "pointing" in opposite directions will have a different ordering of sides depending on which direction they "point." This is the key variable that allows the descriptor to distinguish between enantiomers. To exclude any influence that the order in which atoms are listed in the molecule may play on this directionality scheme, we offset this by incorporating the cross product of the two vectors created from the three atoms. This cross product will swap signs when the atoms are ordered differently thereby eliminating the influence of the order of atoms.

We tested the value of this descriptor by training ANN 3D-QSAR models. In order to provide a basis of comparison with other documented QSAR methods that address stereoselectivity, we used a small dataset of steroids that is commonly used as a benchmark for these types of descriptors. We found that the predictability of our descriptor performed comparably with other stereochemistry-based descriptors when evaluated with this set of 31 steroids (r^2 = 0.89, q^2 = 0.78). Additionally, we assessed the utility of the EMAS descriptor by running vHTS experiment on a publically available dataset (PUBCHEM AID 891). A forward-feature selection analysis that determines the most effective set of descriptors for this dataset was employed and the best set of features included several EMAS functions (seven EMAS of 20 total features). This set of features improved the performance of our models over those that were tested without EMAS functions (enrichment of 4.38 when including EMAS versus enrichment of 3.94 without EMAS).

Although our descriptor performs well with the datasets tested, it is still outperformed by several techniques with the steroid dataset. One difficulty with this dataset is that its small size adds significant noise to the results. Additionally, the cross-validation methods used to analyze the performance of these methods vary and are often more forgiving than ours. Future development of EMAS, however, can provide superior predictions even with smaller datasets and extensions to the current implementation of EMAS are being pursued in our lab. Molecular flexibility is one major avenue in which we are improving our implementation. By design, EMAS currently considers single, static conformations when scoring molecules and this may fail to incorporate widely different conformations seen in highly flexible molecules.

We conclude that the EMAS descriptor encodes stereochemistry thereby providing important information that is not captured in other 3D-QSAR descriptors. There are several published QSAR methods that performed better than ours in the steroid dataset but these methods often require some heuristic for describing the stereocenters within each of the molecules or aligning the 3D structures of these molecules. Our descriptor is not subject to either of these limitations and therefore can be extended to broader applications than those previously described.

2.5 Experimental Methods

Generation of Numerical Descriptors for QSAR Model Creation

3D models of all small molecules were generated using the CORINA software package unless already defined. For feature selection analysis, a set of 2,100 numerical descriptors was generated using the BioChemical Library (BCL) software created in our lab. The descriptors can be classified into five categories, including six scalar descriptors (molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, logP, total charge, and topological surface area), 18 2-dimensional auto-correlation functions, 18 3-dimensional autocorrelation functions, 18 radial distribution functions, and 34 novel molecular asymmetry descriptors. These 34 descriptors included spatially-based asymmetry functions with and without van der Waals (VDW) surface area scaling, 16 property-weighted asymmetry functions based on the multiplicative scheme, and 16 property-weighted asymmetry functions based on the additive scheme. These properties included sigma charge [30-32], pi charge [33-35], Vcharge [36], total charge [30-35], sigma electronegativity [30-32], pi electronegativity [33-35], effective polarizability [37-39], and lone pair electronegativity [33-35] with and without VDW surface area scaling. The control comparison forward feature selection analysis was performed with a feature set that included all features listed above except the novel stereochemistry features. This feature set contains 1,284 features. For steroid binding predictions, descriptor sets were created using only one novel stereochemistry method and those including property weighted used the same properties listed.

Training, Monitoring, and Independent Dataset Generation

Cramer's Steroids

The dataset was split for ANN training into three subsets: training, monitoring, and independent. The monitoring dataset is necessary to prevent over-training. Because of the small size of the dataset, only one molecule was labeled independent. Five molecules were used as the monitoring dataset, 25 for training. The set of five molecules was incremented through the entire dataset for a total of 6 different monitoring sets. Leave-one-out cross validation was performed where each molecule was used as the independent molecule while the remaining 30 molecules were used for training and monitoring. The predictions were averaged across the different monitoring sets to yield the final activity predictions for the entire set of 31 molecules.

PUBMED AID891

AID891 is a publically available dataset that can be found at http://pubchem.ncbi.nlm.nih.gov/. It contains 1,623 active compounds and 7,756 inactive compounds tested for inhibition of cytochrome P450 2D6. This dataset was split into 10 clusters distributed into a training set of eight clusters, a monitoring set of one cluster, and an independent set of one cluster. For cross validation, the monitoring and independent datasets are iterated and then the resulting independent predictions are averaged to give the final list of predicted activities that spans the entire dataset. In order to maximize model performance, the dataset was balanced through oversampling. In other words, the active compounds were represented multiple times so that the number of active compounds roughly equals the number of inactive compounds. This method of balancing has been used to maximize QSAR models in other datasets where the number of active compounds is significantly less than the number of inactive compounds [40].

The pIC_{50} values of each compound within AID891 and the steroid binding data for the Cramer dataset were used as output for the ANN models. For the AID891 dataset, inactive compounds were set to a pIC_{50} value of 3. The root-mean-square deviation (RMSD) between predicted and experimental activities was used as the objective function for training the ANN.

Artifical Neural Network (ANN) Architecture and Training

For the AID891 dataset, the ANN was trained using back propagation and a sigmoid transfer function with a simple weight update of eta = 0.1 and alpha = 0.5. The hidden layer contained eight neurons. For the steroid dataset, the ANN was trained using the same protocol as the AID891 dataset but the number of hidden neurons was reduced to 4 due to the smaller size of the dataset.

Forward-Feature Selection for Optimal Descriptor Set Selection

Descriptor selection was performed to test the novel descriptor against all other implemented descriptors to see if it provided an increase to enrichment over any of the other descriptors. The approach begins with a single descriptor, trains a model with only that descriptor, and then continuously adds more descriptors one at a time, training a new model each round. At the completion of each round, the descriptor set that produced the lowest RMSD score was retained for the next round. All descriptors not present in the retained list of descriptors are then added individually to that retained list of descriptors and the descriptor set producing the best RMSD score is retained for the next round, and so on. At the completion of these iterations, the round that produced the best RMSD score overall is recalled as the top descriptor set. If a descriptor appears in this list of best descriptors, then it suggests that significant information had been gleaned from that descriptor during the ANN training.

Model Evaluation

ANN models using the AID891 datasets were analyzed using receiver operating characteristic (ROC) curves to assess their predictive power. These curves plot the rate of true positives versus the rate of false positives as a fraction of the total number of positives. Therefore, a slope of 1 would reflect random guesses as each true positive would be statistically likely to be followed by a false positive. An increase in slope and area under the curve would indicate an increase in predictive power. The initial section of the ROC curve is often most important because it represents compounds with the highest predicted activity. Therefore, enrichment values are determined based on the slope of the ROC curve comprising the first subset of molecules. Increases in enrichment is often the most important measure for application of virtual screening in drug discovery as it reflects the expected factor at which the fraction of actives will be increased over an unbiased dataset.

Positive predictive value (PPV) is a measure related to enrichment which tracks the model's predictive precision as the fraction of predicted positives (FPP) increases from highest predicted activity to lowest. A model is likely to become less precise as the predicted activities approach the cutoff point and therefore it is common to specify a range of FPP of interest when measuring a PPV. FPP is calculated as the number of true positive predictions plus the number of false positive predictions divided by the size of the dataset. PPV is calculated as the number of true positive predictions divided by the total number of positive predictions (true and false positive).

To determine the statistical significance for the average PPV improvement over the FPP range of 0.005 to 0.05, we compared the average PPV within this FPP range for each combination of training and modeling datasets that went into the cross-validated model. By aligning these datasets between the two models, we were able to perform a two-tailed paired t-test to show a significant improvement for the cross validated model including EMAS features over the cross-validated model without EMAS features.

To evaluate the utility of models trained with the steroid dataset in a way which could be comparable with published methods, the conventional correlation coefficient r^2 of the predicted activities against actual activities and cross validated r^2, also known as q^2 were calculated for each descriptor set.

All predicted values used in these analyses were the average predicted activities from each of the leave-one-out models with the different monitoring datasets. The q^2 is calculated from the equation:

$$q^2 = \frac{SD - press}{SD} \tag{6}$$

Here, SD is the sum of squared deviations of each biological property from their mean and $press$ (predictive residual sum of squares) is the sum of the squared differences between the actual biological property and the cross-validated predicted property.

Implementation

The descriptor generation and ANN algorithms were implemented in the BioChemistryLibrary (BCL) version 2.4. The BCL is a C++ library that includes classes to model small molecules as well as larger molecules such as proteins. It contains force-fields, optimization algorithms, and different prediction approaches such as neural networks and support vector machines to model molecular structures, interactions, and properties. This application will be made freely available for academic use at http://www.meilerlab.org/. The training method used is simple propagation, a supervised learning approach. All C++ ANN trainings were performed on a Dell T3500 workstation equipped with 12GB RAM and an Intel(R) Xeon(R) W3570@3.20GHz running 64-bit CentOS 5.2.

Acknowledgements

Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192) and NSF (Career 0742762). Edward W. Lowe Jr. acknowledges NSF support through the CI-TraCS fellowship (OCI-1122919).

2.6 References

1. Prelog V & Helmchen G (1982) Basic Principles of the CIP-System and Proposals for a Revision. *Angewandte Chemie International Edition in English* 21(8):567-583.
2. Schiffman SS, Clark TB, 3rd, & Gagnon J (1982) Influence of chirality of amino acids on the growth of perceived taste intensity with concentration. *Physiology & behavior* 28(3):457-465.
3. Pai V & Pai N (2007) Recent advances in chirally pure proton pump inhibitors. *Journal of the Indian Medical Association* 105(8):469-470, 472, 474.
4. Mehvar R & Brocks DR (2001) Stereospecific pharmacokinetics and pharmacodynamics of beta-adrenergic blockers in humans. *Journal of pharmacy & pharmaceutical sciences : a publication of the Canadian Society for Pharmaceutical Sciences, Societe canadienne des sciences pharmaceutiques* 4(2):185-200.
5. Gurjar MK (2007) The future lies in chiral purity: a perspective. *Journal of the Indian Medical Association* 105(4):177-178.
6. Anonymous (1992) FDA's policy statement for the development of new stereoisomeric drugs. *Chirality* 4(5):338-340.
7. Beroza P & Suto MJ (2000) Designing chiral libraries for drug discovery. *Drug discovery today* 5(8):364-372.
8. Murakami H (2007) From Racemates to Single Enantiomers - Chiral Synthetic Drugs over the last 20 Years. *Topics in current chemistry* 269:273-299.
9. Golbraikh A, Bonchev D, & Tropsha A (2001) Novel chirality descriptors derived from molecular topology. *Journal of chemical information and computer sciences* 41(1):147-158.
10. Yang C & Zhong C (2005) Chirality Factors and Their Application to QSAR Studies of Chiral Molecules. *QSAR & Combinatorial Science* 24(9):1047-1055.
11. BROWN JB*, et al.* (2010) COMPOUND ANALYSIS VIA GRAPH KERNELS INCORPORATING CHIRALITY. *Journal of Bioinformatics and Computational Biology* 08(supp01):63-81.
12. Lukovits I & Linert W (2001) A topological account of chirality. *Journal of chemical information and computer sciences* 41(6):1517-1520.
13. Marrero-Ponce Y & Castillo-Garit J (2005) 3D-chiral Atom, Atom-type, and Total Non-stochastic and Stochastic Molecular Linear Indices and their Applications to Central Chirality Codification. *J Comput Aided Mol Des* 19(6):369-383.
14. Del Rio A (2009) Exploring enantioselective molecular recognition mechanisms with chemoinformatic techniques. *Journal of Separation Science* 32(10):1566-1584.
15. Benigni R*, et al.* (2000) Deriving a quantitative chirality measure from molecular similarity indices. *Journal of medicinal chemistry* 43(20):3699-3703.

16. Zabrodsky H, Peleg S, & Avnir D (1992) Continuous symmetry measures. *Journal of the American Chemical Society* 114(20):7843-7851.
17. Aires-de-Sousa J & Gasteiger J (2001) New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions. *Journal of chemical information and computer sciences* 41(2):369-375.
18. Aires-de-Sousa J & Gasteiger J (2002) Prediction of enantiomeric selectivity in chromatography. Application of conformation-dependent and conformation-independent descriptors of molecular chirality. *Journal of molecular graphics & modelling* 20(5):373-388.
19. Aires-de-Sousa J, Gasteiger J, Gutman I, & Vidovic D (2004) Chirality codes and molecular structure. *Journal of chemical information and computer sciences* 44(3):831-836.
20. Cramer RD, Patterson DE, & Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 110(18):5959-5967.
21. Verma J, Khedkar VM, & Coutinho EC (2010) 3D-QSAR in drug design--a review. *Current topics in medicinal chemistry* 10(1):95-115.
22. Hemmer MC, Steinhauer V, & Gasteiger J (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* 19(1):151-164.
23. Silverman BD (2000) The Thirty-one Benchmark Steroids Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression. *Quantitative Structure-Activity Relationships* 19(3):237-246.
24. Robert D, Amat L, & Carbo-Dorca R (1999) Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *Journal of chemical information and computer sciences* 39(2):333-344.
25. So SS & Karplus M (1997) Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *Journal of medicinal chemistry* 40(26):4347-4359.
26. Klebe G, Abraham U, & Mietzner T (1994) Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *Journal of medicinal chemistry* 37(24):4130-4146.
27. Maw HH & Hall LH (2001) E-state modeling of corticosteroids binding affinity validation of model for small data set. *Journal of chemical information and computer sciences* 41(5):1248-1254.
28. Liu S-S, Yin C-S, & Wang L-S (2002) Combined MEDV-GA-MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX-2 Inhibitors. *Journal of chemical information and computer sciences* 42(3):749-756.
29. Besalu E, Girones X, Amat L, & Carbo-Dorca R (2002) Molecular quantum similarity and the fundamentals of QSAR. *Accounts of chemical research* 35(5):289-295.
30. Gasteiger J & Marsili M (1978) A new model for calculating atomic charges in molecules. *Tetrahedron Letters* 19(34):3181-3184.
31. Gasteiger J & Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 36(22):3219-3228.
32. Guillen MD & Gasteiger J (1983) Extension of the method of iterative partial equalization of orbital electronegativity to small ring systems. *Tetrahedron* 39(8):1331-1335.
33. Bauerschmidt S & Gasteiger J (1997) Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species. *Journal of chemical information and computer sciences* 37(4):705-714.
34. Streitwieser A (1961) *Molecular orbital theory for organic chemists* (Wiley, New York,) p 489 p.

35. Gasteiger J & Saller H (1985) Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angewandte Chemie International Edition in English* 24(8):687-689.
36. Gilson MK, Gilson HS, & Potter MJ (2003) Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. *Journal of chemical information and computer sciences* 43(6):1982-1997.
37. Gasteiger J & G. Hutchings M (1983) New empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. *Tetrahedron Letters* 24(25):2537-2540.
38. Gasteiger J & Hutchings MG (1984) Quantitative models of gas-phase proton-transfer reactions involving alcohols, ethers, and their thio analogs. Correlation analyses based on residual electronegativity and effective polarizability. *Journal of the American Chemical Society* 106(22):6489-6495.
39. Miller KJ (1990) Additivity methods in molecular polarizability. *Journal of the American Chemical Society* 112(23):8533-8542.
40. Mueller R*, et al.* (2010) Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chemical Neuroscience* 1(4):288-305.

2.7 Supplementary information

Normalization of stereochemistry score

The stereochemistry score is normalized based on the maximum possible stereochemistry score which can be computed assuming $a \geq b \geq c$ and $c = a - b$:

$$f(a,b,c) = -(a-b)(b-c)(c-a) = -a^3\left(1-\frac{b}{a}\right)\left(\frac{b}{a}-\frac{c}{a}\right)\left(\frac{c}{a}-1\right) = a^3\left(1-\frac{b}{a}\right)\left(2\frac{b}{a}-1\right)\left(\frac{b}{a}\right)$$

With a^3 being a constant and $x := \frac{b}{a}$ we find: $f(x) = 3x^2 - x - 2x^3$.

$$\frac{\partial f}{\partial x} = 6x - 1 - 6x^2$$

$$0 = x^2 - x + \frac{1}{6}$$

$$x = \frac{1 \mp \sqrt[2]{\frac{2}{6}}}{2} \quad \rightarrow \quad x = \frac{1}{2} \mp \sqrt[2]{\frac{1}{12}}$$

$$b = 0.211328, c = 0.788675$$

$$\max\{(1-b)(b-c)(c-1)\} = 0.0962243$$

Table S2-1. Complete feature set used in feature selection analysis. Control set included all of the same features without novel EMAS functions.

	Descriptor Name	Description
Scalar descriptors	Weight	Molecular weight of compound
	HbondDonor	Number of hydrogen bonding donors
	HBondAcceptor	Number of hydrogen bonding acceptors
	TopologicalPolarSurfaceArea	Topological polar surface area in [Å^2] of the molecule
	LogP	Octanol/water Partition coefficient calculated by atom-additive method
	TotalCharge	Sum of atomic formal charges across molecule
Vector descriptors	Identity	weighted by atom identities
2D Autocorrelation	SigmaCharge	weighted by σ atom charges
(11 descriptors)	PiCharge	weighted by π atom charges
3D Autocorrelation	TotalCharge	weighted by sum of σ and π charges
(12 descriptors)	SigmaEN	weighted by σ atom electronegativities
Radial Distribution Function	PiEN	weighted by π atom electronegativities
(48 descriptors)	LonePairEN	weighted by lone pair electronegativities
Novel EMAS Function weighted by sum of properties	EffectivePolarizability	weighted by effective atom polarizabilities
(24 descriptors)		
Novel EMAS Function weighted by product of properties	Vcharge	weighted by partial atomic charges accounting for alternate resonance forms
(24 descriptors)		
Every Vector descriptor available with and without van der Waals surface area weighting		

Table S2-2. Feature selection results with and without EMAS features. Novel EMAS features have been highlighted.

Control feature selection (without EMAS)		Novel feature selection (with EMAS)	
Descriptor Type	**Weight**	**Descriptor Type**	**Weight**
Radial Distribution Function	AtomIdentity [surface area scaled]	Radial Distribution Function	AtomIdentity [surface area scaled]
Radial Distribution Function	Vcharge	Radial Distribution Function	Vcharge
Radial Distribution Function	EffectivePolarizability [surface area scaled]	EMAS (product weight)	AtomIdentity [surface area scaled]
3D Autocorrelation	SigmaCharge	2D Autocorrelation	SigmaEN [surface area scaled]
Radial Distribution Function	LonePairEN	Radial Distribution Function	PiEN [surface area scaled]
2D Autocorrelation	SigmaEN	Scalar	HbondDonor
3D Autocorrelation	SigmaEN	EMAS (product weight)	SigmaEN [surface area scaled]
3D Autocorrelation	Vcharge [surface area scaled]	2D Autocorrelation	EffectivePolarizability [surface area scaled]
2D Autocorrelation	Vcharge [surface area scaled]	3D Autocorrelation	Vcharge [surface area scaled]
		Radial Distribution Function	PiEN
		3D Autocorrelation	SigmaCharge
		2D Autocorrelation	EffectivePolarizability
		EMAS (sum weight)	Vcharge [surface area scaled]
		EMAS (product weight)	Vcharge
		EMAS (sum weight)	TotalCharge
		Radial Distribution Function	EffectivePolarizability
		EMAS (sum weight)	LonePairEN
		EMAS (product weight)	PiEN [surface area scaled]
		3D Autocorrelation	PiEN [surface area scaled]
		Radial Distribution Function	SigmaCharge

Chapter 3

Improvements to 3D Autocorrelation Molecular Descriptors in QSAR

Manuscript in preparation

Gregory Sliwoski, Jeff Mendenhall, Jens Meiler: **Improvements to 3D Autocorrelation Molecular Descriptors in QSAR**

3.1 Abstract

Quantitative Structure-Activity Relationship (QSAR) is a branch of computer aided drug discovery (CADD) that relates chemical structures using chemical descriptors to biological activity. One well established QSAR descriptor is three-dimensional autocorrelation (3DA). In this paper we evaluate two variations of 3DA: 3DA_Smooth applies a smoothing functionality to the 3DA descriptor and 3DA_Sign accounts for the sign of atom properties while generating 3DA curves. Splitting unique sign pairs such as negative-negative, positive-positive, and opposite signs avoids information loss when multiplying two negative properties. We evaluate these two variations with models trained on nine datasets spanning a range of drug target classes. 3DA_Smooth did not significantly increase model performance over 3DA, suggesting that the computationally more expensive smoothing does not significantly improve the information content of such a descriptor. Splitting up sign pair variants with 3DA_Sign, however, significantly increased model performance across all datasets. Lastly, we tested a limited 3DA_Sign that encodes atom pair distances up to six angstroms instead of the traditional twelve. We found that focusing this particular style of 3D descriptor to atom pair distances of six angstroms or less significantly increases model performance.

3.2 Introduction

Computer aided drug discovery (CADD) is a multi-faceted approach that implements computational tools into the drug discovery pipeline [1]. CADD can reduce the time and resources required for the development of novel therapeutics. Scientifically, CADD can also provide insights into the complex interaction between small molecule and a biological target protein. Ligand-based CADD (LB-CADD) is one approach that focuses on analyzing the collective chemical properties of a set of active and inactive compounds without leveraging explicit knowledge of the target protein structure. One fundamental principle of LB-CADD is Quantitative Structure-Activity Relationship (QSAR) modeling. The goal of QSAR modeling is to define the relation between chemical structure and biological activity in a quantitative way so that the activity of new molecules can be predicted to prioritize acquisition or synthesis. In general, QSAR can be separated into two major components: a quantitative description of molecular structure (descriptor) and a mathematical model that uses these multidimensional descriptors as input to predict activity. Both components come in a variety of flavors and strategies that vary in performance depending on the specific project. Machine learning techniques are the most commonly applied non-linear mathematical QSAR models [2]. For this study, we use Artificial Neural Networks (ANN) as implemented in BCL::ChemInfo [3] to generate our mathematical models across all conditions.

Descriptors of chemical structure are typically computed as a combination of atomic properties (mass, volume, surface area, partial charge, electro-negativity, polarizability, etc.) that are processed with a translation and rotation invariant geometric function to describe the distribution of these properties in the molecular structure. Descriptors can be grouped into five categories, depending on the 'dimensionality' of the small molecule description required: 1D) Descriptors that can be derived from the molecular formula such as molecular weight by summing up all atom masses or total charge by summing up nominal charges. 2D) Descriptors that depend on constitution such as the number of hydrogen bond donors/acceptors, number of ring systems, topological surface area, and some approximations of volume and surface area. A topological index, for example, encodes which atoms are chemically bonded [4]. 2.5D) Configuration-dependent descriptors that encode, for example, the relation of stereo-centers within a topological index [5]. 3D) Conformation-dependent descriptors including Radial Distribution Functions (RDF) [6] and 3-Dimensional Autocorrelation (3DA) [7] that encode aforementioned atomic properties in a three-dimensional fingerprint. 4D) Descriptors that take conformational flexibility into account such as those derived from low energy conformational ensembles [8].

A descriptor is considered useful when it provides pertinent information about a compound while adding minimal noise to the overall model. In this respect, the most useful descriptors are the ones with the greatest degree of information density (information used by the model divided by total information). A descriptor that provides no useful information is often ignored by statistical models but can sometimes reduce model performance by overwhelming it with noise [9]. The goal of this paper is to evaluate potential improvements to a well-established 3D descriptor known as 3DA [7].

RDF [6] and 3DA [7] are both 3D descriptors that employ slightly different approaches to describe the relative position of atom and their properties in a translation- and rotation-invariant manner. To accomplish this, both RDF and 3DA examine the distances between all pairs of atoms in a molecule. These distances are used to generate a histogram in the case of 3DAs and a probability distribution in the case of RDFs. To extend these descriptors beyond the geometric characteristics of a molecule, atom pair distances may be weighted by multiplication with atom properties such as partial charge, electronegativity, etc. The formal definition of a 3DA is shown in equation 1.

$$3DA(r_a, r_b) = \sum_i^n \sum_j^{n-1} P_i P_j \quad \text{when } r_a <= r_{ij} < r_b\text{; otherwise } 0 \qquad \mathbf{1}$$

To generate the probability distribution, RDFs apply a Gaussian distribution function for each atom pair with a smoothing factor that controls the width of this distribution. The formal definition of an RDF is shown in equation 2.

$$RDF(r) = \sum_i^n \sum_j^{n-1} P_i P_j e^{-\beta(r-r_{ij})^2} \qquad \mathbf{2}$$

In both equations, r_{ij} is the Euclidean distance between atoms *i* and *j* and *n* is the total number of atoms in the molecule. P_i and P_j are the atom properties for atoms *i* and *j* used to weight the 3DA or RDF. In equation 1, r_a and r_b define the lower and upper boundaries of the given distance step. In equation 2, β is the smoothing parameter and *r* is the running variable for the function *RDF(r)*.

The Gaussian smoothing in RDFs is designed to account for atom bond vibration and other sources of position uncertainty [6]. 3DAs lack this smoothing and instead allot every atom pair to the single nearest bin within a histogram. For example, if the distance from one bin to another is 0.5 Å, an atom pair with a separation of 4.0 Å and an atom pair of separation 4.25 Å will both be allotted to a single bin defined by the lower distance boundary 4.0 Å and the upper distance boundary 4.5 Å. In case of the RDF, the atom pair with separation of 4.0 Å will be primarily distributed into the distance center 4.0 Å. However, a small fraction of the signal, inversely proportional to the smoothing factor, will be included in the adjacent distances in the shape of a Gaussian curve centered at 4.0 Å. Atom pairs with a separation of 4.25 Å will be distributed with a similar Gaussian curve centering at 4.25 Å. This strategy, however, leads to a potential drawback with RDFs. In the specific example, the 4.25 Å distance is positioned between two distance centers (4.0 Å and 4.5 Å as defined by the step-size of 0.5 Å). Therefore, the center of the Gaussian curve at 4.25 Å, representing a significant portion of the signal for such an atom pair, will not be encoded. This information loss is especially problematic when the step-size between distance centers or smoothing factor becomes too large. Figure 3.1 details a specific example of this information loss. In Figure 3.1***A***, the step-size between distance centers is 0.1 Å and in Figure 3.1***B*** the step size is 0.5 Å. The greater the step size, the lower the resolution of the curve and the more the RDF suffers from information loss. The present study introduces a descriptor called 3DA_Smooth that mixes the characteristics of 3DAs and RDFs. It applies the Gaussian distribution of the RDF to the 3DA to avoid information loss while still leveraging the smoothening effect.

Second, we introduce a variation named 3DA_Sign that overcomes the information loss when signed properties are multiplied. As mentioned, 3DA and RDF are often weighted with atom properties to encompass information beyond the geometric structure of a molecule. One important signed atom property is partial charge that contains information regarding the distribution of electrons in the molecule. By nature, partial charge can either be positive or negative. Traditionally, when weighting an RDF or 3DA with atom properties, the properties are multiplied as is seen in equations 1 and 2. There is significant information loss when multiplying two signed properties. A pair of atoms both with positive partial charges will be encoded the same as if they both had negative partial charges. This inevitably leads to an overrepresentation of positive charges when more than one atom in the molecule has a negative partial charge. With 3DA_Sign, we separate a single 3DA curve into three: negative-negative, positive-positive, and opposite sign property pairs. Information loss for standard 3DA weighted with atom partial charge is highlighted for an active molecule from dataset AID 435034 in figure 3.1***C***.

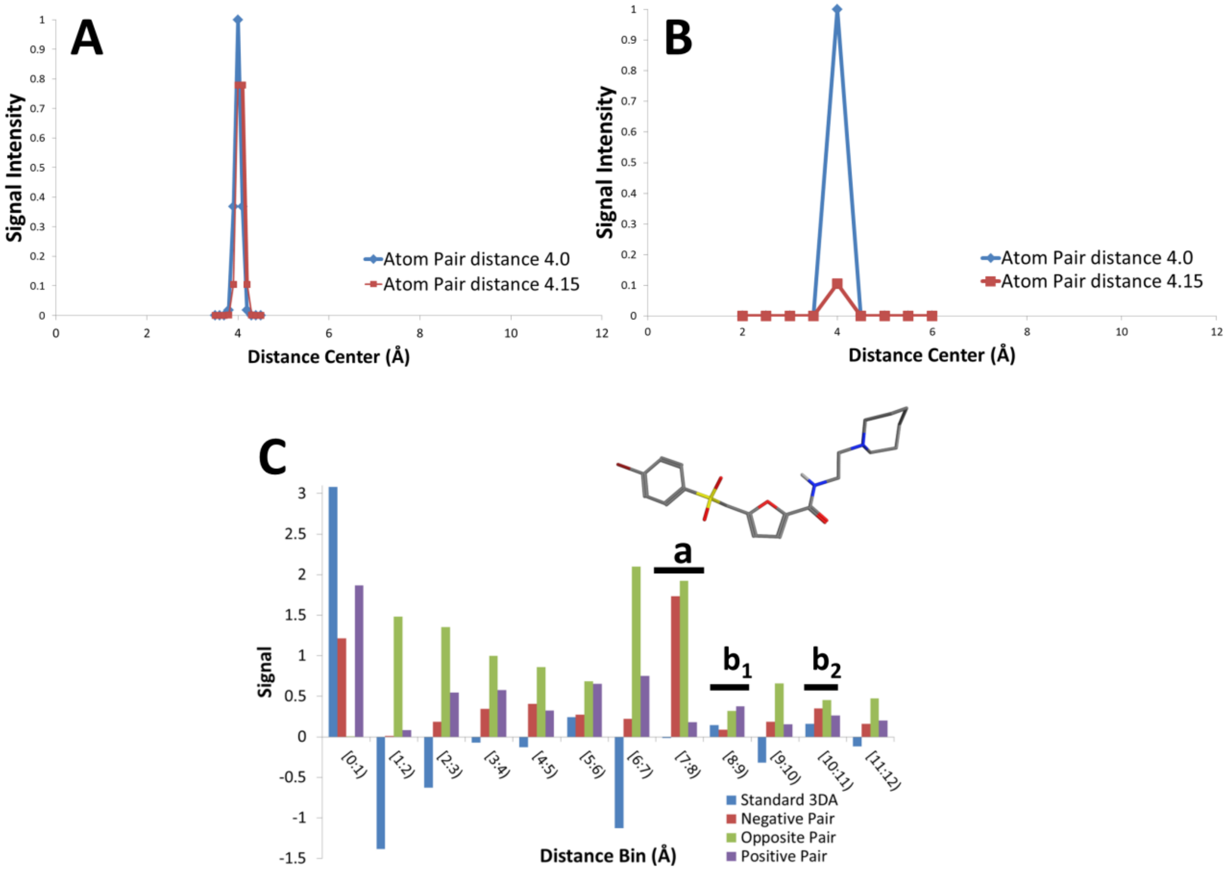


Figure 3.1 3DA_Smooth and 3DA_Sign address two potential sources of information loss with 3D descriptors A) Distribution of two atom pairs in an RDF of resolution 0.1 Å. The atom pair of distance 4.15 (red) lies between distance centers 4.0 and 4.1 and is only slightly under-represented compared with atom distance at 4.0 (blue). B) Distribution of two atom pairs in an RDF of step size 0.5 Å. The atom pair of distance 4.15 (red) lies between distance centers 4.0 and 4.5 and is significantly under-represented compared to the atom pair that lies at distance center 4.0 Å (blue). C) Information loss is revealed when standard 3DA weighted with total atom charge is split into three curves that isolate different sign pairs. 3DA descriptors out to twelve angstroms at a resolution of 1.0 angstroms per bin are compared for an active compound from screen AID 435034. Sections are highlighted including (a) standard 3DA encodes almost no signal for distance bin [7:8), whereas sign pair splitting reveals significant presence of negative sign pairs and opposite sign pairs. (b1) and (b2) standard 3DA encodes equal intensities for bins [8:9) and [10:11), whereas sign pair splitting reveals contributions of negative sign pairs and positive sign pairs are significantly different between bins.

Lastly, by default we use 3DA and RDF descriptors that encode atom pair distances up to 12 Å [10]. This distance is sufficient to capture the maximum width of most small molecules. However, 3D descriptors such as 3DA and RDF are computed from a single predicted conformation of each molecule. The smoothing factor of the RDF can be adjusted to account for some degree of conformational flexibility and uncertainty within a given molecule. However, as inter-atomic distance increases, the degree of flexibility and rotatable bonds may increase. Therefore, higher atom pair distances may come with a higher degree of uncertainty and error. We test a higher resolution 3DA/3DA_Sign variation that is limited to 6 Å instead of 12 Å.

To test whether these variations are useful in training QSAR models, we used a generalizable framework for benchmarking the utility of 3DA_Smooth and 3DA_Sign [10]. With any novel QSAR descriptor, performance evaluation is both important and challenging. In most cases, a predictive model can disregard information that does not increase performance. However, this is not guaranteed and extra descriptors adding too much noise can decrease performance. Additionally, properties that add noise for one dataset may be useful information for another. One approach is to provide the model with as many descriptors as available and perform iterative steps of descriptor selection where those that fail to significantly improve model performance are discarded. However, with an initial set of n descriptors, there are 2^n possible combinations. Coupled with the importance of cross-validation to avoid over-fitting, this process can quickly become time consuming or even intractable. Additionally, any descriptor selection must be repeated for every target of interest or high-throughput screening (HTS) dataset. Several algorithms have been presented to perform efficient descriptor selection [9]. However, as more descriptors and descriptor variations are developed, it is beneficial to use heuristics to eliminate descriptors unlikely to be beneficial. Therefore, we evaluated our descriptors with a rigorous benchmarking protocol that evaluates model performance across a variety of targets and datasets to identify those that consistently improve model performance.

3.3 Results

Developing a standard approach to descriptor benchmarking

The simplest evaluation of a descriptor’s utility is through a one-to-one comparison of models trained with and without the descriptor of interest. To keep the total information provided to QSAR models in either condition constant, it is best to compare models trained with comparable descriptors or variations. Different descriptors may encode comparable information with different approaches. For example, both RDF and 3DA descriptors describe the distribution of atoms and atom properties over a molecule using similar but slightly different algorithms. A meaningful evaluation of 3DA versus RDF utility involves comparing conditions where models have been trained with 3DAs versus models trained with RDFs. To enforce statistical comparability, the resolutions of the two curves are kept constant as well as any atom properties used for weighting. This does not always ensure that the total information provided to models in both conditions is equal. For example, 3DA_Sign splits different sign pair variants by multiplying a single 3DA curve into three. To avoid the possibility that 3DA_Sign outperforms 3DA simply because it supplies more information, we decreased the resolution of 3DA_Sign three-fold to keep the total number of properties consistent across conditions. Any increase in model performance, therefore, will not be due to increased input vector length. Model performance is judged by its ability to predict the activity of compounds it has never seen. Compounds not used for training are evaluated and ranked by their predicted activity. Plotting these predictions as true or false positives generates a receiver operating characteristic (ROC) curve. By computing the area under the curve of a logarithmic x-axis ROC curve, it is possible to score the ratio of true positive predictions to false positive predictions focusing on the high confidence predictions.

When training and evaluating QSAR model performance, large datasets that cover large chemical spaces are preferred [11]. These datasets often come from high-throughput screening (HTS) projects where active compounds have been verified against a single target. Alternatively, smaller, focused datasets may be used to evaluate novel descriptors using leave-on-out (LOO) cross-validation [12]. However, this method of benchmarking can be misleading and tends to rely heavily on the presence of specific geometries rather than more subtle properties [13]. To apply the most generalizable benchmark possible, we used nine HTS datasets curated from PubChem [10]. These datasets target various proteins including G-protein coupled receptors (GPCRs), kinases, and ion channels. The number of compounds in these datasets range from approximately 61,000 to 344,000. These datasets are detailed in table 3-1.

Table 3-1 Nine datasets were used to train models and evaluate model performance across different QSAR descriptor conditions. Dataset curation has been previously described [10].

Pubchem Project Bioassay ID	Target	Active Compounds	Inactive Compounds
1798	M1 muscarinic receptor (agonist)	187	61,646
1843	Kir2.1 potassium channel	172	301,321
2258	KCNQ2 potassium channel	213	302,192
2689	serine threonine kinase 33	172	319,620
435008	orexin 1 receptor	233	217,925
435034	M1 muscarinic receptor (antagonist)	362	61,394
463087	Cav3 calcium channel	703	100,172
485290	tyrosyl-DNA phosphodiesterase 1	281	341,084
488997	choline transporter	252	302,084

Because each 3D descriptor tested can be weighted with a variety of atom properties, we used nine different atom properties with and without accessible van der Waals (VDW) surface area scaling. Accessible VDW surface area accounts for varying accessibility of different atoms in a molecule arising from overlapping and covered VDW surfaces. Additionally, we provide all models with a standard set of descriptors (1D) to achieve a performance baseline that strengthens comparisons. All scalar molecule descriptors and atom properties used for weighting are described in table 3-2.

Table 3-2 Properties used to train ANN models are categorized as molecule (one property per molecule) and atom (one property per atom). Molecule properties are used in every condition as a standard baseline of QSAR information and contain general information regarding overall molecular properties. Atom properties are used in every condition to weight the corresponding descriptor (3DA, RDF, 3DA_Smooth, or 3DA_Sign) with and without VDW surface area scaling. Atom properties that are split into unique sign pairs with the 3da_Sign descriptor are indicated as 'signed.'

Property	Type	Description	Signed
Molecular Weight	Molecule	Total weight of molecule	
HBondDonor	Molecule	Total hydrogen bond donors in molecule	
HBondAcceptor	Molecule	Total hydrogen bond acceptors in molecule	
LogP	Molecule	Octanol/water coefficient; solubility	
TotalCharge	Molecule	Total charge of molecule	
NRotBond	Molecule	Number of rotatable bonds	
NAromaticRings	Molecule	Number of aromatic rings	
NRings	Molecule	Number of closed rings	
TopologicalPolarSurfaceArea	Molecule	Total surface area of molecule that is polar	
BondGirth	Molecule	Maximum number of bonds between two toms	
MaxRingSize	Molecule	Number of atoms in largest ring	
MinRingSize	Molecule	Number of atoms in smallest ring	
AromaticAtoms	Molecule	Number of atoms in aromatic rings	
IntersectionAtoms	Molecule	Number of atoms in ring intersections	
AromaticIntersectionAtoms	Molecule	Number of atoms in aromatic ring intersections	
MaxSigmaCharge	Molecule	Maximum σ charge	
MinSigmaCharge	Molecule	Minimum σ charge	
TotalSigmaCharge	Molecule	Sum of all σ charges	
StDevSigmaCharge	Molecule	Standard deviation of all σ charges	
MaxVcharge	Molecule	Maximum V-charge	
MinVcharge	Molecule	Minimum V-charge	
TotalVcharge	Molecule	Sum of all V-charges	
StDevVcharge	Molecule	Standard deviation of all V-charges	
Girth	Molecule	Widest diameter of molecule	
Identity	Atom	Unweighted; 1 for all atoms	
SigmaCharge[14-16]	Atom	Partial charge localized to σ-electron system	X
PiCharge[17-19]	Atom	Partial charge localized to π-electron system	X
TotalCharge	Atom	Total partial charge of atom	X
Vcharge[20]	Atom	Partial charge accounting for resonance	X
EffectivePolarizability[21-23]	Atom	Responsiveness of electron density to external field	

IsRingIntersection	Atom	1 if atom is at a non-aromatic ring intersection, 0 otherwise
IsInAromaticRing	Atom	1 if atom is within aromatic ring, 0 otherwise
InAromaticRingIntersection	Atom	1 if atom is at an aromatic ring intersection, 0 otherwise

3DA_Smooth: Combining 3DA and RDF

The goal of 3DA_Smooth is to achieve the same smoothing quality of an RDF without losing information for atom pairs that fall between distance centers. As with 3DA and RDF, 3DA_Smooth iterates over all atom pairs in a molecule and bins that atom pair depending on the distance between them. Atom property weighting is handled in an identical manner via the product of the two atom properties. There are two major differences between 3DA_Smooth and 3DA/RDF: atom pairs that lie between two distance centers are distributed into the two nearest centers depending on their distance from each and the Gaussian smoothing is performed after all atom pairs have been distributed. 3DA_Smooth can be divided into two steps:

Step 1

For every atom pair *i,j* whose distance r_{ij} falls between distance centers r_a and r_b:

$$f(r_a) = \sum_{i,j}^{n} \frac{p_i p_j}{e^{-\beta(r_a-r_{ij})^2}+e^{-\beta(r_b-r_{ij})^2}} e^{-\beta(r_a-r_{ij})^2} \qquad \mathbf{3}$$

$$f(r_b) = \sum_{i,j}^{n} \frac{p_i p_j}{e^{-\beta(r_a-r_{ij})^2}+e^{-\beta(r_b-r_{ij})^2}} e^{-\beta(r_b-r_{ij})^2} \qquad \mathbf{4}$$

Every atom pair is distributed into two bins in a method that is similar to the 3DA but has an additional normalization factor $1/(e^{-\beta(r_a-r_{ij})^2}+e^{-\beta(r_b-r_{ij})^2})$ that causes any atom pair whose distance lies directly between two centers to be equally distributed to both and all other atom pairs to be distributed primarily to the closest distance center. As with 3DA/RDF, p_i and p_j are the atom properties of *i* and *j* used to weight and as with RDF, β is the smoothing parameter.

Step 2:

Every calculated intensity *f(r)* is redistributed using the same Gaussian style curve of the RDF:

$$g(s) = \sum_{r}^{d} f(r) e^{-\beta(r-s)^2} \qquad \mathbf{5}$$

Step 1 produces discrete curves that resemble a standard 3DA at lower resolutions save any atom pairs whose distance lies midway between two bins. Step 2 applies a Gaussian smoothing function to these discrete values resulting in a curve that more closely resembles an RDF. In equation 5, *r* is the running distance center variable and *d* is the number of distance centers in the 3DA_Smooth code.

Models trained with 3DA, RDF, or 3DA_Smooth perform comparably

In our nine dataset benchmark, models trained with 3DA, RDF, or 3DA_Smooth perform comparably. Average model performance as scored by the area under the logarithmic ROC curve (logAUC) was 0.343 for 3DA, 0.343 for RDF, and 0.344 for 3DA_Smooth. This suggests that the information loss inherent with RDF does not significantly decrease QSAR model performance. However, these results also suggest that the Gaussian smoothing designed to account for flexibility and uncertainty within RDF and 3DA_Smooth does not provide superior ANN model performance. Therefore, because Gaussian smoothing adds computational expense, 3DA may be preferable over RDF or 3DA_Smooth. This result shows the importance of a generalizable approach to QSAR descriptor benchmarking as neither intuitive improvements increased model performance. Model performance across nine datasets is compared for 3DA, RDF, and 3DA_Smooth in figure 3.2***A***.

3DA_Sign: Separating atom properties by sign

The most common method for weighting 3DA is with the product of atom properties for each atom pair. For signed properties such as partial charge, information can be lost as the product of two negative values cannot be distinguished from the product of two positive values. To avoid this information loss, we modified the 3DA descriptor to allocate atom pairs into one of three vectors depending on the signs of the two atom properties. This descriptor is called 3DA_Sign and contains three times as many properties as a standard 3DA with the same number of steps. Every atom pair is distributed to one of three vectors depending on whether the atom properties are both negative, both positive, or of opposite signs. This improvement is designed specifically for signed descriptors such as partial charge since unsigned properties will solely fill the positive-positive vector. Therefore, when testing the utility of 3DA_Sign, we only apply it with signed properties. All unsigned properties are included with a standard 3DA.

Since we are primarily concerned with information density, we wanted to keep the number of input values constant between conditions. For example, comparable conditions include 3DA with 72 steps of step size 0.167 Å, and 3DA_Sign with 24 steps of step size 0.5 Å. Both conditions contain 36 quantitative properties, but 3DA is of a higher resolution than 3DA_Sign. This decreases the power of discovering an improvement with 3DA_Sign since it is at a lower resolution. However, by adjusting resolution to keep the number of inputs constant, we avoid the potential confounding variable of input vector size.

Despite the lower resolution, 3DA_Sign improved model performance over standard 3DA in all datasets. Average model performance across nine datasets as measured by logAUC was 0.358 when applying signed properties with 3DA_Sign (versus 0.343 with 3DA), an average improvement of 4.4% (paired t-test $p<0.05$). Model performance across nine datasets is compared for 3DA and 3DA_Sign in figure 3.2***B***.

Finally, we tested limiting the maximum atom pair distance encoded to 6 Å instead of 12. By focusing on the first 6 Å at higher resolution, model performance increased significantly from an average performance as measured by logAUC of 0.358 to 0.381 (6.4% improvement, paired t-test $p<.001$). Figure 3.2***B*** compares model performance across nine datasets when encoding atom pair distances up to twelve angstroms versus six angstroms.

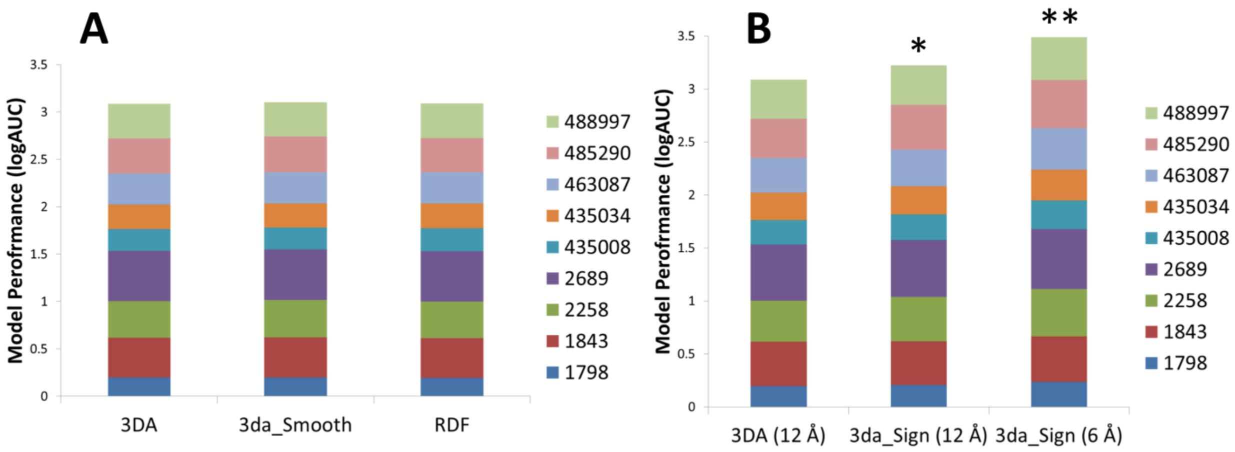


Figure 3.2 Model performance is compared across nine datasets for 3D descriptor modifications A) Model performance is not significantly changed when training with 3DA, 3DA_Smooth, or RDF. Performance is evaluated by the area under the logarithmic ROC curve between 0.001 and 0.01. Nine datasets are indicated by their Pubchem HTS project assay ID. B) Model performance is compared across nine datasets for two 3DA variations. Splitting sign pairs into negative-negative, positive-positive, and opposite signs when weighting 3DA with signed atom properties (3DA_Sign) significantly increases model performance when compared to using standard 3DA with signed properties (*3DA_Sign (12 Å) vs 3DA (12 Å) paired t-test $p<0.05$, n=9). Limiting maximum atom pair distance to 6 Å significantly increases model performance when compared to limiting maximum atom pair distance to 12 Å (**3DA_Sign (12 Å) vs 3DA_Sign (6 Å) paired t-test $p<0.001$, n=9).

3.4 Discussion

This study outlines a general QSAR descriptor benchmarking technique that can be used to evaluate novel descriptors. Three potential 3DA modifications are evaluated using this generalizable benchmark strategy. Descriptors represent small molecules as vectors of numerical properties that can train ANNs to predict small molecule activity towards a specific target. These descriptors come in a continuously growing range of dimensions and information content. Coupled with the high degree of customization for many descriptors, training models using every available descriptor is not only computationally inefficient, but may introduce noise that hinders model performance. Therefore, an evaluation of a novel descriptor is critical before including it with QSAR model application. This evaluation must also be applied across multiple datasets with different targets. By nature, these biological targets may focus on different property demands, thereby making a broad statement of a descriptor's utility helpful.

The first modification tested is a hybrid 3DA/RDF descriptor called 3DA_Smooth. This modification takes advantage of the inherent Gaussian smoothing of RDF while avoiding the information loss that can arise for some atom pair distances. We trained comparable models using 3DA, RDF, or 3DA_Smooth descriptors and evaluated their prediction performance across nine datasets. Surprisingly, models trained with any of these 3D descriptors performed comparably across all datasets. This suggests that the Gaussian smoothing applied to RDF and 3DA_Smooth does not significant improve information content over 3DA. Based on the increased computational overhead for RDF and 3DA_Smooth, 3DAs may be preferable to RDF or 3DA_Smooth for ANN-based QSAR model generation.

Secondly, we tested a variation of standard 3DA that applies to weighting with signed atom properties. Multiplying two negative properties produces the same result as multiplying two equivalent positive properties, leading to misinformation for molecules with two or more atoms with negative properties. To avoid this problem, we introduced 3DA_Sign to replace standard 3DAs when weighting with signed atom properties. 3DA_Sign generates three curves of equal resolution, splitting atom pairs into negative-negative, positive-positive, and opposite signs. Because this modification generates three descriptors for every one corresponding 3DA descriptor, we decreased the resolution of 3DA_Sign threefold in our evaluations to avoid confounding results with different input vector sizes. We found that the replacing 3DA with 3DA_Sign for signed atom properties significantly increased ANN model performance across nine large datasets.

Lastly, we tested a maximum atom pair distance limitation of 6 Å instead of 12. Although 12 Å covers the maximum width of many small molecules, encoding longer atom pair distances can provide false information in cases of high molecular flexibility or bond rotation. A 6 Å limitation, on the other hand, focuses more on fragments within the molecule that are less prone to flexible uncertainty. Additionally, shorter distances can be sampled at a higher resolution without increasing input vector size. We found that limiting the maximum atom pair distance to 6 Å significantly increases performance across nine datasets.

In conclusion, we present three recommendations for ANN-based QSAR descriptor selection: 1) Because RDF and 3DA descriptors produce comparable model performance and 3DA is computationally less expensive than RDF, 3DA is the descriptor of choice for this style of 3D descriptors. 2) Multiplying signed properties when weighting 3DA can significantly hinder model performance. Therefore, splitting sign pairs into negative-negative, positive-positive, and opposite signs can significantly improve model performance. 3) Limiting 3DAs to encode atom pairs up to 6 Å instead of 12 can significantly improve model performance.

3.5 Methods

Generation of numerical descriptors for QSAR model creation

Numerical descriptors and QSAR models were generated and evaluation over nine HTS datasets detailed in table 3-1. The curation of these datasets has been previous outlined [10]. 3D conformations of all small molecules were generated using the CORINA [24] software package.

The BioChemical Library (BCL) software was used to generate all scalar and 3D molecular descriptors tested in this study. All descriptors and atom properties used to weight 3D descriptors are described in table 3-2. When weighting 3D descriptors, all atom properties are represented with and without accessible surface area scaling. All conditions include 1,374 total properties distributed over 39 descriptors.

Artificial neural network model architecture and training

All ANN models were trained using back propagation and a sigmoid transfer function with a simple weight update of $\eta = 0.05$ and $\alpha = 0.5$, a hidden layer of 32 neurons, 0.1 visible neuron dropout, and 0.5 hidden neuron dropout. Each dataset was divided into three sets of compounds: compounds used to train the model (training), compounds used to monitor model performance during training to avoid over-fitting (monitoring), and compounds kept hidden from the model during training to evaluate predictability after training has completed (independent). Five-fold cross-validation was used where 20 individual ANN models were trained for each HTS dataset by rotating which compounds appeared in the training, monitoring, and independent sets. Final active or inactive prediction for each independent compound was taken as a consensus across models for which that compound appeared in the independent set. The objective function used during training was the area under the logarithmic receiver operating characteristic (ROC) curve [25, 26] (logAUC [27]) between false positive rates of 0.001 and 0.01.

ANN model performance evaluation

All models were evaluated with the same objective function used for training. ROC curves with a logarithmic x-axis were generated for consensus predictions sorted by predicted activity and the area under the curve as calculated for the range of 0.001 to 0.01 (the top 1% of predicted compound activities). For all statistical comparisons, two-tailed paired t-tests were performed between descriptor conditions across the nine HTS datasets.

3.6 References

1. Sliwoski G, Kothiwale S, Meiler J, & Lowe EW, Jr. (2014) Computational methods in drug discovery. *Pharmacological reviews* 66(1):334-395.
2. Salt DW, Yildiz N, Livingstone DJ, & Tinsley CJ (1992) The Use of Artificial Neural Networks in QSAR. *Pesticide Science* 36(2):161-170.
3. Butkiewicz M, Lowe EW, & Meiler J (2012) Bcl::ChemInfo - Qualitative analysis of machine learning models for activation of HSD involved in Alzheimer's Disease. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, pp 329-334.
4. Trinajstić N (1992) *Chemical graph theory* (CRC Press, Boca Raton) 2nd Ed p 322 p.
5. Balaban AT (1998) Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR, Similarity/Dissimilarity and Drug Design. *SAR and QSAR in environmental research* 8(1-2):1-21.
6. Hemmer MC, Steinhauer V, & Gasteiger J (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* 19(1):151-164.
7. Broto P, Moreau G, & Vandycke C (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies. Perception of molecules: topological structure and 3-dimensional structure. *European journal of medicinal chemistry* 19(1):61-65.
8. Hopfinger AJ*, et al.* (1997) Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society* 119(43):10509-10524.
9. Shahlaei M (2013) Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. *Chemical Reviews* 113(10):8093-8103.
10. Butkiewicz M*, et al.* (2013) Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database. *Molecules* 18(1):735-756.
11. Kubinyi H, Folkers G, & Martin YC (1998) *3D QSAR in drug design* (Kluwer Academic, Dordrecht ; Boston, Mass) pp v. < 2- >.
12. Kiralj R & Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *Journal of the Brazilian Chemical Society* 20:770-787.
13. Manchester J & Czermiński R (2009) CAUTION: Popular “Benchmark” Data Sets Do Not Distinguish the Merits of 3D QSAR Methods. *Journal of chemical information and modeling* 49(6):1449-1454.
14. Gasteiger J & Marsili M (1978) A new model for calculating atomic charges in molecules. *Tetrahedron Letters* 19(34):3181-3184.
15. Gasteiger J & Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 36(22):3219-3228.
16. Guillen MD & Gasteiger J (1983) Extension of the method of iterative partial equalization of orbital electronegativity to small ring systems. *Tetrahedron* 39(8):1331-1335.
17. Bauerschmidt S & Gasteiger J (1997) Overcoming the limitations of a connection table description: A universal representation of chemical species. *Journal of chemical information and computer sciences* 37(4):705-714.
18. Streitwieser A (1961) Molecular orbital theory for organic chemists.

19. Gasteiger J & Saller H (1985) Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angewandte Chemie International Edition in English* 24(8):687-689.
20. Gilson MK, Gilson HS, & Potter MJ (2003) Fast assignment of accurate partial atomic charges: An electronegativity equalization method that accounts for alternate resonance forms. *Journal of chemical information and computer sciences* 43(6):1982-1997.
21. Gasteiger J & Hutchings MG (1983) New empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. *Tetrahedron Letters* 24(25):2537-2540.
22. Gasteiger J & Hutchings MG (1984) Quantitative models of gas-phase proton-transfer reactions involving alcohols, ethers, and their thio analogs. Correlation analyses based on residual electronegativity and effective polarizability. *Journal of the American Chemical Society* 106(22):6489-6495.
23. Miller KJ (1990) Additivity methods in molecular polarizability. *Journal of the American Chemical Society* 112(23):8533-8542.
24. Sadowski J & Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews* 93(7):2567-2581.
25. Cleves AE & Jain AN (2006) Robust ligand-based modeling of the biological targets of known drugs. *Journal of medicinal chemistry* 49(10):2921-2938.
26. Hristozov DP, Oprea TI, & Gasteiger J (2007) Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *Journal of computer-aided molecular design* 21(10-11):617-640.
27. Clark RD & Webster-Clark DJ (2008) Managing bias in ROC curves. *J Comput Aided Mol Des* 22(3-4):141-146.

Chapter 4

Discovery of Small-Molecule Modulators of the Human Y_4 Receptor

---

manuscript under review

Gregory Sliwoski, Mario Schubert, Jan Stichel, David Weaver, Annette Beck-Sickinger, Jens Meiler:

Discovery of small-molecule modulators of the human Y_4 receptor

4.1 Abstract

The human Y_4 receptor and its native ligand, pancreatic polypeptide, are critically involved in the regulation of human metabolism by signaling satiety and regulating food intake, as well as increasing energy expenditure. Thus, this receptor represents a putative target for treatment of obesity. With respect to new approaches to treat complex metabolic disorders, especially in multi-receptor systems, small molecule allosteric modulators have been in the focus of research in the last years. However, no positive allosteric modulators or agonists of the Y_4 receptor (Y_4R) have been described so far. In this study, small molecule compounds derived from the Niclosamide scaffold were identified by high-throughput screening to increase Y_4R activity. Compounds were characterized for their potency and their effects at the human Y_4R and as well as their selectivity towards Y_1R, Y_2R and Y_5R. These compounds provide a structure-activity relationship profile around this common scaffold and lay the groundwork for hit-to-lead optimization and characterization of positive allosteric modulators of the Y_4R.

4.2 Introduction

Obesity, a major risk factor for diabetes, heart disease, cancer, and mortality is a rising medical concern with doubled worldwide prevalence since 1980, reaching an estimated medical cost of $147 billion in 2008 [1, 2]. Dietary changes and nutritional counseling can be an effective treatment option but results are inconsistent and suffer from poor long-term patient adherence that often ends in weight regain [3]. So far, only invasive treatments such as bariatric surgery show long term success rates, but are limited to patients where the benefits outweigh the risks and costs [4]. Interestingly, several studies suggest that hormonal changes following bariatric surgery significantly contribute to its long term success [2, 5]. Accordingly, the respective hormone receptors may represent promising therapeutic targets. For example, enhanced meal-stimulated glucagon-like peptide-1 (GLP-1) release is thought to participate in the long-term success of bariatric procedures. GLP-1 receptor agonists have been shown to produce weight loss and glucose homeostasis for subjects with type II diabetes. However, the weight loss seen with GLP-1 agonists alone is modest [6].

Two members of the pancreatic polypeptide family including peptide tyrosine tyrosine (PYY) and pancreatic polypeptide (PP) act as satiety factors, inhibit food intake, and modify metabolic homeostasis [7]. Neuropeptide Y (NPY), the third member of this class of hormones, is unlike PP and PYY not a gut derived hormone but acts predominantly in central regions of the nervous system as a neurotransmitter. All three 36 amino acid peptides act through four NPY receptor subtypes in humans, Y_1R, Y_2R, Y_4R, and Y_5R, which thereby are putative targets for treatment of obesity [8]. While the ligands PP and PYY both present promising routes for the treatment of obesity, PP may be preferred as it inhibits feeding in mice more than PYY and PYY-3-36 [9]. Pancreatic polypeptide have also been shown to inhibit food intake in man at low concentrations [10].Further, in contrast to PP, medically relevant doses of PYY induce nausea in humans [10, 11].

PP is released under vagal cholinergic control from F-cells of pancreatic islets in response and proportion to food ingestion [12]. The hormone is furthermore expressed in some endocrine cells of the intestines [13]. Primarily through Y_4R, PP promotes appetite suppression, inhibition of gastric emptying, and increased energy expenditure [14]. The human Y_4R is a 375 amino acid class A G-protein coupled receptor (GPCR) primarily expressed in gastrointestinal tract where it inhibits peristalsis and excretion [15]. Other peripheral organs that express Y_4R include the heart, skeletal muscle, and thyroid gland. In the central nervous system, Y_4R is expressed in the hypothalamus, where it relays anorexigenic signals [16] and inhibits neurotransmitter release [17]. The Y_4R is considered as a putative target for pharmacological treatment of obesity based on its strong anorexigenic potential and studies involving the overexpression or application of PP [16, 18-20].

Our efforts focus on the identification and development of small-molecule positive allosteric modulators (PAMs) of Y_4R. Allosteric ligands represent promising options for treatment of complex metabolic and neurological diseases [21]. Allosteric ligands may adopt a wide range of pharmacological activities including PAMs (including agonism, potentiation, or both), negative allosteric modulation (NAM), inverse agonism, or biased signaling wherein different receptor modulations can favor particular signaling pathways downstream of the GPCR [22]. This provides the opportunity to design compounds that fine tune receptor activity. Additionally, PAMs with little or no intrinsic activity may be safer therapeutics because their dependence on the presence of the endogenous agonist may help to prevent toxicity and other negative side effects [23]. Furthermore, this approach preserves the physiological signaling patterns, which may be critical in complex systems and is not feasible using orthosteric agonists [24]. Allosteric binding sites may be less conserved between receptor subtypes than the orthosteric binding site because they lack the evolutionary pressure to conserve affinity for the orthosteric ligand. Therefore it is often possible to design allosteric modulators with high selectivity [24, 25]. This is of particular importance in multi receptor systems consisting of several subtypes sharing an overlapping preference for a ligand.

Since no small-molecule PAM or agonist has been identified for Y_4R, we used high-throughput screening (HTS) [26] to identify compounds that modulate the Y_4R. Initial hit compounds were validated as PAMs in a complementary set of assays and subtype-selectivity was investigated for all human NPY receptors.

4.3 Results

Identification of Y_4R PAMs

Until now, no small molecule agonists of Y receptors have been described. Based on this lack of any structure activity data, identification of Y_4R PAMs was initiated with a Ca^{2+}-flux-based HTS approach. An initial pilot screen was performed with the ‘Spectrum collection’, a small scale library with 2000 compounds comprised of synthetic small molecules as well as purified natural product covering a range of known biologically active properties. This pilot screen yielded 65 putative PAMs. All initial hits were retested for nonspecific activity in wildtype COS-7 cells and the PAM effect was validated via concentration-dependent potentiation of a submaximal PP response. After eliminating structures that show non-specific effects, seven compounds with potencies in the micromolar range were identified (Figure S4.1) and selected for further investigation. Validation of the Y_4R PAM activity was performed with a well-established assay system for Y receptor activation studies, based on cellular accumulation of inositol phosphate [27, 28]. Furthermore, the compound activity at Y_1R, Y_2R and Y_5R was examined at this stage of Y_4R PAM identification. Y receptor activation with a ligand concentration causing a submaximal response was monitored in presence and absence of compound. This experimental setup allows the detection of potential compound-induced shifts of the concentration response curve and parallel testing of all compounds on all Y receptor subtypes. Cells stably expressing the chimeric G-protein $\Delta 6G_{\alpha qi4myr}$ and one of four different human NPY receptor subtypes (Y_1R, Y_2R, Y_4R, or Y_5R) were treated with test compound at a final concentration of 10 µM followed immediately by stimulation with 1 nM endogenous peptide agonist (Y_4R: PP, $Y_{1,2,5}R$: NPY).

The effect of the compounds on inositol phosphate accumulation was investigated in relation to the control in presence of DMSO (DMSO set to 100%, complete data of all 7 compounds see Figure S4.1). However, only three of the seven HTS PAM hits validated their Y_4R PAM effect also in the inositol phosphate accumulation assay (Figure 4.1). Of all tested compounds, Niclosamide was the strongest Y_4R PAM, significantly increasing IP accumulation following stimulation of Y_4R with 1 nM PP (185.7 ± 23.5 % (SEM) versus 100.00 ± 13.9 %, $p<0.05$; see Figure 4.1). Furthermore, Niclosamide had minor effects on Y_1R (125.4 ± 14.6 % versus 100.0 ± 28.3 %, $p>0.05$), Y_2R (88.5 ± 6.2 % versus 100.0 ± 14.9 %, $p>0.05$), and Y_5R (83.4 ± 5.9 % versus 100.0 ± 27.3 % $p>0.05$) when stimulated with NPY. In addition to Niclosamide, adenosine and compound VU0244224 induced slight increase of the Y_4R signal response in the IP_3 assay (120 ± 11 % and 127 ± 19%, respectively). Whereas compound VU0244224 had no effects on other Y receptor subtypes, adenosine decreased signal transduction at Y_1R and Y_5R. These results show Niclosamide to be the most effective Y_4R PAM hit compound with validated activity in a complementary assay and it alone was selected for further investigation (Figure S4.1). Due to the lack of any viable hits with the exception of Niclosamide, a second screen was performed testing a total of 33,288 compounds. Of these 33,288 compounds, 32,000 were randomly selected from the VICB compound library. Since Niclosamide showed the strongest effect as a Y_4R PAM, the collection of compounds for the second screening was enriched with 1,288 compounds structurally similar to Niclosamide based on Tanimoto coefficients (Table S4-1). All potential Y_4R PAMs were retested for nonspecific effects in wildtype COS-7 cells. Hit compounds that did not show any activity in wildtype COS-7 cells were tested on the Y_4R in a concentration dependent manner for their effect on a PP EC_{20} and EC_{80}. Validated hits were then tested for their effects on histamine and bradykinin receptor-evoked changes in intracellular Ca^{2+} in order to further exclude off-target effects that might, for instance, result from changes in common effectors downstream of the GPCR. In the second HTS, four structurally similar compounds to Niclosamide were identified as Y_4R PAMs. Along with two structurally related but inactive compounds of the VU compound library (VU0357475 and VU0114795), selected to supplement structure-activity relationship studies with YR selectivity, these compounds were further investigated for Y_4R PAM activity and YR subtype selectivity (Figure 4.2).

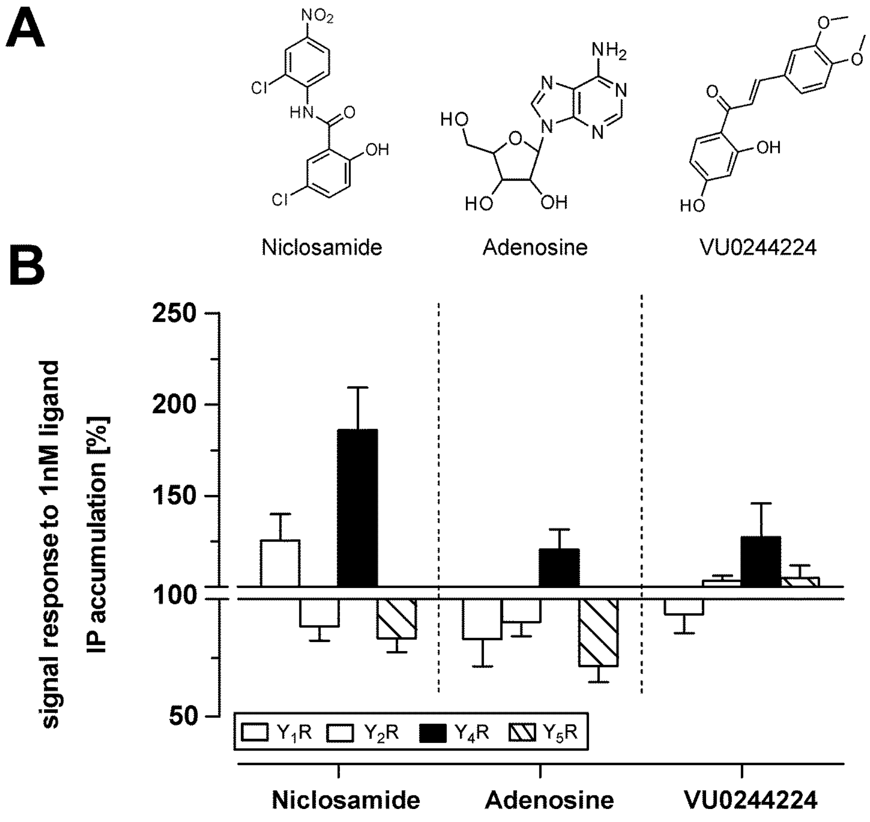


Figure 4.1 Validation of Y4R PAM activity and subtype selectivity of initial Ca2+-flux-based screen hit compounds in an inositol phosphate accumulation assay. Effect of 10 µM compound on submaximal YR activation by 1 nM ligand, which represents EC20-EC60 (Y1,2,5R: NPY; Y4R: PP). Data represent the mean ± SEM of two independent experiments each performed in quadruplicate.

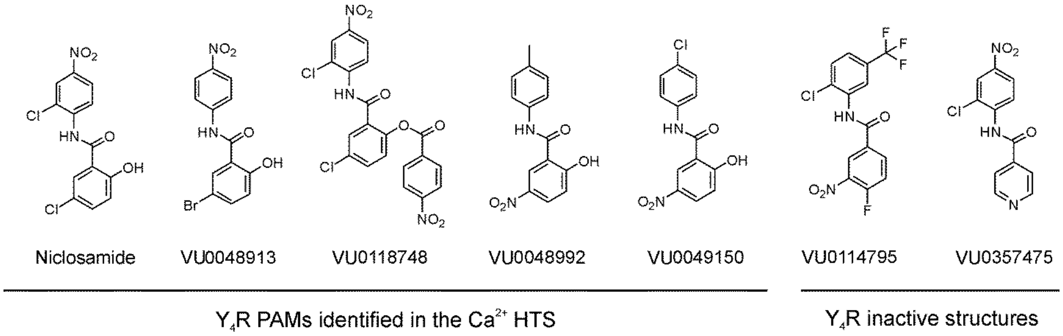
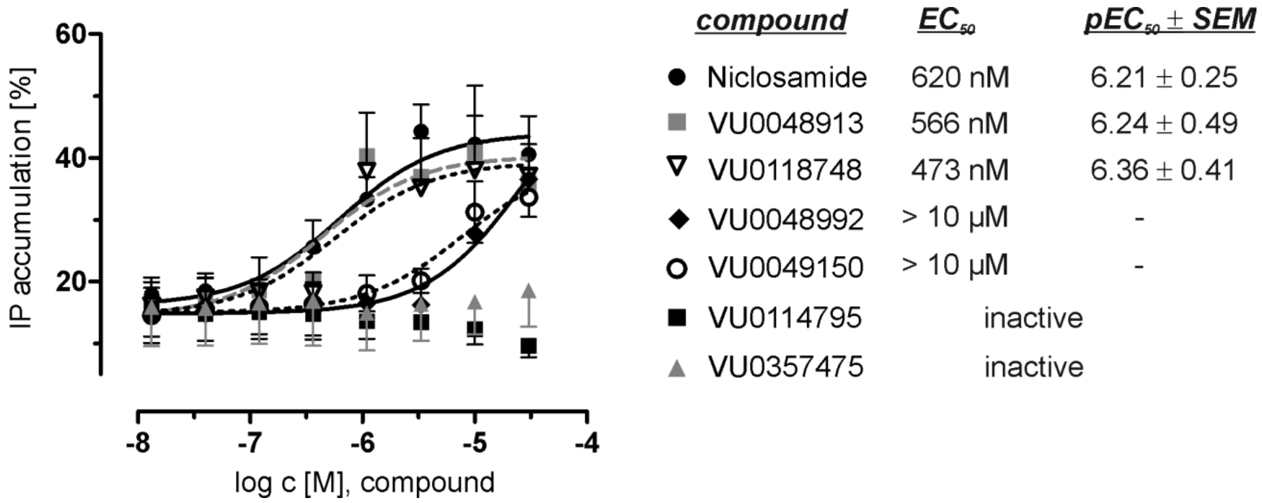

Niclosamide
VU0048913
VU0118748
VU0048992
VU0049150
VU0114795
VU0357475
Y_4R PAMs identified in the Ca^{2+} HTS
Y_4R inactive structures

Figure 4.2 Structures of Y4R PAMs identified by HTS and inactive control compounds chosen for further characterization of Y4R PAM activity and YR subtype selectivity.

Validation and selectivity of Y_4R PAMs

After identification of Niclosamide-like compounds in the HTS, the Y_4R PAM activity was validated using an IP_3 accumulation assay system. Compounds were investigated for potentiation of a PP EC_{20} in a concentration-dependent manner to determine their potency on the Y_4R (Figure 4.3).

Niclosamide, VU0048913, VU0048992, VU0049150 and VU0118748, identified in the Ca^{2+} HTS, also potentiated the PP signal response in the IP_3 assay. In presence of 30 µM of the active compounds, the PP response increased to approximately 40% (Figure 4.3) compared to the PP-evoked signal in the absence of the compounds. However, modifications on the Niclosamide scaffold affected the potency of the compounds. Niclosamide, VU0048913 and VU0118748 potentiated the PP EC_{20} with comparable EC_{50} values of 620 nM, 566 nM and 473 nM, respectively. In contrast, structural modifications in compounds VU0048992, VU004915 and VU0357475 lead to a dramatic loss of Y_4R potency (EC_{50} >10 µM) or completely inactive structures in case of VU0357475 and VU0114795.



	compound	***EC_{50}***	***$pEC_{50} \pm SEM$***
●	Niclosamide	620 nM	6.21 ± 0.25
■	VU0048913	566 nM	6.24 ± 0.49
∇	VU0118748	473 nM	6.36 ± 0.41
◆	VU0048992	> 10 µM	-
○	VU0049150	> 10 µM	-
■	VU0114795	inactive	
▲	VU0357475	inactive	

Figure 4.3 Y_4R PAM activity of Niclosamide-like compounds. Potency of the Y4R PAMs were validated with an inositol phosphate accumulation assay through potentiation of a PP EC20 response. Data have been normalized to the maximum IP accumulation caused by the Emax concentration of PP. A PP concentration resulting in 20% of the maximum signal was used to evaluate the potentiation potency of the candidate compounds. Data represent the mean ± SEM of three independent experiments performed in duplicate.

As noted previously, four subtypes of Y receptors are expressed in the human organism. In order to investigate the Y receptor subtype selectivity, we tested the effect of the Niclosamide-like structures (Figure 4.2) for all four human Y receptors with their native ligands (PP for Y_4R and NPY for Y_1R, Y_2R, and Y_5R). Full concentration-response relationships were determined for each ligand-receptor pair in presence of DMSO vs. 30 µM compound (Figure S4.2), the concentration at which all compounds had a comparable effect on the Y_4R (Figure 4.3). None of the tested compounds had an effect on the basal level and maximum level of the signal response (Figure S4.2). Thus, the influence of the compounds on the agonists EC_{50} (EC_{50}-shift) of the signal response was used as an indicator for selectivity (Figure 4.4).

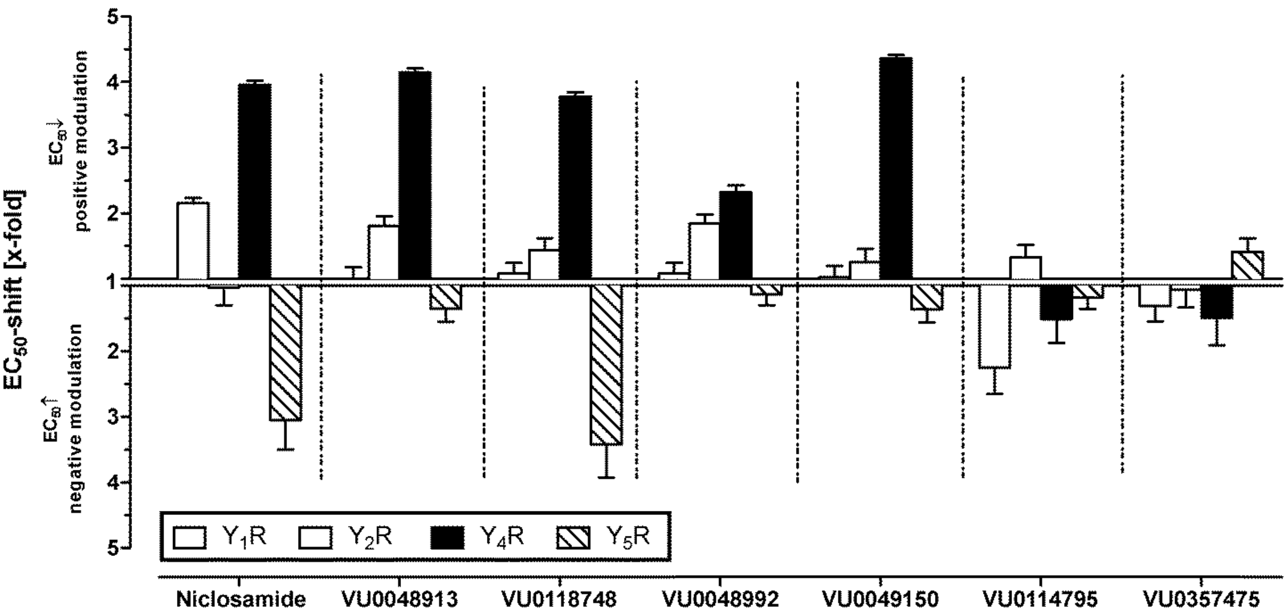


Figure 4.4 YR subtype selectivity of Y4R PAMs Effect of 30 µM compound on the EC50 of Y-receptor agonists in COS-7 cells stable expressing a Y receptor subtype and the chimeric G-protein Gα6qi4myr. Receptors were stimulated with their native ligands (Y_1R, Y_2R, Y_5R: NPY; Y_4R: PP). For Y-axis values, positive modulation represents an increase in the apparent potency of the native agonist and negative modulation represents a decrease in the apparent potency of the native agonist. Data represent the mean ± SEM of at least two independent experiments (for full concentration-response curves see Figure S4.2).

Niclosamide, VU0048913, VU0048992, VU049150 and VU0118748 increase the potency of PP at the Y_4R (Figure 4.4, Figure S4.2) consistent with Y_4R PAM activity observed (Figure 4.3). Testing of other Y receptor subtypes revealed that the compounds are not fully selective for the Y_4R subtype. In addition to the Y_4R activity Niclosamide had a small PAM effect on the Y_1R. However, the structural analogs VU0048913, VU0118748, VU0048992 and VU0049150 had no effect on the Y_1R signal. In contrast to the PAM effects observed on Y_4R, Niclosamide and VU0118748 show a negative allosteric effect on Y_5R. All other tested compounds were inactive on the Y_5R. Whereas Niclosamide had no effect on the Y_2R, compounds VU0048913, VU0118748 and VU0048992 showed a slight PAM effect on the Y_2R. Overall, the effects of the Y_4R PAM compounds on Y_1R, Y_2R, and Y_5R were lower than the PAM effect at the Y_4R.

Niclosamide structure-activity relationships

Potentiation of PP EC_{20} experiments showed that Niclosamide, VU0118748 and VU0048913 have nearly identical potencies on the Y_4R. Accordingly, the loss of the Cl atom (VU0048913) on the aniline ring structure, as well as modification of the OH group in the benzoyl ring (VU0118748) are not affecting

Y_4R PAM activity. In contrast, a major change of the substituents in meta or para position to the hydroxyl function on the benzoyl ring lead to a complete loss of Y_4R PAM activity (VU0357475, VU0114795) or drastically reduce Y_4R potency (VU0048992, VU0049150). As shown by the active compound VU0048913, the Cl on the benzoyl ring structure can be substituted by Br, which underlines the importance of the electron-rich character in this position for the potency to the Y_4R (Figure 4.3 and Figure 4.5).

Furthermore, investigation of PAM activity on Y_1R, Y_2R and Y_5R suggested potential modification sites to control YR selectivity. Removal of the Cl substitution on the aniline ring in VU0048913 reduced Y_1R effects and Y_5R antagonism. As shown by compound VU0118748, selectivity towards Y_1R can also be achieved by the introduction of the nitrobenzoic-acid on the OH position in the benzoyl ring. However, this modification had no effect on the Y_5R NAM activity and Y_4R PAM activity, suggesting this position as a potential site to control subtype selectivity.

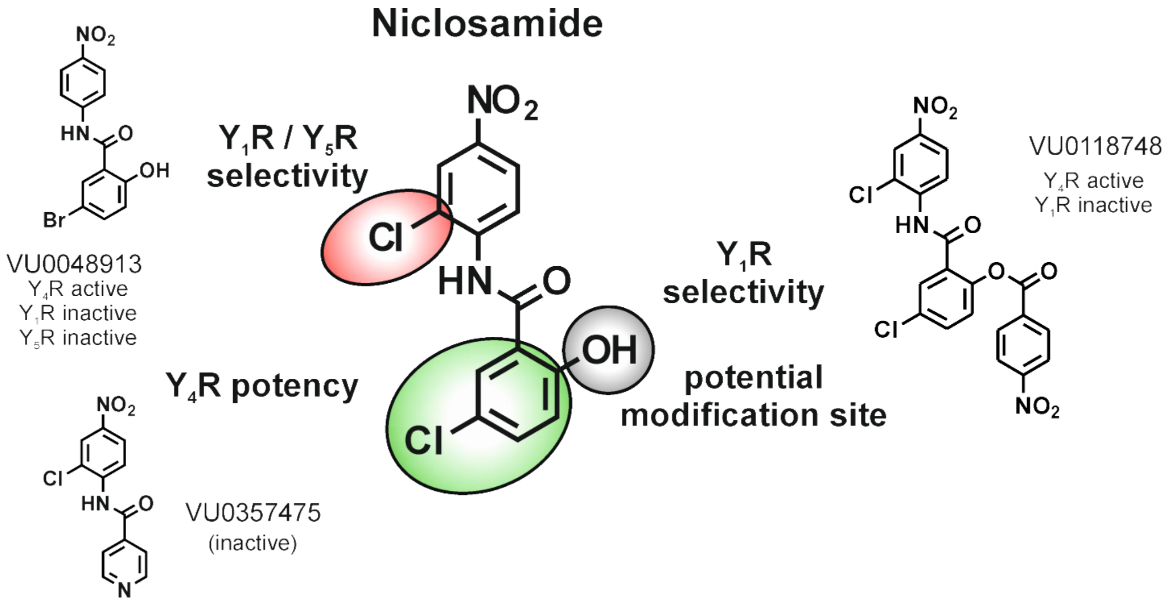


Figure 4.5 Distinct positions of the Niclosamide scaffold were shown to be relevant for Y4R PAM activity and YR selectivity. Substitutions in the benzoyl ring are important for Y4R potency (green), and offer a potential modification site (grey). Modifications in the aniline ring engender selectivity towards Y1R / Y5R subtype (red).

4.4 Discussion

The application of allosteric modulators presents a promising approach for the treatment of complex receptor-ligand systems that regulate sensitive physiological processes such as nervous signal transduction or metabolic regulation [21, 29].

Therapeutic potential

The benefits of Y_4R modulation on obesity and insulin resistance are becoming more important as the number of patients diagnosed with type 2 diabetes rises alongside risk factors such as the prevalence of obesity, physical inactivity, and poor diet [30]. In humans, low circulating PP levels were found in obese children and adults [31]. Hyperphasia in obese patients can be reduced by restoring basal and meal-stimulated PP levels through IV infusion [32]. Additionally, PP is hypothesized to sensitize the liver to insulin through upregulation of the insulin receptor β subunit [33, 34]. In patients with diabetes secondary to chronic pancreatitis, PP administration reduces insulin resistance and improves glucose metabolism [35]. Effects of PP administration and the complex interplay of obesity and diabetes suggest Y_4R modulation may be beneficial for a wide range of metabolic disorders. This is already being seen in preclinical studies of TM-30339, a PP based Y_4R-selective peptide agonist and phase I and II trials of Obineptide, an $Y_{2/4}R$ dual peptide agonist [36].

In this study, we present the first small molecule Y_4R PAM. Niclosamide was identified as an Y_4R PAM in the primary screen of the Spectrum Collection using the Ca^{2+} flux assay and its activity was validated in the alternative IP accumulation assay. The second HTS experiment identified four additional Y_4R PAMs that are structurally similar to Niclosamide, confirming the importance of this scaffold for Y_4R potentiation. YR subtype selectivity was characterized for four Niclosamide-like Y_4R PAMs along with two Y_4R inactive Niclosamide-derived structures.

Y_4R potency and selectivity

Investigation of the potency of the compounds on the Y_4R showed that three compounds, Niclosamide, VU0118748 and VU0048913, had comparable EC_{50} values of around 500 nM. All other compounds lacking the halogen substitution on the benzoyl ring were either completely inactive at the Y_4R or had an at least 10-fold lower EC_{50} for potentiation of a PP EC_{20} (Figure 4.3). These investigations highlight the role of an electron-rich substituent at this position for Y_4R potency. In contrast, the bulky nitro-benzoyl substitution in compound VU0118748 had no influence on Y_4R potency and Y_4R PAM activity. However, this modification lowers the effect at Y_1R. This offers a role for this position as a potential modification site for improving Y_4R selectivity while maintaining the Y_4R PAM activity. Furthermore, this position could be used for linking the modulator to the native ligand PP to create a bitopic ligand, as already performed for mAChRs and adenosine receptors [37-39]. Bitopic ligands can have the advantage over pure allosteric modulators by taking advantage of the increased selectivity without relying on the presence of endogenous agonist.

Selectivity studies of Niclosamide and analogs on all four Y receptor subtypes revealed that the compounds are not fully selective for the Y_4R. Niclosamide and VU0118748, both Y_4R PAMs, showed antagonistic effects on Y_5R. The Y_4R and Y_5R fulfill different actions in the regulation of appetite and food intake. While the Y_4R has an anorexigenic effect by inducing satiety in response to the activation of the native ligand PP, the Y_1R and Y_5R are characterized to have an orexigenic effect[40]. A simultaneous Y_5R antagonism and Y_4R PAM activity thus could contribute to an anti-obesity effect of Niclosamide or VU0118748. However, the different analogs of the Niclosamide scaffold suggest the possibility of developing Y_4R PAMs with a higher degree of specificity relative to other Y receptor subtypes (Figure 4.5). It is not uncommon for compounds to produce a variety of effects at a common binding site. For example, recently a known metabotropic glutamate receptor 4 (mGluR4) PAM/mGluR1 NAM chemotype was converted into a selective mGluR1 PAMs by virtue of a double “molecular switch” [41].

Would an EC_{50} shift of 4 fold be sufficient to cause an *in vivo* effect?

Niclosamide and some related analogs induced a 4-fold increase in the potency of PP at Y_4R. Investigated allosteric modulators of other class A GPCRS, especially modulation of neurotransmitter response, display a stronger allosteric effects with EC_{50} shifts >10 fold, shown for muscarinic receptor 4 (mAChR4) [42]. However, allosteric modulation of the CaSR by cinacalcet shows that even smaller *in vitro* effects can be effective *in vivo* and that clinical efficacy is dependent on the receptor, tissue and the metabolic state that is targeted [43]. This suggests that the comparatively small increase in PP potency caused by Niclosamide may be sufficient to elicit *in vivo* effects.

Niclosamide improves diabetic symptoms in mice

Niclosamide is an FDA approved anthelmintic drug that treats parasitic worm infection through the uncoupling of mitochondria. Interestingly, this compound was recently studied as a potential therapeutic for treatment of type 2 diabetes due to its high tolerability and the benefits of lipid mitochondrial uncoupling for treating diabetes [44]. Tao et al. fed mice the ethanolamine salt form of

Niclosamide and showed it to be efficacious at high nanomolar concentrations (measured with blood sample liquid chromatography–tandem mass spectrometry at various time points) in reducing plasma insulin decline in db/db mice, sensitizing the insulin response, and preventing and treating diabetic symptoms during high fat diet induced obesity in mice. The authors focus on mitochondrial uncoupling as the primary mechanism of action for Niclosamide in the treatment of diabetes symptoms. However, our identification of Niclosamide as an Y_4R PAM suggests that the efficacy of Niclosamide on diabetic symptoms may result from its action on the YR signalling in addition to effects on mitochondrial function.

4.5 Materials and Methods

Cell Culture

COS-7 cells stably expressing $hY_{1/2/4/5}R$_eYFP fusion protein and the $\Delta 6G\alpha_{qi4myr}$ chimeric Gα-protein were prepared as previously described [45]. The $hY_{1/2/4/5}R$_eYFP cDNA was subcloned into MCS1 of a pVitro2-MCS vector carrying a hygromycin resistance gene; the $\Delta 6G\alpha_{qi4myr}$ cDNA was subcloned into MCS1 of a pVitro2-MCS vector carrying a G418 resistance gene (Invivogen).

COS-7 (African Green Monkey kidney) cells stably expressing the $hY_{1/2/4/5}R$_eYFP fusion protein and the $\Delta 6G\alpha_{qi}$4myr chimeric Gα-protein were cultured at 37 °C in high glucose Dulbecco's Modified Eagle Medium (DMEM, Life Technologies) with glutamine and sodium pyruvate (Life Technologies/Lonza) supplemented with 10% FBS (Invitrogen), 1.5 mg/mL G418-sulfate (Amresco) and 133 µg/ml hygromycin (Invivogen).

HTS Calcium Flux Assay

A total of 35,288 compounds were tested for their ability to modulate the Y_4R activation in conjunction with the Vanderbilt HTS facility. In a pilot HTS experiment, 2000 compounds (spectrum collection, MicroSource Discovery Systems, Inc.) were screened for modulation of the Y_4R. This collection is designed to enrich the general hit rate by including drugs with known biological profiles (60%), naturally occurring products with no biological profile (25%), and non-drug compounds with biological profiles (15%). In a following HTS, 33,288 compounds were tested for Y_4R modulatory effects. Thirty-two thousands of these compounds were randomly selected from the Vanderbilt compound library and 1288 were selected based on their similarity to Niclosamide, a PAM discovered in the pilot HTS.

Cells were plated in TC-treated 384-well plates (black, clear bottom, Greiner) in 20 µL cell culture medium using a Multidrop Combi (Thermo Fisher) microplate dispenser (ThermoScientific, Thermo Fisher) at a density of 16,000 cells/well. The cells were incubated for 24 hours at 37 °C in the presence of 5% CO_2. Following incubation, the medium was replaced with 20 µL/well fluorescent dye solution (1.0 µM Fluo-2 AM (TEFlabs), .01% Pluronic Acid F-127 in assay buffer) using an ELx405 cell washer (BioTek). Following a 90 minute incubation at room temperature, fluorescent dye solution was replaced with 20 µL/well assay buffer (HBSS, 20 mM HEPES, and 1.25 mM Probenecid (Sigma-Aldrich)) using the ELx405 and cell plates were loaded into a Functional Drug Screening System (FDSS, Hamamatsu). Once loaded into the FDSS, cell plates were imaged at 1Hz (excitation 470 ± 20 nm, emission 540 ± 30 nm using a 3-addition protocol designed to detect agonists, potentiators, and inhibitors: 1)after collecting 4 seconds of baseline, 20 µl/well of 20 µM test compounds in assay buffer + 0.1% fatty-acid-free bovine serum albumin (Sigma-Aldrich, modified assay buffer) were added 2) following a 150 second delay , 10 µl/well of concentration of 5-fold over the PP EC_{20} (55 ± 27 pM) 3) after 330 seconds, a 13 µl/well addition 5-fold over the PP EC_{80} (836 ± 33 pM) in modified assay buffer was performed. On each screening day, PP EC_{20} and EC_{80} plates were adjusted after a test PP CRC at the beginning of each day to account for minor day to day variations in experimental conditions.

Substructure Search

Substructure searches were performed against the Vanderbilt Institute for Chemical Biology (VICB) library using the ChemCart application (DeltaSoft Inc.) Tanimoto coefficient similarity search. Higher Tanimoto coefficients indicate more similarity based on the shared presence of chemical subgroups. To increase our chemical search space, we altered the amide linker of Niclosamide and repeated the substructure search against the VICB library. Linker alterations included replacing the amide linker with urea, thiourea, δ-lactam, and an extension of the linker by one or two methylene groups. Tanimoto coefficient cut-offs were adjusted to between 0.45 and 0.63 for each substructure search to ensure that approximately 200 to 300 compounds were identified for each scaffold. Table S4-1 lists the chemical structures and parameters used for all similarity searches. Reference structures with alternate backbone constitutions were generated using ChemBioDraw Ultra (PerkinElmer Inc.). Final search results were concatenated and duplicate search hits were removed. The final collection of 1288 compounds were distributed over five plates and tested using the triple-add screen protocol.

IP_3 Assay

Cells were seeded into 48-well plates and incubated for 24 hours at 37°C / 5% CO_2. Next, cells were labeled for at least 16 hours in DMEM+10% FBS containing 2 µCi/ml *myo*-[2- ^{3}H(N)]-inositol (PerkinElmer) at 37°C / 5% CO_2. Labeling solution was aspirated and cells were washed with 20 µL/well DMEM + 10 mM LiCl (Sigma-Aldrich; DMEM/LiCl) and stimulated with the peptide solutions and compounds. Therefore, 50 µl/well DMEM/LiCl were added after washing, followed by addition of 50 µl/well test compound in DMEM/LiCl (3-fold over the final concentration) and addition of 50 µl/well peptide solution (3-fold over the final concentration) in DMEM/LiCl. Stimulation was performed for 2 hours at 37°C / 5% CO_2. Cell lysis, subsequent sample preparation and radiometric detection was performed as described previously [45].

Data analysis

Data analysis was performed with GraphPad Prism 5.03 software (GraphPad Software) using standard non-linear regression (log(agonist) vs. response, three parameters). All data were normalized to the corresponding control curve in the absence of the modulator. EC_{50} ratios were calculated from global concentration response curves (EC_{50} shift function) summarized from the data of at least 3 independent experiments (row means totals function).

Acknowledgments

This research was supported by NIH grant R01 DK097376 to JM, DW and AGB-S, by the DFG (Be1264-16), by the European Union and the Free State of Saxony (100148835 and 1423213128452). All authors certify that the content of this manuscript presents no conflict of interest with any financial organization.

4.6 References

1. Finkelstein EA, Trogdon JG, Cohen JW, & Dietz W (2009) Annual medical spending attributable to obesity: payer-and service-specific estimates. *Health affairs* 28(5):w822-831.
2. Troke RC, Tan TM, & Bloom SR (2014) The future role of gut hormones in the treatment of obesity. *Therapeutic advances in chronic disease* 5(1):4-14.
3. Makris A & Foster GD (2011) Dietary approaches to the treatment of obesity. *The Psychiatric clinics of North America* 34(4):813-827.
4. Kissler HJ & Settmacher U (2013) Bariatric surgery to treat obesity. *Seminars in nephrology* 33(1):75-89.
5. Smith BR, Schauer P, & Nguyen NT (2008) Surgical approaches to the treatment of obesity: bariatric surgery. *Endocrinology and metabolism clinics of North America* 37(4):943-964.

6. Heppner KM & Perez-Tilve D (2015) GLP-1 based therapeutics: simultaneously combating T2DM and obesity. *Frontiers in neuroscience* 9:92.
7. Field BC, Chaudhri OB, & Bloom SR (2010) Bowels control brain: gut hormones and obesity. *Nature reviews. Endocrinology* 6(8):444-453.
8. Pedragosa-Badia X, Stichel J, & Beck-Sickinger AG (2013) Neuropeptide Y receptors: how to get subtype selectivity. *Frontiers in endocrinology* 4:5.
9. Asakawa A*, et al.* (2003) Characterization of the effects of pancreatic polypeptide in the regulation of energy balance. *Gastroenterology* 124(5):1325-1336.
10. Jesudason DR*, et al.* (2007) Low-dose pancreatic polypeptide inhibits food intake in man. *British Journal of Nutrition* 97(03):426-429.
11. le Roux CW*, et al.* (2006) Attenuated peptide YY release in obese subjects is associated with reduced satiety. *Endocrinology* 147(1):3-8.
12. Ekblad E & Sundler F (2002) Distribution of pancreatic polypeptide and peptide YY. *Peptides* 23(2):251-261.
13. Cox HM (2007) Neuropeptide Y receptors; antisecretory control of intestinal epithelial function. *Autonomic neuroscience : basic & clinical* 133(1):76-85.
14. Kojima S*, et al.* (2007) A role for pancreatic polypeptide in feeding and body weight regulation. *Peptides* 28(2):459-463.
15. Holzer P, Reichmann F, & Farzi A (2012) Neuropeptide Y, peptide YY and pancreatic polypeptide in the gut–brain axis. *Neuropeptides* 46(6):261-274.
16. Katsuura G, Asakawa A, & Inui A (2002) Roles of pancreatic polypeptide in regulation of food intake. *Peptides* 23(2):323-329.
17. Acuna-Goycolea C, Tamamaki N, Yanagawa Y, Obata K, & van den Pol AN (2005) Mechanisms of Neuropeptide Y, Peptide YY, and Pancreatic Polypeptide Inhibition of Identified Green Fluorescent Protein-Expressing GABA Neurons in the Hypothalamic Neuroendocrine Arcuate Nucleus. *The Journal of Neuroscience* 25(32):7406-7419.
18. Asakawa A*, et al.* (1999) Mouse pancreatic polypeptide modulates food intake, while not influencing anxiety in mice. *Peptides* 20(12):1445-1448.
19. Misra S, Murthy KS, Zhou H, & Grider JR (2004) Coexpression of Y1, Y2, and Y4 receptors in smooth muscle coupled to distinct signaling pathways. *The Journal of pharmacology and experimental therapeutics* 311(3):1154-1162.
20. Ueno N*, et al.* (1999) Decreased food intake and body weight in pancreatic polypeptide-overexpressing mice. *Gastroenterology* 117(6):1427-1432.
21. Conn PJ, Lindsley CW, Meiler J, & Niswender CM (2014) Opportunities and challenges in the discovery of allosteric modulators of GPCRs for treating CNS disorders. *Nature reviews. Drug discovery* 13(9):692-708.
22. Shonberg J*, et al.* (2014) Biased agonism at G protein-coupled receptors: the promise and the challenges--a medicinal chemistry perspective. *Medicinal research reviews* 34(6):1286-1330.
23. Keov P, Sexton PM, & Christopoulos A (2011) Allosteric modulation of G protein-coupled receptors: A pharmacological perspective. *Neuropharmacology* 60(1):24-35.
24. Kenakin TP (2012) Biased signalling and allosteric machines: new vistas and challenges for drug discovery. *British Journal of Pharmacology* 165(6):1659-1669.
25. Gregory KJ, Dong EN, Meiler J, & Conn PJ (2011) Allosteric Modulation of Metabotropic Glutamate Receptors: Structural Insights and Therapeutic Potential. *Neuropharmacology* 60(1):66-81.
26. Zhang R & Xie X (2012) Tools for GPCR drug discovery. *Acta pharmacologica Sinica* 33(3):372-384.

27. Kaiser A*, et al.* (2015) Unwinding of the C-Terminal Residues of Neuropeptide Y is critical for Y2 Receptor Binding and Activation. *Angewandte Chemie International Edition*:n/a-n/a.
28. Pedragosa-Badia X*, et al.* (2014) Pancreatic polypeptide is recognized by two hydrophobic domains of the human Y4 receptor binding pocket. *The Journal of biological chemistry* 289(9):5846-5859.
29. Wang L, Martin B, Brenneman R, Luttrell LM, & Maudsley S (2009) Allosteric modulators of g protein-coupled receptors: future therapeutics for complex physiological disorders. *The Journal of pharmacology and experimental therapeutics* 331(2):340-348.
30. Ershow AG (2009) Environmental influences on development of type 2 diabetes and obesity: challenges in personalizing prevention and management. *Journal of diabetes science and technology* 3(4):727-734.
31. Yulyaningsih E, Zhang L, Herzog H, & Sainsbury A (2011) NPY receptors as potential targets for anti-obesity drug development. *Br J Pharmacol* 163(6):1170-1202.
32. Berntson GG, Zipf WB, O'Dorisio TM, Hoffman JA, & Chance RE (1993) Pancreatic polypeptide infusions reduce food intake in Prader-Willi syndrome. *Peptides* 14(3):497-503.
33. Seymour NE, Volpert AR, & Andersen DK (1996) Regulation of hepatic insulin receptors by pancreatic polypeptide in fasting and feeding. *The Journal of surgical research* 65(1):1-4.
34. Inui A (2000) Transgenic approach to the study of body weight regulation. *Pharmacological reviews* 52(1):35-61.
35. Brunicardi FC*, et al.* (1996) Pancreatic polypeptide administration improves abnormal glucose metabolism in patients with chronic pancreatitis. *The Journal of clinical endocrinology and metabolism* 81(10):3566-3572.
36. Sato N, Ogino Y, Mashiko S, & Ando M (2009) Modulation of neuropeptide Y receptors for the treatment of obesity. *Expert Opin Ther Pat* 19(10):1401-1415.
37. Lane JR, Sexton PM, & Christopoulos A (2013) Bridging the gap: bitopic ligands of G-protein-coupled receptors. *Trends in pharmacological sciences* 34(1):59-66.
38. Narlawar R*, et al.* (2010) Hybrid Ortho/Allosteric Ligands for the Adenosine A1 Receptor. *Journal of medicinal chemistry* 53(8):3028-3037.
39. Steinfeld T, Mammen M, Smith JAM, Wilson RD, & Jasper JR (2007) A Novel Multivalent Ligand That Bridges the Allosteric and Orthosteric Binding Sites of the M2 Muscarinic Receptor. *Molecular pharmacology* 72(2):291-302.
40. Moreno-Herrera A, Garcia A, Palos I, & Rivera G (2014) Neuropeptide Y1 and Y5 Receptor Antagonists as Potential Anti-Obesity Drugs. Current Status. *Mini reviews in medicinal chemistry*.
41. Cho HP*, et al.* (2014) Chemical modulation of mutant mGlu1 receptors derived from deleterious GRM1 mutations found in schizophrenics. *Acs Chem Biol* 9(10):2334-2346.
42. Chan WY*, et al.* (2008) Allosteric modulation of the muscarinic M(4) receptor as an approach to treating schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 105(31):10978-10983.
43. Davey AE*, et al.* (2011) Positive and Negative Allosteric Modulators Promote Biased Signaling at the Calcium-Sensing Receptor. *Endocrinology* 153(3):1232-1241.
44. Tao H, Zhang Y, Zeng X, Shulman GI, & Jin S (2014) Niclosamide ethanolamine-induced mild mitochondrial uncoupling improves diabetic symptoms in mice. *Nature medicine* 20(11):1263-1269.
45. Mäde V, Bellmann-Sickert K, Kaiser A, Meiler J, & Beck-Sickinger AG (2014) Position and Length of Fatty Acids Strongly Affect Receptor Selectivity Pattern of Human Pancreatic Polypeptide Analogues. *ChemMedChem* 9(11):2463-2474.

4.7 Supplementary Information

Figure S4.1 Validation Y4R PAM activity and YR subtype selectivity of initial Ca2+ HTS hit compounds in an IP3 assay. Structurally different small molecules (A) showed a positive effect on Y4R Ca2+ signal response in an HTS screening of the spectrum collection. Retesting in the IP3 assay as an alternative YR activation readout validated Niclosamide as a Y4R PAM (B) and offered other hits to have additional effects on other YR subtypes. Submaximal activation of Y receptors was observed for stimulation with 1 nM ligand (Y4R: PP, Y1,2,5R: NPY) in presence of 10 µM compound. Data represent the mean ± SEM of two independent experiments performed in quadruplicates.

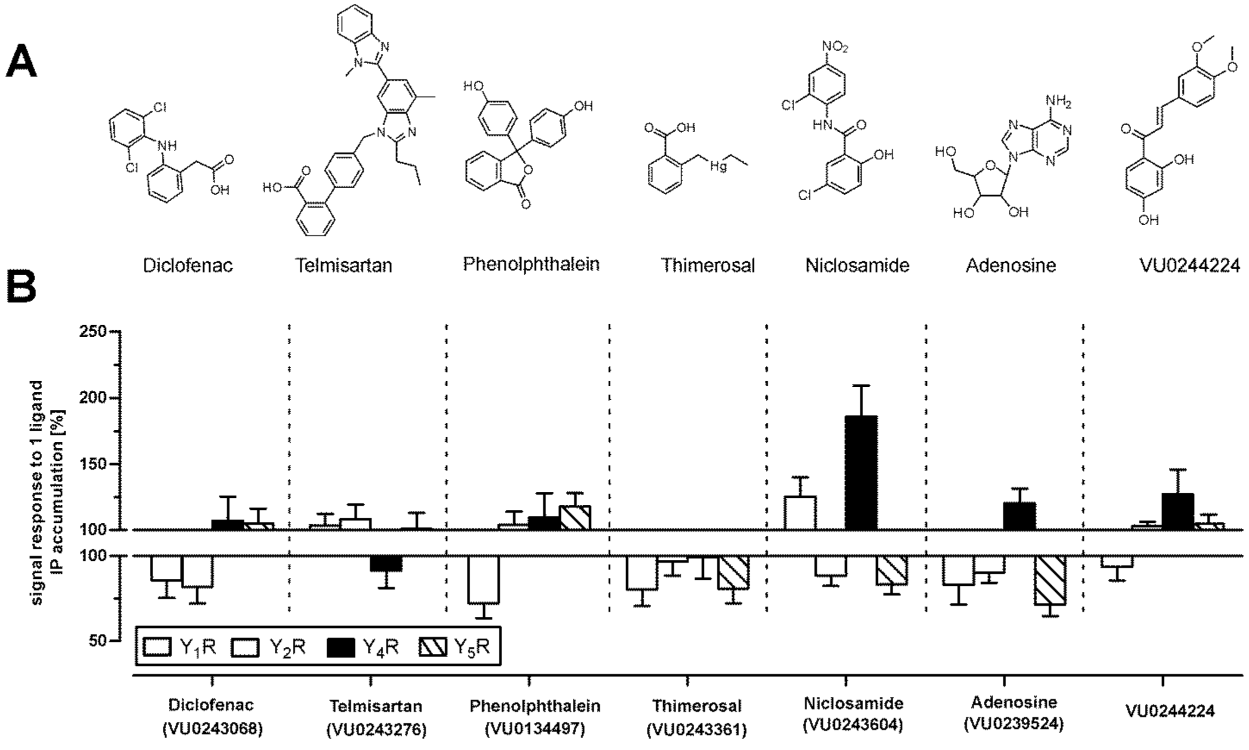


Figure S4.2 Selectivity of Niclosamide-like allosteric modulators among human Y receptors. Receptor activation was investigated with an inositol phosphate accumulation assay in COS-7 cells stably expressing a Y receptor subtype and chimeric G-protein ΔGα6qi4myr. Data represent the mean ± SEM of at least 2 independent experiments.

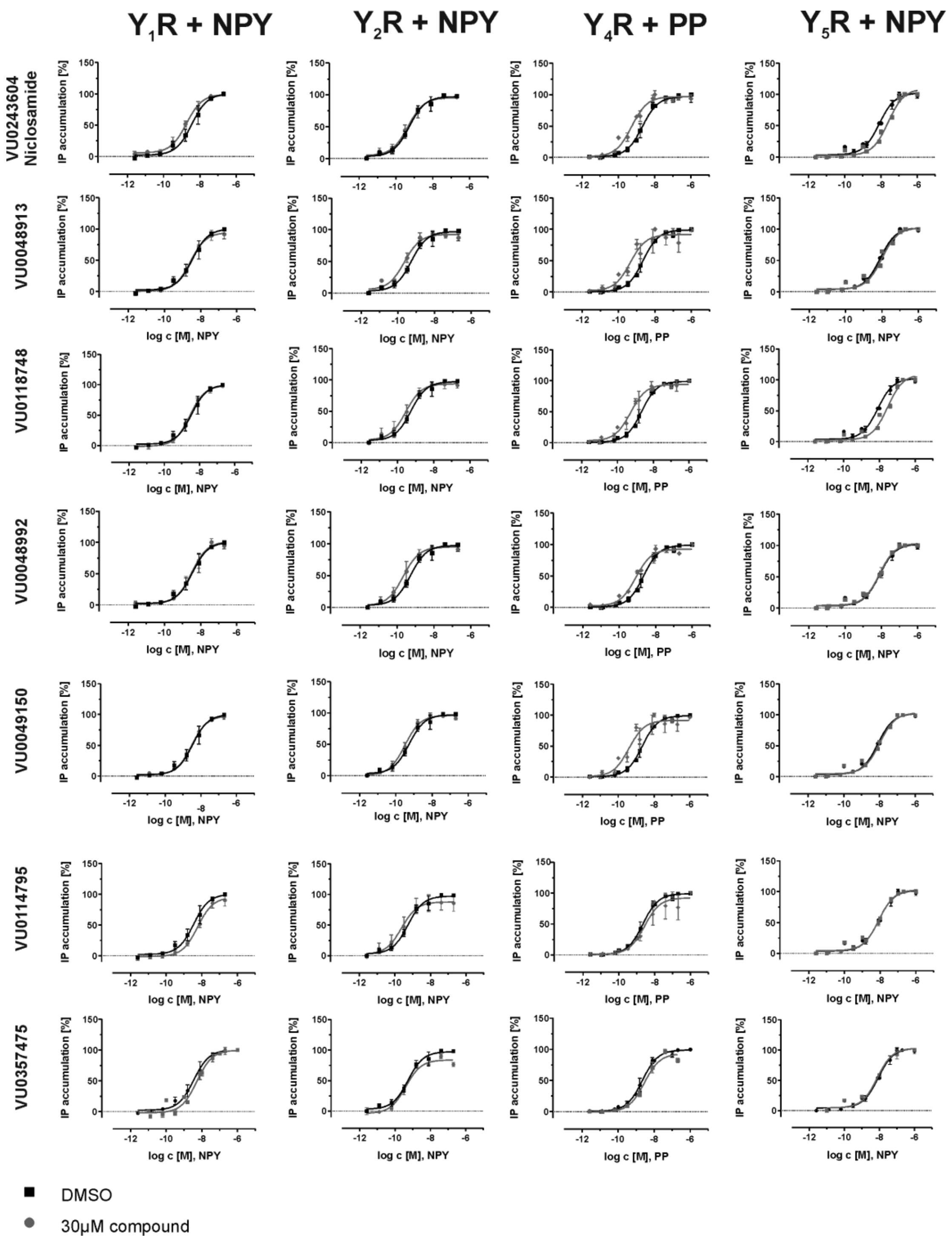
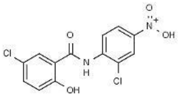
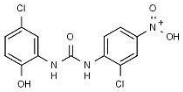
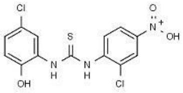
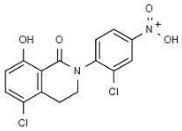
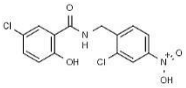
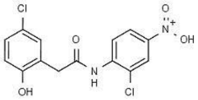


Table S4-1 Substructure search: Niclosamide analogues and results.

Backbone Modification	Search Structure	Tanimoto Cutoff	Total Compounds
None		0.63	380
Urea		0.50	144
Thiourea		0.45	396
δ-lactam		0.57	324
Methylene extension (a)		0.57	232
Methylene extension (b)		0.60	183

Chapter 5

Modeling Interactions of the Human Y_4 Receptor and Pancreatic Polypeptide

---

based on sections of publication in Journal of Biological Chemistry (2014) **9**, 5846-5859

Xavier Pedragosa-Badia, Gregory Sliwoski, Elizabeth Dong Nguyen, Diana Lindner, Jan Stichel, Kristian Kaufmann, Jens Meiler, Annette Beck-Sickinger: **Pancreatic polypeptide is recognized by two hydrophobic domains of the human Y4 receptor binding pocket.**

5.1 Abstract

This chapter begins with detailed background information and computational techniques used to generate comparative models of the Y_4R and subsequently dock PP to these models. Due to the lack of experimental Y receptor structure, comparative modeling with multiple class A GPCR templates was used to generate models of the Y_4R. Secondly, a two-step docking process was performed to model interactions between Y_4R and pancreatic polypeptide. Specific interactions elucidated through *in vitro* experiments used to guide docking include contacts between $Tyr^{2.64}$ of Y_4R and Tyr^{27} of PP, $Asn^{7.32}$ of Y_4R and Arg^{33} of PP, and interactions between $Phe^{7.35}$ and residues Arg^{33} and Tyr^{36} of PP. These experimental findings and high resolution models will contribute to the rational design of ligands with higher affinity and activity at the human Y_4 receptor.

5.2 Introduction

This chapter details the computational methods used to dock the peptide agonist pancreatic polypeptide (PP) to the human Y_4 receptor (Y_4R). In brief, comparative modeling was used to construct 3D models of Y_4R. Then, experimentally-derived restraints were used to guide docking of PP to these Y_4R comparative models.

Modeling the human Y_4 receptor with Rosetta Molecular Modeling Suite

Rosetta Molecular Modeling Suite is a computational modeling suite that contains tools for *de novo* structure prediction, comparative modeling, loop building, protein-protein docking, protein-ligand docking, and protein design [1-4]. It has been successfully applied to various projects including the creation of novel enzymes [5, 6], redesign of metalloenzymes for catalyzing new reactions [7], peptidomimetic design with noncanonical backbones [8], structure elucidation from sparse NMR data [9], NMR refinement [10], homology modeling [11, 12], and ligand bind site elucidation [13]. In favorable cases, Rosetta is capable of refining small proteins to near atomic resolution [14]. In addition to proteins, Rosetta is capable of modeling interactions with small-molecule ligand s[15], RNA [16], and DNA [17]. To improve sampling accuracy and model discrimination, Rosetta accepts experimental restraints from a variety of sources including NMR, EM, EPR, and mutagenesis, all of which have been shown to improve model quality in larger systems [18].

Protein structure prediction involves finding the lowest energy conformation from all potential conformations. This is known as the global minimum in the conformation energy landscape and is thought to represent the native protein structure [19]. Even with approximation, the extent of conformational space available to a polypeptide sequence exceeds current computational resources. To sample this massive conformational space efficiently, Rosetta uses a Monte Carlo (MC) based search designed to explore energy landscapes with multiple minima [20].

In lieu of computationally expensive quantum mechanical calculations, Rosetta uses an energy score approximation that combines statistical data from pre-existing protein structures with simplified physical energy terms. The total energy of a structure is the weighted sum of all energy terms. Although specific scoring terms may vary depending on application, the Rosetta energy score generally includes terms such as solvation (probability of seeing a particular amino acid with a given number of α-carbons within a given distance of another), electrostatics (probability of observing a given distance between amino acids), orientation-dependent hydrogen bonding potential, and 6-12 Lennard-Jones potential for high resolution energy functions [4].

Although powerful, Rosetta's *de novo* folding application is practical only for soluble proteins with 150 amino acids or fewer [21]. This is insufficient for G-protein coupled receptors (GPCRs) such as Y_4R that typically exceed 300 amino acids. A common approach to modeling larger proteins is comparative modeling. This method uses the structure of another, often homologous protein called the "template" to guide tertiary structure prediction. Conformational search space is reduced drastically by providing Rosetta with a scaffold over which to lay the initial backbone structure [22]. Comparative modeling has been successfully applied to structure-function relationship prediction, structure-based drug design, and site-directed mutagenesis [23-25].

Many high quality comparative models have been generated for class A GPCRs despite low sequence identity, especially when more than one template is used [26]. Shared topological characteristics of GPCRs that enable low sequence identity comparative modeling include seven transmembrane α-helices, three extracellular loops, three intracellular loops, an extracellular N-terminus and an intracellular C-terminus [27]. Despite this overall structural similarity, GPCRs can respond to a wide range of stimuli and effect diverse changes across specific cell types, a variability that is in part due to the intrinsically disordered loops and termini [28].

Disordered loop regions present the greatest challenge to comparative modeling GPCRs. These regions comprise the lowest degree of sequence similarity between different class A GPCRs and play important functional roles that help determine a particular receptor's unique behavior. Extracellular loop two (ECL2) presents one of the greatest challenges to class A GPCR comparative modeling due to its high sequence variability and length that often exceeds 12 residues [29]. However, some structural features of ECL2 may be conserved between template and target and used to guide modeling. A disulfide bond typically tethers ECL2 to the third transmembrane helix. Additionally, ECL2 may adopt secondary structure conformations such as the β-strands seen in rhodopsin [30]. Some class A GPCRs contain a second intra-loop disulfide bond in ECL3 that serves to further limit conformational freedom [30].

Rosetta is capable of low resolution loop modeling followed by high resolution refinement. Low resolution modeling in Rosetta refers to the simplification of residue side chains into single-body "centroids" and the specialized low resolution scoring term that models solvation, electrostatics, hydrogen bonding between β-strands and steric clashes [4]. This low resolution representation smooths the energy landscape for improved sampling at the expense of accuracy [4]. Initially, loops are represented as a linear sequence of peptide “fragments” three or nine amino acids long [31]. Fragment conformations are retrieved from a database of experimentally determined conformations using local alignment queries [32]. This allows Rosetta to locally sample a wide variety of local conformational space that reflects known protein structures.

Cyclic coordinate descent (CCD) [33] is used to close loops. Inspired by inverse kinematic applications in robotics, CCD minimizes the sum of the squared distances between three backbone atoms of the loop’s N-terminal and three backbone atoms of the fixed C-terminal anchor. Dihedral backbone angles are adjusted and evaluated iteratively until the loop is closed [3]. This method of loop closure is advantageous because of its speed and ability to close loops in 99% of instances tested and has been shown to outperform loop modeling in other protein modeling applications [22, 34]. Following loop closure, Rosetta uses another robotics inspired method called kinematic closure (KIC) to refine the loop conformation [35].

High resolution models are obtained by replacing centroids with full-atom side chains conformations. Because systematic evaluation of all side chain degrees of freedom is intractable [36], Rosetta limits the number of side chain conformations sampled based on those observed in the Protein Data Bank (PDB). A rotamer is a specific set of chi angles derived from statistical analysis of the PDB that represents a likely side chain conformation [37]. Rotamer libraries define biologically probable conformations for all amino acids and capture likely conformations both within single side chains and between side chains of a given sequence. Experimental evidence may also be used to influence rotamer selection [38]. Sidechain modifications are combined iteratively with backbone modifications to determine the combination of rotamers occupying the global minimum of the energy function [1, 4].

The final step of comparative modeling is a whole-model all-atom refinement called “relax” [39, 40]. The overall goal is to explore local conformational space and move the protein structure into an energetic minimum. During relax, local interactions are improved with iterative side-chain (rotamer) selection and gradient-based minimization. The global conformation of the protein is maintained while random backbone angle perturbations are sampled along with rigid body degrees of freedom and rotamer conformations. This is followed by a gradient minimization over all torsional degrees of freedom including phi, psi, omega, and kappa to resolve clashes and reach an energy minimum.

Rosetta's implicit membrane potential

Membrane proteins such as GPCRs function in a unique environment compared to soluble proteins [41]. This environment directly influences the topological features of the membrane protein fold. Hydrophobic amino acids such as leucine, isoleucine, valine, and phenylalanine favor the lipid exposed environment within the hydrophobic layer of the membrane over the protein-buried environment and small side-chain amino acids such as glycine, alanine, serine, and threonine favor helix-helix interfaces [42]. Additionally, large polar and cation-π interactions are more frequent in helical transmembrane proteins than soluble proteins and inter-helical hydrogen bonds result in tighter packing for transmembrane helices [42].

Rosetta uses an implicit membrane representation defined as a static five-layered, 60 Å wide membrane. These layers approximate the water-exposed, polar, interface, inner and outer hydrophobic layers of the membrane. Computational applications that employ algorithms such as Hidden Markov Models [43] are used to predict the membrane topology of a protein based on its sequence. Rosetta uses this information to predict the membrane spanning segments of the protein and classifies residues into one of eight burial states. Specialized membrane scoring terms evaluate each residue differently depending on their burial and layer state [42].

Contact restraints guide PP docking to Rosetta

Experimental restraints provide critical information for docking PP to the Y_4R comparative models. Experiments performed by Xavier Pedragosa-Badia of the Annette Beck-Sickinger lab identified $Tyr^{2.64}$, $Asp^{2.68}$, $Asn^{6.55}$, $Asn^{7.32}$, and $Phe^{7.35}$ as members of the hY_4R binding pocket [12]. Furthermore, hPP analogs with modifications in residues 27, 33, or 36 revealed these positions as interaction partners with the receptor. The presented work reflects a strong collaboration between the Annette Beck-Sickinger lab at Leipzig University and the Jens Meiler lab at Vanderbilt University. The results presented in this chapter focus on the presenting author's contribution to the computational modeling of Y_4R and interactions with PP. Crucial experimental evidence used to guide these models is detailed in the publication by Pedragosa-Badia, *et al* [12]. However, a summary of specific experimental results contributing to the definition of modeling restraints is included in table 5-1.

Table 5-1 Experimental evidence used to define each restraint as previously published. ND = not determined.

Restraint: Tyr$^{2.64}$ – Tyr^{27}

Y_4R Mutation	hPP Mutation	EC_{50} ratio to wild type
Tyr$^{2.64}$Ala	--	65
Tyr$^{2.64}$Ala	[Ala^{27}]hPP	424
Tyr$^{2.64}$Ala	[Leu^{27}]hPP	63
Tyr$^{2.64}$Ala	[Cha^{27}]hPP	14
Tyr$^{2.64}$Leu	[Ala^{27}]hPP	138
Tyr$^{2.64}$Leu	[Leu^{27}]hPP	21
Tyr$^{2.64}$Leu	[Cha^{27}]hPP	23

Restraint: Asn$^{7.32}$ – Arg^{33}

Y_4R Mutation	hPP Mutation	PP EC_{50} ratio to wild type
Asn$^{7.32}$Ala	--	5
Asn$^{7.32}$Ala	[Lys^{33}]hPP	60
Asn$^{7.32}$Ala	[ADMA^{33}]hPP	979
Asn$^{7.32}$Ala	[SDMA33]hPP	1892
Asn$^{7.32}$Arg	--	18

Restraint: Phe$^{7.35}$ – Arg^{33}

Y_4R Mutation	hPP Mutation	PP EC_{50} ratio to wild type
Phe$^{7.35}$Ala	--	8
Phe$^{7.35}$Ala	[ADMA^{33}]hPP	107
Phe$^{7.35}$Ala	[Lys^{33}]hPP	451
Phe$^{7.35}$Ile	--	41
Phe$^{7.35}$Ile	[ADMA^{33}]hPP	*ND*
Phe$^{7.35}$Ile	[Lys^{33}]hPP	762

Restraint: Phe$^{7.35}$ – Tyr^{36}

Y_4R Mutation	hPP Mutation	PP EC_{50} ratio to wild type
Phe$^{7.35}$Ala	[Ile^{36}]hPP	*ND*
Phe$^{7.35}$Ala	[Phe^{36}]hPP	17
Phe$^{7.35}$Ala	[Cha^{36}]hPP	138
Phe$^{7.35}$Ala	[Nle^{36}]hPP	679

5.3 Results

A final ensemble of Y_4R comparative models draws from all templates

Comparative modeling was performed in parallel for fourteen templates due to the low sequence identity between templates and target. Results in each template were examined to determine whether one or more template provided models with the lowest energy poses and therefore represented the most suitable templates. Table 5-2 details the templates used and lists the resulting low energy model scores within each template. As shown, all templates produced models with comparable energy scores. Based on these results, top models from each template were selected to represent an ensemble of 37 Y_4R model conformations instead of focusing on models from a subset of templates. This ensemble covers a conformational space of an average RMSD of 5.4 ± 0.8 angstroms. As expected, the majority of this conformational variability was found within the loop regions. When comparing the transmembrane helix regions, the average conformation RMSD drops to 1.9 ± 0.6 angstroms. Consistent topological features across all conformations include a very slight bend in helix 1 near $Gly^{1.46}$, a bend in helix 2 near $Pro^{2.59}$, a distortion in helix 4 near $Pro^{4.59}$, a bulge in helix 5 above $Pro^{5.50}$, a bend in helix 6 near $Pro^{6.50}$, and distortions in helix 7 above the conserved NPxxY motif. The general topology of Y_4R comparative models is shown in figure 5.1.

Table 5-2 Fourteen class A GPCR templates show low sequence identity and comparable model performance with Y4R. REU = Rosetta energy units.

PDB ID	Receptor	Y_4R sequence identity (%)	Top pose score (REU)	Average top 100 pose scores (REU)
1u19	Bovine rhodopsin	22	-777.6	-768.7
2rh1	Human β2 adrenergic	23	-830.7	-798.3
2vt4	Turkey β1 adrenergic	22	-788.5	-782.7
3eml	Human A2A	26	-816.0	-803.2
3odu	Human CXCR4	24	-772.8	-763.0
3pbl	Human D3	26	-816.6	-798.1
3rze	Human H1	22	-785.2	-772.5
3uon	Human M2	23	-792.0	-780.8
3v2w	Human S1P1	25	-753.9	-744.2
4daj	Rat M3	24	-799.9	-788.1
4djh	Human κ-opioid	25	-804.3	-787.5
4dkl	Mouse μ-opioid	24	-805.9	-793.2
4ea3	Human N/OFQ opioid	27	-810.2	-782.4
4ej4	Mouse δ-opioid	26	-792.6	-777.4

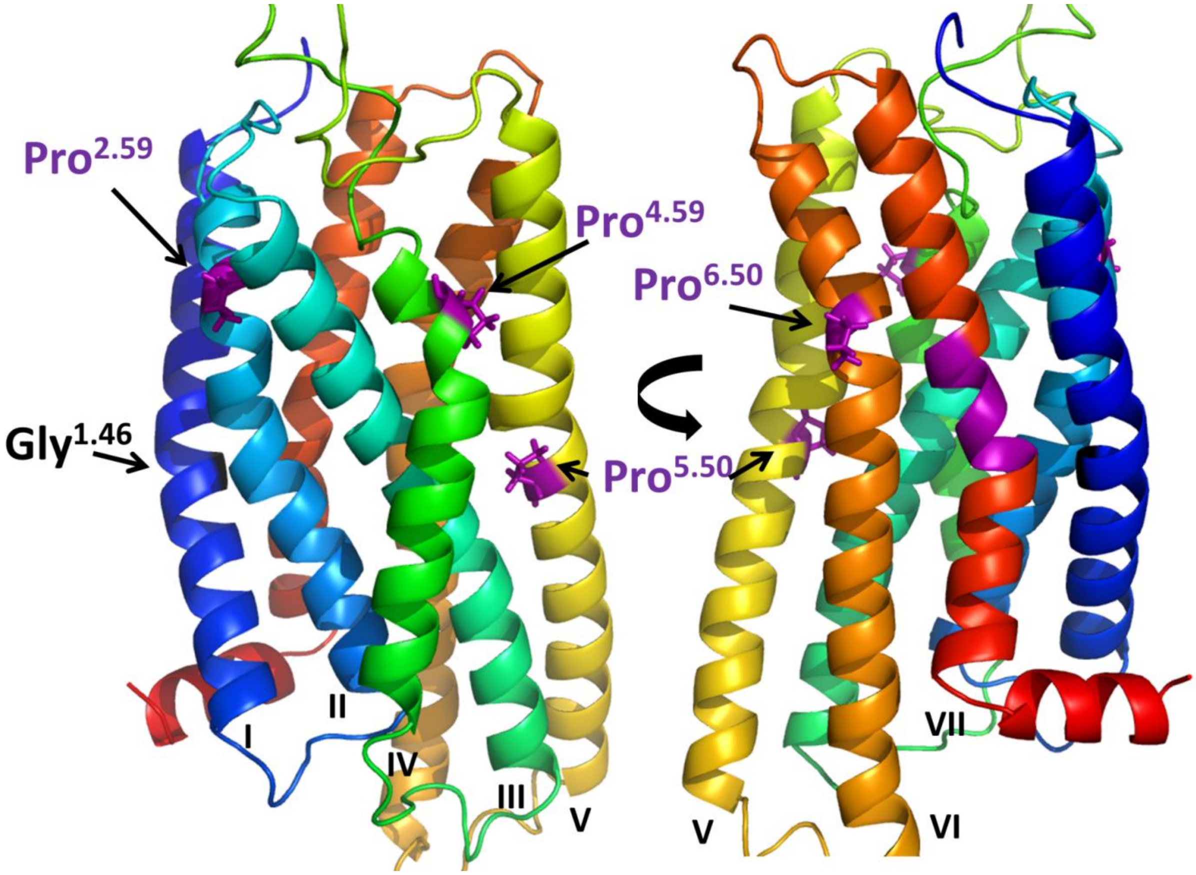


Figure 5.1 Y4R comparative model ensemble contains consistent topological features. Two views of a representative Y4R comparative model are shown and common helix distortions are highlighted along with nearby contributing residue. Helices are numbered at the intracellular end.

Restraints guide pancreatic polypeptide docking

Pancreatic polypeptide was docked into the comparative model of Y_4R to assist interpretation of experimental results. The initial placement of the PP helix was guided specifically by the altered activity of Y_4R $Tyr^{2.64}$ and PP Tyr^{27} mutants (Table 5-1). This restraint was also used to guide docking the helix first due to its rigid conformation compared to the rest of the binding interface. Once the more rigid helix of PP was docked to the more rigid ends of the transmembrane helices of Y_4R, the much more dynamic ECL of Y_4R and PP C-terminal were folded simultaneously. Mutation data outlined in table 5-1 provided several additional interactions that were used to guide the folding of these regions. These interactions specifically include a predicted salt bridge between Y_4R $Asp^{6.59}$ and PP Arg^{35}, a predicted hydrogen bond between Y_4R $Asn^{7.32}$ and PP Arg^{33}, a predicted cation-π interaction between Y_4R $Phe^{7.35}$ and PP Arg^{33}, and an interaction between Y_4R $Phe^{7.35}$ and PP Tyr^{36}.

Experimental results were first represented as low resolution restraints to ensure residue proximity during the low resolution modeling phase. During the high resolution refinement, experimental results were represented as atom-level restraints in an attempt to capture the proposed interactions on an atomic level. The specific restraints imposed and their corresponding steps are described in Table 5-3. Final models fit well with the majority of the experimental results, accurately portraying residues found to affect activity as well as those residues that failed to show any effect on activity. Specifically, the predicted salt bridge between Asp$^{6.59}$ and Arg^{35} is well represented in eight of the nine models. All models show less than a 4.0 Å distance between both inter-residue oxygen-nitrogen pairs, providing possible salt bridge interactions or hydrogen bonding. Six of the nine models demonstrate a distance of less than 3.2 Å between the oxygen in Y$_4$R Asn$^{7.32}$ and amine group in PP Arg^{33}, providing for the possibility of a hydrogen bond between these residues. Y$_4$R Phe$^{7.35}$ and PP Arg^{33} point toward each other in all nine models, which is conducive to the proposed cation-π interaction. Additionally, Y$_4$R Phe$^{7.35}$ and PP Tyr^{36} were oriented toward each other in four models. Finally, Y$_4$R Asp$^{2.68}$ is within 8 Å and points toward the PP helix in five models, suggesting an interaction between the PP helix and Y$_4$R Asp$^{2.68}$. One of the nine models is shown in figure 5-2, ***A*** and ***B***, highlighting the binding site and residues important for PP-Y$_4$R binding.

Table 5-3 Experimental restraints used to guide docking of PP with Y$_4$R

Y$_4$R residue	**PP residue**	**Low resolution restraint**	**High resolution restraint**	**Proposed interaction**	**Steps imposed**
Tyr$^{2.64}$	Tyr^{27}	C-β atoms within 8 Å	None	Unknown	PP helix placement
Asp$^{6.59}$	Arg^{35}	C-β atoms within 8 Å	Asp$^{6.59}$ O-δ and Arg^{35} NH within 4 Å	Salt bridge	PP C-terminal folding (low resolution), Y$_4$R loop building (low resolution), final relaxation (high resolution)
Asn$^{7.32}$	Arg^{33}	C-β atoms within 8 Å	Asn$^{7.32}$ O-δ and Arg^{33} NH within 4 Å	Hydrogen bond	PP C-terminal folding (low resolution), Y$_4$R loop building (low resolution), final relaxation (high resolution)
Phe$^{7.35}$	Arg^{33}	C-β atoms within 8 Å	None	π-cation stacking	PP C-terminal folding (low resolution), Y$_4$R loop building (low resolution), final relaxation (high resolution)
Phe$^{7.35}$	Tyr^{36}	None	Phe$^{7.35}$ CZ and Tyr^{36} CZ within 4 Å	Unknown	Final relaxation (high resolution)

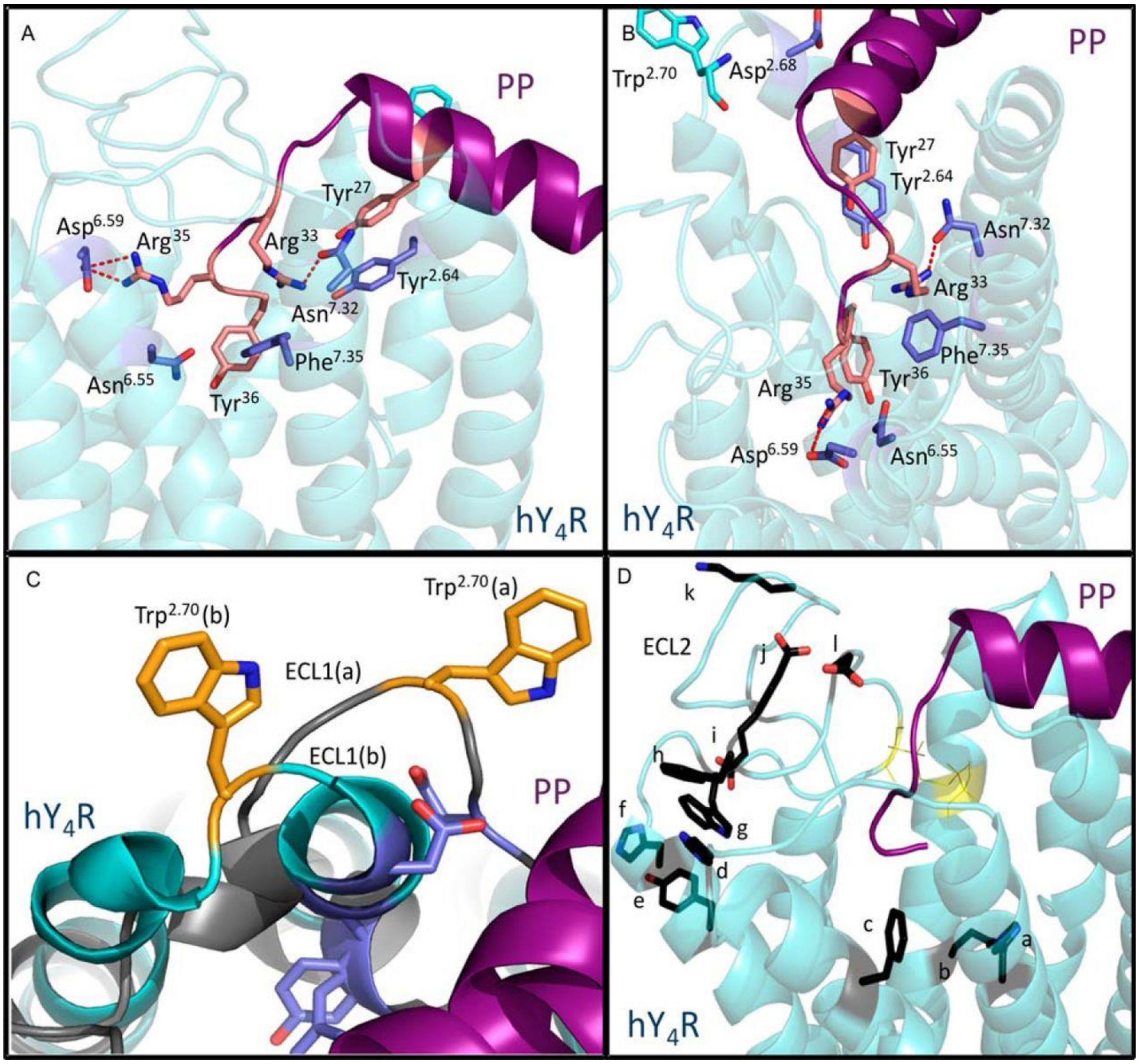


Figure 5.2 Characterization of the binding pocket of PP docked in the hY4R comparative model. A) side view of PP (purple) docked to Y4R (cyan). Residues found to be important in the activation of Y4R by PP are labeled. Predicted interactions are indicated by dotted red lines (salt bridge between Asp6.59 and Arg35 and hydrogen bond between Arg33 and Asn7.32). B) top-down view of the same docked model. C) two docked models show the variability in ECL1. The model shown in gray has a significantly longer ECL1 than that shown in cyan. Trp2.70, which was experimentally shown to be important in Y4R activation by PP, is shown to be in different proximity to PP depending on the size of ECL1. D) side view of the same docked model shown in A and B. Residues experimentally shown to be inactive in the binding of PP to Y4R are indicated in black. The disulfide bond in ECL2 is also shown in yellow. a = His7.39; b = Gln3.32; c = Phe6.54; d = His6.62; e = Tyr5.38; f = His5.34; g = Trp5.29; h = Phe4.80; i = Glu4.67; j = Glu4.79; k = Lys4.72; and l = Asp4.83. Source: [12]

The importance of Y_4R Trp$^{2.70}$ for PP binding is the only previously published experimental finding not well reflected in the models [12]. In all but one of the nine models, it is pointing away and/or not in close proximity to PP. The experimental results regarding this residue may reflect a second site on the receptor that leads to an indirect effect of Trp$^{2.70}$ on binding. Alternatively, model inaccuracy may cause this residue to be oriented improperly. This region of the receptor was highly dynamic in the final model ensemble, suggesting that the models failed to converge on a consistent conformation within this region. The length of TM2 varies across the final model ensemble resulting in ECL1 that varies dramatically from three residues in two models, 9-11 residues in five models, and up to 12-13 residues in two models. This lack of precision in the final model ensemble may result in a drop in accuracy at this region. This discrepancy in loop length is shown in figure 5.2***C***.

Models generally positioned residues that failed to affect activity in published mutational assays away from PP. Most of these residues are located in ECL2 which consistently lies at the edge of the receptor face away from PP. Specifically, Lys$^{4.72}$, Glu$^{4.79}$, Phe$^{4.80}$, Asp$^{4.83}$, His$^{5.34}$, and Phe$^{6.54}$ are not in contact with PP in any models. Gln$^{3.32}$, Glu$^{4.67}$, Trp$^{5.29}$, His$^{6.62}$, and His$^{7.39}$ are within 8 Å of a PP residue in only three of the nine models, and Tyr$^{5.38}$ is within 8 Å of a PP residue in only two of the nine models. ECL2 and the residues not involved in PP binding are shown in figure 5.2***D***.

The ensemble of nine models was analyzed for ligand-receptor interactions. These predictions can serve as hypotheses to direct future mutational assays. Residue pairs between PP and Y_4R with a distance of less than 8 Å were collected across all nine models. The total counts are shown in figure 5.3. This map can serve as a foundation from which to identify the residues that line the binding pocket. For example, five of nine models show that Y_4R Ser$^{5.28}$ and PP Thr^{32} are within 8 Å of each other, suggesting a possible interaction between these two residues.

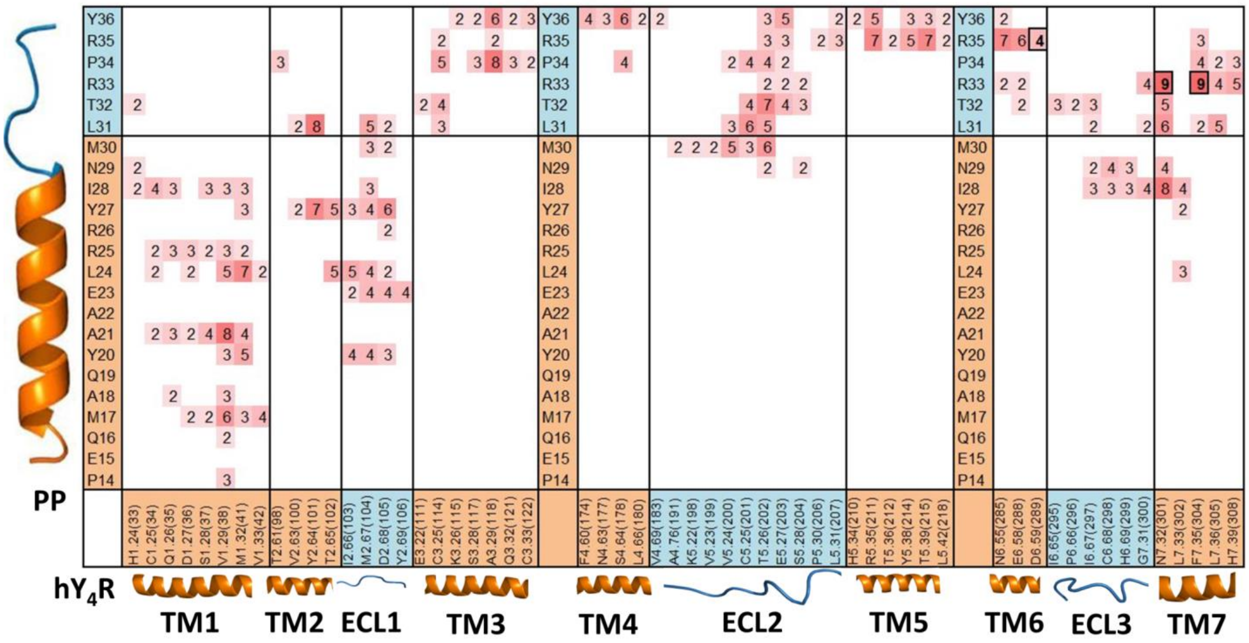


Figure 5.3 Y4R and PP residues within an 8 Å distance (based on C-β atoms) represent possible binding interactions. Neighboring residue pairs were collected across the nine final PP-Y4R docked models and presented as a heatmap indicating the most represented neighbors. Y4R residues are listed on the x axis with their secondary structure indicated (orange = TM and blue = ECL). PP residues are listed on the y axis with similar secondary structure indications. Numbers represent the number of models (out of nine) from which these residue pairs were within 8 Å. TM, transmembrane helix; ECL, extracellular loop. Source: [12]

5.4 Discussion

This chapter presents the computational modeling strategy used in a collaboration that combines *in vitro* experiments with the Rosetta molecular modeling suite to identify and characterize a binding pocket for the Y_4R system that is composed of several residues located on TM2, TM6, and TM7. This text serves to elaborate on the computational results contained in the publication by Pedragosa-Badia, *et al* [12]. The reader is referred to this publication for detailed *in vitro* experimental methods and results. As shown, these models capture the majority of experimental results regarding residues that contact PP as well as many residues that have no effect on activity when mutated. This approach shows the utility of combing comparative modeling with protein docking and *de novo* protein folding to model the interaction between a GPCR and a peptide ligand with regions of different flexibility.

5.5 Methods

Fourteen GPCR Templates were Considered for Y_4R Comparative Modeling

A comparative model of Y_4R was constructed using the protein structure prediction software package Rosetta, version 3.4 [3]. Because Comparative modeling GPCRs has been shown to be more successful when multiple GPCR templates are used instead of one [26], fourteen experimental GPCR structures from the Protein Data Bank (PDB) were considered as possible templates. These structures include the following: rhodopsin (PDB code 1U19) [44]; β_2-adrenergic receptor (PDB code 2RH1) [45]; β_1-adrenergic receptor (PDB code 2VT4) [46]; A_{2A}-adenosine receptor (PDB code 3EML) [47]; CXC chemokine receptor type 4 (PDB code 3ODU) [48]; D3 dopamine receptor (PDB code 3PBL) [49]; H1 histamine receptor (PDB code 3RZE) [50]; M2 muscarinic receptor (PDB code 3UON) [51]; sphingosine 1-phosphate receptor (PDB code 3V2W) [52]; M3 muscarinic receptor (PDB code 4DAJ) [53]; κ-opioid receptor (PDB code 4DJH) [54]; µ-opioid receptor (PDB code 4DKL) [55]; nociceptin/orphanin FQ opioid receptor (PDB code 4EA3) [56], and δ-opioid receptor (PDB code 4EJ4) [57]. Percent identity between Y_4R and each template can be found in table 5-3.

These structures were aligned with MUSTANG [58], and the resulting multiple sequence alignment was aligned with a multiple sequence alignment of hY_1R, hY_2R, Y_4R, and hY_5R using ClustalW [59]. Sequence alignments were adjusted to remove gaps within transmembrane α-helices and ensure that highly conserved residues remain aligned (supplementary figure S5-1). Y_4R residues were threaded onto the three-dimensional coordinates of aligned residues in each of the 14 GPCRs.

Missing Atom Coordinates Were Constructed Using Rosetta Loop Construction Protocols

Missing density and loop regions were reconstructed using Monte Carlo Metropolis fragment replacement and cyclic coordinate descent loop closure algorithms in Rosetta [33]. All models underwent repacking and gradient minimization with RosettaMembrane [42]. An additional constraint was included to account for the expected disulfide bond between Y_4R residues $Cys^{3.25}$ and $Cys^{5.25}$.

The final set of models was clustered based on RMSD using bcl::Cluster [60]. The top scoring models from the five largest clusters were used for docking studies.

Docking of Pancreatic Polypeptide (PP) into the Comparative Model of Y4R

A set of NMR structure conformations of bovine pancreatic polypeptide (PDB code1LJV) [61] was docked into the Y_4R comparative models using RosettaDock. The general design of RosettaDock follows the biophysical theory of encounter followed by transition to bound state[62]. In this algorithm, the helix of PP slides into contact with Y_4R and then Monte Carlo conformational search rotates and translates the helix to find a low energy pose. Bovine pancreatic polypeptide differs only on positions 6 and 23 with respect to PP and has similar affinity for the Y_4R as earlier reported [63, 64]. The use of ILJV provided a guide for the structural distinction between the peptide's helical region and dynamic tail region. The helical region (residues ^{14}PEQMAQYAAELRRYINML^{31}) was first docked into the Y_4R models. Four distinct helix conformations were docked into 37 Y_4R comparative models without ECLs, guided by a predicted interaction between Y_4R Tyr$^{2.64}$ and PP Tyr^{27}.

C-terminal Residues of PP Were Added Using de Novo Folding with Experimental Restraints

The five C-terminal residues of PP (TRPRY) were constructed using Rosetta's low resolution *de novo* folding algorithm where residues are represented as “centroids” [65]. Three experimentally derived restraints between Y_4R and PP residues were used to guide this step using an 8-Å distance cutoff between residues Asp$^{6.59}$ and Arg^{35}, Phe$^{7.35}$ and Arg^{33}, and Asn$^{7.32}$ and Arg^{33} [66, 67]. All restraints are detailed in Table 5-3.

The ECLs were rebuilt as described for the comparative modeling of Y_4R, with the addition of these experimental constraints. Additionally, these models were refined to atomic detail, replacing centroids with side chain rotamers based on a backbone-dependent rotamer library and energy minimization with RosettaMembrane [14, 68, 69].

Models Were Relaxed Using Atomic Resolution Experimental Restraints

Models were again clustered based on RMSD. Top scoring models from the largest clusters were visually inspected for binding poses that preserved the experimental restraints. Selected models underwent an additional relaxation step with constraints adjusted to reflect atomic level interactions between residues Asp$^{6.59}$ and Arg^{35} (3 Å distance between the two δ-oxygen atoms on Asp$^{6.59}$ and the side chain nitrogen atoms on Arg^{35}), and residues Asn$^{7.32}$ and Arg^{33} (4 Å distance between the δ-oxygen atom on Asn$^{7.32}$ and the two side chain nitrogen atoms on Arg^{33}). These constraint distances allow for possible hydrogen bonding and salt bridge interactions. An additional restraint between Y_4R Phe$^{7.35}$ and

PP Tyr^{36} was introduced. Final models were clustered and visually inspected, and nine representative models were selected. The overall workflow for receptor modeling and peptide docking is summarized in figure 5.4.

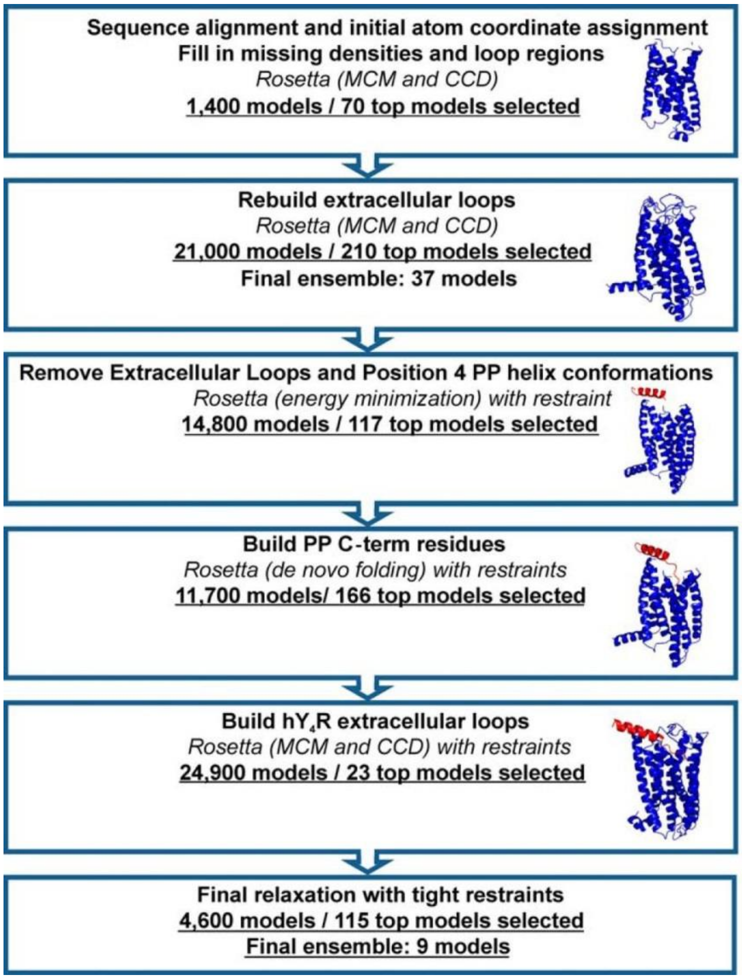

Sequence alignment and initial atom coordinate assignment
Fill in missing densities and loop regions
Rosetta (MCM and CCD)
1,400 models / 70 top models selected
Rebuild extracellular loops
Rosetta (MCM and CCD)
21,000 models / 210 top models selected
Final ensemble: 37 models
Remove Extracellular Loops and Position 4 PP helix conformations
Rosetta (energy minimization) with restraint
14,800 models / 117 top models selected
Build PP C-term residues
Rosetta (de novo folding) with restraints
11,700 models/ 166 top models selected
Build hY4R extracellular loops
Rosetta (MCM and CCD) with restraints
24,900 models / 23 top models selected
Final relaxation with tight restraints
4,600 models / 115 top models selected
Final ensemble: 9 models

Figure 5.4 Y_4R comparative model and PP docking work flow. An ensemble of Y4R comparative models was constructed through several rounds of loop building and energy minimization. Alongside the flowchart are representative models to illustrate the evolution of the comparative model. PP docking was guided by experimentally derived restraints. Source: [12].

5.6 References

1. Rohl CA, Strauss CE, Misura KM, & Baker D (2004) Protein structure prediction using Rosetta. *Methods in enzymology* 383:66-93.
2. Das R & Baker D (2008) Macromolecular modeling with rosetta. *Annual review of biochemistry* 77:363-382.

3. Leaver-Fay A*, et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* 487:545-574.
4. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, & Meiler J (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49(14):2987-2998.
5. Jiang L*, et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319(5868):1387-1391.
6. Siegel JB*, et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329(5989):309-313.
7. Greisen P, Jr. & Khare SD (2014) Computational redesign of metalloenzymes for catalyzing new reactions. *Methods in molecular biology* 1216:265-273.
8. Drew K*, et al.* (2013) Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design. *PloS one* 8(7):e67051.
9. Lange OF*, et al.* (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proceedings of the National Academy of Sciences of the United States of America* 109(27):10873-10878.
10. Mao B, Tejero R, Baker D, & Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *Journal of the American Chemical Society* 136(5):1893-1906.
11. Mittendorf KF, Kroncke BM, Meiler J, & Sanders CR (2014) The homology model of PMP22 suggests mutations resulting in peripheral neuropathy disrupt transmembrane helix packing. *Biochemistry* 53(39):6139-6141.
12. Pedragosa-Badia X*, et al.* (2014) Pancreatic polypeptide is recognized by two hydrophobic domains of the human Y4 receptor binding pocket. *The Journal of biological chemistry* 289(9):5846-5859.
13. Gregory KJ*, et al.* (2013) Probing the metabotropic glutamate receptor 5 (mGlu(5)) positive allosteric modulator (PAM) binding pocket: discovery of point mutations that engender a "molecular switch" in PAM pharmacology. *Molecular pharmacology* 83(5):991-1006.
14. Bradley P, Misura KM, & Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309(5742):1868-1871.
15. Meiler J & Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* 65(3):538-548.
16. Sripakdeevong P*, et al.* (2014) Structure determination of noncanonical RNA motifs guided by (1)H NMR chemical shifts. *Nature methods* 11(4):413-416.
17. Thyme S & Baker D (2014) Redesigning the specificity of protein-DNA interactions with Rosetta. *Methods in molecular biology* 1123:265-282.
18. Lindert S, Meiler J, & McCammon JA (2013) Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol to Improve Model Quality. *Journal of chemical theory and computation* 9(8):3843-3847.
19. Baker D (2014) Centenary Award and Sir Frederick Gowland Hopkins Memorial Lecture. Protein folding, structure prediction and design. *Biochemical Society transactions* 42(2):225-229.
20. Li Z & Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 84(19):6611-6615.
21. Meiler J & Baker D (2003) Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences of the United States of America* 100(21):12105-12110.
22. Combs SA*, et al.* (2013) Small-molecule ligand docking into comparative models with Rosetta. *Nature protocols* 8(7):1277-1298.

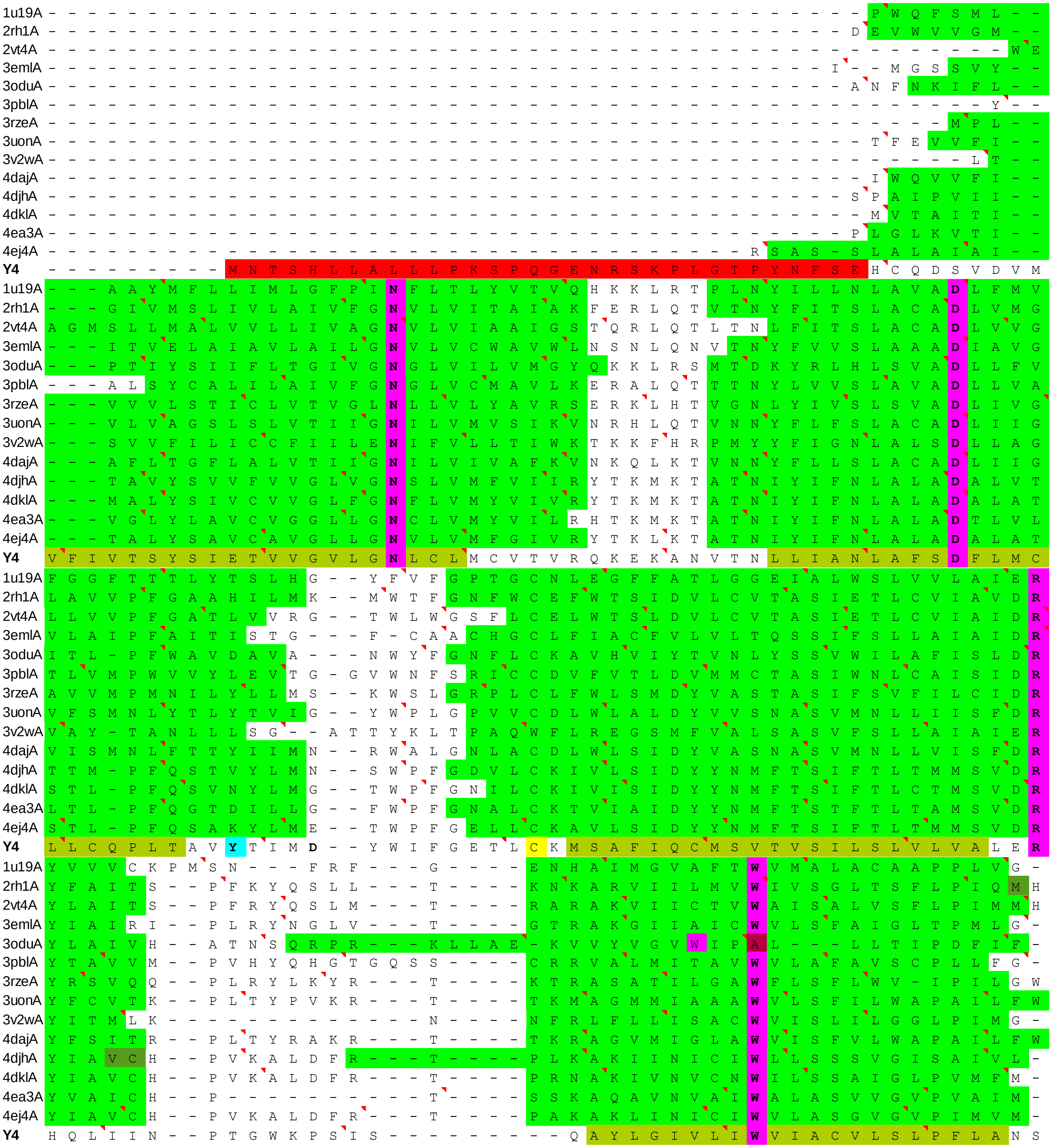
23. Kaufmann KW*, et al.* (2009) Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins* 74(3):630-642.
24. Lees-Miller JP*, et al.* (2009) Interactions of H562 in the S5 helix with T618 and S621 in the pore helix are important determinants of hERG1 potassium channel structure and function. *Biophysical journal* 96(9):3600-3610.
25. Keeble AH*, et al.* (2008) Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *Journal of molecular biology* 379(4):745-759.
26. Latek D, Pasznik P, Carlomagno T, & Filipek S (2013) Towards improved quality of GPCR models by usage of multiple templates and profile-profile comparison. *PloS one* 8(2):e56742.
27. Schertler GF, Villa C, & Henderson R (1993) Projection structure of rhodopsin. *Nature* 362(6422):770-772.
28. Venkatakrishnan AJ*, et al.* (2014) Structured and disordered facets of the GPCR fold. *Current opinion in structural biology* 27:129-137.
29. Taddese B, Simpson LM, Wall ID, Blaney FE, & Reynolds CA (2013) Modeling active GPCR conformations. *Methods in enzymology* 522:21-35.
30. Venkatakrishnan AJ*, et al.* (2013) Molecular signatures of G-protein-coupled receptors. *Nature* 494(7436):185-194.
31. Gront D, Kulp DW, Vernon RM, Strauss CE, & Baker D (2011) Generalized fragment picking in Rosetta: design, protocols and applications. *PloS one* 6(8):e23294.
32. Wang G & Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589-1591.
33. Canutescu AA & Dunbrack RL, Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein science : a publication of the Protein Society* 12(5):963-972.
34. Nguyen ED, Norn C, Frimurer TM, & Meiler J (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. *PloS one* 8(7):e67302.
35. Mandell DJ, Coutsias EA, & Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods* 6(8):551-552.
36. Honig B (1999) Protein folding: from the levinthal paradox to structure prediction. *Journal of molecular biology* 293(2):283-293.
37. Dunbrack Jr RL (2002) Rotamer Libraries in the 21st Century. *Current opinion in structural biology* 12(4):431-440.
38. Alexander NS*, et al.* (2013) RosettaEPR: rotamer library for spin label structure and dynamics. *PloS one* 8(9):e72851.
39. Raman S*, et al.* (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics* 77(S9):89-99.
40. Verma A & Wenzel W (2007) Protein structure prediction by all-atom free-energy refinement. *BMC structural biology* 7:12.
41. Landreh M & Robinson CV (2014) A new window into the molecular physiology of membrane proteins. *The Journal of physiology*.
42. Yarov-Yarovoy V, Schonbrun J, & Baker D (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins* 62(4):1010-1025.
43. Bystroff C & Krogh A (2008) Hidden Markov Models for prediction of protein features. *Methods in molecular biology* 413:173-198.
44. Okada T*, et al.* (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. *Journal of molecular biology* 342(2):571-583.

45. Cherezov V*, et al.* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* 318(5854):1258-1265.
46. Warne T*, et al.* (2008) Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* 454(7203):486-491.
47. Jaakola VP*, et al.* (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* 322(5905):1211-1217.
48. Wu B*, et al.* (2010) Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* 330(6007):1066-1071.
49. Chien EY*, et al.* (2010) Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* 330(6007):1091-1095.
50. Shimamura T*, et al.* (2011) Structure of the human histamine H1 receptor complex with doxepin. *Nature* 475(7354):65-70.
51. Haga K*, et al.* (2012) Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* 482(7386):547-551.
52. Hanson MA*, et al.* (2012) Crystal structure of a lipid G protein-coupled receptor. *Science* 335(6070):851-855.
53. Kruse AC*, et al.* (2012) Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* 482(7386):552-556.
54. Wu H*, et al.* (2012) Structure of the human kappa-opioid receptor in complex with JDTic. *Nature* 485(7398):327-332.
55. Manglik A*, et al.* (2012) Crystal structure of the micro-opioid receptor bound to a morphinan antagonist. *Nature* 485(7398):321-326.
56. Thompson AA*, et al.* (2012) Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* 485(7398):395-399.
57. Granier S*, et al.* (2012) Structure of the delta-opioid receptor bound to naltrindole. *Nature* 485(7398):400-404.
58. Konagurthu AS, Whisstock JC, Stuckey PJ, & Lesk AM (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins* 64(3):559-574.
59. Thompson JD, Gibson TJ, & Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 2:Unit 2 3.
60. Alexander N, Woetzel N, & Meiler J (2011) Bcl::Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pp 13-18.
61. Lerch M*, et al.* (2002) Bovine pancreatic polypeptide (bPP) undergoes significant changes in conformation and dynamics upon binding to DPC micelles. *Journal of molecular biology* 322(5):1117-1133.
62. Chaudhury S*, et al.* (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PloS one* 6(8):e22477.
63. Gehlert DR*, et al.* (1997) [125I]Leu31, Pro34-PYY is a high affinity radioligand for rat PP1/Y4 and Y1 receptors: evidence for heterogeneity in pancreatic polypeptide receptors. *Peptides* 18(3):397-401.
64. Walker MW*, et al.* (1997) A structure-activity analysis of the cloned rat and human Y4 receptors for pancreatic polypeptide. *Peptides* 18(4):609-612.
65. Simons KT, Kooperberg C, Huang E, & Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology* 268(1):209-225.

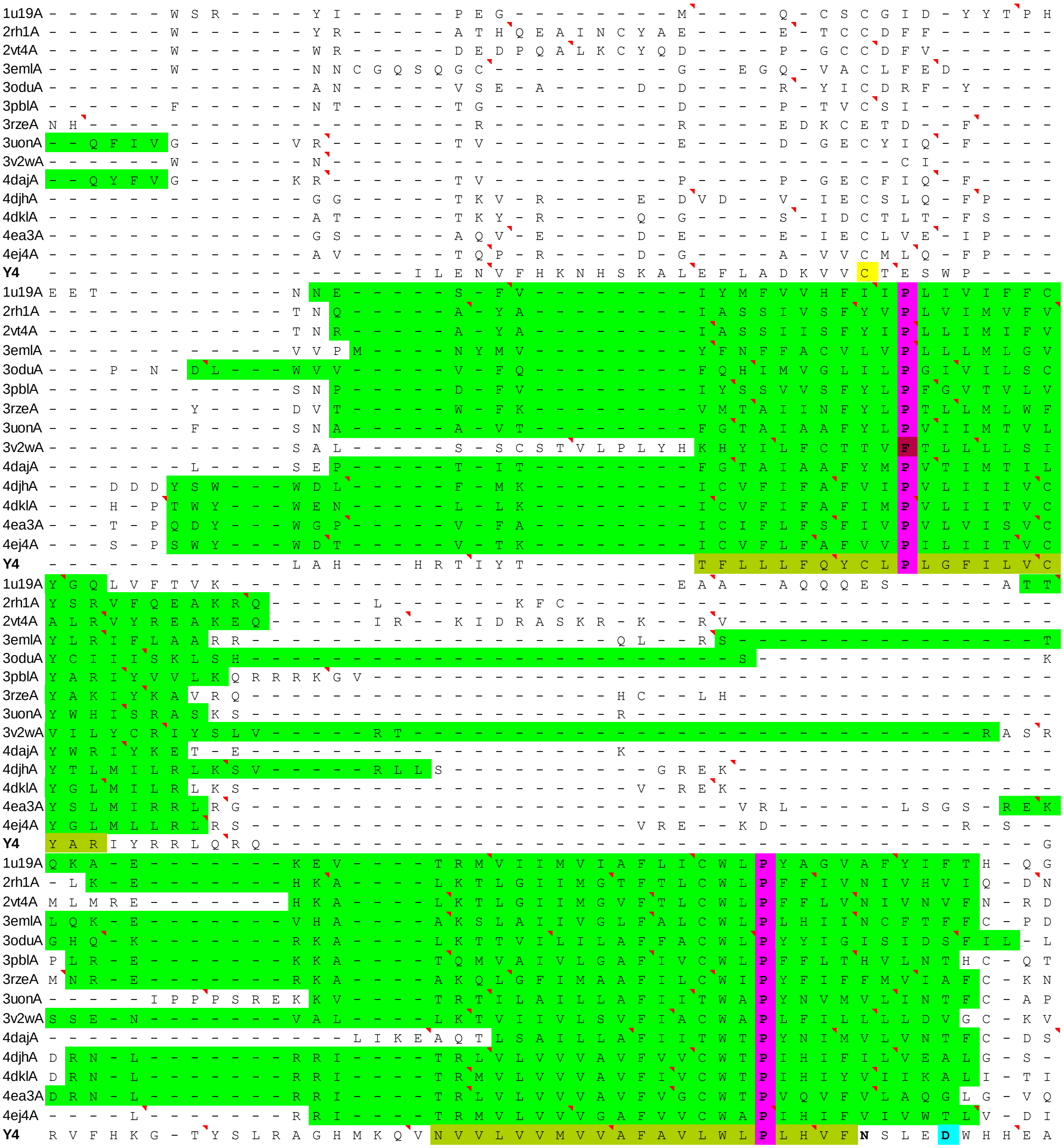
66. Merten N*, et al.* (2007) Receptor subtype-specific docking of Asp6.59 with C-terminal arginine residues in Y receptor ligands. *The Journal of biological chemistry* 282(10):7543-7551.
67. Lindner D, Stichel J, & Beck-Sickinger AG (2008) Molecular recognition of the NPY hormone family by their receptors. *Nutrition* 24(9):907-917.
68. Dunbrack RL & Karplus M (1993) Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction. *Journal of molecular biology* 230(2):543-574.
69. Misura KMS & Baker D (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins-Structure Function and Bioinformatics* 59(1):15-29.

5.7 Supplementary Information

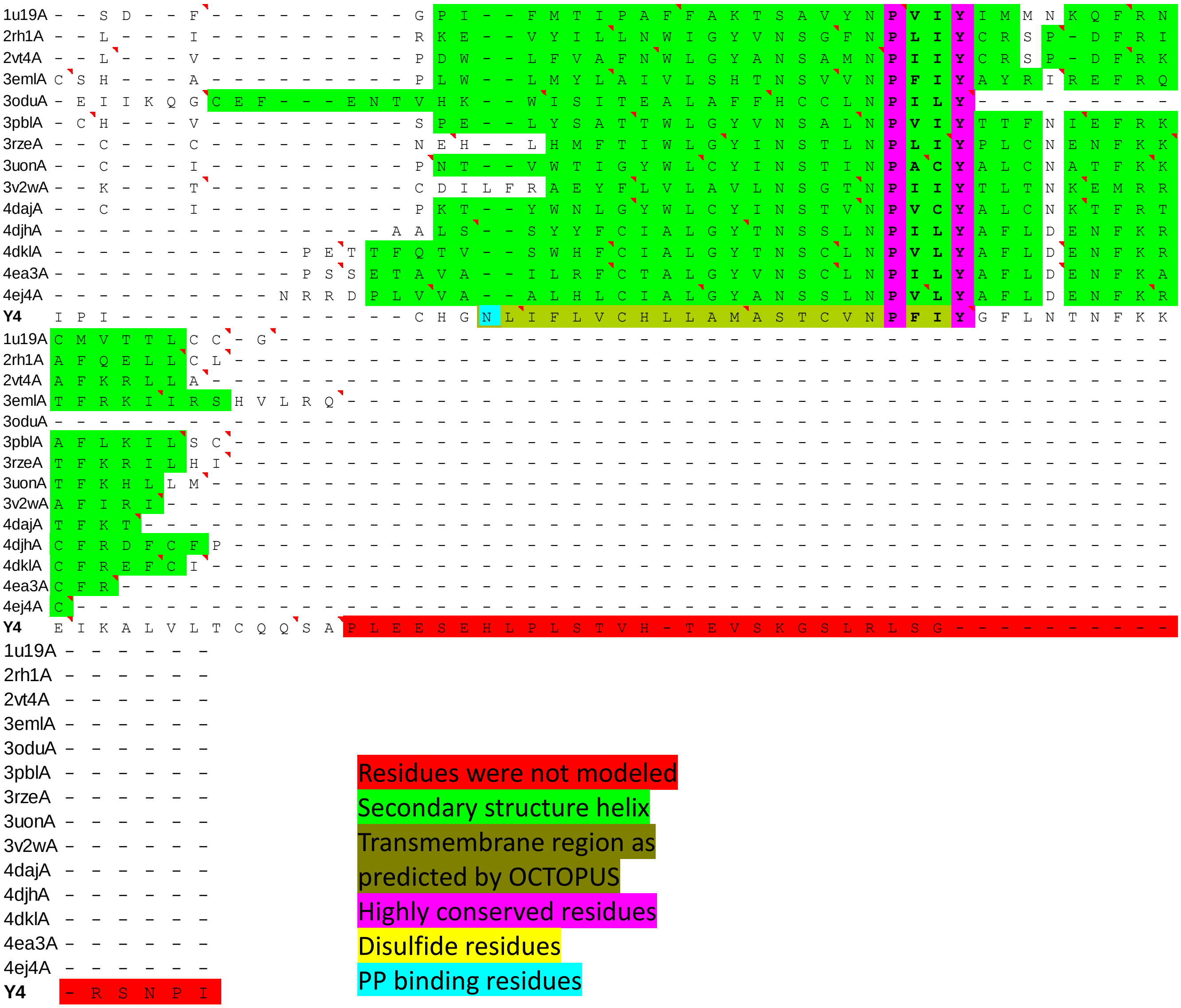
Figure S5-1 Adjusted alignment of Y4R to templates used for threading



```
1u19A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - P W Q F S M L - -
2rh1A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - D E V W V V G M - -
2vt4A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - W E
3emlA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - I - - M G S S V Y - -
3oduA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - A N F N K I F L - -
3pblA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - Y - -
3rzeA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - M P L - -
3uonA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T F E V V F I - -
3v2wA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - L T - -
4dajA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - I W Q V V F I - -
4djhA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - S P A I P V I I - -
4dklA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - M V T A I T I - -
4ea3A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - P L G L K V T I - -
4ej4A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - R S A S - S L A L A I A I - -
Y4    - - - - - - - - - M N T S H L L A L L L P K S P Q G E N R S K P L G T P Y N F S E H C Q D S V D V M
1u19A - - - A A Y M F L L I M L G F P I N F L T L Y V T V Q H K K L R T P L N Y I L L N L A V A D L F M V
2rh1A - - - G I V M S L I V L A I V F G N V L V I T A I A K F E R L Q T V T N Y F I T S L A C A D L V M G
2vt4A A G M S L L M A L V V L L I V A G N V L V I A A I G S T Q R L Q T L T N L F I T S L A C A D L V V G
3emlA - - - I T V E L A I A V L A I L G N V L V C W A V W L N S N L Q N V T N Y F V V S L A A A D I A V G
3oduA - - - P T I Y S I I F L T G I V G N G L V I L V M G Y Q K K L R S M T D K Y R L H L S V A D L L F V
3pblA - - - A L S Y C A L I L A I V F G N G L V C M A V L K E R A L Q T T T N Y L V V S L A V A D L L V A
3rzeA - - - V V V L S T I C L V T V G L N L L V L Y A V R S E R K L H T V G N L Y I V S L S V A D L I V G
3uonA - - - V L V A G S L S L V T I I G N I L V M V S I K V N R H L Q T V N N Y F L F S L A C A D L I I G
3v2wA - - - S V V F I L I C C F I I L E N I F V L L T I W K T K K F H R P M Y Y F I G N L A L S D L L A G
4dajA - - - A F L T G F L A L V T I I G N I L V I V A F K V N K Q L K T V N N Y F L L S L A C A D L I I G
4djhA - - - T A V Y S V V F V V G L V G N S L V M F V I I R Y T K M K T A T N I Y I F N L A L A D A L V T
4dklA - - - M A L Y S I V C V V G L F G N F L V M Y V I V R Y T K M K T A T N I Y I F N L A L A D A L A T
4ea3A - - - V G L Y L A V C V G G L L G N C L V M Y V I L R H T K M K T A T N I Y I F N L A L A D T L V L
4ej4A - - - T A L Y S A V C A V G L L G N V L V M F G I V R Y T K L K T A T N I Y I F N L A L A D A L A T
Y4    V F I V T S Y S I E T V V G V L G N L C L M C V T V R Q K E K A N V T N L L I A N L A F S D F L M C
1u19A F G G F T T T L Y T S L H G - - Y F V F G P T G C N L E G F F A T L G G E I A L W S L V V L A I E R
2rh1A L A V V P F G A A H I L M K - - M W T F G N F W C E F W T S I D V L C V T A S I E T L C V I A V D R
2vt4A L L V V P F G A T L V V R G - - T W L W G S F L C E L W T S L D V L C V T A S I E T L C V I A I D R
3emlA V L A I P F A I T I S T G - - - F - C A A C H G C L F I A C F V L V L T Q S S I F S L L A I A I D R
3oduA I T L - P F W A V D A V A - - - N W Y F G N F L C K A V H V I Y T V N L Y S S V W I L A F I S L D R
3pblA T L V M P W V V Y L E V T G - G V W N F S R I C C D V F V T L D V M M C T A S I W N L C A I S I D R
3rzeA A V V M P M N I L Y L L M S - - K W S L G R P L C L F W L S M D Y V A S T A S I F S V F I L C I D R
3uonA V F S M N L Y T L Y T V I G - - Y W P L G P V V C D L W L A L D Y V V S N A S V M N L L I I S F D R
3v2wA V A Y - T A N L L L S G - - A T T Y K L T P A Q W F L R E G S M F V A L S A S V F S L L A I A I E R
4dajA V I S M N L F T T Y I I M N - - R W A L G N L A C D L W L S I D Y V A S N A S V M N L L V I S F D R
4djhA T T M - P F Q S T V Y L M N - - S W P F G D V L C K I V L S I D Y Y N M F T S I F T L T M M S V D R
4dklA S T L - P F Q S V N Y L M G - - T W P F G N I L C K I V I S I D Y Y N M F T S I F T L C T M S V D R
4ea3A L T L - P F Q G T D I L L G - - F W P F G N A L C K T V I A I D Y Y N M F T S T F T L T A M S V D R
4ej4A S T L - P F Q S A K Y L M E - - T W P F G E L L C K A V L S I D Y Y N M F T S I F T L T M M S V D R
Y4    L L C Q P L T A V Y T I M D - - Y W I F G E T L C K M S A F I Q C M S V T V S I L S L V L V A L E R
1u19A Y V V V C K P M S N - - - F R F - - - G - - - - E N H A I M G V A F T W V M A L A C A A P P L V G -
2rh1A Y F A I T S - - P F K Y Q S L L - - - T - - - - K N K A R V I I L M V W I V S G L T S F L P I Q M H
2vt4A Y L A I T S - - P F R Y Q S L M - - - T - - - - R A R A K V I I C T V W A I S A L V S F L P I M M H
3emlA Y I A I R I - - P L R Y N G L V - - - T - - - - G T R A K G I I A I C W V L S F A I G L T P M L G -
3oduA Y L A I V H - - A T N S Q R P R - - - K L L A E - K V V Y V G V W I P A L - - - L L T I P D F I F -
3pblA Y T A V V M - - P V H Y Q H G T G Q S S - - - - C R R V A L M I T A V W V L A F A V S C P L L F G -
3rzeA Y R S V Q Q - - P L R Y L K Y R - - - T - - - - K T R A S A T I L G A W F L S F L W V - I P I L G W
3uonA Y F C V T K - - P L T Y P V K R - - - T - - - - T K M A G M M I A A A W V L S F I L W A P A I L F W
3v2wA Y I T M L K - - - - - - - - - - - - - N - - - - N F R L F L L I S A C W V I S L I L G G L P I M G -
4dajA Y F S I T R - - P L T Y R A K R - - - T - - - - T K R A G V M I G L A W V I S F V L W A P A I L F W
4djhA Y I A V C H - - P V K A L D F R - - - T - - - - P L K A K I I N I C I W L L S S S V G I S A I V L -
4dklA Y I A V C H - - P V K A L D F R - - - T - - - - P R N A K I V N V C N W I L S S A I G L P V M F M -
4ea3A Y V A I C H - - P - - - - - - - - - - T - - - - S S K A Q A V N V A I W A L A S V V G V P V A I M -
4ej4A Y I A V C H - - P V K A L D F R - - - T - - - - P A K A K L I N I C I W V L A S G V G V P I M V M -
Y4    H Q L I I N - - P T G W K P S I S - - - - - - - - - Q A Y L G I V L I W V I A C V L S L P F L A N S
```



```
1u19A - - - - - - W S R - - - - Y I - - - - - P E G - - - - - - - - M - - - - Q - C S C G I D - Y Y T P H
2rh1A - - - - - - W - - - - - - Y R - - - - - A T H Q E A I N C Y A E - - - - E - T C C D F F - - - - - -
2vt4A - - - - - - W - - - - - - W R - - - - - D E D P Q A L K C Y Q D - - - - P - G C C D F V - - - - - -
3emlA - - - - - - W - - - - - - N N C G Q S Q G C - - - - - - - - - G - - E G Q - V A C L F E D - - - - -
3oduA - - - - - - - - - - - - - A N - - - - - V S E - A - - - - D - D - - - - R - Y I C D R F - Y - - - -
3pblA - - - - - - F - - - - - - N T - - - - - T G - - - - - - - - - D - - - - P - T V C S I - - - - - - -
3rzeA N H - - - - - - - - - - - - - - - - - - - R - - - - - - - - - R - - - - E D K C E T D - - F - - - -
3uonA - - Q F I V G - - - - - V R - - - - - - T V - - - - - - - - - E - - - - D - G E C Y I Q - F - - - -
3v2wA - - - - - - W - - - - - - N - - - - - - - - - - - - - - - - - - - - - - - - - - - - C I - - - - - -
4dajA - - Q Y F V G - - - - - K R - - - - - - T V - - - - - - - - - P - - - - P - G E C F I Q - F - - - -
4djhA - - - - - - - - - - - - - G G - - - - - T K V - R - - - - E - D V D - - V - I E C S L Q - F P - - -
4dklA - - - - - - - - - - - - - A T - - - - - T K Y - R - - - - Q - G - - - - S - I D C T L T - F S - - -
4ea3A - - - - - - - - - - - - - G S - - - - - A Q V - E - - - - D - E - - - - E - I E C L V E - I P - - -
4ej4A - - - - - - - - - - - - - A V - - - - - T Q P - R - - - - D - G - - - - A - V V C M L Q - F P - - -
Y4    - - - - - - - - - - - - - - - - - - I L E N V F H K N H S K A L E F L A D K V V C T E S W P - - - -
1u19A E E T - - - - - - - - - N N E - - - - - S - F V - - - - - - - - I Y M F V V H F I I P L I V I F F C
2rh1A - - - - - - - - - - - - T N Q - - - - - A - Y A - - - - - - - - I A S S I V S F Y V P L V I M V F V
2vt4A - - - - - - - - - - - - T N R - - - - - A - Y A - - - - - - - - I A S S I I S F Y I P L L I M I F V
3emlA - - - - - - - - - - - - V V P M - - - - N Y M V - - - - - - - - Y F N F F A C V L V P L L L M L G V
3oduA - - - P - N - D L - - - W V V - - - - - V - F Q - - - - - - - - F Q H I M V G L I L P G I V I L S C
3pblA - - - - - - - - - - - - S N P - - - - - D - F V - - - - - - - - I Y S S V V S F Y L P F G V T V L V
3rzeA - - - - - - - Y - - - - D V T - - - - - W - F K - - - - - - - - V M T A I I N F Y L P T L L M L W F
3uonA - - - - - - - F - - - - S N A - - - - - A - V T - - - - - - - - F G T A I A A F Y L P V I I M T V L
3v2wA - - - - - - - - - - - - S A L - - - - - S - S C S T V L P L Y H K H Y I L F C T T V F T L L L L S I
4dajA - - - - - - - L - - - - S E P - - - - - T - I T - - - - - - - - F G T A I A A F Y M P V T I M T I L
4djhA - - - D D D Y S W - - - W D L - - - - - F - M K - - - - - - - - I C V F I F A F V I P V L I I I V C
4dklA - - - H - P T W Y - - - W E N - - - - - L - L K - - - - - - - - I C V F I F A F I M P V L I I T V C
4ea3A - - - T - P Q D Y - - - W G P - - - - - V - F A - - - - - - - - I C I F L F S F I V P V L V I S V C
4ej4A - - - S - P S W Y - - - W D T - - - - - V - T K - - - - - - - - I C V F L F A F V V P I L I I T V C
Y4    - - - - - - - - - - - - L A H - - - H R T I Y T - - - - - - - - T F L L L F Q Y C L P L G F I L V C
1u19A Y G Q L V F T V K - - - - - - - - - - - - - - - - - - - - - - E A A - - A Q Q Q E S - - - - - A T T
2rh1A Y S R V F Q E A K R Q - - - - - L - - - - - - K F C - - - - - - - - - - - - - - - - - - - - - - - -
2vt4A A L R V Y R E A K E Q - - - - - I R - - K I D R A S K R - K - - R V - - - - - - - - - - - - - - - -
3emlA Y L R I F L A A R R - - - - - - - - - - - - - - - - - - Q L - - R S - - - - - - - - - - - - - - - T
3oduA Y C I I I S K L S H - - - - - - - - - - - - - - - - - - - - - - - - S - - - - - - - - - - - - - - K
3pblA Y A R I Y V V L K Q R R R K G V - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3rzeA Y A K I Y K A V R Q - - - - - - - - - - - - - - - - - - H C - - L H - - - - - - - - - - - - - - - -
3uonA Y W H I S R A S K S - - - - - - - - - - - - - - - - - - R - - - - - - - - - - - - - - - - - - - - -
3v2wA V I L Y C R I Y S L V - - - - - R T - - - - - - - - - - - - - - - - - - - - - - - - - - - - R A S R
4dajA Y W R I Y K E T - E - - - - - - - - - - - - - - - - - - K - - - - - - - - - - - - - - - - - - - - -
4djhA Y T L M I L R L K S V - - - - - R L L S - - - - - - - - - - G R E K - - - - - - - - - - - - - - - -
4dklA Y G L M I L R L K S - - - - - - - - - - - - - - - - - - - V - R E K - - - - - - - - - - - - - - - -
4ea3A Y S L M I R R L R G - - - - - - - - - - - - - - - - - - - - - - - - V R L - - - - - L S G S - R E K
4ej4A Y G L M L L R L R S - - - - - - - - - - - - - - - - - - - V R E - - K D - - - - - - - - - R - S - -
Y4    Y A R I Y R R L Q R Q - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - G
1u19A Q K A - E - - - - - - - K E V - - - - T R M V I I M V I A F L I C W L P Y A G V A F Y I F T H - Q G
2rh1A - L K - E - - - - - - - H K A - - - - L K T L G I I M G T F T L C W L P F F I V N I V H V I Q - D N
2vt4A M L M R E - - - - - - - H K A - - - - L K T L G I I M G V F T L C W L P F F L V N I V N V F N - R D
3emlA L Q K - E - - - - - - - V H A - - - - A K S L A I I V G L F A L C W L P L H I I N C F T F F C - P D
3oduA G H Q - K - - - - - - - R K A - - - - L K T T V I L I L A F F A C W L P Y Y I G I S I D S F I L - L
3pblA P L R - E - - - - - - - K K A - - - - T Q M V A I V L G A F I V C W L P F F L T H V L N T H C - Q T
3rzeA M N R - E - - - - - - - R K A - - - - A K Q L G F I M A A F I L C W I P Y F I F F M V I A F C - K N
3uonA - - - - - I P P P S R E K K V - - - - T R T I L A I L L A F I I T W A P Y N V M V L I N T F C - A P
3v2wA S S E - N - - - - - - - V A L - - - - L K T V I I V L S V F I A C W A P L F I L L L L D V G C - K V
4dajA - - - - - - - - - - - - - - - L I K E A Q T L S A I L L A F I I T W T P Y N I M V L V N T F C - D S
4djhA D R N - L - - - - - - - R R I - - - - T R L V L V V V A V F V V C W T P I H I F I L V E A L G - S -
4dklA D R N - L - - - - - - - R R I - - - - T R M V L V V V A V F I V C W T P I H I Y V I I K A L I - T I
4ea3A D R N - L - - - - - - - R R I - - - - T R L V L V V V A V F V G C W T P V Q V F V L A Q G L G - V Q
4ej4A - - - - L - - - - - - - R R I - - - - T R M V L V V V G A F V V C W A P I H I F V I V W T L V - D I
Y4    R V F H K G - T Y S L R A G H M K Q V N V V L V V M V V A F A V L W L P L H V F N S L E D W H H E A
```



```
1u19A - - S D - - F - - - - - - - - - G P I - - F M T I P A F F A K T S A V Y N P V I Y I M M N K Q F R N
2rh1A - - L - - - I - - - - - - - - - R K E - - V Y I L L N W I G Y V N S G F N P L I Y C R S P - D F R I
2vt4A - - L - - - V - - - - - - - - - P D W - - L F V A F N W L G Y A N S A M N P I I Y C R S P - D F R K
3emlA C S H - - - A - - - - - - - - - P L W - - L M Y L A I V L S H T N S V V N P F I Y A Y R I R E F R Q
3oduA - E I I K Q G C E F - - - E N T V H K - - W I S I T E A L A F F H C C L N P I L Y - - - - - - - - -
3pblA - C H - - - V - - - - - - - - - S P E - - L Y S A T T W L G Y V N S A L N P V I Y T T F N I E F R K
3rzeA - - C - - - C - - - - - - - - - N E H - - L H M F T I W L G Y I N S T L N P L I Y P L C N E N F K K
3uonA - - C - - - I - - - - - - - - - P N T - - V W T I G Y W L C Y I N S T I N P A C Y A L C N A T F K K
3v2wA - - K - - - T - - - - - - - - - C D I L F R A E Y F L V L A V L N S G T N P I I Y T L T N K E M R R
4dajA - - C - - - I - - - - - - - - - P K T - - Y W N L G Y W L C Y I N S T V N P V C Y A L C N K T F R T
4djhA - - - - - - - - - - - - - - - A A L S - - S Y Y F C I A L G Y T N S S L N P I L Y A F L D E N F K R
4dklA - - - - - - - - - - - P E T T F Q T V - - S W H F C I A L G Y T N S C L N P V L Y A F L D E N F K R
4ea3A - - - - - - - - - - - P S S E T A V A - - I L R F C T A L G Y V N S C L N P I L Y A F L D E N F K A
4ej4A - - - - - - - - - - N R R D P L V V A - - A L H L C I A L G Y A N S S L N P V L Y A F L D E N F K R
Y4    I P I - - - - - - - - - - - - - C H G N L I F L V C H L L A M A S T C V N P F I Y G F L N T N F K K
1u19A C M V T T L C C - G - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
2rh1A A F Q E L L C L - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
2vt4A A F K R L L A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3emlA T F R K I I R S H V L R Q - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3oduA - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3pblA A F L K I L S C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3rzeA T F K R I L H I - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3uonA T F K H L L M - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3v2wA A F I R I - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4dajA T F K T - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4djhA C F R D F C F P - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4dklA C F R E F C I - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4ea3A C F R - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4ej4A C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Y4    E I K A L V L T C Q Q S A P L E E S E H L P L S T V H - T E V S K G S L R L S G - - - - - - - - - -
1u19A - - - - - -
2rh1A - - - - - -
2vt4A - - - - - -
3emlA - - - - - -
3oduA - - - - - -
3pblA - - - - - -
3rzeA - - - - - -
3uonA - - - - - -
3v2wA - - - - - -
4dajA - - - - - -
4djhA - - - - - -
4dklA - - - - - -
4ea3A - - - - - -
4ej4A - - - - - -
Y4    - R S N P I
```

🟥 Residues were not modeled

🟩 Secondary structure helix

Transmembrane region as predicted by OCTOPUS

Highly conserved residues

🟨 Disulfide residues

PP binding residues

Protocol Capture

The following information includes all settings and command lines used for comparative modeling the Y4 receptor and docking PP. The Rosetta software suite is publically available and the license is free for non-commercial users at http://www.rosettacommons.org/

Part 1: Multi-template Y4R receptor comparative modeling

a) Manually generated files (this information may be copied directly into new text files)

y4_truncated.fasta

```
>Y4
HCQDSVDVMVFIVTSYSIETVVGVLGNLCLMCVTVRQKEKANVTNLLIANLAFSDFLMCL
LCQPLTAVYTIMDYWIFGETLCKMSAFIQCMSVTVSILSLVLVALERHQLIINPTGWKPS
ISQAYLGIVLIWVIACVLSLPFLANSILENVFHKNHSKALEFLADKVVCTESWPLAHHRT
IYTTFLLLFQYCLPLGFILVCYARIYRRLQRQGRVFHKGTYSLRAGHMKQVNVVLVVMVV
AFAVLWLPLHVFNSLEDWHHEAIPICHGNLIFLVCHLLAMASTCVNPFIYGFLNTNFKKE
IKALVLTCQQSA
```

y4_truncated.span

```
TM region prediction for y4_truncated.octopus predicted using OCTOPUS
7 312
antiparallel
n2c
  10    35    10    35
  46    68    46    68
  84   110    84   110
 124   144   124   144
 184   209   184   209
 232   255   232   255
 269   290   269   290
```

y4.disulfide

```
82 169
```

loop_build1.options

```
-database /main/database/

-loops:timer #output time spent in seconds for each loop modeling job

-loops:fa_input #input structures are in full atom format
-in:fix_disulf y4.disulfide #read disulfide connectivity information
-in:file:spanfile y4_truncated.span

-loops:relax fastrelax
-loops:extended true #force phi-psi angles to be set to 180 degrees
-loops:frag_sizes 9 3 1
-loops:frag_files aaY4Cut09_05.200_v1_3 aaY4Cut03_05.200_v1_3 none
-loops:remodel quick_ccd
-loops:refine refine_kic
-out:file:silent_struct_type binary #output file type
-out:file:fullatom #output file will be fullatom

-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts
```

loopbuild_final.options

```
-database /main/database

-loops:timer #output time spent in seconds for each loop modeling job
-loops:fa_input #input structures are in full atom format
-in:fix_disulf y4.disulfide #read disulfide connectivity information
-in:file:spanfile y4_truncated.span

-loops:relax fastrelax
-loops:extended true #force phi-psi angles to be set to 180 degrees
-loops:frag_sizes 9 3 1
-loops:frag_files aaY4Cut09_05.200_v1_3 aaY4Cut03_05.200_v1_3 none
-loops:remodel quick_ccd
-loops:refine refine_kic

-out:file:silent_struct_type binary #output file type
-out:file:fullatom #output file will be fullatom

-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts
```

b) Steps and commands: The following steps describe specific command lines and resulting files

	Step	Text	Command	Comment
1	Create GPCR alignments	Create alignment profile for 14 GPCR's from the PDB	mustang -i 1u19A_clean.pdb 2vt4A_clean.pdb 3oduA_clean.pdb 3rzeA_clean.pdb 3v2wA_clean.pdb 4djhA_clean.pdb 4ea3A_clean.pdb 2rh1A_clean.pdb 3emlA_clean.pdb 3pblA_clean.pdb 3uonA_clean.pdb 4dajA_clean.pdb 4dklA_clean.pdb 4ej4A_clean.pdb -F fasta	Gives a pdb and afasta file
2	Create NPY alignments	Create sequence alignment of Y1,Y2, and Y4 (Y5 is excluded due to a very long loop after helix 5)	clustalw -> Sequence input from disc -> npy_1_2_4.fasta -> multiple Alignments	Gives .aln and .dnd files
3	Align NPY to GPCR's	Align the GPCR profile with the NPY sequence alignment	clustalw -> Profile/Structural Alignments -> 1st profile = all_gpcr_profile.afasta; 2nd profile/sequences = npy_1_2_4.aln -> Align sequences to 1st profile (slow/accurate)	Gives .aln and .dnd files
4	Adjust aligments	Manually adjust the alignments to remove gaps that may exist within transmembrane regions so that they exist in loop regions.	N/A	y4_gpcr_profile_adjusted.aln
5	Y4 preparation	Generate secondary structure prediction using the truncated Y4 sequence	http://octopus.cbr.su.se/ and convert output to span file with command perl /home/dongen/scripts/octopus2span.pl y4_truncated.octopus > y4_truncated.span	y4_truncated.octopus and y4_truncated.span
6	Y4 preparation	Generate Y4 fragment files	make_fragments.pl -id Y4 -nohoms -nosam y4_truncated.fasta	jufo, psipred, ss2, v1_3 files output
7	Thread Y4	Thread Y4 over the 14 aligned GPCR structures	/sb/meiler/scripts/sequence_util/thread_pdb_from_alignment.py --template=1u19A -target=Y4 --chain=A --align_format=clustal y4_gpcr_profile_adjusted.aln 1u19A_clean.pdb y4_on_1u19A.pdb	All commands are in thread_all.sh and gives threaded pdbs
8	Generate full Y4 models	Fill in missing densities and relax	/sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/loopmodel.default.linuxgccrelease@loop_build1.options -loops:input_pdb y4_on_1u19A.pdb -loops:loop_file 1u19A.loops -out:file:silent y4_on_1u19A_loop1.out -nstruct25 > y4_on_1u19A_loop1.log	creates 100 structures for each template (out files)
9	Get best models	Extract top 5 models for each template	Gather all scores using grep "SCORE" *.out and select 5 models with lowest energy scores within each template. Extract these with: /sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/score_jd2.default.linuxgccrelease -database /sb/meiler/rosetta/rosetta-3.4/rosetta_database/ -in:file:silent y4_on_1u19A_loop1_0.out -out:pdb -in:file:tags 0001 0002 0003 0004 0005	gathers the requested pdbs from the out files.
10	Final model generation	Build loops and relax models	/sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/loopmodel.default.linuxgccrelease@loopbuild_final.options -loops:input_pdb 1u19A_loop1_0.pdb -loops:loop_file 1u19A.ecloops -out:file:silent 1u19A_final_0.out -nstruct15 > y4_on_1u19A_final.log	Generated 21,000 models
11	Cluster final models	Extract all the pdbs	foreach i (`ls *.out`)/sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/score_jd2.default.linuxgccrelease -database /sb/meiler/rosetta/rosetta-3.4/rosetta_database/ -in:file:silent $i -out:prefix $i -out:pdb -in:file:tagsend	
12	Cluster final models	Cluster pdbs (Do for each template)	bcl.exe PDBCompare -quality RMSD -atom_list CA -pdb_list pdb_list.ls -prefix 1u19A_rmsd -norm100 -aaclass AACaCb -convert_to_natural_aa_type -scheduler PThread12 bcl.exe Cluster -distance_input_file 1u19A__*rmsdRMSD.txt -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster5_1u19A -linkage Average -output_pymol1000 5 100 10000 10 dendogram5* $TEMPLATE.py -remove_internally_similar_nodes5 -pymol_label_output_string -scheduler PThread12	

Part 2: Dock PP helix

a) Manually generated files

dock_helix_y4_1.xml

```
<dock_design>
        <SCOREFXNS> #defines non-standard score functions
                <mem_cen_cst weights=score_membrane >
                        <Reweight scoretype=atom_pair_constraint weight=10/>
                </mem_cen_cst>
                <mem_fa_cst weights=membrane_highres_Menv_smooth >
                        <Reweight scoretype=atom_pair_constraint weight=10/>
                </mem_fa_cst>
        </SCOREFXNS>
        <FILTERS>
        </FILTERS>
        <TASKOPERATIONS>
                <InitializeFromCommandline name=ifcl/>
                <RestrictToRepacking name=rtrp/>
        </TASKOPERATIONS>
        <MOVERS>
                <Docking name=dock score_low=mem_cen_cst
        score_high=mem_fa_cst fullatom=1 local_refine=1
        optimize_fold_tree=1 conserve_foldtree=0 design=0
        task_operations=ifcl/>
                <FastRelax name=fastrlx_all repeats=4 scorefxn=mem_fa_cst />
                <FastRelax name=fastrlx_r1 repeats=1 scorefxn=mem_fa_cst />
                <PackRotamersMover name=repack scorefxn=mem_fa_cst
        task_operations=rtrp/>
                <ConstraintSetMover name=fa_cst
                cst_file=npy4_ensemble_1_noloops.cst />
                <ConstraintSetMover name=lowres_cst
        cst_file=npy4_ensemble_1_noloops.cst />
        </MOVERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover_name=fa_cst/>
                <Add mover_name=lowres_cst/>
                <Add mover_name=dock/>
                <Add mover_name=fastrlx_r1/>
        </PROTOCOLS>
</dock_design>
```

quota.options

```
-in::file::vall vall.apr24.2008.extended.gz
-database /main/database

-frags::scoring::config quota-protocol.wghts
-frags::frag_sizes 9 3
-frags::n_candidates 1000
-frags::n_frags 200

-frags::picking::quota_config_file quota.def
```

dock_helix_y4_1.options

```
-database /main/database
-out
	-output
	-pdb
	-file
		-fullatom
		-silent_struct_type binary
-jd2
	-ntrials 5
-docking
	-dock_pert 4 10
-max_inner_cycles 30
-outer_cycles 1
-membrane
	-normal_cycles 100
	-normal_mag 15
	-center_mag 2
-in
	-file
		-fullatom
-constraints
	-cst_weight 10
	-cst_fa_weight 10
	-viol
	-viol_level 101
-packing
	-ex1
	-ex2
	-repack_only
	-linmem_ig 10
-overwrite
```

quota.def

```
#pool_id    pool_name         fraction
1           psipred           0.6
2           jufo              0.2
3           sam               0.2
```

quota-protocol.flags

```
# Input databases
-in::file::vall vall.apr24.2008.extended.gz
-database /main/database/

# Weights file
-frags::scoring::config quota-protocol.wghts

# we need nine-mers and three-mers
-frags::frag_sizes 9 3

# Select 200 fragments from 1000 candidates.
-frags::n_candidates 1000
-frags::n_frags 200

# Quota.def file defines the shares between difefrent quota pools. The total should be
1.0
-frags::picking::quota_config_file quota.def
```

quota-protocol.wghts

```
# score name         priority  wght   min_allowed  extras
SecondarySimilarity         350          0.5     -            psipred
SecondarySimilarity         300          0.5     -            sam
SecondarySimilarity         250          0.5     -            jufo
RamaScore                   150          1.0      -           psipred
RamaScore                   150         1.0      -            sam
RamaScore                   150          1.0     -            jufo
ProfileScoreL1              200          1.0     -
PhiPsiSquareWell            100          0.0      -
FragmentCrmsd        30           0.0      -
```

npy4_hpp.cst

```
AtomPair CB 69  CB 326 BOUNDED 0.00 8.0 1.0 NOE loose
AtomPair CB 257 CB 334 BOUNDED 0.00 8.0 1.0 NOE loose
AtomPair CB 272 CB 332 BOUNDED 0.00 8.0 1.0 NOE loose
AtomPair CB 269 CB 332 BOUNDED 0.00 8.0 1.0 NOE loose
```

npy4_pp_highrescst.cst

```
AtomPair OD1 257 NH1 334 BOUNDED 0.00 3.0 1.0 NOE loose
AtomPair OD2 257 NH2 334 BOUNDED 0.00 3.0 1.0 NOE loose
AtomPair OD1 269 NH1 332 BOUNDED 0.00 4.0 1.0 NOE loose
AtomPair OD1 269 NH2 332 BOUNDED 0.00 4.0 1.0 NOE loose
AtomPair CB  272 CB  332 BOUNDED 0.00 7.0 1.0 NOE loose
```

build_cterm.options

```
-run
       -reinitialize_mover_for_each_job
-score
       -find_neighbors_3dgrid
-membrane
       -fixed_membrane
       -no_interpolate_Mpair
       -Menv_penalties
       -Membed_init
-abinitio
       -membrane
       -rg_reweight 0.01
       -stage2_patch score_membrane_s2.wts_patch
       -stage3a_patch score_membrane_s3a.wts_patch
       -stage3b_patch score_membrane_s3b.wts_patch
       -stage4_patch score_membrane_s4.wts_patch
-fold_cst
       -force_minimize
-constraints
       -cst_weight 10.0
       -viol
       -viol_level 101
-out
       -output
       -file
              -silent_struct_type binary
-overwrite
```

relax_flags.txt

```
-database /main/database

-relax:fast
-relax:membrane
-constrain_relax_to_start_coords
-constraints
	-cst_fa_file npy4_pp_highrescst.cst
	-cst_fa_weight 10.0
	-viol
	-viol_level 101

-in:file:fullatom

-out:file:silent_struct_type binary #output file type
-out:file:fullatom #output file will be fullatom

-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts
```

npy4_ensemble_1.flags

```
-loops
	-loop_file npy4_ensemble_10.loops
	-extended true
	-relax fastrelax
	-frag_sizes 9 3 1
	-frag_files y4pep_frags_061912.200.9mers
	y4pep_frags_061912.200.3mers none
	-remodel quick_ccd
	-refine refine_kic
-relax
	-membrane
-score
	-weights membrane_highres_Menv_smooth.wts
-in
	-fix_disulf y4pep.disulfide
	-file
		-spanfile y4pep.span
-residues
	-patch_selectors CTERM_AMIDATION
-membrane
	-fixed_membrane
	-no_interpolate_Mpair
	-Menv_penalties
	-Membed_init
-fold_cst
	-force_minimize
-constraints
	-cst_file npy4_hpp.cst
	-cst_weight 10.0
	-cst_fa_file npy4_hpp.cst
	-cst_fa_weight 10.0
	-viol
	-viol_level 101
-out
	-output
	-file
		-fullatom
```

b) Steps and commands

	Step	Text	Command	Notes
1	Prepare Y4 ensemble	Remove EC loops in preparation for docking	Use find_loops.py within pymol and getloops_all.sh and removeloopcoordinates_all.sh	Finds loop residues in pymol, then sets them to coordinates of 0, then removes residues with coordinates of 0
2	Prepare PP helix	Remove c-tail from PP	Got hpp ensemble from PDB (1ljv) and manually cut off c-tail	PP PDB ensemble that contains only residues "PEQMAQYAAELRRYINML"
3	Prepare PP helix	Select 4 different docking start conformations by clustering	bcl.exe PDBCompare -quality RMSD -atom_list CA -pdb_list pdb_list.ls -prefix pphelix_rmsd -norm100 -aaclass AACaCb -convert_to_natural_aa_type -scheduler PThread 12 bcl.exe Cluster -distance_input_file pphelix_rmsdRMSD.txt -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster_pphelix -linkage Average -output_pymol 1000 .8 100 10000 10 dendogram_pphelix.py -remove_internally_similar_nodes .55 -pymol_label_output_string -scheduler PThread 12 grep "Leaf : 1 : " cluster_pphelix.Centers. | sort -nk10	
4	Position helix	Place the helix at the starting position based on experimental findings	Manually add the 4 different helix conformations to npy4 model pdbs.	This creates 37 * 4 total starting models
5	Generate individual span/constraints	Adjust span and constraints for each starting model depending on what loop residues were removed.		
6	Dock helix	Relax models with helix placed using new span and cst files.	/blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_source/build/src/release/linux/2.6/64/x86/gcc/4.5/rosetta_scripts.default.linuxgccrelease @dock_helix_y4_1.options -parser:protocol dock_helix_y4_1.xml -in:file:s npy4_ensemble_1_noloops_helix1.pdb -constraints:cst_fa_file npy4_ensemble_1_noloops.cst -nstruct 100 -out:file:silent docked_npy4_ensemble_1_noloops_helix1.out -out:file:scorefile docked_npy4_ensemble_1_noloops_helix1.sc -spanfile npy4_ensemble_1_noloops.span	Generates 14,800 models total (100 models per receptor ensemble/pphelix combination)
7	Select top models	Top 3 scoring models from each of the 37 starting receptor models	grep for score and extract the top 3 scoring from each.	
8	Prepare cterm	Attach cterm residues to npy4+pphelix models		
9	Prepare files for rebuilding c-term	Create secondary structure predictions and fragment files (Need to do this for each starting model)	runss x.fasta make_fragments.pl -id temp1 npy4_ensemble_1_pphelix.fasta Affix the header to the necessary files so that they can be used: cat header.txt temp1.jufo_ss > npy4_ensemble_1_pphelix1.jufo.ss2 cat header.txt temp1.psipred_ss > npy4_ensemble_1_pphelix1.psipred.ss2 mv status.200_v1_3_aatemp1 > status.200_v1_3_npy4_ensemble_1_pphelix	Generates .jufo.ss2, .psipred.ss2, .checkpoint, .chk, .psipred, .dat, .jufo_ss, frags.200.3mers, frags.200.9mers for each ensemble model.
10	prepare files for rebuilding	Create rdb.ss2 file, lips4, fragment_picker	/blue/meilerlab/apps/scripts/legacy/runsam npy4_ensemble_1_pphelix.fasta cat header.txt npy4_ensemble_1_pphelix.rdb > npy4_ensemble_1_pphelix.rdb.ss2	.rdb.ss2, .lipo, .lips4, .pdb.color, .raw, 3mers, 9mers for each

	c-term part 2	fragments	/sb/meiler/rosetta/rosetta-3.2/rosetta_source/src/apps/public/membrane_abinitio/run_lips.pl npy4_ensemble_1_pphelix.fasta npy4_ensemble_1_noloops_cterm.span /sb/meiler/Linux2/x86/blast/blast-2.2.18/bin/blastpgp /sb/meiler/scripts/sequence_analysis/db/nr /sb/meiler/rosetta/rosetta-3.2/rosetta_source/src/apps/public/membrane_abinitio/alignblast.pl /blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_source/build/src/release/linux/2.6/64/x86/gcc/4.5/fragment_picker.default.linuxgccrelease @quota.options -frags:describe_fragments npy4_ensemble_1_pphelix1_frags.fsc.fsc -out:file:frag_prefix npy4_ensemble_1_pphelix1_frags -in:file:checkpoint npy4_ensemble_1_pphelix.checkpoint -in:file:s npy4_ensemble_1_pphelix1_cterm.pdb -frags:ss_pred npy4_ensemble_1_pphelix.psipred.ss2 psipred npy4_ensemble_1_pphelix.rdb.ss2 sam npy4_ensemble_1_pphelix.jufo.ss2 jufo	ensemble model.
11	Build PP C-term	Use Topology Broker to build the C-term of PP following constraints	/blue/meilerlab/home/hirstsj/rosetta/rosetta_clean/rosetta_source/bin/r_broker.default.linuxgccrelease -database /blue/meilerlab/home/hirstsj/rosetta/rosetta_clean/rosetta_database/ @fold_cterm.options -in:file:s npy4_ensemble_1_pphelix1_cterm.pdb -in:file:spanfile npy4_ensemble_1_noloops_cterm.span -in:file:lipofile npy4_ensemble_1_pphelix.lips4 -in:file:fasta npy4_ensemble_1_pphelix.fasta -in:file:frag3 npy4_ensemble_1_pphelix1_frags.200.3mers -in:file:frag9 npy4_ensemble_1_pphelix1_frags.200.9mers -broker:setup npy4_ensemble_1_pphelix1_cterm_setup.tpb -out:nstruct 100 -out:file:scorefile npy4_ensemble_1_pphelix1.sc -out:file:silent npy4_ensemble_1_pphelix1.out	Generates 11,700 models total (100 models for each of the docked helix models)
12	Add NPY4 loop residues	Add loop residues that were removed for docking PP.	Manually add loop residues at coordinates 0,0,0 or using script. renumber_pdb.py input.pdb output.pdb after each addition of loop segments.	
13	Rebuild loops		/sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/loopmodel.default.linuxgccrelease -database /sb/meiler/rosetta/rosetta-3.4/rosetta_database/ @npy4_ensemble_1.flags -loops:input_pdb npy4_ensemble_1_hpp1_full.pdb -out:prefix npy4_1_hpp1_final" -out:nstruct 100 -out:file:scorefile npy4_1_hpp1_final.sc -out:file:silent npy4_1_hpp1_final.out	Generates 100 models per initial file (16,600 models total)
15	Cluster models	cluster models within each receptor ensemble model. Collect top scoring models within the 3 largest clusters.	bcl.exe PDBCompare -quality RMSD -atom_list CA -pdb_list pdb_list.ls -prefix 1_rmsd -norm100 -aaclass AACaCb -convert_to_natural_aa_type -scheduler PThread 12 bcl.exe Cluster -distance_input_file 1_rmsdRMSD.txt -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster_1 -linkage Average -output_pymol 1000 4 100 10000 10 dendogram5_1.py -remove_internally_similar_nodes 4 -pymol_label_output_string -scheduler PThread 12	
16	High resolution constraint relax	Perform relaxation using adjusted high-resolution constraints and addition of new constraint.	/sb/meiler/rosetta/rosetta-3.4/rosetta_source/bin/relax.default.linuxgccrelease @post_relax_flags.txt -in:file:s final_1.pdb -in:fix_disulf y4pep.disulfide -in:file:spanfile y4pep.span -out:file:silent final_relax_1.out -out:prefix final_relax_1 -nstruct 100	Generates 2300 models (100 for each starting model)
18	Contact map analysis	Calculate number of models in final ensemble that contain residue pairs between NPY4 and PP closer than 8 angstroms.	rosetta_scripts.default.linuxgccrelease -parser:protocol contact.xml -database /sb/meiler/rosetta/rosetta-3.4/rosetta_database/ -no_output -in:file:s final_relax_final_11.pdb_7final_11_0016_0001.pdb	Generates .contacts file for each model which contains a list of all residue pairs between NPY4 and PP and lists a 1 if they are within the cutoff and 0 if they are not based on Cbeta atom distance.

Chapter 6

Modeling Interactions of the Melanocortin 4 Receptor and α-MSH

---

manuscript in preparation

6.1 Abstract

Melanocortin 4 receptor (MC4R) haploinsufficiency is the most common known form of monogenic obesity, making this system an attractive target for the treatment of obesity and related diseases. However, efforts to develop treatments that target the MC4R and its endogenous peptide agonist α-MSH have seen little success. Therefore, a detailed understanding of the interaction between MC4R and α-MSH may help researchers develop drugs that effectively potentiate MC4R signaling. The present chapter uses comparative modeling and flexible peptide docking with Rosetta Molecular Modeling Suite to model the interaction between MC4R and α-MSH. A variety of experimental evidence suggests specific interactions used to guide docking. Additionally, docked models were analyzed and predicted residue contacts showed good agreement with additional mutation data.

6.2 Introduction

The melanocortin 4 receptor (MC4R) is one of five class A G-protein coupled receptors (GPCRs) in the melanocortin (MC) system [1]. Named for its combination of melanotropic and adrenocorticotropic activities [2], the MC system participates in the regulation of feeding, bone metabolism, cardiovascular function, erectile function, drug abuse, inflammation, adiposity, energy homeostasis, pigmentation, and neuromuscular regeneration [2-10]. Endogenous agonists of the MC system primarily result from post-translational modification of the prohormone proopiomelanocortin (POMC) into four melanocortins: α-, β-, and γ- melanocyte stimulating hormone (MSH) and adrenocorticotrophic hormone (ACTH), the opioid β-endorphin, and β- and γ- lipotropin (LPH) [1, 2, 5]. POMC is primarily expressed in the central nervous system within the nucleus of the solitary tract (NST), the arcuate nucleus of the hypothalamus (Arc), and the pituitary [11, 12], but is also expressed in the skin, spleen, thyroid, and GI tract [13]. In addition to the peptide agonists, the MC system also contains two endogenous peptide competitive antagonists: agouti and agouti related protein (AgRP) [2, 5]. Additionally, the N-terminal region of MC4R is thought to act as a tethered agonist [14].

MC4R haploinsufficiency is the most common known form of monogenic obesity [15]. Since the first report of obesity-associated human MC4R mutations in 1998 [16, 17], approximately 150 naturally occurring MC4R gene mutations have been identified among patients [15] contributing to between 0.5% (Italian and Belgian) and 6% (British Caucasian) of early-onset morbid obesity [18, 19]. A recent analysis of obese European individuals found a prevalence of MC4R loss of function in 1.8% obese children and 1.6% obese adults [20]. Human MC4R deficiency is characterized by hyperphagia, increased longitudinal growth, increased adiposity, and severe hyperinsulinemia, with more severe symptoms in homozygous carriers [1]. Taken together, over 103 residues of MC4R have seen mutations in patient cohorts covering 31% of the total receptor [21]. Interestingly, two naturally occurring mutations in MC4R have been negatively correlated with obesity: V103I polymorphism appears to reduce the risk of obesity by 20% while I251L polymorphism appears to reduce the risk of obesity by 50% [22].

Genetic studies in mice have provided extensive insights into the role of the MC4R in appetite and energy homeostasis. One of the oldest known genetic models of obesity is the agouti mouse which constitutively expresses agouti in all tissues and causes a yellow coat and obesity through antagonism of MC1R and MC4R [23, 24]. The MC4R knockout mouse, generated in 1997, reflects many phenotypes of the agouti mouse [25] including maturity onset obesity characterized by hyperphagia, increased adiposity, increased longitudinal growth, normal lean body mass, hyperinsulinaemia, and hyperleptinaemia [4].

The hypothalamic MC system appears to be a convergence center for peripheral and central factors that regulate feeding behavior and metabolism [5]. In the simplest sense, when anorexigenic POMC neurons are activated in a well fed state, POMC cleavage products α- and β-MSH activate MC4R receptors to decrease feeding behavior. When orexogenic NPY/AgRP neurons are activated in a fasting state, AgRP antagonizes MC4R to increase feeding behavior [26]. Many endogenous factors contribute to the regulation of these neurons including leptin, insulin, glucose, ghrelin, peptide YY, NPY, β-endorphin, serotonin, GABA, melanin-concentrating hormone, and orexins [5, 27]. Leptin, for example, is released from adipocytes in amounts directly related to body fat mass and activates POMC neurons of the Arc while at the same time inhibiting nearby NPY/AgRP neurons [3]. Ghrelin, on the other hand, appears to oppose this action of leptin [28].

Due to the prevalence of MC4R defective mutations in obesity, development of MC4R specific drugs are actively being pursued [4]. However, the MC system has proven to be a difficult therapeutic target. Local injection of α-MSH into the PVN results in reduction of food intake in mice and rats but this effect fades over time [4]. To date no MC4R agonist has progressed past phase I except MK-0493, which subsequently failed in phase II [29, 30]. A comprehensive understanding of the interaction between peptide and non-peptide modulators and MC4R, therefore, can provide useful tools for structure-based drug discovery and lead improvement.

MC4R binds α-MSH and ACTH with approximately equal affinity and β- and γ-MSH with slightly lower affinity [5, 31]. However, hypothalamic α-MSH appears to be the most important endogenous MC4R agonist involved in energy regulation [32]. α-MSH is a linear thirteen amino acid peptide capable of adopting many conformations owing to its high flexibility [33]. As with all MSH peptides, the conserved four residue motif (His-Phe-Arg-Trp) is critical for activation of MC4R by α-MSH. This fragment alone is capable of activating MC4R *in vitro* [34]. Therefore, many additional MCR ligands have been developed based on similarity with this pharmacophore. One of the most important of these ligands is [Nle^4,dPhe^7]-α-MSH or NDP-MSH, a more stable and potent version of α-MSH [35].

Modifications of α-MSH reveal that, despite its flexibility, a common reverse β-turn around the core residues 5-9 is important for activity [36-39]. The aromatic ring of Phe^7 has been suggested as the most important pharmacophore element for melanocortin receptor activation [40]. Substitution of Phe^7

with larger aromatic groups such as d-Nal(2’) in the case of SHU9119 switches the behavior from agonist to antagonist further underscoring the importance of this residue [41].

There are currently no experimentally determined MCR structures available. Therefore, a comprehensive understanding of the interaction between MC4R and α-MSH must involve an approach that combines comparative modeling and peptide ligand binding. Fortunately, a wealth of mutagenesis data is available to guide and evaluate the modeling of MC4R with α-MSH.

MC4R is a 332 amino acid GPCR encoded by a single exon gene [42]. Although much of the characteristic class A GPCR architecture is present, MC4R is predicted to have some unique structural properties. The highly conserved proline in transmembrane helix five is replaced with methionine in MC4R and the asparagine in the conserved NPxxY motif of helix 7 is replaced with aspartic acid. Another major difference with MC4R is unusually short intracellular loops and extracellular loop 2 (ECL2) which is missing a conserved cysteine that normally forms a disulfide bond with the top of transmembrane helix three. Instead, MC4R contains two putative disulfide bonds: Cys^{271} forms an intra-loop disulfide bond with Cys^{277} in ECL3 and a disulfide bond between Cys^{40} and Cys^{279} connects the N-terminal with ECL3 [43, 44].

Several direct interactions between α-MSH and MC4R have been experimentally elucidated. Consistent mutagenesis results have elucidated two potential binding sites that engage the different sides of the core tetrapeptide’s β-turn. An acidic binding pocket formed by negatively charged residues including $Glu^{2.60}$, $Asp^{3.25}$, and $Asp^{3.29}$ is thought to directly interact with one side of the turn [36, 45-49]. Specifically, Arg^8 is predicted to form direct ionic interactions with $Asp^{3.25}$ and $Asp^{3.29}$. Mutations in these residues decrease affinity for Arg^8 containing agonists but not Nle^8 containing antagonists [45, 48]. A second hydrophobic binding pocket, specifically residue $Phe^{6.51}$ is thought to interact with Phe^7 [36, 48, 50, 51].

The present chapter uses the common topology of class A GPCRs to model the structure of MC4R. Due to the unique properties of MC4R described above, a multi-template comparative modeling approach called RosettaCM [52] is used to combine low energy fragments from different templates instead of relying on a single template. This model is then used to dock α-MSH with special attention to the flexibility of the peptide and the extracellular region of MC4R to which it binds. The two consensus binding pockets previously established to interact with Arg^8 and Phe^7 of α-MSH are used to guide the docking process. Finally, the models are used to predict the interface between α-MSH and MC4R and this interface is evaluated in terms of all additional mutagenesis data that implicates residues beyond those used to guide docking. These models may also propose residues of interest that have not yet been characterized in terms of α-MSH binding.

6.3 Results

A conformational ensemble captures flexibility of MC4R-α-MSH interface

Based on the flexibility of the linear peptide α-MSH and the flexible MC4R loop regions of its binding environment, an ensemble of MC4R conformations was used for docking instead of a single MC4R conformation. RosettaCM multi-template comparative modeling was used to generate an ensemble of 20 MC4R models. To quantitate the degree of conformational variability represented by the ensemble, a pair-wise RMSD analysis was performed. The highest degree of similarity between models in this ensemble (lowest RMSD) is 2.87 Å and the greatest dissimilarity is 5.34 Å. As expected, the conformational difference between models is found primarily in the disordered loop regions. When comparing transmembrane helix regions only, the most similar models within the ensemble differ by an RMSD of 1.20 angstroms and the most dissimilar models differ by an RMSD of 2.8 angstroms. MC4R comparative models are shown in figure 6.1.

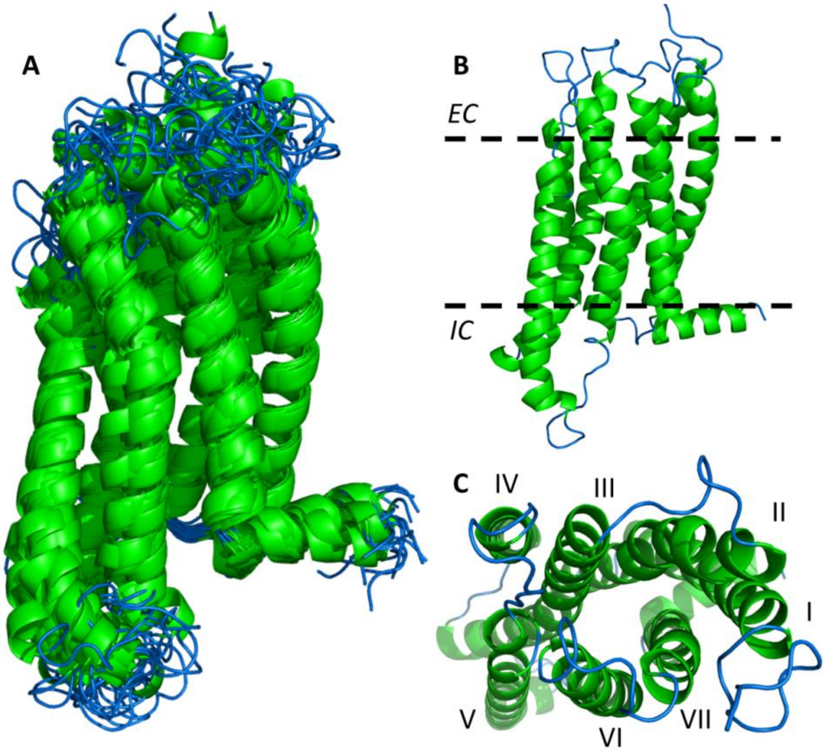


Figure 6.1 MC4R comparative modeling ensemble. A) Ensemble of 20 comparative models of MC4R used for α-MSH docking. Transmembrane helix regions converge on a similar topology across 20 receptor models, whereas loop regions in blue show high conformational variability. B) Representative model of MC4R from a side view with transmembrane region indicated (EC: extracellular region; IC: intracellular region). C) Representative model of MC4R looking down from the extracellular surface. Transmembrane helices are numbered in order from N-terminus to C-terminus.

Docking α-MSH to the ensemble of MC4R comparative models yielded an ensemble of 21 complex conformations with similar interface energy scores. In general, a consensus binding mode was found in which the core tetrapeptide of α-MSH docked into a shallow region of the transmembrane pore and both termini of α-MSH extended away from the membrane, reaching outside of the loops. A second pair-wise model RMSD analysis was conducted, focusing on the conformational variability of α-MSH across these 21 docked poses. The conformational space covered by α-MSH in these models was moderate, the most dissimilar α-MSH poses having an RMSD of 6.16 Å and the most similar α-MSH poses having an RMSD of 1.86 Å. This difference was primarily in the terminal loop residues of α-MSH as opposed to the tetrapeptide core critical for binding and activation. The average RMSD for the tetrapeptide core alone was 1.17 Å and the most similar tetrapeptide poses had an RMSD of 0.26 Å. This suggests that the docking procedure converged on a general position of α-MSH with high degrees of flexibility in the most terminal regions of the peptide. This is expected given the high flexibility of α-MSH and the flexible extracellular loops of MC4R that surround these residues. A representative ensemble of poses show the different degrees of conformational variability between the two regions of alpha-MSH in figure 6-2***A***.

Despite a general convergence of tetrapeptide core position, there was some degree of variability in the central binding mode of the 21 models. Examination of these core residues reveals two major modes of variability: His^6/Phe^7 positioning and Trp^9 side-chain conformation. The greatest mode of variability concerns the positioning of His^6 and Phe^7 and specifically which residue occupies a pocket formed by MC4R residues of TM6 and TM7. In six of the 21 models, the side chain of Phe^7 occupies this pocket, whereas in 15 models, His^6 occupies it. Comparing the interface scores of these two binding poses suggests that poses in which His^6 occupies this pocket are more energetically favorable than poses in which Phe^7 occupies it. Models with His^6 in the pocket have an average interface score of -45.1 Rosetta Energy units (REU), whereas models in which Phe^7 occupies this core have an average interface score of -39.2 REU (t-test, $p < .05$). Because models with Phe^7 occupying this pocket are less energetically favorable, these models were removed from the final pose ensemble. A comparison of these slightly different binding poses is shown in figure 6-2***B***.

In six of the 21 models, the side chain conformation of Trp^9 is oriented out of the receptor pore (upwards) and away from the membrane. In the other 15 models, Trp^9 points into the pore (downwards) and lies closer to the membrane. However, since both side chain conformations produce comparable interface binding energies, neither conformation can be ruled out. Models containing downwards facing Trp^9 conformations have an average interface score of -43.1 REU, whereas models containing an upwards conformation of Trp^9 show an average binding interface score of -44.2 REU ($p > 0.7$). A comparison of these different side chain conformations is shown in figure 6-2***C***.

Following the removal of less energetically favorable poses, average RMSD of the tetrapeptide core across the final ensemble of 15 complex models dropped slightly to 1.09 Å.

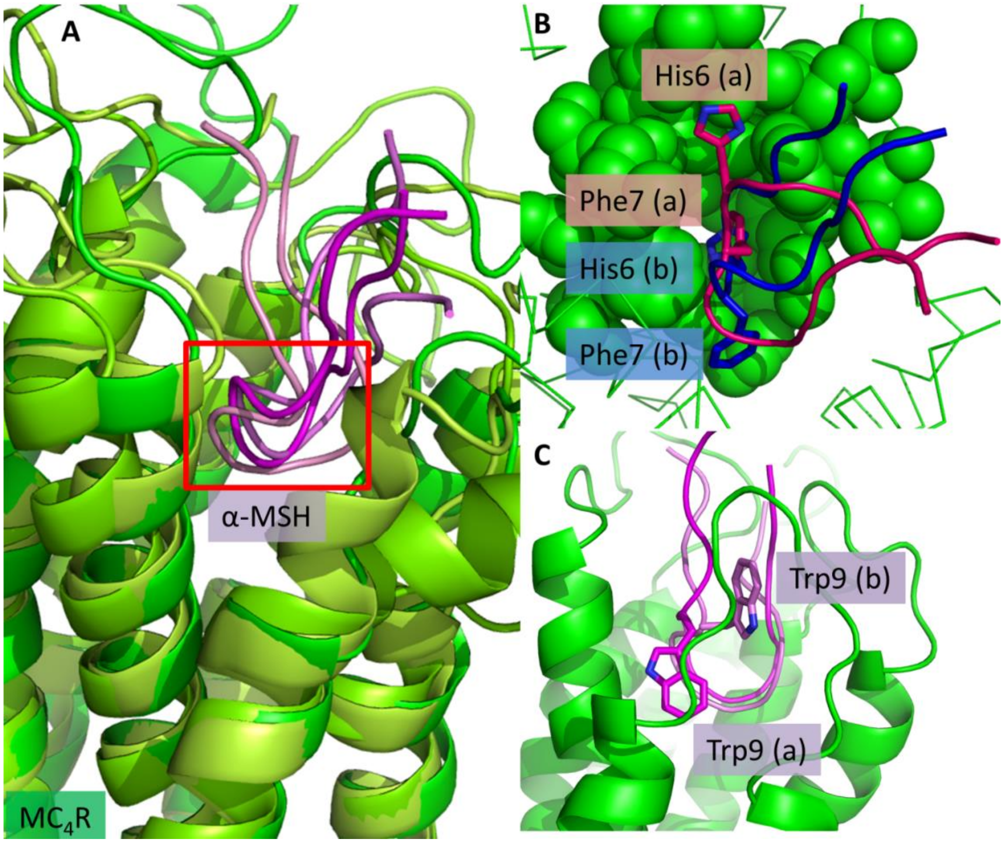


Figure 6.2 α-MSH variability across 21 model complexes. A) Overall convergence is seen at the central core of α-MSH (indicated by a red box) and variability increases towards the termini residues. A representative sample of three models is shown for clarity. B) Significant variability of two core α-MSH residues are highlighted. Green spheres indicate the binding site of MC4R where this variability occurs. The red pose indicates a less energetically favorable pose where the side-chain of Phe7 (a) points into this binding site and His6 (a) points "up" and away. In the blue pose, His6 (b) replaces Phe7 (a) in the center of this pocket and Phe7 (b) points "down" into the transmembrane pore of MC4R. C) Variability in the side-chain conformation of Trp9 in docked alpha-MSH poses. Both poses show equal interface energy scores.

Contacts used to guide docking were captured in final model ensemble

A consensus approach was used to elucidate the predicted binding site and contacts between α-MSH and MC4R across final conformations. This approach assumes that within the conformational space of the final ensemble of 15 docked models, the most important residue contacts should be retained across models even if these contacts occur in slightly different positions or orientations.

Four potential interactions were used to guide the docking of the α-MSH to MC4R. These interactions span two well defined binding sites: an ionic binding site that includes interactions between

α-MSH Arg^{8} and MC4R residues Glu$^{2.60}$, Asp$^{3.25}$, and Asp$^{3.29}$ and a hydrophobic binding site that includes a direct interaction between Phe$^{6.51}$ of MC4R and Phe^{7} of α-MSH. These interactions were used to guide docking only through atom-pair distance constraints. Final models were selected, in part, due to their adherence to these distance constraints. To elucidate whether interactions between these residues actually contribute to the overall interface energy, a residue-pair analysis of interface energy score was performed. Residue pairs between α-MSH and MC4R that contribute favorable negative energy scores to the overall interface score consistently across most or all of the docked models suggest direct contacts between α-MSH and MC4R with higher confidence.

An interaction between Arg^{8} and Glu$^{2.60}$ was captured in 13 models with an average score of -2.2 REU. An interaction between Arg^{8} and Asp$^{3.25}$ was captured in 8 models with an average score of -1.9 REU. An interaction between Arg^{8} and Asp$^{3.29}$ was captured in 13 models with an average score of -4.2 REU. All models individually captured at least two of the three Arg^{8} interactions. An interaction between Phe^{7} and Phe$^{6.51}$ was captured in all 15 models with an average score of -1.1 REU. The modeled binding pose of α-MSH and MC4R is represented by a single model with the lowest interface energy score in figure 6.3. The three atom-pair distance constraints involving Arg^{8} are highlighted for this model in figure 6.3***B*** and the atom-pair distance constraint involving Phe^{7} is highlighted in figure 6.3***C***.

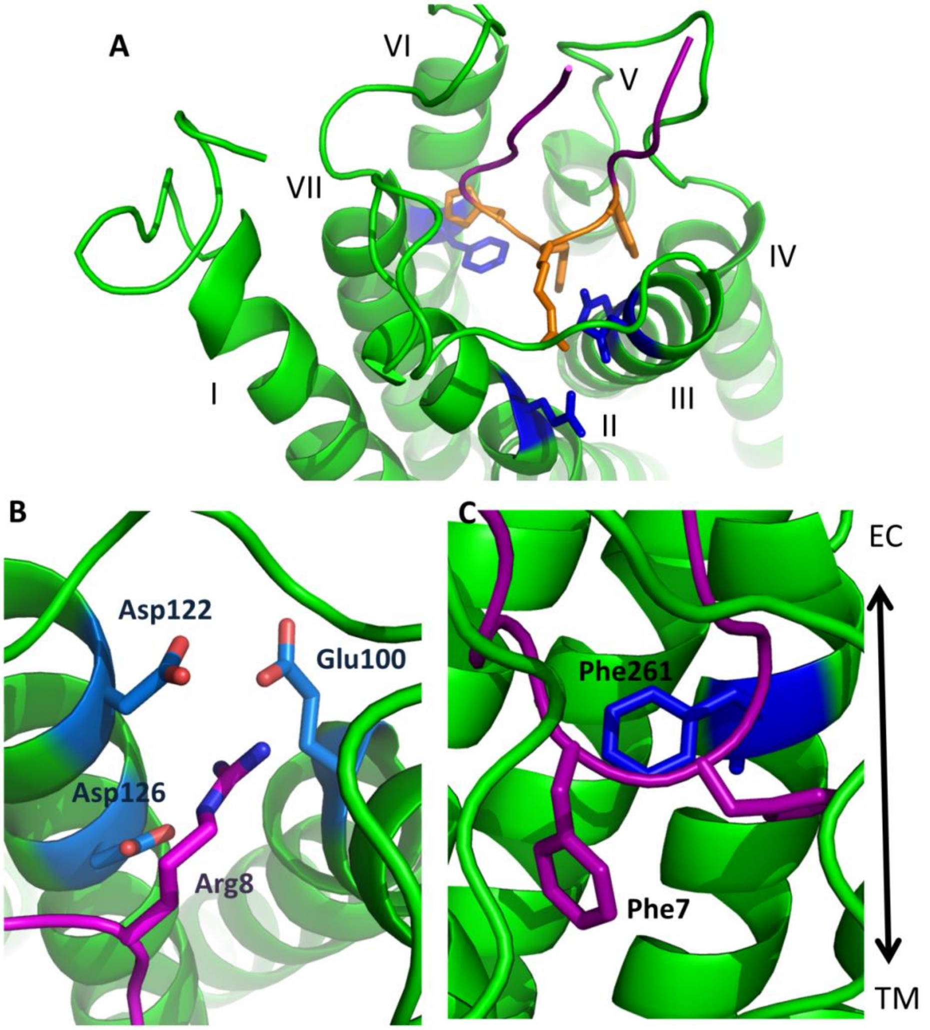


Figure 6.3 α-MSH docked to MC4R comparative model. A) Representative model of α-MSH (purple) docked to MC4R (green). Blue residues indicate receptor residues used to guide docking and orange residues indicate critical tetrapeptide residues docked during the first phase of docking. TM helices are numbered from N- to C-terminus. Viewpoint is from extracellular side looking “down” towards membrane. B) Experimentally established binding site with Arg8 involves up to three residues. Potential contacts between Arg8 and three MC4R residues including Glu100, Asp122, and Asp126 were captured in a single binding site for Arg8 and TM helices 2 and 3. C) Experimentally established contact between Phe7 of α-MSH and Phe261 of MC4R was captured in docked models. In both images, α-MSH is shown in purple, MC4R is shown in green, and the receptor contact residues are shown in blue. Orientation is indicated with EC: extracellular, TM: transmembrane.

The predicted interface between α-MSH and MC4R can be divided into three regions

A consensus approach was used to predict the overall binding interface between α-MSH and MC4R. Within a single model, residues of MC4R are considered as potential binding interface residues if they fall within 6 Å of any residue in α-MSH. Computing the percentage of agreement across the 15 ensemble models gives a measure of confidence for each of these residues. Residues are considered to potentially line the interface with α-MSH if they appear within 6 Å of α-MSH across 75% or more of the docked models. This analysis yielded three distinct binding sites that face a different portion of the α-MSH core. All residues predicted to line the three binding regions with α-MSH are illustrated in figure 6.4 and listed in table 6-1.

Table 6-1 Prediction binding site between α-MSH and MC4R can be broken into three sections based on nearby α-MSH residue. *Residues are numbered according to Ballesteros-Weinstein numbering and full sequence number in parenthesis. **Residue is also within 6 Å of Phe7.

Binding Region 1 (α-MSH: Arg8)		**Binding Region 2 (α-MSH: Glu5+His6)**		**Binding Region 3 (α-MSH: Trp9)**	
Residue*	**Location**	**Residue***	**Location**	**Residue***	**Location**
Glu 2.60 (100)	TM2	Phe 6.51 (261)	TM6	Phe4.60 (184)	TM4
Thr 2.61 (101)	TM2	Phe 6.52 (262)	TM6	Ile4.61 (185)	TM4
Val 3.22 (119)	TM3	His 6.54 (264)	TM6	Tyr4.63 (187)	TM4
Asn 3.23 (120)	TM3	Leu6.55 (265)**	TM6	Ser4.64 (188)	ECL2
Asp 3.25 (122)	TM3	Phe6.57 (267)	TM6	Asp4.65 (189)	ECL2
Asn 3.26 (123)	TM3	Tyr6.58 (268)	TM6	Ser4.66 (190)	ECL2
Ile 3.28 (125)	TM3	Ile 6.59 (269)	TM6	Ser5.37 (191)	ECL2
Asp 3.29 (126)	TM3	Cys 6.61 (271)	ECL3	Ala5.38 (192)	ECL2
Ile 3.32 (129)**	TM3	Gln 6.63 (273)	ECL3	Val5.39 (193)	ECL2
Cys 3.33 (130)**	TM3	Tyr 7.27 (276)	ECL3	Cys5.42 (196)	TM5
		Cys 7.28 (277)	ECL3	Leu5.43 (197)	TM5
		Val 7.29 (278)	ECL3		
		Cys 7.30 (279)	TM7		
		Phe 7.31 (280)	TM7		
		Met 7.32 (281)	TM7		
		Ser 7.33 (282)	TM7		
		His 7.34 (283)	TM7		
		Phe 7.35 (284)**	TM7		
		Asn 7.36 (285)	TM7		
		Tyr 7.38 (287)	TM7		
		Leu 7.39 (288)	TM7		

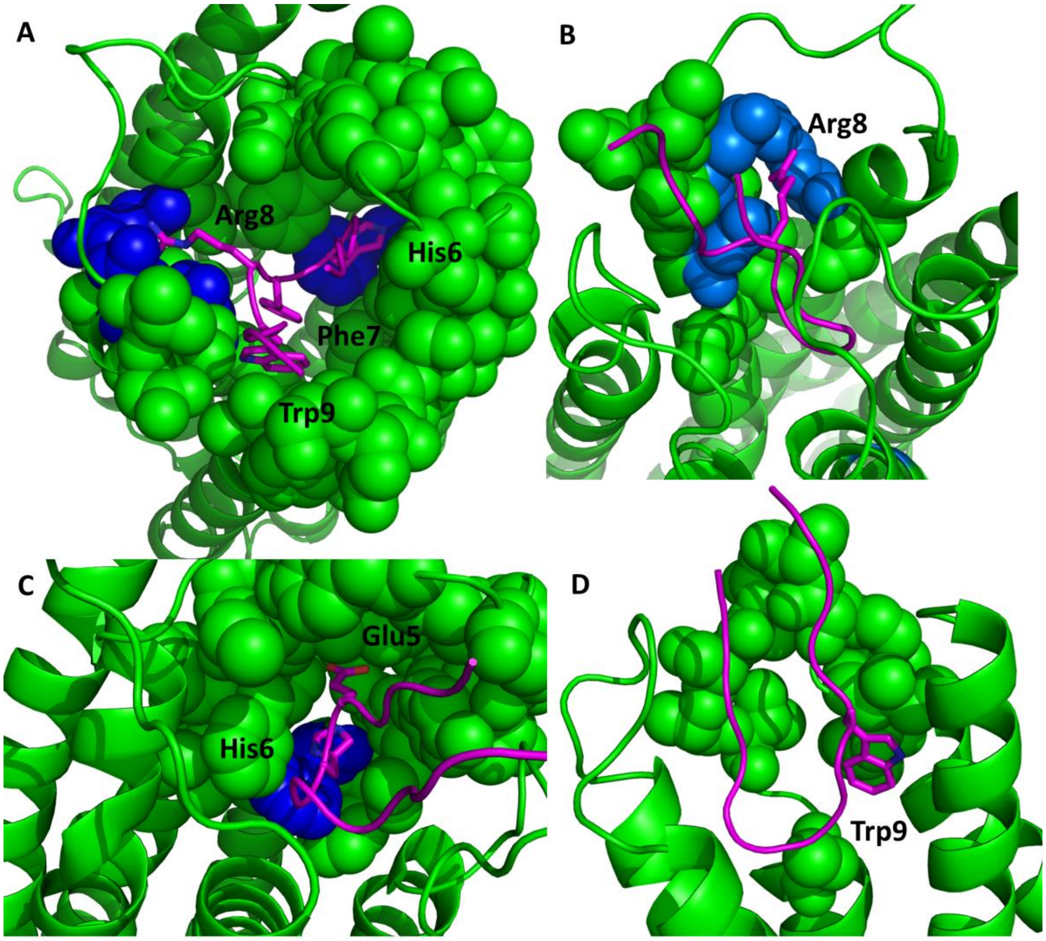


Figure 6.4 Three MC4R binding interfaces face different sections of the α-MSH. A) Residues appearing in the interface between α-MSH and MC4R in 13 or more of the 15 models are shown as spheres. Full side chains of α-MSH core tetrapeptide are shown in purple. MC4R residues used as constraints to guide docking are shown in blue. Image is orientated so that view is from extracellular surface looking "down" into transmembrane core. Specific residues for each of these binding sites are listed in table 4. B) Section of binding interface facing Arg8 of α-MSH. Full side chain of Arg8 only is shown. C) Section of binding interface facing Glu5 and His6 of α-MSH. Full side chains of Glu5 and His6 are shown. D) Section of binding pocket facing Trp9 of α-MSH. Full side chain of Trp9 only is shown.

The first binding region includes residues of TM2 and TM3 and involves the three contacts with Arg^{8} used to guide docking. Residues that line this binding site all face Arg8 and include Glu$^{2.60}$, Thr$^{2.61}$, Val$^{3.22}$, Asn$^{3.23}$, Asp$^{3.25}$, Asn$^{3.26}$, Ile$^{3.28}$, Asp$^{3.29}$, Ile$^{3.32}$, and Cys$^{3.33}$. This binding site is illustrated in figure 6.4***B***.

The second binding region consists of residues in the upper regions of TM6 and TM7 and face α-MSH residues Glu^{5} and His^{6}. These residues include Phe$^{6.51}$, Phe$^{6.52}$, His$^{6.54}$, Leu$^{6.55}$, Phe$^{6.57}$, Tyr$^{6.58}$, Ile$^{6.59}$, Cys$^{6.61}$, Gln$^{6.63}$, Tyr$^{7.27}$, Cys$^{7.28}$, Val$^{7.29}$, Cys$^{7.30}$, Phe$^{7.31}$, Met$^{7.32}$, Ser$^{7.33}$, His$^{7.34}$, Phe$^{7.35}$, Asn$^{7.36}$, Tyr$^{7.38}$ and Leu$^{7.39}$. His^{6} appears to enter a pocket formed by the lower residues, whereas Glu^{5} faces the residues at the extracellular ends of TM6 and TM7. This binding site is illustrated in figure 6***C***.

The third binding region involves residues of TM4, TM5, and ECL2. These residues include Phe$^{4.60}$, Ile$^{4.61}$, Tyr$^{4.63}$, Ser$^{4.64}$, Asp$^{4.65}$, Ser$^{4.66}$, Ser$^{5.37}$, Ala$^{5.38}$, Val$^{5.39}$, Cys$^{5.42}$, and Leu$^{5.43}$. Trp^{9} faces this binding site in all 15 models. However, the docked ensemble does not converge on a single side-chain conformation of Trp^{9} and the area of this binding site is likely over-estimated due to the varibility. This binding site is illustrated in figure 6***D***.

Phe7 of α-MSH interfaces with two of the three binding regions

As mentioned previously, Phe^{7} is a critical α-MSH residue thought to play important roles in binding and activation of MC4R. In all 15 models, Phe$_{7}$ was oriented in the approximate center of the transmembrane helix pore with the side chain pointed towards the intracellular side of the receptor. In this position and conformation, Phe^{7} appears to be shared by two of the three binding interfaces. The intracellular-most residues of the His^{6} binding site were found within 6.0 Å of Phe^{7}, including the expected contact Phe$^{6.51}$, and residues Leu$^{6.55}$ and Phe$^{7.35}$. Several residues close to Arg^{8} were also within 6.0 Å of Phe^{7}, including Ile$^{3.32}$ and Cys$^{3.33}$.

Sixteen MC4R residues contribute to the binding energy score with α-MSH

In addition to the four residue contacts used to guide docking, potential direct interactions between α-MSH and MC4R can be predicted by examining all consensus residue-pair energy scores within the interface. Across 60% of the models, 12 additional MC4R residues paired with α-MSH residues to contribute favorable negative energy scores. Several MC4R residues paired with multiple α-MSH residues to contribute to the binding energy score. All consensus interactions involved one of the core tetrapeptide residues except for a single weak interaction between Ser^{4} and Met$^{7.32}$, which appeared in 11 models with an average score of -0.3 REU and several interactions involving Glu^{5}. Interactions with Glu^{5} include Tyr$^{6.58}$ (93% models, average score -1.5 REU), Phe$^{7.31}$ (87% models, -0.8 REU), and Met$^{7.32}$ (73%, -0.7 REU).

Interactions that involve His^{6} include Phe$^{6.51}$ (100% models, average score -0.9 REU), His$^{6.54}$ (93% models, -1.0 REU), Leu$^{6.55}$ (100% models, -0.9 REU), Tyr$^{6.58}$ (100% models, -0.9 REU), Phe$^{7.31}$ (100% models, -0.7 REU), Phe$^{7.35}$ (100%, -1.4 REU), and Tyr$^{7.38}$ (100%, -0.2 REU).

In addition to the expected interaction with Phe$^{6.51}$, interactions that involved Phe^{7} include Ile$^{3.32}$ (100% models, -0.7 REU), Cys$^{3.33}$ (100% models, -0.5 REU), Leu$^{6.55}$ (60% models, -0.9 REU), and Phe$^{7.35}$ (100% models, -1.4 REU).

Arg8 did not contribute to the binding energy score outside of the three interactions used to guide docking (Glu$^{2.60}$, Asp$^{3.25}$, and Asp$^{3.29}$).

Interactions involving Trp^{9} include Phe$_{4.60}$ (80% models, -1.0 REU), Ile$^{4.61}$ (60% models, -1.2 REU), and Tyr$^{4.63}$ (60% models, -0.8 REU).

6.4 Discussion

In this study, a specialized multi-template comparative modeling technique called RosettaCM was used to combine fragments from 20 GPCR templates to model the MC4 receptor. A two stage process was used to dock α-MSH to MC4R: the core tetrapeptide (His^{6}-Phe^{7}-Arg^{8}-Trp^{9}) was first docked into the putative binding site guided by residue pair restraints reflecting previous modeling and mutagenesis data. Secondly, the terminal residues were modeled simultaneously with the extracellular loops of MC4R to simulate the flexibility of these regions. Additionally, the final model analysis focused on potential conformational flexibility of the final docked pose. Therefore, rather than selecting a single model over which to predict and characterize the binding site, an ensemble of 15 models was compared to identify a consistent binding interface across models with comparable interface energy scores.

This modeling approach identified a binding pose of α-MSH that involves three receptor regions: Arg^{8} faces residues of TM2 and TM3, Glu^{5} and His^{6} face residues of TM6 and TM7, and Trp^{9} faces residues of TM4, TM5, and ECL2. The α-MSH residue Phe^{7} point into the transmembrane poor and engages residues from both the Arg^{8} and Glu^{5}/His^{6} facing regions. Beyond Glu^{5} and the core tetrapeptide, residues from either terminal of α-MSH did not appear to consistently interact with residues of MC4R aside from a potential weak interaction between Ser^{4} and MC4R residue Met$^{7.32}$.

Previously identified contacts can be used to evaluate modeling

Many naturally occurring mutations have been identified with MC4R, largely due to their potential role in obesity. These mutations have been characterized based on their phenotypes, generally described as belonging to one of five classes: mutations causing decreased expression, intracellular receptor retention, impaired ligand affinity, impaired ligand efficacy and/or potency, and unknown effects. Some residues have been highly characterized in terms of their direct role in ligand binding. However, many are implicated in binding without a precise description of their role. Therefore, models can be used to characterize the specific role of residues in MC4R-α-MSH binding. Taken further, these models can be used to predict participating residues not previously implicated in ligand binding or activity.

As mentioned, four putative interactions were used to guide the docking process and the protocol was capable of capturing these interactions consistently across the final docked models. Three potential interactions involve an ionic binding site that interacts with Arg8 of α-MSH. Mutations of Glu$^{2.60}$, Asp$^{3.25}$, and Asp$^{3.29}$ across multiple studies reveal that this residue is critical for binding of both α-MSH and NDP-MSH in mouse and human systems [19, 27, 36, 44, 48, 51, 53, 54]. In the present models, all three interactions were captured in the majority of the models, with the strongest and most consistently captured interaction between Arg^{8} and Asp$^{3.29}$. The fourth potential interaction used to guide modeling was between Phe^{7} of α-MSH and Phe$^{6.51}$ of MC4R. A variety of mutagenesis studies in both mouse and human systems reveal that Phe$^{6.51}$ is critical for α-MSH binding to MC4R[36, 44, 45, 48, 51, 54]. An interaction between Phe^{7} and Phe$^{6.51}$ was reflected in all of the final MC4R-α-MSH models.

Interactions with Phe^{7} can predict differences in α-MSH and NDP-MSH binding

One of the major differences between α-MSH and NDP-MSH, the higher affinity modification of α-MSH, is the change in stereochemistry from L- to D- at Phe^{7} in NDP-MSH. Therefore, modeled interactions with α-MSH Phe^{7} that correlate with residues that, when mutated, disrupt α-MSH binding and not NDP-MSH binding, present potential agreement between modeling and experimental results.

An interaction between Phe^{7} and MC4R residue Ile$^{3.32}$ was captured in all models. This interaction agrees with I129A mutations that decreased α-MSH binding and efficacy but not NDP-MSH [44, 55, 56]. An interaction between Phe^{7} and Cys$^{3.33}$ was also captured in all models. However, it is unclear whether this is supported with current mutagenesis data. C130A shows normal NDP-MSH binding but α-MSH binding data is unavailable [36, 48, 54]. Additionally, Phe$^{7.35}$ is proposed to interact with either His^6 or Phe^7 in the current models. Previous mutagenesis data suggest a direct interaction between $Phe^{7.35}$ and Phe^7 due to a strong effect on α-MSH binding but not NDP-MSH binding [45, 51, 57].

Eight Predicted Interactions are Supported with Experimental Data

In addition to the four residues involved in constraints used to guide α-MSH docking, twelve residues were predicted to interact with one or more residue of α-MSH. Many of these residues have been characterized for their mutated effects on α-MSH and/or NDP-MSH binding but do not have sufficient evidence to suggest a direct interaction with the ligand. Comparing predicted interaction with published mutagenesis data helps to validate the presented models and propose direct interactions between these residues and specific ligand residues. Of these twelve predicted interactions, eight are supported by mutagenesis experiments that revealed a decrease in α-MSH affinity. These interactions and their corresponding experimental evidence are listed in table 6-2.

Models predict residue not previously tested for ligand binding to MC4R

In addition to residues participating in the interface energy between MC4R and α-MSH with previously published experimental data, the current models propose the involvement of one MC4R residue in binding α-MSH not yet been characterized with mutagenesis approaches. Binding studies similar to those used to characterize many other mutations within MC4R may support the identification of additional residues involved in α-MSH binding to MC4R. This interaction includes $Met^{7.32}$ predicted to interact with Ser^4 or Glu^5 on α-MSH. Taken together, the degree of corroboration between models and experimental data not used to guide docking improves the confidence in the presented models.

Table 6-2 Eight predicted interactions have supporting mutagenesis data in literature *mutagenesis data included for corresponding mouse MC4R residue **Green color indicates mutagenesis data supports prediction; red color indicates mutagenesis data does not support prediction; yellow color indicates previously untested residue. REU = Rosetta Energy Units

Residue	Average interface score (REU)	Potential α-MSH partner(s)	Experimental Evidence**
Ile3.32 (129)	-0.7	Phe7	🟩 I129A normal NDP-MSH binding and decreased α-MSH binding [55]
Cys3.33 (130)	-0.5	Phe7	🟥 C130M/A Normal α-MSH and NDP-MSH binding [36, 48]
Phe4.60 (184)	-1.0	Trp9	🟩 F184A normal binding [44]; F176S* abolished NDP-MSH binding (mouse) [48]; F184L normal NDP binding, decreased α-binding [55]
Ile4.61 (185)	-1.2	Trp9	🟩 F177K* decreased NDP binding (mouse) [48]
Tyr4.63 (187)	-0.8	Trp9	🟩 Y179C* deceased α-MSH potency, normal NDP-MSH binding (mouse) [48]
His6.54 (264)	-1.0	His6	🟩 H264A decreased α-MSH and NDP-MSH binding [36, 44, 54]
Leu6.55 (265)	-0.9	His6	🟥 L265A decreased small molecule binding; normal α-MSH and NDP-MSH binding [44]
Tyr6.58 (268)	-1.5/-0.9	Glu5/His6	🟩 Y268A normal NDP-MSH binding, decreased α-MSH binding; Y268F normal NDP-MSH and α-MSH binding [55]
Phe7.31 (280)	-0.8/-0.7	Glu5/His6	🟥 F280L normal α-MSH binding [58]
Met7.32 (281)	-0.3/-0.7	Ser4/Glu5	🟨 Unknown
Phe7.35 (284)	-1.4/-1.4	His6/Phe7	🟩 F284A decreased α-MSH binding [51, 55]; F284A normal NDP-MSH binding [55]
Tyr7.38 (287)	-0.2	His6	🟩 Y287A decreased α-MSH and NDP-NSH binding; Y287F = normal binding [55]

6.5 Methods

The structure of MC4R was modeled using multi-template comparative modeling with the RosettaCM application in the Rosetta Molecular Modeling Suite [52, 59, 60]. Since no MCR crystal structures are available, twenty GPCR crystal structures were selected as templates based on resolution and conformational variability. This provided the widest possible conformational space over which the MC4R sequence could be threaded. Details regarding all templates are outlined in table 6-3.

Table 6-3 Twenty templates were used to model the MC4R receptor. *constitutively active mutation in complex with carboxy terminus of transducin. **Partially active conformation. ***Active conformation in complex with terniary Gs. IA = inverse agonist; AG = agonist; AN = Antagonist.

PDB ID	Protein	Ligand	Resolution	Identity to MC4R
1u19	Bovine Rhodopsine	Retinal (IA)	2.2	13.5
2rh1	Human β2-AR	Carazolol (IA)	2.4	26.1
2vt4	Turkey β1-AR	Cyanopindolol (AN)	2.7	25.7
2x72	Bovine Rhodopsin	None*	3.0	13.4
2y03	Turkey β1-AR	Isoprenaline (AG)**	2.85	25.4
3eml	Human A2A	ZM241385 (AN)	2.6	27.7
3odu	Human CXCR4	IT1t (AN)	2.5	19.7
3pbl	Human D3	Eticlopride (AN)	2.89	23.9
3qak	Human A2A	UK-432097 (AG)**	2.71	27.9
3rze	Human H1	Doxepin (AN)	3.1	22.5
3sn6	Bovine β2-AR	P0G (AG)***	3.2	26.9
3uon	Human M2	3-quinuclidinyl-benzilate (AN)	3	23.1
3v2w	Human S1P1	ML5 (AN)	3.35	30.6
4daj	Rat M3	Tiotropium (IA)	3.4	26.3
4djh	Human κ-opioid	JDTic (AN)	2.9	18.3
4dkl	Mouse µ-opioid	BF0 (AN)	2.8	17.0
4ea3	Human N/OFQ opioid	C-24 (AN)	3.01	19.1
4ej4	Mouse δ-opioid	Naltrindole (AN)	2.7	18.1
4iar	Human 5HT-1B	Ergotamine (AG)**	2.7	25.2
4ib4	Human 5HT-2B	Ergotamine (AG)**	2.7	21.3

An initial sequence alignment of MC1R, MC2R, MC3R, MC4R, and MC5R was performed using ClustalW [61] and a profile alignment of the GPCR templates was performed using the structural-alignment tool MUSTANG [62]. A profile-profile alignment was performed in ClustalW between the MCR sequence alignment and the GPCR template structure alignment.

Manual adjustments were made to each template-target alignment to ensure that no gaps were present within transmembrane helices. In addition to the removal of inter-helical gaps, minor adjustments were made to ensure the alignment of highly conserved residues and structurally critical residues such as prolines and glycines when possible. Finally, terminal residues of MC4R were removed to ease the modeling process. Specifically, the first 27 residues of the N-terminal and the last 11 residues of the C-terminal were removed. The final MC4R models, therefore, represent residues 28 through 321 of the human MC4R sequence. The adjusted alignment used for threading can be found in the supplemental figure S6.1.

The partial-thread application in Rosetta was used to assign coordinates from each template onto the aligned residue in MC4R. This generated twenty incomplete MC4R models that vary based on the individual template-target alignments. Any residues in MC4R not assigned coordinates from a template were filled in during the RosettaCM hybridize phase.

Several predictions were used to guide the generation of comparative MC4R models. The transmembrane regions of MC4R were predicted using the online OCTOPUS prediction tool [63]. This tool uses artificial neural networks trained on known protein structures to predict stretches likely to lie within the membrane. Residues predicted to lie within the membrane are scored according to Rosetta's membrane scoring terms. Transmembrane segments were defined as the following: TM1 = 47-67, TM2 = 82-102, TM3 = 125-145, TM4 = 166-186, TM5 = 194-213, TM6 = 246-266, TM7 = 281-301.

Based on experimental data, two disulfide bonds are predicted within MC4R. These disulfide bonds include one between N-term C40 and the top of transmembrane helix C279 and an intra-loop disulfide bond in ECL3 between cysteine C271 and C277.

RosettaCM uses fragment-based ab initio folding to fill in missing densities and smooth the connections between pieces of different templates. Three and nine residue fragments were compiled using the truncated MC4R sequence with the online Robetta fragment library server [64].

The RosettaCM [52] protocol was used to generate full-atom models by combining low-energy fragments from different templates to generate the most energetically favorable structure based on a “hybrid” template. Twenty template-threaded partial models were supplied to RosettaCM's hybridize mover with equal weighting. Scoring terms included standard RosettaCM scoring terms previously describe with the addition of membrane-specific scoring terms to account for the transmembrane environment of MC4R.

In brief, RosettaCM consists of initialization and three main stages. Initialization aligns all twenty partial models in Cartesian space and fragments them based on secondary structure elements. Coordinate constraints are generated based on each template to preserve the tertiary structure of the templates. During the first stage, low resolution scoring terms are gradually phased in while fragments are randomly inserted to generate a complete low-resolution model. A total of 10,000 fragment insertions are attempted during this phase, divided equally between template-based and *de novo* fragments. The second stage consists of a two-step Monte Carlo search that randomly swaps fragments to efficiently explore conformational space beyond the starting template structures. Fragment swaps are performed for 1000 template-based segments and 500 de novo fragments. Selection of de novo fragments is biased towards regions of poor geometry and segment boundaries. This fragment replacement is followed by energy minimization and move evaluation. The final stage of RosettaCM involves a side-chain optimization and relaxation to replace all centroids with full-atom side-chains and arrive at an energetic minimum.

An additional relaxation step was performed following the standard RosettaCM protocol to enforce disulfide bonds and an implicit membrane potential [65].

MC4R comparative model selection

RosettaCM was used to build 20,000 multi-template comparative models. A model's pose energy score in Rosetta Energy Units (REUs) is an approximate description of the model's energy state. Low negative energy scores represent favorable energy poses with the lowest energy models expected to correlate with the native protein structure. The 5000 models with the lowest overall pose energy were clustered using the BCL clustering application [66]. A cut-off of four angstroms was selected to arrive at the largest clusters representing approximately 10% of the 5000 models each. The top five scoring models from the five largest clusters were selected for further analysis. These models were combined with the top twenty scoring models overall to produce a set of 33 models. Models that contained excessively malformed helix regions, loops that dipped far into and grossly obscured the central transmembrane pore, and models that failed to preserve the disulfide bonds were discarded. Additionally, residues used to guide α-MSH docking were inspected to ensure that at least three of the four residues were oriented into the binding pocket to facilitate docking. This left a final ensemble of 20 MC4R models. The overall comparative modeling strategy is outlined in figure 6.5.

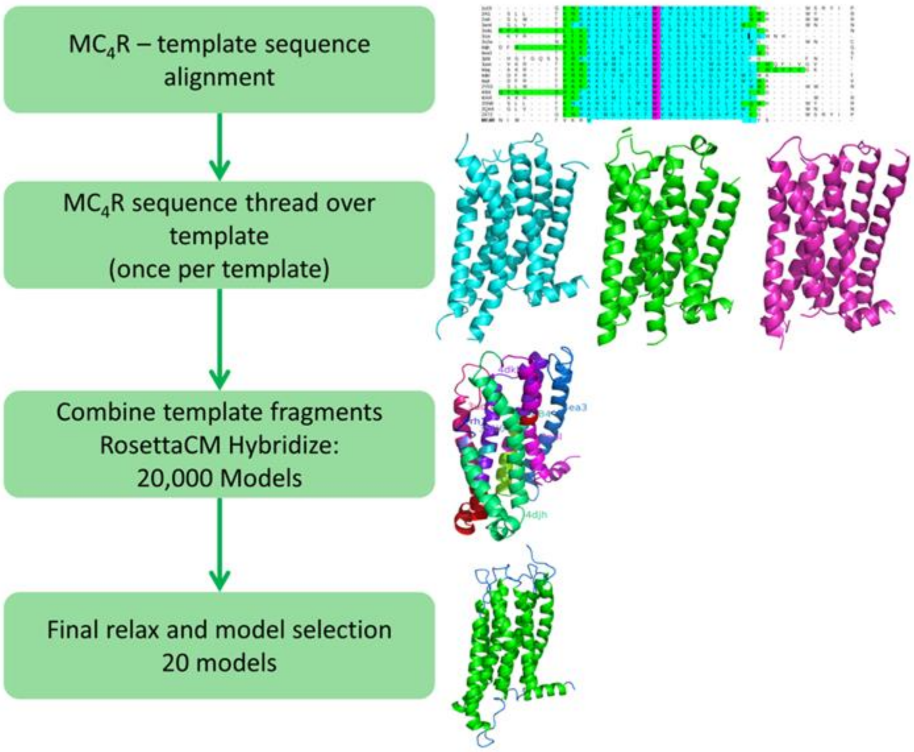

MC$_4$R – template sequence alignment
MC$_4$R sequence thread over template (once per template)
Combine template fragments RosettaCM Hybridize: 20,000 Models
Final relax and model selection 20 models
4djh

Figure 6.5 MC4R multi-template comparative modeling with RosettaCM.

Docking α-MSH to MC4R Comparative Models

α-MSH was docked into the twenty MC4R comparative models in two steps. The overall docking strategy is outlined in figure 6.6***A*** and the predicted active conformation of α-MSH is shown in figure 6.6***B***.

Step 1. Experimental evidence shows the tetrapeptide His-Phe-Arg-Trp is critical and sufficient for activation of MC4R [34]. Therefore, this tetrapeptide was isolated and docked to the MC4R comparative models. Rosetta's FlexPepDock [67] protocol was used to dock His-Phe-Arg-Trp. This protocol allows full flexibility and rigid body orientation for the peptide's backbone and side chain flexibility for both the peptide and receptor. Four constraints were used during this step to reflect previously determined contacts between alpha-MSH and MC4R. Atom pair constraints are detailed in table 6-4. All FlexPepDock models were refined with 200 cycles of full-model relaxation under high resolution atom pair constraints, disulfide bond constraints, and the implicit membrane potential defined by the membrane spanning predictions used during comparative modeling.

The tetrapeptide His-Phe-Arg-Trp was docked to each of the twenty MC4R comparative models 1,000 times, generating a total of 20,000 models. The top scoring models in terms of overall pose energy, interface energy between the tetrapeptide and MC4R, and compliance with atom pair constraints were selected for analysis. Forty-six models were visually inspected for structural inconsistencies such as unlikely loop or helix conformations and a final set of 16 models was carried to step two.

Step 2. The ECL regions of MC4R and the remaining nine residues (5 N-terminal residues SYSME and 4 C-terminal residues GKPV) of α-MSH were built and refined simultaneously to account for the high degree of flexibility expected in the receptor-peptide interface. An additional constraint was introduced during this step to account for the experimentally verified β-turn structure of α-MSH. Cα atoms of peptide residues Ser^{3} and Gly^{10} were kept within 6 Å based on experiments with cyclic α-MSH analogues. Loop closure was performed with Rosetta's cyclic coordinate descent (CCD) protocol [68]. This was followed by a kinematic loop modeling (KIC) [69] refinement. Finally, side chain optimization and full model relaxation within the implicit membrane potential produced 14,200 full-atom high resolution models.

Table 6-4 Five atom-pair constraints were used to guide docking. *Constraint within α-MSH to enforce active β-turn conformation

MC4R Residue	α-MSH Residue	Low Resolution Distance (Å)	Low Resolution Atoms	High Resolution Distance (Å)	High Resolution Atoms
E100	Arg8	8.0	Cβ-Cβ	6.0	Oε1-NH2
D122	Arg8	8.0	Cβ-Cβ	6.0	Oδ1-NH2
D126	Arg8	8.0	Cβ-Cβ	6.0	Oδ2-Cβ
P261	Phe7	8.0	Cβ-Cβ	6.0	CZ-CZ
	Ser3 + Gly4*	6.0	Cα-Cα	6.0	Cα-Cα

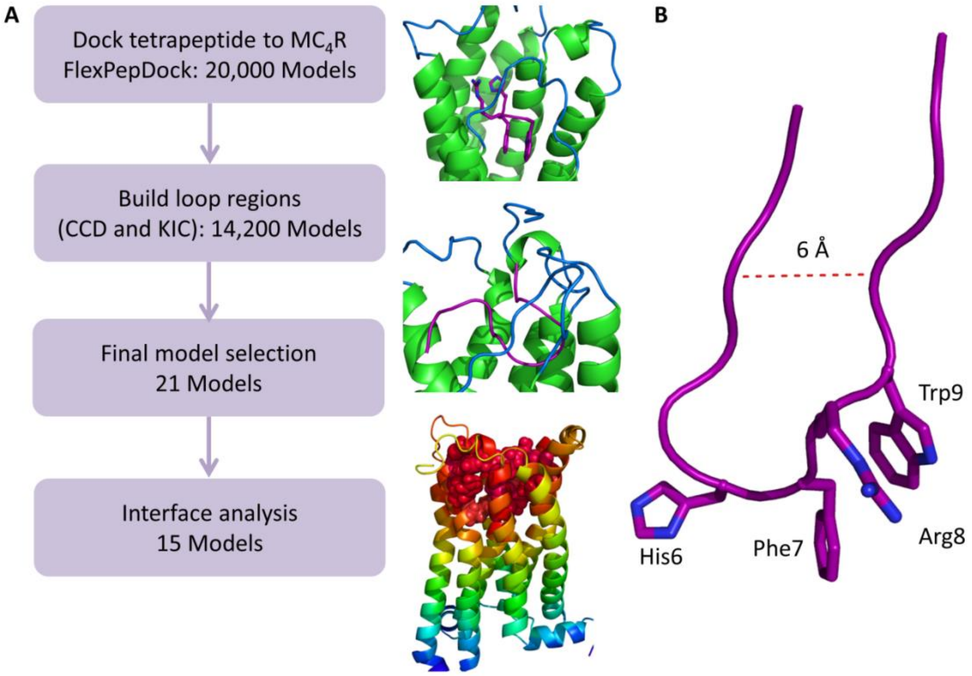


Figure 6.6 Docking α-MSH to MC4R comparative models. A) Two step docking protocol is outlined. B) Hairpin structure of α-MSH is shown. Critical tetrapeptide residues docked during the first round are indicated. A dashed line illustrates the 6 Å restraint used to enforce the β-turn feature of active α-MSH during the loop building step.

Final MC4R α-MSH complex model selection

All 14,200 high-resolution models were analyzed with Rosetta's InterfaceAnalyzer that scores several interface-specific metrics between the peptide ligand and receptor. Specifically, the binding energy is calculated as the change in energy when the peptide and receptor are separated. Models with poor overall pose energies, weak interface energies, or failure to retain atom-pair constraints used for docking were discarded, leaving a total of 412 models with good pose energy and atom-pair constraint compliance. These models were visually inspected for structurally inconsistencies and a large ensemble of 330 models remained. This large ensemble contained many highly similar models that could be traced to models stemming from common low-resolution intermediate CCD models. Models with highly similar conformations were removed to produce an ensemble representing the lowest scoring models covering the greatest conformational space with minimal redundancy. The final ensemble contained 21 MC4R + α-MSH models.

6.6 References

1. Girardet C & Butler AA (2014) Neural melanocortin receptors in obesity and related metabolic disorders. *Biochimica et biophysica acta* 1842(3):482-494.
2. Olney JJ, Navarro M, & Thiele TE (2014) Targeting central melanocortin receptors: a promising novel approach for treating alcohol abuse disorders. *Frontiers in neuroscience* 8:128.
3. Wikberg JE & Mutulis F (2008) Targeting melanocortin receptors: an approach to treat weight disorders and sexual dysfunction. *Nature reviews. Drug discovery* 7(4):307-323.
4. Adan RA*, et al.* (2006) The MC4 receptor and control of appetite. *Br J Pharmacol* 149(7):815-827.
5. Gantz I & Fong TM (2003) The melanocortin system. *American journal of physiology. Endocrinology and metabolism* 284(3):E468-474.
6. Jeong JK, Kim JG, & Lee BJ (2014) Participation of the central melanocortin system in metabolic regulation and energy homeostasis. *Cellular and molecular life sciences : CMLS* 71(19):3799-3809.
7. Elefteriou F*, et al.* (2005) Leptin regulation of bone resorption by the sympathetic nervous system and CART. *Nature* 434(7032):514-520.
8. Li SJ*, et al.* (1996) Melanocortin antagonists define two distinct pathways of cardiovascular control by alpha- and gamma-melanocyte-stimulating hormones. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 16(16):5182-5188.
9. Wessells H*, et al.* (2003) Ac-Nle-c[Asp-His-DPhe-Arg-Trp-Lys]-NH2 induces penile erection via brain and spinal melanocortin receptors. *Neuroscience* 118(3):755-762.
10. Baran K*, et al.* (2002) Chronic central melanocortin-4 receptor antagonism and central neuropeptide-Y infusion in rats produce increased adiposity by divergent pathways. *Diabetes* 51(1):152-158.
11. Joseph SA, Pilcher WH, & Bennett-Clarke C (1983) Immunocytochemical localization of ACTH perikarya in nucleus tractus solitarius: evidence for a second opiocortin neuronal system. *Neuroscience letters* 38(3):221-225.
12. Hadley ME & Haskell-Luevano C (1999) The proopiomelanocortin system. *Ann N Y Acad Sci* 885:1-21.
13. Smith AI & Funder JW (1988) Proopiomelanocortin processing in the pituitary, central nervous system, and peripheral tissues. *Endocrine reviews* 9(1):159-179.
14. Ersoy BA*, et al.* (2012) Mechanism of N-terminal modulation of activity at the melanocortin-4 receptor GPCR. *Nat Chem Biol* 8(8):725-730.
15. Valette M*, et al.* (2013) Eating behaviour in obese patients with melanocortin-4 receptor mutations: a literature review. *International journal of obesity* 37(8):1027-1035.
16. Yeo GS*, et al.* (1998) A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature genetics* 20(2):111-112.
17. Vaisse C, Clement K, Guy-Grand B, & Froguel P (1998) A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nature genetics* 20(2):113-114.
18. Santini F*, et al.* (2009) Melanocortin-4 receptor mutations in obesity. *Advances in clinical chemistry* 48:95-109.
19. Tao YX (2010) The melanocortin-4 receptor: physiology, pharmacology, and pathophysiology. *Endocrine reviews* 31(4):506-543.
20. Stutzmann F*, et al.* (2008) Prevalence of melanocortin-4 receptor deficiency in Europeans and their age-dependent penetrance in multigenerational pedigrees. *Diabetes* 57(9):2511-2518.
21. Tao YX (2009) Mutations in melanocortin-4 receptor and human obesity. *Progress in molecular biology and translational science* 88:173-204.

22. Xi B, Chandak GR, Shen Y, Wang Q, & Zhou D (2012) Association between common polymorphism near the MC4R gene and obesity risk: a systematic review and meta-analysis. *PloS one* 7(9):e45731.
23. Wolff GL, Roberts DW, & Mountjoy KG (1999) Physiological consequences of ectopic agouti gene expression: the yellow obese mouse syndrome. *Physiological genomics* 1(3):151-163.
24. Lu D*, et al.* (1994) Agouti protein is an antagonist of the melanocyte-stimulating-hormone receptor. *Nature* 371(6500):799-802.
25. Huszar D*, et al.* (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* 88(1):131-141.
26. Dores RM*, et al.* (2014) Molecular evolution of GPCRs: Melanocortin/melanocortin receptors. *Journal of molecular endocrinology* 52(3):T29-42.
27. Rediger A*, et al.* (2012) MC4R dimerization in the paraventricular nucleus and GHSR/MC3R heterodimerization in the arcuate nucleus: is there relevance for body weight regulation? *Neuroendocrinology* 95(4):277-288.
28. Ellacott KL & Cone RD (2006) The role of the central melanocortin system in the regulation of food intake and energy homeostasis: lessons from mouse models. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361(1471):1265-1274.
29. Fosgerau K, Raun K, Nilsson C, Dahl K, & Wulff BS (2014) Novel alpha-MSH analog causes weight loss in obese rats and minipigs and improves insulin sensitivity. *The Journal of endocrinology* 220(2):97-107.
30. Fani L, Bak S, Delhanty P, van Rossum EF, & van den Akker EL (2014) The melanocortin-4 receptor as target for obesity treatment: a systematic review of emerging pharmacological therapeutic options. *International journal of obesity* 38(2):163-169.
31. Eves PC & Haycock JW (2010) Melanocortin signalling mechanisms. *Advances in experimental medicine and biology* 681:19-28.
32. Pritchard LE, Turnbull AV, & White A (2002) Pro-opiomelanocortin processing in the hypothalamus: impact on melanocortin signalling and obesity. *The Journal of endocrinology* 172(3):411-421.
33. Haslach EM, Schaub JW, & Haskell-Luevano C (2009) Beta-turn secondary structure and melanocortin ligands. *Bioorganic & medicinal chemistry* 17(3):952-958.
34. Haskell-Luevano C, Holder JR, Monck EK, & Bauzo RM (2001) Characterization of melanocortin NDP-MSH agonist peptide fragments at the mouse central and peripheral melanocortin receptors. *Journal of medicinal chemistry* 44(13):2247-2252.
35. Tarnow P, Schoneberg T, Krude H, Gruters A, & Biebermann H (2003) Mutationally induced disulfide bond formation within the third extracellular loop causes melanocortin 4 receptor inactivation in patients with obesity. *The Journal of biological chemistry* 278(49):48666-48673.
36. Yang YK*, et al.* (2000) Molecular determinants of ligand binding to the human melanocortin-4 receptor. *Biochemistry* 39(48):14900-14911.
37. Sawyer TK, Hruby VJ, Darman PS, & Hadley ME (1982) [half-Cys4,half-Cys10]-alpha-Melanocyte-stimulating hormone: a cyclic alpha-melanotropin exhibiting superagonist biological activity. *Proceedings of the National Academy of Sciences of the United States of America* 79(6):1751-1755.
38. Knittel JJ, Sawyer TK, Hruby VJ, & Hadley ME (1983) Structure-activity studies of highly potent cyclic [Cys4,Cys10]Melanotropin analogues. *Journal of medicinal chemistry* 26(2):125-129.
39. Al-Obeidi F, Castrucci AM, Hadley ME, & Hruby VJ (1989) Potent and prolonged acting cyclic lactam analogues of alpha-melanotropin: design based on molecular dynamics. *Journal of medicinal chemistry* 32(12):2555-2561.

40. Chen M, Georgeson KE, Harmon CM, Haskell-Luevano C, & Yang Y (2006) Functional characterization of the modified melanocortin peptides responsible for ligand selectivity at the human melanocortin receptors. *Peptides* 27(11):2836-2845.

41. Fan W, Boston BA, Kesterson RA, Hruby VJ, & Cone RD (1997) Role of melanocortinergic neurons in feeding and the agouti obesity syndrome. *Nature* 385(6612):165-168.

42. Loos RJ (2011) The genetic epidemiology of melanocortin 4 receptor variants. *Eur J Pharmacol* 660(1):156-164.

43. Holst B & Schwartz TW (2003) Molecular mechanism of agonism and inverse agonism in the melanocortin receptors: Zn(2+) as a structural and functional probe. *Ann N Y Acad Sci* 994:1-11.

44. Pogozheva ID*, et al.* (2005) Interactions of human melanocortin 4 receptor with nonpeptide and peptide agonists. *Biochemistry* 44(34):11329-11341.

45. Nickolls SA*, et al.* (2003) Molecular determinants of melanocortin 4 receptor ligand binding and MC4/MC3 receptor selectivity. *The Journal of pharmacology and experimental therapeutics* 304(3):1217-1227.

46. Wilczynski A*, et al.* (2004) Identification of putative agouti-related protein(87-132)-melanocortin-4 receptor interactions by homology molecular modeling and validation using chimeric peptide ligands. *Journal of medicinal chemistry* 47(9):2194-2207.

47. Chai BX*, et al.* (2005) Receptor-antagonist interactions in the complexes of agouti and agouti-related protein with human melanocortin 1 and 4 receptors. *Biochemistry* 44(9):3418-3431.

48. Haskell-Luevano C, Cone RD, Monck EK, & Wan YP (2001) Structure activity studies of the melanocortin-4 receptor by in vitro mutagenesis: identification of agouti-related protein (AGRP), melanocortin agonist and synthetic peptide antagonist interaction determinants. *Biochemistry* 40(20):6164-6179.

49. Yang Y, Dickinson C, Haskell-Luevano C, & Gantz I (1997) Molecular basis for the interaction of [Nle4,D-Phe7]melanocyte stimulating hormone with the human melanocortin-1 receptor. *The Journal of biological chemistry* 272(37):23000-23010.

50. Yang Y (2011) Structure, function and regulation of the melanocortin receptors. *Eur J Pharmacol* 660(1):125-130.

51. Fleck BA*, et al.* (2005) Molecular interactions of nonpeptide agonists and antagonists with the melanocortin-4 receptor. *Biochemistry* 44(44):14494-14508.

52. Song Y*, et al.* (2013) High-resolution comparative modeling with RosettaCM. *Structure* 21(10):1735-1742.

53. Lagerstrom MC*, et al.* (2003) High affinity agonistic metal ion binding sites within the melanocortin 4 receptor illustrate conformational change of transmembrane region 3. *The Journal of biological chemistry* 278(51):51521-51526.

54. Chen M*, et al.* (2007) Contribution of the conserved amino acids of the melanocortin-4 receptor in [corrected] [Nle4,D-Phe7]-alpha-melanocyte-stimulating [corrected] hormone binding and signaling. *The Journal of biological chemistry* 282(30):21712-21719.

55. Hogan K*, et al.* (2006) Mapping the Binding Site of Melanocortin 4 Receptor Agonists: A Hydrophobic Pocket Formed by I3.28(125), I3.32(129), and I7.42(291) Is Critical for Receptor Activation. *Journal of medicinal chemistry* 49(3):911-922.

56. Chen M, Celik A, Georgeson KE, Harmon CM, & Yang Y (2006) Molecular basis of melanocortin-4 receptor for AGRP inverse agonism. *Regulatory Peptides* 136(1–3):40-49.

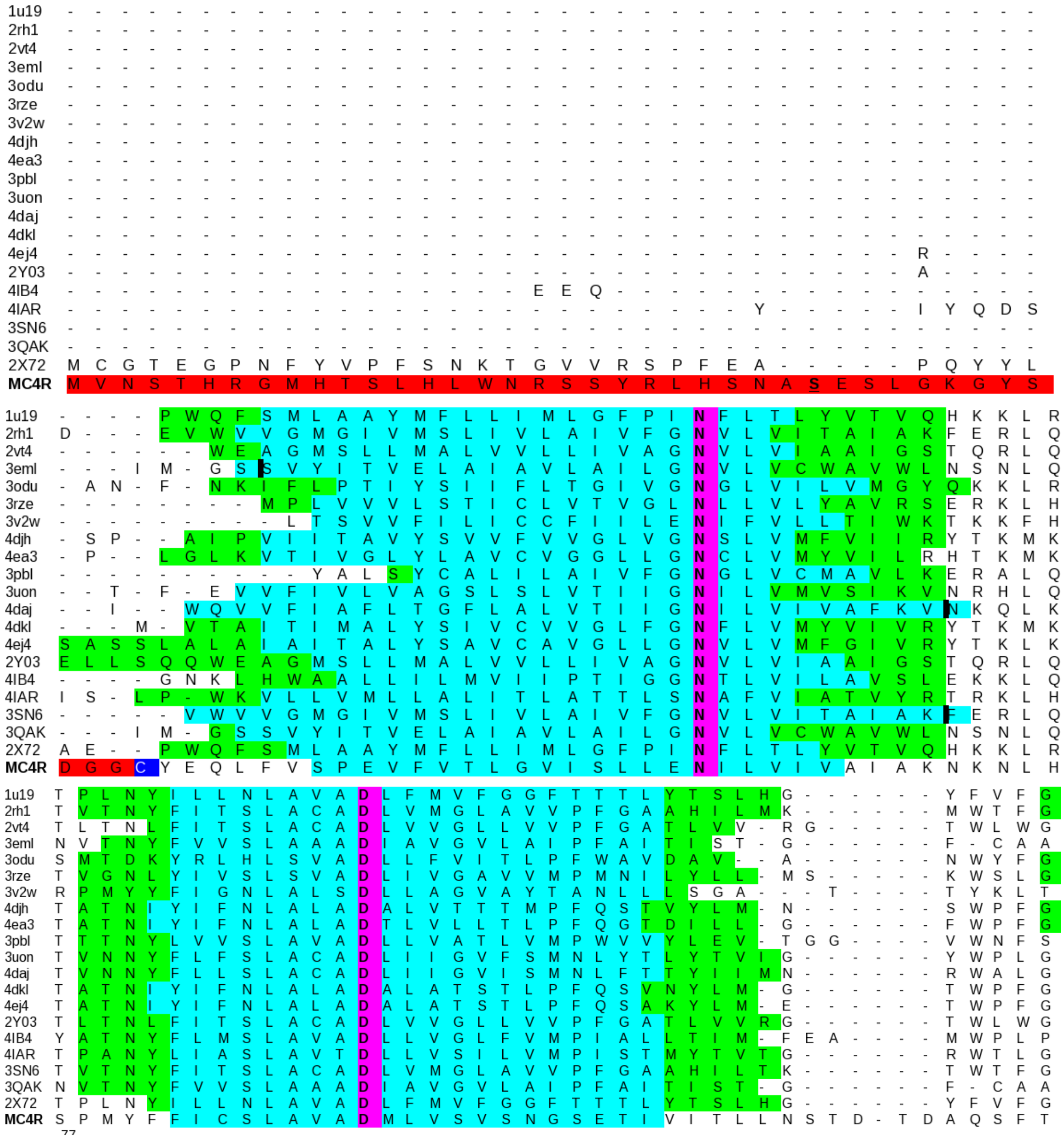
57. Fleck BA, Ling N, & Chen C (2007) Substituted NDP-MSH peptides paired with mutant melanocortin-4 receptors demonstrate the role of transmembrane 6 in receptor activation. *Biochemistry* 46(37):10473-10483.

58. Wang ZQ & Tao YX (2011) Functional studies on twenty novel naturally occurring melanocortin-4 receptor mutations. *Biochimica et biophysica acta* 1812(9):1190-1199.

59. Rohl CA, Strauss CE, Misura KM, & Baker D (2004) Protein structure prediction using Rosetta. *Methods in enzymology* 383:66-93.
60. Leaver-Fay A*, et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* 487:545-574.
61. Larkin MA*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.
62. Konagurthu AS, Whisstock JC, Stuckey PJ, & Lesk AM (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins* 64(3):559-574.
63. Viklund H & Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24(15):1662-1668.
64. Kim DE, Chivian D, & Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic acids research* 32(Web Server issue):W526-531.
65. Yarov-Yarovoy V, Schonbrun J, & Baker D (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins* 62(4):1010-1025.
66. Alexander N, Woetzel N, & Meiler J (2011) Bcl::Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pp 13-18.
67. Raveh B, London N, Zimmerman L, & Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PloS one* 6(4):e18934.
68. Canutescu AA & Dunbrack RL, Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein science : a publication of the Protein Society* 12(5):963-972.
69. Stein A & Kortemme T (2013) Improvements to robotics-inspired conformational sampling in rosetta. *PloS one* 8(5):e63090.

6.7 Supplementary Information

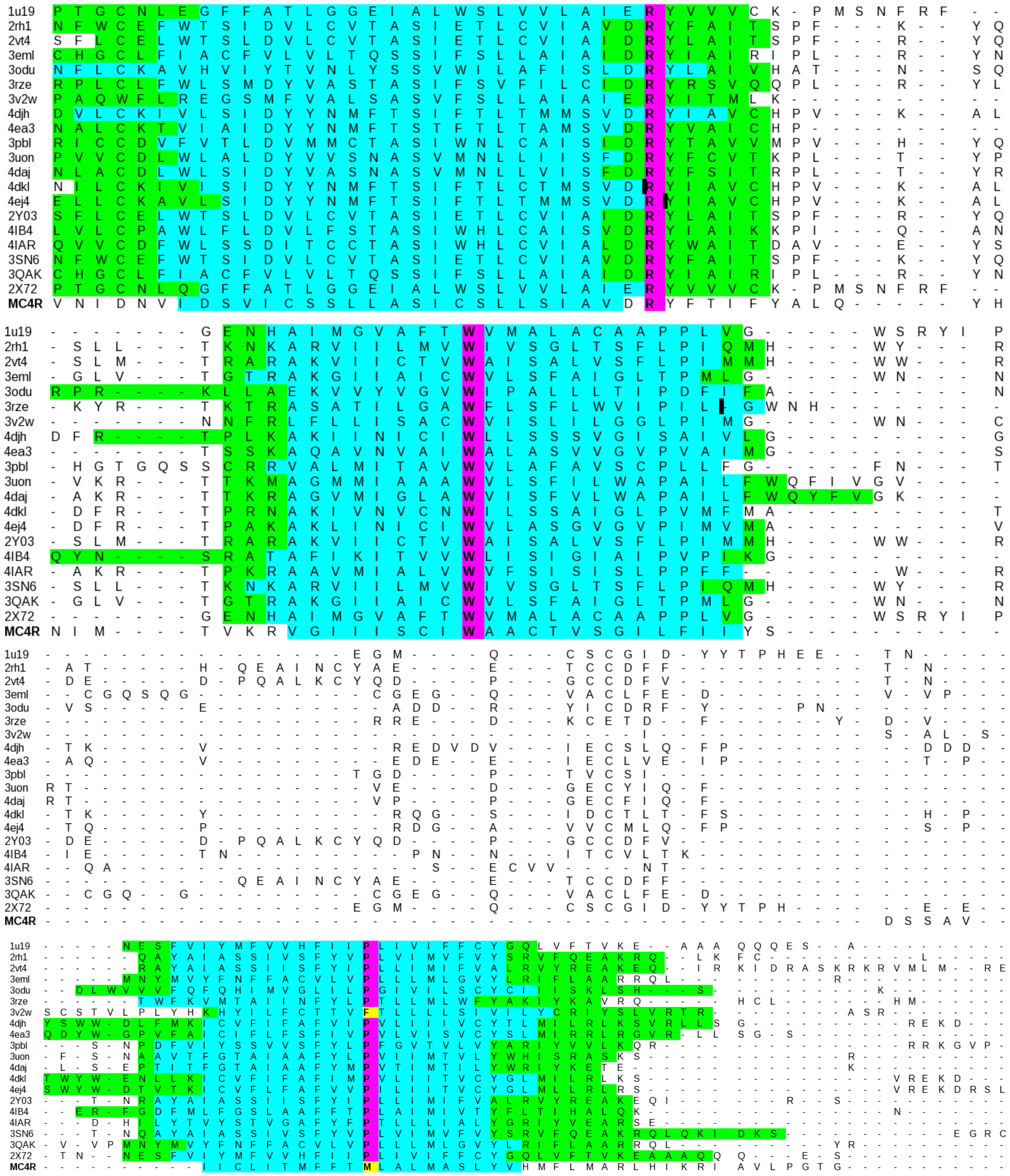
Figure S6.1 adjusted alignment used for template threading.



```
1u19  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
2rh1  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
2vt4  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3eml  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3odu  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3rze  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3v2w  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4djh  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4ea3  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3pbl  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3uon  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4daj  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4dkl  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
4ej4  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - R - - - -
2Y03  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - A - - - -
4IB4  - - - - - - - - - - - - - - - - - E E Q - - - - - - - - - - - - - - - -
4IAR  - - - - - - - - - - - - - - - - - - - - - - - - - Y - - - - - I Y Q D S
3SN6  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
3QAK  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
2X72  M C G T E G P N F Y V P F S N K T G V V R S P F E A - - - - - P Q Y Y L
MC4R  M V N S T H R G M H T S L H L W N R S S Y R L H S N A S E S L G K G Y S

1u19  - - - - P W Q F S M L A A Y M F L L I M L G F P I N F L T L Y V T V Q H K K L R
2rh1  D - - - E V W V V G M G I V M S L I V L A I V F G N V L V I T A I A K F E R L Q
2vt4  - - - - - - W E A G M S L L M A L V V L L I V A G N V L V I A A I G S T Q R L Q
3eml  - - - I M - G S S V Y I T V E L A I A V L A I L G N V L V C W A V W L N S N L Q
3odu  - A N - F - N K I F L P T I Y S I I F L T G I V G N G L V I L V M G Y Q K K L R
3rze  - - - - - - - - M P L V V V L S T I C L V T V G L N L L V L Y A V R S E R K L H
3v2w  - - - - - - - - - L T S V V F I L I C C F I I L E N I F V L L T I W K T K K F H
4djh  - S P - - A I P V I I T A V Y S V V F V V G L V G N S L V M F V I I R Y T K M K
4ea3  - P - - L G L K V T I V G L Y L A V C V G G L L G N C L V M Y V I L R H T K M K
3pbl  - - - - - - - - - - Y A L S Y C A L I L A I V F G N G L V C M A V L K E R A L Q
3uon  - - T - F - E V V F I V L V A G S L S L V T I I G N I L V M V S I K V N R H L Q
4daj  - - I - - W Q V V F I A F L T G F L A L V T I I G N I L V I V A F K V N K Q L K
4dkl  - - - M - V T A I T I M A L Y S I V C V V G L F G N F L V M Y V I V R Y T K M K
4ej4  S A S S L A L A I A I T A L Y S A V C A V G L L G N V L V M F G I V R Y T K L K
2Y03  E L L S Q Q W E A G M S L L M A L V V L L I V A G N V L V I A A I G S T Q R L Q
4IB4  - - - - G N K L H W A A L L I L M V I I P T I G G N T L V I L A V S L E K K L Q
4IAR  I S - L P - W K V L L V M L L A L I T L A T T L S N A F V I A T V Y R T R K L H
3SN6  - - - - - V W V V G M G I V M S L I V L A I V F G N V L V I T A I A K F E R L Q
3QAK  - - - I M - G S S V Y I T V E L A I A V L A I L G N V L V C W A V W L N S N L Q
2X72  A E - - P W Q F S M L A A Y M F L L I M L G F P I N F L T L Y V T V Q H K K L R
MC4R  D G G C Y E Q L F V S P E V F V T L G V I S L L E N I L V I V A I A K N K N L H

1u19  T P L N Y I L L N L A V A D L F M V F G G F T T T L Y T S L H G - - - - - - Y F V F G
2rh1  T V T N Y F I T S L A C A D L V M G L A V V P F G A A H I L M K - - - - - - M W T F G
2vt4  T L T N L F I T S L A C A D L V V G L L V V P F G A T L V V - R G - - - - - T W L W G
3eml  N V T N Y F V V S L A A A D I A V G V L A I P F A I T I S T - G - - - - - - F - C A A
3odu  S M T D K Y R L H L S V A D L L F V I T L P F W A V D A V - - A - - - - - - N W Y F G
3rze  T V G N L Y I V S L S V A D L I V G A V V M P M N I L Y L L - M S - - - - - K W S L G
3v2w  R P M Y Y F I G N L A L S D L L A G V A Y T A N L L L S G A - - - T - - - - T Y K L T
4djh  T A T N I Y I F N L A L A D A L V T T T M P F Q S T V Y L M - N - - - - - - S W P F G
4ea3  T A T N I Y I F N L A L A D T L V L L T L P F Q G T D I L L - G - - - - - - F W P F G
3pbl  T T T N Y L V V S L A V A D L L V A T L V M P W V V Y L E V - T G G - - - - V W N F S
3uon  T V N N Y F L F S L A C A D L I I G V F S M N L Y T L Y T V I G - - - - - - Y W P L G
4daj  T V N N Y F L L S L A C A D L I I G V I S M N L F T T Y I I M N - - - - - - R W A L G
4dkl  T A T N I Y I F N L A L A D A L A T S T L P F Q S V N Y L M - G - - - - - - T W P F G
4ej4  T A T N I Y I F N L A L A D A L A T S T L P F Q S A K Y L M - E - - - - - - T W P F G
2Y03  T L T N L F I T S L A C A D L V V G L L V V P F G A T L V V R G - - - - - - T W L W G
4IB4  Y A T N Y F L M S L A V A D L L V G L F V M P I A L L T I M - F E A - - - - M W P L P
4IAR  T P A N Y L I A S L A V T D L L V S I L V M P I S T M Y T V T G - - - - - - R W T L G
3SN6  T V T N Y F I T S L A C A D L V M G L A V V P F G A A H I L T K - - - - - - T W T F G
3QAK  N V T N Y F V V S L A A A D I A V G V L A I P F A I T I S T - G - - - - - - F - C A A
2X72  T P L N Y I L L N L A V A D L F M V F G G F T T T L Y T S L H G - - - - - - Y F V F G
MC4R  S P M Y F F I C S L A V A D M L V S V S N G S E T I V I T L L N S T D - T D A Q S F T
       77
```

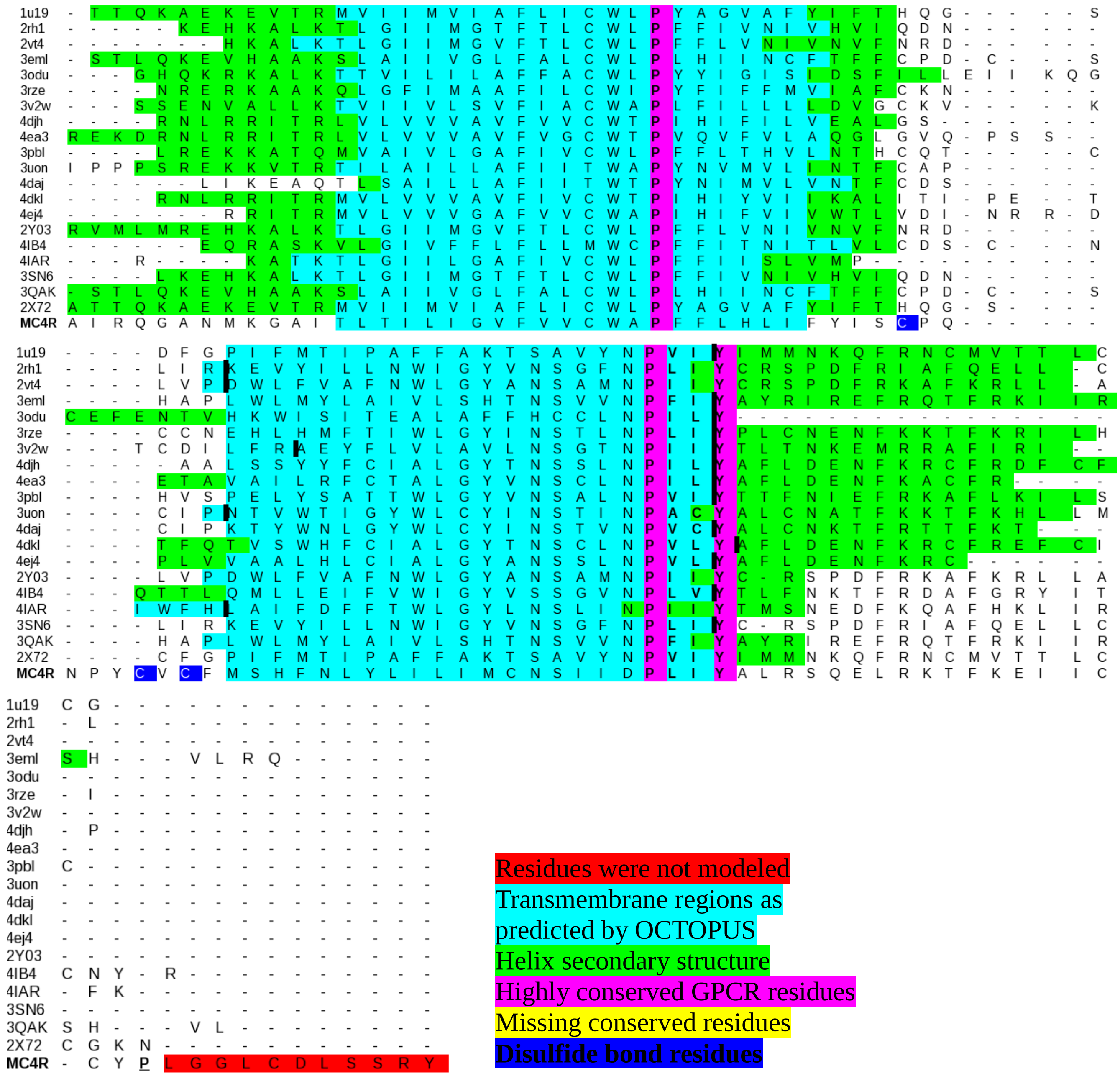


```
1u19  P T G C N L E G F F A T L G G E I A L W S L V V L A I E R Y V V V C K - P M S N F R F - -
2rh1  N F W C E F W T S I D V L C V T A S I E T L C V I A V D R Y F A I T S P F - - - K - - Y Q
2vt4  S F L C E L W T S L D V L C V T A S I E T L C V I A I D R Y L A I T S P F - - - R - - Y Q
3eml  C H G C L F I A C F V L V L T Q S S I F S L L A I A I D R Y I A I R I P L - - - R - - Y N
3odu  N F L C K A V H V I Y T V N L Y S S V W I L A F I S L D R Y L A I V H A T - - - N - - S Q
3rze  R P L C L F W L S M D Y V A S T A S I F S V F I L C I D R Y R S V Q Q P L - - - R - - Y L
3v2w  P A Q W F L R E G S M F V A L S A S V F S L L A I A I E R Y I T M L K - - - - - - - - - -
4djh  D V L C K I V L S I D Y Y N M F T S I F T L T M M S V D R Y I A V C H P V - - - K - - A L
4ea3  N A L C K T V I A I D Y Y N M F T S T F T L T A M S V D R Y V A I C H P - - - - - - - - -
3pbl  R I C C D V F V T L D V M M C T A S I W N L C A I S I D R Y T A V V M P V - - - H - - Y Q
3uon  P V V C D L W L A L D Y V V S N A S V M N L L I I S F D R Y F C V T K P L - - - T - - Y P
4daj  N L A C D L W L S I D Y V A S N A S V M N L L V I S F D R Y F S I T R P L - - - T - - Y R
4dkl  N I L C K I V I S I D Y Y N M F T S I F T L C T M S V D R Y I A V C H P V - - - K - - A L
4ej4  E L L C K A V L S I D Y Y N M F T S I F T L T M M S V D R Y I A V C H P V - - - K - - A L
2Y03  S F L C E L W T S L D V L C V T A S I E T L C V I A I D R Y L A I T S P F - - - R - - Y Q
4IB4  L V L C P A W L F L D V L F S T A S I W H L C A I S V D R Y I A I K K P I - - - Q - - A N
4IAR  Q V V C D F W L S S D I T C C T A S I W H L C V I A L D R Y W A I T D A V - - - E - - Y S
3SN6  N F W C E F W T S I D V L C V T A S I E T L C V I A V D R Y F A I T S P F - - - K - - Y Q
3QAK  C H G C L F I A C F V L V L T Q S S I F S L L A I A I D R Y I A I R I P L - - - R - - Y N
2X72  P T G C N L Q G F F A T L G G E I A L W S L V V L A I E R Y V V V C K - P M S N F R F - -
MC4R  V N I D N V I D S V I C S S L L A S I C S L L S I A V D R Y F T I F Y A L Q - - - - - - Y H

1u19  - - - - - - - - G E N H A I M G V A F T W V M A L A C A A P P L V G - - - - - - W S R Y I   P
2rh1  - S L L - - - - T K N K A R V I I L M V W I V S G L T S F L P I Q M H - - - - - W Y - - - - R
2vt4  - S L M - - - - T R A R A K V I I C T V W A I S A L V S F L P I M M H - - - - - W W - - - - R
3eml  - G L V - - - - T G T R A K G I I A I C W V L S F A I G L T P M L G - - - - - - W N - - - - N
3odu  R P R - - - - - K L L A E K V V Y V G V W I P A L L L T I P D F I F A - - - - - - - - - - - N
3rze  - K Y R - - - - T K T R A S A T I L G A W F L S F L W V I P I L - G W N H - - - - - - - - -
3v2w  - - - - - - - - N N F R L F L L I S A C W V I S L I L G G L P I M G - - - - - - W N - - - - C
4djh  D F R - - - - - T P L K A K I I N I C I W L L S S S V G I S A I V L G - - - - - - - - - - G
4ea3  - - - - - - - - T S S K A Q A V N V A I W A L A S V V G V P V A I M G - - - - - - - - - - S
3pbl  - H G T G Q S S C R R V A L M I T A V W V L A F A V S C P L L F G - - - - - - F N - - - - T
3uon  - V K R - - - - T T K M A G M M I A A A W V L S F I L W A P A I L F W Q F I V G V - - - - -
4daj  - A K R - - - - T T K R A G V M I G L A W V I S F V L W A P A I L F W Q Y F V G K - - - - -
4dkl  - D F R - - - - T P R N A K I V N V C N W I L S S A I G L P V M F M A - - - - - - - - - - T
4ej4  - D F R - - - - T P A K A K L I N I C I W V L A S G V G V P I M V M A - - - - - - - - - - V
2Y03  - S L M - - - - T R A R A K V I I C T V W A I S A L V S F L P I M M H - - - - - W W - - - - R
4IB4  Q Y N - - - - - S R A T A F I K I T V V W L I S I G I A I P V P I K G - - - - - - - - - - -
4IAR  - A K R - - - - T P K R A A V M I A L V W V F S I S I S L P P F F - - - - - - - W - - - - R
3SN6  - S L L - - - - T K N K A R V I I L M V W I V S G L T S F L P I Q M H - - - - - W Y - - - - R
3QAK  - G L V - - - - T G T R A K G I I A I C W V L S F A I G L T P M L G - - - - - - W N - - - - N
2X72  - - - - - - - - G E N H A I M G V A F T W V M A L A C A A P P L V G - - - - - - W S R Y I   P
MC4R  N I M - - - - - T V K R V G I I I S C I W A A C T V S G I L F I I Y S - - - - - - - - - - -

1u19  - - - - - - - - - - - - - - - - E G M - - - - Q - - - C S C G I D - Y Y T P H E E - -   T N - - - - -
2rh1  - A T - - - - - H - Q E A I N C Y A E - - - - E - - - T C C D F F - - - - - - - - - -   T - N - - - -
2vt4  - D E - - - - - D - P Q A L K C Y Q D - - - - P - - - G C C D F V - - - - - - - - - -   T - N - - - -
3eml  - - C G Q S Q G - - - - - - - - - C G E G - - Q - - - V A C L F E - D - - - - - - - -   V - V P - - -
3odu  - V S - - - - - E - - - - - - - - - A D D - - R - - - Y I C D R F - Y - - - - P N - -   - - - - - - -
3rze  - - - - - - - - - - - - - - - - - R R E - - - D - - - K C E T D - - F - - - - - - Y -   D - V - - - -
3v2w  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - I - - - - - - - - - - -   S - A L - S -
4djh  - T K - - - - - V - - - - - - - - - R E D V D V - - - I E C S L Q - F P - - - - - - -   - - D D D - -
4ea3  - A Q - - - - - V - - - - - - - - - E D E - - E - - - I E C L V E - I P - - - - - - -   - - T - P - -
3pbl  - - - - - - - - - - - - - - - - T G D - - - - P - - - T V C S I - - - - - - - - - - -   - - - - - - -
3uon  R T - - - - - - - - - - - - - - - V E - - - - D - - - G E C Y I Q - F - - - - - - - -   - - - - - - -
4daj  R T - - - - - - - - - - - - - - - V P - - - - P - - - G E C F I Q - F - - - - - - - -   - - - - - - -
4dkl  - T K - - - - - Y - - - - - - - - - R Q G - - S - - - I D C T L T - F S - - - - - - -   - - H - P - -
4ej4  - T Q - - - - - P - - - - - - - - - R D G - - A - - - V V C M L Q - F P - - - - - - -   - - S - P - -
2Y03  - D E - - - - - D - P Q A L K C Y Q D - - - - P - - - G C C D F V - - - - - - - - - -   - - - - - - -
4IB4  - I E - - - - - T N - - - - - - - - - P N - - N - - - I T C V L T K - - - - - - - - -   - - - - - - -
4IAR  - - Q A - - - - - - - - - - - - - - - - S - - E C V V - - - - N T - - - - - - - - - -   - - - - - - -
3SN6  - - - - - - - - - - Q E A I N C Y A E - - - - E - - - T C C D F F - - - - - - - - - -   - - - - - - -
3QAK  - - C G Q - - G - - - - - - - - - C G E G - - Q - - - V A C L F E - D - - - - - - - -   - - - - - - -
2X72  - - - - - - - - - - - - - - - - E G M - - - - Q - - - C S C G I D - Y Y T P H - - - -   - - E - E - -
MC4R  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -   D S S A V - -

1u19  - - - - - N E S F V I Y M F V V H F I I P L I V I F F C Y G Q L V F T V K E - - A A A   Q Q Q E S - - A - - - - - - - - - -
2rh1  - - - - - - Q A Y A I A S S I V S F Y V P L V I M V F V Y S R V F Q E A K R Q - - L K   F C - - - - - - - - - - L - - - - - -
2vt4  - - - - - - R A Y A I A S S I I S F Y I P L L I M I F V A L R V Y R E A K E Q - - I R   K I D R A S K R K R V M L M - - R E
3eml  - - - - - M N Y M V Y F N F F A C V L V P L L L M L G V Y L R I F L A A R R Q L - - -   - - - - - - - R - - - - - - - - - - -
3odu  - - D L W V V V F Q F Q H I M V G L I L P G I V I L S C Y C I I S K L S H - - - S -   - - - - - - - - K - - - - - - - - - -
3rze  - - - - - - T W F K V M T A I I N F Y L P T L L M L W F Y A K I Y K A V R Q - - - - - -   H C L - - - - - - - H M - - - - - -
3v2w  S C S T V L P L Y H K H Y I L F C T T V F T L L L L S I V I L Y C R I Y S L V R T R -   - - - - - - - A S R - - - - - - - -
4djh  Y S W W - D L F M K I C V F I F A F V I P V L I I I V C Y T L M I L R L K S V R L L S   G - - - - - - - - - - R E K D - - -
4ea3  Q D Y W - G P V F A I C I F L F S F I V P V L V I S V C Y S L M I R R L R G V R - L L   S G - S - - - - - - - - - - - - - - -
3pbl  - - - S - N P D F V I Y S S V V S F Y L P F G V T V L V Y A R I Y V V L K Q R - - - - -   - - - - - - - - - R R K G V P -
3uon  - F - S - N A A V T F G T A I A A F Y L P V I I M T V L Y W H I S R A S K S - - - - - -   - - - - - - R - - - - - - - - - -
4daj  - L - S - E P T I T F G T A I A A F Y M P V T I M T I L Y W R I Y K E T E - - - - - -   - - - - - - K - - - - - - - - - -
4dkl  T W Y W - E N L L K I C V F I F A F I M P V L I I T V C Y G L M I L R L K S - - - - -   - - - - - - - - - V R E K D - - -
4ej4  S W Y W - D T V T K I C V F L F A F V V P I L I I T V C Y G L M L L R L R S - - - - -   - - - - - - - - - V R E K D R S L
2Y03  - - - T - N R A Y A I A S S I I S F Y I P L L I M I F V A L R V Y R E A K E Q I - - -   - R - - S - - - - - - - - - - - -
4IB4  - - E R - F G D F M L F G S L A A F F T P L A I M I V T Y F L T I H A L Q K - - - - -   - - - - - - - - - N - - - - - - -
4IAR  - - - D - H I L Y T V Y S T V G A F Y F P T L L L I A L Y G R I Y V E A R S E - - - -   - - - - - - - - - - - - - - - -
3SN6  - - - T - N Q A Y A I A S S I V S F Y V P L V I M V F V Y S R V F Q E A K R Q L Q K I   D K S - - - - - - - - - - E G R C
3QAK  - V - V P M N Y M V Y F N F F A C V L V P L L L M L G V Y L R I F L A A R R Q L - - -   - - - - - - - Y R - - - - - - - - - -
2X72  - T N - - N E S F V I Y M F V V H F I I P L I V I F F C Y G Q L V F T V K E A A A Q Q   Q - - - E - S - - - - - - - - - -
MC4R  - - - - - - - - - - I I C L I T M F F T M L A L M A S L Y V H M F L M A R L H I K R I   A V L P G T G - - - - - - - - - -
```



1u19	- T T Q K A E K E V T R M V I I M V I A F L I C W L P Y A G V A F Y I F T H Q G - - - - - S
2rh1	- - - - - K E H K A L K T L G I I M G T F T L C W L P F F I V N I V H V I Q D N - - - - - -
2vt4	- - - - - - - H K A L K T L G I I M G V F T L C W L P F F L V N I V N V F N R D - - - - - -
3eml	- S T L Q K E V H A A K S L A I I V G L F A L C W L P L H I I N C F T F F C P D - C - - - S
3odu	- - - G H Q K R K A L K T T V I L I L A F F A C W L P Y Y I G I S I D S F I L L E I I K Q G
3rze	- - - - N R E R K A A K Q L G F I M A A F I L C W I P Y F I F F M V I A F C K N - - - - - -
3v2w	- - - S S E N V A L L K T V I I V L S V F I A C W A P L F I L L L L D V G C K V - - - - - K
4djh	- - - - R N L R R I T R L V L V V V A V F V V C W T P I H I F I L V E A L G S - - - - - - -
4ea3	R E K D R N L R R I T R L V L V V V A V F V G C W T P V Q V F V L A Q G L G V Q - P S S - -
3pbl	- - - - L R E K K A T Q M V A I V L G A F I V C W L P F F L T H V L N T H C Q T - - - - - C
3uon	I P P P S R E K K V T R T I L A I L L A F I I T W A P Y N V M V L I N T F C A P - - - - - -
4daj	- - - - - - L I K E A Q T L S A I L L A F I I T W T P Y N I M V L V N T F C D S - - - - - -
4dkl	- - - - R N L R R I T R M V L V V V A V F I V C W T P I H I Y V I I K A L I T I - P E - - T
4ej4	- - - - - - - R R I T R M V L V V V G A F V V C W A P I H I F V I V W T L V D I - N R R - D
2Y03	R V M L M R E H K A L K T L G I I M G V F T L C W L P F F L V N I V N V F N R D - - - - - -
4IB4	- - - - - - E Q R A S K V L G I V F F L F L L M W C P F F I T N I T L V L C D S - C - - - N
4IAR	- - - R - - - - K A T K T L G I I L G A F I V C W L P F F I I S L V M P - - - - - - - - - -
3SN6	- - - - L K E H K A L K T L G I I M G T F T L C W L P F F I V N I V H V I Q D N - - - - - -
3QAK	- S T L Q K E V H A A K S L A I I V G L F A L C W L P L H I I N C F T F F C P D - C - - - S
2X72	A T T Q K A E K E V T R M V I I M V I A F L I C W L P Y A G V A F Y I F T H Q G - S - - - -
MC4R	A I R Q G A N M K G A I T L T I L I G V F V V C W A P F F L H L I F Y I S C P Q - - - - - -

1u19	- - - - D F G P I F M T I P A F F A K T S A V Y N P V I Y I M M N K Q F R N C M V T T L C
2rh1	- - - - L I R K E V Y I L L N W I G Y V N S G F N P L I Y C R S P D F R I A F Q E L L - C
2vt4	- - - - L V P D W L F V A F N W L G Y A N S A M N P I I Y C R S P D F R K A F K R L L - A
3eml	- - - - H A P L W L M Y L A I V L S H T N S V V N P F I Y A Y R I R E F R Q T F R K I I R
3odu	C E F E N T V H K W I S I T E A L A F F H C C L N P I L Y - - - - - - - - - - - - - - - -
3rze	- - - - C C N E H L H M F T I W L G Y I N S T L N P L I Y P L C N E N F K K T F K R I L H
3v2w	- - - T C D I L F R A E Y F L V L A V L N S G T N P I I Y T L T N K E M R R A F I R I - -
4djh	- - - - - A A L S S Y Y F C I A L G Y T N S S L N P I L Y A F L D E N F K R C F R D F C F
4ea3	- - - - E T A V A I L R F C T A L G Y V N S C L N P I L Y A F L D E N F K A C F R - - - -
3pbl	- - - - H V S P E L Y S A T T W L G Y V N S A L N P V I Y T T F N I E F R K A F L K I L S
3uon	- - - - C I P N T V W T I G Y W L C Y I N S T I N P A C Y A L C N A T F K K T F K H L L M
4daj	- - - - C I P K T Y W N L G Y W L C Y I N S T V N P V C Y A L C N K T F R T T F K T - - -
4dkl	- - - - T F Q T V S W H F C I A L G Y T N S C L N P V L Y A F L D E N F K R C F R E F C I
4ej4	- - - - P L V V A A L H L C I A L G Y A N S S L N P V L Y A F L D E N F K R C - - - - - -
2Y03	- - - - L V P D W L F V A F N W L G Y A N S A M N P I I Y C - R S P D F R K A F K R L L A
4IB4	- - - Q T T L Q M L L E I F V W I G Y V S S G V N P L V Y T L F N K T F R D A F G R Y I T
4IAR	- - - I W F H L A I F D F F T W L G Y L N S L I N P I I Y T M S N E D F K Q A F H K L I R
3SN6	- - - - L I R K E V Y I L L N W I G Y V N S G F N P L I Y C - R S P D F R I A F Q E L L C
3QAK	- - - - H A P L W L M Y L A I V L S H T N S V V N P F I Y A Y R I R E F R Q T F R K I I R
2X72	- - - - C F G P I F M T I P A F F A K T S A V Y N P V I Y I M M N K Q F R N C M V T T L C
MC4R	N P Y C V C F M S H F N L Y L I L I M C N S I I D P L I Y A L R S Q E L R K T F K E I I C

1u19	C G - - - - - - - - - - - - -
2rh1	- L - - - - - - - - - - - - -
2vt4	- - - - - - - - - - - - - - -
3eml	S H - - - V L R Q - - - - - -
3odu	- - - - - - - - - - - - - - -
3rze	- I - - - - - - - - - - - - -
3v2w	- - - - - - - - - - - - - - -
4djh	- P - - - - - - - - - - - - -
4ea3	- - - - - - - - - - - - - - -
3pbl	C - - - - - - - - - - - - - -
3uon	- - - - - - - - - - - - - - -
4daj	- - - - - - - - - - - - - - -
4dkl	- - - - - - - - - - - - - - -
4ej4	- - - - - - - - - - - - - - -
2Y03	- - - - - - - - - - - - - - -
4IB4	C N Y - R - - - - - - - - - -
4IAR	- F K - - - - - - - - - - - -
3SN6	- - - - - - - - - - - - - - -
3QAK	S H - - - V L - - - - - - - -
2X72	C G K N - - - - - - - - - - -
MC4R	- C Y P L G G L C D L S S R Y

Residues were not modeled
Transmembrane regions as predicted by OCTOPUS
Helix secondary structure
Highly conserved GPCR residues
Missing conserved residues
Disulfide bond residues

Protocol Capture

The following information includes all settings and command lines used for comparative modeling the MC4 receptor and docking α-MSH. The Rosetta software suite is publically available and the license is free for non-commercial users at http://www.rosettacommons.org/

Multi-template MC4R comparative modeling with RosettaCM

a) Manually generated files (this information may be copied directly into new text files)

MC4R_truncated.fasta

```
>MC4R
SESLGKGYSDGGCYEQLFVSPEVFVTLGVISLLENILVIVAIAKNKNLHSPMYFFICSLA
VADMLVSVSNGSETIVITLLNSTDTDAQSFTVNIDNVIDSVICSSLLASICSLLSIAVDR
YFTIFYALQYHNIMTVKRVGIIISCIWAACTVSGILFIIYSDSSAVIICLITMFFTMLAL
MASLYVHMFLMARLHIKRIAVLPGTGAIRQGANMKGAITLTILIGVFVVCWAPFFLHLIF
YISCPQNPYCVCFMSHFNLYLILIMCNSIIDPLIYALRSQELRKTFKEIICCYP
```

MC4R_truncated.span

```
TM region prediction for MC4R_truncated.octopus predicted using OCTOPUS
7 294
antiparallel
n2c
  20    40    20    40
  55    75    55    75
  98   118    98   118
 139   159   139   159
 167   187   167   187
 219   239   219   239
 254   274   254   274
```

MC4R.disulfide

```
13 252
244 250
```

flags_membrane

```
-in:file:fasta MC4R_truncated.fasta
-parser:protocol rosetta_cm_membrane.xml
-in:detect_disulf true
-relax:minimize_bond_angles
-relax:minimize_bond_lengths
-relax:jump_move true
-default_max_cycles 200
-relax:min_type lbfgs_armijo_nonmonotone
-relax:jump_move true
-score:weights stage3_rlx_membrane.wts
-use_bicubic_interpolation
-hybridize:stage1_probability 1.0
-sog_upper_bound 15
-membrane
-in:file:spanfile MC4R_truncated.span
-membrane:no_interpolate_Mpair
-membrane:Menv_penalties
-rg_reweight .1
```

rosetta_cm_membrane.xml

```
<dock_design>
        <TASKOPERATIONS>
        </TASKOPERATIONS>
        <SCOREFXNS>
                <stage1 weights=stage1_membrane symmetric=0>
                        <Reweight scoretype=atom_pair_constraint weight=1/>
                </stage1>
                <stage2 weights=stage2_membrane symmetric=0>
                        <Reweight scoretype=atom_pair_constraint weight=0.5/>
                </stage2>
                <fullatom weights=stage3_rlx_membrane symmetric=0>
                        <Reweight scoretype=atom_pair_constraint weight=0.5/>
                </fullatom>
        </SCOREFXNS>
        <FILTERS>
        </FILTERS>
        <MOVERS>
        <Hybridize name=hybridize stage1_scorefxn=stage1 stage2_scorefxn=stage2 fa_scorefxn=fullatom
        batch=1 stage1_increase_cycles=1.0 stage2_increase_cycles=1.0 linmin_only=1>
        <Fragments 3mers="aaMC4RA03_05.200_v1_3"     9mers="aaMC4RA09_05.200_v1_3"/>
        <Template pdb="1u19_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="2rh1_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="2vt4_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="2X72_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="2Y03_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3eml_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3odu_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3pbl_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3QAK_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3rze_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3SN6_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3uon_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="3v2w_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4daj_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4djh_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4dkl_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4ea3_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4ej4_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4IAR_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        <Template pdb="4IB4_clean.pdb.pdb" cst_file="AUTO" weight=1.000 />
        </Hybridize>
        </MOVERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover=hybridize/>
        </PROTOCOLS>
</dock_design>
```

relax.options

```
-database /main/database/

-in:fix_disulf MC4R.disulfide #read disulfide connectivity information
-in:file:spanfile MC4R_truncated.span
-relax:membrane #set up membrane environment for relax
-relax:dualspace
-relax:minimize_bond_angles #setting used with dualspace relax
-set_weights cart_bonded .5 pro_close 0 #dualspace specific setting
-default_max_cycles 200
-out:file:fullatom #output file will be fullatom
-out:pdb
-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts
```

b) Steps and commands: The following steps describe specific command lines and resulting files

	Step	Text	Command	Comment
1	Create GPCR alignments	Create alignment profile for 20 GPCR's from PDB	mustang -i 1u19A_clean.pdb 2rh1A_clean.pdb 2vt4A_clean.pdb 3emlA_clean.pdb 3oduA_clean.pdb 3rzeA_clean.pdb 3v2wA_clean.pdb 4djhA_clean.pdb 4ea3A_clean.pdb 3pblA_clean.pdb 3uonA_clean.pdb 4dajA_clean.pdb 4dklA_clean.pdb 4ej4A_clean.pdb 2Y03_clean.pdb 4IB4_clean.pdb 4IAR_clean.pdb 3SN6_clean.pdb 3QAK_clean.pdb 2X72_clean.pdb -F fasta	Gives a pdb and .afasta file with structural alignment
2	Create MCR alignments	Create sequence alignment of MC1, MC2, MC3, MC4, and MC5	clustalw -> Sequence input from disc -> MCR_all.fasta -> multiple alignments	Gives .aln and .dnd files
3	Align MCR and GPCRs	Align GPCR structural alignment profile with MCR sequence alignment	clustalw -> Profile/Structural alignments -> 1st profile = gpcr_all_mustang.afasta; 2nd profile/sequences = MCR_all.aln -> Align sequences to 1st profile (slow/accurate)	Gives .aln and .dnd files Convert to Gishin format manually
4	Thread templates	Thread MC4 sequence over each template using Gishin alignment	partial_thread.linuxgccrelease -database ROSETTA_DATABASE_PATH/ -in:file:fasta MC4R_truncated.fasta -in:file:alignment mc4_TEMPLATE.aln -in:file:template_pdb TEMPLATE.pdb	Generates one threaded pdb per template for a total of 20 pdbs.
5	Generate MC4 fragment files	Use Robetta online fragment server to generate MC4 fragment files from truncated MC4 sequence	http://robetta.bakerlab.org/fragmentsubmit.jsp	Download and save fragment files aaMC4RA03_05.200_v1_3 and aaMC4RA09_05.200_v1_3
6	Generate span file	Predict membrane spanning region with OCTOPUS and convert to .span file	http://octopus.cbr.su.se/	MC4R_truncated.span
7	Hybridize	Run RosettaCM hybridize protocol	rosetta_scripts.linuxgccrelease @flags_membrane -database ROSETTA_DATABASE_PATH/ -out:prefix hybridize_ -nstruct 1000 > hybridize.log	Generates 1000 models per run.
8	Relax models	Run final relax over hybridize models including membrane spanning and disulfide definitions	relax.linuxgccrelease @relax.options -s MODEL -nstruct 1	Generates 1 model per hybridize model
9	Cluster pdbs	Cluster models using BCL	ls *.pdb > pdb_list.ls bcl.exe PDBCompare -quality RMSD -atom_list CA -pdb_list pdb_list.ls -prefix MC4R_rmsd -aaclass AACaCb -convert_to_natural_aa_type bcl.exe Cluster -distance_input_file MC4R_*rmsdRMSD.txt -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster5_MC4R -linkage Average* -remove_internally_similar_nodes 5 grep "Leaf : 1 : " cluster5_MC4R.Centers. | sort -nk10 | (Lists the top 5 clusters)	Generates cluster center and row files.

Chapter 7

Future Directions and Concluding Remarks

7.1 Future Directions

3D-QSAR descriptor improvements

Chapters 2 and 3 discuss several improvements to 3D-QSAR descriptors that increase QSAR model performance. Additional avenues of improvement remain and several potential modifications are currently being pursued.

In chapter 3, decreasing the maximum 3DA atom pair distance cutoff from 12 Å to 6 Å significantly improved QSAR model performance. One possible explanation is that capturing molecule fragments instead of the entire structure avoids problems arising from conformational flexibility. Encoding a single conformation has presented a major drawback to many 3D-QSAR descriptors and several approaches to incorporating conformation flexibility have been published under the general category called 4D-QSAR (see chapter 1). One approach to account for conformational flexibility is to encode all low energy conformations as individual molecules with equal activities for the different conformations of the same molecule. This approach has several drawbacks that make it an unappealing approach. Many datasets used to train effective QSAR models include over one hundred thousand compounds. Therefore, the size of these datasets quickly becomes computationally inefficient when considering all low-energy conformations.

Several more intuitive approaches are currently being explored with the BioChemical Library (BCL) to incorporate conformational flexibility in 3D-QSAR descriptors. If the reduced sensitivity to conformational flexibility is the cause for increase model performance at a distance limit of 6 Å, then future improvements may allow a distance cutoff to once again encapsulate the entire molecule.

One approach is replacing the static Gaussian smoothing of 3DA_Smooth with a dynamic smoothing that correlates the curve width to the atom pair distance. The width of distribution is easily controlled with the smoothing coefficient of standard curve equation and therefore, instead of using a constant smoothing coefficient throughout the entire 3DA_Smooth, this coefficient can be adjusted for different atom pair distances. Shorter atom pair distances, being less susceptible to conformational flexibility, can be distributed in a narrow curve, while longer atom pair distances can be distributed over a wider range of distance centers. More sophisticated correlations are being explored that take into account the number of rotatable bonds separating two atoms and atom types that can influence flexibility when adjusting smoothing coefficients.

Chapter 3 also presents an improvement to 3DA called 3DA_Sign that separates atom pairs based on signs. The improved performance of 3DA_Sign is likely due to the avoidance of information loss inherent with combining two atom properties into a single coefficient. This inevitably leads to the speculation of additional forms of information loss inherent with this treatment of atom properties. For example, when combining two atom properties of significantly different proportions, it becomes difficult to encode which atom of the pair contributes more to the property coefficient than the other. We are currently exploring modifications to 3DA that mix atom properties in different ways in a new descriptor called MultiProperty3DA. This descriptor can take up to two different atom properties to encode property distances including the distances between hydrogen bond donors and receivers. Additionally, this descriptor allows for one of the two atoms for every pair to go unweighted, thereby avoiding the potential mixing of atom properties with significantly different intensities

Discovering positive allosteric modulators of Y_4R

In chapter 5, high throughput screening is used to discovery five small molecule positive allosteric modulators (PAMs) of Y_4R. These compounds all share a common scaffold. This is both beneficial and limiting. A common scaffold makes molecular alignment straightforward and provides early insight into small molecule structure-activity relationships. Importantly, IP accumulation assays against all neuropeptide Y receptor subtypes extended this relationship to subtype selectivity. However, sharing a common scaffold severely limits the chemical space covered by these hits. One of the benefits of unbiased HTS is its ability to identify structurally unique hits. On the other hand, the four structurally similar compounds identified with the help of Tanimoto similarity to the initial hit underscore the utility of CADD in enhancing a traditional HTS approach to improve hit rates. Therefore, while future HTS studies will benefit from LB-CADD guidance, care must be taken to avoid limiting chemical search space to this single scaffold.

One potential approach is to train non-linear models using QSAR descriptors including those described in chapters 2 and 3. The data generated for chapter 4 presents a dataset of approximately 35,000 compounds over which to train models. Despite being relatively small, this dataset is of sufficient size to train QSAR models that can screen virtual compound libraries and prioritize compounds to be screened. This approach, however, significantly suffers from the extremely small population of active compounds. Additionally, the common scaffold of these active compounds will significantly restrict the conformational space explored by these models. Models will be trained to recognize and prioritize this scaffold and positive predictions outside of this scaffold will be rare if at all. This particular LB-CADD approach will be better served after a larger set of diverse Y_4R PAMs has been compiled.

With the present set of active compounds, a more appropriate approach is with ligand-based pharmacophores (see chapter 1). The shared scaffold simplifies structural alignment, a prerequisite for ligand-based pharmacophore generation. Additionally, the inclusion of similar inactive compounds can significantly refine and enhance a pharmacophore hypothesis. Most importantly, pharmacophore mapping is specifically powerful in cases where all compounds share a common scaffold because it is capable of discovering novel scaffolds through a process called “scaffold hopping.” Because pharmacophores represent the spatial distribution of physicochemical properties independent of overall molecular geometry, generating a pharmacophore hypothesis that can be used to screen virtual compound libraries allows for the identification of structurally distinct compounds that share a similar distribution of properties.

Pharmacophore hypothesis generation is not currently implemented in the BCL framework. However, chapter 1 lists several software packages and instances where ligand-based pharmacophores were used to identify new scaffolds in similar situations. Commercially available software packages such as Catalyst [1] contain multiple tools such as HipHop for the development of ligand-based pharmacophore models. More recent tools such as HypoGen can enhance the pharmacophore hypotheses for niclosamide-similar compounds with the inclusion of the structurally similar inactive compounds. Once one or more pharmacophore hypotheses are generated, they can be compared against the Vanderbilt Institute for Chemical Biology compound library to select compounds and prioritize a third screen. Successful identification of structurally unique hits in a third screen may provide the necessary chemical space to train high quality non-linear QSAR capable of screening virtual libraries of millions of compounds. Those predicted to be active can either be synthesized or purchased, extending the high throughput screening well beyond the scope of the on-site compound library.

Y_4R-PP models: application to drug discovery

Chapter 4 describes the application of LB-CADD to enhancing the hit rate for screening small molecule modulators of Y_4R and chapter 5 describes the application of SB-CADD to modeling the interaction of Y_4R with its endogenous peptide agonist PP. Combining LB-CADD and SB-CADD with discoveries from both projects may provide additional routes for discovering Y_4R modulators and provide significant insights into the modulation of PP signaling at Y_4R by niclosamide and related compounds.

As discussed, ligand-based pharmacophores can significantly enhance future screening projects with Y_4R. However, combining structure-based and ligand-based pharmacophores has been shown to increase the number of chemotypes retrieved [2]. Despite the absence of an experimentally derived structure of Y_4R, results from chapter 5 can be combined with hit compounds of chapter 4 to generate structure-based pharmacophores.

The first step towards developing a structure-based pharmacophore of these Y_4R PAMs is the identification of a common binding pocket for this particular scaffold. However, several caveats to this approach must be considered. To date, few structures of GPCRs in complex with allosteric modulators have been experimentally elucidated. One structure reveals the binding pose of LY2119620, an allosteric modulator of M2 receptor that involves extracellular portions of the receptor [3]. Recently, the crystal structure of P2Y1 in complex with BPTU, an allosteric antagonist, has been published which reveals a completely distinct binding site on the receptor surface within the lipid bilayer [4]. Additionally, studies of the chemokine receptors reveal at least two potential allosteric binding sites including transmembrane and cytoplasmic binding sites [5]. This makes computationally modeling the binding site of Y_4R PAMs challenging without any experimental evidence to guide binding site prediction.

Y_4R-PP models may help to identify the binding site of niclosamide at Y_4R. Unlike crystalized M2 receptor agonists, PP is a 36 amino-acid peptide that binds to the extracellular surface, reaching its C-terminal into the transmembrane pore and occluding significant portions of the extracellular surface of Y_4R. This information can restrict the conformational search space to sites that can accommodate niclosamide in the absence and presence of PP.

An appendix chapter has been included that presents preliminary docking studies of niclosamide and Y_4R-PP. This initial approach focuses on the common small molecule orthosteric binding site in class A GPCRs which, unlike the position of LY2119620 on M2, would not obscure the binding site of PP on Y_4R. Small molecule docking tools in Rosetta were used to dock niclosamide to Y_4R models with and without PP. This strategy is designed to identify low energy binding poses that incorporate similar residues from both sets of models. Since no experimental information was used to guide these models, clustering analyses were performed to generate large conformation ensembles containing models with good binding energies in different poses. Examining the pairwise interaction energies between niclosamide and every residue of these models revealed several residues that interact with niclosamide across different poses. These residues present targets of interest for future targeted mutagenesis studies aimed at loss of niclosamide affinity and/or activity.

Preliminary results are encouraging as several residues interacting with niclosamide across most models appear to participate in the GPCR agonist binding and activation. For example, $Trp^{6.48}$ shows favorable interactions with niclosamide across over 70% of docked models. Movements at this conserved position have been shown to be involved in activation of M2 and A2A receptors [6, 7]. Therefore, interaction between niclosamide and this residue may participate in potentiation of Y_4R activation from PP.

If *in vitro* verification of these predicted interactions reveals one or more residues that significantly decrease binding of niclosamide to Y_4R, these experimental results may be used to guide a second round of niclosamide docking that can focus conformational search space to model a binding pose of higher confidence. Alternatively, if mutagenesis studies show that these predicted residues have little to no effect on niclosamide binding, a second round of docking may be focused on allosteric binding sites outside of the transmembrane pore.

Once the binding site has been elucidated with confidence through iterative rounds of modeling and *in vitro* experiments, two potential strategies may be used to aid in virtually screening for additional Y_4R modulators. The first strategy, as mentioned, involves the combination of ligand-based and structure-based pharmacophores that has already been shown to outperform single-strategy pharmacophores. Residues participating in the modeled binding site of niclosamide can be used to construct the structure based pharmacophores.

A second potential strategy is to combine ligand-based and structure-based approaches to generate a pseudoreceptor [8]. A pseudoreceptor constructs a high-resolution 3D model of only the binding site that can be used for virtual screening. Peptide-based pseudoreceptors construct a binding site around a known ligand using amino acids found to contact the ligand, including directional interactions defined as vectors. Pseudoreceptor models have been used successfully to identify high affinity ligands for 5HT1A and binding determinants for cocaine derivatives [9, 10]. Both studies found that averaging potential states of ligands over multiconformer models outperformed single pseudoreceptor models. This finding fits well with the modeling approach used for docking PP and niclosamide to Y_4R that uses conformation ensembles to predict interactions rather than a single binding pose.

MC4R-α-MSH models

Chapter 6 models an interaction between α-MSH and MC4R. [Nle4,dPhe7]-MSH (NDP-MSH) is an analogue of α-MSH that shows dramatically improved stability and binding affinity. The altered stereochemistry of Phe^7 is considered to be the most important factor for the superior affinity of NDP-MSH. Modeling the underlying interactions owing to this change in affinity may be useful for designing high affinity drugs that target MC_4R. Several interactions captured in the presented models suggest possible interactions that differ between α-MSH and NDP-MSH binding. Specifically, two residues that interacted with Phe^7 in our models showed decreased binding to α-MSH but not NDP-MSH when mutated including $Ile^{3.32}$ and $Phe^{7.35}$.

Currently, integration of non-canonical amino acids, including D-amino acids has limited support in Rosetta. Implementation is limited to side chain rotamer optimization and energy minimization. This means that the tools employed to model the interaction between α-MSH and MC4R including FlexPepDock and loop modeling and refinement are not current applicable to docking NDP-MSH. However, there are multiple researchers in different Rosetta laboratories that are extending the implementation of non-canonical amino acids and D-amino acids to include a variety of different applications including FlexPepDock and loop modeling. Once these methods have been successfully implemented, it will be useful to dock NDP-MSH to MC4R and rigorously compare all interaction profiles between the two model ensembles.

In addition to the interactions supported with mutagenesis data, models predicted a potential interaction between $Met^{7.32}$ of MC4R and Ser^4 or Glu^5 of α-MSH. This residue has not been previously explored with mutagenesis in α-MSH binding assays. Therefore, targeted mutagenesis can be used to determine whether $Met^{7.32}$ is an important residue for α-MSH binding.

7.2 Closing Remarks

At the interface between extracellular signals and intracellular changes, GPCRs present one of the most important pharmacological targets in medicine. Modern advancements in experimental techniques provide insight into the behavior of these proteins and the psychochemical properties underlying their ligand interactions. The utility of such discoveries is evident in their direct contribution to many sophisticated treatments capable of targeting specific and complex physiological symptoms. However, as with all science, every discovery prompts new questions, new ideas, and new hypotheses. Computer aided drug discovery gives scientists the means to combine information and new discoveries into systems of extraordinary complexity. With these models, it becomes possible to explore events involving hundreds to thousands of simultaneous processes. Not only can these models help scientists paint individual discoveries into comprehensive biological murals, but computational models produce environments where questions can be prioritized, ideas can be explored, and hypotheses tested in a rapid and cost-effective framework.

The benefits of scientific collaboration can't be overstated and computational techniques are not an exception but a prominent example of such benefits. This conversation between experimental and computational approaches fosters spectacular insight that can be directly applied to improving human life. Much of the present work is only possible through the strong collaboration between experts in the field of cell signaling, biochemistry, and protein function represented by members of the Annette Beck-Sickinger lab of Leipzig University and experts in computational structure biology and biophysics represented by members of the Jens Meiler lab of Vanderbilt University.

During the generation of this work, many exciting discoveries and advancements were made regarding the conformational changes accompanying GPCR activation. In the past few years, the first experimentally derived structures of active GPCRs in complex with G-proteins have become available, contributing to a comprehensive model of GPCR activation that covers shared conformational changes across the common topology and those unique to specific receptors. The time scale and complexity of such conformational changes typically overwhelm available computational resources. However, recent hardware advancements and techniques that accelerate molecular dynamics simulations have allowed computation techniques to join experimental techniques in modeling these important processes. For example, the recent application of sophisticated processors designed for molecular dynamics allowed for the atomic-level molecular dynamic simulations of microsecond timescale events such as G-protein nucleotide exchange [11]. Additionally, novel tools in cloud computing allowed researchers to combine individual nanosecond molecular dynamic simulations into Markov state model microsecond timescale simulations of the activation of the β2 adrenergic receptor [12, 13].

One exciting future for the presented work with Y_4R involves modeling the overall conformational changes of Y_4R following binding of PP and allosteric modulators. With the application of these new computational techniques, Y_4R conformational spaces can be explored that include the resting unbound states, PP bound states, and positive allosteric modulator bound states. Comparing these populations and the elucidating shared transition states can help explain how compounds such as niclosamide potentiate activation of Y_4R by PP. Understanding these conformational lines of communication can improve small molecule therapeutics targeting subtle and specific aspects of receptor activation. Modeling the activation of MC4R by α-MSH, on the other hand, can help to explain the effects of specific mutations on MC4R signaling that contribute to monogenic obesity, contributing to the growth of personalized medicine.

7.3 References

1. Kurogi Y & Guner OF (2001) Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Current medicinal chemistry* 8(9):1035-1055.
2. Cross S & Cruciani G (2010) Grid-derived structure-based 3D pharmacophores and their performance compared to docking. *Drug Discovery Today: Technologies* 7(4):e213-e219.
3. Kruse AC*, et al.* (2013) Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* 504(7478):101-106.
4. Zhang D*, et al.* (2015) Two disparate ligand-binding sites in the human P2Y1 receptor. *Nature* 520(7547):317-321.
5. Yanamala N & Klein-Seetharaman J (2010) Allosteric Modulation of G Protein Coupled Receptors by Cytoplasmic, Transmembrane and Extracellular Ligands. *Pharmaceuticals* 3(10):3324-3342.
6. Miao Y, Nichols SE, Gasper PM, Metzger VT, & McCammon JA (2013) Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the National Academy of Sciences of the United States of America* 110(27):10982-10987.
7. Li J, Jonsson AL, Beuming T, Shelley JC, & Voth GA (2013) Ligand-dependent activation and deactivation of the human adenosine A(2A) receptor. *Journal of the American Chemical Society* 135(23):8749-8759.
8. Tanrikulu Y & Schneider G (2008) Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. *Nature reviews. Drug discovery* 7(8):667-677.
9. Guccione S, Doweyko AM, Chen H, Barretta GU, & Balzano F (2000) 3D-QSAR using 'multiconformer' alignment: the use of HASL in the analysis of 5-HT1A thienopyrimidinone ligands. *J Comput Aided Mol Des* 14(7):647-657.
10. Srivastava S & Crippen GM (1993) Analysis of cocaine receptor site ligand binding by three-dimensional Voronoi site modeling approach. *Journal of medicinal chemistry* 36(23):3572-3579.
11. Dror RO*, et al.* (2015) SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* 348(6241):1361-1365.
12. Shukla D, Lawrenz M, & Pande VS (2015) Elucidating Ligand-Modulated Conformational Landscape of GPCRs Using Cloud-Computing Approaches. *Methods in enzymology* 557:551-572.
13. Feixas F, Lindert S, Sinko W, & McCammon JA (2014) Exploring the role of receptor flexibility in structure-based drug discovery. *Biophysical chemistry* 186:31-45.

Appendix

Modeling Interactions of the Human Y_4 Receptor and Niclosamide

Introduction

In chapter 4, several positive allosteric modulators of Y_4R were discovered with a common scaffold. Understanding the interaction of these molecules with Y_4R and potentiation of PP signaling at Y_4R can help discover modulators of Y_4R with higher affinity and activity. Chapter 5 introduces a comprehensive model of Y_4R and PP interaction. This appendix chapter presents initial results for modeling the interaction site of niclosamide on Y_4R based on the models generated in chapter 5. These results can be used to suggest targeted mutagenesis studies that can, in turn, be applied to a second round of niclosamide docking to Y_4R. This introduction presents the overall strategy of ligand docking in Rosetta and the rationale behind initial ligand placement.

Ligand docking into GPCR comparative models is shown to be relatively successful depending on the information known prior to modeling [1]. However, this approach can be especially difficult when no information regarding the binding site is known. Rosetta does not perform binding site detection but relies on the user to specify the area around which ligand docking will be sampled. One option is to use information from templates to guide the docking process. Class A GPCRs share a relatively conserved orthosteric binding site occupying the extracellular portion of TM3, TM5, TM6 and to a lesser extent TM2, TM7, and some of ECL2 [1]. Topologically similar residues across a variety of class A GPCRs participate in orthosteric binding including residues 3.32, 3.33, 3.36, 6.48, 6.51 and 7.39 [2]. Additionally, binding pockets within receptors of the same type may be practically identical. For example, β1 and β2 adrenergic receptors both share a salt bridge at $Asp^{3.32}$ and hydrogen bonds at $Ser^{5.42}$ and $Asn^{7.39}$ with their small molecule ligands [3].

Despite the orthosteric binding site similarities across many class A GPCRs, the depth of this binding pocket can vary significantly, especially when comparing small molecule and peptide ligands. Peptide ligands tend to bind closer to the extracellular surface [4]. Chemokine receptors, for example, bind ligands at the top of the transmembrane bundle and interactions involve mainly extracellular domains [5]. The crystal structure of NT8-13 with NT1 receptor reveals a shallow binding cavity that does not penetrate the receptor with a pose almost perpendicular to the membrane [6].

Allosteric compounds present additional challenges for docking since they bind to sites topologically distinct from endogenous ligands [5]. Experimental evidence shows a trend for class A GPCR allosteric modulators to bind at shallow cavities that include extracellular regions [3]. The crystal structure of M2 in complex with the positive allosteric modulator LY2119620, for example, reveals a binding site at the extracellular portions of the receptor [7]. The residues that create this binding pocket appear to be shared across other muscarinic acetylcholine receptors [8]. However, a shallow, extracellular binding pocket is not the only position allosteric ligands may bind. For example, two allosteric binding sites have been proposed for chemokine receptors: a transmembrane binding site similar to the allosteric binding site seen in class B GPCRs and a cytoplasmic allosteric binding site [8]. Therefore, additional care must be taken when analyzing models involving allosteric compounds when little or no information of the allosteric binding site is known. For docking niclosamide to Y_4R, information is gained from the fact that PP obscures much of the extracellular surface of the receptor. In the absence of any experimental restraints other than the fact that niclosamide bound Y_4R must accommodate PP binding to the extracellular surface, an initial search at the general class A GPCR small-molecule orthosteric binding pocket is a logical choice.

Small molecule docking in the Rosetta Molecular Modeling Suite

RosettaLigand is a modification of its predecessor RosettaDock [9] that docks small molecule ligands with proteins [10]. Its unique design samples the flexibility of both the small molecule ligand and protein target simultaneously [11, 12]. RosettaLigand has been shown to successfully predict binding modes of small molecules in ten different comparative GPCR structures [13]. When compared with other ligand-docking programs including Dock, FlexX, Glide, GOLD, MOE, and others, RosettaLigand performance was comparable or better across 136 ligands and eight target receptors [14].

Rosetta uses two strategies to sample small molecule flexibility. A pre-generated ensemble of ligand conformations is provided that will be randomly sampled during the docking process and Rosetta internally defines bonds that may be rotated during ligand conformational sampling [15].

RosettaLigand begins by translating the ligand within the user-defined sphere until its geometric center does not clash with any atoms in the receptor. This is followed by random rotation through all rotational degrees of freedom. Only rotations that significantly change the pose and pass the Lennard-Jones attractive and repulsive score filter are stored. One pose is randomly selected for high resolution docking. Initial translation and rotation cycles allow Rosetta to sample hard to find poses and tight binding cavities [15].

The high resolution stage of RosettaLigand combines stochastic receptor side-chain rotamer and ligand conformer sampling with small ligand movements evaluated with Monte Carlo simulated annealing [1]. This combination of ligand movement, conformation, and side chain rotamer sampling is designed to capture the simultaneous flexibility of ligand and receptor during the binding event [16]. The final stage of RosettaLigand employs a stringent gradient-based minimization while exploring minor changes in ligand position and orientation as well as receptor side chain and backbone torsion sampling with a hard repulsive VDW potential that creates a rugged energy landscape tuned to discriminate native from nonnative binding modes [16].

Evaluating receptor-ligand models focuses on specific binding interface scores rather than the overall energy of the model. Several binding interface-specific scoring terms predict binding free energy and critical interactions [17]. The 'interface_delta' score is a commonly used approximation of binding free energy in Rosetta that identifies the contribution of the total model score for which the ligand is responsible. Interface_delta is calculated as the difference of total model scores with and without the ligand [18]. Clustering models based on ligand pose is often used to discern native from nonnative poses with comparable interface_delta scores.

Results and Discussion

When niclosamide was docked to Y_4R alone, 25 residues contributed attractive energy scores to the interface energy of niclosamide and Y_4R in at least 30% of the top scoring models. This cutoff of 30% was purposely low to account for the fact that the location of niclosamide binding is largely unknown. As expected, docking niclosamide to Y_4R+PP limited the space available for niclosamide and 19 residues contributed attractive energy scores to the interface energy in at least 30% of the top scoring models. Contribution of all residues involved in niclosamide binding to Y_4R in both conditions is plotted in figure A.1. This plot reflects the greater degree of variability in poses lacking PP. Taken together, 11 residues contributed to the interaction of niclosamide and Y_4R across at least 30% of top models in both conditions. These residues are ranked in accordance with the average percentage of models across the two conditions that they contributed to niclosamide binding energy in table A-1. Residues of particular interest include conserved residues identified as contributors to activation of M2 and A2A such as $Trp^{6.48}$ and $Met^{7.43}$ [19, 20]. A similar low-energy binding pose of niclosamide docked to Y_4R alone and Y_4R+PP is illustrated in figure A.2.

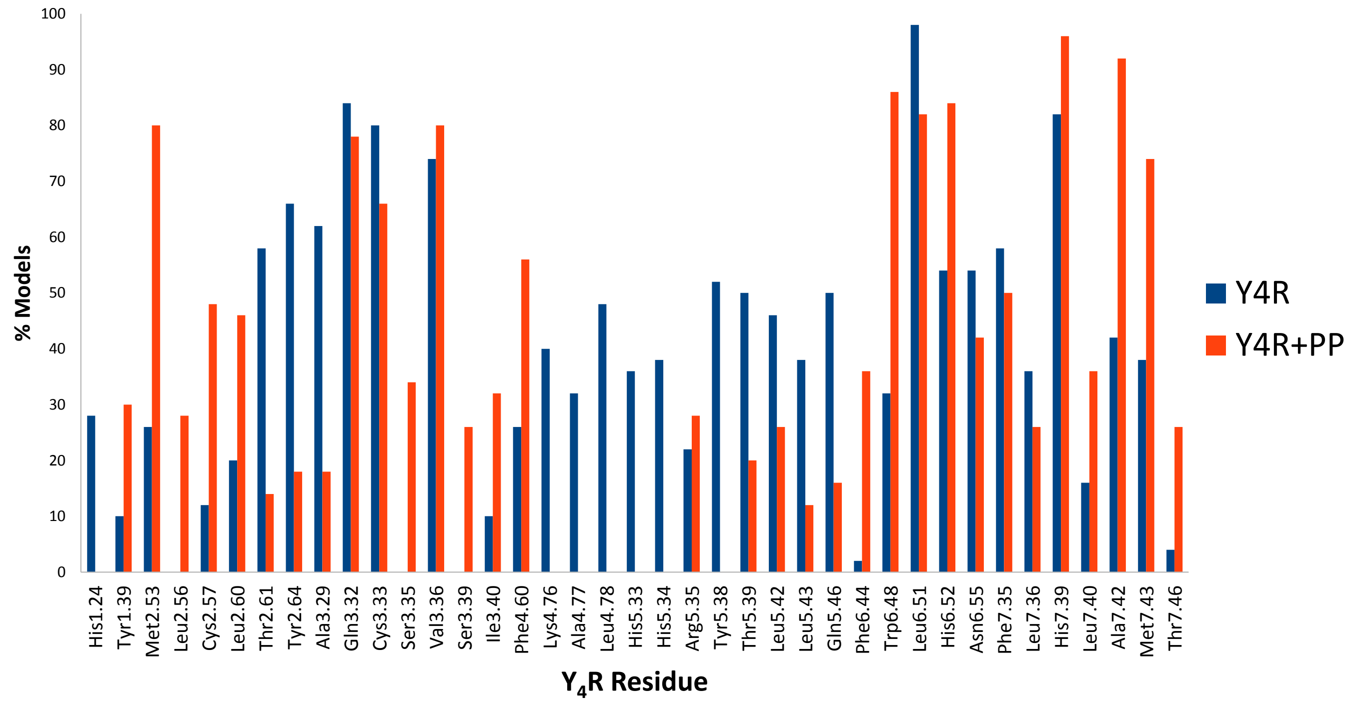


Figure A.1 Y4R residues interacting with niclosamide in docked models. Specific residues contributing to interface energy with niclosamide in at least 20% of models within each docking project are listed along the x-axis. Percent of models within specific condition is tracked in y-axis. Limited conformation space sampling in Y$_4$R+PP condition is highlighted by the lack of residues including Lys$^{4.76}$ through Gln$^{5.46}$.

Table A-1 Highly represented residues contributing to niclosamide interface energy in at least 30% of final models.

	Residue	Final Y$_4$R Models (out of 50)	Final Y$_4$R+PP Models (out of 50)	Average %
1	Leu6.51	49	41	90
2	His7.39	41	48	89
3	Gln3.32	42	39	81
4	Val3.36	37	40	77
5	Cys3.33	40	33	73
6	His6.52	27	42	69
7	Ala7.42	21	46	67
8	Trp6.48	16	43	59
9	Met7.43	19	37	56
10	Phe7.35	29	25	54
11	Asn6.55	27	21	48

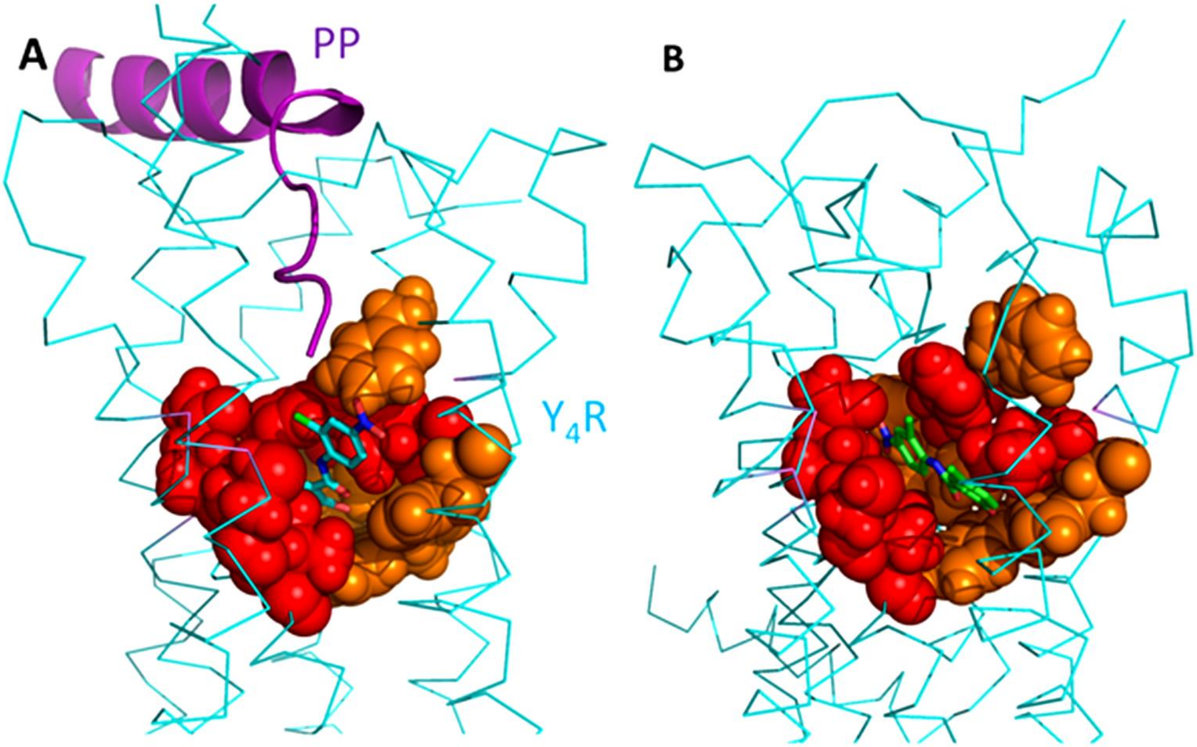


Figure A.2 Similar binding poses between Y4R+PP and Y4R models. A) Low energy binding pose from niclosamide docked to Y4R+PP B) Low energy binding pose from niclosamide docked to Y4R alone Cyan ribbon illustrates Y4R comparative model and purple cartoon illustrates PP. Red spheres illustrate residues 1-5 in table A-1 and orange spheres illustrate residues 6-10 in table A-1.

This small molecule docking strategy is designed to achieve the highest predictive power capable with the limited experimental evidence. This is the first step of an iterative workflow that passes information back and forth between computational modeling and *in vitro* studies. These models, therefore, represent potential binding poses based on the class A GPCR orthosteric binding pocket.

Predicted interactions in table A-1 were gathered from the ensemble of potential binding poses for prioritizing *in vitro* studies. Residues that consistently interact with niclosamide across the majority of the final models may be *suggested* participants in niclosamide binding. It is critical to test these residues or protein segments *in vitro* through mutagenesis, protein chimeras, or other techniques to verify these predictions.

After these predictions have been experimentally verified or rejected, a second round of modeling can be performed with the inclusion of additional experimental evidence. Naturally, this will limit the conformational search space and increase the confidence of future predictions. A second round of predictions can be made that may focus on elucidating other residues in the binding pocket or proposing specific interactions between the ligand and receptor. This iterative pipeline alternating *in vitro* and *in silico* applications has the greatest chance of producing high quality models that describe the binding mode at atomic level detail.

Methods

Two individual docking projects were run in parallel using RosettaLigand to dock niclosamide to Y_4R with and without PP. Results were combined for final analysis and binding interaction predictions.

Niclosamide conformation generation

Conformations of niclosamide were generated with LowModeMD conformational search in MOE using the MMFF94X force field and Born solvation. A conformation RMSD limit of 0.5 Å yielded three unique low-energy conformations. These conformations were used to define niclosamide parameterization files for RosettaLigand.

Docking niclosamide to Y_4R and Y_4R+PP

Niclosamide was docked to fifteen Y_4R comparative models and nine models of PP docked to Y_4R. Ligand starting coordinates were determined as the geometric center of niclosamide positioned within the class A GPCR orthosteric binding pocket as defined by the ligand overlap across the fourteen class A GPCR crystal structures used as templates for Y_4R comparative modeling. See chapter 5 for template structure details.

Low resolution docking included 50 cycles of ligand translation with a radius of six angstroms around the starting coordinates followed by 500 cycles of 360 degree rotation and a slide-together step to identify the low resolution binding pose. High resolution docking included six cycles of side chain refinement and minimization that included simulated potential solvent interactions at the binding site.

Ten thousand models were generated each for niclosamide docked to Y_4R and niclosamide docked to Y_4R+PP.

Model analysis and selection

The top 5,000 models by interface_delta score within each docking project (Y_4R and Y_4R+PP) were clustered independently using BCL::Cluster [21]. Models were clustered based on niclosamide RMSD with a cut-off of 3.5 Å. The top five models from the five largest clusters were combined with the top twenty five models by interface energy score to produce an ensemble of 50 models for each docking project. Models that produced the best interface energy scores and fell within the top clusters were permitted as duplicates in the final ensemble to allow for a higher influence in the prediction of binding interactions.

Binding site prediction

Pairwise energy scores were isolated for niclosamide and specific interacting residues using the Rosetta application Residue Energy Breakdown. Repulsive interactions were filtered and attractive energy potentials between specific residues and niclosamide were tallied within each docking project. Consistent attractive interactions were compared between projects and consensus binding poses were used to predict specific residues involved in the binding of niclosamide to Y_4R.

References

1. Nguyen ED, Norn C, Frimurer TM, & Meiler J (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. *PloS one* 8(7):e67302.
2. Venkatakrishnan AJ*, et al.* (2013) Molecular signatures of G-protein-coupled receptors. *Nature* 494(7436):185-194.
3. Shonberg J, Kling RC, Gmeiner P, & Lober S (2014) GPCR crystal structures: Medicinal chemistry in the pocket. *Bioorganic & medicinal chemistry*.
4. Kenakin T & Miller LJ (2010) Seven transmembrane receptors as shapeshifting proteins: the impact of allosteric modulation and functional selectivity on new drug discovery. *Pharmacological reviews* 62(2):265-304.
5. Wootten D, Christopoulos A, & Sexton PM (2013) Emerging paradigms in GPCR allostery: implications for drug discovery. *Nature reviews. Drug discovery* 12(8):630-644.
6. Egloff P*, et al.* (2014) Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 111(6):E655-662.
7. Kruse AC*, et al.* (2013) Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* 504(7478):101-106.
8. Yanamala N & Klein-Seetharaman J (2010) Allosteric Modulation of G Protein Coupled Receptors by Cytoplasmic, Transmembrane and Extracellular Ligands. *Pharmaceuticals* 3(10):3324-3342.
9. Chaudhury S*, et al.* (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PloS one* 6(8):e22477.
10. Meiler J & Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* 65(3):538-548.

11. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, & Meiler J (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49(14):2987-2998.
12. Allison B*, et al.* (2014) Computational design of protein-small molecule interfaces. *Journal of structural biology* 185(2):193-202.
13. Kaufmann KW & Meiler J (2012) Using RosettaLigand for Small Molecule Docking into Comparative Models. *PloS one* 7(12):e50769.
14. Davis IW, Raha K, Head MS, & Baker D (2009) Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein science : a publication of the Protein Society* 18(9):1998-2002.
15. Lemmon G & Meiler J (2012) Rosetta Ligand docking with flexible XML protocols. *Methods in molecular biology* 819:143-155.
16. Davis IW & Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. *Journal of molecular biology* 385(2):381-392.
17. Combs SA*, et al.* (2013) Small-molecule ligand docking into comparative models with Rosetta. *Nature protocols* 8(7):1277-1298.
18. Kaufmann KW*, et al.* (2009) Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins* 74(3):630-642.
19. Miao Y, Nichols SE, Gasper PM, Metzger VT, & McCammon JA (2013) Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the National Academy of Sciences of the United States of America* 110(27):10982-10987.
20. Li J, Jonsson AL, Beuming T, Shelley JC, & Voth GA (2013) Ligand-dependent activation and deactivation of the human adenosine A(2A) receptor. *Journal of the American Chemical Society* 135(23):8749-8759.
21. Alexander N, Woetzel N, & Meiler J (2011) Bcl::Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pp 13-18.

a) Manually generated files

dock_nic.options

```
-database /Rosetta/main/database/
-in:fix_disulf y4.disulfide
-in:file:extra_res_fa NIC.params
-packing:ex1, ex2
-parser:protocol dock_nic.xml
-out:pdb
```

dock_nic.xml

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <ligand_soft_rep weights=ligand_soft_rep>
                        <Reweight scoretype=fa_elec weight=0.42/>
                        <Reweight scoretype=hbond_bb_sc weight=1.3/>
                        <Reweight scoretype=hbond_sc weight=1.3/>
                        <Reweight scoretype=rama weight=0.2/>
                </ligand_soft_rep>
                <hard_rep weights=ligand>
                        <Reweight scoretype=fa_intra_rep weight=0.004/>
                        <Reweight scoretype=fa_elec weight=0.42/>
                        <Reweight scoretype=hbond_bb_sc weight=1.3/>
                        <Reweight scoretype=hbond_sc weight=1.3/>
                        <Reweight scoretype=rama weight=0.2/>
                </hard_rep>
        </SCOREFXNS>
        <LIGAND_AREAS>
                <docking_sidechain_X    chain=X    cutoff=6.0    add_nbr_radius=true    all_atom_mode=true
                minimize_ligand=10/>
                <final_sidechain_X chain=X cutoff=6.0 add_nbr_radius=true all_atom_mode=true/>
                <final_backbone_X     chain=X     cutoff=7.0     add_nbr_radius=false     all_atom_mode=true
                Calpha_restraints=0.3/>
        </LIGAND_AREAS>
        <INTERFACE_BUILDERS>
                <side_chain_for_docking ligand_areas=docking_sidechain_X/>
                <side_chain_for_final ligand_areas=final_sidechain_X/>
                <backbone ligand_areas=final_backbone_X extension_window=3/>
        </INTERFACE_BUILDERS>
        <MOVEMAP_BUILDERS>
                <docking sc_interface=side_chain_for_docking minimize_water=true/>
                <finalsc_interface=side_chain_for_finalbb_interface=backbone minimize_water=true/>
        </MOVEMAP_BUILDERS>
        <MOVERS>
                <StartFrom name=start_from_X chain=X>
                        <Coordinates x=36.0856 y=7.93917 z=16.4311/>
                </StartFrom>
                <CompoundTranslate name=compound_translate randomize_order=false allow_overlap=false>
                        <Translate chain=X distribution=uniform angstroms=6.0 cycles=50/>
                </CompoundTranslate>
                <Rotate name=rotate_X chain=X distribution=uniform degrees=360 cycles=500/>
                <SlideTogether name=slide_together chains=X/>
                <HighResDocker  name=high_res_docker  cycles=6  repack_every_Nth=3  scorefxn=ligand_soft_rep
                movemap_builder=docking/>
                <FinalMinimizer name=final scorefxn=hard_rep movemap_builder=final/>
                <InterfaceScoreCalculator name=add_scores chains=X scorefxn=hard_rep/>
                <ParsedProtocol name=low_res_dock>
                        <Add mover_name=start_from_X/>
                        <Add mover_name=compound_translate/>
                        <Add mover_name=rotate_X/>
                        <Add mover_name=slide_together/>
                </ParsedProtocol>
                <ParsedProtocol name=high_res_dock>
                        <Add mover_name=high_res_docker/>
                        <Add mover_name=final/>
                </ParsedProtocol>
        </MOVERS>
        <PROTOCOLS>
                <Add mover_name=low_res_dock/>
                <Add mover_name=high_res_dock/>
                <Add mover_name=add_scores/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

b) Specific commands

	Step	Text	Command	Comment
1	Generate params file	Convert conformations generated in MOE to Rosetta-compatable parameterization file	molfile_to_params.py niclosamide_conformations.sdf	Generates LG_0001.pdb, LG_0002.pdb, LG_0003.pdb and LG.params
2	Complete params file	Combine confirmations and add ensemble to parameterization file	cat LG_000*.pdb > NIC_confs.pdb mv LG.params NIC.params Manually add 'PDB_ROTAMERS NIC_confs.pdb' to end of NIC.params	Keep NIC_confs.pdb and NIC.params
4	Run RosettaLigand	Dock niclosamide to Y4R and Y4R+PP models using starting coordinates defined in step 3.	rosetta_scripts.default.linuxgccrelease @dock_nic.options -s Y4R_1_nic.pdb -out:prefix Y4R_1_nic_ -nstruct 150	This command generates 150 models per run. Run parallel trajectories to produce as many models as necessary. Run for each Y4R and Y4R+PP model.
5	Cluster models	Cluster top 5000 scoring models by interface_delta within Y4R and Y4R+PP docked models	bcl.exe molecule:Compare all_ligands.sdf -method RealSpaceRMSD -output all_model_rmsd -bcl_table_format bcl.exe Cluster -distance_input_file "all_model_rmsd" -input_format TableLowerTriangle -output_format Rows Centers -output_file Y4_nic_cluster35 -linkage Average -remove_internally_similar_nodes 3.5	Y4_nic_cluster35.Centers lists all clusters and can be used to identify the largest cluster nodes. Y4_nic_cluster35.Rows lists all models and their corresponding nodes.
6	Analyze residue-pairs	Calculate all residue-pair energies within selected models.	residue_energy_breakdown.default.linuxgccrelease -database /Rosetta/main/database/ -in:file:extra_res_fa NIC.params -s DOCKED_MODEL -out:file:silent energy_breakdown_DOCKED_MODEL.out	
7	Isolate niclosamide participating energies	Isolate residue pair interactions that participate in niclosamide binding to Y4R	grep 'LG' energy_breakdown_*.out > nic_interactions.tab	nic_interactions.tab can be opened in spreadsheet or other analysis tools to examine common interactions across models and conditions.

Publications and Presentations

Publications

Sliwoski G., Kothiwale S., Meiler J., and Lowe EW Jr.: **Computational methods in drug discovery**. *Pharmacol Rev* (2013), 66: 334-395.

Sliwoski, G.R., Lowe, E.W., Butkiewicz, M., and Meiler, J. **BCL::EMAS – enantioselective molecular asymmetry descriptor for 3D-QSAR**. *Molecules* (2012), 17: 9971-9989.

Pedragosa-Badia X., Sliwoski G.R., Dong Nguyen E., Lindner D., Stichel J., Kaufmann K.W., Meiler J., and Beck-Sickinger A.G.: **Pancreatic polypeptide is recognized by two hydrophobic domains of the human Y4 receptor binding pocket**. *J Biol Chem* (2014), 289: 5846-5859.

Kaufmann, K., Romaine, I., Days, E., Pascual C., Malik, A., Yang, L., Zoe, B., Du, Y., Sliwoski, G., Morrison, R.D., Denton, J., Niswender, C.M., Daniels, J.S., Sulikowski, G.A., Xie, X.S., Lindsley, C.W., and Weaver, C.D. **ML297 (VU0456810), the first potent and selective activator of the GIRK potassium channel, displays antiepileptic properties in mice.** *ACS Chem Neurosci* (2013), 4: 1278-1286.

Poster Presentations

Sliwoski, G., Stichel, J., Pedragosa, X., Weaver, C.D., Beck-Sickinger, A.G., Meiler, J.: **Discovery of novel neuropeptide Y4 modulators for the treatment of obesity: computer-aided drug design, in vitro cellular assays, and high-throughput screening**, *Vanderbilt Diabetes Day 2012*, 11/2012, Nashville, TN (USA)

Sliwoski, G., Stichel, J., Pedragosa, X., Weaver, C.D., Beck-Sickinger, A.G., Meiler, J.: **Discovery of novel neuropeptide Y4 modulators for the treatment of obesity: computer-aided drug design, in vitro cellular assays, and high-throughput screening**, *Vanderbilt Diabetes Day 2013*, 11/2013, Nashville, TN (USA)

Sliwoski, G., Stichel, J., Weaver, C., Beck-Sickinger, A., Meiler, J.: **Discovery of novel neuropeptide Y4 modulators for the treatment of obesity**, *66th Southester Regional Meeting of the American Chemical Society (SERMACS)*, 10/2014, Nashville, TN (USA)

Sliwoski, G., Schubert, M., Stichel, J., Weaver, C.D., Beck-Sickinger, A.G., Meiler, J.: **Discovery of novel small-molecule modulators of the human Y4 receptor**, *11th International NPY Meeting*, 08/2015, Leipzig (Germany)

Oral Presentations

Computational docking of pancreatic polypeptide in the Y4 receptor comparative model, *10th International NPY Meeting*, 2012, Montreal (Canada)

Comparative modeling of the NPY4 receptor with Rosetta, *Lab Seminar,* 2012, Nashville, TN (USA)

The case of the lazy receptor and other short stories, *Lab Seminar,* 2013, Nashville, TN (USA)

3D enantioselective descriptors for QSAR, *Lab Seminar,* 2012, Nashville, TN (USA)

Tutorial 4: Comparative Modeling, *Rosetta Workshop*, 2013, Nashville, TN (USA)

Tutorial 2: Comparative Modeling, *Rosetta Workshop*, 2014, Nashville, TN (USA)

Modeling MC4 with Rosetta, *Lab Seminar,* 2014, Nashville, TN (USA)

RDF: Behind the data, *Lab Seminar*, 2015, Nashville, TN (USA)

Curriculum Vitae

Personal

Surname: Sliwoski

Name: Gregory

Date of birth: 30.07.1981

Place of birth: New Hampshire, USA

Nationality: USA

Education

Since 08/2012	Ph.D. position at the Institute of Biochemistry, Leipzig University, Leipzig, Germany under supervision of Prof. Dr. Annette G. Beck-Sickinger and Prof. Dr. Jens Meiler
Since 08/2012	Research technician at Chemistry Department, Vanderbilt University, Nashville, TN, USA under supervision of Prof. Dr. Jens Meiler
06/2010 – 08/2012	Master of Science in Interdisciplinary Pharmacology, Vanderbilt University Master project in Pharmacology Department, under supervision of Prof. Dr. Jens Meiler
03/2005 – 06/2009	Senior technical research assistant, McLean Hospital, Belmont, MA, USA Molecular Pharmacology Laboratory, under supervision of Dr. Edgar Buttner
10/2001 – 06/2004	Research assistant, Carnegie Mellon University, Pittsburgh, PA, USA Center For Cognitive Brain Imaging, under supervision of Prof. Marcel Just
09/1999 – 06/2004	Bachelor of Science in Psychology and Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit versichere ich die vorliegende Dissertation selbstständig und ohne unerlaubte fremde Hilfe angefertigt zu haben. Ich habe keinen anderen als die im Literaturverzeichnis angeführten Quellen genutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden und alle Angaben, die auf mündliche Auskünfte beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche kenntlich gemacht.

Unterstützungsleistungen bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskriptes habe ich von Frau Prof. Dr. Annette G. Beck-Sickinger (Universität Leipzig) und Prof. Dr. Jens Meiler (Vanderbilt University) erhalten.

Darüber hinaus versichere ich, dass außer den genannten Personen bei der geistigen Herstellung der vorliegenden Arbeit keine weiteren Personen, insbesondere keine Promotionsberater, beteiligt waren und dass weder unmittelbar noch mittelbar geldwerte Leistungen an Dritte vergeben wurden.

Die vorliegende Arbeit ist weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt worden. Ich habe keine früheren erfolglosen Promotionsversuche unternommen.

Leipzig, den 24.08.2015

Gregory Richard Sliwoski

Acknowledgements

My foremost thank you goes to Prof. Dr. Annette G. Beck-Sickinger and Prof. Dr. Jens Meiler. Working at the collaborative interface between the two labs has provided me with a unique and enriching graduate experience. I am fortunate to be guided by mentors who are generous with their time, effective in their guidance, and brilliant in their fields. My growth as a student of science under such impressive teachers is equaled only by my growth as a person. Thanks to Dr. Beck-Sickinger and Dr. Meiler, I experienced more of science and of the world than I could have ever hoped.

Thank you to Dr. Jan Stichel. Long high throughput screening nights with logic puzzles, moving cameras, and makeshift magnet dart games to pass the time between plate reads gave me some of my fondest graduate school memories. Your generosity and great personality helped my first stays in Leipzig comfortable and enjoyable. I am grateful for your guidance but more so I am grateful to call you a friend.

Thank you to Dr. Will Lowe, to whom I owe not only a great deal of my knowledge regarding artificial intelligence and computational techniques, but much of my graduate student sanity.

Thank you to Dr. David Weaver for help with screening, cell assays, and valuable advice. Thank you to Dr. Seva Gurevich for your help with pharmacology and manuscript reviews.

Thank you to the members of the Beck-Sickinger lab who welcomed and helped me with generosity. I am glad to have known such a range of unique and wonderful people. Thank you to Xavi for being a great friend and a great coworker. Thank you to Mario, a great person and a great scientist. Your concentration response curves are unparalleled and I am fortunate to have gotten to work with you and to know you. Thank you to Rene, Sven, Rayk, Steffi, Annette K., Verena, and many more. Thank you to Anett for all your help coordinating my stays in Leipzig and showing me around.

Thank you to Axel Fischer for the help with German translations, great conversations ranging from quantum mechanics to hot water heaters, and help with more projects and failed attempts than I can list.

Thank you to members of the Meiler lab. Thank you to Liz Dong who helped me get started with Rosetta and Rocco Moretti who helped me learn its many details. Thank you to my friends and coworkers Sam and Steph, Kristian and Thuy, Mariusz, Ralf, Jan, Alex, Steven, Sten, Brittany, and Alyssa. Thank you to

Eric for helping me ignore academia and science long enough to play some good music, Joe, Tyler, Evan, Jessica, Ashley, Kate, and many more who made my free time enjoyable and fun.

Thank you to my brother and sister. Thank you Amelia for showing me just how powerful the love of a big sister can be. A heart such as yours is rare and I am grateful for the things in life you have given me that can never be explained with science. Thank you to Jeremy for giving me a friend for life. Time has tested every friendship I have ever known but brotherhood is a friendship tested by nothing. You've made me a better person. Thank you to Heather for putting up with Jeremy. Amelia Bedelia and Germy the Germ from Germany: the two greatest gifts a person can ask for.

Thank you to my mother Susan Sliwoski. Every decision was made a little easier, every mistake was made a little less severe, and every thought was made a little clearer with a mother as wonderful, as kind, and as loving as you.

Thank you to my father Richard Sliwoski. Of all the people who have contributed to this work, you have contributed the most. Thank you for demanding academic success, for helping me become a man, and for giving me the single most important piece of advice I have ever received: “Knowledge is the only thing that can never be taken away from you; it is yours forever.”

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: **Computational methods in drug discovery**

Journal: **Pharmacological Reviews, 2013 (accepted). doi: 10.1124/pr.112.007336. Print 2014.**

Authors: Gregory R. Sliwoski, Sandeep Kothiwale, Jens Meiler, and Edward Will Lowe

Gregory Sliwoski:

- contributed to manuscript outline
- wrote introduction, ligand-based drug design, ADMET, and conclusion sections
- revised structure-based drug design section

Sandeep Kothiwale:

- contributed to manuscript outline
- wrote structure-based drug design section
- revised remaining sections of manuscript

Jens Meiler:

- revised the manuscript

Edward Will Lowe:

- supervised manuscript outline and direction
- wrote subsection on feature selection
- revised the manuscript

Gregory R. Sliwoski

Sandeep Kothiwale

Jens Meiler

Edward Will Lowe

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: **BCL::EMAS – enantioselective molecular asymmetry descriptor for 3D-QSAR**

Journal: **Molecules, 2012 (accepted). Doi: 10.3390/molecules17089971**

Authors: Gregory Sliwoski, Edward Will Lowe, Mariusz Butkiewicz, and Jens Meiler

Gregory R. Sliwoski:

- designed and implimented BCL::EMAS within BioChemical Library framework
- performed and analyzed benchmark analyses
- wrote the manuscript

Edward Will Lowe:

- guided code implementation and artificial neural network model generation
- revised the manuscript

Mariusz Butkiewicz:

- created feature selection pipeline used for benchmarking descriptor performance
- compiled high-throughput screen datasets used for descriptor benchmark

Jens Meiler:

- project idea
- project funding
- project supervision
- revised the manuscript

Gregory R. Sliwoski

Edward Will Lowe

Mariusz Butkiewicz

Jens Meiler

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: **Improvements to 3D Autocorrelation Molecular Descriptors in QSAR**

***Journal:* XXX**

Authors: Gregory R. Sliwoski, Jeff Mendenhall, and Jens Meiler

Gregory R. Sliwoski:

- designed comparative descriptor benchmark pipeline
- performed descriptor benchmarks
- analyzed benchmark results
- wrote the manuscript

Jeff Mendenhall:

- implemented 3da_Smooth and 3da_Sign descriptors in BioChemical Library framework
- helped design benchmark conditions and overall strategy
- contributed to artificial neural network paramterization

Jens Meiler:

- project idea
- project supervision
- project funding
- revised the manuscript

Gregory R. Sliwoski

Jeff Mendenhall

Jens Meiler

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: **Discovery of small molecule modulators of the human Y4 receptor**

Journal: ACS Chemical Biology (submitted)

Authors: Gregory R. Sliwoski, Mario Schubert, Jan Stichel, David Weaver, Annette G. Beck-Sickinger, and Jens Meiler

Gregory R. Sliwoski:

- contributed to high-throughput screening data generation and analysis
- contributed to concentration-response curve and selectivity data generation and analysis
- performed molecular similarity searches for compound library enrichment
- co-wrote the manuscript

Mario Schubert:

- contributed to concentration-response curve and selectivity data generation and analysis
- co-wrote the manuscript
- generated manuscript figures

Jan Stichel:

- designed and planned all high-throughput screening experiments
- performed pilot high-throughput screen experiments
- contributed to all high-throughput screening follow up data generation and analysis
- generated cell lines used in study
- co-wrote the manuscript

David Weaver:

- supervised high-throughput screening design and execution
- provided resources and instruction necessary for high-throughput screening
- contributed to data analysis
- revised the manuscript

Annette G. Beck-Sickinger:

- project idea
- project funding
- project supervision
- revised the manuscript

Jens Meiler:

- project funding
- project supervision
- revised the manuscript

Gregory R. Sliwoski

Mario Schubert

Jan Stichel

David Weaver

Annette G. Beck-Sickinger

Jens Meiler

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: Modeling Ineractions of the Human Y4 Receptor and Pancreatic Polypeptide

Authors: Gregory Sliwoski

The author confirms the authorship for the aforementioned chapter.

Gregory Sliwoski

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: Modeling Interactions of the Melanocortin 4 Receptor and α-MSH

Authors: Gregory Sliwoski

The author confirms the authorship for the aforementioned chapter.

Gregory Sliwoski

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: Future Directions and Concluding Remarks

Authors: Gregory Sliwoski

The author confirms the authorship for the aforementioned chapter.

Gregory Sliwoski

Author Contribution Statement, Gregory Sliwoski

Computer aided drug discovery: descriptor improvement and application to obesity-related therapeutics

Title: Modeling Interactions of the Human Y4 Receptor and Niclosamide

Authors: Gregory Sliwoski

The author confirms the authorship for the aforementioned chapter.

Gregory Sliwoski