Multivariate Design of Molecular Docking Experiments

An Investigation of Protein-Ligand Interactions

David Andersson

Doctorial Thesis

Department of Chemistry
Umeå University
Umeå 2010

ISBN: 978-91-7459-065-4
Tryck/Printed by VMC-KBC Umeå
Umeå, Sweden 2010

Author

David Andersson

Title

Multivariate Design of Molecular Docking Experiments - An Investigation of Protein-Ligand Interactions

Abstract

To be able to make informed descicions regarding the research of new drug molecules (ligands), it is crucial to have access to information regarding the chemical interaction between the drug and its biological target (protein). Computer-based methods have a given role in drug research today and, by using methods such as molecular docking, it is possible to investigate the way in which ligands and proteins interact. Despite the acceleration in computer power experienced in the last decades many problems persist in modelling these complicated interactions. The main objective of this thesis was to investigate and improve molecular modelling methods aimed to estimate protein-ligand binding. In order to do so, we have utilised chemometric tools, *e.g.* design of experiments (DoE) and principal component analysis (PCA), in the field of molecular modelling. More specifically, molecular docking was investigated as a tool for reproduction of ligand poses in protein 3D structures and for virtual screening. Adjustable parameters in two docking software were varied using DoE and parameter settings were identified which lead to improved results. In an additional study, we explored the nature of ligand-binding cavities in proteins since they are important factors in protein-ligand interactions, especially in the prediction of the function of newly found proteins. We developed a strategy, comprising a new set of descriptors and PCA, to map proteins based on their cavity physicochemical properties. Finally, we applied our developed strategies to design a set of glycopeptides which were used to study autoimmune arthritis. A combination of docking and statistical molecular design, synthesis and biological evaluation led to new binders for two different class II MHC proteins and recognition by a panel of T-cell hybridomas. New and interesting SAR conclusions could be drawn and the results will serve as a basis for selection of peptides to include in *in vivo* studies.

Keywords

Molecular docking, chemometrics, multivariate analysis, principal component analysis, PCA, design of experiments, DoE, partial least-square projections to latent structures, PLS, scoring functions, ligand-binding cavity, major histocompatibility complex, MHC, glycopeptide, T-cell.

Sammanfattning

Läkemedel består generellt av små molekyler (ligander) vars verkan i kroppen är en effekt av dess interaktion med kroppens proteiner. För att utveckla av nya läkemedel är det viktigt att kunna bestämma på vilket sätt liganderna binder till specifika proteiner och hur stark denna bindning är. Olika beräkningsmetoder, där ibland t.ex. dockning, har utvecklats just för detta syfte. I denna avhandling har vi undersökt olika dockningsprogram och deras förmåga att återskapa ligandernas geometrier och beräkna bindningsstyrka i komplex mellan ligander och proteiner med känd 3D-struktur. Men också, och kanske viktigare, dockningsprogrammens förmåga att identifiera nya ligander till ett protein. För att planera beräkningsexperiment och för att analysera resultaten har vi använt oss av kemometriska metoder. Dessa syftar till att minimera antalet experiment som behöver genomföras utan att information förloras på vägen, samt att hantera stora datamängder via olika projektionsmetoder som underlättar tolkningen av resultaten. Våra resultat visar att val av dockningsprogram och olika inställningar i dessa har stor betydelse för vilka resultat man får. Vidare kan man med ”smart” experimentell planering finna inställningar som är optimala när det gäller att identifiera geometrin hos en ligand i olika typer av protein-ligand komplex. Vi har också utvecklat en metod för att beskriva och gruppera proteiner med avseende på de kemiska egenskaperna hos de ligand-bindande ytorna. Vi kunde visa att det är möjligt att prediktera funktionen hos ett protein med hjälp av denna beskrivning. Vi har även applicerat de nya metoderna vi utvecklat för att designa nya ligander (glykopeptider) för en typ av protein involverat i uppkomsten av sjukdomen autoimmun artrit. Med hjälp av dockning och statistisk molekyl-design konstruerade vi 20 glykopeptider. Dessa syntetiserades och utvärderades i biologiska testsystem för att bestämma deras bindningskapacitet till två typer av proteiner och deras förmåga att inducera respons hos immunförsvarsceller (T-cellshybridom). Genom denna studie kunde vi dra slutsatser kring vilka egenskaper hos glykopeptiderna som är viktiga för deras bindnigskapacitet. Detta kommer att ligga till grund för beslut kring vilka peptider som ska inkluderas i framtida vaccinationsstudier.

“Corpora non agunt nisi fixata”

“Bodies do not work when they are not bound”

Paul Ehrlich

1913

CONTENTS

LIST OF PAPERS

I. **C. David Andersson**, Elin Thysell, Anton Lindström, Max Bylesjö, Florian Raubacher and Anna Linusson. A Multivariate Approach to Investigate Docking Parameters' Effects on Docking Performance. *Journal of Chemical Information and Modeling*, **2007**; 47(4): 1673-1687.

II. **C. David Andersson**, Brian Y. Chen and Anna Linusson. Mapping of Ligand-Binding Cavities in Proteins. *Proteins: Structure, Function, and Bioinformatics*, **2010**; 78: 1408-1422.

III. **C. David Andersson**, Brian Y. Chen and Anna Linusson. Design of Target-Tailored Virtual Screening Experiments. (Submitted)

IV. Ida E. Andersson,# **C. David Andersson**,# Tsvetelina Batsalova, Balik Dzhambazov, Rikard Holmdahl, Jan Kihlberg, and Anna Linusson. Design of Glycopeptide Chemical Probes Used to Investigate Multiresponses Associated with Autoimmune Arthritis. (Submitted)

Papers I and II have been reprinted with kind permission from the publishers.

# These authors contributed equally to this work.

ABBREVIATIONS

ΔG	Change in Gibbs free energy
ΔH	Change in enthalpy
ΔS	Change in entropy
2D	Two Dimensional
3D	Three Dimensional
Å	Ångström
ACE	Angiotensin-Converting Enzyme
AChE	Acetylcholinesterase
ANOVA	Analysis of Variance
AUC	Area Under Curve
CDK2	Cyclin-Dependent Kinase2
cf.	*confer* (Latin for "compare")
CGO	Chemical Gaussian Overlay
CIA	Collagen Induced Arthritis
DoE	Design of Experiments
DOOD	Determinant-Optimal Onion Design
D-optimal	Determinant-Optimal
DUD	Directory of Useful Decoys
e.g.	*exempli gratia* (Latin for "for example")
EF	Enrichment Factor
FD	Full Factorial Design
FFD	Fractional Factorial Design
FGFr1	Fibroblast Growth Factor Receptor 1
FRED	Fast Rigid Exhaustive Docking
FXa	Coagulation Factor Xa
g	Gram
GA	Genetic Algorithm
GB	Generalized Born
Gln	Glutamine
GOLD	Genetic Optimization for Ligand Docking
i.e.	*id est* (Latin for "that is")
Ile	Isoleucine
LogP	Logarithm of the Partition coefficient

MD	Molecular Dynamics
MHC	Major Histocompatibility Complex
MLR	Multiple Linear Regression
NMR	Nuclear Magnetic Resonance spectroscopy
OPLS	Orthogonal Projections to Latent Structures
PB	Poisson-Boltzmann
PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
Phe	Phenylalanine
Plp	Piecewise Linear Potential
PLS	Projections to Latent Structures
QSAR	Quantitative Structure-Activity Relationship
RA	Rheumatoid Arthritis
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristics
SA	Surface Area
SAR	Structure-Activity Relationship
SCOP	Structural Classification of Proteins
SCREEN	Surface Cavity Recognition and Evaluation
SEK	Svenska kronor
SMD	Statistical Molecular Design
UV	Unit Variance
vdW	van der Waals
VS	Virtual Screening

1. INTRODUCTION

1.1 Molecular Interactions

The studies presented in this thesis have to a large extent dealt with describing and quantifying non-covalent bonds between molecules. Put loosely, one of the main conclusions that can be drawn from the results obtained is that the attractive and repulsive forces that govern molecular interactions are both intriguing and puzzling. Drugs are typically small molecules that interact with proteins in our bodies by binding to important areas on or inside the protein. Thereby they can inhibit the protein's function or hinder its interaction with other proteins. How is it possible for a small (drug) molecule to find its way through the chaotic cellular environment that exists within our bodies and finally end up inside a specific target protein? Obviously the small molecule needs to overcome many obstacles on the way to its intended target and charting these obstacles is beyond the scope of this thesis. However, there must clearly be some kind of attractive force that causes the two to bind to one-another. So to be able to design a molecule which is intended to bind to a specific protein, we need to be able to distinguish molecules that will bind to the protein from those that will not. The famous lock and key metaphor[1] illustrates the challenge: we must design a key that will fit the lock perfectly and which is also able to open the lock (*i.e.* produce the desired biological effect). This principle is illustrated in Figure 1a which depicts the surface of a protein (the lock) and a ligand (the key) which binds to the ligand binding site.

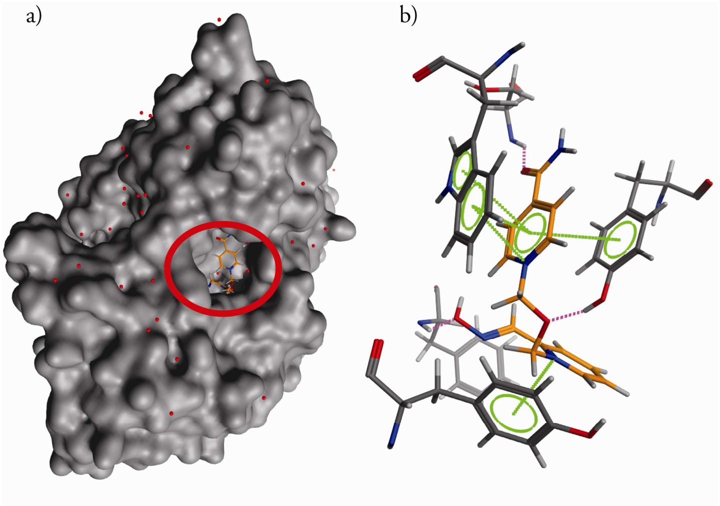


Figure 1. a) The surface of a protein (acetylcholinesterase, AChE (PDB 2gyu)) with an inhibitor (HI-6) in the ligand-binding site marked with a red circle. b) HI-6 and its non-covalent bonds with amino acids in AChE. HI-6 carbons are orange, hydrogen bonds are indicated with purple lines and π-stacking and π-cation interactions are indicated with green lines.

Binding affinity experiments, computational methods based on quantum mechanics or molecular mechanics, and crystallography have all contributed to our current understanding of the factors that affect binding between proteins and small molecules (ligands).[2] It is believed that the most important types of non-covalent bonds involved include ion-ion and ion-dipole interactions, Van der Waals (vdW) forces (Keesom forces and London forces), hydrogen bonds,[3] and π system interactions;[4] the last two of these are illustrated in Figure 1b. Common interactions found between ligands and proteins have been assembled in comprehensive libraries of interactions[2] and reviews have been written on the subject[5, 6] Even though these interactions are quite well understood and it is possible to estimate the contribution made by each type of bond to the overall ligand-protein binding affinity, calculated binding affinities often do not correlate well with measured affinities.[5, 7, 8] Obviously more factors need to be considered; one of the most important is the curious property called entropy.[9] The energy of a system can be described in terms of its Gibbs free energy, *G*. Upon binding, the ligand and the protein form a complex with a lower free energy than the additive energy of the two separate species. The change in free energy can be expressed according to

$$\Delta G^o = \Delta H^o \text{-} T\Delta S \qquad (1)$$

where ΔG^o is the change in standard free energy, ΔH^o and ΔS^o are the changes in the standard enthalpy and entropy of the system respectively, and T is the temperature in Kelvin. The formation and breaking of the above-mentioned non-covalent bonds contribute to the enthalpy change, while the change in entropic energy is largely dependent on changes in the freedom of movement of the molecules. This is where the importance of considering the environment surrounding the two molecules becomes clear. Most ligand-protein interactions take place in an environment that largely consists of water, and water molecules form loosely-organised hydration shells around the two species. All of these molecules are in constant motion and some are bound together, which is entropically unfavourable. In simple terms, the binding of a ligand to a protein causes water to be 'squeezed out' from between the interacting surfaces of the two molecules. While the conversion of ordered water molecules on the surface of a solute to free water in solution is entropically favourable, the conversion of a freely-moving protein and ligand to a single supramolecular entity is not. All of these factors contribute to the entropy change on binding. The $T\Delta S$ term in equation (1) can be of considerable magnitude, especially when the ligand and protein have complementary lipophilicc surfaces, but is generally more complicated to estimate computationally than are the enthalpic elements.[8, 10] Furthermore, there is a negative correlation between ΔH^o and ΔS^o, known as enthalpy-entropy compensation;[11] while the free energy change of ligand binding is often small, the enthalpy and entropy changes can vary widely. Therefore, small errors in the predicted values of either ΔH^o or ΔS^o can have vast effects on the calculated ΔG^o, and this must be accounted for when calculating binding affinities.[5, 8]

1.2 How to Estimate Molecular Interactions

The binding affinity between a ligand and a protein can be quantified experimentally using a biological test system (assay). For instance, changes in protein activity following the addition of a ligand can be measured. Alternatively, one can use a competitive assay, in which the ability of the ligand of interest to displace one of the protein's native ligands is measured; this approach was adopted in **Paper IV**. Affinity can also be measured by a range of other methods, including microcalorimetry.[12] Binding can be studied using nuclear magnetic resonance (NMR) spectroscopy[13] and x-ray crystallography (vide infra).[14] All these experimental methods require access to pure samples of the protein (or the protein source) and ligands, which is not always achievable. There are many computational methods for predicting the geometry of protein-ligand complexes and for estimating their binding affinity. These methods can generally be classified as belonging to one of four groups:[9] molecular docking and scoring, approximate free energy methods,[15-17] and relative-[18, 19] and absolute[20, 21] binding free energy methods.[22, 23] Of these methods, molecular docking is the least computationally demanding and also the least precise, trading accuracy for speed. Hence, it can be used to rapidly evaluate the binding of many ligands to a protein, making it particularly valuable when screening large databases of molecules in a virtual screen (VS). Scoring functions are relatively simple mathematical expressions or regressions that estimate the strength of protein-ligand interactions. Some of them attempt to predict the enthalpy and entropy of binding or only the enthalpy; others make predictions on the basis of experimental data. In this thesis, we have investigated docking strategies and scoring functions (see the segment on scoring functions) and their ability to predict ligand binding poses and ranks on the basis of their predicted binding capabilities.

1.3 The Molecule’s 3D Structure

Modern drug design is facilitated by the knowledge of the 3D structures of ligands, proteins, and protein-ligand complexes (Figure 1a). Individual molecules are much too small to be identified by methods that rely on light in the visible spectrum, even when they are as large as proteins, which consist of thousands of atoms. Therefore, it is not possible to *see* a molecule. That being the case, it is remarkable that many (or at least, many chemists) have quite a clear picture of what they look like. To be able to see molecules, we need to use radiation of wavelengths other than those of visible light, such as x-rays, as in the case of x-ray crystallography,[14] or radio waves, as in the case of NMR spectroscopy.[13] These are the two most commonly-used methods for deriving the 3D-structures of proteins and smaller molecules. 3D structures of proteins can also be derived by computational methods such as comparative modelling.[24] The structures derived via this method are based on previously determined structures of proteins having similar sequences to the protein of interest; a protein whose 3D structure is unknown is modelled or "threaded" upon a structure with high sequence similarity, and a hypothetical 3D structure is derived after refinement and energy minimization. We adopted this approach to obtain a 3D structure for the class II major histocompatibility complex (MHC) A^q protein, which was then used in the structure-based design of glycopeptides (**Paper IV**). Typically, newly solved 3D protein structures are deposited in the publically available RCSB protein data bank,[25] which contains roughly 65300 such structures as of the time of writing. All of the x-ray crystallographic protein structures (x-ray structures) investigated in these studies were obtained either from the protein data bank or from one of its subsidiary collections (specifically, the PDBbind database[26, 27] and the Directory of useful decoys, DUD[28]).

It is important to realize that x-ray structures are models based on experimental data, and that the "quality" of these models is sensitive to the experimental conditions and computational methods used. The quality of a 3D structure has implications, especially in structure-based design, where structural details such as the lengths of specific protein-ligand bonds are used to form conclusions and to design new and improved ligands. One measure of the 'quality' of an x-ray structure is its resolution, reported in Ångström (Å). For instance, in an x-ray structure with a resolution of 2.5 Å the standard deviation in the atomic coordinates may be as high as 0.4 Å.[5, 29] Considering that the distance between heavy (non-hydrogen) atoms in a hydrogen bond is typically 2.4-2.8 Å[3] it is evident that analysis of structures in which the standard deviation in the positions of heavy atoms exceeds 0.4 Å may result in faulty conclusions concerning the nature of the such interactions. Furthermore, at a resolution of > 2.5 Å the model details in the structure are more subjective and more dependent on the modelling strategy.[30] In this work, we have only used x-ray structures with a resolution of more than 2.5 Å.

1.4 Designing Drug Molecules

A plethora of computational methods have proved useful in modern drug design.[31, 32] They are mostly used for calculating the physicochemical properties of molecules, estimating binding affinities (for instance, by calculating free energies of binding), predicting binding poses (*e.g.* docking), molecular dynamics (MD) simulations (simulations of the movement of molecules),[33, 34] and different search methods (such as pharmacophore matching[35] and scaffold hopping[36, 37]). It has been estimated that it takes ten years and a billion dollars (~10 billion SEK) to develop a drug and bring it to market[38] so the drug development process has much to gain in terms of time, costs and innovation by using computational methods. Methods like molecular docking, which is the primary topic of this thesis, can be used to identify potential ligands for a specific protein target (lead generation), and to assist chemists in identifying chemical modifications that might improve a molecule's pharmaceutical properties (lead optimization), reducing experimental costs. Furthermore, the value of experimental planning and statistical molecular design (SMD)[39-43] in drug evolution should not be underestimated, as we have shown previously in the design of antibacterial Type III secretion inhibitors[44] and as is further demonstrated in this thesis.

1.4.1 Ligand-Based Design

In the absence of a 3D structure for the protein of interest, it is possible to rationally design modified ligands by studying the structure of ligands known to produce the desired biological response. The *design* of molecules is a broad concept. Generally, any non-random structural modification of a molecule is considered to be designed. This thesis employs a more stringent definition of 'design' and uses more specific terminology when discussing different design strategies. For instance, **Paper IV** describes a ligand-based approach to design, in which SMD was used to construct a library of peptides composed of a selection of physicochemically-diverse amino acids; the amino acids employed were chosen on the basis of a statistical experimental design. The strengths of SMD are discussed in more detail below. Alternative ligand-based design strategies include similarity search methodologies,[45] which identify molecules having similar structures (*e.g.* 2D fingerprints[46] or 3D pharmacophores)[35] to known bioactive ligands, and scaffold hopping, which focuses on the exchange of individual substructures of bioactive molecules for similar fragments.[36, 37]

1.4.2 Structure-Based Design

If a 3D structure of the protein of interest is available, preferably complexed with a ligand, it is possible to perform a structure-based design of new ligands. This design strategy involves the rational modification of a ligand on the basis of the protein-ligand interactions revealed in the 3D structure. This can be done by visual inspection of the 3D structure or by the analysis of protein-ligand interactions identified within the structure by computational methods such as molecular docking. Alternatively, it is possible to apply fragment-based[47, 48] and *de novo* design[49, 50] in which new ligands are designed by connecting fragments that bind to specific residues within the protein or by 'growing' a new ligand within the active site, respectively. An important and limiting aspect of structure-based design is that it does not generally consider the flexibility of proteins, which is perhaps natural since the designs are based on rigid protein-ligand complexes. Protein flexibility (or the lack thereof) is discussed below; Teague has written an excellent review of the literature on the impact of flexibility on drug design.[51]

1.5 Molecular Docking

One method for exploring the interactions between a ligand and a protein is to synthesize the ligand, co-crystallize it with the protein and then try to obtain an x-ray structure of the complex. Although both synthesis and crystallography can sometimes be quite unpredictable and time-consuming, the method may be viable for small collections of ligands. If synthesis or crystallization fails, or if the aim is to screen many ligands for binding to the protein, computational molecular docking is often the method of first choice, and has become popular within both academia and industry.[52] Furthermore, docking can be valuable when forming hypotheses regarding the way a ligand binds to the protein, or for modelling parts of the ligand whose structure or conformation when bound have not been successfully determined by crystallography. More than 60 docking programs have been reported, of which roughly 10 are widely used.[53]

Docking requires a 3D structure of the protein as input. Typically, the software will generate 3D conformations of the ligands and optimize their interactions with the protein by computing the binding affinity (scoring) between the two. In most docking programs used today, the ligand is treated as a flexible structure but the conformation of the protein is treated as being (mostly) rigid, and water molecules are typically not considered at all. Obviously, both of these approximations constitute major simplifications of the real environment in which ligands and proteins interact. Still they are useful because of the immense amount of computation that would be necessary to accurately model the effects of water and protein flexibility – imagine the difficulty of modelling a lock and key that are constantly changing shape, in aqueous solution, and trying to measure the interactions between the two! However, these simplifications are thought to be the two most important reasons why docking fails to correctly predict the affinity of a ligand for a protein, and the pose the ligand will adopt on binding (see segment on scoring),[54] and this failure has plagued docking since its birth in the 1980s. In 2006, Leach *et al.* stated that docking had reached a plateau in its development and was waiting for a breakthrough.[55] Gratifyingly, reports of numerous studies seeking to address the problem of protein flexibility have since appeared, and a range of different methods including ensemble docking,[56-60] coarse graining,[61] flexibility trees,[62] elastic potential grids,[63] and genetic algorithms have been investigated.[64] Methods for dealing with protein flexibility have been reviewed by Alonso et al.,[65] B-Rao et al.,[66] and Henzler and Rarey.[67] The inclusion of structural water has also been intensively studied, and recent studies suggest that this does improve the accuracy of docking.[68-70]

The accuracy of docking software packages is often tested by so-called redocking experiments. These involve docking one or a set of ligands back into the binding site of the native 3D structure of the protein; the docking is judged to be successful if the software is able to reproduce the experimentally-observed ligand pose. Redocking results are commonly evaluated by calculating the root mean square deviations (RMSD) between the native ligand conformation (as observed in the x-ray structure) and the ligand conformation suggested by the docking software (docking poses). RMSD is a measure of the average deviation in the positions of the heavy atoms of the ligand between the two complexes. The native pose is typically judged to have been "successfully" reproduced if the RMSD is below 2.0 Å,[71-73] although such fixed limits should be treated with caution in some cases.

Despite (and perhaps because of) the many drawbacks of docking, we and our fellow scientists continue to further develop and investigate techniques to improve upon it, and the many reported success stories concerning the use of docking are a major driving force in this development.[74-80] Many of these successes originate from the field of virtual screening (see below) which has identified many new and unexpected molecules as ligands for various proteins, or from lead optimization studies aiming to rationalize the binding of designed molecules.

1.6 Scoring Functions

In docking, the binding affinity, or rather the complementarity, between the ligand and protein is assessed by scoring functions. These can generally be classified into one of three categories: empirical, force field-based and knowledge-based.[81] Empirical scoring functions are the most common;[53] they estimate binding affinities by dividing ΔG^o into scalable contributions from individual types of protein-ligand interactions such as hydrophobic effects, hydrogen bonding, and constraints upon movement imposed by binding. The equations involved are exemplified by a simplified version of the Chemscore scoring function:[82]

$$\Delta G_{bind} = \Delta G_{H\text{-}bond} + \Delta G_{metal} + \Delta G_{lipo} + \Delta G_{rotHrot} + \Delta G_{0} \qquad (2)$$

where ΔG_{bind} is the estimated free energy of binding, and the remaining terms are contributions to ΔG_{bind} from hydrogen bonds ($\Delta G_{H\text{-}bond}$), metal interactions (ΔG_{metal}), lipophilic interactions (ΔG_{lipo}), frozen rotatable ligand bonds ($\Delta G_{rot}H_{rot}$) and non-specific interactions (ΔG_{0}). The coefficients for the individual ΔG terms were derived using multiple linear regressions (MLR). Other empirical scoring functions used in some of the studies discussed in in this thesis are Piecewise linear potential (Plp)[54] and Screenscore.[83] One of the major drawbacks of functions in this category is that their predictive ability is limited by the scope and quality of the 'training set' of complexes used when developing the function.

Force field based scoring functions, represented in our studies by Goldscore[71], estimate binding affinities as the sum of electrostatic and vdW interaction energies (which are often modelled using Lennard-Jones potentials)[84] calculated using molecular mechanics force fields. Atoms are treated as single particles and the force fields contain information regarding the nature and behaviour of different atoms, including their vdW area and partial charges. The force fields are parameterized on the basis of experiments and quantum mechanics calculations. These scoring functions tend to overemphasise polar interactions, although these effects can be compensated for to some extent.[85]

The third category of scoring functions includes those based on the knowledge gained from the ever increasing number of protein-ligand complexes, hence the name knowledge-based scoring functions.[86] These functions rely on pairwise atom potentials calculated from statistical analyses of bonds that are frequently observed between ligand and protein atoms. The final score is then calculated as the sum of all the pairwise interactions between the ligand and protein (within a defined distance cut off). A drawback of this method is that some rare types of interactions (*e.g.* interactions with halogens) may be less well parameterized.

Other types of scoring functions that do not belong to any of these three classes have also been developed. For instance, Gaussian scoring functions,[87] represented by Chemgauss3, Shapegauss, and chemical Gaussian overlay (CGO) in this thesis. These functions use smooth Gaussian functions to represent atoms and to evaluate steric clashes and beneficial interatomic interactions;[87] these functions may incorporate additional terms to describe hydrogen bonding, desolvation, and metal interactions (Chemgauss3) or to account for overlap with a bound ligand (CGO). Finally, rescoring using a more computationally demanding physics-based approach, such as molecular mechanics Poisson-Boltzmann/Generalized Born surface area (MM-PB/GB-SA)[15, 88, 89] has become more popular in recent years because of the promising results.[16, 17, 90-92] The increase in available computational power allow these powerful but computationally demanding calculations to be performed in reasonable timeframes.

One may ask whether this lack of correlation between calculated and measured affinities is due to the failure of the models to account for protein flexibility,[93] to the failure to account for the presence of water,[68] or simply to poor descriptions of the interactions between the ligand and the protein. In all likelihood, the answer is that all of these factors contribute to some extent; it may be the case that scoring is more a measure of the complementarity between a ligand and a protein rather than an estimate of affinity. Nevertheless, scoring functions have been surprisingly accurate in many cases, especially in ranking active compounds in VS[75] and in the identification of accurate binding poses[7, 73, 94, 95] and they are continuously being refined and improved. The development of target-specific or "tailor-made" scoring functions has become increasingly popular,[57, 96-98] and the development and application of methods for selecting an appropriate scoring function are described in this thesis.

1.7 Virtual Screening

The aim of VS is to find molecules (hits) not previously identified as ligands (actives) for a specific protein. In drug discovery, these new ligands should preferably be structurally and/or physicochemically different from the already-known ligands and also quite small in size. This is because hits from the VS will go through a structural evolution as they are turned into potentially viable drugs, and smaller molecules can be more extensively modified and can have more additional chemical groups incorporated before they reach "non-drug like" sizes (*i.e.* molecular weight > 500 g/mol).[99] VS has proved to be a very useful techniques in lead generation[75, 76, 100] and it has been estimated that new ligands have been found for more than 50 different proteins.[100] Several successful docking-based VS campaigns have been reported,[74, 77, 101-103] many of which have been reviewed elsewhere.[104]

If a VS tool is able to assign high ranks to active compounds from a library of potential ligands, *i.e.* more actives can be found among the top ranked molecules than among the low-ranked molecules, it is said to "enrich" the library. Moitessier *et al.*[53] claim that "in virtually every case, it is worth running a VS to guide the development of a focused library as enrichment is likely to be obtained". Essentially, this means that it is desirable to use every bit of information we have regarding the target protein and its possible interactions with small molecules. We adopted this concept in **Paper IV**, in which a VS was performed against a comparative model of a protein.

A range of different methods can be employed in VS, including ligand-based and structure-based methods.[105] Ligand-based methods include strategies where molecules are compared to 2-dimensional representations of known ligands (*e.g.* topology fingerprints/descriptors, as reviewed by Hert *et al.*)[106] or to 3D representations. Examples of the 3D approach include ligand pharmacophore modelling[35, 107] and screening based on ligand shape.[108] Docking is the most commonly-used structure-based method, although other methods such as protein-ligand complex pharmacophores have been developed.[35, 109-112] Comparisons of ligand- and structure-based methods have not given consistent answers as to which are the most reliable. In some cases ligand-based methods work best;[113] in others, both give comparable results.[114] It has been suggested that combining the different methods[114, 115] or applying post-VS pharmacophore filters[116, 117] might be useful in achieving good VS results.

Choosing an appropriate VS tool is challenging because the tool strongly affects the outcome of a VS. Furthermore, the quality of the 3D ligand and/or protein structures and the scoring functions employed also influence the outcome. Different VS tools can be evaluated using what we prefer to call “simulated VS”, in which a VS is performed against a protein using a database of "decoy" molecules that are presumed to be inactive, but which has been spiked with known active ligands. If the tool is able to enrich this database, *i.e.* if it identifies and assigns high ranks to the active ligands, it may be suitable for use in a real VS focusing on that protein. Valuable guides for setting up VS experiments have been written by Kirchmair *et al.*[76] and Nicholls.[118] The performance of a tool in simulated VS can be evaluated using the so-called Enrichment Factor (EF), which provides a measure of the enrichment achieved, or by using the Receiver Operating Characteristics Area Under Curve (ROC-AUC).[119, 120] The EF focuses on ligands in an arbitrarily chosen high percentile of the ranked database, and its value is dependent on the ratio of actives to decoys in the database. The sensitivity to the precise percentile chosen and the active/decoy ratio can makes comparisons between reported EF values difficult. All of the simulated VS experiments performed in our study (**Paper III**) had very similar active/decoy ratios; for comparative purposes, we examined relative (normalized) EF values[121]. ROC-AUC is a more general measure of enrichment than EF in the sense that it measures enrichment in the whole database, not just in some high percentile. ROC-AUC compares the ratio of correctly identified actives to total identified actives (*i.e.* real actives plus false positives) and the ratio of correctly identified decoys to total identified decoys (*i.e.* real decoys plus false negatives). Using this criterion, good enrichment has been achieved if the ratio of actives compared to that of decoys is high in the beginning of the ranked database and the ROC-AUC value is close to 1. A ROC-AUC of 0.5 indicates a random distribution of actives in the ranked database and a ROC-AUC below 0.5 indicates a negative enrichment. An attempt to shed some light on the effectiveness of different tools and scoring functions in VS using a small set of diverse proteins is presented in **Paper III**.

1.8 Chemometrics in Drug Discovery

Multivariate data analysis and Design of Experiments (DoE), which are the two pinnacles of chemometrics, have proved to be very useful in drug discovery and development.[122-124] We view chemometrics as a concept, or a tool box, which contains tools that can be used to effectively plan and evaluate experiments within almost any area of research, *i.e.* despite what its name might imply, its applicability is not restricted to chemistry. Both computational and “wet” experiments usually generate a lot of information but this information is encoded in some sort of data, such as spectroscopic read-outs from analyses of samples or the calculated molecular properties of a set of ligands. Information in the data that is relevant to the questions at hand is often intertwined with non-relevant information; in the case of spectroscopic data, this may be due to fluctuations in the measuring equipment (noise), while in the case of ligand properties, it may be due to the calculation of non-relevant properties. Importantly, the identification of relevant information is not just something to be done using multivariate methods after the experiment has been conducted: one can use the statistical method for experimental design when *planning* the experiments to maximise the amount of relevant information in the data gathered.[125]

1.9 Chemometric Methods

Variation is an important concept in chemometrics. Using a collection of molecules as an example, the collection will exhibit variation if its members have a range of different molecular properties. Many of the methods applied in chemometrics, such as principal component analysis (PCA)[126-128] and partial least-square projections to latent structures (PLS)[129, 130] are said to "extract variation" in a dataset, but what does this really mean? Molecular features can be described in various ways, for example using calculated descriptors or spectroscopic data. Such data are referred to as descriptor data. By quantifying or extracting the main variation in the descriptor data we gain information embedded in the descriptor data concerning the similarities and differences between the molecules. In essence, variation is information, although it can be irrelevant or non-systematic and hence hard to model and interpret. In the simplest case, variation can be statistically verified by calculation of the *variance*, and one can thereby determine whether one molecule is statistically different from another based on the properties by which they are described. Variability can be extracted by multivariate methods such as factor analysis[128] and PCA, where the aim is to identify relationships and patterns among the observations (*e.g.* molecules), or by PLS and MLR which make it possible to identify variability among observations which is connected to a response. PCA and PLS were used extensively used in this work and are described in the following paragraphs.

1.9.1 Design of Experiments

For a simple example of the value of DoE, consider a one-step synthesis of a molecule where you want to optimize the yield of the product. The choice of solvent, reaction temperature and catalyst are all *factors* influencing the yield and it may seem intuitive to start by testing different solvents to find the optimal one, followed by identifying the optimal temperature and then the optimal catalyst. However, this approach is flawed in that it relies on the assumption that all of the factors are uncorrelated, *i.e.* (for example) that the effect of the catalyst is independent of solvent and that the solubility of the reactants in a certain solvent is independent of the reaction temperature. Obviously, this is generally not the case.

The existence of dependencies between the factors that influence the outcome of experiments is common in all fields of research, including purely computational fields (for example, the optimal values of the parameters in the docking software investigated in this thesis are interdependent). By applying DoE, in which several factors are investigated at the same time, it is possible to determine how important individual factors are and to identify combination effects (dependencies) involving multiple factors. DoE relies on the use of experiments in which every factor is tested at two or more levels (*e.g.* two or more different temperatures) in every possible combination, at least in the case of full factorial designs (FD). Alternatively (and more practically), one may conduct a subset of experiments that preserve statistical balance, for example by using a fractional factorial design (FFD),[125] or D-optimal design.[131, 132] One problem with DoE, which is especially pronounced for "wet" experiments, is that the number of experiments required to elucidate the effects of all of the factors may be too great to be practical. However, elucidating effects without DoE may be even more impractical, and might lead one to perform numerous experiments that generate little or no new information. To avoid this, one can use fractional designs when there are many factors to be investigated, although these have the drawback that it may be harder to identify combination effects.

DoE has proved useful in both optimization and experimental planning within drug discovery,[123] and has been further extended to the design of molecules via SMD;[43, 133, 134] in this case, the chemical features of molecules are the factors which are explored. This is of particular interest in drug design, where the aim is to identify (quantitative) structure-activity relationships ((Q)SAR), *i.e.* to identify relationships between the structural features of a molecule and its biological effects (*e.g.* inhibition of enzymes or antibacterial properties). In the 1990s, SMD was primarily used for the design of combinatorial libraries;[133, 135, 136] this thesis focuses on our efforts to extend the applicability of SMD to encompass the design of libraries of small molecules[41, 44]and peptides.[42, 137] The main advantage of SMD is that it generates a set of molecules that exhibits variation in the factors (*i.e.* the chemical features) thought to be important in generating biological responses. This in turn means that a range of biological responses will be observed, which is necessary for statistically-valid conclusions to be drawn from the SAR.

DoE can be applied to introduce, or ensure, variation in experiments and independence between factors influencing the outcome of the experiments. Factorial and D-optimal designs were employed in several of the studies described in this thesis. Prior to a DoE, it is common practise to select the factors one would like to investigate; in our case, these were changeable parameters in various docking programs. Next, the span of these factors' values and the number of levels of each factor that will be investigated must be determined. Finally, in a FD, one generates an exhaustive list of all possible combinations of the factors and their respective levels, generating *N* experiments, where $N = x^k$, x is the number of factor levels, and k is the number of factors. For example, in **Paper III**, three factors were investigated at two levels (*e.g.* the *clash scale* parameter was set to either 0.25 or 0.75) giving rise to $2^3 = 8$ experiments (Figure 2).

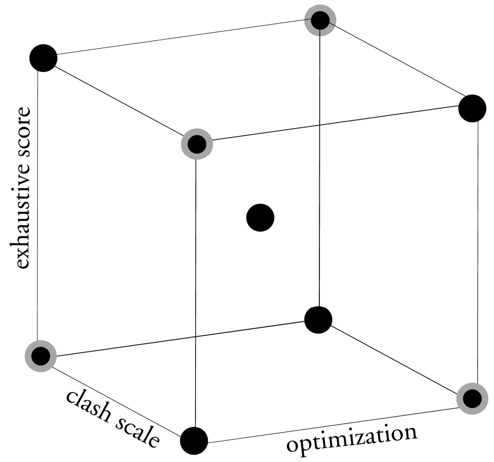


Figure 2. A graphical representation of a 2^3 full factorial design with eight experiments and one central point experiment. Factors are indicated in the axis and each factor is varied at a high and low setting. Points with gray boarders correspond to a fractional factorial design.

It is possible to reduce the number of experiments via FFD which creates a subset of the experiments used in a FD according to $N = x^{k-n}$ where x is the number factor levels, k is the number of factors, and n is the reduction factor. Hence, in the example from **Paper III**, the reduced design consisted of the corners in Figure 2 highlighted with red borders. In addition it is important to include centre point experiments, to be able to identify non-linear relationships between factors and responses in subsequent regression modelling. Preferably, three centre points should be added to be able to determine the experiment reproducibility.

Other DoE strategies applied in this thesis include space-filling designs[138] and D-optimal designs. The aim of both these design strategies is generally to select a predefined number of objects in a multi-dimensional space (*e.g.* a matrix consisting of molecules described by several physicochemical descriptors) such that the objects span the space as well as possible. The procedure for generating space-filling designs is iterative; the goal is to select an evenly-distributed subset of the x^k experiments from an FD design by maximizing the minimum Euclidean distance between the selected objects. In D-optimal designs, the aim is to select a subset of objects that span the space "D-optimally". Geometrically, this means that the selected subset should span the greatest possible volume of space. Obviously, it is possible that several subsets of different objects may span the same volume, and so the subsets are further distinguished by the designs' condition number (a measure of the sphericity of the space spanned by the subset) The D-optimal onion design (DOOD) divide the space into layers and a D-optimal design is performed in each layer, leading to a balanced selection throughout the space.[40] The benefit of these three design types is that the user can impose restrictions on the selected subsets and specify the number of experiments or selected objects.

1.9.2 Principal Component Analysis

An important tool in the chemometrics toolbox is the unsupervised multivariate modelling method known as PCA.[126-128] Data derived from chemical, biological and computational experiments can be very information rich, especially if DoE has been employed. PCA is a data analysis tool that extracts the main variation in the data, reduces its complexity, and allows the visualisation of data structures, simplifying the interpretation of the data. For example, molecules can be described by physicochemical descriptors, which may be determined experimentally or computationally. These descriptors include properties such as molecular weight, volume, hydrophobicity (LogP), and charges. A set of molecules may be described by hundreds of these types of descriptors (as was the case in **Papers I**, **II** and **IV**) in an effort to create a unique physicochemical description of each molecule in the data matrix and thence to relate the biological response generated by the molecules to their physicochemical properties. Simply looking at a descriptor data matrix containing hundreds of dimensions does not tell us how and to what extent the molecules are similar or different (*i.e.* how they vary), in which physicochemical features these differences and similarities are most pronounced, or to what extent the variables are correlated. PCA can extract the main variation (the *principal properties* in which the molecules are most diverse) in a data matrix by calculation of principal components (PCs) as shown in Figure 3.

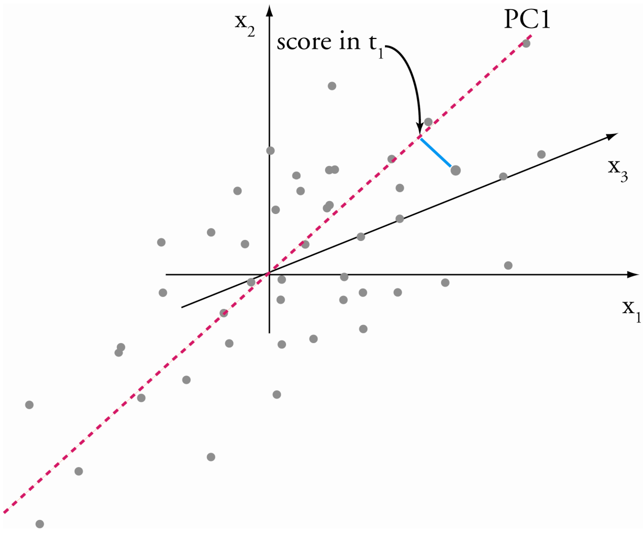


Figure 3. A geometric representation of the extraction of the first PC in PCA exemplified by a matrix of observations (gray dots) with three variables (x_1, x_2 and x_3). The red dotted line corresponds to an eigenvector (PC1) placed in the direction of the main variation in the data. Each observation is projected orthogonally (blue line) down onto the new vector and receives a new value (score-value in score-vector $\mathbf{t}_1$).

The PCs are eigenvectors; the first PC ($\mathbf{t}_1$) is aligned in the direction of the bulk of the variation in the descriptor matrix. Each object is then assigned a score value along $\mathbf{t}_1$ by projecting its old position in the descriptor space onto the new PC

(Figure 3).The contribution of a specific descriptor to the orientation of the new PC is described by its loading value, *p*. These PCs gives rise to a new decomposition matrix according to:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \ldots + \mathbf{t}_A\mathbf{p}_A' + \mathbf{E} \quad (3)$$

where **T** is the score vector matrix, **P′** is the transposed loading vector matrix, **E** is the residual, **t** is a score vector, **p′** is a loading vector and *A* is the number of extracted PCs. In our experience, the number of PC extracted from matrices of molecular descriptors ranges from three to seven, with each PC corresponding to one of the principal properties of the molecules. The main variation (PC 1) usually separate the molecules based on size which is a property that is captured in size-related descriptors such as molecular weight and volume. The interrelations between the molecules can be identified by studying the score plots and explained by studying the loading plots. Prior to modelling, it is common practise to scale to unit variance (UV-scale) and centre the data; this is particularly important when the variables in X are of different magnitudes. This thesis discusses the use of PCA in extracting the principal properties of ligands (**Papers I** and **II**), ligand-binding cavities in proteins (**Papers II** and **III**), and in separating amino acids on the basis of their principal properties and docking scores (**Paper IV**).

1.9.3 Response Modelling

A response is a numerical description of the outcome of an experiment such as the yield of a reaction, a biological effect, or the outcome of a docking experiment, and is typically stored in a response matrix, **Y**. The relationship between the response and the factors influencing the response (*e.g.* features of a ligand that affect protein-ligand affinity) is commonly estimated by a regression model; MLR[125] or PLS[129, 130] are often employed for this purpose in chemometrics. These methods correlate matrix X, which contains experimental data with matrix Y, which contains response data, via linear regression according to:

$$y_i = \sum_{k=1}^{K} x_{ik} b_k + f_i \quad , \quad (4)$$

where y_i is the *i*th response, x_{ik} is the *i*th experiment/molecule described by $k = 1 \ldots K$ factors, b_k is the model coefficient for each factor *k*, and f_i is the residual of the *i*th response. Correlation between matrices X and Y is observed if there is common or shared variance (covariance) between the two; PLS extracts this common variation. Specifically, PLS determines the inner relation (Figure 4) by connecting the decomposition matrix of X and Y. Hence the first PLS-component (analogous to the first PC in PCA) is a line in X-space and a line in Y-space that approximates each data-point in X and Y *and* that provides a good correlation between the scores in $\mathbf{t}_1$ and $\mathbf{u}_1$ (Figure 4). The inner relation can thus be used to predict the **Y**-values that will be associated with specific new observations in **X**. Unlike MLR, PLS can be used to analyse numerous **X**-variables, which may be correlated and noisy. An additional benefit of PLS is that the method can model several responses simultaneously.

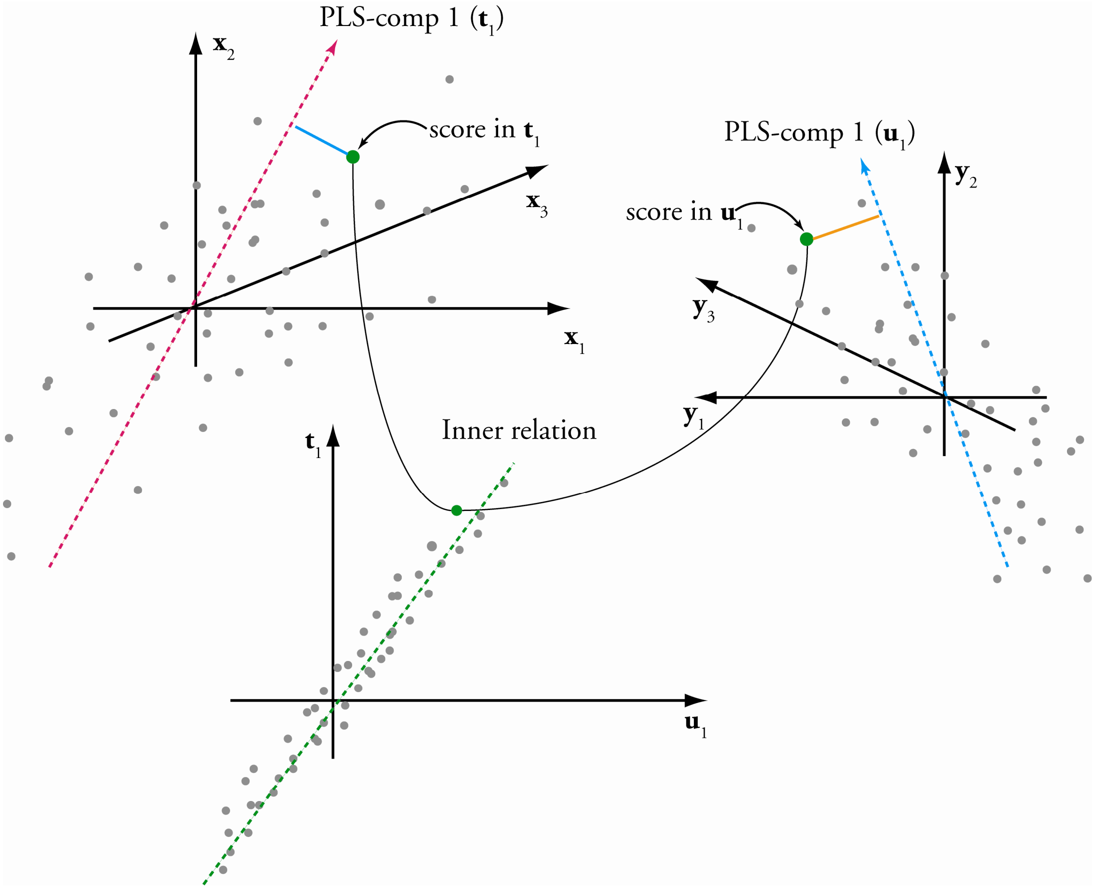


Figure 4. A geometric representation of the extraction of the first PLS-component in PLS exemplified by a matrix of observations (gray dots) with three variables ($\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$) and three responses ($\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$). The first PLS-component consists of a one score-vector ($\mathbf{t}_1$) in variable space and one in response-space ($\mathbf{u}_1$), and these are oriented so that the correlation between $\mathbf{t}_1$ and $\mathbf{u}_1$ is optimised.

As previously mentioned, the aim of DoE is to introduce variation in, and independence between, factors which are believed to have an effect on a response. Since DoE commonly results in the generation of a dataset with a broad variety of observed responses, the regression coefficients (see equation 4) can be determined with increased certainty. The benefit of having a broad variation in the responses is shown schematically in Figure 5. In Figure 5a, there is little variation in the factor x_1, which gives rise to little variation in the response y. This leads to uncertainty about the nature of the regression line and the coefficient (*i.e.* the slope of the line).

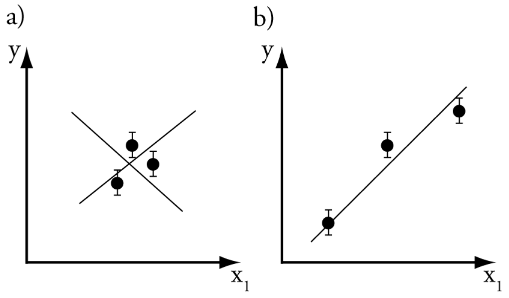


Figure 5. Schematic picture of the benefit of a large variation in variable x1 and response y. a) A small variation in x1 typically led to a small variation in y making it hard to verify the relationship between the two leading to uncertainty in the regression coefficient. b) A large variation in x1 and y will lead to a more well-determined coefficient.

In Figure 5b, the variation in the response data is greater, and so the coefficient is more precisely determined. Furthermore, predictions regarding the response associated with new molecules/experiments are likely to be more reliable since the range of the factors is larger and there is less risk of unreasonable extrapolation. PLS is somewhat restricted to the modelling of linear relationships between variables and responses; if the relationship is non-linear, PLS may perform less well in accurately correlating **X** with **Y**. Provided that a sufficiently large number of experiments has been run, non-linearity in this relationship can be identified and compensated for to some extent by the introduction of interaction or squared terms of the original variables. Non-linear relationships may also be modelled by methods such as support vector regression machines,[139] or neural networks.[140, 141] This thesis discusses the use of response modelling using PLS in the analysis of the relationship between docking parameters and docking performance (**Papers I** and **III**) and in QSAR modelling of peptides that bind to MHC proteins (**Paper IV**).

1.9.4 Validation of Multivariate Models

Validation is a very important aspect of the modelling process, and we have strived to validate our PCA, PLS and design models in a careful and transparent manner. The goodness of fit (R^2) is a valuable measure that reveals the percentage of the original variation in **X** (or in **Y**) that is explained by the model. Another measure of a model's quality is its cross validation, Q^2, which provides an estimate of the model's internal predictive capability, *i.e.* its ability to predict the data from which it was built.[127, 142] In practice, this is achieved by leaving out one or a subset of observations and rebuilding the model based on the remaining data. The new model is then used to predict the values of the variables from the excluded observations; an estimation error is calculated from the difference between these predictions and the true values. Both R^2 and Q^2 range between 0 and 1, where a value of 1 indicates a perfect model fit and internal prediction capability (which, if encountered in real life, would cause suspicion). Analysis of Variance (ANOVA) can be used to determine model significance and lack of fit by analysis of residual variation and replicate errors. A model's ability to predict the properties of objects or responses can be validated using external test-sets. For instance, new molecules can be created based on conclusions drawn from a QSAR model. Preferably, the test set (molecules) should not have been used in building the model, and should only be introduced when testing the finished model. In addition, a model builder can identify outliers (strongly deviating objects) using measures such as the distance to model in **X** (DModX) and Hotelling statistics.[143, 144] PLS models can be validated by performing permutation experiments.[145, 146] In these experiments, the response matrix **Y** is scrambled and new models are created using these distorted matrices. Large values of R^2 and Q^2 for the permutated models would indicate that the original model lacks significance.

2. SCOPE OF THE THESIS

The main objective of this thesis was to investigate and improve molecular modelling methods for the assessment of protein-ligand interactions. In general, all of the studies performed involved the use of multivariate methods such as statistical design, PCA, and PLS when setting up experiments and interpreting results. More specifically, molecular docking was investigated as a tool for the reproduction of ligand poses in protein structures (**Papers I** and **IV**) and for virtual screening (**Papers III** and **IV**). Statistical design was used to vary and optimise the values of adjustable parameters in the docking software that was used, resulting in the identification of values that give improved results. The nature of ligand-binding cavities, which play crucial roles in protein-ligand interactions, was investigated in **Paper II**. Similarities and differences in the properties of 239 different protein cavities were evaluated by calculating a set of physicochemical descriptors for each one. In addition, the biological function of a set of proteins structurally unrelated to those studied was correctly predicted. The strategies developed in **Papers I** and **II** were applied in **Paper III** to select a set of physicochemically dissimilar proteins which were used in a virtual screen with various docking parameters. This resulted in the identification of scoring functions that are likely to be useful (and some which are not so useful) in virtual screens against these proteins. Finally, the strategies developed in **Papers I** and **III** were applied in the design of a set of glycopeptides which were used to study autoimmune arthritis (**Paper IV**). We designed 20 glycopeptides (using docking and SMD) which were synthesized and biologically evaluated both as ligands for two different class II MHC proteins and in terms of their recognition by a panel of T-cell hybridomas. New and interesting SAR conclusions regarding the binding preferences of A^q and DR4 were drawn, and the T-cell activation results will serve as the basis for the selection of a set of glycopeptides for *in vivo* studies.

3. AN INSIGHT INTO DOCKING (Paper I)

3.1 Factors Influencing Docking

Setting up a docking experiment is, in practise, quite easy nowadays. The development of intuitive graphical interfaces has made the software more accessible and user friendly. Nevertheless, docking software manuals tell us very little about how to interpret the results we get from docking or about the limitations of the software. Consequently, users may be disappointed when software does not live up to their expectations and may regrettably even end up dismissing docking as a valuable technique altogether. In some cases docking may not be the best choice for tackling the problem at hand and indeed, many factors influence the outcome of a docking experiment. Fortunately there are measures that can be taken in order to elucidate how some of these factors affect results and how to control them. As mentioned in the introduction, these factors include the representation of the protein (*i.e.* the quality of the 3D structure and how well it represents reality), the representation of the ligand, the docking software, and the scoring function used. In **Paper I** we investigated the way in which one of these factors, namely the values of the user-specified parameters in the docking program, can influence the outcome. Although programs have vendor-specified default values for these parameters, these defaults will not necessarily be optimal for docking any specific protein-ligand complex.

3.2 A Study of the "Tunability" of Docking Software

Paper I describes a DoE-based approach to the investigation and optimisation of variable parameters in the docking programs FRED[147] and GOLD.[71, 148, 149] Both programs are commonly used in drug research but they differ significantly in the way they address the problem of docking. Their primary differences have to do with the mechanisms they use to generate ligand conformations: FRED relies on rigid docking of pre-generated ligand conformations while GOLD uses a genetic algorithm (GA) to generate ligand conformations in the protein's ligand-binding site.

Five parameters in FRED and 10 in GOLD were subjected to a factorial and D-optimal design respectively. This gave rise to 243 experiments in FRED and 126 in GOLD, with each experiment employing a unique combination of parameter settings. A set of 68 ligands were redocked into their target proteins with the different parameter sets, and the RMSD between the docked ligand and the x-ray ligand was calculated for the 15 top-ranked solutions in each case. The protein-ligand complexes included in the study were physicochemically diverse in terms of the properties of the ligands, allowing us to identify optimal settings for a broad range of ligand types. This high diversity was achieved by selecting ligands from a principal property space that was constructed by means of PCA compression of a matrix of physicochemical ligand descriptors using a space-filling design.

3.3 Results

We set out to compare the results of docking using the programs' default settings to those obtained using the settings that gave the best results in terms of RMSD. 66 ligands were evaluated in docking parameter variation experiments using FRED and 65 in GOLD. The top 15 poses for each ligand were compared to their pose in the x-ray structure of their protein-ligand complex by calculating the RMSD values for the top-ranked pose and for the best pose (*i.e.* the pose with the lowest RMSD relative to the x-ray ligand, selected from the 15 most highly-ranked poses for that protein-ligand pair). Using FRED, 32 of the ligands had at least one highly-ranked pose with an RMSD < 2.0 Å in at least one of the designed experiments; using GOLD, this number rose to 45. These dockings were considered to be successful (Figure 6). Using its default settings, FRED successfully docked 17 ligands (*i.e.* for these ligands, it generated at least one highly-ranked pose with an RMSD < 2.0 Å) while 29 ligands were successfully docked using individually-optimised settings (Figure 6a). Using GOLD, the corresponding numbers were 25 ligands with the default settings and 45 with tuned settings. Hence, a substantial number of ligands can be docked with RMSD values below 2.0 Å if one uses tuned parameters. We also noted that increasing the number of GA runs in GOLD from 20 to 100 did not result in a drastic difference in the results obtained when using the default settings (Figure 6b).

Although parameter tuning can have a large impact for individual ligands, our results demonstrate that for a set of ligands with a broad range of properties, the default settings are a good choice, *i.e.*, no other parameter setting that we tested significantly improved docking on average. This can clearly be seen in the PLS models used to relate the parameter settings to the docking outcomes as described by RMSD: the score plots show that the default settings are consistently positioned close to the optimal settings (see **Paper I**).

Analysis of the PLS models' regression coefficients revealed which parameters had the largest impact on the outcome of docking and are thus most worth tuning. In FRED, the identity of the exhaustive scoring function had the greatest influence; *clash checking* had some lesser influence, and the rest of the parameters were of little importance. On average, Chemgauss was the best exhaustive scoring function and a low *clash checking* value was preferable for our set of ligands. In GOLD, the *number* *of operations* were clearly the most important parameter, with optimal values ranging between 50500 and 100 000. Niche size was the second most important parameter in GOLD; for this parameter, a setting of 3 proved beneficial.

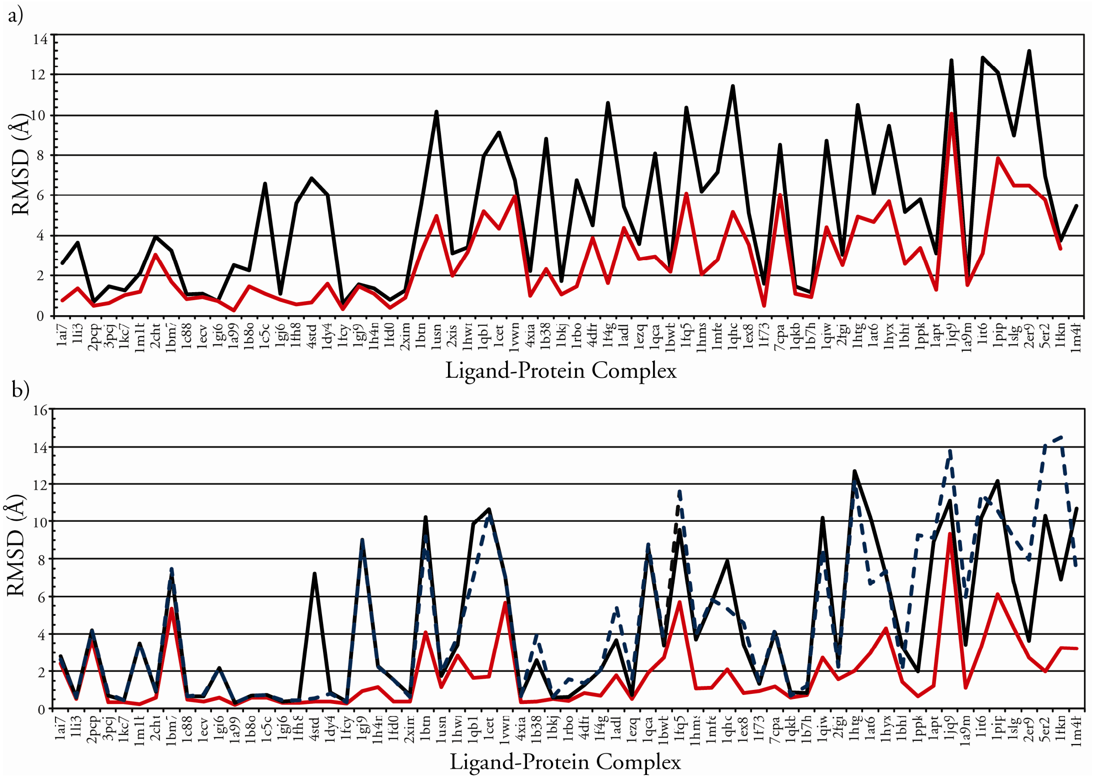


Figure 6. Docking results obtained for individual protein-ligand complexes presented as RMSD-values for the top ranked pose. Ligands are ordered from left to right according to the number of rotatable bonds present in them. Results when the default settings were used are shown by a black line and the docking results based on the best parameter sets for each complex are shown as a red line. a) FRED results. b) GOLD results. The dotted line corresponds to the results obtained with default settings and 100 GA runs.

3.4 Summary of Paper I

- Statistical experimental design together with PLS modelling is a viable and straightforward way of elucidating the impact of changing docking parameters on the outcome of docking.
- For roughly 25% of the ligands, it was possible to reduce the RMSD to less than 2.0 Å by using tuned settings instead of the defaults. However, use of the default settings in FRED and GOLD gave reasonably good results with all of the ligands examined.
- Ligands with many rotatable bonds tended to be poorly docked, although even ligands with as many as 30 rotatable bonds could be successfully docked using suitable parameter settings.

- Of the different parameters one might vary when seeking to improve the results of redocking experiments using FRED, changing the identity of the *exhaustive scoring* function and the value of the *clash checking* parameter are most likely to be useful. With GOLD, one should look to vary the *number of operations* and *niche sizes.* In both programs, these changes may also reduce the time taken for a docking run.

4. MAPPING LIGAND-BINDING CAVITIES IN PROTEINS (Paper II)

4.1 Why are Ligand-Binding Cavities Interesting?

In this thesis, we sought to develop an understanding of the interactions between ligands and proteins by using different computational tools and experiments. Obviously, elucidating the nature of the protein (and especially the area in the protein where the ligands bind) is of fundamental importance when trying to understand protein-ligand interactions. We use protein crystal structures in our calculations; as discussed in the introduction, these static structures do not provide a complete picture of the protein's behaviour under 'real' conditions. Nevertheless, we can obtain insights into protein-ligand interactions by describing both binding partners and comparing them. In theory, the properties of a ligand should be complementary to those of the protein it binds to. The use of physicochemical descriptors to describe ligands is well-established, and this thesis contains several examples of this approach. Furthermore, over the course of the last decade, physicochemical descriptors for the ligand-binding cavities of proteins have been developed. We wished to find ways to correlate the properties of proteins with those of their ligands, and to select a subset of proteins exhibiting high diversity in the natures of their binding sites from larger datasets. To this end, we developed a method for describing proteins' cavities which is similar to those used in describing and selecting diverse ligands; this method was presented in **Paper II**.

4.2 Charting Ligand-Binding Cavity Property Space

In **Paper II**, a set of ligand-binding cavities were described by calculating their physicochemical properties; the matrix of calculated properties was then compressed with PCA to provide an overview of the relationships between the proteins as judged by the properties of their cavities. The training set consisted of 239 ligand-binding cavities (representing 121 unique Structural Classification of Proteins (SCOP)[150] protein domains) The cavities' properties were determined by calculating 239 descriptors for each one using the Surface Cavity Recognition and Evaluation (SCREEN)[151] program. These descriptors include measures of the cavity's size, polarity, hydrogen bonding potential, amino acid content, shape, electrostatics, flexibility and secondary structural features. The predictive ability of the model was tested using two sets of cavities that had not been used in the modelling procedure. Set A consisted of 13 cavities from protein domains which were represented in the set used when building the PCA model (except for 1FDQ), while set B consisted of cavities from two protein domains, chymotrypsinogen and subtilisin, represented by five cavities each. Notably, the PCA model did not include any representatives of the subtilisin SCOP family used in set B.

4.3 Results

The use of PCA to visualize the inter-relations between proteins on the basis of the properties of their ligand-binding cavities was successful in the sense that the groups and deviators identified among the proteins could be rationalized by considering the proteins' structures and the SCOP to which they belonged. The PCA of the cavities' properties identified 11 PCs which described 54 % of the original variation with a cross-validated Q^2 of 0.42. Although 11 PCs were statistically significant, only the six first were evaluated in detail, mainly because PC1-6 accounted for the bulk of the variation present in the majority of cavities. Score and loading plots of PC1 versus PC2 are shown in Figure 7.

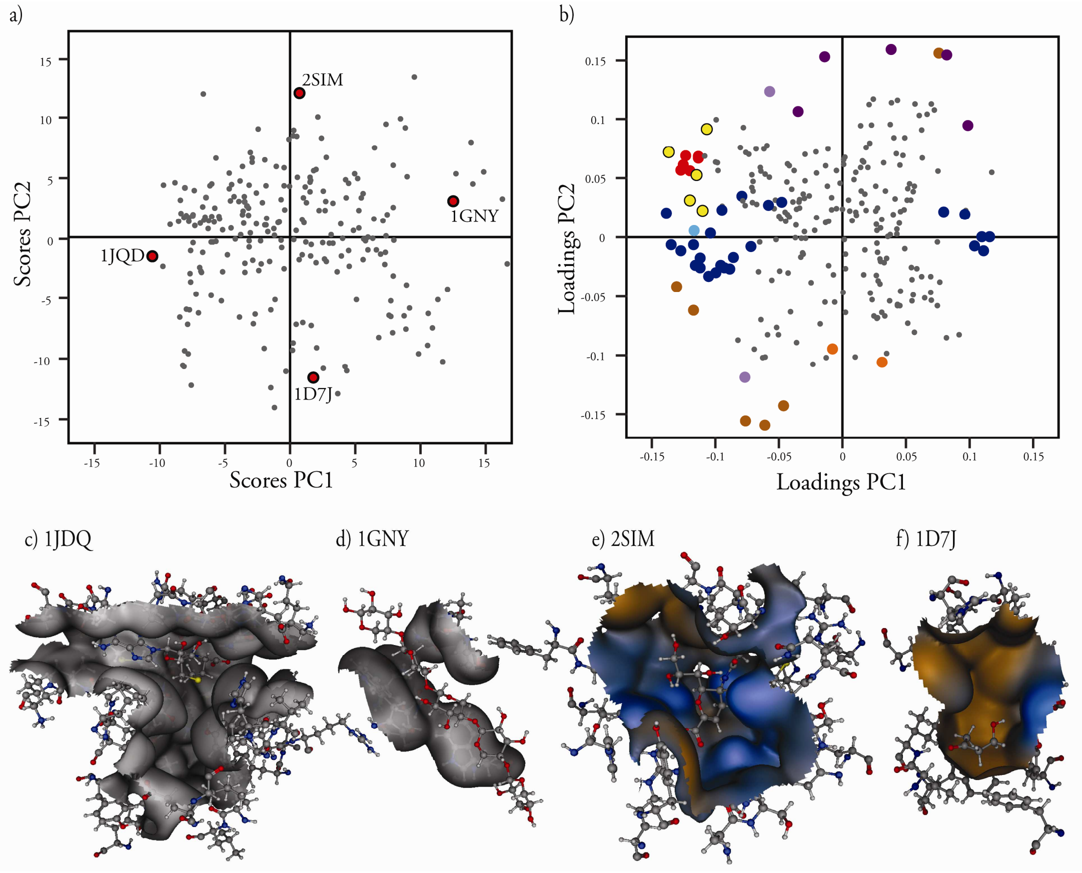


Figure 7. PCA score- and loading plots of PC1versus PC2 and surfaces of cavities. a) Score plot of PC1 versus PC2; each dot represents a ligand binding cavity. In red are proteins which are opposites in PC1 and PC2 and for which cavities are shown below. b) Loading plot of PC1 versus PC2; each dot represents a surface descriptor. The most influential descriptors in each PC are colour-coded as follows. PC1: size (red), depth (dark blue), shape (yellow), flexibility (light blue), and polarity (brown). PC2: charge (orange), polarity (brown), electrostatic field (light purple), and hydrogen-bond density (dark

purple). c-f) Ligand-binding cavities of 1JQD, 1GNY, 2SIM and 1D7J, respectively. Colours in the cavities are related to polarity: brown areas are lipophilic and blue areas are hydrophilic.

The positions of individual cavities in the score plot (Figure 7a) are determined by their principal properties (loading values in Figure 7b). Size descriptors and descriptors correlated to size were most influential in PC1 and size is thus the most influential principal property in discriminating between the cavities. The proteins 1JQD (histamine methyltransferase) and 1GNY (xylanase 10c) shown in Figure 7c-d provide examples of cavities which are well-separated along PC1: the former has a large and deep cavity while the latter has a small and shallow cavity. The second principal property differentiating the cavities was their polarity; descriptors affected by this property are the primary constituents of PC2 (Figure 7b). Two proteins whose cavities are located at the extreme of PC2 are 2SIM (salmonella sialidase, Figure 7e) which has a very hydrophilic cavity, and 1D7J (FK-506 binding protein, Figure 7f) which has a lipophilic cavity. Subsequent PCs described the cavities' charges (PC3), depth/shape (PC4), electrostatic fields (PC5), and aromaticity/flexibility (PC6).

We wanted to investigate the relationships between the physicochemical properties of the ligand-binding cavities and those of their ligands. To do this, we calculated physicochemical descriptors for the ligands and extracted their principal properties using PCA. Interestingly, although the same properties were dominant in the two first PCs of both the ligand and the cavity PCAs (*i.e.* PC1 corresponded to size and PC2 to polarity in both cases) there was no evidence of correlation between the two property spaces. This indicates that proteins do not completely embrace their ligands, which has implications for the design of new ligands. It suggests that that it may be possible to develop ligands with very different binding modes compared to 'natural' ligands, which do not necessarily fill the cavities in which they bind.

There is a great need for new methods for the prediction of proteins' functions in drug discovery and in related fields because new proteins whose purpose is unclear are frequently identified. The function of newly discovered proteins can often be assigned by comparison with proteins whose functions are known because it is often the case that similar proteins have similar roles in biological systems. Although the prediction of functionality was not the primary objective of this study, our model proved to have reasonable predictive ability in assigning the functions of unknown proteins. We defined the cavities' principal property space in terms of the first six PC from the PCA and created a map (PCA clustering tree) of the cavities' positions within this six-dimensional space (Figure 8). By calculating and visualising the distances between cavities within this space, we were able to draw conclusions about the relationships between the cavities and the proteins to which they belong. Proteins close to each other on the branches of the tree have similar cavity-properties. By simply studying the tree, we could identify segregated and coherent domain distributions, and we could draw conclusions about the physicochemical basis of these patterns by analyzing the score and loading plots from the underlying PCA. The predictive ability of the PCA tree was assessed by using it to predict the functions of the cavities in test sets A and B. All of the cavities in set A were found to cluster close to their respective domains while those in set B were located on the branch containing the serine proteases (urokinase-type plasminogen activators and trypsinogens) found in the lower half of the tree in Figure 8. The cavities of set B do indeed have similar functions to these domains, but interestingly, the subtilisins of set B have a completely different fold to the serine proteases.

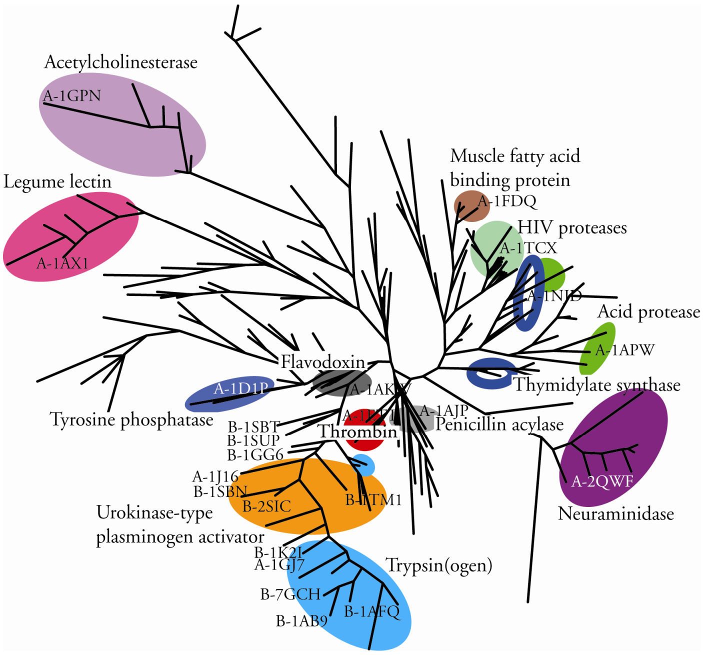


Figure 8. PCA clustering tree of a subset of the modelled cavities. Proteins are positioned at the ends of branches and the distances along the branches correspond to the Euclidean distance between the proteins in the PC space spanned by the first six PCs. Examples of positions of protein domains are indicated with coloured ellipses. Prediction sets A and B are indicated by prefix A or B and the PDB-code.

4.4 Summary of Paper II

- PCA together with SCREEN descriptors proved to be a viable alignment-independent method for elucidating the relationships between proteins on the basis of the properties of their ligand-binding cavities.
- The biggest differences between the cavities were due to variation in their sizes. Progressively smaller differences were observed in terms of their polarity, charges, depth/shape, electrostatic fields, and aromaticity/flexibility.
- In general, there is no clear correlation between the main properties of the ligand-binding cavities and those of their ligands, indicating that it may be possible to design more suitable ligands in some cases.

- Differences in the properties of the protein ligand-binding cavities within a domain were typically attributable to differences in their conformations. This should be borne in mind when selecting protein 3D structures for use in structure-based design, the results of which are highly sensitive to such structural details.
- Two proteins with substantial differences between their sequences may nevertheless have cavities with similar physicochemical properties. This may be important in predicting their functionality and cross-reactivity.

5. AN INSIGHT INTO VIRTUAL SCREENING (Paper III)

5.1 Virtual Screening

Much like the single ligand docking experiments investigated in **Paper I**, VS results are heavily dependent on both the software and the scoring-functions used. The objective of VS is to screen a database of molecules to find ligands that will fit in the protein's ligand-binding cavity. The proposed binders are then tested in biological experiments; ideally, the experiments will confirm their affinity for the protein. Although many successful VS campaigns have been described, it is still very hard to plan a VS experiment. As discussed in the introduction, one must carefully choose the protein crystal structure to be examined as well as the docking software and scoring function. We know that all these factors influence the results, so how should we go about finding “optimal” conditions which will lead to new exciting molecules with high affinities for the targeted proteins? A comprehensive answer to this question is beyond the scope of this thesis. However, our attempts to address some aspects of this question are discussed in **Paper III**. We applied the methods described in **Papers I** and **II** to design a method that can be used to identify factors that affect VS targeting specific proteins.

5.2 Design of Virtual Screening Experiments

The ability of a docking program to find known ligands for a protein can be investigated using data mining techniques which we call “simulated VS”. These experiments rely on the use of a database of molecules that are unlikely to bind to the protein (so-called "decoys") that has been spiked with a small number of known binders (ligands, or "actives"). The spiked database is used to evaluate the docking program's ability to identify good ligands by assessing its ability to assign high ranks to the known binders *i.e.* its ability to enrich the database. We used this strategy in conjunction with spiked databases created specifically for six different proteins to test the ability of FRED[152] and GOLD[71, 148, 153] to correctly identify actives. The docking parameters and scoring functions used with FRED were varied according to DoE to investigate their influence on the results of the VS. There are significant differences between the physicochemical properties of the binding cavities of the different proteins used in this study; the proteins were selected to maximise these differences because it is known that different kinds of proteins require different VS conditions. In this way, it would be possible to relate the properties of the different cavities to the performance of the VS.

The six proteins were obtained from the DUD database[28] which contains proteins and associated databases of molecules comprising both known binders (~50-150) and decoys (~1600-5700). Cavity descriptors were calculated for the DUD proteins using SCREEN[151] and then subjected to compression by PCA; six proteins were selected from the directory on the basis of their positions in the resulting score-plots. The properties of the decoys in the DUD have, to some extent, been matched to those of the ligands so that the ligands are similar to the decoys, making it more of a challenge for the docking software to identify ligands that will fit to the protein. The ligands from the databases were docked to their respective proteins using five different combinations of parameter settings in FRED and one in GOLD. Five different scoring functions were used to rank the output databases. Ligand enrichment in the ranked databases was assessed in terms of their EF and ROC-AUC.[119] Finally, the relationships between the docking software settings and the VS results were analysed in terms of their EF and ROC-AUC values using a PLS regression model.

5.3 Results

Six proteins with physicochemically diverse cavities were examined: angiotensin-converting enzyme (ACE), acetylcholinesterase (AChE), cyclin-dependent kinase2 (CDK2), fibroblast growth factor receptor 1 (FGFr1), coagulation factor Xa (FXa), and trypsin. A broad range of enrichments was observed over the 25 VS experiments with FRED and five with GOLD, depending on the precise settings and scoring functions used. The choice of post-docking scoring function was found to have the largest impact on the results, and different protein targets were found to have different optimal scoring functions for enriching active binders. In general, enrichment was highest for FXa, CDK2 and trypsin. The Chemgauss3 and Plp scoring functions were optimal for finding active binders of FXa, while Chemscore was optimal for CDK2 and trypsin. Furthermore, different scoring functions tended to assign high ranks to different kinds of active molecules. For example, of the molecules screened against FXa, 21 of those in the 95th percentile as ranked by Chemgauss3 were not ranked highly by Plp. Similarly, Plp identified nine active molecules that did not feature in the 95th percentile as calculated using Chemguss3. This suggests that by using multiple scoring functions, it may be possible to identify actives with a broader range of properties than would be obtained if one focused exclusively on a single function.

One point of concern in this study was the finding that the DUD database which had simply been sorting on ligand vdW volume generated EF values greater than or equal to those obtained with the various 'real' scoring functions for all of the proteins examined. This enrichment was only really apparent when considering EF values; the ROC-AUC values achieved by sorting in this way were rather lower than those achieved in the docking-based VS. By design, the DUD should contain 36 similar decoy molecules for each ligand,[28] so it was somewhat surprising that the vdW volumes of so many of the decoys were so dissimilar to those of the real ligands. One consequence of this was that good enrichments were obtained with Shapegauss, which focuses exclusively on volume complementarity.

We also found weak indications of a relationship between the identity of the optimal scoring functions for a given protein and the properties of the protein's ligand-binding site. Thus, Chemgauss3 and Plp seemed to give higher enrichment for lipophilic cavities such as those found in AChE and FXa, while Chemscore was better at enriching actives for target cavities with many ionic amino acids and potential hydrogen bonding groups such as trypsin (and to some extent CDK2 and FGFr1).

5.4 Summary of Paper III

- We used DoE and PLS to elucidate the effects of various docking parameters and scoring functions on the outcome of simulated VS.
- The choice of scoring function was found to be the single most important factor influencing the outcome. Importantly, good (and bad!) choices of parameter settings and scoring functions could be identified for all of the targets used in the study.
- Different scoring functions assigned high ranks to different kinds of active molecules, indicating that it may be possible to obtain a set of binders having a broader range of properties by screening with multiple scoring functions.
- Trends in the properties of the actives and decoys in the databases used for studies such as this are highly important and can influence the results. Much works remains to be done in balancing the properties of molecules in the reference databases to reduce bias introduced by things like differences in size of the actives compared to the decoys.

6. PEPTIDE DESIGN FOR THE STUDY OF AUTOIMMUNE ARTHRITIS (Paper IV)

6.1 Background

Rheumatoid arthritis (RA) is a chronic inflammatory disease. It primarily affects the joints, and leads to disfigurement, loss of function and pain, especially in the hands and feet.[154] RA is thought to be an autoimmune disease but its cause is unknown. Its symptoms are due to an abnormal immune system response which destroys the connective cartilage in the joints. Collagen-induced arthritis (CIA) is a form of arthritis that can be induced in mice using rat collagen, and is used as a model for the study of RA. The class II MHC, a protein found in antigen presenting cells such as macrophages, plays a key role in the onset of arthritis; it is involved in the presentation of self and non-self protein fragments to T-cells (Figure 9).

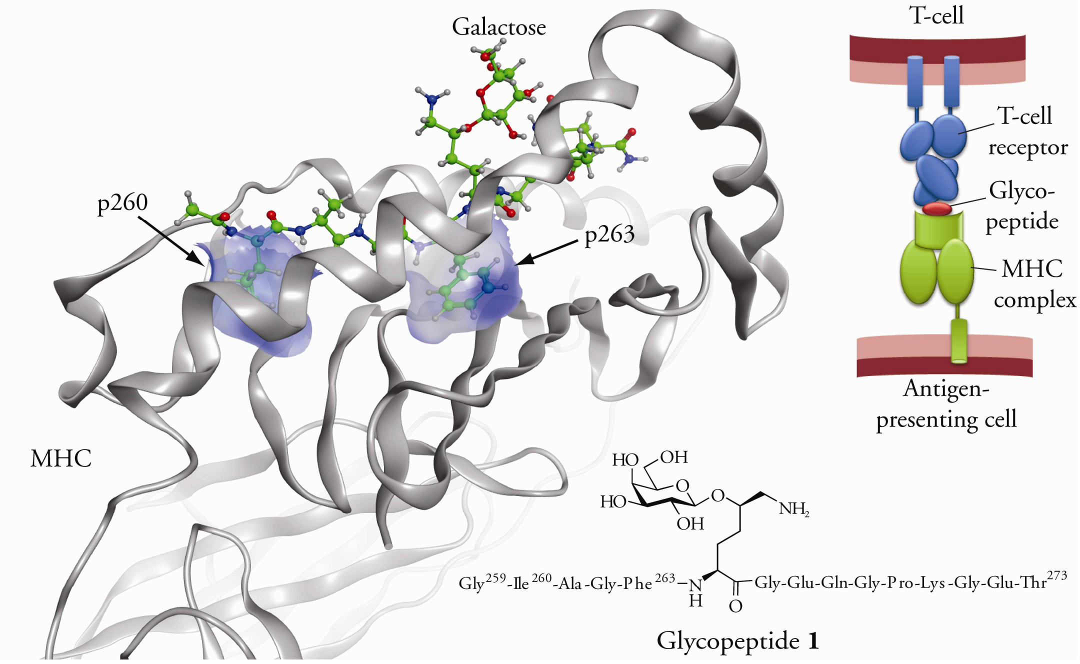


Figure 9. Upper right: schematic picture of the T-cell/MHC interaction mediated by the glycopeptide. Left: a ribbon representation of the ligand-binding region of A^q. The glycopeptide (CII260-267) is in green carbons and the anchor amino acids (Ile^{260} and Phe^{263}) are pointing down into the anchoring pockets P1 and P4 in A^q. The pockets are coloured in gray/blue where dark blue areas are more hydrophilic. Glycopeptide **1** (CII259-273, to the lower right) is a known binder to A^q, and position 260 and 263 in the peptide has been modified as part of this study.

A comparative model of the peptide-binding region of the mouse MHC protein A^q is shown in Figure 9.[155] Glycopeptide **1** (Figure 9) is a known binder of A^q; when bound, its galactose moiety, which is recognized by the T-cells, projects outwards from the MHC.[156, 157] The Ile^{260} and Phe^{263} residues of glycopeptide **1** (which are located in peptide positions p260 and p263, respectively) are referred to as "anchoring residues", and are important in the binding of the peptide to the MHC. Previous studies have shown that it is possible to prevent and even reverse CIA in mice by vaccination with glycopeptides such as **1** or with glycopeptide-MHC complexes.[158, 159]

We designed a set of 21 novel glycopeptides (including **1**) and studied their binding to MHC proteins (mouse A^q and its human counterpart, DR4) and the response of a range of T-cell hybridomas to these glycopeptide-MHC complexes. The glycopeptides were prepared using solid-phase peptide synthesis, and incorporated various natural and unnatural amino acids. The affinity of the glycopeptides for the A^q and DR4 proteins was evaluated *in vitro*, as was the ability of the glycopeptide-MHC complexes to promote T-cell activation with six T-cell hybridomas associated with A^q and two that are associated with DR4. The study generated novel and useful insights into the relationships between glycopeptide structure and the binding preferences of A^q and DR4, and into the SAR associated with T-cell responses. These findings will be taken into consideration when we select a group of glycopeptides to include in future *in vivo* studies.

6.2 Results

6.2.1 Glycopeptide Design

A two-step process was used when designing the glycopeptides. The first step involved a docking-based VS of 11025 truncated versions of **1** against the A^q protein. The truncated peptides consisted of the minimal epitope between Ile^{260} and Gln^{267}, and did not incorporate the sugar moiety. Various different amino acids were incorporated at p260 and p263 to assess the impact of variation in these positions on binding. The aim of this first step was to filter out amino acids that were likely to result in peptides that could interact with the anchoring pockets in the active site of A^q. The second step involved the design of new glycopeptides by means of a SMD strategy using the amino acids identified in the VS. The SMD resulted in a library of 21 peptides whose members collectively exhibit a broad range of chemical properties at positions p260 and p263.

6.2.2 Virtual Screening and Docking Software Tuning

In the first step, a VS was performed using both the FRED[152] and GOLD[71, 148, 153] docking programs. The use of two software packages was motivated by the fact that different programs sometimes arrive at different conclusions regarding putative binders. Initial redocking experiments using the peptide from the comparative model of the A^q–glycopeptide complex (henceforth referred to as the "native peptide") and the default software settings failed to reproduce the binding pose observed in the model. It has been suggested that in order for a VS to be fruitful, the docking software needs to be able to reproduce the pose of the ligand in the x-ray structure; we agree with this suggestion.[160] Docking (non-drug like) molecules of the size of the native peptide (30 rotatable bonds) is complicated due to the huge number of conformations that must be evaluated. Nevertheless, even with large molecules, it is sometimes possible to optimise the docking process by varying the docking parameters, as was shown in **Paper I**.

Therefore, new protocols were designed for the generation of conformations in OMEGA[161] and for docking in FRED and GOLD, inspired by the parameter optimization strategies presented in **Papers I** and **III**. First, a docking constraint was introduced to restrict the rotation of the peptide. Second, DoE (in the form of a FFD) was used to tune the parameters of the docking software. The OMEGA settings recommended by Kirchmair and co-workers were initially adopted,[162] but later experiments showed that with further parameter optimisation, OMEGA can produce conformations with an RMSD of only 1.60 Å relative to the native peptide; the default parameters produce conformations with an RMSD of 2.74 Å (unpublished results). These optimized settings in OMEGA (*maxconfs* = 250, *ewindow* = 35 and *rms* = 0.8) have an additional benefit compared to the default settings: because the value of *maxconfs* is reduced, the docking calculations are faster. Tuning of the settings resulted in a slight improvement in the results obtained with FRED and in a rather greater improvement in those obtained with GOLD, which ultimately generated docking solutions with an RMSD of 1.59 Å, compared to the RMSD of 2.80 Å obtained with the default settings. In addition, the use of optimised settings reduced the time required for docking with both programs.

A VS was performed in FRED, using the tuned software settings, on 11025 virtual peptides. A physicochemically-representative subset of 2916 peptides selected by DOOD was docked using GOLD. Post-docking filtration was performed to remove peptides whose RMSD relative to the native peptide structure was > 3 Å, and the remaining peptides were subjected to an energy minimisation. Peptides that had an RMSD < 1.5 Å relative to the native peptide after this re-minimisation (of which there were 1540 from FRED and 741 from GOLD) were rescored using Chemgauss2, Chemgauss3, Chemscore,[82] Goldscore,[71] Plp,[54] Screenscore,[83]and Shapegauss[87] to assess their binding to A^q.

By applying PCA to the re-scoring results, we developed a strategy by which it was possible to use the scores for whole peptides to assess the influence of individual amino acids at the p260 and p263 positions on the overall affinity of the peptides for the MHC. For example, of the 1540 peptides that were reoptimised after FRED dockings and then re-scored, 29 incorporated a phenylalanine (Phe) residue at p263, making Phe one of the most commonly-observed residues in this position. The average Chemgauss3 score value for all peptides containing Phe was -124, with a standard deviation of 6.8, and the best score value for a peptide containing Phe was -134, making it the 12th most highly-ranked amino acid. The data from the re-scoring was used to create descriptors for the 'performance' of specific amino acids at p260 and p263. Thus, at p263, Phe was characterised by a frequency of 29 and by the average, standard deviation, and maximum of the scores assigned to all of the peptides incorporating Phe at p263, by all of the scoring functions examined. This effectively allowed us to evaluate the 'consensus' on the performance of Phe in position p263. On the basis of the score plot from the PCA, a subset of 46 highly-scored and frequently-observed p260 amino acids and 52 highly-scored and frequently-observed p263 amino acids were selected for inclusion in a SMD of new glycopeptide ligands for the MHC.

6.2.3 Statistical Molecular Design of Glycopeptides

Physicochemical descriptors were calculated for the amino acids that had been selected for use in the SMD, and PCA were performed to extract and visualise the variability in their properties. On the basis of the PCA, two subsets of seven physicochemically-diverse amino acids were selected, one for p260 and another for p263, for incorporation into a small library of new glycopeptides (Figure 10).

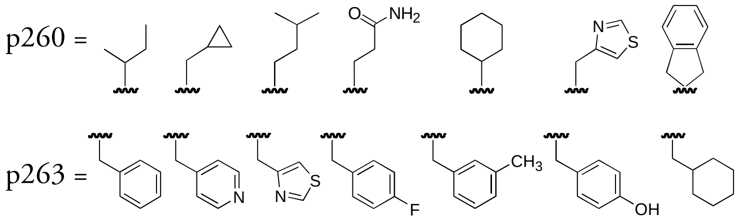


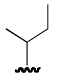
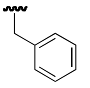
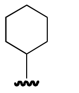
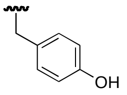
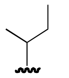
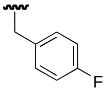
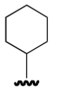
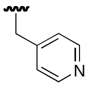
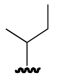
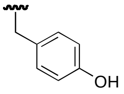
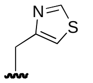
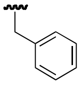
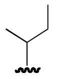
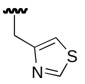
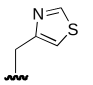
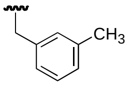
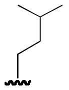
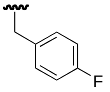
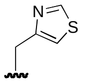
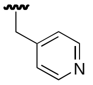
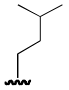
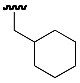
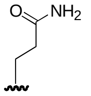
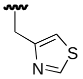
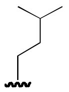
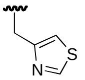
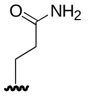
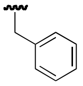
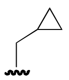
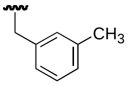
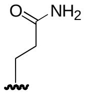
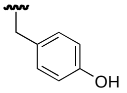
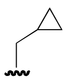
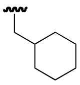
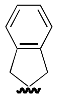
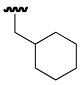
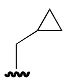
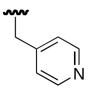
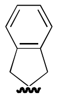
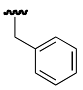
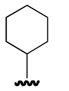
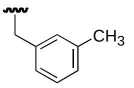
Figure 10. Seven amino acids (whose side chains are displayed here) for p260 and p263 were selected from PCA scores. The combination of these amino acids gave rise to 49 theoretical peptides from which 20 were selected by DOOD to be synthesized.

The amino acids selected for p260 exhibited diversity in terms of their size, hydrophobicity and flexibility, whereas the variance in those selected for p263 was largely restricted to differences in hydrophobicity and density. With two sets of seven amino acids, the maximum possible number of unique glycopeptides in the library was 49; of these, a subset of 20 was selected using DOOD. Importantly, each amino acid was incorporated into two or (more often) three glycopeptides, facilitating a robust interpretation of their individual biological effects.

6.2.4 Affinity Evaluation of the Designed Glycopeptides

The 20 designed glycopeptides and glycopeptide **1**, were synthesized, and their affinities for the A^q and DR4 proteins were assessed, along with their recognition by T-cell hybridomas. (Table 1)

Table 1. Competitive inhibition of biotinylated CII259-273 binding to the A^q and DR4 protein by glycopeptides **1**-**21** substituted in p260 and p263.

glyco- pept.	p260 side chain	p263 side chain	A^q % inhib.[a]	DR4 % inhib.[b]	glyco- pept.	p260 side chain	p263 side chain	A^q % inhib.[a]	DR4 % inhib.[b]
1			90 ± 1	72 ± 4	12			-	73 ± 3
2			87 ± 1	73 ± 6	13			-	52 ± 3
3			28 ± 3	65 ± 4	14			20 ± 4	74 ± 1
4			86 ± 1	31 ± 10	15			41 ± 1	76 ± 2
5			29 ± 2	77 ± 1	16			-	23 ± 11
6			46 ± 1	84 ± 4	17			-	26 ± 11
7			39 ± 2	34 ± 6	18			-	60 ± 6
8			55 ± 3	70 ± 3	19			-	80 ± 5
9			38 ± 6	75 ± 0	20			-	82 ± 3
10			-	19 ± 15	21			-	84 ± 5
11			29 ± 3	76 ± 5					

[a] *Inhibition at 100 µM. Glycopeptides assigned with" –" were identified as inactive (< 30% inhibition) in a preliminary A^q assay and were not included in the second assay presented here.* [b] *Inhibition at 500 µM.*

In the model of the complex formed between A^q and the 'natural' glycopeptide CII259-273, the P1 pocket accommodates an Ile residue, while P4 plays host to the side chain of a Phe (Figure 9). The P4 pocket is larger and deeper than P1 and both pockets are mainly lipophilic but they contain a hyrdrophilic region close to the pocket floor. The inhibitory effects of the designed glycopeptides are shown in Table 1; most of them were found to show affinity for A^q. Binding to A^q was relatively insensitive to modifications at p263. In contrast, modifications at p260 were not well tolerated; the incorporation of bulkier or more polar side chains than that of isoleucine lead to a total loss of binding (peptides **12**, **13**, **17**-**21**).

The use of smaller side chains such as cyclopropylalanine also led to a decrease in binding, suggesting that isoleucine has “optimal” properties for interacting with this pocket.With the exception of the 4-pyridylalanine side chain (which was detrimental to binding; see peptides **10**, **13**, and **16**), most modifications at p263 were well tolerated. This means that it may be worth investigating the incorporation of even more diverse amino acids at this position in future studies. On the basis of the comparative model of A^q, we hypothesised that the P4 pocket might be able to accommodate Phe-type residues bearing small substituents in the *para-* and/or *meta*-positions, and this indeed proved to be the case. Both 4-fluorophenylalanine (*cf.* **2** with **1**) and *m*-methylphenylalanine (*cf.* **15** with **14**) exhibited the same affinity for A^q as the native peptide. The more bulky 3-cyclohexylalanine (*cf.* **6** with **7**) is also a good substitute for Phe, as is 4-thiazolylalanine. However, replacing phenylalanine with tyrosine led to a reduction in affinity for A^q (*cf.* **3** with **1**).

DR4 also binds to **1**, and the glycopeptide residue Phe^{263} is most heavily involved in this binding is.[163] The anchoring pocket in the DR4 binding can accommodate larger amino acids than Phe.[164, 165] Our binding experiments showed that some of the designed peptides have a greater affinity for DR4 than does the native peptide (Table 1).The binding evaluations clearly show that the DR4 anchoring pocket is very tolerant of a broad range of amino acids: essentially all of the glycopeptides except those incorporating 4-thiazolylalanine or 4-pyridylalanine at p263 bound well to DR4. Varying the amino acids at p260 caused only small changes in affinity, which suggests that it may be interesting to examine the impact of introducing residues bearing larger and more hydrophobic side chains at this position.

6.2.5 Comparison of Affinities and Docking Scores

With the biological results in hand, we had an opportunity to evaluate the performance of the different scoring functions in predicting the influence of the different amino acids on the peptides' affinity for A^q. Our first observation was that different scoring functions differed in their rankings of amino acids, and that the default scoring function in FRED, Chemgauss3, did not reliably assign high ranks to amino acids that contribute usefully to binding. Weak correlations between the binding results and the rankings obtained with Chemscore and Plp were observed for amino acids at p260 and p263. Chemscore recognized the inactive amino acids Gln and 2-indaneglycine as weak binders, ranking them in the lower half of the 105 amino acids docked to the p260 position. Plp was the only scoring function to rank the inactive 4-pyridylalanine last among the biologically tested amino acids.

6.2.6 Evaluation of T-cell Responses

The ability of the appropriate T-cells to recognise the complexes formed between the new glycopeptides and the MHC proteins was measured. The results clearly showed that there is a substantial correlation between T-cell recognition and MHC binding (see **Paper IV**). This was anticipated since the peptides need to bind to the MHC in order for it to be recognized by the T-cells. However, the results also demonstrate that by itself, a high peptide-MHC affinity is not sufficient for strong T-cell recognition. All of the T-cells responded well to MHC complexes with glycopeptide **1** but, interestingly, the complex formed between A^q and glycopeptide **4** (which has a high affinity for A^q) was only weakly recognized by two of the six T-cell hybridomas examined, indicating that changing the amino acid at p263 can affect T-cell recognition. Subsequent MD simulations revealed that the variation in RMSD was larger for A^q/glycopetide complexes which induced stronger T-cell responses (**1** and 7) compared to those that induced weaker responses (**6** and **9**). Further MD simulations may reveal if there is a general connection between A^q/glycopetide complex flexibility and T-cell recognition.

In the case of DR4, some of the designed glycopeptides generated even stronger T-cell responses than did glycopeptide **1** (*cf.* **1** with **5**, **8** and **11**). Furthermore, certain amino acids (*e.g.* cyclohexylalanine) were found to consistently generate stronger responses towards one of the hybridomas (mDR17.2) when incorporated at p263. Finally, when compared to the native ligand, some of the designed species exhibited enhanced recognition with respect to one hybridoma but diminished recognition by the others (*cf.* **1** and **4**).

6.3 Summary of Paper IV

- Anchoring residues in a glycopeptide that binds to MHC proteins were exchanged for various natural and unnatural amino acids, resulting in peptides whose affinities for MHC proteins are similar to or better than those of the native ligand and which generate a range of different T-cell responses.

- The use of docking software parameters tuned by DoE facilitated the successful redocking of the native peptide.
- A combination of docking-based virtual screening and SMD allowed us to create a set of glycopeptides that exhibited a range of affinities for A^q and DR4, allowing robust SAR conclusions to be drawn.
- The P1 pocket of A^q is more sensitive to peptide modifications than is the P4 pocket. The anchoring side chain in p260 should be aliphatic and of medium size (*i.e.* it should not contain more than four carbons), while the side chain in p263 should be a non-substituted or *meta*-substituted phenylalanine- or 3-cyclohexylalanine derivative.
- The Chemscore and Plp scoring functions produced the best predictions of the affinities of the peptides for A^q and may be generally useful for docking peptides to A^q.
- The designed glycopeptides elicited a range of T-cell responses and we were able connect some of these effects to changes in the epitope seen in MD simulations. This augurs well for the use of related compounds in future *in vivo* studies, where they may prove useful in elucidating the biological effects of the responses of specific T-cell types.

7. PERSPECTIVE AND FUTURE WORK

Understanding and predicting the interactions between ligands and proteins are two of the great challenges of modern chemistry, and good solutions to these problems would be of great value, not least in the context of drug discovery. Molecular modelling tools that can satisfactorily meet these challenges have yet to be developed. It is necessary to consider a huge number of factors when assessing binding, and workers in this area have only recently begun to address issues such as molecular flexibility and the effects of water. Much work remains to be done in the development of more accurate and efficient tools. In the absence of such new tools, we need to exploit the full potential of existing modelling tools and to develop strategies to effectively describe proteins. This work described in this thesis represents a contribution towards this latter challenge.

In our experience, the software used to perform molecular modelling tasks such as docking is often 'tuned' by individual researchers to improve its performance in the problems that have attracted their interest. However, the manner by which the tuning has been performed is not often discussed in the literature. Furthermore, no comprehensive reports have been presented regarding the possibility of optimizing software for use in the study of specific types of proteins, although it is clear that the default settings of most programs are tuned for applicability to drug-like molecules. We have shown in this thesis that multivariate designs can be used to investigate factors that influence docking results, both in redocking experiments and in virtual screening. These methods could easily be applied to the tuning of scoring functions, or to the identification of optimal default settings for any molecular modelling tool with adjustable parameters. In addition, relatively simple methods based on varying the parameters used in docking might provide attractive approaches to generating ligand poses for subsequent scoring with more advanced methods such as PB/GB-SA. We adopted this strategy in a recent study where we correlated calculated free energies of binding with experimentally-determined binding data for a set of small-molecule ligands of the FXa and MHC proteins.[92]

Unfortunately, in recent years, the tendency has been to hide the parameters used in docking programs from the end-user, making it more difficult change them and to investigate their influence. We have also noticed that software packages increasingly tend to focus on certain specific target protein classes. Why not offer one software package that can be tuned for use with different targets at will? Such a program would probably be attractive to a large number of users, especially to those who are studying non-drug like compounds, and to academics, whose funds are typically somewhat limited.

With regard to tuning software for virtual screening, we feel that our results show that multiple different scoring functions can be suitable for enriching actives against different target proteins. Although this has been stated by others, our results complement these prior findings in that they provide a method by which suitable scoring functions can be efficiently identified. It is evident from our results and from those of other studies that the evaluation of the effectiveness of a VS tool is heavily dependent on the availability of bias-free databases. However, no such databases have yet been made publically available. If one sought to create a truly physicochemically-balanced dataset containing highly similar ligands and decoys, one could calculate descriptors and use PCA to identify and quantify this similarity. Finally, the transferability of results obtained from simulated VS to real-world problems remains to be investigated.

When dealing with protein-ligand interactions it is important to have a clear picture of the properties of both molecules in order to understand how and why they interact. For instance, are the properties of the ligand and the ligand-binding cavity complementary, and if so, is it possible to identify new ligands for a protein solely on the basis of the physicochemical properties of the ligand-binding site? Given that the number of new potential drug targets with unknown ligands is rising and the importance of predicting off-target actions of existing drugs, questions such of these are of great interest to members of the bioinformatics community who focus on drug research. The results described in this thesis suggest that one cannot identify new ligands for a protein solely on the basis of similarities between their properties and those of the binding cavity; the property spaces of the ligands and the cavities are not directly correlated. However, it may be possible to use this approach to match the properties of ligands and proteins to find potential binders, although the descriptors we have used will probably need to be augmented with information on interatomic and inter-functional group distances. Our strategy can also be used to select target proteins in structure-based design and to predict the purpose of proteins whose function is unknown by relating the properties of their cavities to those of proteins whose function is known.

Finally, some of the work described in this thesis focused on the study of autoimmune arthritis. We showed that combining structure-based and ligand-based design is an efficient way to incorporate known data on ligand- and protein-structures into one's experimental designs. The method is useful in ensuring that the molecules one ultimately designs will be sufficiently diverse to allow meaningful conclusions to be drawn from their SAR. We developed this strategy for non-drug like molecules but it is equally applicable to the design of molecules other than peptides. Furthermore, the results from the MD simulation studies were encouraging, and they may help to unveil the structural changes that occur on binding in peptide-MHC complexes, which may be related to changes in their T-cell recognition. The results obtained will be used to identify peptides suitable for use in *in vivo* vaccination studies looking at the effects of MHC binding and T-cell responses on the development of the disease.

8. ACKNOWLEDGEMENTS

Tack till…

Min handledare **Anna Linusson** för att du gav mig chansen att jobba med alla dessa spännande forskningsprojekt, trots at jag inte riktigt trodde på mig själv där i början. Du har en otrolig förmåga att inspirera och peppa vilket har varit väldigt värdefullt (kanske helt avgörande) för mig under de här åren. Jag önskar dig och familjen all lycka i framtiden!

Min rumskamrat och vän **Anton Lindström**. Tack för all hjälp och för att du och **Hanna**, **Lilly** och **Knodden 2.0** har välkomnat mig i er familj. **Ida Andersson**, för din positivism och entusiasm. Du är en fantastisk samarbetspartner och en super-kemist!

Brian Chen and **Florian Raubacher** for great collaborations. Both of you have contribute with very insightful comments on our publications and I hope we will be able to collaborate in the future as well. Sees-Editing Ltd. (**John**, **David** and **Joelle**) for the lagua….lagnuag….language editing.

Lotta Edvinsson, min nuvarande rumskamrat. Tack för alla trevlig pratstunder både kring jobb och annat och jag ser fram emot fler! **Mia** för värdefullt jobbsnack, trevliga konferensresor och välbehövda yoga-stunder.

Fredrik Almqvist, **Mikael Elofsson** och **Andreas Larsson** för att ni inspirerade mig att stanna kvar här på kemi efter min grundutbildning och doktorera.

Carina, **LG**, **Pelle och Barbro** för att ni alltid tar er tid med alla mina frågor och för allt stöd med praktiska saker

Johan Trygg, **Elisabeth Sauer-Eriksson** och **Frippe** för att ni ställt upp på mina uppföljningsmöten och bidragit med insiktsfulla kommentarer till mitt arbete.

Anna K, **Sara S**, **Ida A**, **Caroline Z**, **Sussi J**, **Lotta E**, **Sofi E**, **Maria P**, **Hanna F** och alla andra på Kemin som förgyllt dagarna med fika och påhitt. Tycker vi ska fortsätta bjuda varandra på konditoriska läckerheter! Särskilt tack till **Anna K** för ditt stöd under min tid som nyfiken men lite rädd examensarbetare.

Markus D för alla pratstunder om film och spel och allt annat som ibland tar alldeles för mycket tid i anspråk. **Mattias H** och **Hans A** för alla otroliga äventyr i grottor och slott bland alver och orcher. Jag vet inte hur många gånger jag räddat livhanken på er…jag säger bara WOW.

Max B, **Elin S**, **Elin C**, **Anna W**, **Henke** och hela kemometri-gänget för trevliga fikastunder och intressanta ”multivariata” diskussioner.

Hiba A, **Eva N**, **Lucas R**, **Joel E**, **Cecilia L** för era bidrag till den här avhandlingen genom era examensarbeten och projekt (for your contribution to this thesis through your master thesis work and projects).

Dig som jag glömt nämna här och som bidragit till trevnad och gott arbete under min doktorandtid.

Mamma, **Pappa**, **Mormor**, **Farmor**, **Erik**, **Joel** och resten av familjen som stöttat och nog aldrig riktigt fattat vad jag pysslat med på jobbet, vilket har varit ganska skönt. Nu hoppas jag vi kan ses mer!

Sist men allra främst vill jag tacka älskade **Pedher** för att du gett mig perspektiv på tillvaron. Jag vet nu vad som verkligen betyder något. Tack för ditt tålamod och ditt stöd och en sak är säker, jag kommer aldrig mer ha en tråkig stund så länge du finns med mig.

9. REFERENCES

1. Fisher, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, 27, 2985-2993.
2. Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S., *et al.* IsoStar: a library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, 11, 525-537.
3. Gilli, G.; Gilli, P. Towards an unified hydrogen-bond theory. *J. Mol. Struct.* **2000**, 552, 1-15.
4. Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem. Int. Ed.* **2003**, 42, 1210-1250.
5. Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.* **2002**, 41, 2645-2676.
6. Ermondi, G.; Caron, G. Recognition forces in ligand-protein complexes: blending information from different sources. *Biochem. Pharmacol.* **2006**, 72, 1633-1645.
7. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G., *et al.* Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, 47, 3032-3047.
8. Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomolec. Struct.* **2007**, 36, 21-42.
9. Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R., *et al.* Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.* **2009**, 5, 350-358.
10. Murray, C. W.; Verdonk, M. L. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput.-Aided Mol. Des.* **2002**, 16, 741-753.
11. Dunitz, J. D. Win some, lose some - enthalpy-entropy compensation in weak intermolecular interactions. *Chem. Biol.* **1995**, 2, 709-712.
12. Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal. Biochem.* **1989**, 179, 131-137.
13. Middleton, D. A. NMR methods for characterising ligand-receptor and drug-membrane interactions in pharmaceutical research. *Annu. Rep. NMR Spectrosc.* **2007**, 60, 39-75.
14. Mcpherson, A.; Malkin, A. J.; Kuznetsov, Y. G. The science of macromolecular crystallization. *Structure* **1995**, 3, 759-768.
15. Honig, B.; Sharp, K.; Yang, A. S. Macroscopic models of aqueous-solutions - biological and chemical applications. *J. Phys. Chem.* **1993**, 97, 1101-1109.
16. Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, 48, 4040-4048.
17. Guimaraes, C. R. W.; Cardozo, M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.* **2008**, 48, 958-970.
18. Steinbrecher, T.; Case, D. A.; Labahn, A. A multistep approach to structure-based drug design: studying ligand binding at the human neutrophil elastase. *J. Med. Chem.* **2006**, 49, 1837-1844.
19. Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. Accurate and efficient corrections for missing dispersion interactions in molecular Simulations. *J. Phys. Chem. B* **2007**, 111, 13052-13063.

20. Helms, V.; Wade, R. C. Computational alchemy to calculate absolute protein-ligand binding free energy. *J. Am. Chem. Soc.* **1998,** 120, 2710-2713.

21. Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B* **2003,** 107, 9535-9551.

22. Tembe, B. L.; Mccammon, J. A. Ligand receptor interactions. *Comput. Chem.* **1984,** 8, 281-283.

23. Miyamoto, S.; Kollman, P. A. Absolute and relative binding free-energy calculations of the interaction of biotin and Its analogs with streptavidin using molecular-dynamics free-energy perturbation approaches. *Proteins Struct. Funct. Bioinf.* **1993,** 16, 226-245.

24. Cavasotto, C. N.; Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* **2009,** 14, 676-683.

25. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.*, et al.* The protein data bank. *Nucleic Acids Res.* **2000,** 28, 235-242. www.pdb.org.

26. Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004,** 47, 2977-2980.

27. Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.*, et al.* The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005,** 48, 4111-4119.

28. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006,** 49, 6789-6801.

29. Kossiakoff, A. A.; Randal, M.; Guenot, J.; Eigenbrot, C. Variability of conformations at crystal contacts in Bpti represent true low-energy structures - correspondence among lattice packing and molecular-dynamics structures. *Proteins* **1992,** 14, 65-74.

30. Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed.* **2003,** 42, 2718-2736.

31. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004,** 303, 1813-1818.

32. Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* **2010,** 10, 127-141.

33. Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002,** 12, 190-196.

34. Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Nat. Acad. Sci. U.S.A.* **2005,** 102, 6679-6685.

35. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010,** 53, 539-558.

36. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **1999,** 38, 2894-2896.

37. Bergmann, R.; Linusson, A.; Zamora, I. SHOP: Scaffold HOPping by GRID-based similarity searches. *J. Med. Chem.* **2007,** 50, 2708-2717.

38. Adams, C. P.; Brantner, V. V. Spending on new drug development. *Health Econ.* **2010,** 19, 130-141.

39. Linusson, A.; Wold, S.; Norden, B. Statistical molecular design of peptoid libraries. *Mol. Divers.* **1998,** 4, 103-114.

40. Olsson, I. M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab. Syst.* **2004,** 73, 37-46.

41. Dahlgren, M. K.; Kauppi, A. M.; Olsson, I. M.; Linusson, A., *et al.* Design, synthesis, and multivariate quantitative structure-activity relationship of Salicylanilides-potent inhibitors of type III secretion in Yersinia. *J. Med. Chem.* **2007**, 50, 6177-6188.

42. Holm, L.; Frech, K.; Dzhambazov, B.; Holmdahl, R., *et al.* Quantitative structure-activity relationship of peptides binding to the class II major histocompatibility complex molecule A(q) associated with autoimmune arthritis. *J. Med. Chem.* **2007**, 50, 2049-2059.

43. Linusson, A.; Elofsson, M.; Andersson, I. E.; Dahlgren, M. K. Statistical molecular design of balanced compound libraries for QSAR modeling. *Curr. Med. Chem.* **2010**, 17, 2001-2016.

44. Kauppi, A. M.; Andersson, C. D.; Norberg, H. A.; Sundin, C., *et al.* Inhibitors of type III secretion in Yersinia: design, synthesis and multivariate QSAR of 2-arylsulfonylamino-benzanilides. *Bioorg. Med. Chem.* **2007**, 15, 6994-7011.

45. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 983-996.

46. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, 11, 1046-1053.

47. Miranker, A.; Karplus, M. Functionality maps of binding-sites - a multiple copy simultaneous search method. *Proteins* **1991**, 11, 29-34.

48. Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, 6, 211-219.

49. Bohm, H. J. Ludi - rule-based automatic design of new substituents for enzyme-inhibitor leads. *J. Comput.-Aided Mol. Des.* **1992**, 6, 593-606.

50. Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, 4, 649-663.

51. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003**, 2, 527-541.

52. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, 3, 935-949.

53. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J., *et al.* Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Brit. J. Pharmacol.* **2008**, 153, S7-S26.

54. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A., *et al.* Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, 14, 731-751.

55. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, 49, 5851-5855.

56. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, 124, 5632-5633.

57. Breu, B.; Silber, K.; Gohlke, H. Consensus adaptation of fields for molecular comparison (AFMoC) models incorporate ligand and receptor conformational variability into tailor-made scoring functions. *J. Chem. Inf. Model.* **2007**, 47, 2383-2400.

58. Huang, S. Y.; Zou, X. Q. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins Struct. Funct. Bioinf.* **2007**, 66, 399-421.

59. Amaro, R. E.; Baron, R.; McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput.-Aided Mol. Des.* **2008**, 22, 693-705.

60. Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: the role of the receptor structure and ensembles in accurate docking. *Proteins Struct. Funct. Bioinf.* **2008**, 73, 566-580.

61. Gohlke, H.; Thorpey, M. F. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* **2006**, 91, 2115-2120.

62. Zhao, Y.; Sanner, M. F. FLIPDock: docking flexible ligands into flexible receptors. *Proteins Struct. Funct. Bioinf.* **2007**, 68, 726-737.

63. Kazemi, S.; Kruger, D. M.; Sirockin, F.; Gohlke, H. Elastic potential grids: accurate and efficient representation ofintermolecular interactions for fully flexible docking. *Chemmedchem* **2009**, 4, 1264-1268.

64. Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, 47, 435-449.

65. Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, 26, 531-568.

66. B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discov. Today* **2009**, 14, 394-400.

67. Henzler, A. M.; Rarey, M. In pursuit of fully flexible protein-ligand docking: modeling the bilateral mechanism of binding. *Mol. Inf.* **2010**, 29, 164-173.

68. Li, Z.; Lazaridis, T. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.* **2007**, 9, 573-581.

69. Huang, N.; Shoichet, B. K. Exploiting ordered waters in molecular docking. *J. Med. Chem.* **2008**, 51, 4862-4865.

70. Roberts, B. C.; Mancera, R. L. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.* **2008**, 48, 397-408.

71. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R., *et al.* Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, 267, 727-748.

72. Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins Struct. Funct. Bioinf.* **1999**, 37, 228-241.

73. Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, 20, 115-144.

74. Boehm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H., *et al.* Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **2000**, 43, 2664-2674.

75. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, 11, 580-594.

76. Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G., *et al.* Enhancing drug discovery through in silico screening: Strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008**, 15, 2040-2053.

77. Cournia, Z.; Leng, L.; Gandavadi, S.; Du, X., *et al.* Discovery of Human Macrophage Migration Inhibitory Factor (MIF)-CD74 Antagonists via Virtual Screening. *J. Med. Chem.* **2009**, 52, 416-424.

78. Frederick, R.; Robert, S.; Charlier, C.; Wouters, J., *et al.* Mechanism-based thrombin inhibitors: design, synthesis, and molecular docking of a new selective 2-oxo-2H-1-benzopyran derivative. *J. Med. Chem.* **2007**, 50, 3645-3650.

79. Fousteris, M. A.; Papakyriakou, A.; Koutsourea, A.; Manioudaki, M., *et al.* Pyrrolo[2,3-a]carbazoles as potential cyclin dependent kinase 1 (CDK1) inhibitors. Synthesis, biological evaluation, and binding mode through docking simulations. *J. Med. Chem.* **2008,** 51, 1048-1052.

80. Jozwiak, K.; Ravichandran, S.; Collins, J. R.; Moaddel, R., *et al.* Interaction of noncompetitive inhibitors with the alpha 3 beta 2 nicotinic acetylcholine receptor investigated by affinity chromatography and molecular docking. *J. Med. Chem.* **2007,** 50, 6279-6283.

81. Schulz-Gasch, T.; Stahl, M. Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discov. Today Tech.* **2004,** 1, 231-239.

82. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V., *et al.* Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997,** 11, 425-445.

83. Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001,** 44, 1035-1042.

84. Jones, J. E. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc. Lond. A* **1924,** 106, 463-477.

85. Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999,** 34, 4-16.

86. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000,** 295, 337-356.

87. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A., *et al.* Gaussian docking functions. *Biopolymers* **2003,** 68, 76-90.

88. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990,** 112, 6127-6129.

89. Lee, M. C.; Yang, R.; Duan, Y. Comparison between Generalized-Born and Poisson-Boltzmann methods in physics-based scoring functions for protein structure prediction. *J. Mol. Model.* **2005,** 12, 101-110.

90. Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.* **2001,** 123, 5221-5230.

91. Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J. Chem. Inf. Model.* **2006,** 46, 243-253.

92. Lindstrom, A.; Edvinsson, L.; Johansson, A.; Andersson, C. D., *et al.* Post-processing of docked protein-ligand complexes using implicit solvation models. **2010,** (Manuscript).

93. Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? *J. Chem. Inf. Model.* **2009,** 49, 1568-1580.

94. Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003,** 46, 2287-2303.

95. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B., *et al.* A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006,** 49, 5912-5931.

96. Seifert, M. H. J. Targeted scoring functions for virtual screening. *Drug Discov. Today* **2009,** 14, 562-569.

97. Fischer, B.; Fukuzawa, K.; Wenzel, W. Receptor-specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery. *Proteins Struct. Funct. Bioinf.* **2008**, 70, 1264-1273.

98. Knox, A. J. S.; Meegan, M. J.; Sobolev, V.; Frost, D., *et al.* Target specific virtual screening: optimization of an estrogen receptor screening platform. *J. Med. Chem.* **2007**, 50, 5301-5310.

99. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliver. Rev.* **1997**, 23, 3-25.

100. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, 432, 862-865.

101. Sarmiento, M.; Wu, L.; Keng, Y. F.; Song, L., *et al.* Structure-based discovery of small molecule inhibitors targeted to protein tyrosine phosphatase 1B. *J. Med. Chem.* **2000**, 43, 146-155.

102. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P., *et al.* Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, 45, 2213-2221.

103. Shen, J.; Tan, C. F.; Zhang, Y. Y.; Li, X., *et al.* Discovery of potent ligands for estrogen receptor beta by structure-based virtual screening. *J. Med. Chem.* **2010**, 53, 5361-5365.

104. Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, 7, 1047-1055.

105. McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **2007**, 11, 494-502.

106. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P., *et al.* Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, 2, 3256-3266.

107. Sun, H. M. Pharmacophore-based virtual screening. *Curr. Med. Chem.* **2008**, 15, 1018-1024.

108. Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, 50, 74-82.

109. Laggner, C.; Schieferer, C.; Fiechtner, B.; Poles, G., *et al.* Discovery of high-affinity ligands of sigma(1) receptor, ERG2, and emopamil binding protein by pharmacophore modeling and virtual screening. *J. Med. Chem.* **2005**, 48, 4754-4764.

110. Wolber, G.; Langer, T. LigandScout: 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, 45, 160-169.

111. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F., *et al.* A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* **2007**, 47, 279-294.

112. Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M., *et al.* High-throughput virtual screening of proteins using GRID molecular interaction fields. *J. Chem. Inf. Model.* **2010**, 50, 155-169.

113. McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C., *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, 47, 1504-1519.

114. Kruger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *Chemmedchem* **2010**, 5, 148-158.

115. Tan, L.; Geppert, H.; Sisay, M. T.; Gutschow, M., *et al.* Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *Chemmedchem* **2008**, 3, 1566-1571.
116. Muthas, D.; Sabnis, Y. A.; Lundborg, M.; Karlen, A. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. *J. Mol. Graphics Modell.* **2008**, 26, 1237-1251.
117. Kontoyianni, M.; Madhav, P.; Suchanek, E.; Seibel, W. Theoretical and practical considerations in virtual screening: A beaten field? *Curr. Med. Chem.* **2008**, 15, 107-116.
118. Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, 22, 239-255.
119. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn.* **2006**, 27, 861-874.
120. Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, 22, 201-212.
121. von Korff, M.; Freyss, J.; Sander, T. Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **2009**, 49, 209-231.
122. McKay, B.; Hoogenraad, M.; Damen, E. W. P.; Smith, A. A. Advances in multivariate analysis in pharmaceutical process development. *Curr. Opin. Drug Discovery Dev.* **2003**, 6, 966-977.
123. Tye, H. Application of statistical 'design of experiments' methods in drug discovery. *Drug Discov. Today* **2004**, 9, 485-491.
124. Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A. The utility of multivariate design in PLS modeling. *J. Chemom.* **2004**, 18, 156-165.
125. Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for experiments, an introduction to design, data analysis, and model building*. John Wiley & Sons, Inc.: 1978.
126. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, 2, 559-572.
127. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, 2, 37-52.
128. Jackson, J. E. *A user's guide to principal components*. John Wiley & sons, Inc.: New York, 1991.
129. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The collinearity problem in linear-regression - the partial least-squares (Pls) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, 5, 735-743.
130. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, 58, 109-130.
131. St. John, R. C.; Draper, N. R. D-optimality for regression designs: a review. *Technometrics* **1975**, 17, 15-23.
132. DuMouchel, W.; Jones, B. A simple bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics* **1994**, 36, 37-47.
133. Martin, E.; Blaney, J.; Siani, M.; Spellmeyer, D. Measuring diversity - experimental-design of combinatorial libraries for drug discovery. *Abstr. Papers Am. Chem. Soc.* **1995**, 209, 32-Comp.
134. Linusson, A.; Gottfries, J.; Olsson, T.; Ornskov, E., *et al.* Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors. *J. Med. Chem.* **2001**, 44, 3424-3439.
135. Drewry, D. H.; Young, S. S. Approaches to the design of combinatorial libraries. *Chemom. Intell. Lab. Syst.* **1999**, 48, 1-20.

136. Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.* **2000,** 43, 1320-1328.

137. Larsson, A.; Johansson, S. M. C.; Pinkner, J. S.; Hultgren, S. J.*, et al.* Multivariate design, synthesis, and biological evaluation of peptide inhibitors of FimC/FimH protein-protein interactions in uropathogenic Escherichia coli. *J. Med. Chem.* **2005,** 48, 935-945.

138. Marengo, E.; Todeschini, R. A new algorithm for optimal, distance-based experimental design. *Chemom. Intell. Lab. Syst.* **1992,** 16, 37-44.

139. Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.*, et al.* Support vector regression machines. *Adv. Neural Inf. Proc. Syst.* **1996,** 9, 155-161.

140. Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear qsar data using neural networks. *J. Med. Chem.* **1994,** 37, 3758-3767.

141. Niculescu, S. P. Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struct-Theochem* **2003,** 622, 71-83.

142. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. B* **1974,** 36, 111-147.

143. Hotelling, H. The generalization of student's ratio. *Ann. Math. Statist.* **1931,** 2, 360-378.

144. SIMCA-P+, 12.0; Umetrics AB: Box 7960, Umeå, Sweden, **2008**.

145. Eriksson, L.; Verboom, H. H.; Pejnenburg, W. J. G. M. Multivariate QSAR modelling of the rate of reductive dehalogenation of haloalkanes. *J. Chemom.* **1996,** 10, 483-492.

146. Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.*, et al.* Model validation by permutation tests: applications to variable selection. *J. Chemom.* **1996,** 10, 521-532.

147. FRED, 2.0.1; Openeye Scientific Software Inc.: 3600 Cerrillos Road, Suite 1107, Santa Fe, NM 87507, **2004**.

148. Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995,** 245, 43-53.

149. GOLD, 2.2; The Cambridge Crystallographic Datacenter: 12 Union Road, Cambridge, CB2 1EZ, U.K, **2004**.

150. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995,** 247, 536-540.

151. Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins Struct. Funct. Bioinf.* **2006,** 63, 892-906.

152. FRED, 2.2.3; Openeye Scientific Software Inc.: 3600 Cerrillos Road, Suite 1107, Santa Fe, NM 87507, **2007**.

153. GOLD, 3.1.1; The Cambridge Crystallographic Datacenter: 12 Union Road, Cambridge, CB2 1EZ, U.K, **2006**.

154. Arnett, F. C.; Edworthy, S. M.; Bloch, D. A.; Mcshane, D. J.*, et al.* The american-rheumatism-association 1987 revised criteria for the classification of rheumatoid-arthritis. *Arthritis Rheum.* **1988,** 31, 315-324.

155. Andersson, I. E.; Dzhambazov, B.; Holmdahl, R.; Linusson, A.*, et al.* Probing molecular interactions within class II MHC A(q)/Glycopeptide/T-cell receptor complexes associated with collagen-induced arthritis. *J. Med. Chem.* **2007,** 50, 5627-5643.

156. Broddefalk, J.; Backlund, J.; Almqvist, F.; Johansson, M.*, et al.* T cells recognize a glycopeptide derived from type II collagen in a model for rheumatoid arthritis. *J. Am. Chem. Soc.* **1998,** 120, 7676-7683.

157. Holm, B.; Backlund, J.; Recio, M. A. F.; Holmdahl, R., *et al.* Glycopeptide specificity of helper T cells obtained in mouse models for rheumatoid arthritis. *Chembiochem* **2002**, 3, 1209-1222.

158. Backlund, J.; Treschow, A.; Bockermann, R.; Holm, B., *et al.* Glycosylation of type II collagen is of major importance for T cell tolerance and pathology in collagen-induced arthritis. *Eur. J. Immunol.* **2002**, 32, 3776-3784.

159. Dzhambazov, B.; Nandakumar, K. S.; Kihlberg, J.; Fugger, L., *et al.* Therapeutic vaccination of active arthritis with a glycosylated collagen Type II peptide in complex with MHC class II molecules. *J. Immunol.* **2006**, 176, 1525-1533.

160. Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M., *et al.* Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 793-806.

161. OMEGA, 2.2.0; OpenEye Scientific Software Inc.: 3600 Cerrillos Road, Suite 1107, Santa Fe, NM 87507, **2007**.

162. Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative performance assessment of the conformational model generators omega and catalyst: A large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **2006**, 46, 1848-1861.

163. Andersson, E. C.; Hansen, B. E.; Jacobsen, H.; Madsen, L. S., *et al.* Definition of MHC and T cell receptor contacts in the HLA-DR4-restricted immunodominant epitope in type II collagen and characterization of collagen-induced arthritis in HLA-DR4 and human CD4 transgenic mice. *Proc. Nat. Acad. Sci. U.S.A.* **1998**, 95, 7574-7579.

164. Bolin, D. R.; Swain, A. L.; Sarabu, R.; Berthel, S. J., *et al.* Peptide and peptide mimetic inhibitors of antigen presentation by HLA-DR class II MHC molecules. Design, structure-activity relationships, and X-ray crystal structures. *J. Med. Chem.* **2000**, 43, 2135-2148.

165. Boots, A. M. H.; Hubers, H.; Kouwijzer, M.; Zandbrink, L. D., *et al.* Identification of an altered peptide ligand based on the endogenously presented, rheumatoid arthritis-associated, human cartilage glycoprotein-39(263-275) epitope: an MHC anchor variant peptide for immune modulation. *Arthritis Res. Ther.* **2007**, 9.