

# **Development and application of computational methods for the design of bioactive molecules**

**Pablo Englebienne**

Department of Chemistry, McGill University

Montreal, QC, Canada

June 2009

*A thesis submitted to McGill University in partial fulfilment  
of the requirements of the degree of Doctor of Philosophy*

© Pablo Englebienne 2009



## Copyright statement

Some of the material included in the following thesis is adapted from published papers and is under copyright:

Chapter 1 reproduces material published in Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7-S26. © The Authors.

Chapter 2 reproduces material published in Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N., Evaluation of Docking Programs for Predicting Binding of Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160-176. © Wiley-Liss Inc. Used with permission.

Chapter 3 reproduces material published in 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. © American Chemical Society. The ACS grants blanket permission for authors to use their work in their dissertation, as described at the following URL: <http://pubs.acs.org/copyright/forms/dissertation.pdf>

Chapter 4 is a draft of a manuscript to be submitted for publication.

Chapter 5 reproduces material published in Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F., A platinum supramolecular square structure as a G-quadruplex interactive agent. *J. Am. Chem. Soc.* **2008**, *130*, 10040-10041. © American Chemical Society. The ACS grants blanket permission for authors to use their work in their dissertation, as described at the following URL: <http://pubs.acs.org/copyright/forms/dissertation.pdf>.

## Abstract

Drug development is a time-consuming, expensive area of research that requires the collaboration of different fields of expertise. In this context, computer-aided methods have the potential to critically shorten the times and monetary expense required for preclinical development. Very early in a drug discovery effort, docking programs are used to virtually screen large libraries of compounds searching for a lead compound that can be developed into an efficient and safe drug. One of the critical components of a docking program is the one providing an estimate of the binding affinity of the formed complex: the scoring function. The central goal of this work was to develop methods to predict the binding affinity of potential drugs for application in virtual screening campaigns. Towards this goal, we first examined the current status of docking programs and scoring functions applied to a metalloprotein target relevant for medicinal chemistry, Golgi  $\alpha$ -mannosidase II. This triggered a more in-depth analysis of scoring functions, which led us to assemble a set of protein complexes and analyze the performance of different available scoring functions in flexible and solvated proteins. We then set out to develop a scoring function based on molecular mechanical force fields and additional parameters accounting for entropic costs and solvation. We also reached into the development of molecules binding to nucleic acids by developing a hybrid docking/molecular dynamics method to study transition metal complexes binding to G-quadruplexes. Our docking program, FITTED, was prepared for the application to virtual screening campaigns with the implementation of the developed scoring functions and the development of additional tools to manipulate, select and prepare large libraries of ligands. With this goal, we developed a module for FITTED, termed SMART, to prepare ligands prior to the docking. The core of SMART was used to build two other modules, REACTOR and SELECT, that have applications in the preparation and clustering, respectively, of virtual libraries of ligands.



## Résumé

Le développement de médicaments est un domaine de recherche exigeant une quantité considérable de financement et de temps, ainsi que la collaboration étroite de spécialistes dans diverses disciplines des sciences de la vie. Dans ce contexte, les méthodes assistées par ordinateur ont le potentiel d'améliorer la situation. Par exemple, des programmes de docking moléculaire sont souvent utilisés pour faire des criblages virtuels sur de grandes bibliothèques de molécules, à la recherche de candidats qui peuvent être convertis en médicaments efficaces et sécuritaires. L'une des composantes critiques d'un programme de docking est celui chargé de faire l'évaluation de l'affinité du complexe formé entre le ligand et la cible: la fonction de score. L'objectif central de ce travail est d'étudier des méthodes pour estimer l'affinité de médicaments potentiels pour des biopolymères, avec l'intention de les utiliser pour le criblage virtuel. Nous avons tout d'abord examiné la performance des fonctions de score appliquées à une métalloprotéine importante pour le développement de médicaments, la  $\alpha$ -mannosidase II du Golgi. Ensuite, nous avons approfondi cette analyse en sélectionnant un groupe de complexes de protéines qui a été utilisé pour prédire leur affinité tout en considérant la flexibilité conformationnelle et le solvant. Nous avons ensuite développé une fonction de score sur la base des calculs de mécanique moléculaire et de paramètres additionnels prenant en compte l'entropie et la solvation. De plus, nous avons étudié des molécules que interagissent avec des acides nucléiques en développant une méthode hybride de docking/dynamique moléculaire pour l'étude des complexes de métaux de transition avec des G-quadruplexes. Notre programme de docking, FITTED, a été modifié pour entreprendre des études de criblages virtuels avec les nouvelles fonctions de score et le développement de nouveaux outils pour la manipulation, sélection et préparation de grandes bibliothèques de molécules. À cet effet, nous avons préparé les programmes SMART, pour préparer les ligands avant le docking, et REACTOR et SELECT, qui peuvent préparer et filtrer, respectivement, des bibliothèques virtuelles.

## Acknowledgements

None of the work presented in the current manuscript would have been possible without the support of a variety of individuals and organizations. Here goes my deep thanks to them:

- to my supervisor Nicolas Moitessier, for his guidance, patience, support and encouragement;
- to Chris Corbeil, my labmate, for his help with a multitude of issues raised in the almost five years we worked together;
- to the rest of my co-workers in the Moitessier lab through the years: Anupama Haran, Stéphane Bourg, Janice Lawandi, Devin Lee, Rana Bilbeisi, Sabine Thielges, Sylvestre Toumieux, Joris DeSchutter, Mitch Huot, Melanie Burger, Rodrigo Mendoza Sánchez and Eric Therrien;
- to our collaborators: Hélène Fiaux, Lucienne Juillerat-Jeanneret, Hanadi Sleiman, Douglas Kuntz, Constantin Yannopoulos and Jim Gleason, for bringing us interesting problems where we could help;
- to CIHR and the Strategic Training in Chemical Biology, for providing me with funding for a 2-year fellowship and the opportunity to attend an international conference, as well as exposure to first-class speakers through invited seminars;
- to the J.W. McConnell Foundation, for providing me with funding for a McGill Majors Fellowship for 3 consecutive years;
- to the departmental staff, particularly Chantal Marotte and Sandra Aerssen, for efficiently oiling up the rusty machinery when it was needed;
- to the supercomputer centers that generously provided us with CPU time to run many of the simulations reported in this work: CLUMEQ (McGill

University), CERMM (Concordia University) and RQCHP (Université de Sherbrooke);

- to my friends in Montreal: Carolina, Rodrigo, Nick S., Nicolas E., Julie, for the fun we had that made it bearable;
- to my fellow BWIMers: Juan Carlos R.P., Pablo H., Andrés Z., Gustavo M., Marcos M., Germán S., for making those late Tuesday nights worth staying up for;
- to my family, especially my mother, Elizabeth and George, for giving me the courage and the support to reach my goals.

And finally, I would like to very specially thank my wife Roxanne: you gave me strength when I was weak, you gave me words when I was stuck, and most importantly, you gave me love always. *TQM*.

## Contribution of authors

The present thesis contains five original research chapters (Chapters 2-6). Of these, Chapters 2 and 3 have been published as full articles, while Chapter 4 is a draft to be submitted for publication. Chapter 5 contains material published as a communication.

Professor Nicolas Moitessier, as my supervisor, is a co-author of all the articles presented.

Chapter 2: I prepared, with assistance from H       Fiaux, all the structures of the mannosidase complexes for docking and virtual screening, run the calculations and analyzed the data; Chris Corbeil ran the FITTED calculations. H       Fiaux (with supervision from Sandrine Gerber-Lemaire) synthesized the new mannosidase inhibitors reported, while Douglas Kuntz and David Rose performed the crystallographic studies.

Chapter 3: I prepared all the structures of the sets of complexes described, ran all the calculations and analyzed the data.

Chapter 4: I prepared the structures of the sets of complexes described for the development of RankScore 2 and 3, ran the calculations and analyzed the data. I wrote Python scripts for the analysis of the correlation of scoring functions and for the tuning of the parameter set.

Chapter 5: I developed the MM parameters from *ab initio* data, run the MD simulations and analyzed the data. Roxanne Kieltyka (with supervision from Prof. Hanadi Sleiman, Dept. Chemistry, McGill University) synthesized the platinum complexes reported and performed the biophysical assays, while Johans Fakhoury (with supervision from Chantal Autexier, Dept. Biomedical Engineering, McGill University) performed the biochemical assays.

Chapter 6: The initial development of SMART was performed in collaboration with Christopher Corbeil and Nicolas Moitessier. The application of REACTOR to the

development of HDAC inhibitors is being performed in collaboration with Melanie Burger and Rodrigo Mendoza Sánchez (with supervision from Prof. Jim Gleason, Dept. Chemistry, McGill University).

# Table of contents

## Preface

Copyright statement .....	i
Abstract.....	ii
Résumé.....	iii
Acknowledgements.....	iv
Contribution of authors .....	vi
Table of contents.....	viii
List of figures.....	xii
List of tables.....	xviii
List of equations.....	xx
List of abbreviations .....	xxii

<b>Chapter 1: Introduction .....</b>	<b>1</b>
<b>1.1 Drug discovery and design .....</b>	<b>1</b>
1.1.1 Drugs.....	1
1.1.2 Drug discovery and development.....	1
1.1.3 Experimental measurement of binding affinities .....	3
1.1.4 High throughput screening.....	3
1.1.5 Computer-aided drug design .....	4
1.1.6 Virtual screening.....	4
<b>1.2 Energetics of ligand-biomacromolecule binding.....</b>	<b>5</b>
1.2.1 Thermodynamical considerations.....	5
1.2.2 Electrostatic contributions .....	6
1.2.3 van der Waals forces .....	7
1.2.4 Hydrogen bonds .....	8
1.2.5 $\pi$ interactions.....	10
1.2.6 Solvation effects .....	11
1.2.7 Entropic considerations.....	12
<b>1.3 Docking and scoring.....</b>	<b>13</b>
1.3.1 Empirical scoring functions.....	13
1.3.2 Force-field based scoring functions.....	15
1.3.3 Knowledge-based scoring functions.....	15
<b>1.4 Free energy calculations .....</b>	<b>16</b>
1.4.1 Free energy perturbation (FEP) and thermodynamic integration (TI) ..	17
1.4.2 Linear interaction energy (LIE) .....	18
1.4.3 MM-PB/SA and MM-GB/SA .....	19
<b>1.5 Quantum mechanics-based binding affinity calculations .....</b>	<b>19</b>
<b>1.6 Thesis objectives.....</b>	<b>20</b>
1.6.1 General objectives .....	20
1.6.2 Evaluation of docking programs for Golgi $\alpha$ -mannosidase inhibitors....	20
1.6.3 Assessment of scoring functions for flexible docking.....	21
1.6.4 Development of scoring functions for protein-ligand interactions .....	21
1.6.5 Modelling of platinum complexes as G-quadruplex binders .....	22
1.6.6 Development of SMART, REACTOR and SELECT.....	22
<b>1.7 References.....</b>	<b>23</b>
<b>Chapter 2: Case study: crystallographic and docking studies of Golgi <math>\alpha</math>-mannosidase II inhibitors.....</b>	<b>33</b>

<b>2.1</b>	<b>Introduction.....</b>	<b>33</b>
2.1.1	Protein glycosylation.....	33
2.1.2	Golgi $\alpha$ -mannosidase and cancer .....	33
2.1.3	Goals of this research .....	35
<b>2.2</b>	<b>Materials and methods .....</b>	<b>35</b>
2.2.1	Enzyme assays and crystallography.....	35
2.2.2	Preparation of structures for docking .....	36
2.2.3	Glide.....	36
2.2.4	Glide VS .....	37
2.2.5	FlexX .....	37
2.2.6	AutoDock.....	37
2.2.7	eHiTS.....	38
2.2.8	GOLD.....	38
2.2.9	LigandFit.....	39
2.2.10	FITTED.....	39
<b>2.3</b>	<b>Results and discussion .....</b>	<b>39</b>
2.3.1	Docking data set .....	39
2.3.2	Crystallography .....	40
2.3.3	Docking of GMII inhibitors: general considerations .....	48
2.3.4	Docking methods used in this comparative study and parameterization.....	50
2.3.5	Application to the docking of mannosidase inhibitors.....	54
2.3.6	Metal ligation .....	57
2.3.7	Docking to conformational ensembles and flexible proteins .....	58
2.3.8	Scoring accuracy.....	60
2.3.9	Virtual screening using Glide .....	61
<b>2.4</b>	<b>Conclusions .....</b>	<b>64</b>
<b>2.5</b>	<b>References.....</b>	<b>65</b>
 <b>Chapter 3: Assessment of the performance of scoring functions on</b>		
	<b>complexes with flexible and solvated proteins .....</b>	<b>73</b>
<b>3.1</b>	<b>Introduction.....</b>	<b>73</b>
<b>3.2</b>	<b>Methods.....</b>	<b>75</b>
3.2.1	Training set selection criteria.....	75
3.2.2	Correlation metrics and statistical significance .....	76
3.2.3	Training set preparation.....	77
3.2.4	Cross-scoring training set .....	81
<b>3.3</b>	<b>Results and discussion .....</b>	<b>81</b>
3.3.1	Accuracy of the selected scoring functions on the entire set.....	81
3.3.2	Accuracy of the selected scoring functions within protein classes.....	89
3.3.3	Hydropathicity and accuracy.....	90
3.3.4	Scoring function correlation .....	91
3.3.5	Consensus scoring.....	93
3.3.6	Impact of protein conformation and water molecules .....	94
<b>3.4</b>	<b>Conclusions .....</b>	<b>98</b>
<b>3.5</b>	<b>Experimental section.....</b>	<b>99</b>
3.5.1	Selection of the training set structures .....	99
3.5.2	Preparation of the training set.....	100
3.5.3	Scoring.....	103
3.5.4	Bootstrap analysis .....	104
<b>3.6</b>	<b>References.....</b>	<b>104</b>

<b>Chapter 4: Development of scoring functions for virtual screening from force field data .....</b>	<b>109</b>
<b>4.1 Introduction.....</b>	<b>109</b>
<b>4.2 Results and discussion.....</b>	<b>110</b>
4.2.1 RankScore .....	110
4.2.2 Screened force fields .....	111
4.2.3 Training sets and scoring function.....	112
4.2.4 General considerations.....	112
4.2.5 Force field energy terms.....	113
4.2.6 RankScore2, 3 and 4: novel scoring functions.....	114
4.2.7 RankScore formalism.....	115
4.2.8 Development of RankScore2 and RankScore3.....	119
4.2.9 Development of RankScore4.....	122
4.2.10 Application of the RankScore scoring functions to benchmark sets ....	123
<b>4.3 Conclusions.....</b>	<b>124</b>
<b>4.4 Experimental Section .....</b>	<b>125</b>
4.4.1 Preparation of the training set structures .....	125
4.4.2 Derivation of additional parameters for force fields.....	125
4.4.3 Force field charges .....	126
4.4.4 Development and validation of RankScore2 and RankScore4.....	126
<b>4.5 References.....</b>	<b>127</b>
<b>Chapter 5: Modeling of supramolecular compounds binding to guanine quadruplex DNA structures.....</b>	<b>131</b>
<b>5.1 Introduction.....</b>	<b>131</b>
5.1.1 Cancer .....	131
5.1.2 Telomeres.....	131
5.1.3 Telomerase .....	132
5.1.4 Guanine quadruplexes.....	132
5.1.5 G-quadruplex binders as telomerase inhibitors.....	133
5.1.6 Molecular modeling of G-quadruplex binders .....	133
5.1.7 Summary of work presented in this chapter .....	134
5.1.8 Acknowledgments.....	134
<b>5.2 Results and discussion.....</b>	<b>135</b>
5.2.1 Development of molecular mechanical parameters for platinum (II) complexes with heteroaromatic ligands .....	135
5.2.2 Platinum (II) square complex.....	137
5.2.3 Extended $\pi$ -surface platinum (II) complexes as G-quadruplex binders.....	141
5.2.4 Hydrogen-bonded ligands for an even larger $\pi$ -surface.....	145
<b>5.3 Conclusions.....</b>	<b>148</b>
<b>5.4 Methods.....</b>	<b>149</b>
5.4.1 Development of molecular mechanics parameters for Pt complexes..	149
5.4.2 Docking.....	150
5.4.3 Molecular dynamics simulations .....	150
5.4.4 Binding affinity calculations.....	150
<b>5.5 References.....</b>	<b>151</b>
<b>Chapter 6: Development of programs for the handling of ligands for virtual screening: SMART, REACTOR and SELECT .....</b>	<b>157</b>
<b>6.1 Introduction.....</b>	<b>157</b>



6.1.1	Ligand treatment for docking.....	157
6.1.2	Molecular mechanical force fields .....	157
6.1.3	Atomic charges.....	157
6.1.4	Virtual libraries of ligands .....	158
6.1.5	Combinatorial libraries .....	158
6.1.6	Virtual SAR (VSAR) .....	159
6.1.7	Fingerprints and similarity metrics .....	159
6.1.8	Goals .....	160
<b>6.2</b>	<b>Methods.....</b>	<b>163</b>
6.2.1	SMART.....	163
6.2.2	REACTOR.....	170
6.2.3	SELECT.....	176
<b>6.3</b>	<b>Application to the design of HDAC inhibitors.....</b>	<b>177</b>
<b>6.4</b>	<b>Conclusions .....</b>	<b>179</b>
<b>6.5</b>	<b>Future developments and applications.....</b>	<b>180</b>
<b>6.6</b>	<b>References.....</b>	<b>181</b>
<b>Chapter 7:</b>	<b>Contributions to knowledge.....</b>	<b>187</b>
<b>7.1</b>	<b>Conclusions .....</b>	<b>187</b>
7.1.1	Evaluation of docking programs for Golgi $\alpha$ -mannosidase .....	187
7.1.2	Assessment of scoring functions for flexible docking.....	187
7.1.3	Development of scoring functions for protein-ligand interactions .....	187
7.1.4	Modelling of platinum complexes as G-quadruplex binders .....	188
7.1.5	Development of programs for the handling of ligands for virtual screening: SMART, REACTOR and SELECT .....	188
<b>7.2</b>	<b>Papers and conference presentations .....</b>	<b>189</b>
7.2.1	Papers published.....	189
7.2.2	Book chapters.....	190
7.2.3	Forthcoming papers .....	190
7.2.4	Conference presentations .....	190
<b>Appendix</b>	<b>193</b>	
<b>A.1</b>	<b>Supplementary information for Chapter 3 .....</b>	<b>193</b>
<b>A.2</b>	<b>Supplementary information for Chapter 4 .....</b>	<b>210</b>
<b>A.3</b>	<b>Supplementary information for Chapter 5 .....</b>	<b>211</b>
A.3.1	Fluorescence resonance energy transfer assay (FRET) experiment details .....	211
A.3.2	TRAP assay protocol .....	212
<b>A.4</b>	<b>Circular dichroism (CD) study .....</b>	<b>213</b>
<b>A.5</b>	<b>References.....</b>	<b>216</b>

## List of figures

Figure 1.1. Lennard-Jones potential. Depicts the interaction of a pair of oxygen atoms with $\sigma=3.816$ Å and $\epsilon=0.1094$ kcal/mol. ....	8
Figure 1.2. Hydrogen bonding interactions in biopolymers. Top: generic arrangement for hydrogen bonding; D is a HBD, A is a HBA. Middle: peptide hydrogen bonding; backbone hydrogen bonding, His-Asp/Glu side chain hydrogen bonding. Bottom: nucleic acid base pairing; A-T base pair, G-C base pair. ....	9
Figure 1.3. Hydrogen bonding interactions in biopolymers. a: peptide $\beta$ -sheet; b: peptide $\alpha$ -helix; c: DNA duplex (G-C and A-T base pair indicated). This figure was prepared with PyMol. ....	9
Figure 1.4. $\pi$ - $\pi$ interactions: parallel stacking (shown with 30° displacement), edge-to-face, C-H/ $\pi$ , cation- $\pi$ . ....	10
Figure 1.5. Sample potential of mean force for (left) a positively charged nitrogen interacting with a negatively charge oxygen and (right) a pair of aliphatic carbons. ....	16
Figure 1.6. Thermodynamical cycle of alchemical transformation of protein-ligand complexes; P is a protein, A and B are ligands. $\Delta\Delta G_{A-B}=\Delta G_A-\Delta G_B=\Delta G_1-\Delta G_2$ . ....	17
Figure 2.1 Electron density representation of the inhibitors 8, 9 and 10 bound in the active site of dGMII. Maps are simulated annealing omit maps ( $F_o-F_c$ ) of only the inhibitors contoured at 3.5 $\sigma$ . For orientation purposes the active site zinc ion is represented as a magenta ball. This figure was generated with PyMOL. ....	40
Figure 2.2 a) Interaction of 8 with residues in the active site of dGMII. Interactions closer than 3.2 Å are indicated with cyan dotted lines; interactions with the zinc ion are indicated in magenta. Water molecules appear as orange balls. Distances are presented in Table 2.3. This figure was generated with PyMOL. ....	45
Figure 2.3 Comparison of the conformation of the active site residues in the complexes of 8 (ligand green, protein grey) and 10 (ligand and protein orange) with dGMII. The zinc atoms appear as magenta balls. This figure was generated with PyMOL. ....	47
Figure 2.4 Overlays of the binding of 8 with (a) swainsonine 1 (PDB code 1HWW) and (b) the diastereomer of salacinol 7 (PDB code 1TQT). 8 is drawn in green, 7 and 1 are drawn in grey. The active site zinc is represented as a magenta ball. ....	48
Figure 2.5 Rigid protein docking. Accuracy for the 7 programs assessed to dock the 10 ligands to the 10 protein structures (self-docking and cross-docking) measured as the RMSD between the docked poses and the crystal structure binding mode for (a) all ligand heavy atoms; (b) metal-binding atoms. ....	55

Figure 2.6 Glide docked vs. crystallographically observed binding mode of compound 8 (a) and 7 (b). The crystal structure appears in green, the docked structure in orange. This figure was generated with PyMOL.....	57
Figure 2.7 Superposition of 1HWW, 1HXK, 1R34, 1TQT and 2F18 protein structures (backbone green, side chains grey), and ligand 8 (green). Only selected residues of each structure are shown to illustrate the largest displacements. This figure was generated with PyMOL. ....	59
Figure 2.8 Docking to protein ensembles. Accuracy for the 7 programs assessed to dock the 10 ligands to an ensemble of the 10 protein structures, measured as the RMSD between the docked poses and the crystal structure binding mode for all ligand heavy atoms.....	60
Figure 2.9 Number of known active compounds recovered as a function of the percentage of the ranked list. Dark blue: standard protonation states, light blue: best and worse scores with various protonation states, brown: random.....	63
Figure 3.1. Properties of the training set. a) Profile of molecular weight and binding affinity dependence; b) distribution of ligand MW; c) distribution of ligand binding affinities.....	78
Figure 3.2. Spearman (a) and Kendall (b) coefficients for three subsets of the training set: in blue, 209 complexes (whole set); in red, 5 outliers removed (204 complexes); in green, all transition metal-containing proteins removed (188 complexes). ....	85
Figure 3.3. Comparison of the Spearman correlation for ten scoring functions. Blue columns indicate the correlation of the “no outliers” set (204 complexes); red bars denote the Spearman coefficients as reported by Wang <i>et al.</i> <sup>12</sup> .....	86
Figure 3.4. Comparison of Spearman (a) and Kendall (b) coefficients for the scoring functions considered. Error bars calculated through bootstrap method (see Methods section) .....	87
Figure 3.5. Accuracy ( $\tau$ ) of the scoring functions on 3 subsets of set 1. ....	90
Figure 3.6. Accuracy (Kendall $\tau$ correlation) of the scoring functions on three subsets of Set 1, classified by the hydrophaticity of the binding sites. Error bars calculated with the bootstrap method. ....	91
Figure 3.7. Ranked-list correlation coefficients ( $\tau$ ) calculated between predicted ranking lists of SFs; the darker the shading, the higher the correlation (see key). The numbers 1-18 represent the scoring functions as specified in the top row. ....	93
Figure 3.8. $\tau$ calculated for combinations of scoring functions. The numbers 1-18 on the second row and leftmost columns represent the scoring functions as	

described in the top row of Figure a). (a) Pairs of scoring functions: darker boxes represent increase of the combined scoring functions over the individual scoring functions (see key; value in key indicates the increase of the value of  $\tau$  over the value for the individual scoring function in the same row). The yellow boxes correspond to the individual scoring functions. (b) Groups of three scoring functions: darker boxes represent increase with respect to XScore alone (yellow box)..... 94

Figure 3.9. Accuracy ( $\tau$ ) of the scoring functions on set 2 with different water considerations. Waters were kept for docking and scoring (wet/wet, blue), kept for docking and removed for scoring (dry/dry, red), removed for docking and scoring (dry/dry, green) or made displaceable (yellow). For comparison, the correlation of the scores obtained for the native structures in wet/wet conditions is shown (light blue)..... 96

Figure 3.10. Accuracy of the scoring functions ( $\tau$ ) on the complete set 2 when protein conformational ensembles (*best scored complex among the cross-docked ones*) were considered..... 98

Figure 4.1 Correlation of van der Waals and electrostatic terms for the different force fields in use..... 114

Figure 4.2. Ligand/protein complex formation ..... 115

Figure 4.3. Frozen bond determination. Atoms 1 and 3 have van der Waals contacts (distance between van der Waals surfaces < 0.5 Å) with the protein (purple surface), while atoms 2 and 4 do not. In this case, the bond between atoms 1 and 2 would have a *frozen* value of 0.75, while the bond between atoms 2 and 4 would have a value of 0.25. .... 118

Figure 4.4. Flexibility of side chain atoms. The further away from the peptide backbone an atom is, the more flexible it is considered, therefore the entropic penalty upon binding will be larger. .... 119

Figure 4.5. Procedure used to derive RankScore2 and RankScore3. In each case, 25,000 loops were performed, yielding an identical amount of sets of weights with their associated Kendall  $\tau$  value. .... 121

Figure 4.6. Distribution of scaling factors for van der Waals (blue) and electrostatic (red) interactions. .... 121

Figure 4.7. Pearson correlation and Kendall  $\tau$  for a variety of scoring functions when applied to the testing set of 93 complexes. .... 124

Figure 5.1. Guanine quadruplex hydrogen-bonding structure showing Hoogsteen base-pairing. .... 133

Figure 5.2. Sample force field modification file (frcmod) for a platinum (II) square planar complex containing two aromatic nitrogen ligands and two aliphatic nitrogen ligands. ....	137
Figure 5.3. Structure of molecular square 1. ....	138
Figure 5.4. Docking of platinum square 1 to G-quadruplex structure 1kf1. (a) Representative docked conformations; (b) stacked conformation. This pictures were produced with PyMol. <sup>42</sup> .....	139
Figure 5.5. FRET stabilization curve of square 1 with quadruplex (red) and duplex DNA (blue). ....	140
Figure 5.6. TRAP assay of complex 1, showing ladders generated by the action of telomerase on a TS primer (PCR amplified). The lower band is an internal control primer (ITAS). ....	141
Figure 5.7. Structures of platinum (II) complexes: platinum ethylenediamine 2,2'-bipyridyl, Pt(en)bipy, 2; platinum ethylenediamine phenylphenanthroimidazole, Pt(en)PIP, 3; platinum ethylenediamine naphthylphenanthroimidazole, Pt(en)PIN, 4. ....	142
Figure 5.8. Foldings of the human G-quadruplex considered. a: parallel structure (PDB code 1kf1); b: anti-parallel structure (PDB code 143d). This pictures were produced with PyMol. <sup>42</sup> .....	142
Figure 5.9. Schematic depiction of the binding modes considered as starting structures for complexes interacting with different foldings of a G-quadruplex. (a) Poses interacting with 1kf1; yellow: top; purple: bottom; teal: loop; red: groove. (b) Poses interacting with 143d; yellow: handles; purple: bottom; red: groove. ....	143
Figure 5.10. Binding affinities for stacking modes of complexes 2, 3 and 4. In blue, binding to parallel G-quadruplex folding (PDB code 1kf1); in red, binding to anti-parallel G-quadruplex folding (PDB code 143d). Error bars correspond to one standard deviation for the binding energy as calculated from the molecular dynamics snapshots. ....	145
Figure 5.11. Structures of platinum (II) complexes: platinum ethylenediamine indolylphenanthroimidazole, Pt(en)PII, 5; platinum ethylenediamine quinolylphenanthroimidazole, Pt(en)PIQ, 6. ....	146
Figure 5.12. The biphenyl torsional angle is shown in blue. ....	147
Figure 5.13 Binding affinities for complexes 1-5 calculated with the MM-PBSA method. Error bars are one standard deviation. ....	148

Figure 6.1. Molecular fingerprints. The two-dimensional structures of neuraminidase inhibitors oseltamivir 1 and zanamivir 2 are converted into a fingerprint based on their molecular features. ....	160
Figure 6.2. Ligand properties assigned by SMART.....	161
Figure 6.3 Information flow in REACTOR. ....	162
Figure 6.4. SELECT. Left to right: given a query molecule and a Tanimoto cutoff, the program can retrieve a subset of compounds from a library. Right to left: given a library and a Tanimoto cutoff, SELECT can extract a representative library of compounds with increased diversity.....	162
Figure 6.5. Hydrogen bond donor and acceptor groups. Left: hydrogen bond donor atoms shown in blue; right: hydrogen bond acceptor atoms coloured red. R groups can be hydrogen, alkyl, aryl or acyl, among other possibilities. Double-bonded O and S can be attached (*) to C, S, N or P.....	168
Figure 6.6. Rotatable bond assignment. Rotatable bonds assigned for docking are shown with red arrows; bonds considered rotatable for scoring are shown with blue arrows. ....	169
Figure 6.7. Bond order assignment in a molecule containing only C, H and O atoms. Bonds with (yet) unassigned order are pictured with dashed bonds, bonds assigned after each step are coloured red.....	170
Figure 6.8. Bond order assignment in a heterocycle containing nitrogen atoms. The same colour convention as in previous figure applies, and in addition a bond order guess is coloured blue.....	170
Figure 6.9. Functional group depiction: a central atom is bound to accessory atoms. ....	171
Figure 6.10. Sample <code>rules_definitions.txt</code> file.....	172
Figure 6.11. Example of a rule for peptide bond formation. Top: content of <code>rules.txt</code> for a peptide bond formation (in bold, rules matched by bottom structures). Bottom: schematic representation of compounds matching Rule1 and Rule2, with atoms colour-coded for their function in the rule.....	173
Figure 6.12. Example for a rule specified with dummy atoms. content of <code>rules.txt</code> for formation of a bond between fragments with X1 and X2 dummy atoms respectively. Bottom: schematic representation of compounds matching Rule1 and Rule2, with atoms colour-coded for their function in the rule; * specifies any atom. ....	174
Figure 6.13. Schematic of the algorithm of REACTOR. ....	174

Figure 6.14. Assignment of new torsions upon product formation. Top and middle: the values for the torsions where the new bond is terminal (i.e., 1-2-3-4 and 3-4-5-6) are assigned the value of a torsion from the reactant (i.e., 1-2-3- $X_1$ and $X_2$ -4-5-6 respectively). Bottom: the torsion around the newly formed bond (i.e., 2-3-4-5) is scanned to minimize clashes. ....	175
Figure 6.15. Histone deacetylase inhibitors. Top: HDACis trichostatin A 1 and suberoylanilide hydroxamic acid (SAHA) 2. Bottom: schematic representation of a library of HDAC inhibitors. In yellow, a capping group; in green, a hydrophobic linker of varying length; in yellow, a zinc-chelating moiety. ....	177
Figure 6.16. Workflow for the generation of a combinatorial library from libraries of 2D or 3D fragments. Libraries of three-dimensional fragments are fed to reactor, which outputs a combinatorial library with all possible products. The redundancy of this library can be reduced by removing compounds below a Tanimoto cutoff with SELECT. Finally, the filtered library can be fed to SMART and then docked with FITTED. ....	178
Figure A.1. Titration of complex 1 (0.1 to 1.25 $\mu$ M) with G-quadruplex DNA. ....	210
Figure A.2. Percent activity of telomerase upon addition of increasing amounts of complex 1. ....	211
Figure A.3. Titration of complex 1 with a 3 $\mu$ M solution of G-quadruplex in a sodium phosphate buffer. The solid black line represents the G-quadruplex with no complex added. There is a slight decrease in the peak near 300 nm and an increase in the signal around 250 nm. ....	212
Figure A.4. Titration of complex 1 with a 3 $\mu$ M solution of G-quadruplex in potassium phosphate buffer. Solid black line is the G-quadruplex without complex 1. Successive additions of complex 1 lead to a slight decrease in the peak at 295 nm with each aliquot. ....	213
Figure A.5. Heating and cooling of complex 1 and the G-quadruplex in a sodium phosphate buffer. The solid black line represents G-quadruplex after annealing. The dotted line is the result of the ....	213
Figure A.6. Heating and cooling of platinum square 1 and G-quadruplex in a potassium phosphate buffer. Black solid line represents the annealed G-quadruplex. The dotted line shows the effect of addition of the platinum molecular square 1 to the annealed G-quadruplex. The dashed line is the result after heating the platinum molecular square with the G-quadruplex and cooling the mixture to room temperature. There appears to be a slight change in the conformational preference from the original quadruplex structure upon the addition of the platinum square complex 1. However, after heating and cooling this signal is also maintained. ....	214

## List of tables

Table 2.1 Selection of structures of $\alpha$ -mannosidase/inhibitor complexes. The inhibitory activity ( $IC_{50}$ ) was measured on dGMII (EC 3.2.1.114) at 37°C, pH = 5.75. All protein crystal structures correspond to <i>Drosophila melanogaster</i> GMII. ....	41
Table 2.2 Data collection and structural refinement statistics.....	43
Table 2.3 Summary of interatomic distances (Å) between the inhibitors and dGMII. Distances in bold represent distances greater than 3.5 Å, where no significant hydrogen bonding is expected to occur.....	46
Table 2.4 Structures of $\alpha$ -mannosidase inhibitors used for virtual screening evaluation. The inhibitory activity is given as $IC_{50}$ on <i>Drosophila</i> GMII unless otherwise noted. ....	62
Table 2.5 Results for the virtual screening of mannosidase inhibitors. The ranking corresponds to the order of the compound in the sorted list of scores; the score is the GlideScore value of the best docked pose for the compound. ....	64
Table 3.1. Proteins represented more than once. Set 1 is the self-docking set while Set 2 includes the cross-docked structures. The complete sets are given in the appendix (Tables A.1 and A.2). ....	79
Table 3.2. Selected scoring functions used in this study.....	83
Table 3.3. Accuracy of the scoring functions on the complete Set 1 and on two reduced sets compared to previously reported data. <sup>a</sup> .....	88
Table 4.1. AUC for the docking of libraries containing about 1000 ligands and decoys to 11 proteins (6 in the training set and 5 in the testing set).....	124
Table 5.1. Binding affinity of platinum (II) complexes to human telomere G-quadruplex parallel folding (1kf1) and antiparallel folding (143d) by the MM-PBSA method on 5 different initial conformations for each. Errors are one standard deviation from the average. ....	144
Table 5.2 MM-PBSA binding affinities for the different platinum(II) complexes to the human telomere G-quadruplex motif in a stacking binding mode. Energies in kcal/mol.....	148
Table 6.1. GAFF atom types defined by SMART.....	163
Table 6.2. List of functional groups recognized by SMART. In the figures, blue lowercase indicates a GAFF atom type, while a green uppercase label denotes an element. ....	165
Table A.1. Description of the 209 complexes included in Set 1. ....	191



Table A.2. Listing of 87 complexes included in Set 2.....	199
Table A.3. Accuracy of the scoring functions on 5 subsets of set 1. Each cell presents the ranked correlation coefficient $\tau$ for each scoring function within a subset of proteins.....	202
Table A.4. Accuracy ( $\tau$ ) of the scoring functions on three subsets of Set 1.....	204
Table A.5. Accuracy of the scoring functions ( $\tau$ ) on set 2 when waters are kept for docking and scoring, kept for docking and removed for scoring, removed for both docking and scoring, or made displaceable.....	206
Table A.6. Accuracy of the scoring functions ( $\tau$ ) on the complete set 2 when protein conformational ensembles were considered. ....	207
Table A.7. Correlations between predicted score and experimental binding affinity for the test set of 93 complexes.....	208

## List of equations

Equation 1.1. Chemical equilibrium for drug binding. L is a ligand, R a receptor, C a complex.....	5
Equation 1.2. Association constant for ligand L to receptor R. ....	5
Equation 1.3. Dissociation constant for ligand L to receptor R.....	5
Equation 1.4. Relationship between Gibbs free energies of formation ( $\Delta G^\circ$ ) and equilibrium constants ( $K$ ).....	5
Equation 1.5. Relationship between free energies ( $\Delta G^\circ$ ), enthalpies ( $\Delta H^\circ$ ) and entropies ( $\Delta S^\circ$ ) of formation. ....	6
Equation 1.6. Coulomb's law. $q_i$ and $q_j$ are the point atomic charges, $\epsilon$ denotes the dielectric constant of the medium and $r_{ij}$ is the interatomic distance.....	7
Equation 1.7. Lennard-Jones potential with two alternate sets of parameters. $A_{ij}$ , $B_{ij}$ and $\sigma_{ij}$ are parameters (fit to experimental data) specific for each (i,j) atom pair; $r_{ij}$ is the interatomic distance. ....	7
Equation 1.8. Empirical scoring functions, exemplified by functional form of ChemScore. $\Delta G_i$ are coefficients obtained by regression. Subindices denote different interaction types; <i>HB</i> : hydrogen bonds, <i>met</i> : metal interaction, <i>lipo</i> : lipophilic interaction, <i>rot</i> : entropic penalty for frozen rotors. $f(\Delta r)$ is a certain function of interatomic distance, $f(\Delta \alpha)$ is a certain function of torsional angle, $N'_{rot}$ is a count of rotatable bonds (see text) .....	14
Equation 1.9. Force field-based scoring functions. See captions for Equation 1.6 and Equation 1.7 for a description of the symbols. ....	15
Equation 1.10. Free energy perturbation theory. $\Delta E = E_B - E_A$ , $\langle \rangle$ denotes an ensemble average.....	17
Equation 1.11. Thermodynamic integration. $\lambda$ denotes a parameter varying between 0 and 1 for states A and B respectively; $H_\lambda$ is the energy of the system as a function of $\lambda$ ; $\langle \rangle$ denotes an ensemble average. ....	18
Equation 1.12. Linear interaction energy method. $\alpha$ and $\beta$ are regression-based parameters modulating the conformational ensemble average van der Waals and electrostatic contributions. ....	18
Equation 1.13. MM-PBSA and MM-GBSA methods. $G$ is the free energy of binding, $E_{MM}$ is the MM energy as calculated with a given forcefield, $E_{solv}$ is the desolvation energy calculated by either GB/SA or PB/SA methods, $TS$ is the entropic energy. $\langle \rangle$ denotes an ensemble average.....	19

Equation 4.1. Free energy of binding.....	115
Equation 4.2. Decomposition of the free energy complex formation.....	115
Equation 4.3. Change of enthalpy of complex for formation of the ligand-protein complex, calculated as the difference of force field energies of the complex, unbound ligand and protein with bound waters ( $E_i^{FF}$ ).....	116
Equation 4.4. The entropic contribution on the ligand side is calculated as a function of the number of rotatable bonds ( $N_{rot}$ ), affected by the polarity of the bond and the buriedness of the bond, as estimated by the numbe of contacts with the protein (see text). .....	117
Equation 4.5. The entropic contribution to the binding from the protein is estimated by scaling down the interaction with the ligand ( $E_{prot-lig}$ ) calculated through a force field by a factor $\lambda$ .....	119
Equation 4.6. Free energy of solvation is calculated as a function of the solvent-accessible surface area (SASA) and the generalized Born approach.....	119
Equation 4.7. Functional form of RankScore2. The different terms relate the different components ( $E^{vdW}$ , $E^{elec}$ , $E^{HB}$ ) of the intermolecular energy in the complex ( $E_{complex}$ ), the generalized Born polar solvation energy ( $\Delta G_{GB}$ ), the non-polar solvation energy proportional to the SASA ( $\Delta G_{SASA}$ ), the number of captured water molecules ( $N_{wat}$ ) and the number of rotatable bonds ( $N_{rot}$ ). ..	121
Equation 4.8. Functional form of RankScore4. See caption to Equation 4.7 for description of the terms. ....	123
Equation 6.1. Tanimoto coefficient. $T(A,B)$ is the Tanimoto similarity between A and B, that is, the ratio between the bits that are “on” in both A and B ( $A \cap B$ ) and the total number of bits ( $A \cup B$ ). In the third term, $\chi_{iA}$ is the i-th bit of item A ( $\chi=0$ if off, $\chi=1$ if on), and analogously for $\chi_{iB}$ .....	160

## List of abbreviations

AMBER	Assisted model building with energy refinement
CADD	Computer-aided drug design
CD	Circular dichroism
CLS	Crystallographic ligand structure
DFT	Density functional theory
dGMII	<i>Drosophila melanogaster</i> Golgi $\alpha$ -mannosidase II
DNA	Deoxyribonucleic acid
ER	Estrogen receptor
FF	Force field
FITTED	Flexibility induced through targeted evolutionary description
FRET	Fluorescence resonance energy transfer
G4	Guanine quadruplex
GA	Genetic algorithm
GAFF	General Amber force field
GB	Generalized Born
GMII	Golgi $\alpha$ -mannosidase II
HBA	Hydrogen bond acceptor
HBD	Hydrogen bond donor
HDAC	Histone deacetylase
hGMII	Human Golgi $\alpha$ -mannosidase II
HIVP	Human immunodeficiency virus 1 (HIV-1) protease
HTS	High-throughput screening
IC <sub>50</sub>	Inhibitory concentration for 50% of the population
LJ	Lennard-Jones
MC	(Metropolis) Montecarlo

MD	Molecular dynamics
MLS	Energy-minimized ligand structure
MM	Molecular mechanics
MMP	Matrix metalloprotease
MM-PBSA	Molecular mechanics/Poisson-Boltzmann/Surface area
MMFF	Merck molecular force field
MW	Molecular weight
N <sub>rot</sub>	Number of rotatable bonds
PB	Poisson-Boltzmann
PDB	Protein Data Bank
PROCESS	Protein conformational ensemble system setup
QSAR	Quantitative structure-activity relationship
REACTOR	Rapid enumeration by an automated combinatorial tool for organic reactions
RMSD	Root-mean square deviation
SASA	Solvent-accessible surface area
SF	Scoring function
SMART	Small molecule atomtyping and rotatable torsion assignment
TRAP	Telomere repeat amplification protocol
VDR	Vitamin D receptor
vdW	van der Waals
VS	Virtual screening

# Chapter 1: Introduction

## 1.1 Drug discovery and design

### 1.1.1 Drugs

A *drug*, in the widest definition of the term, is a chemical substance used in the treatment, cure, prevention, or diagnosis of a condition or used to otherwise enhance physical or mental well-being.<sup>1</sup> As such, they have been present in human life from the beginning of history, having been linked to spiritual and religious use through the ages. In molecular terms, and in the context of the work presented in this thesis, a drug is a chemical substance that exerts a biological effect on an organism by *binding* to, or otherwise *activating*, a biomacromolecular receptor (most commonly a protein or a nucleic acid), leading to a measurable biological response, or lack thereof (e.g., in the case of receptor antagonists).

Traditionally herbs, plant leaves and roots, as well as mushrooms, have been used as treatment and cures for ailments, as well as for the mind- and perception-altering states they provoke. More recently, it has been shown that chemicals isolated from extracts of these organisms have drug qualities; these are the so-called *natural products*. Natural product chemistry is an endless source of inspiration for organic chemists, out of the desire to reproduce complex chemical structures found in nature,<sup>2,3</sup> as well as for medicinal chemists, for the very desirable biological activities they exhibit.<sup>4,5</sup>

### 1.1.2 Drug discovery and development

#### 1.1.2.1 Serendipity

Many of the early drugs used in modern pharmacology were discovered by random events. Alexander Fleming, for example, reportedly discovered the antibiotic penicillin when he noticed the inhibition of bacterial growth around a mold contaminant (a strain of *Penicillium notatum*) in a *Staphylococcus aureus* gel culture

plate in 1928, a discovery that earned him the 1945 Nobel Prize in Medicine or Physiology.<sup>6</sup> Dorothy C. Hodgkin, using X-ray crystallography, solved the chemical structure of penicillin only in 1949 (this accomplishment partly led to her being awarded the Nobel Prize in Chemistry in 1964).<sup>7</sup> The mechanism of action of penicillin troubled scientists for some more time, until Park and Strominger defined it in 1957.<sup>8</sup> Unsatisfactory pharmacokinetics (rapid drug clearance through the kidneys) as well as the narrow spectrum of activity (it was only active against certain Gram-positive bacterial strains) were thought to be the drug's demise, although chemical alterations of the side-chains around the  $\beta$ -lactam core led to clinically useful antibiotics.<sup>9</sup>

### **1.1.2.2 Rational drug design**

Although serendipitous discoveries yielded important advances in drug development, they are an unreliable source of new treatments. Rational drug design, on the other hand, attempts to follow a scientific method for the development of novel bioactive molecules. As a field of research, it is strongly related to the concept of "magic bullet" developed by German scientist Paul Ehrlich, co-winner of the 1908 Nobel Prize in Physiology or Medicine. An experienced histologist, Ehrlich recognized that the biological effect of a chemical agent was given simultaneously by the chemical structure of the agent, as well as by the cellular target on which it acts.<sup>10</sup> This gave rise to the *lock-and-key* concept (first proposed by Emil Fischer in 1894<sup>11</sup>) by which a drug and the receptor it binds to share complementary features.<sup>12</sup> With this in mind, drug design can be viewed as, given a *lock* (a biomolecular target), trying to design a *key* (a drug) to act on it.

Rational drug design is an iterative and sequential process requiring, first of all, the identification of a plausible *target* to be acted on.<sup>13</sup> The latter can be of a varied nature; drugs have exploited enzymes, protein receptors, DNA sequences, organelles and membranes (or other organisms' boundaries, such as cell walls, lipopolysaccharides or viral capsides) as molecular targets. Once a target has been determined and validated, the rational drug design protocol starts from a *hit*

compound (e.g., the natural ligand or substrate of the target, or an unrelated compound identified by screening), which is found to be at least moderately active against the target. Increase of the effectiveness through chemical modification turns the hit into a *lead* compound, which is further transformed into a *drug candidate* by improving its pharmacological (i.e., pharmacokinetics and pharmacodynamics) and safety profiles on cell cultures and animal models. A promising drug candidate would be assayed in clinical trials, upon which demonstration of effectiveness and safety would lead to its consideration as a *drug*.

### **1.1.3 Experimental measurement of binding affinities**

Binding of a small molecule to a biomolecule can be quantified in different ways. Biochemical methods act by measuring a biological effect, for example the loss of enzymatic activity, the inhibition of bacterial growth or reduction in tumour size. The result of these experiments is usually reported as the amount of drug required to provoke the effect on half of the systems observed (e.g.,  $IC_{50}$ ,  $LD_{50}$ , etc.). On the other hand, biophysical methods attempt to detect binding by measuring changes in some physical properties of the system. Isothermal titration calorimetry (ITC)<sup>14</sup> and surface plasmon resonance (SPR),<sup>15</sup> among others (e.g., spectrometric), fall into this category. With these methods, thermodynamical parameters of the binding ( $\Delta G$ ,  $\Delta H$ ,  $\Delta S$ ) can be directly measured or calculated.

### **1.1.4 High throughput screening**

Beginning in the mid-1980's, the pharmaceutical industry implemented radical changes in the assaying of biological activity for drug design.<sup>16</sup> These developments involved a drastic reduction in sample volumes and the use of multiple-well plates instead of single tubes, leading to a large increase in assay throughput. In the late 1990's and early 2000's, further miniaturization and automation, together with developments in combinatorial synthesis revolutionized the way drug development was done.<sup>17</sup> In high throughput screening (HTS) campaigns, libraries of compounds (in the  $10^4$ - $10^6$  range) are quickly assayed for activity against a biological target. Once potentially active compounds are identified, slow-throughput assaying at



multiple concentrations is performed. One of the weakest points of HTS is the large amount (~10% or higher<sup>17</sup>) of false positives (*i.e.*, compounds that are incorrectly flagged as active) reported; besides some of them appearing as an artefacts of the assay itself, aggregation caused by promiscuous inhibitors has been shown to be a relevant matter.<sup>18,19</sup>

### **1.1.5 Computer-aided drug design**

Modern drug design efforts rely heavily on the use of computers at some stage.<sup>20,21</sup> Computational methods were first used for drug design in the 1960's, when X-ray crystal structures of drug-protein complexes were first solved with the use of computers. Computer-aided drug design (CADD) methods can, in the broadest sense, be separated into structure-based drug design, or receptor-based methods, and ligand-based methods, mostly represented by quantitative structure-activity relationship (QSAR) techniques and chemical similarity search methods.<sup>22</sup> While the former require the availability of structural information about the receptor (in decreasing order of confidence: 3-D crystallographic models, NMR structures, pharmacophores, homology models), the latter can be applied when little or no information about the receptor is available, but binding affinities for families of compounds are on hand.

### **1.1.6 Virtual screening**

Besides classical HTS (see above), CADD methods can be applied to identify new biologically active small molecules, in virtual screening (VS) campaigns.<sup>23-27</sup> In the latter, virtual libraries of compounds are assessed by CADD methods for their affinity towards targets of interest, with molecular docking being the most popular technique (see section 1.3). With around 500 proteins currently being targeted by available drugs and an estimate 10,000 druggable targets (*i.e.*, their activity can be modulated by small molecules),<sup>28</sup> the increase in screening throughput attainable by screening libraries of compounds virtually with computer-aided techniques might pave the way for the treatment of new diseases. Besides these differences at the

receptor space level, the ligand chemical space explorable by virtual methods is orders of magnitude larger than the one attainable by synthetic efforts.<sup>29,30</sup>

## 1.2 Energetics of ligand-biomacromolecule binding

### 1.2.1 Thermodynamical considerations

Most drugs (only about 5% of drugs in the market are covalent) bind non-covalently to their intended target, which allows for the formation of a chemical equilibrium between the ligand, the biomolecular receptor and the complex (Equation 1.1).

**Equation 1.1.** Chemical equilibrium for drug binding. L is a ligand, R a receptor, C a complex.



**Equation 1.2.** Association constant for ligand L to receptor R.

$$K_a = \frac{[C]}{[L] \times [R]}$$

**Equation 1.3.** Dissociation constant for ligand L to receptor R.

$$K_d = \frac{[L] \times [R]}{[C]}$$

The equilibrium constant of this transformation ( $K_a$ , or association constant, Equation 1.2) dictates the free energy of binding (or binding affinity) of the ligand for this specific receptor (Equation 1.4); the higher the constant, the higher the binding affinity. A more common representation of the binding affinity is given by the reciprocal of the association constant,  $K_d$  (Equation 1.3), termed the dissociation constant. The advantage of the latter is that the value of  $K_d$  has units of concentration (e.g., mol/L), and represents the concentration of free ligand such that 50% of the binding sites are occupied. A change in  $K$  of an order of magnitude corresponds to  $\sim 1.4$  kcal/mol change in free energy of binding at room temperature (from Equation 1.4).

**Equation 1.4.** Relationship between Gibbs free energies of formation ( $\Delta G^\circ$ ) and equilibrium constants ( $K$ ).

$$\Delta G^0 = -RT \ln K$$

**Equation 1.5.** Relationship between free energies ( $\Delta G^\circ$ ), enthalpies ( $\Delta H^\circ$ ) and entropies ( $\Delta S^\circ$ ) of formation.

$$\Delta G^0 = \Delta H^0 - T \times \Delta S^0$$

The free energy of binding has itself enthalpic and entropic contributions (Equation 1.5); a more favourable (more negative) change in enthalpy of binding or an increase in entropy will decrease the free energy of binding. It is important to recall that the binding process takes place in aqueous solutions, hence the participation of water molecules in the energetics of binding cannot be neglected.

In many fields of biomolecular simulations, such as docking, the free energy of binding is decomposed in a series of additive terms, what Southall et al. call the BIPSE model: “Break Into Pieces, Sum the Energies”.<sup>31</sup> The following sections describe the different components of the free energy of binding from the BIPSE perspective.

### 1.2.2 Electrostatic contributions

Biomacromolecules (proteins and nucleic acids) are charged molecules at physiological pH. In proteins, ionizable side chains (Asp, Glu, His, Lys, Arg) are charged at physiological pH (with the possible exception of His,  $pK_a \sim 7$ ), while in nucleic acids the phosphate diester backbone makes the chain highly negatively charged. In addition to formal charges, the presence of heteroatoms in aminoacids and nucleotides leads to the existence of permanent and induced dipoles. With this in mind, the electrostatic interactions among ions, permanent dipoles and induced dipoles make a large contribution to the binding affinity in a drug-target complex. Coulomb’s law (Equation 1.6) is the most common treatment of electrostatic interactions in biomolecular simulations; the energy of the electrostatic interaction is calculated as a pair-wise potential, proportional to the product of the charges of both atoms and inversely proportional to their distance, with the proportionality constant being given by the dielectric constant  $\epsilon$ .

**Equation 1.6.** Coulomb's law.  $q_i$  and  $q_j$  are the point atomic charges,  $\epsilon$  denotes the dielectric constant of the medium and  $r_{ij}$  is the interatomic distance.

$$E_{coul} = \sum_{i,j \neq i} \frac{q_i \times q_j}{\epsilon \times r_{ij}}$$

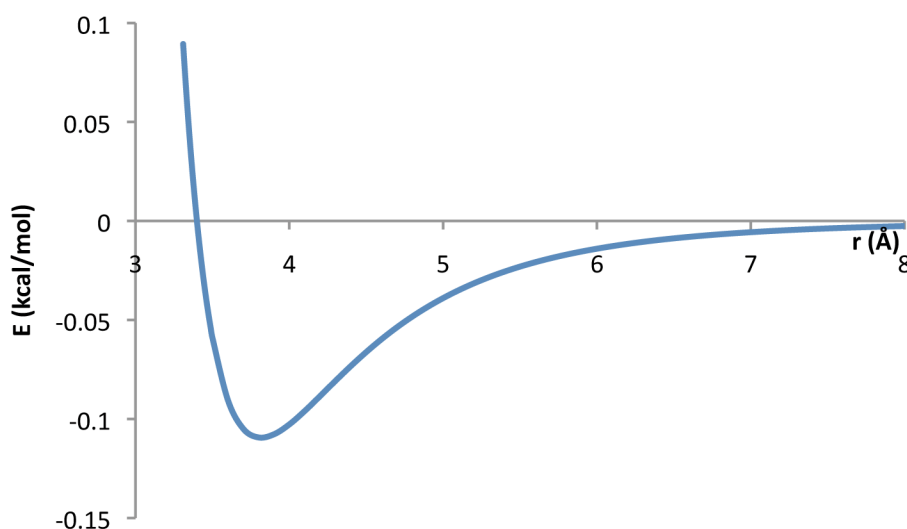
The source of the point charges for molecular mechanical simulations is a delicate topic, as force field energies are meant to be self-consistent with one charge derivation mechanism. Charges are usually derived by fitting the electrostatic potential generated by point charges to the one obtained at a high-level of theory, most commonly DFT (B3LYP/6-31G), HF or semiempirical methods such as AM1-BCC.<sup>32,33</sup> A large enough parameter set can allow for charges among pairs of atoms to be converted to bond increments,<sup>34</sup> therefore reducing the computational expense of the charge assignment greatly.

### 1.2.3 van der Waals forces

These are short-range interactions that arise from the interpenetration of the electron clouds when atoms become in contact through London forces. They are also referred to as “non-polar” forces as they are the dominating contribution in the interaction among non-polar groups; the low polarizability of electronegative atoms such as oxygen and nitrogen when compared to a methylene group partly accounts for this effect.

**Equation 1.7.** Lennard-Jones potential with two alternate sets of parameters.  $A_{ij}$ ,  $B_{ij}$  and  $\sigma_{ij}$  are parameters (fit to experimental data) specific for each (i,j) atom pair;  $r_{ij}$  is the interatomic distance.

$$E_{vdW} = \sum_{i,j \neq i} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) = \sum_{i,j \neq i} \epsilon_{ij} \times \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \times \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

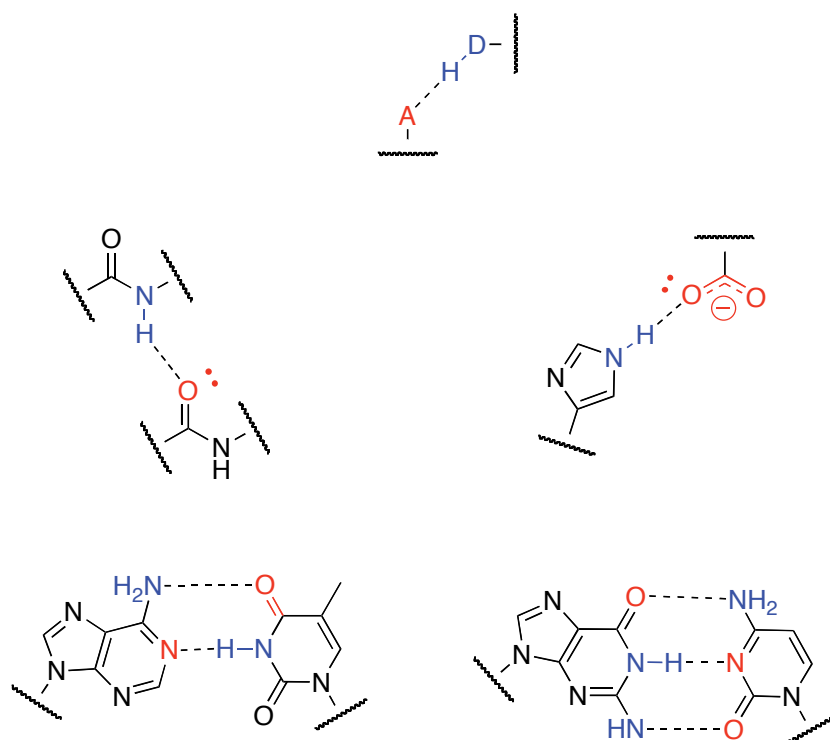


**Figure 1.1.** Lennard-Jones potential. Depicts the interaction of a pair of oxygen atoms with  $\sigma=3.816$  Å and  $\epsilon=0.1094$  kcal/mol.

The interaction is weakly attractive at relatively large distances, while at short distances it becomes highly repulsive. Most commonly, this effect is modelled as a combination of terms of a Lennard-Jones potential (Equation 1.7): an unfavourable (repulsive) term acting at short distances (usually scaling as  $r^{-12}$ ), and a favourable (attractive) term that becomes dominant at larger distances (scaling as  $r^{-6}$ , or as an exponential).

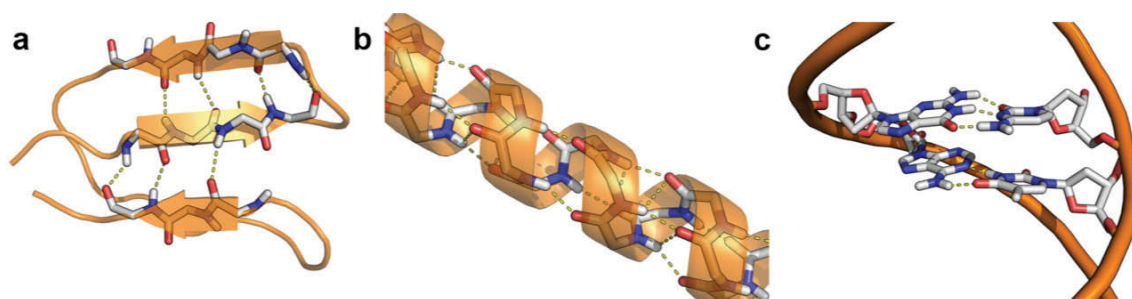
#### 1.2.4 Hydrogen bonds

A hydrogen bond formed between a hydrogen atom covalently bound to an electronegative atom (e.g., -OH in alcohol, water and carboxylic acids; -NH in amines, amides, ammoniums and heterocycles) that is shared with another electronegative atom having a free pair of electrons (e.g., oxygen in carbonyls, alcohols and water; nitrogen in aromatic heterocycles). The former is termed the hydrogen bond donor, HBD, while the latter is referred to as the hydrogen bond acceptor, HBA.



**Figure 1.2.** Hydrogen bonding interactions in biopolymers. Top: generic arrangement for hydrogen bonding; D is a HBD, A is a HBA. Middle: peptide hydrogen bonding; backbone hydrogen bonding, His-Asp/Glu side chain hydrogen bonding. Bottom: nucleic acid base pairing; A-T base pair, G-C base pair.

Biopolymers exhibit a number of polar moieties capable of acting as hydrogen bonds donors and/or acceptors. In fact, hydrogen bonding is the main feature leading to the secondary structure of proteins (e.g.,  $\alpha$ -helix,  $\beta$ -sheet) and nucleic acids (e.g., double helix, quadruplexes).



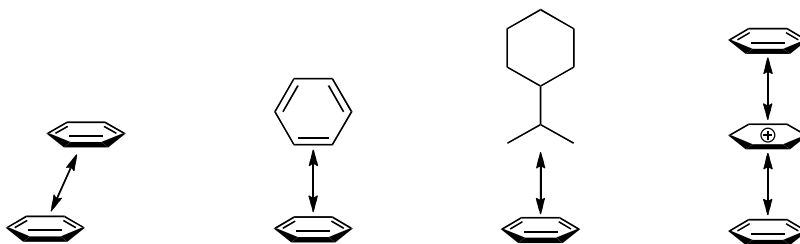
**Figure 1.3.** Hydrogen bonding interactions in biopolymers. a: peptide  $\beta$ -sheet; b: peptide  $\alpha$ -helix; c: DNA duplex (G-C and A-T base pair indicated). This figure was prepared with PyMol.

Molecules binding to biopolymers usually make extensive use of the possibility of hydrogen bonding with their target, as the short range and the directionality (i.e.,

optimal orbital overlaps favour strong hydrogen bonds) enforced by this type of interaction convey them a powerful mechanism for selectivity.<sup>35</sup> A single hydrogen bond between neutral groups has been calculated to add 0.5-1.5 kcal/mol to the binding affinity, while the interaction involving an ionic group can add up to 4.7 kcal/mol.<sup>36,37</sup> Weaker hydrogen bonds (~1 kcal/mol) are feasible when the hydrogen bond acceptor is an aromatic ring.<sup>38</sup>

### 1.2.5 $\pi$ interactions

The flat, delocalized cloud of electrons of aromatic groups confers them distinct properties.<sup>39</sup> Aromatic rings play an important role in the structural properties of proteins and nucleic acids, and are a key component of the molecular recognition process.<sup>40,41</sup> The stacking of rings can lead to orientations that are especially favoured: the “sandwich” or parallel stacking, and the T-shaped or edge-to-face (Figure 1.4).



**Figure 1.4.**  $\pi$ - $\pi$  interactions: parallel stacking (shown with 30° displacement), edge-to-face, C-H/ $\pi$ , cation- $\pi$ .

Additionally to the interactions between  $\pi$ -systems, interactions with entities of a different nature are also relevant (Figure 1.4). A weak type of hydrogen bond (~0.5-1.5 kcal/mol) is found between the non-traditional carbon donor and an aromatic acceptor, in the termed C-H/ $\pi$  interactions.<sup>42</sup> Cation- $\pi$  interactions are responsible for the orientation of basic aminoacids (Arg, Lys and His) in proximity of aromatic residues (Phe, Trp and Tyr).<sup>43</sup> These cation- $\pi$  interactions have been measured to be up to 18 kcal/mol in the case of  $K^+$  and benzene, in the same range of the water solvation energy for  $K^+$ . Aromatic systems are also able to act as hydrogen bond acceptors (*vide supra*).

### 1.2.6 Solvation effects

It is important to bear in mind that, cells being composed mostly of water, most interactions between small molecules and biopolymers occur in aqueous solution; the role of water in ligand binding is, therefore, not trivial. On one hand, its large dielectric constant modulates the electrostatic interactions between charged groups. On the other hand, the property of water molecules to act as hydrogen bond donors and acceptors means that they occupy the sites available for interaction with a ligand; this requires the solvent to be displaced prior to the formation of hydrogen bonds between the biopolymer and a ligand. With this in mind, one could define two types of solvent molecules: bulk water and point water.

There are two ways to consider solvation effects in computer simulations: explicitly or implicitly. In the former, atomistic models of solvent molecules are considered part of the system, usually with limitations in order to reduce the complexity of the system. Multiple explicit water models (e.g., SPC,<sup>44</sup> TIP3P,<sup>45</sup> TIP4P,<sup>45</sup> TIP5P<sup>46</sup>) have been developed over the years; they differ in the geometry of the molecules (O-H bond length, H-O-H angle) and the non-bonded parameters (Lennard-Jones parameters, partial charge values and placement). Implicit solvation, also known as continuum-solvent methods, involves simulating the presence of the solvent by treating the electrostatics of the system in a special way, mainly by solving the Poisson-Boltzmann equation (PB) or by using the generalized Born model (GB). Both methods of solvation have their advantages and disadvantages. On one hand, the explicit solvent molecules add computational expense to the simulations; on the other hand, the use of implicit solvation leads to the loss of detail on the specific water interactions with ligand and biopolymers.

Besides the direct interactions with water, mostly leading to polar interactions, the desolvation of hydrophobic groups is favoured by exchanging disfavoured interactions between water molecules and the hydrophobic groups while forming favoured water-water interactions and the formation of new favoured van der Waals contacts between the non-polar portions. The interaction of hydrophobic



groups between a ligand and a receptor gives rise to a contribution to binding affinity of about  $28 \text{ cal mol}^{-1} \text{ \AA}^{-2}$ .<sup>47</sup> This is an overly simplistic model, as the energetics of non-polar contacts depend not only on the surface area, but also on the structure and geometry of the interface.<sup>48,49</sup> However, the solvent-accessible surface area (SASA) has been found to be proportional to the hydrophobic binding affinity.<sup>50</sup>

### 1.2.7 Entropic considerations

The factors considered so far have an impact mostly in the enthalpic term of the free energy of binding, however the changes in entropy upon ligand binding are also relevant.<sup>51</sup> Entropy is a measure of the chaos of a system; the more configurations a system is able to achieve, the higher its entropy will be. With this in mind, as binding is a process that brings two molecules together, it reduces the degrees of freedom of the system. On the other hand, the solvent molecules displaced upon binding partially counteract this effect. As a result, the change in entropy of binding is mostly negative, hence opposed to the process. The enthalpic contribution to the free energy of binding is usually a fairly large, favourable amount, while the entropic term ( $T\Delta S$ ) is unfavourable, and about the same order of magnitude. This fact, known as enthalpy-entropy compensation, results in binding free energies being orders of magnitude smaller than the individual contributions.<sup>52</sup>

The most obvious entropic contribution to model is the one arising from the loss of degrees of freedom from the torsions that become frozen upon binding. When in solution ligand torsions are mostly free to rotate, but when interacting with a biopolymer the restrictions imposed by the binding site topology lead to much of this freedom being lost. With this in mind, a term proportional to the number of rotatable bonds present in a molecule has been used as an estimator for the entropic penalty upon binding.<sup>53,54</sup> In multi-conformational assessment of structures, a quasi-harmonic analysis can be used to estimate the entropy of binding.<sup>55</sup>

### 1.3 Docking and scoring

One of the easiest-to-grasp, graphical and powerful methods in computational drug design is molecular docking.<sup>56-61</sup> Essentially, it consists of finding the best relative orientation (i.e., the best *pose*) of a ligand in a macromolecular binding site, thus constituting a global optimization problem. To achieve this, potential binding modes have to be generated and their fitness assessed. In the first implementation of a docking program (DOCK),<sup>62</sup> the ligand and the receptor were considered both rigidly, that is, the program would evaluate the match of a single conformation of the ligand (considered rigid, but translating the centre of mass and rotating it in space) and the receptor at a time. With increasing computing power, conformational sampling engines were embedded into docking programs, hence increasing the search space many-fold and leading to better predictions.

The best binding mode is assessed by an estimate of the binding affinity on a single pose by a *scoring function*. The goal of the latter is two-fold: i) it should assign a better score to the native (i.e., experimentally observed) pose of a ligand; ii) it should assign a better score to a stronger binder. From this perspective, it is clear that the two goals can be better achieved by two different functions: one (simpler, faster) that can distinguish among *different conformers of the same compound*, thus guiding the docking; another function (more complex, slower) that can distinguish final docked poses of *different compounds*, thus providing a way to discriminate compounds based on their predicted affinity. The following sections describe scoring functions according to their derivation.

#### 1.3.1 Empirical scoring functions

Since the pioneering work of Böhm in the development of LUDI,<sup>53</sup> a significant amount of work has been devoted to the development and improvement of empirical scoring functions (SFs). With empirical SFs, the evaluation of the energetics of the ligand binding (mostly derived from protein/ligand crystal structures) is decomposed into simpler, scalable contributions arising from, for example, hydrogen bonds, metal ligation, hydrophobic effects and freezing of

rotatable bonds (Equation 1.8). The various scaling factors ( $\Delta G_i$  in Equation 1.8) are then defined by regression to fit experimentally determined protein-ligand affinities.

**Equation 1.8.** Empirical scoring functions, exemplified by functional form of ChemScore.  $\Delta G_i$  are coefficients obtained by regression. Subindices denote different interaction types; *HB*: hydrogen bonds, *met*: metal interaction, *lipo*: lipophilic interaction, *rot*: entropic penalty for frozen rotors.  $f(\Delta r)$  is a certain function of interatomic distance,  $f(\Delta \alpha)$  is a certain function of torsional angle,  $N'_{rot}$  is a count of rotatable bonds (see text)

$$\Delta G_{bind} = \Delta G^0 + \Delta G_{HB} \sum_{HB} f(\Delta r) f(\Delta \alpha) + \Delta G_{met} \sum_{met} f(\Delta r) + \Delta G_{lipo} \sum_{lipo} f(\Delta r) + \Delta G_{rot} N'_{rot}$$

Among the most commonly used SFs is ChemScore<sup>54</sup> (Equation 1.8), which has been implemented in various docking programs (e.g., GOLD,<sup>63,64</sup> FRED<sup>65</sup>). Standalone SFs have also been devised and include X-Score,<sup>66</sup> DrugScore,<sup>67,68</sup> VALIDATE<sup>69</sup> and HINT.<sup>70</sup> Each empirical SF differs by the number and nature of the terms used to make up its equation. For instance, several include an explicit directional hydrogen bond energy term (e.g., ChemScore, X-Score and the SFs implemented in eHiTS,<sup>71</sup> FlexX,<sup>72</sup> and Surflex<sup>73</sup>), while only a few include an explicit directional metal-ligand interaction term (e.g., eHiTS, Surflex and X-Score). Functions such as the eHiTS and PLP<sup>74</sup> SFs evaluate the internal energy of the ligand in its bound conformation, while solvation and/or predicted captured water molecules (within GlideScore<sup>75</sup>) are computed in a different manner. Many empirical scoring functions take into account the hydrophobic effect in the binding, mostly either by computing the hydrophobic surface buried in the complex (e.g., SCORE1/2,<sup>53,76</sup> LigScore<sup>77</sup>), or by evaluating the match of the hydrophobicity of an atom with its environment (e.g., FlexX, SCORE,<sup>78</sup> SLIDE<sup>79</sup>), while several combine both approaches (e.g., eHiTS, GlideScore, HammerHead,<sup>73</sup> X-Score<sup>66</sup>). On the other hand, HINT<sup>70</sup> computes the logP of the ligand as a measure of its water solvation. The entropic contribution to the binding energy due to the freezing of torsional degrees of freedom upon binding is often estimated by a term proportional to the number of sp<sup>3</sup>-sp<sup>3</sup> and sp<sup>2</sup>-sp<sup>3</sup> rotatable bonds. In some cases (e.g., ChemScore,<sup>54</sup> GlideScore,<sup>75</sup> VALIDATE,<sup>69</sup> X-Score<sup>66</sup>), the environment of a bond is taken into consideration to assess the extent of its effect,

while RankScore<sup>80</sup> attempts to include the freezing of protein side chains by scaling the interaction with flexible side chains.

### 1.3.2 Force-field based scoring functions

Force fields (FFs) were originally developed to reproduce conformational behaviour and thermodynamic and kinetic properties of small molecules and macromolecules. When applied to protein-ligand complexes, FFs are often found to significantly overestimate the binding affinity (Equation 1.9) even when applied in conjunction with highly accurate, time-consuming techniques (e.g., Linear Interaction Energy method), which consider the bulk water either explicitly or implicitly.<sup>81</sup> Scaling factors applied to the non-bonded terms (van der Waals and electrostatics) were found to restore part of the predictiveness of FFs in this area.<sup>82,83</sup>

**Equation 1.9.** Force field-based scoring functions. See captions for **Equation 1.6** and **Equation 1.7** for a description of the symbols.

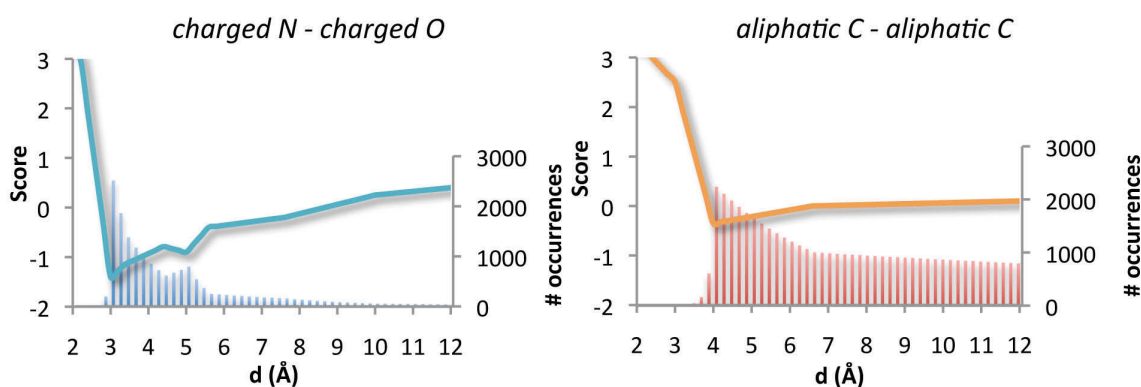
$$\Delta G_{bind} = \sum_i^{lig} \sum_j^{rec} \left[ \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + 332.0 \frac{q_i q_j}{\epsilon r_{ij}} \right]$$

When compared to empirical SFs, a smaller number of functions were developed exclusively from FFs. More commonly, non-bonded FF terms (illustrated in Equation 1.9) are combined with terms from empirical SFs, such as the solvation and ligand entropy terms in the AutoDock SF.<sup>84</sup> The choice of force field parameters is varied: AutoDock,<sup>84</sup> DOCK<sup>85</sup> and RankScore<sup>80</sup> SFs combine the van der Waals, electrostatic and hydrogen bond interaction energy computed using the AMBER force field (sometimes with additional non-FF terms), while GoldScore<sup>64</sup> makes use of the Tripos FF parameters and ICM<sup>86</sup> implements a hybrid AMBER-ECEPP/2 approach.

### 1.3.3 Knowledge-based scoring functions

Other popular SFs, such as DrugScore<sup>67,68</sup> and PMF,<sup>87,88</sup> have been developed from statistical analysis of crystal structures of ligand-protein complexes. These analyses report the distribution of ligand-protein atom type pairs (histograms in Figure 1.5) and convert this data into pairwise potentials (curves in Figure 1.5). When

considering the interaction between charged species (left), there is often a sharp preference (maximum in the histogram; minimum in the score) at a relatively close distance, and a secondary extreme at a larger separation that accounts for the interaction *via* a bridging water molecule. In contrast, the potential for a pair of aliphatic carbons (right) shows little preference over a wide range of interatomic distances. The score is calculated by the sum of all interaction pairs between each ligand and protein atom lying within a sphere of a given cutoff (usually 6-12 Å). Although these functions are expected to capture all the data needed for predicting the free energy of binding, some of the interactions are underrepresented in the available crystal structures (e.g. interactions with metals and/or halogens) and are not well parameterized. As for force field-based SFs, correcting/additional terms were implemented as exemplified by the solvation term included in DrugScore.<sup>68</sup>



**Figure 1.5.** Sample potential of mean force for (left) a positively charged nitrogen interacting with a negatively charge oxygen and (right) a pair of aliphatic carbons.

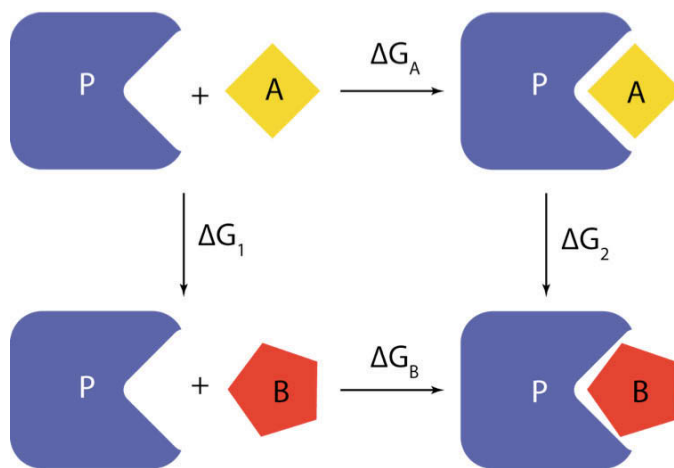
## 1.4 Free energy calculations

Scoring functions attempt to predict the binding affinity of a pair of molecules by assessing a single conformation; this is clearly a simplification of the problem, as the binding process is a dynamic one. The binding energy is in reality determined by a Boltzmann distribution of binding modes. Additionally, most scoring functions

disregard the effect of water on binding (both as a solvent and as relaying interactions).

Molecular dynamics (MD) simulations<sup>89</sup> provide a time-evolution of a system following Newton's equation of motion, while Monte Carlo (MC) simulations<sup>90</sup> provide a stochastic sampling of the different conformations attainable by a system. Both methods can yield an ensemble of binding modes considering explicit solvent molecules.

#### 1.4.1 Free energy perturbation (FEP) and thermodynamic integration (TI)



**Figure 1.6.** Thermodynamical cycle of alchemical transformation of protein-ligand complexes; P is a protein, A and B are ligands.  $\Delta\Delta G_{A-B} = \Delta G_A - \Delta G_B = \Delta G_1 - \Delta G_2$ .

MD or MC simulations on alchemical systems (Figure 1.6) can be used to calculate relative binding affinities by gradually converting one ligand into the other. Two statistical mechanics-derived formalisms can be applied for this purpose: FEP (Equation 1.10) and TI (Equation 1.11).<sup>91</sup> In both methods, the relative binding affinity of two ligands, i.e.,  $\Delta\Delta G_{A-B} = \Delta G_B - \Delta G_A$ , is calculated through the thermodynamical cycle depicted in Figure 1.6. The free energy changes  $\Delta G_1$  and  $\Delta G_2$  are calculated by performing simulations where one ligand is gradually converted into the other, by altering the MM parameters describing them.

**Equation 1.10.** Free energy perturbation theory.  $\Delta E = E_B - E_A$ ,  $\langle \rangle$  denotes an ensemble average.

$$\Delta G_{A \rightarrow B} = G_B - G_A = -RT \ln \left\langle e^{-\Delta E / RT} \right\rangle$$

**Equation 1.11.** Thermodynamic integration.  $\lambda$  denotes a parameter varying between 0 and 1 for states A and B respectively;  $H_\lambda$  is the energy of the system as a function of  $\lambda$ ;  $\langle \rangle$  denotes an ensemble average.

$$\Delta G_{A \rightarrow B} = G_B - G_A = \int_A^B \left\langle \frac{\partial H_\lambda}{\partial \lambda} \right\rangle_\lambda d\lambda$$

The advantage of these methods is that they explicitly consider the sampling of conformational ensembles of both protein and ligand, as well as the effect of solvent molecules upon binding. One of the main drawbacks of these methods is the high computational cost involved, as the two simulations required may be hard to converge (i.e., provide a realistic conformational ensemble). By the same token, the applicability is limited to closely related ligands, such as ones differing in one atom or pseudo-atom (e.g., a CH<sub>2</sub> group to O or NH groups). Despite this, these methods have been successfully applied to predict relative ligand binding affinities on different systems, such as streptavidin<sup>92</sup> and the estrogen receptor.<sup>93</sup>

#### 1.4.2 Linear interaction energy (LIE)

The LIE method relies on the scaling of the average van der Waals and electrostatic interaction energies on a conformational ensemble obtained from a pair of molecular dynamics simulations, one of the bound ligand and another of the free species.<sup>82</sup> Huang and Caflisch modified this method to include a continuum electrostatics term calculated by the solving the Poisson-Boltzmann equation, and used it on a single energy-minimized structure instead of sampling a set of multiple conformers.<sup>94</sup> Their results showed that, at least for BACE-1 and HIV-1 protease, the MD sampling was indeed not required for accurate predictions, but the parameters derived for one system were not transferable to the other.

**Equation 1.12.** Linear interaction energy method.  $\alpha$  and  $\beta$  are regression-based parameters modulating the conformational ensemble average van der Waals and electrostatic contributions.

$$\Delta E_{bind} = \alpha \cdot \langle \Delta E_{vdW} \rangle + \beta \cdot \langle \Delta E_{elec} \rangle$$

### 1.4.3 MM-PB/SA and MM-GB/SA

**Equation 1.13.** MM-PBSA and MM-GBSA methods.  $G$  is the free energy of binding,  $E_{MM}$  is the MM energy as calculated with a given forcefield,  $E_{solv}$  is the desolvation energy calculated by either GB/SA or PB/SA methods,  $TS$  is the entropic energy.  $\langle \rangle$  denotes an ensemble average.

$$\langle G \rangle = \langle E_{MM} \rangle + \langle E_{solv} \rangle - TS$$

These methods use explicit solvent MD simulations of the free ligand and receptor as well as the bound species to generate conformational ensembles.<sup>95</sup> Post-processing of snapshots from the simulations involves the removal of solvent molecules and computation of the potential energies on the three systems (complex, ligand and receptor) with a molecular mechanical force field, the solvation energies with a continuum model (either Poisson-Boltzmann or Generalized Born), and calculation of the solvent-accessible surface area. An estimate of the entropic contribution to binding can also be obtained from a quasi-harmonic analysis of the snapshots.<sup>55</sup> The requirement for three independent simulations poses a dual challenge of increased computational cost and difficulty in convergence,<sup>96</sup> hence an interesting modification involves the use of a single simulation on the bound system, which is then stripped of either ligand or biomolecule to calculate the averages on the free systems.

## 1.5 Quantum mechanics-based binding affinity calculations

Methods relying on first principles have seen much more limited application in drug development problems, mostly due to the sheer increase in computational cost associated with them.<sup>97</sup> One important application is in QM/MM simulations, where one part of the system is described with molecular mechanics and another quantum mechanically; these techniques have proven useful in the study of enzymatic mechanisms.<sup>98</sup> Additionally, a combination of docking, QM/MM and molecular dynamics was applied successfully to a set of MMP-9 inhibitors.<sup>99</sup> Recently, Caflisch and co-workers observed that considering QM interactions (using the RM1 semiempirical hamiltonian) improved the predictions on highly charged systems, likely accounting for polarization effects.<sup>100</sup> Likewise, recent results by different groups points to the benefits of using QM/MM methods linked to SBDD



techniques,<sup>101,102</sup> as a way of overcoming the lack of predictivity of the latter for lead optimization.<sup>103</sup>

## **1.6 Thesis objectives**

### **1.6.1 General objectives**

The central goal of this work was to develop methods to predict the binding affinity of potential drugs for application in virtual screening campaigns. Towards this goal, we first examined the current status of docking programs and scoring functions applied to a metalloprotein target relevant for medicinal chemistry, Golgi  $\alpha$ -mannosidase II (Chapter 2). This triggered a more in-depth analysis of scoring functions, which led us to assemble a set of protein complexes and analyze the performance of different available scoring functions in flexible and solvated proteins (Chapter 3). We then set out to develop a scoring function to apply in VS, based on MM force fields and additional parameters accounting for entropic costs and solvation (Chapter 4). We reached into the development of molecules binding to nucleic acids by developing a hybrid docking/molecular dynamics method to study transition metal complexes binding to G-quadruplexes (chapter 5). After implementation of the developed scoring functions, our docking program FITTED was set to be used in VS campaigns. However, additional tools were required to manipulate, select or prepare large libraries of ligands. With this goal in mind, we developed a module for FITTED, termed SMART, to prepare ligands prior to the docking (Chapter 6). The core of SMART was used to build two other modules, REACTOR and SELECT, that have applications in the preparation and clustering, respectively, of virtual libraries of ligands (also in Chapter 6).

### **1.6.2 Evaluation of docking programs for Golgi $\alpha$ -mannosidase inhibitors**

Chapter 2 describes a first approximation to the current state of docking techniques, where the performance of diverse docking programs was assessed in the context of metalloenzymes. It is well established that metal-containing receptors are a challenging case for docking, both for the prediction of the binding mode when

multiple coordination of the ligand is possible, as well as for the evaluation of the binding affinity including the metal ligation. Through the evaluation of seven available docking programs, including an early development version of FITTED, we found that Glide was the most successful one at pose prediction in this system. A simulated VS with this software resulted in the recovery of 80% of the seeded actives in the top 15% of a library of 1000 members.

### **1.6.3 Assessment of scoring functions for flexible docking**

Chapter 3 deepens the study of the performance of SFs by assessing the effect of protein flexibility and solvation on a panel of 18 scoring functions acting on a challenging training set of protein-ligand complexes. In some sets of protein-ligand complexes, ligand descriptors such as molecular weight or number of heavy atoms were better predictors of the binding affinity than the score calculated by a SF. We assembled a set of 209 protein-ligand complexes exhibiting little correlation between ligand MW and binding affinity, among other stringent selection criteria. The evaluation of the ability of the evaluated scoring functions to reproduce the experimentally observed ranking of complexes resulted in the eHiTS SF being the most predictive, followed by DrugScore and ChemScore. We also found that the consideration of explicit water molecules did not affect the scoring, and that the accuracy dropped significantly when considering docking to non-native receptor conformations.

### **1.6.4 Development of scoring functions for protein-ligand interactions**

Chapter 4 describes our efforts towards the development of force field-based scoring functions for implementation in a docking program for VS applications. While assessing the ability of different force fields to reproduce binding affinities, we found that the energies predicted by different force fields were highly correlated among each other. We considered a scoring function including MM terms, a GB/SA term, and a term accounting for the entropic penalty due to the loss of ligand torsional degrees of freedom upon binding. The different terms were tuned by an iterative approach, optimizing the correlation of the calculated scores with

experimentally determined binding affinities on a set of 209 complexes. Validation on an independent set of protein-ligand complexes confirmed the good accuracy of the SF. Additionally, we produced a SF derived from a VS-like approach, where the fitness to optimize was the recovery of known actives in a library containing decoys. This SF was validated against a separate library of ligands/decoys on different protein targets.

### **1.6.5 Modelling of platinum complexes as G-quadruplex binders**

Chapter 5 describes the development and application of a hybrid docking/MD technique for the evaluation of transition metal complexes as potential G-quadruplex binders and telomerase inhibitors. G-quadruplex structures are secondary structures of DNA observed in guanine-rich sequences, most notably telomeric DNA. Stabilization of this secondary structure of DNA can lead to inhibition of telomerase activity, which is sought as an anticancer therapy. In collaboration with researchers from the Sleiman and Autexier labs, we undertook the modelling of platinum (II) complexes for use as G-quadruplex binders. We developed MM force field parameters for the platinum (II) centers from *ab initio* data, which were then applied to the docking of these complexes onto two different foldings of the G-quadruplexes: parallel and anti-parallel. The resulting poses were used as starting points for MD simulations in explicit solvent, which were then processed to calculate the binding affinity of the compounds by the MM-PB/SA method. The calculated values were found to be in good agreement with experimental data from biophysical and biological sources.

### **1.6.6 Development of SMART, REACTOR and SELECT**

Chapter 6 describes the development of programs for the handling of ligands for a docking-based VS campaign. The first program is SMART, developed as a module of FITTED, which assigns the generalized Amber atom types used in FITTED for the scoring of the poses and marks rotatable bonds whose degrees of freedom are to be scanned. Additionally, a set of descriptors for toxic and reactive groups as well as various properties (partially based on Lipinski's *rule of five*) was implemented,

where the presence of these groups is used as a filter (i.e., skipping the ligand in the docking run) in order to reduce the time spent in a VS campaign. Furthermore, the ability to assign point charges was implemented in order to streamline the docking process. The framework of this program was used for the development of a pair of programs to facilitate the production of structures for VS. A first program, REACTOR, takes virtual libraries of reactants and combines them following user-defined rules to generate virtual libraries of compounds. The second one, SELECT, features a similarity search algorithm that allows for the selection of representative structures from a virtual library (filtering) and the extraction of a set of compounds similar to a query molecule from a virtual library (analog search).

## 1.7 References

1. Dictionary.com "Drug" definition.  
<http://dictionary.reference.com/browse/drug> (accessed May 05, 2009).
2. Nicolaou, K. C.; Sorensen, E. J., *Classics in total synthesis*. VCH: New York, 1996.
3. Nicolaou, K. C.; Snyder, S. A., *Classics in total synthesis II : more targets, strategies, methods*. Wiley-VCH: Weinheim, 2003.
4. Newman, D. J.; Cragg, G. M.; Snader, K. M., Natural products as sources of new drugs over the period 1981-2002. *J. Nat. Prod.* **2003**, 66, 1022-1037.
5. Grabowski, K.; Baringhaus, K. H.; Schneider, G., Scaffold diversity of natural products: Inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008**, 25, 892-904.
6. Fleming, A., The antibacterial action of cultures of a *Penicillium*, with special reference to their use in the isolation of *B. influenzae*. *Br. J. Exp. Pathol.* **1929**, 10, 226-36.
7. Hodgkin, D. C., X-ray analysis of the structure of penicillin. *Advancement Sci.* **1949**, 6, 85-9.
8. Park, J. T.; Strominger, J. L., Mode of action of penicillin: Biochemical basis for the mechanism of action of penicillin and for its selective toxicity. *Science* **1957**, 125, 99-101.

9. Rolinson, G. N.; Geddes, A. M., The 50th anniversary of the discovery of 6-aminopenicillanic acid (6-APA). *Int. J. Antimicrob. Agents* **2007**, *29*, 3-8.
10. Ehrlich, P., Aus Theorie und Praxis der Chemotherapie. *Folia Serologica* **1911**, *7*, 697-714.
11. Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Deutschen Chemischen Gesellschaft* **1894**, *27*, 2985-2993.
12. Gund, P.; Andose, J. D.; Rhodes, J. B.; Smith, G. M., Three-dimensional molecular modeling and drug design. *Science* **1980**, *208*, 1425-1431.
13. Cohen, S. S., Strategy for Chemotherapy of Infectious-Disease. *Science* **1977**, *197*, 431-432.
14. Ward, W. H.; Holdgate, G. A., Isothermal titration calorimetry in drug discovery. *Prog. Med. Chem.* **2001**, *38*, 309-376.
15. Schuck, P., Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules. In *Annu. Rev. Biophys. Biomol. Struct.*, 1997; Vol. 26, pp 541-566.
16. Pereira, D. A.; Williams, J. A., Origin and evolution of high throughput screening. *Br. J. Pharmacol.* **2007**, *152*, 53-61.
17. Houston, J. G.; Banks, M. N., High-Throughput Screening for Lead Discovery. In *Burger's Medicinal Chemistry and Drug Discovery*, John Wiley & Sons, Inc.: 2002.
18. Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. In *Hits, Leads and Artifacts from Virtual and High Throughput Screening*, Molecular Informatics: Confronting Complexity, Bozen, Italy, May 13-16, 2002; Hicks, M. G.; Kettner, C., Eds. Beilstein-Institut: Bozen, Italy, 2002.
19. Shoichet, B. K., Screening in a spirit haunted world. *Drug Discovery Today* **2006**, *11*, 607-615.
20. Jorgensen, W. L., The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813-1818.
21. Stahl, M.; Guba, W.; Kansy, M., Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discovery Today* **2006**, *11*, 326-333.

22. McLeish, M. J.; Kenyon, G. L., Approaches to the Rational Design of Enzyme Inhibitors. In *Burger's Medicinal Chemistry and Drug Discovery*, John Wiley & Sons, Inc.: 2003.
23. Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual screening - an overview. *Drug Discovery Today* **1998**, *3*, 160-178.
24. Lyne, P. D., Structure-based virtual screening: An overview. *Drug Discovery Today* **2002**, *7*, 1047-1055.
25. Muegge, I.; Enyedy, I., Virtual Screening. In *Burger's Medicinal Chemistry and Drug Discovery*, John Wiley & Sons, Inc.: 2003.
26. Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862-865.
27. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580-594.
28. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727-730.
29. Lipinski, C.; Hopkins, A., Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855-861.
30. Fink, T.; Reymond, J.-L., Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342-353.
31. Southall, N. T.; Dill, K. A.; Haymet, A. D. J., A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2002**, *106*, 521-533.
32. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623-1641.
33. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132-146.

34. Halgren, T. A., Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520-552.
35. Taylor, R.; Kennard, O., Hydrogen-bond geometry in organic crystals. *Acc. Chem. Res.* **1984**, *17*, 320-326.
36. Fersht, A. R.; Shi, J. P.; Knill-Jones, J., Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **1985**, *314*, 235-238.
37. Williams, D. H.; Searle, M. S.; Mackay, J. P.; Gerhard, U.; Maplestone, R. A., Toward an estimation of binding constants in aqueous solution: Studies of associations of vancomycin group antibiotics. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 1172-1178.
38. Levitt, M.; Perutz, M. F., Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.* **1988**, *201*, 751-754.
39. Hunter, C. A.; Sanders, J. K. M., The nature of  $\pi$ - $\pi$  interactions. *J. Am. Chem. Soc.* **1990**, *112*, 5525-5534.
40. Burley, S. K.; Petsko, G. A., Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science* **1985**, *229*, 23-28.
41. 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.
42. Brandl, M.; Weiss, M. S.; Jabs, A.; Sühnel, J.; Hilgenfeld, R., CH- $\pi$  interactions in proteins. *J. Mol. Biol.* **2001**, *307*, 357-377.
43. Ma, J. C.; Dougherty, D. A., The cation- $\pi$  interaction. *Chem. Rev.* **1997**, *97*, 1303-1324.
44. Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J., Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, Pullman, B., Ed. Elsevier: Dordrecht, Netherlands, 1981; pp 331-342.
45. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.

46. Mahoney, M. W.; Jorgensen, W. L., A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910-8922.
47. Hansch, C., Quantitative structure-activity relationships and the unnamed science. *Acc. Chem. Res.* **1993**, *26*, 147-153.
48. Lazaridis, T.; Karplus, M., Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139-145.
49. Southall, N. T.; Dill, K. A., The mechanism of hydrophobic solvation depends on solute radius. *J. Phys. Chem. B* **2000**, *104*, 1326-1331.
50. Hasel, W.; Hendrickson, T. F.; Still, W. C., A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Computer Methodology* **1988**, *1*, 103-116.
51. Mobley, D. L.; Dill, K. A., Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009**, *17*, 489-498.
52. Lumry, R.; Rajender, S., Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous property of water. *Biopolymers - Peptide Science Section* **1970**, *9*, 1125-1227.
53. Böhm, H. J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243.
54. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425-445.
55. Srinivasan, J.; Cheatham III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401-9409.
56. Sottriffer, C.; Klebe, G.; Stahl, M.; Böhm, H.-J., Docking and Scoring Functions/Virtual Screening. In *Burger's Medicinal Chemistry and Drug Discovery*, John Wiley & Sons, Inc: 2003.



57. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409-443.
58. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151-166.
59. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 15-26.
60. Kroemer, R. T., Structure-based drug design: Docking and scoring. *Curr. Protein Pept. Sci.* **2007**, *8*, 312-328.
61. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7-S26.
62. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269-288.
63. 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.
64. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
65. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K., Gaussian docking functions. *Biopolymers* **2003**, *68*, 76-90.
66. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11-26.
67. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337-356.
68. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore<sup>CSD</sup>-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296-6303.

69. Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R., VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959-3969.
70. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A., Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002**, *45*, 2469-2483.
71. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P., eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198-212.
72. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
73. Jain, A. N., Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427-440.
74. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T., Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317-324.
75. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750-1759.
76. Böhm, H.-J., Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309-323.
77. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M., LigScore: A novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395-407.
78. Wang, R.; Liu, L.; Lai, L.; Tang, Y., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **1998**, *4*, 379-394.

79. Schnecke, V.; Kuhn, L. A., Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discov. Des.* **2000**, *20*, 171-190.
80. Moitessier, N.; Therrien, E.; Hanessian, S., A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic  $\beta$ -secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885-5894.
81. Michel, J.; Verdonk, M. L.; Essex, J. W., Protein-ligand binding affinity predictions by implicit solvent simulations: A tool for lead optimization? *J. Med. Chem.* **2006**, *49*, 7427-7439.
82. Aqvist, J.; Medina, C.; Samuelsson, J. E., A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385-391.
83. Hansson, T.; Marelus, J.; Åqvist, J., Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27.
84. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.
85. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D., Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505-524.
86. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488-506.
87. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791-804.
88. Muegge, I., PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895-5902.
89. Alder, B. J.; Wainwright, T. E., Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **1959**, *31*, 459-466.
90. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.
91. Kollman, P., Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395-2417.

92. 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., Genet.* **1993**, *16*, 226-245.
93. Oostenbrink, C.; Van Gunsteren, W. F., Free energies of ligand binding for structurally diverse compounds. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6750-6754.
94. Huang, D.; Caflisch, A., Efficient evaluation of binding free energy using continuum electrostatics solvation. *J. Med. Chem.* **2004**, *47*, 5791-5797.
95. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.
96. Gilson, M. K.; Zhou, H. X., Calculation of protein-ligand binding affinities. In *Annu. Rev. Biophys. Biomol. Struct.*, 2007; Vol. 36, pp 21-42.
97. Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz Jr, K. M., The role of quantum mechanics in structure-based drug design. *Drug Discovery Today* **2007**, *12*, 725-731.
98. Lodola, A.; Woods, C. J.; Mulholland, A. J., Applications and Advances of QM/MM Methods in Computational Enzymology. In *Annual Reports in Computational Chemistry*, Wheeler, R. A.; Spellmeyer, D. C., Eds. Elsevier: 2008; Vol. Volume 4, pp 155-169.
99. Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S., A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* **2005**, *48*, 5437-5447.
100. Zhou, T.; Huang, D.; Caflisch, A., Is quantum mechanics necessary for predicting binding free energy? *J. Med. Chem.* **2008**, *51*, 4280-4288.
101. Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R., Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26*, 915-931.
102. Gleeson, M. P.; Gleeson, D., QM/MM calculations in drug discovery: A useful method for studying binding phenomena? *J. Chem. Inf. Model.* **2009**, *49*, 670-677.

103. Enyedy, I. J.; Egan, W. J., Can we use docking and scoring for hit-to-lead optimization? *J. Comput.-Aided Mol. Des.* **2008**, 1-8.

## **Chapter 2: Case study: crystallographic and docking studies of Golgi $\alpha$ -mannosidase II inhibitors**

### **2.1 Introduction**

#### **2.1.1 Protein glycosylation**

Glycoproteins and glycolipids are major components of the outer surface of mammalian cells and the majority of cell surface and secreted proteins of eukaryotes are glycosylated. The carbohydrates are commonly bound to an asparagine residue within the sequence Asn-X-Ser (or Thr) through an *N*-glycosidic linkage.<sup>1</sup> The glycosylation process corresponds to a post-translational modification involving a large panel of specific glycosidases and glycosyltransferases, and is responsible for proper processing of proteins. The biosynthesis of Asn-linked glycoproteins<sup>2-4</sup> starts in the endoplasmic reticulum then progresses in the Golgi apparatus to produce the mature glycosylated structure on the nascent protein.<sup>5,6</sup> This second step is highly dependent on species, tissues and cells, thus resulting in the diverse nature of the final branched oligosaccharides. The Golgi  $\alpha$ -mannosidase II (GMII) is responsible for the specific trimming of 2 mannose residues from the branched GlcNAcMan<sub>5</sub>GlcNAc<sub>2</sub> mannose intermediate, with retention of sugar anomeric configuration, and is therefore a key enzyme of the Golgi processing pathway.

#### **2.1.2 Golgi $\alpha$ -mannosidase and cancer**

In various tumor cells, the distribution of cell surface *N*-linked oligosaccharides is altered and correlates with disease progression, metastasis and poor prognosis.<sup>7-10</sup> GMII has consequently been viewed as a potential target in the development of new anti-cancer therapies. In clinical trials, swainsonine, a natural inhibitor of GMII featuring a 4-amino-4-deoxy-mannofuranoside unit<sup>11,12</sup> has been shown to reduce certain tumors and hematological dysfunctions.<sup>13,14</sup> However, the co-inhibition of lysosomal mannosidases prevents further development of this compound towards

medicinal treatments. It is thus highly important to find highly specific inhibitors of GMII that would exhibit anti-cancer activity.

In an ideal scenario, medicinal chemists use three-dimensional structures of the various protein targets to design potent and selective inhibitors. In fact, docking studies performed on various mannosidases would aid in the computational evaluation of the selectivity of the inhibitors. Unfortunately, mammalian mannosidases are difficult to purify in suitable quantities and to date only a single structure of bovine lysosomal mannosidase is available with a suboptimal resolution of 2.70 Å, which hinders the possibility of performing accurate docking experiments on it.

Jack bean  $\alpha$ -mannosidase was first found to be a readily available and reliable model enzyme for assaying inhibition of mammalian GMII, however its crystal structure and primary sequence has not yet been determined.<sup>15</sup> More recently, *Drosophila melanogaster* GMII (dGMII), which displays high sequence identity with human GMII (hGMII), 40% identity and 70% homology, was used as a valid model of the structural and functional features of the mammalian enzyme.<sup>16-18</sup> In particular, it has been shown that the exposed residues in the active site cavity are almost completely conserved between hGMII and dGMII.<sup>17</sup> As a consequence, the latter was used in place of hGMII in various crystallographic studies and has recently provided a series of crystal structures of GMII:inhibitor complexes.<sup>19</sup> This newly available structural data could now be the starting point for the structure-based design of potentially active and selective GMII inhibitors. For this purpose, we naturally turned our attention to the available computational structure-based drug design methods.

For the last two decades, computational methods for structure-based drug design have evolved significantly. Their increasing accuracy has been followed with growing interest by the pharmaceutical industry. Among these methods, docking techniques have been extensively investigated and exploited in medicinal chemistry projects. Unfortunately, no universal method (i.e., applicable to any protein target) has been discovered and the choice of the software has to be done wisely.<sup>20</sup> While

the accuracy of existing methods is increasing, remaining limitations have been identified. The flexibility of the enzyme and the presence of key water molecules are major issues yet to be addressed.<sup>21,22</sup> To account for the induced-fit of the protein upon binding of a ligand, several strategies have been proposed.<sup>23</sup>

### **2.1.3 Goals of this research**

We report herein our efforts in the structural determination of dGMII:inhibitor complexes and their use in docking studies. In particular, three new structures of dGMII:inhibitor complexes are presented. The present work had two main goals, the additional validation of our recently developed software FITTED and the identification of accurate software for designing and screening potential GMII inhibitors. Thus, a large section of this report will be devoted to a comparative study of the most accurate docking programs, namely Glide,<sup>24</sup> GOLD,<sup>25</sup> FlexX,<sup>26</sup> LigandFit,<sup>27</sup> eHiTS,<sup>28</sup> AutoDock<sup>29</sup> and FITTED<sup>30</sup> in combination with a large panel of scoring functions. This study was not intended to fully evaluate these docking programs but to find the best one in the context of mannosidase inhibition. A last section will describe the assessment of the accuracy of Glide in a virtual screening study.

## **2.2 Materials and methods**

### **2.2.1 Enzyme assays and crystallography**

Measurement of inhibition, crystallization, data collection and structural refinement were carried out essentially as outlined by Kuntz *et al.*<sup>31</sup> with the exceptions noted below. Crystals of dGMII were grown overnight, washed with phosphate buffered reservoir solution (as per Shah *et al.*<sup>32</sup>) and soaked with 10 mM **8** and **9** for at least 3 hours. In the case of **10**, crystals were soaked in Tris-buffered reservoir solution without phosphate washing. Data were collected on Beamline 191D at the Advanced Photon Source for crystals of **8** and **9** and at Beamline A1 at the Cornell High Energy Synchrotron Source for **10**. 400 frames with 0.5 degree oscillation/frame were collected. To obtain a data set with good completeness, data on 2 crystals of **9** were collected and the data merged with Scalepack.



### 2.2.2 Preparation of structures for docking

The structures of the dGMII complexes were retrieved from the Protein Data Bank (PDB codes: 1HWW, 1HXX, 1PS3, 1R33, 1R34, 1TQS, 1TQT, as well as the newly determined 2F18, 2F1A, 2F1B presented here) and prepared using Maestro 7.0<sup>33</sup> from Schrödinger as follows. Water molecules were removed and the resulting proteins were aligned based on the  $\alpha$ -carbon trace. Hydrogen atoms were added to both proteins and ligands, and bonds to the zinc atom were broken. The atom types and partial charges were first assigned automatically. Charges of the catalytic site residues were corrected following a DFT calculation at the B3LYP/6-31G\*\* level of theory (Jaguar 6.0<sup>34</sup>) on a truncated site consisting of His90, His471, Asp92, Asp204 and Zn; Mulliken populations were considered as the source of the charges. These charges were assigned at different stages depending on the docking software used. Appropriate zinc atom van der Waals parameters were obtained from the literature<sup>35</sup> and implemented in each program requiring these parameters. The structures of the ligands were optimized through energy minimization (Tripos force field) prior to the docking (MLS, minimized ligand structures); the original crystal structure (CLS, crystal-derived ligand structures) conformations were kept for RMSD measurements.

### 2.2.3 Glide

Neutral zones (as defined in Glide) of 10-20 Å around the ligands were first defined and the inhibitor/proteins were refined using the local optimization procedure proposed in Glide. The zinc parameters were added to the OPLS2003 force field definition and specific charges were assigned to the catalytic residues and zinc atom. The rest of the protein and the ligands were assigned Macromodel/OPLS2003 charges and atom types. Grids were prepared for each protein with the exact same center and a size of 40 Å. A constraint that forced the interaction with the metal ion was included. A specific keyword (CMAE) was employed for the DFT-derived partial charges (see above) to be maintained during the grid preparation. Preparation, refinement and grid calculation took about 4 hours per protein on an SGI R16K. The ligands were minimized using the OPLS2003 forcefield and submitted to Glide for

docking. In order to ensure convergence, an exhaustive search was secured by using search parameters set to their maximum values and a set of 25 runs. Using this exhaustive search led to the docking of the ten inhibitors on a single receptor in an average time of 44 minutes on an SGI R16K. The following parameters were used: ligvdwscale factor 1.0; maxkeep 50,000; maxconf 10,000; nreport 5,000; maxref 4,000; scorecut 100. A specific keyword (reference) was needed to report the RMSD.

#### **2.2.4 Glide VS**

Default parameters were used to dock with Glide; the time-consuming exhaustive search described above was discarded in order to better simulate a realistic virtual screening study. The protein target 1PS3 was used as prepared for the Glide docking described above. Ligands **11-17** were prepared for docking as described in Preparation of structures for docking.

#### **2.2.5 FlexX**

The input structures were prepared using the Sybyl 7.0 interface.<sup>36</sup> Binding sites based on proteins truncated at 7.0 Å around the inhibitors were used; ligand MLS structures were used as recommended in the FlexX User Manual. The RIGID\_RING mode was selected in order to keep the input conformation as the sole ring conformation (as for Glide and AutoDock). The torsion-standard.dat library of torsions was used. All the other default parameters for the various incremental construction stages or for the input file setup were used. An additional metal pharmacophore-filter using the FlexX-Pharm module was used. All the successful poses were kept for further analysis. Rescoring with the scoring functions implemented in CScore of both the non-relaxed (FlexX output) and relaxed (Tripos force field) docked poses was performed on a receptor featuring the free coordination sites of the Zn atom as dummy atoms (Tripos atom type “Du”).

#### **2.2.6 AutoDock**

The AutoDock Tool interface was used to prepare the ligands and the proteins. Kollman charges and solvation parameters were assigned to the protein. The

created pdbqs proteins files were modified to account for the presence of a metal (called M) and the calculated charges for the catalytic site residues and the zinc ion. Grids of identical size and center as for the Glide study were computed. Gasteiger-Marsili charges were assigned to the inhibitors whenever possible.<sup>37</sup> Otherwise, when non-parameterized groups (e.g., sulfur cation) were present, partial charges were computed using the MOPAC semi-empirical method (Mulliken charges). Both the CLS and MLS were used as input. 25 runs with a maximum of 1,000,000 energy evaluations and a population size of 100 individuals were performed. The same calculations were also done with a maximum of 5,000,000 energy evaluations and a population of 200 individuals but did not show any improvement. This last set of computations indicates a good convergence using a maximum of 1,000,000 energy evaluations and a population size of 100 individuals. The default parameters for the Solis and Wets optimization and the genetic operators were used.

#### **2.2.7 eHiTS**

No interface is provided with eHiTs. The protein input structures were given as pdb files and the ligands as mol2 files. The receptor was truncated keeping any residue with at least one atom within 7.0 Å from any of the inhibitors. The CLS and MLS were alternatively used. The docking was performed using the default parameters for the docking (fragment docking, graph matching algorithm and pose optimization) and scoring. The highest accuracy was selected.

#### **2.2.8 GOLD**

The protein and ligand (MLS) mol2 files prepared previously were used for this study. Most of the optimized parameters were set as defaults (population size of 100 individuals, 5 islands, niche size of 2 and a selection pressure of 1.1). However, in order to ensure an exhaustive search for each ligand, the following parameters were used: 200,000 as maximum operations allowed and a binding site defined using a radius of 18 Å. A substructure constraint (alcohol functional group within 2.5 Å from the zinc atom with a spring constant of 5.0 kcal/Å<sup>2</sup>) was also used in a set of runs. The docking terminated when the top three solutions were within 1.0 Å, otherwise

25 runs were carried out. ChemScore and GoldScore scoring functions were used alternatively. Although the metal coordination can be automatically determined, we overruled the automatic definition and set two possible metal coordination geometries (trigonal bipyramidal and octahedral).

### **2.2.9 LigandFit**

The protein and ligand (MLS) structures prepared previously were loaded on Cerius2.<sup>38</sup> Self docking with the zinc atom defined as a “feature” was first attempted with the three scoring functions (PLP, CFF, Dreiding); cross-docking was carried out only with PLP. Default parameters were used.

### **2.2.10 FITTED**

The protein and ligands were processed using PROCESS and SMART, two modules in FITTED.<sup>30</sup> Population sizes of 100 individuals were used and a maximum of 100 generations were carried out. All other default parameters were used as defined elsewhere.<sup>30</sup> In order to avoid a strong bias of the docking, a sphere as large as 6.0 Å centered on the zinc atom was used to orient the metal-binding moieties in the docking.

## ***2.3 Results and discussion***

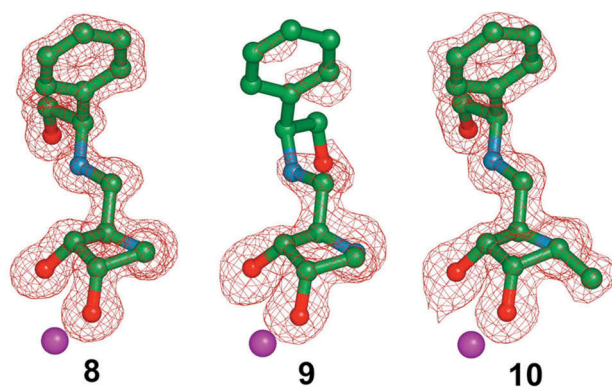
### **2.3.1 Docking data set**

Seven structures of dGMII:inhibitor complexes were selected from available crystal structures in the Protein Data Bank (PDB)<sup>39</sup> and added to the three described below. Structures with ligands that are not competitive inhibitors or structures with resolution worse than 2.0 Å were discarded. Ligands **1-7** (Table 2.1) were selected to represent a wide range of activity against dGMII. The seven dGMII:inhibitor complexes were imported from the Protein Data Bank: 1HWW (swainsonine),<sup>17</sup> 1HXX (1-deoxymannojirimycin),<sup>17</sup> 1PS3 (kifunensine),<sup>32</sup> 1R33,<sup>40</sup> 1R34,<sup>40</sup> 1TQS (salacinol),<sup>31</sup> and 1TQT.<sup>31</sup> We also included in this study three polyfunctionalized pyrrolidine derivatives (**8-10**) related to the family of  $\alpha$ -mannosidase inhibitors developed by Vogel and co-workers<sup>41</sup> which were co-crystallized in the active site of

dGMII (2F18, 2F1A, 2F1B). The latter inhibitors could be described as consisting of a pyrrolidine head coupled to a phenylglycinol tail.

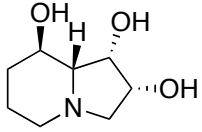
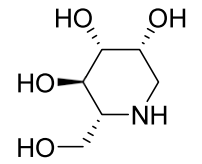
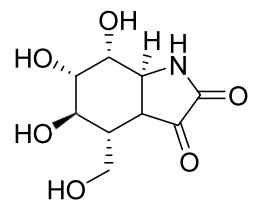
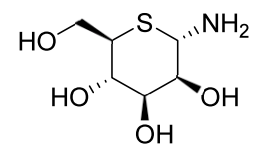
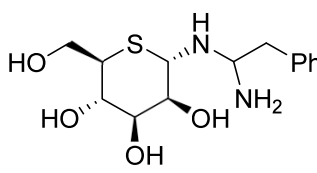
### 2.3.2 Crystallography

We report here the information retrieved from the analysis of the crystal structures of dGMII complexed with inhibitors **8**, **9** and **10**. Resolution of the synchrotron collected data was 1.30-1.45 Å and  $R_{\text{free}}$  was 18-19% for the three structures. Detailed data collection and refinement statistics are presented in Table 2.2. Figure 2.1 shows the quality of the electron density around the bound inhibitors. The density for inhibitors **8** and **10** was much cleaner in the “tail” region of the aromatic ring than the similar region of inhibitor **9**. For **9**, the electron density of the aromatic ring was only visible when the contour levels of the maps were lowered significantly. The average temperature factors for the aromatic groups of **8** and **10** were 16 and 17.6, respectively, while for **9**, it was 32.5. This indicated that this region of **9** was in an unfavorable location, and might be oscillating between numbers of positions so that it did not show up clearly in the electron density map. This lack of good density correlated well with the poorer inhibitory activity for **9** ( $IC_{50} = 720 \mu\text{M}$  vs.  $IC_{50} = 80 \mu\text{M}$  for **8**). The configuration of the phenylglycinol residue was thus a determinant of the recognition process.

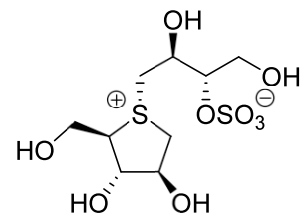


**Figure 2.1** Electron density representation of the inhibitors **8**, **9** and **10** bound in the active site of dGMII. Maps are simulated annealing omit maps ( $F_o - F_c$ ) of only the inhibitors contoured at 3.5  $\sigma$ . For orientation purposes the active site zinc ion is represented as a magenta ball. This figure was generated with PyMOL.

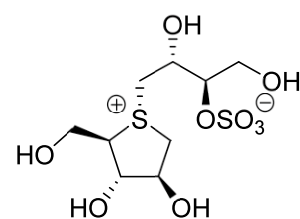
**Table 2.1** Selection of structures of  $\alpha$ -mannosidase/inhibitor complexes. The inhibitory activity ( $IC_{50}$ ) was measured on dGMII (EC 3.2.1.114) at 37°C, pH = 5.75. All protein crystal structures correspond to *Drosophila melanogaster* GMII.

Compound	$IC_{50}$ ( $\mu$ M)	PDB code	Resolution ( $\text{\AA}$ )	R/ $R_{\text{free}}$	Inhibitor Structure
<b>1</b> (Swainsonine)	0.017 <sup>16</sup>	1HWW <sup>17</sup>	1.87	0.18/0.21	
<b>2</b> (DMNJ)	400 <sup>19</sup>	1H XK <sup>17</sup>	1.50	0.20/0.22	
<b>3</b> (Kifunensine)	5200 <sup>a,35</sup>	1PS3 <sup>35</sup>	1.80	0.20/0.22	
<b>4</b>	70 <sup>36</sup>	1R33 <sup>36</sup>	1.80	0.16/0.19	
<b>5</b>	900 <sup>36</sup>	1R34 <sup>36</sup>	1.95	0.15/0.20	

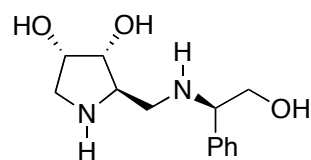
<b>6</b> (Salacinol)	7500 <sup>37</sup>	1TQS <sup>37</sup>	1.30	0.16/0.18
-------------------------	--------------------	--------------------	------	-----------



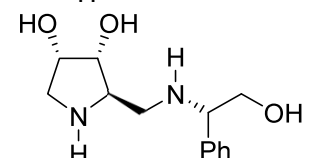
<b>7</b>	7500 <sup>37</sup>	1TQT <sup>37</sup>	1.90	0.15/0.18
----------	--------------------	--------------------	------	-----------



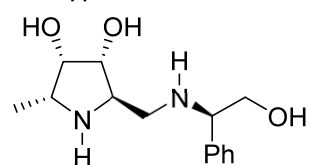
<b>8</b>	80 <sup>b</sup>	2F18 <sup>b</sup>	1.30	0.17/0.18
----------	-----------------	-------------------	------	-----------



<b>9</b>	720 <sup>b</sup>	2F1A <sup>b</sup>	1.45	0.17/0.19
----------	------------------	-------------------	------	-----------



<b>10</b>	1000 <sup>b</sup>	2F1B <sup>b</sup>	1.45	0.17/0.19
-----------	-------------------	-------------------	------	-----------



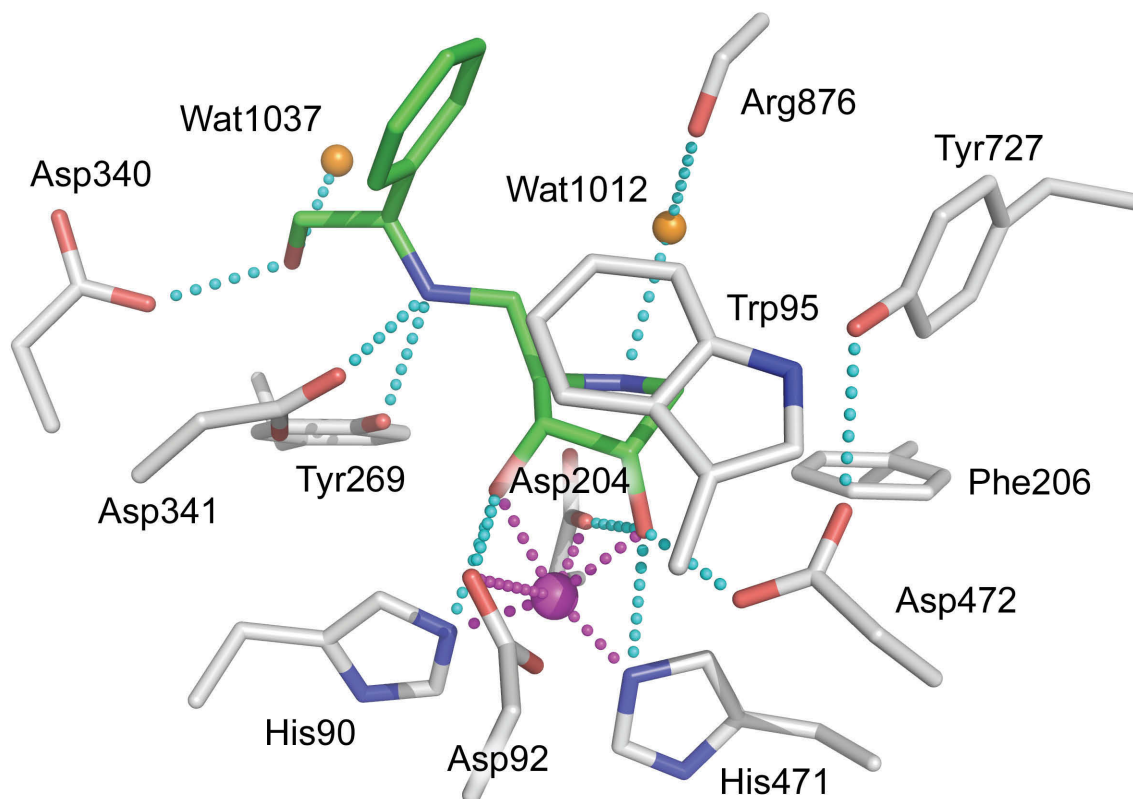
<sup>a</sup> K<sub>i</sub> value; <sup>b</sup> This work.

**Table 2.2** Data collection and structural refinement statistics

Compound	8	9	10
PDB code	<b>2F18</b>	<b>2F1A</b>	<b>2F1B</b>
HET symbol	GB1	GB2	GB3
<b>Data Collection</b>			
X-ray Source	APS	APS	CHESS
Cell dimensions (Å)	68.97 X 109.7 X 138.9	68.90 X 109.4 X 138.6	69.43 X 110.6 X 139.9
<b>Data Processing (Denzo/Scalepack)</b>			
Resolution (Å)	30-1.30/	30-1.45/	30-1.45
(overall/hi_res)	1.35-1.30	1.50-1.45	1.48-1.45
Redundancy			
(overall/hi_res)	10.8/5	12.5/10	3.9/3
I/sigma			
(overall/hi_res)	39/5.1	18.5/3.7	15/2.8
% Completeness			
(overall/hi_res)	99.9/99.6	96.3/88.0	95.8/97.7
R merge			
(overall/hi_res)	0.066/0.35	0.06/0.57	0.066/0.50
<b>Refinement (CNS)</b>			
R <sub>test</sub> /R <sub>free</sub>	0.168/0.180	0.168/0.189	0.173/0.194
Amino Acids	1014	1015	1014
Alternate			
Conformations	33	29	35
Water Molecules	1172	1099	1158
rmsd bonds (Å)	0.019	0.016	0.023
rmsd angles (°)	2.2	1.8	1.9
<b>Average B-Factors (Å<sup>2</sup>)</b>			
Overall	15.3	20.9	17.6
Protein Main Chain	12.5	18.0	14.7
Protein Side Chain	14.9	20.9	17.2
Water	26.0	31.3	28.6
Inhibitor	12.7	25.5	15.0
(MPD,NAG,PO4,Zn)	30.5	38.0	33.0



The binding of the highly active inhibitor **8** in the active site of dGMII is illustrated in Figure 2.2 and a list of interactions between the three inhibitors and the protein where the interaction distance was less than 3.2 Å is given in Table 2.3. For comparison, the interaction distances with swainsonine **1** (IC<sub>50</sub>=17 nM) are also indicated. These distances were derived from a high resolution (1.30 Å) dGMII:**1** co-crystal structure.<sup>17</sup> The major changes between the three pyrrolidine-based inhibitors occurred in the interaction with the terminal hydroxyl group (OH-9). The importance of the two hydroxyl groups on the pyrrolidine ring was clear. There were tight interactions between the hydroxyl moieties and the active site zinc as well as interactions with His90, Asp92, Asp204, His471 and Asp472. An unusual feature of the inhibitors presented in this article was that there was no hydroxyl group occupying the space between Asp472 OD1 and Tyr727 OH (Figure 2.2). In all the structures that we have previously examined there has been a hydroxyl group on the bound compound sitting between Asp472 OD1 and Tyr727 OH. In the case of phosphate-washed crystals where no compound was bound in the active site, the position was occupied by a water molecule.<sup>42</sup> As a result of this absence of hydroxyl group, the Tyr727 OH was shifted about 0.6 Å towards the Asp472 OD1. Other interactions with the inhibitors were also seen, in particular with N7 in the “tail” region. Trp95 made two important interactions with the inhibitors. There was a T-shaped interaction between the aromatic ring in the inhibitor and the indole in the tryptophan side-chain, the two planes being at near right angles of each other. Trp95 also made stacking hydrophobic interactions with the pyrrolidine ring, a common feature of GMII complexes.<sup>17</sup>



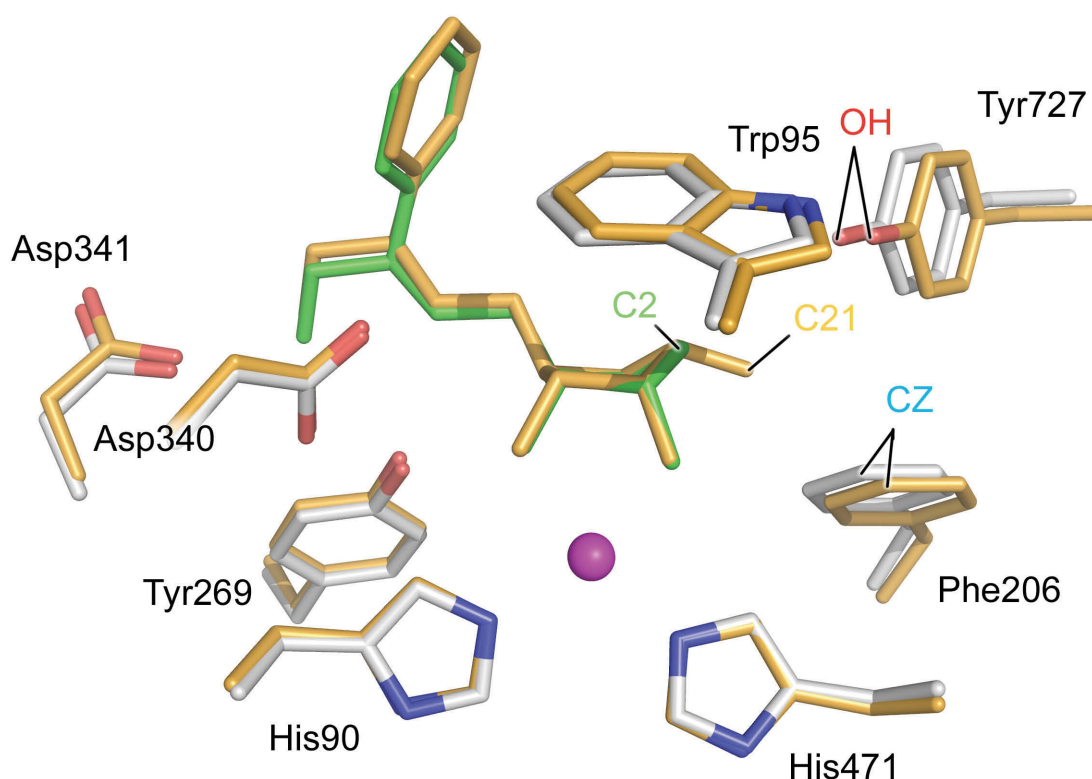
**Figure 2.2** a) Interaction of **8** with residues in the active site of dGMII. Interactions closer than 3.2 Å are indicated with cyan dotted lines; interactions with the zinc ion are indicated in magenta. Water molecules appear as orange balls. Distances are presented in **Table 2.3**. This figure was generated with PyMOL.

In the dGMII:**8** complex, the terminal oxygen (O9) made numerous hydrogen bonds. It was 2.7 Å from Asp340 OD1, 2.8 Å from a water molecule, 3.3 Å from Tyr269 OH and 3.5 Å from Asp341 OD1. An almost identical bonding pattern was seen for inhibitor **10**, although the interaction distances were slightly longer. The oxygen (O9) was 2.7 Å from Asp340 OD1, 2.8 Å from a water molecule, 3.5 Å from Tyr269 OH and 3.7 Å from Asp341 OD1. Because of the different stereochemistry of the inhibitor **9**, the O9 oxygen sat in a different location and was now 4.8 Å from Asp340 OD1 and 4.5 Å from Asp341 OD1. Nevertheless, there were still interactions with Tyr269 and two water molecules (Table 2.3).

**Table 2.3** Summary of interatomic distances (Å) between the inhibitors and dGMII. Distances in bold represent distances greater than 3.5 Å, where no significant hydrogen bonding is expected to occur.

Zinc Interactions									
Compound		1		8		9		10	
PDB		1HWW		2F18		2F1A		2F1B	
Protein or Inhibitor		Distance		Distance		Distance		Distance	
Atom		(Å)		(Å)		(Å)		(Å)	
H90 NE2		2.10		2.12		2.11		2.14	
D92 OD1		2.24		2.13		2.16		2.17	
D204 OD1		2.17		2.09		2.08		2.11	
H471 NE2		2.09		2.11		2.09		2.13	
OH-1 (1) / OH-3 (8-10)		2.20		2.18		2.20		2.20	
OH-2 or OH-4		2.13		2.26		2.35		2.31	
Protein/ligand Interactions									
Compound		1		8		9		10	
PDB		1HWW		2F18		2F1A		2F1B	
Protein	Atom	Distance	Atom	Distance	Atom	Distance	Atom	Distance	
Atom		(Å)		(Å)		(Å)		(Å)	
D92 OD1	OH-1	3.03	OH-3	2.99	OH-3	2.98	OH-3	3.03	
	OH-2	2.92	OH-4	3.11	OH-4	3.18	OH-2	3.12	
D92 OD2	OH-2	2.54	OH-4	2.69	OH-4	2.61	OH-2	2.59	
D204 OD1	OH-1	2.83	OH-3	2.76	OH-3	2.81	OH-3	2.77	
	OH-2	2.97	OH-4	2.93	OH-4	2.96	OH-4	2.98	
	N-4	2.75	N-1	2.81	N-1	2.88	N-1	2.81	
D204 OD2	N-4	3.45	N-1	3.34	N-1	3.36	N-1	3.38	
Y269 OH			N-7	3.00	N-7	2.88	N-7	3.14	
			OH-9	3.30	OH-9	2.97	OH-9	<b>3.52</b>	
			OH -9	2.67	OH -9	<b>4.82</b>	OH -9	2.66	
D340 OD1			OH-9	3.49	OH-9	<b>4.48</b>	OH-9	<b>3.70</b>	
D341 OD1			OH-9	3.49	OH-9	<b>4.48</b>	OH-9	<b>3.70</b>	
	D341 OD2	OH-8	2.56	N-7	2.78	N-7	2.85	N-7	2.80
OH-4				3.42	OH-4	3.49	OH-4	3.45	
D472 OD1			-	-	-	-	-	-	
D472 OD2	OH-1	2.60	OH-3	2.49	OH-3	2.46	OH-3	2.59	
Y727 OH	OH-8	2.64	-	-	-	-	-	-	
WATERS			OH-9	2.78	OH-9	2.63	OH-9	2.78	
			N-1	2.91	N-1	2.94	N	2.96	

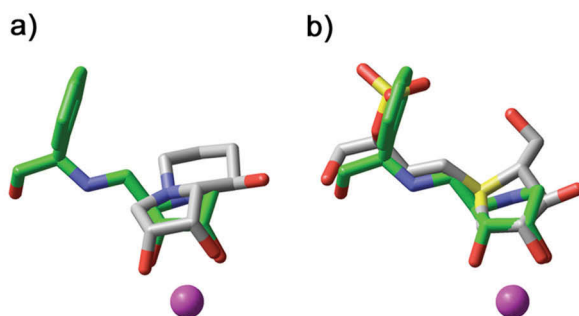
Inhibitors **8** and **10** exhibited virtually identical binding modes in the active site of dGMII (Figure 2.3), despite the additional methyl group on the pyrrolidine ring in **10** (labeled C21 in the PDB file). In the protein, however, the active site region was opened up in the structure of **10** compared to the structure of **8**, Phe206 and Tyr727 being pushed away from the inhibitor. The distance between C2 in the inhibitors and CZ of Phe206 was 4.8 Å in the **10**:dGMII complex, while it was 4.3 Å in the **8**:dGMII adduct. The Tyr727 OH was shifted 0.4 Å away from C21. The additional methyl group in **10** was in a position normally occupied by a polar hydroxyl in other inhibitors; the conformational stress put on the enzyme to accommodate its binding might be a reason for the poorer affinity of **10** relative to **8** ( $IC_{50} = 1000 \mu M$  vs.  $IC_{50} = 80 \mu M$ ).



**Figure 2.3** Comparison of the conformation of the active site residues in the complexes of **8** (ligand green, protein grey) and **10** (ligand and protein orange) with dGMII. The zinc atoms appear as magenta balls. This figure was generated with PyMOL.

The position of the aromatic group was similar to that occupied by the ends of the chain of the salacinol analog complexes.<sup>31</sup> An overlay of the crystal structure binding

mode of the diastereomer of salacinol (**7**) with **8** is shown in Figure 2.4b. Both of these inhibitors exhibited very clear density in their "tail" regions. This suggested that these parts of the inhibitors were occupying a favorable area of the active site space. Nevertheless, a major difference was that the zinc ion was bound to two hydroxyl groups of the inhibitor **8**, but only to one hydroxyl group of the diastereoisomer of salacinol (**7**). This might account for the poor inhibitory properties of **7** in comparison with **8** ( $IC_{50} = 7.5 \text{ mM}$  vs.  $IC_{50} = 80 \text{ }\mu\text{M}$ ).



**Figure 2.4** Overlays of the binding of **8** with (a) swainsonine **1** (PDB code 1HWW) and (b) the diastereomer of salacinol **7** (PDB code 1TQT). **8** is drawn in green, **7** and **1** are drawn in grey. The active site zinc is represented as a magenta ball.

The binding modes of **8** and swainsonine (**1**) are compared in Figure 2.4b. It was obvious that although the interatomic distances were almost identical (Table 2.3), there was a slight shift in the binding position of **8**. This shift along with the presence of the long and flexible aromatic tail (leading to entropy loss upon binding), the additional ammonium group (high desolvation cost upon binding) as well as the lack of a third hydroxyl interacting with the protein, might account for its weaker inhibitory potency ( $IC_{50} = 80 \text{ }\mu\text{M}$  for **8** vs.  $IC_{50} = 17 \text{ nM}$  for **1**).

### 2.3.3 Docking of GMII inhibitors: general considerations

A close look at the crystal structures revealed challenges for the accurate docking of GMII inhibitors. First, a zinc atom was involved in the catalytic activity of this enzyme and in the strong binding of the selected inhibitors. Second, a few of the active inhibitors featured solvent-exposed moieties that were not specifically interacting with any of the enzyme residues as exemplified by the electron density

of compound **9** in Figure 2.1. Third, there were bound water molecules that made important interactions with the inhibitors (in the case of **8**, at the amino group in the pyrrolidine ring and the O9 oxygen in the tail region). It was therefore expected that docking accuracy would be linked to a proper handling of the zinc ligation and solvation/desolvation by both the conformational sampling engine and the scoring function. Water molecules were present in some of the complexes; however their locations were not conserved among the different complexes. As it would have been impossible to keep them without biasing the self-docking of an inhibitor to its natural solvated receptor, they were not considered for the present docking study. Ideally, when performing virtual screening or *de novo* design of enzyme inhibitors, water molecules should be properly located or displaced by the docking program, a feature not implemented or in development in most of the available software. To date, only FITTED can move and displace water molecules<sup>43</sup> while GOLD can toggle their presence on or off.<sup>25</sup>

After addition of the hydrogen atoms to the crystal structures, the ligands were removed and the proteins were charged and prepared for their use in the subsequent docking study. Special attention was given to the catalytic site including two histidines, two aspartates and a zinc ion. Although the formal charge of the zinc atom is +2, it was clear that this charge was delocalized onto the chelating residues.<sup>44</sup> The charges for the catalytic site residues and for the zinc atom were derived from density functional theory (DFT) calculations of a truncated binding site and used for AutoDock, Glide, LigandFit and FITTED. The van der Waals parameters for the zinc atom required for both the protein preparation and the docking study were obtained from the literature.<sup>44</sup> For the following studies, the crystal-derived ligand structure (CLS) and/or optimized energy-minimized ligand structure (MLS) were used as input. A comparative study aiming at identifying a program for future virtual screening of drug design could not provide useful results if CLS's were required as input. In a real drug design scenario, one does not know the final pose and would guess an input structure as MLS. As we will describe in the following

sections, selection of the input structures has an impact on the accuracy of some of the docking programs.

#### **2.3.4 Docking methods used in this comparative study and parameterization**

Many comparative studies discussed in a recent review have shown Glide and GOLD to be amongst the most accurate docking programs.<sup>20</sup> For instance, Rognan and co-workers<sup>45</sup> ranked GOLD, Glide and Surflex as the most accurate docking programs for their set, followed by FlexX. We have previously obtained good results for the docking of zinc-containing enzyme inhibitors with AutoDock.<sup>46,47</sup> In addition, a recent review of eHiTS indicates that this docking program is a new candidate of interest.<sup>48</sup> We decided to assess GOLD, Glide, FlexX, AutoDock, LigandFit and eHiTS to see which one, if any, would provide accurate docking results for GMII inhibitors. In order to add to the validation of our own software, the current version of FITTED<sup>30</sup> was also assessed. It is also worth noting that Glide, FlexX (FlexE), AutoDock and FITTED have versions where flexibility of the protein is accounted for. Induced-fit docking using Glide/Prime relies on a combination of rigid protein docking and homology modeling techniques to construct the backbone and residue side chains.<sup>49</sup> FlexE relies on docking to composite structures,<sup>50</sup> while AutoDock grids can be combined using appropriate weighting schemes into virtual conformational ensembles.<sup>51</sup> The flexibility of the protein / ligand complexes in FITTED relies on the use of chromosomes to describe the whole complexes.<sup>30,52</sup> A section on the use of these functionalities is included in this manuscript.

The assessed programs covered a variety of conformational search methods: genetic algorithms (GOLD, AutoDock and FITTED), incremental construction (FlexX), rigid fragment docking and linking (eHiTS), Monte Carlo/matching algorithm (LigandFit) and multi-level search (Glide) and a large panel of scoring functions (e.g, ChemScore, GlideScore, GoldScore, F-Score, AutoDock scoring function, PMF, RankScore). Some of the evaluated methods came with an interface that was used to prepare the protein and ligand structure and initial keyword files. However, in order to obtain the best performance from each docking program, the standalone versions with

optimized parameters were used. In addition, although the computation of the atomic root mean square deviations (RMSD) was part of the output of Glide, FlexX and FITTED, we made our own scripts to compute the RMSD's for the AutoDock and eHiTS studies. The deviations were evaluated using the CLS's as references, taking into account only the coordinates of the heavy atoms but accounting for equivalent atoms that can be exchanged by rotation. One of the main issues addressed is the evaluation of the metal coordination as all the assessed programs treat the metal / ligand interaction differently. Therefore the deviations of the docked ligand structures from the CLS's were also computed for the metal ligation (only for the one or two oxygen atoms bound to the metal center).

As all these programs have been reported, we will describe them succinctly, emphasizing the way they treat binding to metal ions and solvation and their application to GMIL.

**Glide 3.5.**<sup>24,53</sup> Glide uses a funnel-type of approach to search the conformational space and the best poses are scored using GlideScore,<sup>54</sup> a scoring function derived from ChemScore.<sup>55</sup> Among the many terms of this scoring function is a term accounting for metal-ligand interactions. However, this term is restricted to anionic ligands, single atom ligation (we will consider diols) and does not account for the specific geometry of metal-ligand complexes. In the present study, the ligands were neutral, and the metal ligation would therefore be modeled as purely electrostatic. GlideScore also includes a unique solvation term which accounts for solvation of solvent-exposed moieties as well as water molecules captured in hydrophobic protein pockets. Glide also proposes the use of constraints to force specific interactions. The user-defined constraints add an additional filter to the hierarchical filtering. In order to evaluate the impact of constraints on accuracy, two studies - with and without constraints- were carried out. Unexpectedly, filtering off the poses where no groups were in close proximity to the zinc ion did not significantly increase the accuracy of the binding mode prediction.



**FlexX 1.13**<sup>26,58,56</sup>/**CScore**<sup>36</sup>. FlexX is conceptually very different to Glide. Rather than using grids, FlexX models the protein with an all-atom representation of the binding site, and while Glide uses an exhaustive conformational search, FlexX builds up the ligand within the binding site by incremental construction.<sup>57</sup> One similarity between Glide and FlexX is the possibility of using pharmacophore-like constraints to force specific interactions between the ligands and the receptor.<sup>58</sup> Poses that do not satisfy the pharmacophoric constraints are removed. In the present work, we imposed the requirement that at least one zinc-binding atom of the ligand should interact with the zinc ion. The incremental construction algorithm accounts for the geometry of ligand-metal interactions but not for metal coordination geometry.

The poses proposed by FlexX were submitted to rescoring by a panel of four scoring functions implemented in CScore, namely PMF, GoldScore, DockScore and ChemScore. ChemScore includes a specific (but non-directional) term for metal/ligand binding while there is no specific term for metal/ligand interaction in the FlexX, PMF, GoldScore, and DockScore scoring functions.

**AutoDock 3.0.**<sup>29,59</sup> The inhibitors, modeled using a united atom representation, are flexibly docked into the grids modeling the proteins by means of a Lamarckian genetic algorithm (LGA). The LGA optimizes the ligand pose to find the local minimum by perturbing the genes of the ligand. The scoring function available with the version 3.0 of this suite of programs does not include any specific term for metal-ligand interaction, the metals being treated as charged spheres with no specific coordination geometry. In contrast to other methods used herein, in the current version (version 4.0 is under validation<sup>60</sup>), no constraint can be specified.

**GOLD 3.0.**<sup>25,61</sup> GOLD also samples the ligand conformational space using a genetic algorithm (GA). However, in contrast to AutoDock, it uses the atomic description of the protein (or a truncated binding site) and allows for the hydroxyl (Ser, Thr, Tyr) and ammonium (Lys) hydrogen atoms to relax upon ligand docking. A large number of parameters can be optimized to improve the accuracy of the conformational search, although only a few were optimized in the present work according to the

observed poses of initial runs (e.g., torsion angle distribution databases, constraints to direct the docking towards the observed binding modes, parameters for the GA, and user-defined scoring functions). Upon docking, GOLD makes use of virtual coordination points (which can be specified by the user) to model the chelation of metals, a potentially very important feature for GMI inhibitor docking. In this study, two possible metal coordination geometries were assessed (trigonal bipyramidal and octahedral). The GOLD scoring function (GoldScore) as well as ChemScore do not consider the metal coordination geometry, the metal-ligand interaction strength being only distance dependent. As for Glide and FlexX, GOLD allows for the use of constraints (harmonic constraints) to direct and speed up the docking process.

**eHiTS.**<sup>28,62</sup> A recent software review revealed the high accuracy of eHiTS for docking small molecules to rigid proteins.<sup>48</sup> This review together with a comparative study from the developers<sup>63</sup> prompted us to evaluate eHiTS in the prospect of accurately docking  $\alpha$ -mannosidase inhibitors. The conformational search sampling is performed by rigid fragment docking followed by linkage and optimization of the reconstructed ligands. This approach, claimed to be “truly exhaustive”, is quite different from the methods used by the other programs described herein. The original scoring function is an empirical scoring function with many terms, including a specific term for angle-dependent metal coordination. However, although angles around the ligand atoms are considered, the metal coordination geometry is not optimized. This piece of software is still under development and only a very few parameters are accessible for modification by the user. While this manuscript was in preparation, a new and more accurate scoring function was implemented but has not been assessed in our study.<sup>64</sup>

**LigandFit.**<sup>27,38</sup> LigandFit first defines the binding site as a series of grid points used to locate the binding site, defines its shape and further docks the flexible ligands. The docking is performed by generating sets of ligand conformations using a Monte Carlo technique and matching these ligand conformations to the binding site partitions. Three scoring options are available, namely CFF and Dreiding force fields, as well as the PLP scoring function. Neither these two force fields nor PLP include a

specific treatment of metal ligation. LigandFit also allows for additional constraints using filters. The LigandFit constraint matches a feature atom such as a polar hydrogen with a complementary ligand atom such as a hydrogen bond acceptor oxygen.

**FITTED 1.0.**<sup>30</sup> To complete the comparative study, FITTED, a program recently developed in our laboratory has been assessed.<sup>30</sup> FITTED exploits a LGA to model the flexibility of both the ligands and proteins. It also includes a specific function for displaceable water molecules. Unlike AutoDock, FITTED optimizes the ligand pose through a Fletcher-Reeves conjugate gradient minimization.<sup>65</sup> The scoring function (RankScore) does not include any specific term for metal/ligand interactions which are treated as hydrogen bonds.<sup>52</sup> Constraints are implemented in FITTED to direct the binding to a specific atom or group of atoms and will be used to select poses with metal chelation.

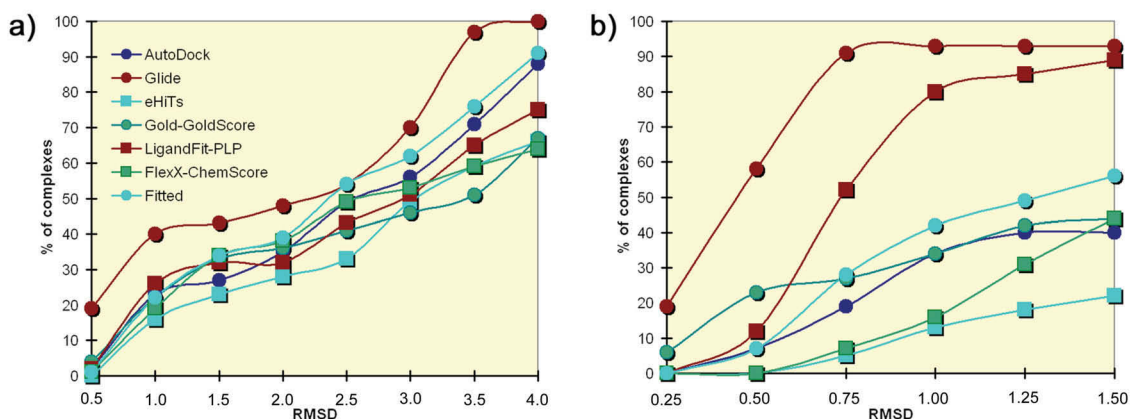
### 2.3.5 Application to the docking of mannosidase inhibitors

The ten selected weak to strong ligands were docked to the ten protein structures using each of the assessed programs for a total of 100 docking runs (10 self-docking runs and 90 cross-docking runs) for each program. Thus, Glide/GlideScore, FlexX/FlexXScore, FlexX/PMF, FlexX/GoldScore, FlexX/ChemScore, FlexX/DockScore, AutoDock/AutoDock scoring function, eHiTs/eHiTs scoring function, GOLD/GoldScore, GOLD/ChemScore, LigandFit/PLP, LigandFit/CFF, LigandFit/Dreiding and FITTED/RankScore were successively assessed. Figure 2.5 summarizes the collected data.

The FlexX docking engine allowed us to evaluate the generated poses. Thus, when we considered the use of FlexX to dock inhibitors **1-10** into GMII, all the docked poses were kept for post-docking analysis and rescoring with 4 other scoring functions (DockScore,<sup>66</sup> PMF,<sup>67</sup> GoldScore,<sup>25</sup> ChemScore<sup>55</sup>) included in the CScore module of Sybyl. Only the results with the ChemScore scoring function are shown as the other four functions were less accurate. In the present study, a pose with an RMSD below 2.0 Å was generated in only 75% on the cases and was identified by the

scoring function in less than 40% of the cases. This revealed that a significant portion of the failures was due to poor conformational sampling and not only to inaccurate scoring.

Within GOLD, GoldScore was found to be marginally better than ChemScore. Interestingly, the use of ChemScore with GOLD and FlexX led to similar results while the enhanced version of ChemScore, GlideScore, led to significantly better accuracy when used in conjunction with Glide. LigandFit was found to be inaccurate when CFF or Dreiding force fields were used to score the generated poses. Unexpectedly, CLS provided significantly better results than MLS with eHiTs. When a simple rotation in space or a change in one torsion angle was applied to the CLS used as input structures, the accuracy dropped.



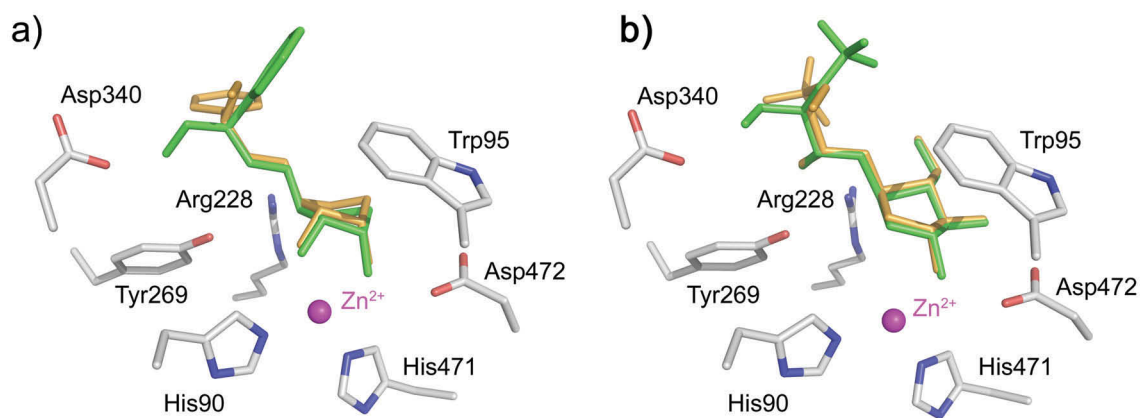
**Figure 2.5** Rigid protein docking. Accuracy for the 7 programs assessed to dock the 10 ligands to the 10 protein structures (self-docking and cross-docking) measured as the RMSD between the docked poses and the crystal structure binding mode for (a) all ligand heavy atoms; (b) metal-binding atoms.

Glide clearly appeared as the most accurate program in this comparative study. The ligands were docked with RMSD's below 1.0 Å in 40% of the cases and below 2.0 Å in 48%. eHiTs was much less accurate with an accuracy of 28% at 2.0 Å RMSD, while the other five programs showed accuracies ranging from 32% (LigandFit) to nearly 40% (Fitted). We were also pleased to see that the accuracy of Fitted was equivalent to that of Glide at RMSD = 2.5 Å.

In general, the computed RMSD's indicated better accuracy for the self-docking of the smaller ligands **1**, **2**, **3**, **4**, and **5** with all the programs. For instance, **1-5** were docked back to their protein structure (self-docking) with RMSD's below 1.1 Å when the MLS's were used as input with the Glide program. These data are consistent with the results reported in the original Glide publication from Friesner and co-workers where nearly 50% of the compounds from the testing set were docked back with RMSD's below 1.0 Å.<sup>24</sup> In that same study, about a third of the compounds were docked with RMSD between 1.0-2.0 Å, although in our case ligands **6** to **10** were self-docked with RMSD's higher than 2.5 Å.

To further evaluate the apparent poor docking accuracy observed for ligands **6-10**, we inspected the docked and experimentally observed conformations. First, we noticed that the experimentally observed zinc chelation by the diol or alcohol moieties was correctly predicted in almost all of the one hundred complexes when Glide was used, but with poorer accuracy when the other programs were used (Figure 2.5b). Second, a close look at the docked structures indicated that the docked pose was for the most part correct in many cases, the main deviation coming from the solvent-exposed moieties and the orientation of the aromatic groups. For instance, compound **8**'s phenyl group was predicted to interact with Arg228 through  $\pi$  interactions, while it was involved in a T-shaped  $\pi$  interaction with Trp95 in the crystal structure (see Figure 2.6). In fact, none of the programs predicted the T-shaped interaction between the phenyl ring of **8** and Trp95. This was probably due to the poor description of this type of interaction by the commonly used scoring functions. Nevertheless, the pyrrolidine ring was perfectly oriented (Figure 2.6a) in all cases. Similar conclusions were drawn for compounds **9** and **10**. As discussed above, the complex with **9** showed large B-factors for the phenyl ring and a poorly defined electron density map for this solvent-exposed group. The chelation by a single alcohol of **6** and **7** was predicted to occur by Glide (Figure 2.6b), but the main deviation arose from the positioning of the solvent-exposed sulfate predicted to interact with Tyr267 and/or with Arg228 (Figure 2.6b). In the crystal structures, this moiety was solvated while not specifically interacting with a protein residue.

Owing to the solvation/desolvation term of GlideScore, this situation was predicted in a few cross-docking situations (e.g., ligand **7** docked into 1HXK protein structure).



**Figure 2.6** Glide docked vs. crystallographically observed binding mode of compound **8** (a) and **7** (b). The crystal structure appears in green, the docked structure in orange. This figure was generated with PyMOL.

When using minimized ligands as input, the ring conformation in the input structure was already different from the crystal structure. This structure optimization stage contributed to a small fraction of the RMSD ( $< 0.3$  Å). Compounds in the crystal structures of dGMII were not in their lowest energy conformation but rather in a higher energy conformation induced by environmental constraints in the active site.<sup>32</sup> The effect of the zinc ion and the active site environment on inhibitor distortion has recently been discussed.<sup>68</sup>

### 2.3.6 Metal ligation

In order to further evaluate the program's ability to dock the GMII inhibitors, a closer inspection of the poses around the Zn cation was carried out. A clear indication of the predictive power of Glide was given by the computed RMSD's of the metal ligating groups (Figure 2.5b). In 54% and 91% of the cases (cross- and self-docking respectively), the chelating alcohols or diols were positioned within 0.5 Å and 1.0 Å of the observed positions respectively, even though this docking/scoring method did not account for coordination geometry. AutoDock suffered from the lack of constraints and in a few cases, **8** did not even interact with the zinc cation. The other programs offered the use of constraints to direct the docking and the zinc was

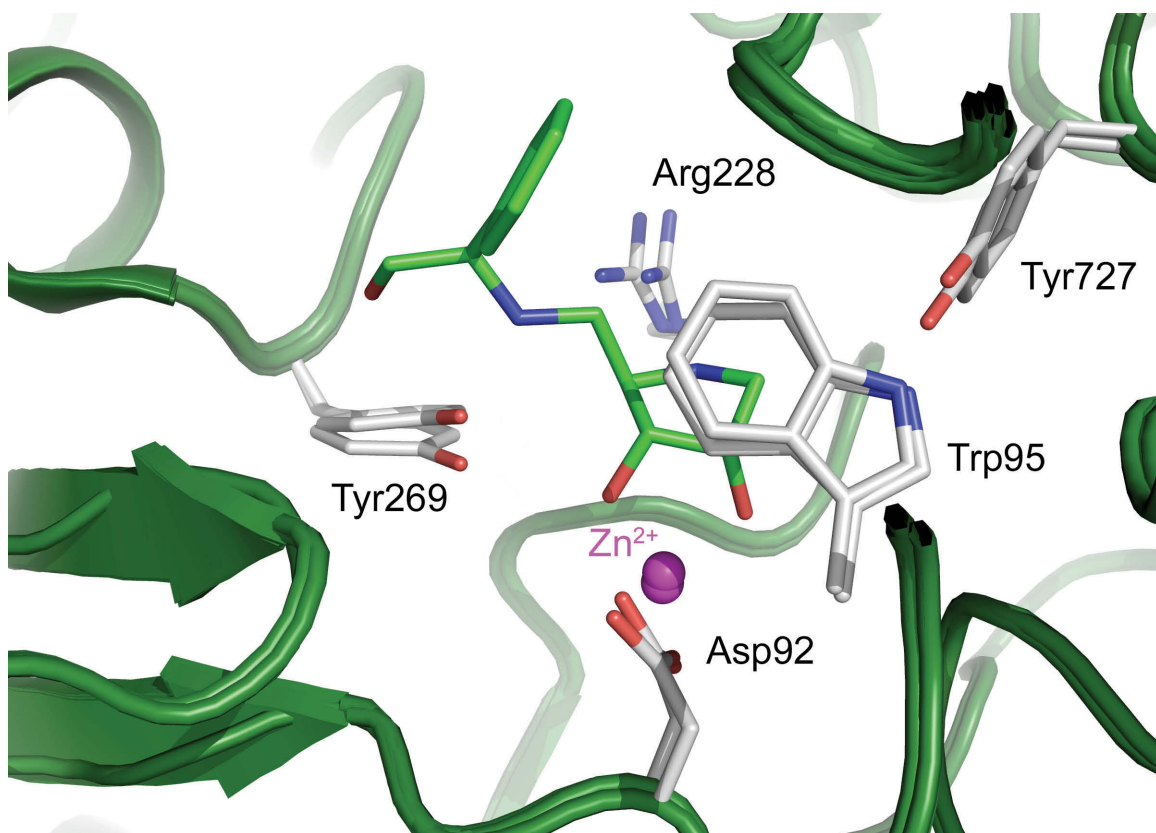
chelated in most of the cases. Compounds **1**, **3** and **5** were generally docked in a binding mode similar to the crystal structure but in a few cases a single alcohol was interacting with the metal, the second alcohol interacting with Asp204, or a different second hydroxyl group was chelating the zinc. Similarly, **2** was often docked with a single alcohol (e.g., O6) interacting with the metal. Interestingly, although the metal chelation was treated as a standard non-bonded interaction in AutoDock, only 5% of the docked structures did not involve the binding of at least one alcohol to the zinc atom. These observations were consistent with our previous report.<sup>46</sup> However, the detailed geometry was not accurately predicted, with only 35% of the ligation being predicted with a deviation of less than 1.0 Å relative to the crystal structure. Surprisingly, even though LigandFit was not among the most accurate docking programs, the metal ligation was fairly well predicted as shown in Figure 2.5b. In contrast, FITTED was not very predictive when considering the metal chelation even though overall it ranked second. The lack of a specific metal binding term in the FITTED scoring function may be responsible for this observation.

Overall, Glide clearly outperformed the other programs. As discussed above the apparent poor accuracy of nearly 50% should be taken with great care as most of the compounds were docked properly, the main deviation being attributed to the solvent exposed moieties. When these moieties were not considered, accuracies higher than 90% were recorded with Glide. The other programs poorly predicted both the metal chelation and the location of the solvent exposed groups.

### **2.3.7 Docking to conformational ensembles and flexible proteins**

Comparing the self-docking data to the cross-docking data revealed the sensitivity of the seven programs for the protein structure. Figure 2.7 shows a superposition of 5 of the 10 selected crystal structures. In this figure only the largest side chain moves are shown. For these five residues (Tyr727, Trp95, Arg228, Tyr269 and Asp92), moves of about 1 Å for specific atoms were observed. The location of the zinc atom also varied. Although 1 Å can be seen as negligible, it is roughly the size of an atom

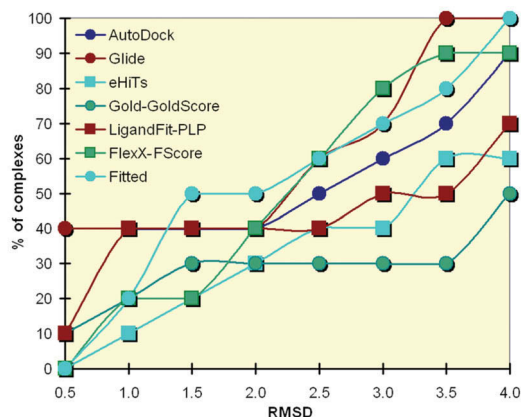
and can preclude the proper binding of a functional group of the ligand, significantly affecting the docking accuracy.



**Figure 2.7** Superposition of 1HWW, 1HXK, 1R34, 1TQT and 2F18 protein structures (backbone green, side chains grey), and ligand 8 (green). Only selected residues of each structure are shown to illustrate the largest displacements. This figure was generated with PyMOL.

The clear dependence of the docking accuracy of ligands (i.e., **7** and **10**) on the protein structure led us to consider docking to flexible proteins. One obvious method was to dock the compounds to the 10 protein structures. Then, the best scoring pose among the ten docked structures for each inhibitor was considered. Using this approach (docking to multiple conformations), slight to good improvement was observed (Figure 2.8). Again, Glide and FITTED appeared as the best two programs. However, these data were collected for only 10 ligands and should not be considered as representative of the overall accuracy of the assessed programs. At low RMSD's (below 1.0 Å), LigandFit/PLP and Glide outperformed the other programs. Their ability to predict the zinc chelation allowed these two programs to very accurately predict the binding mode of the smallest inhibitors.





**Figure 2.8** Docking to protein ensembles. Accuracy for the 7 programs assessed to dock the 10 ligands to an ensemble of the 10 protein structures, measured as the RMSD between the docked poses and the crystal structure binding mode for all ligand heavy atoms.

We next looked at the flexible versions of the programs. An induced-fit docking protocol has recently been made available by Schrödinger.<sup>49</sup> However, the average CPU time required for a single run exceeds 10 hours and would therefore be a major limitation to the use of this protocol for drug design. We decided not to include this method in our study. Flexible protein versions of AutoDock<sup>51</sup> and of FlexX, namely FlexE,<sup>50</sup> have been proposed and were used together with FITTED<sup>30</sup> as a complement to the cross-docking studies. However, no significant improvement comparatively to the docking to multiple conformations was observed (data not shown). This indicated that the failures observed were not due to the flexibility of the protein, but to inherent limitations of the docking/scoring methods used by each of the programs evaluated.

### 2.3.8 Scoring accuracy

When looking at the predicted activities of these weak to strong inhibitors with any of the programs assessed, the scores did not correlate well with the observed activities. The most active swainsonine (**1**) was predicted to be one of the least active inhibitors of the set with all the scoring functions. This revealed that the scoring functions used can accurately discriminate between the different poses, hence predicting reasonable binding modes, but not between compounds of different affinities. In fact, it is well known that the accuracy of current scoring

functions for small compounds is still poor,<sup>69,70</sup> with large compounds being often assigned higher scores than small compounds by most of the scoring functions.<sup>71</sup>

### 2.3.9 Virtual screening using Glide

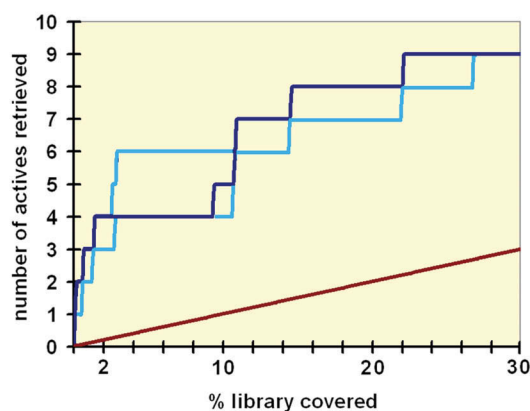
We next turned our attention to the use of docking-based virtual screening (VS) tool to screen potential  $\alpha$ -mannosidase inhibitors. As discussed above, Glide/GlideScore was more accurate in predicting the correct binding mode than any of the other programs assessed. We therefore decided to restrict the VS study to Glide alone, as VS data from a program unable to provide the correct binding mode would be hard to interpret. However, although the overall accuracy of Glide was excellent, the apparent poor accuracy of the scoring function may be a hurdle in virtual screening. In order to evaluate the performance of Glide for VS, we seeded a library of decoys with previously reported GMII inhibitors and docked the complete library to a single dGMII protein structure. The protein structure from complex 1PS3 was found to be a fair representative (by visual inspection and comparison of active site side chain RMSD) of the set of 10 proteins used in the docking study, hence its selection as the target for the VS. Then, a set of 1000 decoys used by the Schrödinger team to benchmark Glide<sup>54</sup> was seeded with ten active compounds shown in Table 2.4. Compound **3** (protein 1PS3 natural ligand) was purposely not selected to prevent biasing self-docking. The activity for some of the compounds was measured on jack bean GMII, a common model for hGMII as discussed in the introduction. At first sight these active compounds **11-17**<sup>72-78</sup> can be viewed as poor inhibitors, however low micromolar inhibitors are considered strong inhibitors of GMII.<sup>79,80</sup> In addition, one needs to bear in mind that the primary goal of a VS study is to discover micromolar lead compounds from large libraries and not nanomolar compounds, which are often developed through lead optimization.

**Table 2.4** Structures of  $\alpha$ -mannosidase inhibitors used for virtual screening evaluation. The inhibitory activity is given as  $IC_{50}$  on *Drosophila* GMII unless otherwise noted.

Compound	$IC_{50}$ ( $\mu M$ )	Inhibitor structure
1	0.017	
4	70	
8	80	
11	0.8 <sup>†,71</sup>	
12	9 <sup>†,72</sup>	
13	10 <sup>*,73</sup>	
14	12 <sup>*,74</sup>	
15	0.5 <sup>†,75</sup>	
16	40 <sup>†,76</sup>	
17	20 <sup>†,77</sup>	

<sup>†</sup> K<sub>i</sub> on jack bean GMII; \* IC<sub>50</sub> on jack bean GMII

Figure 2.9 shows the number of known hits retrieved when increasing the fraction of the ranked list. A charge of 1 or 2 (protonated amines) was assigned to each compound and alternate protonation modes were considered. For instance, compound **17** had to be monoprotinated to be able to properly chelate the zinc cation. The tetrazole ring of compound **13** could be protonated at different locations, or not at all, having an estimated *pK<sub>a</sub>* of around 5. The dark blue line in Figure 2.9 shows the accuracy of docking all the compounds with all the amines protonated and the tetrazole ring of **13** protonated at position 2. The light blue lines represent the results based on the best or worst scores of each compound when considering all the possible protonation states. Obviously, the final score did not take into account the energy required to change the protonation state from the one stable in solution to the one in the protein binding site. As can be seen in Figure 2.9, half of the active compounds were retrieved in the top 10% regardless of the protonation states selected. However, between 3 and 5 compounds were recovered in the top 2% depending on the chosen protonation states. We noticed that the protonation state was a critical factor to consider, as compounds with two amines (*i.e.*, **1**, **16** and **17**) provided quite different scores with alternative protonation states.



**Figure 2.9** Number of known active compounds recovered as a function of the percentage of the ranked list. Dark blue: standard protonation states, light blue: best and worse scores with various protonation states, brown: random.

Overall, this was a very promising result that correlated well with the good performance of Glide in virtual screening studies.<sup>54</sup> Using the same decoy set, Halgren and co-workers recovered from 50 to 90% of the actives in the top 10% with most of the proteins studied and less than 50% with p38 MAP kinase. Clearly, Glide is a promising tool for the virtual screening of  $\alpha$ -mannosidase inhibitors considering the known poor performance of docking methods with metal-containing proteins. As can be seen in Table 2.5, the most active compounds **1**, **11** and **15** were assigned lower scores than less active inhibitors, which confirmed the lack of accuracy of GlideScore to rank some of the actives. Nevertheless, having 8 out of 10 seeded actives in the top 15% of the ranked library demonstrated that GlideScore is indeed appropriate for the retrieval of mannosidase inhibitors.

**Table 2.5** Results for the virtual screening of mannosidase inhibitors. The ranking corresponds to the order of the compound in the sorted list of scores; the score is the GlideScore value of the best docked pose for the compound.

Compound	Ranking	Score	IC <sub>50</sub> /K <sub>i</sub> ( $\mu$ M)
<b>4</b>	2	-11.99	70
<b>8</b>	3	-10.72	80
<b>14</b>	8	-10.16	12*
<b>13</b>	27	-9.94	10*
<b>12</b>	15	-9.82	9*
<b>17</b>	30	-9.38	20*
<b>15</b>	110	-8.25	0.5*
<b>11</b>	147	-8.04	0.8*
<b>1</b>	222	-7.71	0.017
<b>16</b>	362	-7.22	40*

\* Activity value for jack bean GMII

## 2.4 Conclusions

In medicinal chemistry, crystallography and computational chemistry are well established tools at the lead generation stage. We have described herein three new three-dimensional structures of *Drosophila melanogaster* Golgi  $\alpha$ -mannosidase II:inhibitor complexes and their application to assessing the ability of seven available docking programs to predict the binding mode and binding affinity of  $\alpha$ -mannosidase II inhibitors. Overall, Glide outperformed the other docking programs

followed by FITTED, GOLD, AutoDock and FlexX. eHiTs was found to be the least accurate. Unexpectedly, the prediction of the metal coordination geometry appeared to be best with Glide/GlideScore even compared to other programs that included a specific term for metal ligation and coordination geometries.

Although the docked poses were often close to the observed binding modes, the predicted binding constants were not well correlated with the observed inhibition data. The highly active inhibitor swainsonine (**1**) was ranked among the least actives in most of the cases. We believe that metal ligation and solvation were not adequately evaluated in the tested scoring functions, and that large molecules were over-scored. In order to evaluate the impact of this apparently poor scoring on the performance of Glide, we carried out a VS study and were pleased to obtain enrichment factors in the range observed with non-metal containing proteins. As previously observed, swainsonine, the most active inhibitor considered, was found to be the outlier of the set with a score much lower than less active inhibitors.

In summary, using Glide, small inhibitors were docked with excellent accuracy (RMSD < 1.1 Å), while larger inhibitors with solvent-exposed polar and nonpolar functional groups were docked with good accuracy (RMSD ~ 2.5 Å). More specifically, the zinc ligation was well predicted in most cases of self and cross-docking, while the largest deviations arose from the solvent-exposed moieties. Finally, we believe that the application of Glide to the VS of large libraries of compounds is a promising strategy for the discovery of novel GMII inhibitors.

## **2.5 References**

1. Dwek, R. A., Glycobiology: Toward understanding the function of sugars. *Chem. Rev.* **1996**, 96, 683-720.
2. Kornfeld, R.; Kornfeld, S., Assembly of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **1985**, VOL. 54, 631-664.
3. Roth, J., Protein N-glycosylation along the Secretory Pathway: Relationship to organelle topography and function, protein quality control, and cell interactions. *Chem. Rev.* **2002**, 102, 285-303.

4. Herscovics, A., Importance of glycosidases in mammalian glycoprotein biosynthesis. *Biochim. Biophys. Acta* **1999**, *1473*, 96-107.
5. Moremen, K. W.; Trimble, R. B.; Herscovics, A., Glycosidases of the asparagine-linked oligosaccharide processing pathway. *Glycobiology* **1994**, *4*, 113-125.
6. Sears, P.; Wong, C. H., Enzyme action in glycoprotein synthesis. *Cell. Mol. Life Sci.* **1998**, *54*, 223-252.
7. Goss, P. E.; Baker, M. A.; Carver, J. P.; Dennis, J. W., Inhibitors of carbohydrate processing: A new class of anticancer agents. *Clin. Cancer Res.* **1995**, *1*, 935-944.
8. Dennis, J. W.; Granovsky, M.; Warren, C. E., Protein glycosylation in development and disease. *BioEssays* **1999**, *21*, 412-421.
9. Dennis, J. W.; Granovsky, M.; Warren, C. E., Glycoprotein glycosylation and cancer progression. *Biochim. Biophys. Acta* **1999**, *1473*, 21-34.
10. Couldrey, C.; Green, J. E., Metastases: The glycan connection. *Breast Cancer Res.* **2000**, *2*, 321-323.
11. Elbein, A. D.; Molyneux, R. D., Inhibitors of Glycoprotein Processing. In *Iminosugars as Glycosidase Inhibitors*, Stütz, A. E., Ed. Wiley-VCH: Weinheim, 1999; pp 216-251.
12. Colegate, S. M.; Dorling, P. R.; Huxtable, C. R., Spectroscopic Investigation of Swainsonine - Alpha-Mannosidase Inhibitor Isolated from Swainsona-Canescens. *Aust. J. Chem.* **1979**, *32*, 2257-2264.
13. Kino, T.; Inamura, N.; Nakahara, K., Studies of an immunomodulator, swainsonine. II. Effect of swainsonine on mouse immunodeficient system and experimental murine tumor. *J. Antibiot.* **1985**, *38*, 936-940.
14. Goss, P. E.; Reid, C. L.; Bailey, D.; Dennis, J. W., Phase IB clinical trial of the oligosaccharide processing inhibitor swainsonine in patients with advanced malignancies. *Clin. Cancer Res.* **1997**, *3*, 1077-1086.
15. Howard, S.; Braun, C.; McCarter, J.; Moremen, K. W.; Liao, Y. F.; Withers, S. G., Human lysosomal and jack bean  $\alpha$ -mannosidases are retaining glycosidases. *Biochem. Biophys. Res. Commun.* **1997**, *238*, 896-898.

16. Rabouille, C.; Kuntz, D. A.; Lockyer, A.; Watson, R.; Signorelli, T.; Rose, D. R.; Van Den Heuvel, M.; Roberts, D. B., The Drosophila GMII gene encodes a Golgi  $\alpha$ -mannosidase II. *J. Cell Sci.* **1999**, *112*, 3319-3330.
17. Van den Elsen, J. M. H.; Kuntz, D. A.; Rose, D. R., Structure of Golgi  $\alpha$ -mannosidase II: A target for inhibition of growth and metastasis of cancer cells. *EMBO J.* **2001**, *20*, 3008-3017.
18. Numao, S.; Kuntz, D. A.; Withers, S. G.; Rose, D. R., Insights into the Mechanism of Drosophila melanogaster Golgi  $\alpha$ -Mannosidase II through the Structural Analysis of Covalent Reaction Intermediates. *J. Biol. Chem.* **2003**, *278*, 48074-48083.
19. Rose, D. R.; Kuntz, D. A.; van den Elsen, J. M. H., Crystal Structure of Drosophila  $\alpha$ -Mannosidase II and Swainsonine Complexes its Use for Identifying Mannosidase II Activity-Modulating Ligands and Therapeutic Applications. *US Pat. Appl.* **2002**, 2002172670.
20. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325-332.
21. Rester, U., Dock around the clock - Current status of small molecule docking and scoring. *QSAR Comb. Sci.* **2006**, *25*, 605-615.
22. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580-594.
23. Cavasotto, C. N.; Orry, A. J. W.; Abagyan, R. A., The Challenge of Considering Receptor Flexibility in Ligand Docking and Virtual Screening. *Curr. Comput.-Aided Drug. Des.* **2005**, *1*, 423-440.
24. Friesner, R. A.; Knoll, E. H.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Shaw, D. E.; Shenkin, P. S.; Shelley, M.; Perry, J. K.; Francis, P., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
25. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748.



26. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
27. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289-307.
28. Zsoldos, Z.; Szabo, I.; Szabo, Z.; Johnson, A. P., Software tools for structure based rational drug design. *J. Mol. Struct.:THEOCHEM* **2003**, *666-667*, 659-665.
29. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J.; Halliday, R. S.; Hart, W. E.; Belew, R. K., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.
30. 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.
31. Kuntz, D. A.; Ghavami, A.; Johnston, B. D.; Pinto, B. M.; Rose, D. R., Crystallographic analysis of the interactions of *Drosophila melanogaster* Golgi  $\alpha$ -mannosidase II with the naturally occurring glycomimetic salacinol and its analogues. *Tetrahedron: Asymmetry* **2005**, *16*, 25-32.
32. Shah, N.; Kuntz, D. A.; Rose, D. R., Comparison of Kifunensine and 1-Deoxymannojirimycin Binding to Class I and II  $\alpha$ -Mannosidases Demonstrates Different Saccharide Distortions in Inverting and Retaining Catalytic Mechanisms. *Biochemistry* **2003**, *42*, 13812-13816.
33. *Maestro*, 7.0; Schrödinger, LLC: Portland, OR, 2004.  
<http://www.schrodinger.com>
34. *Jaguar*, 6.0; Schrödinger, LLC: Portland, OR, 2004.
35. Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W., Parallel Ab Initio and Molecular Mechanics Investigation of Polycoordinated Zn(II) Complexes with Model Hard and Soft Ligands: Variations of Binding Energy and of Its Components with Number and Charges of Ligands. *J. Comput. Chem.* **2000**, *21*, 1011-1039.
36. *Sybyl*, 7.0; Tripos, Inc.: St. Louis, MO, 2005.

37. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity--a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-3228.
38. *Cerius2*, 4.10; Accelrys, Inc.: San Diego, CA, 2005.
39. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Bourne, P. E.; Shindyalov, I. N., The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
40. Kavlekar, L. M.; Kuntz, D. A.; Wen, X.; Johnston, B. D.; Svensson, B.; Rose, D. R.; Pinto, B. M., 5-Thio-D-glycopyranosylamines and their amidinium salts as potential transition-state mimics of glycosyl hydrolases: Synthesis, enzyme inhibitory activities, X-ray crystallography, and molecular modeling. *Tetrahedron: Asymmetry* **2005**, *16*, 1035-1046.
41. Popowycz, F.; Gerber-Lemaire, S.; Rodriguez-García, E.; Schütz, C.; Vogel, P., Synthesis and  $\alpha$ -mannosidase inhibitory evaluation of (2R,3R,4S)- and (2S,3R,4S)-2-(aminomethyl)pyrrolidine-3,4-diol derivatives. *Helv. Chim. Acta* **2003**, *86*, 1914-1948.
42. Kuntz DA, unpublished results.
43. Moitessier, N.; Westhof, E.; Hanessian, S., Docking of aminoglycosides to hydrated and flexible RNA. *J. Med. Chem.* **2006**, *49*, 1023-1033.
44. Hoops, S. C.; Anderson, K. W.; Merz Jr, K. M., Force Field Design for Metalloproteins. *J. Am. Chem. Soc.* **1991**, *113*, 8262-8270.
45. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D., Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225-242.
46. Hanessian, S.; Moitessier, N.; Therrien, E., A comparative docking study and the design of potentially selective MMP inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 873-881.
47. Hanessian, S.; MacKay, D. B.; Moitessier, N., Design and synthesis of matrix metalloproteinase inhibitors guided by molecular modeling. Picking the S1 pocket using conformationally constrained inhibitors. *J. Med. Chem.* **2001**, *44*, 3074.

48. Kerwin, S. M., Computer Software Review: eHiTS 5.1.6, SimBioSys Inc. *J. Am. Chem. Soc.* **2005**, *127*, 8899-8900.
49. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R., Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534-553.
50. Claußen, H.; Buning, C.; Rarey, M.; Lengauer, T., FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377-395.
51. Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S., Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 34-40.
52. Moitessier, N.; Therrien, E.; Hanessian, S., A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic  $\beta$ -secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885-5894.
53. *Glide*, 3.5; Schrödinger, LLC: Portland, OR, 2004.
54. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750-1759.
55. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425-445.
56. Kramer, B.; Rarey, M.; Lengauer, T., Evaluation of the FlexX incremental construction algorithm for protein- ligand docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228-241.
57. Rarey, M.; Kramer, B.; Lengauer, T., Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369-384.
58. Hindle, S. A.; Rarey, M.; Buning, C.; Lengauer, T., Flexible docking under pharmacophore type constraints. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129-149.
59. *AutoDock*, 3.0; The Scripps Research Institute: La Jolla, CA, 1998.

60. AutoDock 4 beta release. [http://autodock.scripps.edu/news\\_files/autodock-4-beta-release](http://autodock.scripps.edu/news_files/autodock-4-beta-release) (accessed Jan 25, 2007)
61. *GOLD*, 3.0; CCDC: Cambridge, UK, 2005.
62. *eHiTS*, 5.1; SimBioSys: Toronto, ON, Canada, 2005.
63. Reid D, personal communication.
64. For more information see: [http://www.simbiosys.ca/ehits/ehits\\_newfeatures.html](http://www.simbiosys.ca/ehits/ehits_newfeatures.html) (accessed Jan 25, 2007).
65. Fletcher, R.; Reeves, C. M., Function Minimization by Conjugate Gradients. *Comp. J.* **1964**, *7*, 149-154.
66. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269-288.
67. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791-804.
68. Kuntz, D. A.; Liu, H.; Bols, M.; Rose, D. R., The role of the active site Zn in the catalytic mechanism of the GH38 Golgi  $\alpha$ -mannosidase II: Implications from neuromycin inhibition. *Biocatal. Biotransform.* **2006**, *24*, 55-61.
69. Wang, R.; Lu, Y.; Wang, S., Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287-2303.
70. Ferrara, P.; Gohlke, H.; Price, D. J.; Brooks, C. L., III; Klebe, G., Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032-3047.
71. Jacobsson, M.; Karlén, A., Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1334-1343.
72. Picasso, S.; Chen, Y.; Vogel, P., Glycosidase inhibition by 1,5-dideoxy-1,5-iminoctitols and 1,2,6,7,8-pentahydroxyindolizidines. *Carbohydrate Letters* **1994**, *57*, 1-8.
73. Papandreou, G.; Tong, M. K.; Ganem, B., Amidine, amidrazone, and amidoxime derivatives of monosaccharide aldonolactams: Synthesis and evaluation as glycosidase inhibitors. *J. Am. Chem. Soc.* **1993**, *115*, 11682-11690.
74. Davis, B. G.; Brandstetter, T. W.; Hackett, L.; Winchester, B. G.; Nash, R. J.; Watson, A. A.; Griffiths, R. C.; Smith, C.; Fleet, G. W. J., Tetrazoles of manno- and

rhamno-pyranoses: Contrasting inhibition of mannosidases by [4.3.0] but of rhamnosidase by [3.3.0] bicyclic tetrazoles. *Tetrahedron* **1999**, *55*, 4489-4500.

75. Tatsuta, K.; Mijura, S.; Ohta, S.; Gunji, I., Syntheses and glycosidase inhibiting activities of nagstatin analogs [2]. *J. Antibiot.* **1995**, *48*, 286-288.

76. Eis, M. J.; Rule, C. J.; Wurzburg, B. A.; Ganem, B., Synthesis of 1,4,6-trideoxy-1,4-imino-D-mannitol: A potent  $\alpha$ -mannosidase inhibitor. *Tetrahedron Lett.* **1985**, *26*, 5397-5398.

77. Limberg, G.; Lundt, I.; Zavilla, J., Deoxyiminoalditols from aldonic acids VI. Preparation of the four stereoisomeric 4-amino-3-hydroxypyrrolidines from bromodeoxytetronic acids. Discovery of a new  $\alpha$ -mannosidase inhibitor. *Synthesis* **1999**, 178-183.

78. Le Merrer, Y.; Poitout, L.; Depezay, J. C.; Dosbaa, I.; Geoffroy, S.; Foglietti, M. J., Synthesis of azasugars as potent inhibitors of glycosidases. *Bioorg. Med. Chem.* **1997**, *5*, 519-533.

79. Asano, N.; Nash, R. J.; Molyneux, R. J.; Fleet, G. W. J., Sugar-mimic glycosidase inhibitors: Natural occurrence, biological activity and prospects for therapeutic application. *Tetrahedron: Asymmetry* **2000**, *11*, 1645-1680.

80. Ganem, B., Inhibitors of Carbohydrate-Processing Enzymes: Design and Synthesis of Sugar-Shaped Heterocycles. *Acc. Chem. Res.* **1996**, *29*, 340-347.

## **Chapter 3:      Assessment of the performance of scoring functions on complexes with flexible and solvated proteins**

### ***3.1 Introduction***

As the timeframe and costs of the traditional drug discovery approach are ever-increasing, computational/virtual approaches arose as promising techniques in the medicinal chemist toolkit. In particular, docking-based virtual screening (VS) methods are becoming increasingly popular as a fast, cost-effective alternative or complement to classical high throughput screening (HTS).<sup>1</sup> The purpose of docking methods in this context is twofold: i. prediction of the binding mode of a ligand to a given biologically relevant target receptor, enzyme or nucleic acid, and ii. prediction of the binding affinity of the complex formed.<sup>2</sup> To carry out these two tasks, docking programs rely on two major components: a conformational search algorithm and a scoring function. The former samples the translational, rotational and conformational space of the complex, and several approaches that accurately dock flexible ligands to proteins have been disclosed.<sup>3,4</sup> As soon as a putative binding mode of the ligand -referred to as a pose- is proposed, its binding affinity must be predicted. Scoring functions often provide a fast evaluation of the free energy of binding during and after the conformational sampling stage.<sup>5</sup> Despite the intense effort in the area of docking methods in recent years, the moderate performance of scoring functions revealed by independent comparative studies remains the chief issue to address in the improvement of docking methods<sup>2</sup> and strategies have been developed to increase the identification of active compounds.<sup>6</sup>

Commonly used scoring functions can be classified into force-field based, knowledge-based and regression-based, even though some functions may combine two approaches (e.g., regression and FF-based AutoDock scoring function). Several issues need to be considered when assessing or selecting scoring functions for docking and scoring of potential drugs. First, force fields (e.g., AMBER in DOCK

scoring function) are known to overestimate the binding affinities and the calculated intermolecular energy values need to be scaled down for more accurate predictions. An example is the Linear Interaction Energy (LIE) method.<sup>7</sup> In addition, force fields only approximate the enthalpic interaction energy, disregarding some contributions to the free energy of binding such as entropy and solvation. Second, knowledge-based potentials are developed from statistical analysis of protein/ligand complexes regardless of their affinities. Third, regression-based scoring functions, also called empirical scoring functions, are also trained against a set of protein/ligand complexes that are related to known 3D structures and affinities. Clearly, the scoring functions from the last two categories are strongly dependent on the training set used to derive them and rely on the accuracy of the binding affinity data, crystallographic experiments, model fitting and transferability of the parameters to other complexes not present in the training sets.<sup>8</sup> To partially address this issue, large data sets (i.e., the whole PDB database) can be used. However, as these training sets do not contain compounds that are too large to fit into the binding site or are otherwise inactive, the potentials derived from these sets may fail to discriminate inactive compounds from actives thus leading to the occurrence of false positives.

We have recently reported the development of FITTED (versions 1.0, 1.5 and 2.6), a docking program accounting for protein flexibility and essential water molecules.<sup>9-11</sup> In parallel, we have reported the evaluation of the impact of ligand and protein structures and presence of water molecules on the pose prediction accuracy of major docking programs.<sup>11</sup> In order to complement this previous study and later develop a scoring function that accounts for these two aspects, we have investigated the impact of protein flexibility and water molecules on the accuracy of several commonly used scoring functions including our current version of RankScore.<sup>11</sup> Thus we report herein the development of two sets of protein/ligand complexes, their selection criteria, their preparation as well as their use as testing sets for the evaluation of 18 commonly used scoring functions.

## 3.2 Methods

### 3.2.1 Training set selection criteria

The accuracy of a scoring function is largely dependent on the training set used to calibrate it. It has been reported that several commonly used scoring functions are less accurate than the ligand molecular weight (MW) used as a descriptor.<sup>12,13</sup> Indeed, a close look at training sets used in the development and evaluation of scoring functions shows that there is sometimes a strong correlation between binding affinities and MW, an artifact that should be considered. It is well known that truncating a large ligand often results in a significant loss of binding affinity, a property clearly captured by several scoring functions. However, as VS is often carried out with libraries of drug-like, lead-like and even fragment-like molecules with similar molecular weights, these scoring functions do not perform as well. The current challenge is therefore to develop a scoring function able to discriminate between actives and inactives of similar sizes. Some training sets also lack diversity in protein and ligand structures. For instance, scoring functions may be developed from training sets excluding metalloenzymes and/or highly polar enzymes (e.g., neuraminidase), therefore being poorly transferable to these classes of targets. In addition, most of these sets are developed from crystal structures and therefore trained to perform well in self-docking experiments. This is a significant limitation as cross-docking is a more realistic experiment simulating a VS situation. More recently, Verdonk and co-workers have reported a training set carefully prepared using strict criteria.<sup>14,15</sup> In order to evaluate the state-of-the-art in the development of scoring functions, we propose herein to report our training/testing sets that follow a number of restrictions and conditions:

- i. The training set should be large enough to be statistically relevant. We targeted a minimum of 200 complexes in order to guarantee an accurate evaluation of existing scoring functions and a good predictivity of the scoring function that will be eventually developed.<sup>16</sup>



- ii. The ligands should be as diverse as possible (both in shape, bioactivity and functional groups) to assess the transferability of scoring functions and eventually ensure transferability of the scoring function to be developed.
- iii. The ligand molecular weight should be higher than 250, and not exceed 700.
- iv. The affinity of the ligand towards the co-crystallized receptor should be known.
- v. Crystal structures should be available at a good resolution ( $\leq 2.5$  Å). Although this criterion is not strict enough to evaluate the “quality” of the complexes, it is easily accessible.<sup>17</sup>
- vi. Some proteins should appear more than 5 times in this set so that cross-docked structures can be considered.
- vii. Proteins with both hydrophobic and hydrophilic binding sites should be included.
- viii. Different aspects of protein-ligand binding, such as water-mediated binding and metal binding should be represented.
- ix. Correlation between ligand molecular weight and binding affinities should be as small as possible.
- x. Metal-containing and covalently bound ligands should not be included as they may be poorly defined in scoring functions and would require specific terms.
- xi. Binding affinities should cover a range as wide as possible without overweighting one range of affinities and should include as many poor binders as possible.

Due to these many criteria, the set developed herein is very different from the set we have previously used to assess docking programs.<sup>11</sup>

### **3.2.2 Correlation metrics and statistical significance**

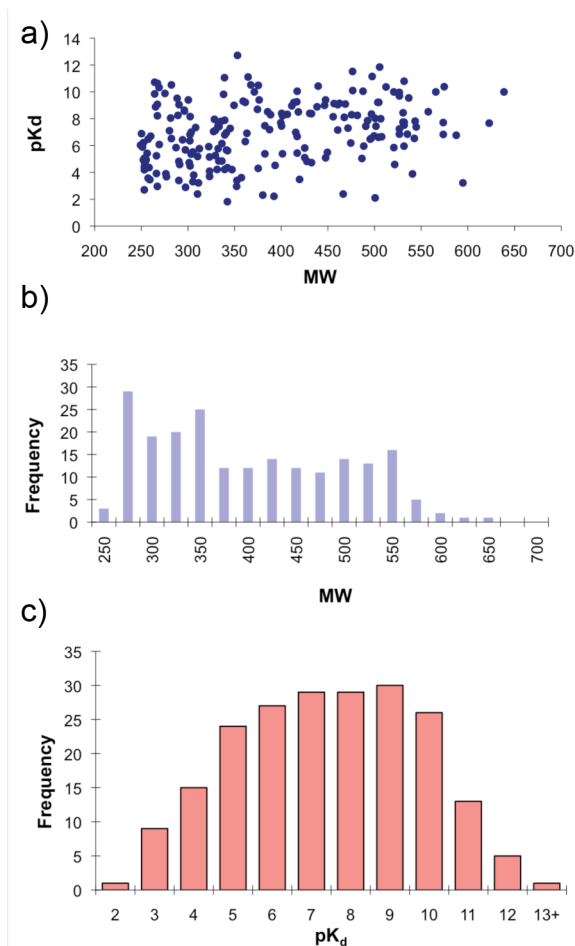
Most commonly, the square of Pearson’s correlation coefficient ( $r^2$ ) and Spearman’s  $\rho$  are used to assess the correlation between observed and predicted affinities.

While  $r^2$  is the traditional correlation metrics, measuring the correlation between experimental binding affinities and *scores*,  $\rho$  is a non-parametric measure of the correlation between the *ranked lists* of experimental binding affinities and predicted scores. A  $\rho$  of  $\pm 1.0$  corresponds to a perfect match between the two ranked lists (but a negative  $\rho$  indicates an inverse order for one of the lists), while a value of 0.0 is consistent with random ordering. As suggested by Nicholls and Jain, and a reviewer of this manuscript, we also considered Kendall's tau ( $\tau$ ) as an alternative to Spearman's for the assessment of the rank-ordered correlation.<sup>17</sup> Kendall's has the advantage of being more robust than Spearman's, while also being easier to interpret, as it is an estimate of the probability of having the same trend between two sets of ranks. We used the bootstrap technique to assess the statistical significance of the correlation coefficients: random subsets of the data set were drawn, allowing for duplicates, and the correlation of each subset was calculated. The range containing 95% of the values is taken as the confidence interval for the given descriptor (either  $\rho$  or  $\tau$ ).

### 3.2.3 Training set preparation

A meticulous preparation of the complexes was believed to be essential to provide objective results. As described above, the ligands were chosen to be varied in shape, size, bioactivity and functional groups, with known activity towards a given receptor. To follow the criterion vi, a careful selection of complexes from the PDB led to the selection of 58 complexes of five highly studied proteins: HIV-1 protease, thrombin, trypsin and matrix metalloproteases (MMP-3 - stromelysin 1 and MMP-8 - collagenase 2). This set was next expanded through the addition of complexes from the PDDBind database,<sup>18,19</sup> taking into account the chemical diversity of the ligands (criteria ii and x), an even distribution of the binding affinity of the complexes (criterion xi) and a proper selection of proteins (criteria vii and viii). After filtering out the complexes that were not well characterized (i.e., missing residues, multiple ligands in binding site), a set of 223 complexes was obtained. In order to meet criterion ix, another 14 complexes were removed and led to a final set of 209 complexes with 82 proteins represented (Table 3.1, see also Appendix Table A.1).

This first set –referred to as Set 1- was used to evaluate SFs in a self-docking situation as described below. Proteins for which 5 or more structures were found were further processed. For these systems, protein and ligand structures were swapped to generate cross-docked structures. This second set of nearly 1,000 complexes was used to evaluate SFs in a cross-docking context (Set 2).



**Figure 3.1.** Properties of the training set. a) Profile of molecular weight and binding affinity dependence; b) distribution of ligand MW; c) distribution of ligand binding affinities.

**Table 3.1.** Proteins represented more than once. Set 1 is the self-docking set while Set 2 includes the cross-docked structures. The complete sets are given in the appendix (Tables A.1 and A.2).

<b>Proteins</b>	<b>Number of ligands</b>	<b>Sets</b>
Thrombin	22	1, 2
Trypsin	13	1, 2
HIV-1 protease	10	1, 2
MMP-8	8	1, 2
Factor Xa	7	1, 2
Purine nucleoside phosphorylase	6	1 <sup>a</sup>
Scytalone dehydratase	6	1, 2
Urokinase-type plasminogen activator	6	1, 2
Carbonic anhydrase	5	1, 2
MMP-3	5	1, 2
PTP-1b	5	1, 2
Acetylcholinesterase	4	1
Neuraminidase	4	1
Retinoic acid receptor gamma-1	4	1
Xylanase beta-1	4	1
2,2-Dialkylglycine decarboxylase	3	1
Cyclin dependent kinase 2	3	1
Glutathione s-transferase	3	1
Ribonuclease a	3	1
Thymidylate synthase	3	1
Carboxypeptidase	2	1
Orotidine 5'-monophosphate decarboxylase	2	1
Serine/threonine-protein kinase chk1	2	1
Sex hormone-binding globulin	2	1

<sup>a</sup> The PNP proteins were from different species and thus were not included in set #2.

As shown in Figure 3.1 (and Table 3.3), only little correlation between MWs and affinities was obtained ( $r^2=0.109$ ;  $\rho=0.330$ ;  $\tau=0.230$ ). In addition, the values of

affinity constants span 9 orders of magnitude, with an even distribution in the 0.1  $\mu\text{M}$  – 0.1 nM range. Unfortunately a very low number of millimolar ligands fulfilling the constraints defined above were found in the PDB, and this extreme range of activities is underrepresented (criterion ix). In fact, in order for a crystal structure of a protein/ligand complex to be solvable, a ligand needs to present a measurable binding affinity. In addition, crystal structures with highly active compounds provide more information to medicinal chemists and are prioritized by crystallographers.

Careful manipulations were necessary to set up the complexes for calculations. For most of the steps, the preparation was not fully automated in order to reduce potential errors. First, critical water molecules were carefully selected. To do so, crystallographic water molecules were removed unless they were involved in at least 3 hydrogen bonds with the ligand and the protein simultaneously. Next, ligand bond orders (not present in the source PDB files) were properly set, hydrogens were added and atom types and partial charges were assigned. Special attention was given to the protonation state of ionizable groups in the complexes. For instance, the catalytic dyads in most aspartyl proteases (such as HIV-1 protease) are required to be monoprotonated for the catalytic mechanism to proceed.<sup>20</sup> However, X-ray crystallography and modeling studies indicate that fully protonated states are also observed with some diol ligands.<sup>21</sup> Klebe and co-workers suggested that ligands with an ammonium group facing the catalytic dyad might stabilize the deprotonated state.<sup>22</sup> Histidine protonation was also carefully assigned by optimizing the hydrogen bond network with neighboring residues. On the ligand side, reasonable protonation states of ionizable groups were assumed. The next step was the full optimization of the hydrogen positions and ligand position (see Experimental Section) through energy minimization. Initial attempts to fully relax the complexes led to unreliable structures often far from the crystal structure. Freezing the protein and water heavy atoms and constraining ( $K = 5 \text{ kcal/mol}\cdot\text{\AA}^2$ ) ligand heavy atoms was necessary to restrict large motion of some of the ligands, thus keeping the poses close to the experimentally observed ones. As pointed out by

a reviewer, the accuracy of scoring functions may be dependent on the method of preparation of the systems. We assessed the scoring functions on unoptimized (i.e., raw PDB files) structures and observed poorer predictions. In the following sections, we will consider only the optimized poses.

#### **3.2.4 Cross-scoring training set**

To explore the sensitivity of scoring functions to the protein conformation, we decided to score every ligand/protein combinations for proteins in set 2 (see Table A.2, Appendix). First, all complexes within a family were superimposed by aligning the alpha carbons on the proteins. Then new complexes were constructed by swapping ligand and protein structures. To relieve any undesired clashes between ligand, protein and waters in these manually docked structures, a local conformational search of the ligand was performed on each complex. For this purpose, we have implemented a local search mode in our docking program FITTED. This conformational search mode was designed to optimize the intermolecular interactions among ligand, protein and waters, without greatly disturbing the initial conformation of the ligand. The RMSD (vs. crystal structure) of the resulting ligand conformations was below 1.5 Å in 84% of the cases while in the case of ligands in their cognate receptor, all poses were below 1.5 Å RMSD from the experimentally observed ones. In order to evaluate the impact of water molecules in the final scores, the cross-docked structures were optimized using the local search algorithm keeping the water molecules, and then scored with or without the water molecules present in the binding site. Alternatively, the ligand poses were also optimized with no waters included and then scored.

### ***3.3 Results and discussion***

#### **3.3.1 Accuracy of the selected scoring functions on the entire set**

With the MW-unbiased set (Set 1) in hand, we evaluated the accuracy of well-established scoring functions. Thus, two implementations of PMF (Cerius2 and CScore),<sup>23</sup> PLP1/2,<sup>24</sup> LigScore1/2,<sup>25</sup> ChemScore,<sup>26</sup> GoldScore,<sup>27</sup> XScore,<sup>16</sup> six

different versions of DrugScore,<sup>13</sup> GlideScore,<sup>28</sup> the eHiTS scoring function,<sup>29</sup> the Surflex scoring function<sup>30</sup> and its predecessor, the Hammerhead scoring function, and our first version of RankScore<sup>31</sup> were used to predict the binding affinities of the ligand set (Table 3.2). These scoring functions have been derived using different training sets, as shown in the rightmost column of Table 3.2. At this stage, we can only discuss the reported training sets used to derive these scoring functions, although we are aware that some changes may have been made to the latest releases. Knowledge-based scoring functions (PMF, DrugScore) feature the largest training sets, although in these cases the training sets are not used to adjust coefficients by regression. Of the disclosed training sets used by empirical scoring functions, XScore features the largest training set, followed by the eHiTS scoring function. The latter, includes an additional tunable property: the set of regression coefficients has been calibrated against multiple subsets of the whole Protein Data Bank, and a specific scoring function is selected for each complex to be scored.

We next looked at the overlap between these sets and our sets. In fact, our Set 1 has little in common with the previous training sets used in the development of these regression-based scoring functions: either one (FlexXScore, Hammerhead/Surflex), three (LigScore), five (ChemScore) or seven (XScore) complexes were shared. In addition, although the sets used to train DrugScore and the eHiTS scoring functions are larger and may share a greater number of structures with our Set 1, the use of most of the available PDB structures in their training dilutes the effect of a single one.

The relative ranking of the ligands for their binding affinity were also computed. The correlation between experimental and calculated data was next evaluated and compared to the reported accuracies computed with a more MW-biased testing set (Wang's set).<sup>12</sup> In the upcoming sections, we may suggest that a scoring function is better than another one although in many cases the differences are within the error bars.

**Table 3.2.** Selected scoring functions used in this study.

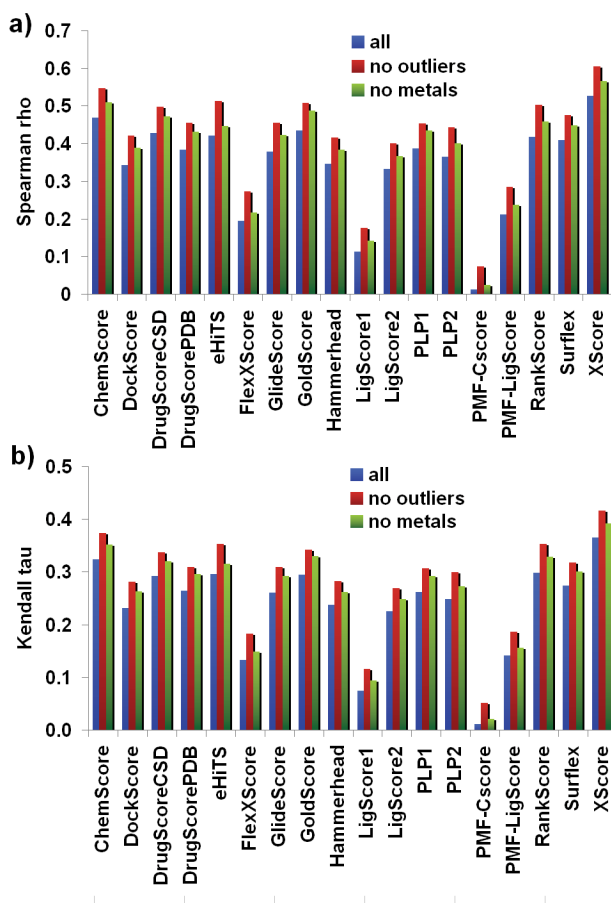
Scoring function	Implementation	Class	Training sets used for development <sup>a</sup>
ChemScore	Sybyl	Empirical	82 complexes in 5 classes
DockScore	Sybyl	FF-based	No training set
DrugScore <sup>CSD</sup>	standalone	Knowledge-based	28642 small molecules
DrugScore <sup>PDB</sup>	standalone	Knowledge-based	6026 complexes
eHiTS SF	eHiTS	Knowledge-based	133 complexes + extended training set, protein class-specific
FlexXScore	Sybyl	Empirical	45 complexes (LUDI SCORE1)
GlideScore	Glide	Empirical	82 complexes in 5 classes (ChemScore)
GoldScore	Sybyl	FF-based	No training set
Hammerhead	Cerius2	Empirical	34 complexes
LigScore1	Cerius2	Empirical	50 complexes
LigScore2	Cerius2	Empirical	112 complexes
PLP1	Cerius2	Empirical	3 complexes (DHFR, FKBP, HIV-1P)
PLP2	Cerius2	Empirical	3 complexes (DHFR, FKBP, HIV-1P)
PMF	Sybyl	Knowledge-based	697 complexes
PMF	Cerius2	Knowledge-based	697 complexes
RankScore	Fitted	FF-based	50 BACE-1 inhibitors and 4 complexes
Surflex SF	Surflex	Empirical	34 complexes
XScore	standalone	Empirical/consensus	200 complexes

<sup>a</sup> These training sets are those reported. In some cases, improved versions have been released but the training sets used to derive them have not been reported.



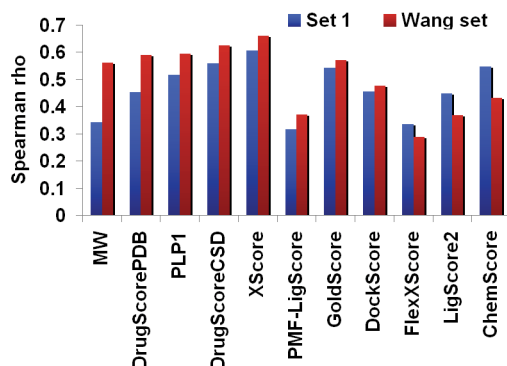
As previously noted, some reported biological activities can be highly dependent on the experimental conditions of the assay.<sup>2</sup> In order to remove part of the noise due to experimental errors, we removed the worst 5 predictions for each scoring function. A close look at these reduced sets showed that 1bn4, 1bnn, 1m0n, 1m0o, 1m0q and 1osv are the most frequently found within the selected 5 outliers and their activities were poorly predicted with all the scoring functions. We also looked at a reduced set with no metalloenzymes as it is known that metal coordination is often poorly scored and/or scoring functions not trained on metalloenzymes.<sup>32</sup>

The correlation between the scores computed with the selected scoring functions and the experimental binding affinities is low (see Table 3.3), with XScore ( $r^2=0.320$ ) being the most predictive followed by ChemScore ( $r^2=0.276$ ), DrugScore<sup>CSD</sup> ( $r^2=0.241$ ), GoldScore ( $r^2=0.235$ ), the eHiTS scoring function ( $r^2=0.228$ ) and RankScore ( $r^2=0.216$ ). However, although this correlation is a good indicator of the potential of scoring functions, we –as others-<sup>33</sup> believe that the correlation between the ranked lists using either Spearman  $\rho$  or Kendall  $\tau$  provides a more useful indication of the predictive power of scoring functions in the context of virtual screening.  $\rho$  and  $\tau$  also identify XScore ( $\rho=0.606, \tau=0.416$ ), as the most accurate scoring functions followed by ChemScore ( $\rho=0.547, \tau=0.374$ ), the eHiTS scoring function ( $\rho=0.487, \tau=0.353$ ) RankScore ( $\rho=0.482, \tau=0.353$ ), GoldScore ( $\rho=0.543, \tau=0.342$ ), DrugScore<sup>CSD</sup> ( $\rho=0.559, \tau=0.337$ ), and PLP1 ( $\rho=0.516, \tau=0.306$ ). Clearly,  $r^2$ ,  $\rho$  and  $\tau$  identified the same scoring functions as the most accurate with insignificant changes in the ranking. Among these top-scoring functions, the eHiTS scoring function is the least sensitive to the presence of outliers while GlideScore and LigScore1 are very sensitive to the presence of metalloenzymes and outliers respectively. When correlations were computed on non-metal containing enzymes, XScore remains the most accurate scoring function.



**Figure 3.2.** Spearman (a) and Kendall (b) coefficients for three subsets of the training set: in blue, 209 complexes (whole set); in red, 5 outliers removed (204 complexes); in green, all transition metal-containing proteins removed (188 complexes).

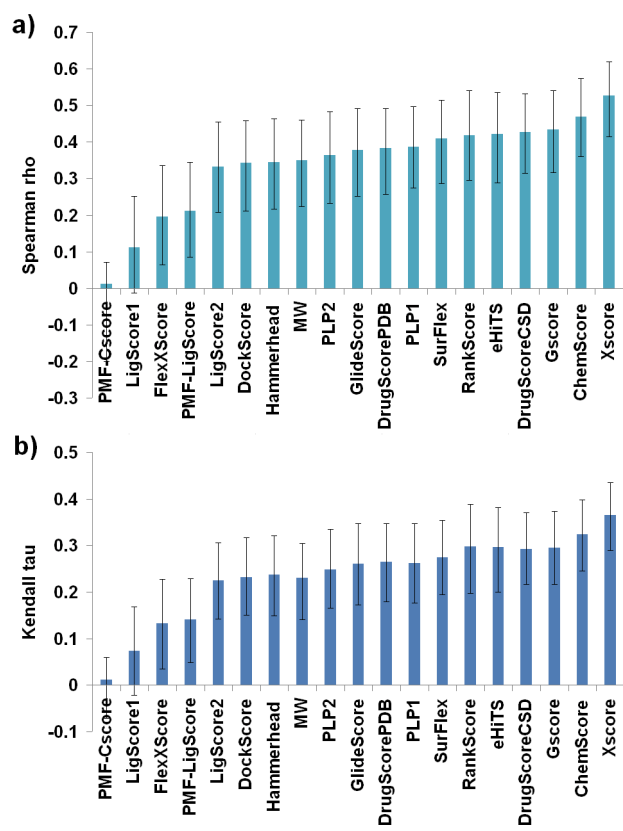
Interestingly, the collected data summarized in Table 3.3 and Figure 3.3 reveals the change in accuracy when going from Wang's set to ours and also confirms the variety of accuracies measured with this set of scoring functions. Comparing Wang's data to our data indicates that the accuracy of all the scoring functions but ChemScore, LigScore and FlexXScore were affected by the reduced dependence on MW. In fact, the major difference between the scoring functions accuracy on our set and on Wang's set is the data collected with ChemScore and LigScore2. These scoring functions were previously found to be poorer than MW used as a descriptor while in the present work, they were found to perform better, clearly capturing some of the binding aspects other than ligand size. It is not clear why these two scoring functions perform better with our more challenging set.



**Figure 3.3.** Comparison of the Spearman correlation for ten scoring functions. Blue columns indicate the correlation of the “no outliers” set (204 complexes); red bars denote the Spearman coefficients as reported by Wang *et al.*<sup>12</sup>

When considering the error bars arising from bootstrapping the data set with either correlation coefficient (Figure 3.4), XScore appears to be better than most of the other scoring functions by a marginal value, while LigScore1, FlexXScore and both implementations of PMF poorly rank-ordered the set by scores. As a matter of fact, with the CScore implementation of PMF and LigScore1, one cannot rule out the possibility of a chance correlation, as a null correlation coefficient is included in the respective interval of confidence for both scoring functions.

Although the eHiTS scoring function and DrugScore have been trained on large sets, their accuracy does not exceed the ones observed with some regression-based scoring functions. We believe that the uneven number of protein complexes represented in the Protein Data Bank (e.g., several hundreds of thrombin/ligand complexes, 1% of the PDB) may lead to the overtraining of these scoring functions for these specific proteins. As a result, increasing the number of structures in the training set if not accompanied by an increase in the diversity is not expected to improve the training and transferability of the developed scoring functions.



**Figure 3.4.** Comparison of Spearman (a) and Kendall (b) coefficients for the scoring functions considered. Error bars calculated through bootstrap method (see Methods section)

**Table 3.3.** Accuracy of the scoring functions on the complete Set 1 and on two reduced sets compared to previously reported data.<sup>a</sup>

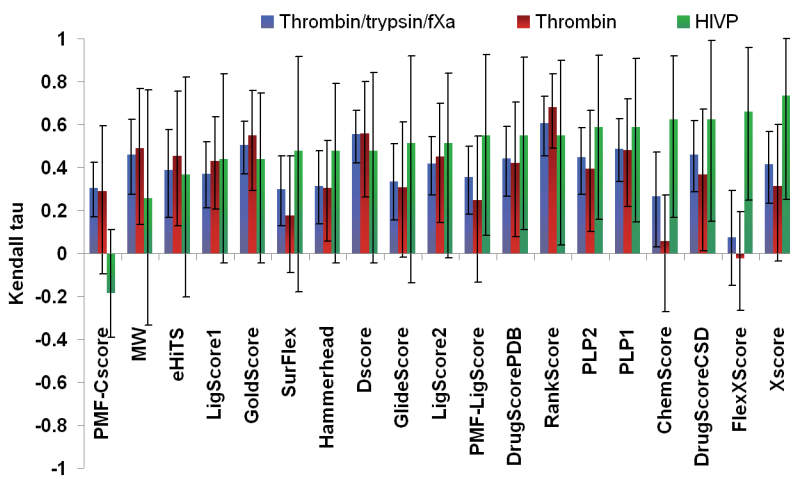
	Entire set (209 complexes)			5 outliers removed (204 complexes)			Metalloenzymes removed (188 complexes)			From ref. <sup>12</sup> and <sup>13</sup>
	r <sup>2</sup>	ρ	τ	r <sup>2</sup>	ρ	τ	r <sup>2</sup>	ρ	τ	ρ
ChemScore (Sybyl)	<b>0.197</b>	<b>0.469</b>	<b>0.324</b>	<b>0.276</b>	<b>0.547</b>	<b>0.374</b>	<b>0.224</b>	<b>0.510</b>	0.351	0.431 / 0.432
DockScore (Sybyl)	0.098	0.342	<b>0.231</b>	<b>0.166</b>	0.421	<b>0.281</b>	<b>0.123</b>	<b>0.389</b>	0.263	0.475 / 0.476
DrugScore <sup>CSD</sup>	<b>0.176</b>	<b>0.427</b>	<b>0.292</b>	<b>0.241</b>	<b>0.498</b>	<b>0.337</b>	<b>0.251</b>	<b>0.472</b>	<b>0.320</b>	<b>0.624</b>
DrugScore <sup>PDB</sup>	<b>0.121</b>	<b>0.383</b>	<b>0.264</b>	<b>0.169</b>	<b>0.455</b>	<b>0.309</b>	<b>0.148</b>	<b>0.431</b>	<b>0.296</b>	<b>0.587 / 0.589</b>
eHiTS SF	0.208	<b>0.421</b>	<b>0.296</b>	<b>0.228</b>	<b>0.513</b>	<b>0.353</b>	<b>0.228</b>	<b>0.446</b>	0.315	-
FlexXScore	0.038	0.195	0.133	0.094	0.272	0.183	0.046	0.216	0.149	0.283 / 0.287
GlideScore	<b>0.114</b>	<b>0.378</b>	<b>0.26</b>	<b>0.176</b>	<b>0.454</b>	<b>0.309</b>	<b>0.365</b>	<b>0.423</b>	0.292	-
GoldScore (Sybyl)	<b>0.169</b>	<b>0.434</b>	<b>0.295</b>	<b>0.235</b>	<b>0.507</b>	<b>0.342</b>	<b>0.205</b>	<b>0.487</b>	<b>0.330</b>	<b>0.569 / 0.570</b>
Hammerhead (Cerius2)	<b>0.115</b>	0.345	<b>0.237</b>	<b>0.166</b>	0.415	<b>0.282</b>	<b>0.144</b>	<b>0.383</b>	0.262	-
LigScore1 (Cerius2)	0.011	0.112	0.074	0.030	0.176	0.116	0.017	0.141	0.094	-
LigScore2 (Cerius2)	0.096	0.332	<b>0.225</b>	<b>0.141</b>	0.401	<b>0.269</b>	<b>0.122</b>	<b>0.367</b>	0.248	0.363 / 0.368
PLP1 (Cerius2)	<b>0.139</b>	<b>0.387</b>	<b>0.262</b>	<b>0.190</b>	<b>0.453</b>	<b>0.306</b>	<b>0.173</b>	<b>0.434</b>	<b>0.292</b>	<b>0.592 / 0.593</b>
PLP2 (Cerius2)	<b>0.116</b>	<b>0.364</b>	<b>0.248</b>	<b>0.185</b>	<b>0.443</b>	<b>0.299</b>	<b>0.122</b>	<b>0.400</b>	<b>0.273</b>	-
PMF (Sybyl)	0.000	0.012	0.011	0.011	0.073	0.051	0.000	0.023	0.021	-
PMF (Cerius2)	0.050	0.212	0.141	0.093	0.284	0.186	0.054	0.236	0.156	0.369 / 0.370
RankScore	<b>0.148</b>	<b>0.418</b>	<b>0.298</b>	<b>0.216</b>	<b>0.503</b>	<b>0.353</b>	<b>0.177</b>	<b>0.458</b>	0.328	-
Surflex SF	<b>0.143</b>	<b>0.409</b>	<b>0.274</b>	<b>0.161</b>	<b>0.476</b>	<b>0.317</b>	<b>0.167</b>	<b>0.448</b>	0.301	-
XScore	<b>0.239</b>	<b>0.526</b>	<b>0.365</b>	<b>0.320</b>	<b>0.605</b>	<b>0.416</b>	<b>0.259</b>	<b>0.566</b>	<b>0.392</b>	<b>0.660 / 0.660</b>
MW	0.109	0.349	0.230	0.117	0.422	0.277	0.110	0.350	0.229	0.560

<sup>a</sup> AutoDock scoring function has not been included in the present study. In bold if better than MW.

### 3.3.2 Accuracy of the selected scoring functions within protein classes

A scoring function is of interest for VS applications (e.g., hit compound discovery) only if it can discriminate between active and inactive compounds in a given library against a particular target. In structure-activity relationship (lead optimization), a scoring function should rank compounds with subtle structural changes. To evaluate this ability, one has to investigate families of protein/ligand complexes. We used the collected data and extracted the scores obtained with serine proteases (thrombin, trypsin and factor Xa, 42 complexes), HIV-1 protease (11 complexes) and thrombin alone (22 complexes) (Figure 3.5). Although XScore was found to be the most accurate on the entire set, RankScore, GoldScore and DockScore were found to be the most predictive with the selected serine proteases inhibitors and the only 3 scoring functions that are likely more predictive than the MW descriptor. Similarly, ChemScore demonstrates very good accuracy with HIV-1 protease with  $\tau = 0.62$ , while it is not predictive at all with thrombin ( $\tau < 0.1$ ) or the serine proteases ( $\tau = 0.27$ ). It is worth noting that affinities of these subset ligands are more MW-dependent than the complete set and that only three of the assessed functions (RankScore, GoldScore and DockScore) were consistently more accurate than MW (although within the computed errors) used as a descriptor with these families of proteins. In contrast to Wang's report, we found XScore and ChemScore reliable with HIV-1 protease. In fact, many scoring functions were at least marginally predictive against this protein class, while MW was likely a chance correlation. We relate this better accuracy to the care taken to prepare the systems: the protonation state of this aspartic protease was carefully assigned to each of the complexes and the essential water molecule was kept when necessary (see Experimental Section). These two features ensured that the scores reflect the binding energies of the complexes. Interestingly, FlexXScore was found among the most accurate with HIV-1 protease and the least accurate with the thrombin and serine protease ligands, further demonstrating the poor transferability of some scoring functions and the need for a broad set when carrying a comparative study or developing a scoring function or for protein-specific scoring functions. When we attempted to expand this

analysis to other families of proteins, we found that their representation in our set prevented us from making statistically significant predictions (that is, discarding the possibility of a chance correlation).

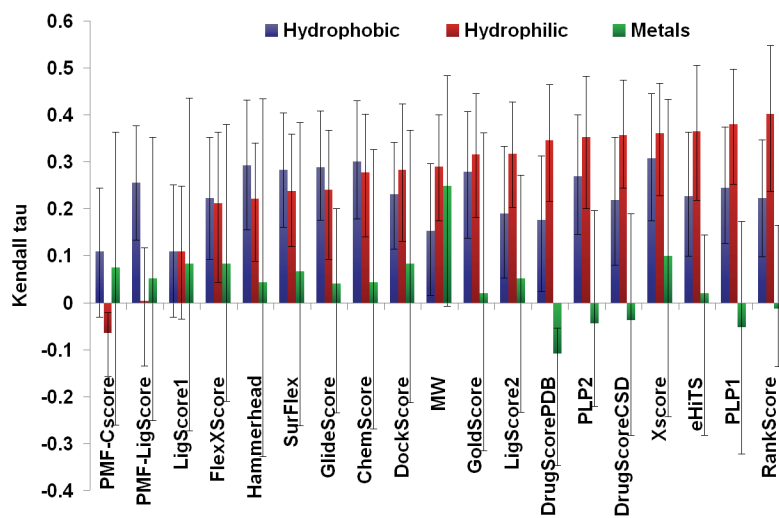


**Figure 3.5.** Accuracy ( $\tau$ ) of the scoring functions on 3 subsets of set 1.

### 3.3.3 Hydropathicity and accuracy

When a computational medicinal chemist starts a docking study, a recurrent question is always about the selection of the docking/scoring program best suited for the ongoing study. To partially answer this question, we previously investigated the accuracy of docking programs as a function of the hydropathicity of the proteins,<sup>11</sup> while herein we looked at the accuracy of scoring functions. The hydropathicity of the binding sites was evaluated by considering all the residues within 6.0 Å of the ligands on a combination of the Hopp-Woods, Kyle-Doolittle and Grantham scales of hydrophobicity.<sup>34-36</sup> From the 209 initial complexes, the 85 most hydrophobic and the 93 most hydrophilic were selected (Figure 3.6). Three classes of scoring functions emerged from this study. First, XScore performed well with both hydrophobic and hydrophilic protein classes. Second, RankScore, DrugScore, PLP1 and the eHiTS scoring functions were found to be more predictive with hydrophilic proteins. This is an indication for further improvement of FITTED, as we have also found that FITTED docks small molecules very accurately to hydrophilic proteins but performed poorly with hydrophobic proteins.<sup>11</sup> These observations are

most likely related to the prediction of hydrophobic interactions on one side and hydrogen bonds and ionic interactions on the other side, the latter being easier to identify and quantify. At the other end of the spectrum, GlideScore is more accurate with hydrophobic proteins. Nevertheless, as these subsets demonstrate different dependence between MW and affinities, drawing conclusions might require considering larger sets. Finally, the data collected for the 23 metalloenzymes does not allow one to rule out a chance correlation for any of the scoring functions, except for DrugScore<sup>PDB</sup> that exhibited an unexpected modest anti-correlation. These observations clearly confirmed that current scoring functions are not reliable when metal chelation is the key interaction, as none of the scoring functions considered were more predictive than the ligand molecular weight.



**Figure 3.6.** Accuracy (Kendall  $\tau$  correlation) of the scoring functions on three subsets of Set 1, classified by the hydrophobicity of the binding sites. Error bars calculated with the bootstrap method.

### 3.3.4 Scoring function correlation

As discussed above, the accuracy of scoring functions varies significantly from one protein class to another and is very dependent on the testing set used (as shown by the error bars). A closer look at the data revealed that some of the scoring functions varied following the same patterns. In order to further investigate this trend, we computed the correlations between scoring functions regardless of the protein target and binding affinities (Figure 3.7). Some of the scoring functions are highly



correlated, with values of  $\tau$  computed for each pair often over 0.60, with a maximum of 0.75 between XScore and ChemScore. In these cases, they are more correlated to each other than to the observed binding energies, as the best coefficient obtained between scoring functions and observed binding affinities was 0.37 (with XScore). In fact, the four scoring functions previously identified as the most accurate in Table 3.3 (XScore, GoldScore, ChemScore and DrugScore<sup>CSD</sup>) are all highly correlated ( $\tau$  ranging from 0.50 to 0.75). Our scoring function, RankScore, correlates with 8 scoring functions with  $\tau$  greater than 0.50, while GoldScore correlates with 11 scoring functions at the same threshold of  $\tau$ . On the other side, PMF and FlexXScore, which were found to be poorly accurate, are not highly correlated with the other scoring functions assessed. It is striking that scoring functions derived by different groups in very different manners exhibit this high level of correlation. It is worth recalling that DrugScore is a knowledge-based scoring function, while XScore is an empirical/consensus scoring function and ChemScore is empirical. In addition, they are made up of terms accounting for different properties (e.g., no entropy term in DrugScore). Interestingly, GlideScore has been originally derived from ChemScore but does not show a high degree of correlation with it.

	ChenScore	DrugScoreCSD	DrugScorePDB	DockScore	eHTS	FlexXScore	GlideScore	GoldScore	Hammerhead	LigScore1	LigScore2	PLP1	PLP2	PMF-Cerius2	PMF-Sybyl	RankScore	Surflex	XScore
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00	0.57	0.53	0.36	0.45	0.26	0.36	0.51	0.44	0.10	0.30	0.44	0.43	0.21	0.00	0.39	0.45	0.75
2	0.57	1.00	0.74	0.46	0.42	0.19	0.37	0.60	0.45	0.28	0.51	0.60	0.56	0.22	0.01	0.56	0.44	0.62
3	0.53	0.74	1.00	0.48	0.46	0.26	0.40	0.61	0.49	0.32	0.52	0.74	0.67	0.12	0.05	0.62	0.41	0.63
4	0.36	0.46	0.48	1.00	0.41	0.37	0.44	0.60	0.44	0.42	0.57	0.52	0.57	0.31	0.23	0.56	0.40	0.41
5	0.45	0.42	0.46	0.41	1.00	0.37	0.42	0.46	0.37	0.22	0.38	0.43	0.44	0.15	0.04	0.46	0.38	0.48
6	0.26	0.19	0.26	0.37	0.37	1.00	0.37	0.28	0.43	0.34	0.33	0.34	0.37	0.09	0.16	0.35	0.33	0.19
7	0.36	0.37	0.40	0.44	0.42	0.37	1.00	0.44	0.41	0.28	0.38	0.44	0.45	0.19	0.11	0.51	0.55	0.38
8	0.51	0.60	0.61	0.60	0.46	0.28	0.44	1.00	0.57	0.38	0.56	0.60	0.57	0.31	0.12	0.59	0.52	0.58
9	0.44	0.45	0.49	0.44	0.37	0.43	0.41	0.57	1.00	0.43	0.46	0.49	0.50	0.23	0.11	0.45	0.60	0.43
10	0.10	0.28	0.32	0.42	0.22	0.34	0.28	0.38	0.43	1.00	0.61	0.37	0.39	0.20	0.20	0.47	0.34	0.12
11	0.30	0.51	0.52	0.57	0.38	0.33	0.38	0.56	0.46	0.61	1.00	0.54	0.55	0.31	0.19	0.60	0.41	0.36
12	0.44	0.60	0.74	0.52	0.43	0.34	0.44	0.60	0.49	0.37	0.54	1.00	0.84	0.14	0.02	0.67	0.41	0.53
13	0.43	0.58	0.67	0.57	0.44	0.37	0.45	0.57	0.50	0.39	0.55	0.84	1.00	0.18	0.06	0.64	0.40	0.49
14	0.21	0.22	0.12	0.31	0.15	0.09	0.19	0.31	0.23	0.20	0.31	0.14	0.18	1.00	0.55	0.14	0.26	0.22
15	0.00	0.01	0.05	0.23	0.04	0.16	0.11	0.12	0.11	0.20	0.19	0.02	0.06	0.55	1.00	0.06	0.16	0.05
16	0.39	0.56	0.62	0.56	0.46	0.35	0.51	0.59	0.45	0.47	0.60	0.67	0.64	0.14	0.06	1.00	0.47	0.45
17	0.45	0.44	0.41	0.40	0.38	0.33	0.55	0.52	0.60	0.34	0.41	0.41	0.40	0.26	0.16	0.47	1.00	0.44
18	0.75	0.62	0.63	0.41	0.48	0.19	0.38	0.58	0.43	0.12	0.36	0.53	0.49	0.22	0.05	0.45	0.44	1.00
Average	0.38	0.45	0.47	0.44	0.37	0.30	0.38	0.49	0.43	0.32	0.45	0.48	0.48	0.23	0.13	0.47	0.41	0.42
Minimum	0.00	0.01	0.05	0.23	0.04	0.09	0.11	0.12	0.11	0.10	0.19	0.02	0.06	0.09	0.00	0.06	0.16	0.05

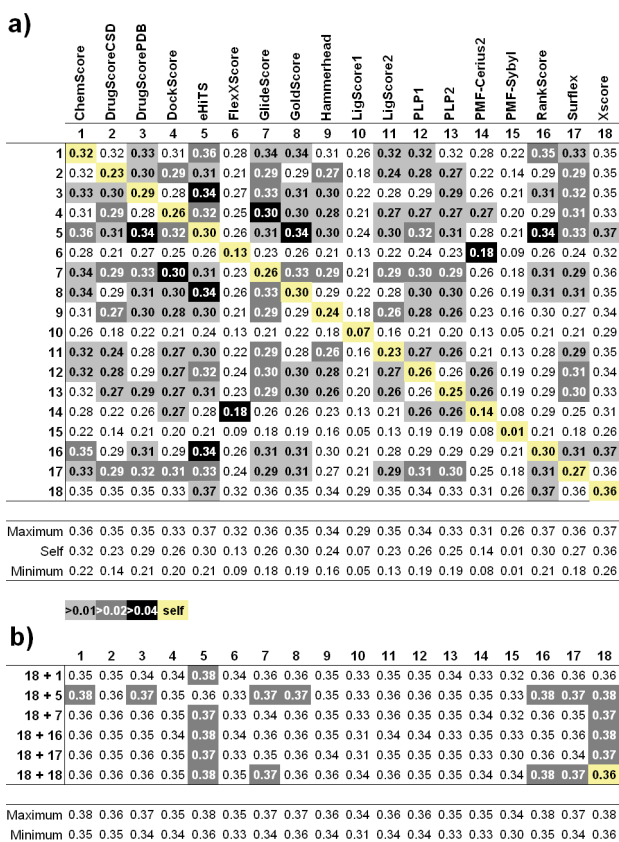
>0.30>0.40>0.50>0.60

**Figure 3.7.** Ranked-list correlation coefficients ( $\tau$ ) calculated between predicted ranking lists of SFs; the darker the shading, the higher the correlation (see key). The numbers 1-18 represent the scoring functions as specified in the top row.

### 3.3.5 Consensus scoring

A scoring function aims to predict the binding affinities of ligands for proteins and/or to compute the free energy of binding. The results from the previous section demonstrated that the best performing scoring functions capture the same information, showing high correlation between their scores (often with  $\tau > 0.60$ ), while the moderate correlation ( $0.20 < \tau < 0.35$ ) between these functions and the observed binding affinities also indicates that these functions may disregard the same aspects of the binding process. From these conclusions, we hypothesized that consensus scoring may not lead to significantly better accuracy and that only a very few scoring functions can be considered in this context. In order to test this hypothesis, the accuracy for each pair of scoring functions was assessed. In order to normalize the data, we combined the ranks computed with each scoring function and not the scores. As illustrated in Figure 3.8, most of the combinations (shown in white) led to  $\tau$  coefficients that are not better than each of the  $\tau$  coefficients computed for each scoring function. More interestingly, there was no case where the predictiveness of a pair of functions was lower than both individual scoring

functions. In all cases, combinations that led to the best  $\tau$  values included either XScore, the eHiTS scoring function, RankScore or ChemScore, which were already found to be among the four most predictive scoring functions with  $\tau$  greater than 0.30. This data validates our hypothesis and demonstrates that consensus scoring using a combination of traditional scoring functions can at best provide a moderate increase in accuracy and that consensus scoring should be developed from more different scoring approaches<sup>37</sup> or include additional information.<sup>38</sup>



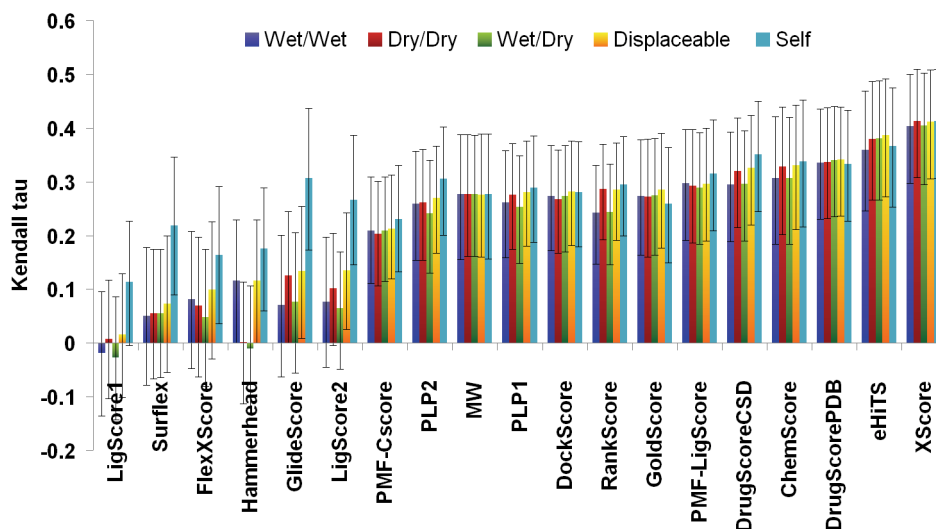
**Figure 3.8.**  $\tau$  calculated for combinations of scoring functions. The numbers 1-18 on the second row and leftmost columns represent the scoring functions as described in the top row of Figure a). (a) Pairs of scoring functions: darker boxes represent increase of the combined scoring functions over the individual scoring functions (see key; value in key indicates the increase of the value of  $\tau$  over the value for the individual scoring function in the same row). The yellow boxes correspond to the individual scoring functions. (b) Groups of three scoring functions: darker boxes represent increase with respect to XScore alone (yellow box)

### 3.3.6 Impact of protein conformation and water molecules

Cross-docking is a more appropriate experiment than self-docking when one wants to mimic virtual screening studies. A set of cross-docked ligands (i.e., Set 2) was therefore assembled and used to assess scoring functions in this context. With this second set, we aim to assess the impact of the selected protein conformation and presence/absence of water molecules on scoring function accuracy. Each cross-docked complex was first optimized in presence of water molecules and then assigned two scores: one corresponding to the scoring considering water molecules (wet/wet) and one corresponding to the scoring with no waters included (wet/dry). As a third subset, we performed the local optimization of the ligands without the water molecules and scored the resulting complexes without any water molecules (dry/dry). This resulted in three subsets: one with the key water molecules retained for both docking and scoring, one with all the key water molecules retained for docking but removed for scoring, and one with all the water molecules removed for both docking and scoring. As each of the nearly 1000 complexes can be found in the three subsets, each complex was assigned three scores. With all this data in hand, one can simulate the displacement of the water molecule, by selecting the best score out of the three for each complex (water displaceable).

At this stage, each of the 92 ligands had been scored with all the structures (native and non-native) of the same protein. In order to evaluate the impact of the selection of the protein conformation on the scoring accuracy, scripts were written to evaluate the correlation coefficients with various random selections of cross-docked complexes. For this purpose, a protein structure (e.g., HIV-1 protease1a30) out of the non-native protein structures of the same protein was randomly selected for each ligand (e.g., HIV-1 protease ligand 1b6l) and the resulting complex scored. From this set of predictions, the ranking of ligands by predicted binding affinity was computed, compared to the observed ranking and a value of  $\tau$  was generated. Then the process was reiterated with a different random selection (e.g., HIV-1 protease ligand 1b6l alternatively cross-docked to all the HIV-1 protease conformations but the 1b6l native protein conformation). This process was iterated 10,000 times with

a different population of cross-docked structures each time, thus providing a range of values for the correlation coefficient  $\tau$  (see Figure 3.8). The median values for the  $\tau$  obtained under the different conditions are given in Table A.5 (see Appendix) and graphed in Figure 3.9.

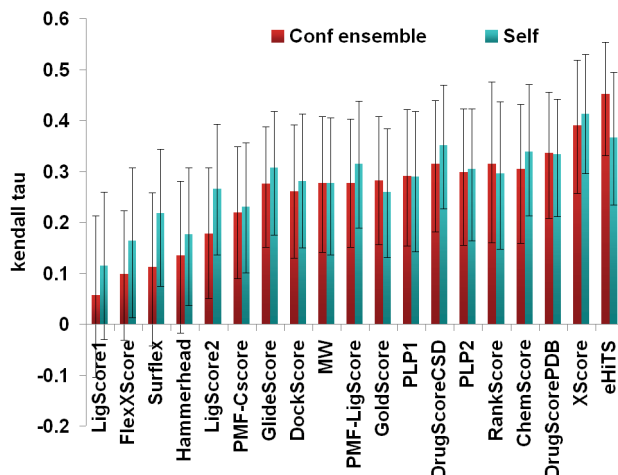


**Figure 3.9.** Accuracy ( $\tau$ ) of the scoring functions on set 2 with different water considerations. Waters were kept for docking and scoring (wet/wet, blue), kept for docking and removed for scoring (dry/dry, red), removed for docking and scoring (dry/dry, green) or made displaceable (yellow). For comparison, the correlation of the scores obtained for the native structures in wet/wet conditions is shown (light blue).

Although the presence of water molecules bridging the intermolecular interaction between ligands and biomacromolecules is known to be critical for binding in some cases, docking and scoring programs do not commonly handle water molecules. Two aspects can be affected by water molecules: the binding mode can be optimized differently whether the waters are kept or not, and the score of the same pose can be different if the scoring function scores the water-mediated interactions. As can be seen in Figure 3.8 (and Table A.5 in Appendix), XScore, DrugScore, and the eHiTS scoring function demonstrated the best correlations when the waters are kept, although within errors from more than half of the other scoring functions. When comparing the Kendall coefficients for the different water treatments, it is clear that none of the scoring function is significantly affected by the presence or absence of water molecules. It is worth mentioning that RankScore has been optimized to account for the presence of water molecules and demonstrates slightly enhanced

accuracy when displaceable water molecules are used. Overall, making the water molecules displaceable increases the accuracy of most of the scoring functions as expected although by only a small increment.

Next we looked at the drop in accuracy when going from native protein structures to cross-docked complexes. It clearly appears (Figure 3.10, see also Table A.6 in Appendix) that some of the scoring functions including GlideScore and RankScore are greatly affected by the protein structures considered. The selection of the protein conformation when more than one crystal structure is available should be done with care. Once more, XScore and the eHiTS scoring function were the most accurate, likely an outcome of the larger training sets used in the parameterization of these scoring functions. Alternatively, each ligand was docked to all the non-native protein conformations available for each protein and the best score was retained. The selected score corresponds to a docking experiment carried out on a conformational ensemble of protein structures, therefore considering the protein flexibility (conformational ensemble, see Figure 3.10). By moving to conformational ensembles, we expected to restore part of the accuracy lost when moving from self- to cross-docked structures. In fact, while most of the scoring functions are negatively affected by the use of conformational ensembles, the accuracy of the eHiTS scoring function significantly increases when conformational ensembles are used in place of the native protein conformation. A closer look at the data did not allow any explanation for this behavior. We believe that scoring functions based on soft proteins are less sensitive to the protein conformation. In fact, the steep Lennard-Jones 12-6 employed by some of the scoring functions (e.g., RankScore) is very sensitive to subtle moves while the soft function used by eHiTS scoring function is not (Figure 3.10).



**Figure 3.10.** Accuracy of the scoring functions ( $\tau$ ) on the complete set 2 when protein conformational ensembles (*best scored complex among the cross-docked ones*) were considered.

### 3.4 Conclusions

We have carefully developed two sets of protein/ligand complexes with reduced correlation between MW and binding affinities. The first set was scanned for accuracy in scoring of native poses (analogous to self-docking) using a large number of available scoring functions. This screening revealed the good accuracy of XScore, the eHiTS scoring function, DrugScore (PDB- and CSD-derived), GoldScore and ChemScore. Analysis of the results on subsets of complexes indicated a large dependence on the protein, analogously to the one observed previously with docking programs.<sup>11</sup> This can be in part explained by the training sets used in the derivation of these scoring functions: eHiTS and XScore are the empirical scoring functions with the largest training sets considered in this study. In particular, the eHiTS scoring function implements specific parameters based on the protein class involved, which might (at least in part) account for this enhanced accuracy. In a subsequent section, we demonstrated that consensus scoring can only lead to moderate increase in accuracy and that other strategies should be proposed (i.e., smart post-processing of poses) or novel (i.e., more predictive) scoring functions should be developed to address the scoring issue in docking methods. Finally, we have shown that some scoring functions lose all predictive power when applied to cross-docked structures, demonstrating the need to incorporate protein flexibility in

docking programs. From this information we draw the conclusion that using softer proteins or conformational ensembles should lead to more predictive scoring functions for use in docking-based virtual screening. More surprisingly, the consideration of ligand-water molecule interactions has been shown to be of little importance for scoring. However, most of these scoring functions were trained on “dry” proteins and often do not consider the water molecules even if present. Our scoring function RankScore performed well in the self-docking experiment and when using conformational ensembles (as does our docking program FITTED), but poorly with cross-docked structures. In addition, although we have found that considering displaceable water molecules improves the docking accuracy, the present work showed that it affects the scoring. In subsequent work, we will develop a more accurate scoring function for these situations.

### ***3.5 Experimental section***

Scripts were required in order to automate many repetitive tasks in each of the interfaces; to this effect Python scripts were used in Maestro, SPL scripts in Sybyl and BCL scripts in InsightII. Awk, shell and Python scripts were written to pre- and post-process the structures, input and output files from each of the scoring functions as necessary. Calculations were run on SGI Fuel workstations with a single R16000 processor and Linux workstations (AMD Opteron and/or Intel Core2 processors).

#### **3.5.1 Selection of the training set structures**

##### ***3.5.1.1 PDB***

Queries on the Protein Data Bank<sup>39</sup> were performed looking for X-ray crystal structures of either HIV-1 protease (HIVP), thrombin or matrix metalloproteases (MMPs) in complexes with small molecule ligands. The structures found were filtered by keeping the ones with resolution better than 2.5 Å, and among those, the ones containing non-covalent ligands with reported activity. A set of 20 complexes for each protein was selected ensuring chemical diversity and a homogeneous representation of 7 orders of magnitude of  $K_i$ . For HIVP, the 20 complexes included



11 wild-type structures and 9 mutants, with all the point mutations located away from the first layer of residues in contact with the ligand. For thrombin, the 20 structures correspond to wild-type human thrombin; the MMP training set is composed of 6 structures of MMP-3 (stromelysin-1), 2 structures of MMP-7 (matrilysin), 9 structures of MMP-8 (collagenase-2) and 3 structures of MMP-13 (collagenase-3).

### **3.5.1.2 PDBBind**

Two hundred and twenty complexes from the refined PDBBind database<sup>18,19</sup> were selected for chemical diversity and even distribution of affinity spanning 9 orders of magnitude (from pKd=3 to 12). Naturally, the affinity ranges between pKd=5 to 10 (the most common affinity of a good lead) are the better represented. The compounds selected have a lead-like molecular weight of between 250 and 600, and the complexes for which there was more than one and different affinities reported within the PDBBind database were not selected.

## **3.5.2 Preparation of the training set**

### **3.5.2.1 Generalities**

Preparation of the complexes for further calculations was performed with Maestro 7.0 (Schrödinger, Inc.) and MacroModel 9.0. (Schrödinger, Inc.). Succinctly, it involved completion of the side-chains missing from the PDB structure (exposed to solvent); capping of the protein termini as either ammoniums or carboxylates; assignment of bond orders and protonation states in the ligand and active site residues; and removal of all extraneous molecules (e.g., ions far away from the ligand binding site, ethyleneglycol). In the cases where more than one pose for the ligand was present in the PDB file, the one with the highest occupancy was chosen; otherwise the first pose described was used. Water molecules were treated as described in the following section. Hydrogen atoms were added and minimized with all other atoms fixed (MMFFs94, up to 500 steps of conjugate gradient). The binding mode of the ligands was relaxed by an energy minimization in which only ligand atoms and hydrogens bound to heteroatoms were allowed to move (MMFF94s, up to

2000 steps of conjugate gradient), and heavy atoms were constrained by a harmonic potential ( $k = 5 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ ) to their crystal structure positions.

#### **3.5.2.2 Treatment of water molecules**

All explicit water molecules were removed from the complexes, except for the ones that were contained in the intersection of a volume  $3.0 \text{ \AA}$  around all atoms of the ligand and a volume  $3.0 \text{ \AA}$  around all the atoms of the receptor. These remaining water molecules (varying in number between 0 and 20) were examined more closely, and the ones capable of forming at least 3 hydrogen bonds with both the ligand and the receptor were kept.

#### **3.5.2.3 Assignment of protonation states for aspartyl proteases**

The two catalytic aspartyl residues in HIV-1 protease (Asp25) are considered to exist in a monoprotonated (monoionized) state for catalytic activity.<sup>20</sup> The exception has to be made for ligands binding to the catalytic dyad by means of a 1,2-diol, where NMR and X-ray data points to both Asp25 being protonated, formed a tight hydrogen bonding network with the two hydroxyls in the ligand,<sup>21</sup> as well as ligands with an ammonium group facing the catalytic dyad, which might stabilize the deprotonated state.<sup>22</sup> A careful observation of the environment around the Asp25, defining how the hydrogen bond network could be formed, together with the coplanar (or not) orientation of the Asp25 carboxylates led us to define the protonation state of each of the complexes.

#### **3.5.2.4 Thrombin**

Crystal structures for thrombin usually contain 3 chains: the 2 chains (L and H) resulting from the self-cleavage of the protein, plus a hirugen peptide, which binds to an alternate binding site in the protein. Given that the hirugen peptide and the low-molecular weight chain of thrombin lie far away (more than  $15 \text{ \AA}$ ) from any atoms of the ligand, they were removed in all cases, and only the heavy chain was used for the calculations. The protonation state of all residues in the protein was assigned as expected at pH 7, except for the artificial terminal groups resulting from

the missing loop in the crystal structures between residues 146 and 149E, which were considered neutral (COOH and NH<sub>2</sub>) for the sake of not adding artificial charges in the binding site. All ligands contain at least one protonated basic moiety (ammonium, amidinium, guanidinium), which interacts with Asp45 in the receptor.

#### ***3.5.2.5 Metalloenzymes***

Zinc-containing proteins (e.g., MMP-3, MMP-8, carbonic anhydrase) were prepared by breaking all bonds between the metal ion and heteroatoms in ligands and protein, and specifying a formal charge of +2.

#### ***3.5.2.6 PDBBind complexes***

The complexes retrieved from the PDBbind database were prepared in a way analogous to the previous complexes. Ligand protonation states were checked and corrected where applicable (e.g., in some cases nitrogens attached to aryl groups were incorrectly protonated); aspartyl proteases (e.g., penicillopepsin, endothiapepsin, SIV and HIV-1 protease) were identified as such and the catalytic dyad treated as described above; metalloenzymes' Zn atoms were treated as in MMPs. In the case of multimeric complexes, the minimum number of chains necessary to describe the complex was kept. The number of atoms for the proteins was kept under 10,000 by removing residues far away (i.e., > 20 Å) from the ligand binding site if necessary.

#### ***3.5.2.7 Preparation of cross-scoring set (set 2)***

Proteins represented with at least 5 complexes in set 1 were selected (with the exception of PNP, for which not all proteins were from the same source species). Crystal structures of the complexes were prepared as described above, and all complexes from the same protein superimposed to the alpha carbon trace of one of them. FITTED was used in local search mode to adjust the binding mode of the ligands to each protein binding site separately, with each protein being treated rigidly. Due to the presence of flexible side chains, the maximum allowed translation

in the generation of the initial population was set as 5Å. The resulting binding modes were used as input for all scoring methods.

### **3.5.3 Scoring**

#### **3.5.3.1 CScore**

The stand-alone CScore module from Sybyl v7.3 was used, with default parameters for the DScore (DockScore), GScore (GoldScore), PMF and ChemScore scoring functions.

#### **3.5.3.2 GlideScore**

Glide v4.5 was used in all calculations. Grids were generated in a box of 20Å around the ligand, with default parameters. Scoring was performed in place.

#### **3.5.3.3 eHiTS**

The score.sh script supplied with eHiTS v6.2 was used. Two ligands, 1h22 and 1h23, were not assigned a score by eHiTS for having more than 10 rotatable bonds in a linear fragment; a few others failed to be optimized, hence the non-optimized score was considered.

#### **3.5.3.4 Surflex**

Surflex v2.301 was used. Protomol files for the protein structures were first generated with the “proto” option, and then scores were calculated with the “score\_list” option. The non-optimized score was considered, as the optimized one gave poorer correlations.

#### **3.5.3.5 Cerius2**

Cerius2 v4.10 was used on all calculations. PDB-formatted protein structure files and SD-formatted ligand files were used as input for the LigScore1/2, PLP1/2, PMF and Jain (Hammerhead) scoring functions.

#### **3.5.3.6 XScore**

X-Tool v.1.2.1 was used. Protein and ligand structures were first prepared with the “-fixpdb” and “-fixmol2” options respectively, prior to running the score computation with the “-score” option.

#### **3.5.3.7 DrugScore**

Executables of DrugScore<sup>PDB</sup> and DrugScore<sup>CSD</sup> v1.2 were used under IRIX.

#### **3.5.3.8 RankScore**

The scores were calculated with the FITTED 2.6 docking program, after pre-processing the protein and ligand structures with PROCESS and SMART, respectively.

### **3.5.4 Bootstrap analysis**

The scores for each protein/ligand complex with every scoring function, as well as the molecular weight and the experimental binding affinities were organized in a CSV file and processed with a Python script. A random subset of observations of the same size as the original (repetition was allowed) was selected, and the correlation coefficients for each scoring function were calculated using the functions provided in the SciPy module; this process was iterated 10,000 times. For each scoring function, the range of correlation coefficients spanning 95% of the obtained in the previous trials was taken as the uncertainty in the correlation coefficient for the original set. In the case of the cross-docked structures, a random protein was selected for each ligand on each iteration and the correlation calculated with one complex for each ligand.

## **3.6 References**

1. Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. In *Hits, Leads and Artifacts from Virtual and High Throughput Screening*, Molecular Informatics: Confronting Complexity, Bozen, Italy, May 13-16, 2002; Hicks, M. G.; Kettner, C., Eds. Beilstein-Institut: Bozen, Italy, 2002.

2. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7-S26.
3. Rester, U., Dock around the clock - Current status of small molecule docking and scoring. *QSAR Comb. Sci.* **2006**, *25*, 605-615.
4. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 15-26.
5. Jain, A. N., Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407-420.
6. Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G., Enhancing drug discovery through in silico screening: Strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008**, *15*, 2040-2053.
7. Aqvist, J.; Medina, C.; Samuelsson, J. E., A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385-391.
8. Pham, T. A.; Jain, A. N., Customizing scoring functions for docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 269-286.
9. 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.
10. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of FITTED 1.5 to the Virtual Screening of Potential HCV Polymerase Inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 902-909.
11. Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997-1009.
12. Wang, R.; Lu, Y.; Wang, S., Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287-2303.

13. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore<sup>CSD</sup>-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296-6303.
14. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W., Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726-741.
15. Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W., Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214-2225.
16. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11-26.
17. Jain, A. N.; Nicholls, A., Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133-139.
18. 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.
19. Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S., The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111-4119.
20. Kulkarni, S. S.; Kulkarni, V. M., Structure Based Prediction of Binding Affinity of Human Immunodeficiency Virus-1 Protease Inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1128-1140.
21. Yamazaki, T.; Nicholson, L. K.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Eyermann, C. J.; Hodge, C. N.; Lam, P. Y. S.; Torchia, D. A.; et al., NMR and X-ray Evidence That the HIV Protease Catalytic Aspartyl Groups Are Protonated in the Complex Formed by the Protease and a Non-Peptide Cyclic Urea-Based Inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791-10792.
22. Czodrowski, P.; Sottriffer, C. A.; Klebe, G., Atypical Protonation States in the Active Site of HIV-1 Protease: A Computational Study. *J. Chem. Inf. Model.* **2007**, *47*, 1590-1598.

23. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791-804.
24. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T., Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317-324.
25. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M., LigScore: A novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395-407.
26. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425-445.
27. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 609-623.
28. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
29. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P., eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198-212.
30. Jain, A. N., Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281-306.
31. Moitessier, N.; Therrien, E.; Hanessian, S., A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic  $\beta$ -secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885-5894.
32. Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N., Evaluation of Docking Programs for Predicting Binding of Golgi



alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160-176.

33. Wang, R.; Lu, Y.; Fang, X.; Wang, S., An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114-2125.

34. Grantham, R., Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862-864.

35. Hopp, T. P.; Woods, K. R., Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 3824-3828.

36. Kyte, J.; Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105-132.

37. Renner, S.; Derksen, S.; Radestock, S.; Mörchen, F., Maximum common binding modes (MCBM): Consensus docking scoring using multiple ligand information and interaction fingerprints. *J. Chem. Inf. Model.* **2008**, *48*, 319-332.

38. Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H., SeleX-CS: A new consensus scoring algorithm for hit discovery and lead optimization. *J. Chem. Inf. Model.* **2009**, *49*, 623-633.

39. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Bourne, P. E.; Shindyalov, I. N., The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

## Chapter 4: Development of scoring functions for virtual screening from force field data

### 4.1 Introduction

For the last decade or so, docking methods have been widely used in structure-based drug design. However, although many successful studies have been reported, scoring of the docked ligands is still often poorly predictive and highly target dependent.<sup>1</sup> In practice, the prediction of binding affinities in protein-ligand complexes can be achieved with various levels of accuracy and speed. On one side of the spectrum, molecular dynamics (MD) simulations (e.g., free energy perturbation and linear interaction energy (LIE)<sup>2,3</sup> methods) take into account the average interaction energy of a Boltzmann distribution of conformers in explicit or implicit aqueous media. On the other end of the spectrum, scoring functions implemented in docking programs, and/or used in virtual screening studies, deal with a single ligand pose and often *in vacuo*. This certainly yields a different time frame for the calculation but with a concomitant decrease in accuracy. These scoring functions are traditionally classified as being either force-field based, empirical (regression-based) or based on a potential of mean force.

The application of molecular mechanical force fields to the scoring of docked poses (i.e., scoring functions implemented in docking programs) has been limited to the AMBER force field (AutoDock,<sup>4</sup> Dock,<sup>5</sup> FITTED<sup>6</sup>), the Tripos force field (Gold<sup>7</sup>) as well as the Dreiding and CFF force fields (LigandFit<sup>8</sup>). In these cases, the selection of a specific force field was often based on its availability (e.g., Amber parameters are publicly available). To date, force fields have never been assessed in detail to evaluate their ability to reproduce binding free energies of ligands to proteins. In fact, although force fields have been developed to reproduce a number of thermodynamic and kinetic properties of molecules, they have not been specifically developed to predict binding free energies and often overestimate them.

We have recently developed a docking program that accounts for protein flexibility and bridging water.<sup>6,9,10</sup> The scoring function used with this software is a force field-based scoring function reported earlier.<sup>9</sup> Although this scoring function was found to extract new active compounds from large libraries,<sup>10</sup> and rank compounds by activity with accuracy similar to the state-of-the art scoring functions,<sup>11</sup> we believe that the number of false positives can be reduced by the use of a more advanced scoring function. We report herein our efforts toward the development of a new force field-based scoring function starting from an exhaustive comparative study of the ability of popular force fields to predict the binding affinity of ligands to proteins. In a second section, we describe the development of a scoring function derived from crystallographic structures and docked complexes, as well as another scoring function trained from virtual screening data. Finally, the validation of the developed scoring functions on benchmark sets of protein/ligand complexes and sets of active compounds and decoys is described. As with our docking program, we focused on a scoring function that would be suited for flexible proteins including bridging water molecules.

## ***4.2 Results and discussion***

### **4.2.1 RankScore**

Our previous version of RankScore (referred to as RankScore1 in this publication) was derived from a set of docked ligands and crystal structures of BACE-1 inhibitors.<sup>9</sup> In the present study, we envisioned three different approaches. A scoring function as implemented in docking programs can have two major applications. First, scoring functions can be used to rank potential ligands by their predicted binding affinity (i.e., as in a lead optimization problem). Second, they can be used to discriminate active compounds from inactive compounds (i.e., as in a hit identification problem).

Based on these premises, we exploited our previously reported set of 209 protein/ligand crystal structures<sup>11</sup> to derive a scoring function (referred to as RankScore2). This second version of RankScore was, consequently, developed from

active compounds (exhibiting weak to strong binding affinities) and will have potential application in the ranking of actives. However, this scoring function was based on the assumption that the protein conformation is fine-tuned to each of the ligands. To address this issue, we also developed a scoring function, namely RankScore3, from our set of 946 cross-docked ligand/protein complexes. Finally, we will describe the development of a third scoring function (RankScore4) from large sets of active and inactive compounds. This last variant will have application in virtual screening (discrimination between binders and non-binders).

#### **4.2.2 Screened force fields**

When considering the development of a force field-based scoring function, there is no rationale for the selection of one force field over another. The first goal of this research project was therefore to assess various class I and class II molecular mechanics force fields to identify the one(s) that would be better suited to make a quick yet accurate estimation of the binding energy between a ligand and a receptor and to compare their accuracy to that of commercially available scoring functions. We selected 5 force fields implemented in Discover (Accelrys), namely CVFF, CFF91, CFF, ESFF and AMBER84; 6 force fields implemented in MacroModel (Schrödinger), namely OPLS2001, OPLS2003, MMFF94, MMFF94s, MM2\*, MM3\* and AMBER\*; and 3 force fields implemented in Sybyl (Tripos), namely Tripos force field, AMBER99 and AMBER02 (which is the only polarizable force field used in this study). MM3 and MM4 have been developed with more accurate but more computationally expensive terms such as dipole-dipole electrostatic term, stretch-bend cross terms and a Buckingham potential term for the van der Waals interaction energy evaluation. As the use of these force fields in VS would significantly reduce the throughput, they were not evaluated in the present work. Although some of these force fields are overall very similar, their parameterization is inherently different (e.g., experimental data such as microwave and NMR spectroscopy, neutron diffraction for AMBER<sup>12</sup> and high quality –MP2/6-31G\*– ab initio data for MMFF<sup>13</sup>) and the mathematical functions used vary from one force field to another. For instance, the van der Waals interaction energy is often computed using a Lennard-

Jones (LJ) potential, most commonly with 6-12 exponents, but 6-9 (CFF, ESFF) and buffered 7-14 (MMFF) exponents are also used. All the force fields use the Coulomb equation to calculate the electrostatic interaction between point charges centered on the nuclei. Older versions of AMBER (such as the AMBER84 implemented in Discover) and the MMx\* family add an explicit term for hydrogen bonding, in the form of a 10-12 Lennard-Jones potential.

### 4.2.3 Training sets and scoring function

The starting point for our comparative study was the two training sets of protein-ligand structures reported in the preceding manuscript of this series.<sup>11</sup> Efforts were devoted to the development of an unbiased training set, showing little correlation between binding affinities and ligand molecular weights. In this previous report, we also applied 18 commonly used scoring functions to evaluate their accuracy with these testing sets. This study shed light on the poor to moderate accuracy of some of these scoring functions when a challenging testing set is used. It also set the lower limit for the development of an accurate scoring function as XScore and ChemScore were found to be the most accurate, with Kendall  $\tau$  coefficients never exceeding 0.37.<sup>11</sup>

### 4.2.4 General considerations

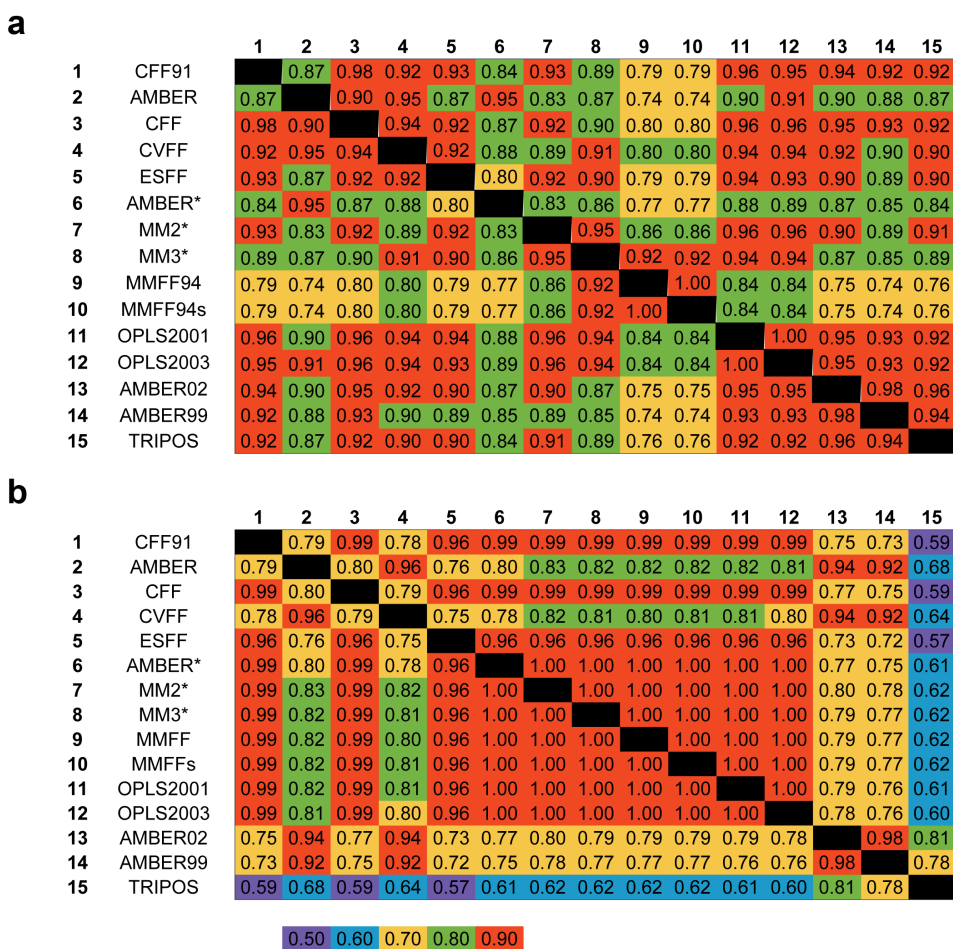
A first set of calculations was performed with the force fields as they were shipped in the corresponding software packages, which in some cases prevented the calculation to proceed because of the lack of parameters. With the addition of parameters to the force field definitions (see *Derivation of additional parameters for force fields*), all systems were run with all the force fields. As expected, preliminary runs with HIV-1 protease inhibitors have shown that the inclusion of the key water molecules (the so-called water 301) was critical for a greater predictive power of the method, and only these results will be discussed. The hydrogen-bond terms arising from the force field energy were taken into account, in the cases where they existed (AMBER84, MM2\*, MM3\*). The first step of the computation was to reconcile the structures with the assessed force fields. These relaxation steps when

performed with each the force fields led to large deviation in some of the cases. In order to address this issue, the ligands were energy-minimized in the binding sites with the ligand heavy atoms constrained to the crystal structure coordinates. As expected, the resulting structures were much closer to the crystal structures and were further processed, with a global average RMSD of 0.32 Å, a standard deviation of 0.20 Å and a maximum RMSD of 1.6 Å for 1okl when minimized with the Amber99 force field. A closer look at the energy-minimized structures indicated that the deviations were only slightly force field dependent, with the Sybyl-implemented force fields exhibiting larger RMSDs than the Discover and MacroModel ones. However, it was seen that the distribution was highly complex-dependent: 81 out of the 209 ligands had RMSDs of less than 0.5 Å with all the force fields, while only 14 ligands displayed RMSD's higher than 1.0 Å with at least one force field. This may indicate either weaknesses of the force fields to reproduce crystal structures, the impact of computation *in vacuo*, inaccuracies in the fitting of the models to the electron density in the crystal structures or the effect of crystal packing. In crystals, protein/ligand complexes are packed and using a single structure may lead to a misinterpretation of the binding mode. For example, we may believe that some of the ligands are exposed to the aqueous medium, (which may exhibit a high ionic strength) while they may interact with a second complex (next cell of the crystal). Observing a large change in the potentially solvent-exposed portions of ligands is therefore expected.

#### **4.2.5 Force field energy terms**

The first step of our study was to compare the various force fields and their implementations for their ability to reproduce binding affinities. We therefore looked at the correlations between force fields van der Waals and coulombic energies as computed with each of the force fields. For this, we calculated the respective components of the binding affinity as the difference between the non-bonded interactions observed in the complex and the ligand and protein, and computed the Pearson correlation for each pair of values. As can be seen in Figure 4.1, the van der Waals and electrostatic terms of all the investigated force fields

considered were highly correlated. The only force field that appeared to have less of a correlation is Tripos. This was not unexpected, as the recommended charge treatment for its use is through formal charges, while all other force fields consider some kind of partial charges. MMFF94 also appeared to produce van der Waals energies that are less correlated to the other force fields. The van der Waals functional form, a buffered 14-7 Lennard Jones, is different from the more traditional Lennard Jones 12-6 or 9-6 used by the other force fields.

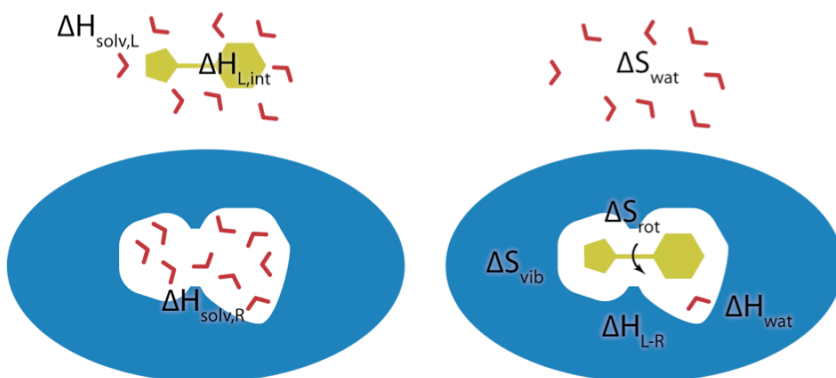


**Figure 4.1** Correlation of van der Waals and electrostatic terms for the different force fields in use.

#### 4.2.6 RankScore2, 3 and 4: novel scoring functions

From that initial study, it became clear that any force field would most probably perform as well when implemented in a more complex scoring function. We therefore turned our attention to the publicly available AMBER/GAFF force field

already implemented in our docking program FITTED 2.6.<sup>6</sup> An ideal scoring function would accurately predict the binding affinity of any given ligand for any given protein or nucleic acid target. In practice, this can be reduced to the prediction of the free energy of binding ( $\Delta G_{\text{binding}}$ ) representing the complex formation illustrated on Figure 4.2. When developing such a function, many researchers relied on the additivity of contributions.<sup>9,14-18</sup> Following this approximation, we broke apart the free energy of binding into the change in entropy and enthalpy measured upon complex formation. As this Michaelis complex forms in water, these contributions are accompanied by a change in solvation (Equation 4.1).



**Figure 4.2.** Ligand/protein complex formation

**Equation 4.1.** Free energy of binding.

$$\Delta G_{\text{binding}} = \Delta G_{\text{complex formation}} + \Delta G_{\text{solvation}}$$

**Equation 4.2.** Decomposition of the free energy complex formation.

$$\begin{aligned} \Delta G_{\text{complex formation}} = & \Delta H_{\text{complex}} - \Delta H_{\text{ligand}} - \Delta H_{\text{protein}} - \\ & -T\Delta S_{\text{complex}} + T\Delta S_{\text{ligand}} + T\Delta S_{\text{protein}} \end{aligned}$$

#### 4.2.7 RankScore formalism

Each of these contributions to the free energy of binding had next to be computed as accurately and as quickly as possible. To do so, we implemented various approaches into FITTED and tried a number of combinations. The first three terms of the right hand side of Equation 4.2 were computed using the GAFF force field (Equation 4.3). In the present study, we assume the protein potential energy (internal energy) to be constant ( $C$  in Equation 4.3) as the energetic aspect of protein conformational



changes would be difficult to evaluate accurately with high throughput. It is clear that  $C$  should be different from one complex to the next as the protein may adjust its conformation upon ligand binding.

**Equation 4.3.** Change of enthalpy of complex for formation of the ligand-protein complex, calculated as the difference of force field energies of the complex, unbound ligand and protein with bound waters ( $E_i^{FF}$ ).

$$\begin{aligned}\Delta H_{\text{complex formation}} &= \Delta H_{\text{complex}} - \Delta H_{\text{ligand}} - \Delta H_{\text{protein}} \\ &= E_{\text{complex}}^{FF} - E_{\text{ligand (unbound)}}^{FF} - E_{\text{protein + water}}^{FF} + C\end{aligned}$$

In practice, these contributions were computed following these steps: i. Optimization through conjugate gradient energy minimization of the ligand pose within the protein binding site using FITTED and computation of the ligand internal energy and intermolecular energy between protein and water molecules, and the ligand. ii. Optimization of the ligand *in vacuo* and computation of the ligand potential energy. As the developed scoring function is to be used with FITTED, we thought that optimizing the pose using the function implemented in FITTED would be more representative of a docked pose.

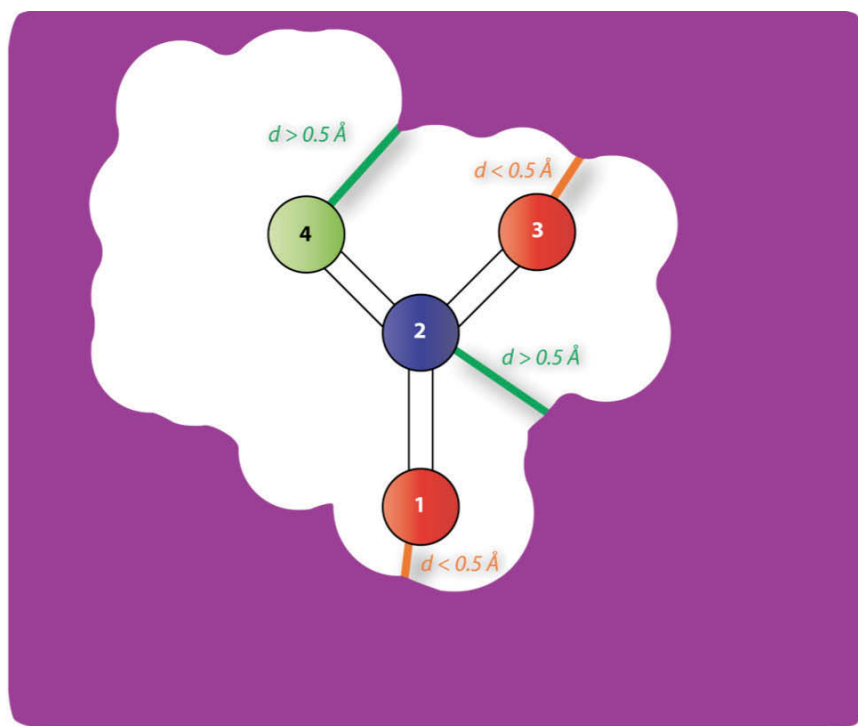
The fourth term of Equation 4.2 was computed by penalizing the number of rotatable bonds ( $N_{\text{rot}}$  in Equation 4.4) that are frozen upon binding, defined as all single bonds not in rings. Many scoring functions give each rotatable bond the same penalty. However, as already mentioned by Eldridge et al. some bonds become “more frozen” than others when a ligand binds.<sup>17</sup> First, it is well known that a long hydrophobic chain does not move as freely as a polar chain in water and is less frozen than hydrogen-bonded groups when bound to the protein. Second, a portion of a ligand in close contact with a protein cannot move as much as a solvent-exposed group. In order to account for these two aspects, we developed a function (Equation 4.4) where the penalty given to the ligand is a function of the number of rotatable bonds, the polarity of the bonds and their deepness in the protein.

**Equation 4.4.** The entropic contribution on the ligand side is calculated as a function of the number of rotatable bonds ( $N_{rot}$ ), affected by the polarity of the bond and the buriedness of the bond, as estimated by the number of contacts with the protein (see text).

$$\Delta S_{lig} = f(N_{rot}, polarity, contact)$$

Two strategies accounting for bond polarity were evaluated. Either the atom types of the rotatable bonds (identified as “*polar*”, “*semipolar*” or “*apolar*”) or the solvation energy of the entire ligand were used as polarity descriptors. Only the first of these two descriptors led to an increase in accuracy when added to the scoring function.

The freezing of the bond is next evaluated and defined by the value “*frozen*”. Once more, two options were evaluated: i. the number of protein atoms within a certain distance of the “*frozen*” bond (e.g., 10 Å) or ii. the presence or absence of a contact with the protein. The former value was used as a descriptor of the deepness of the rotatable bond in the protein binding site, while the latter is a measure of the freezing effect of the neighboring protein atoms. Although the former had a negligible effect on the scoring function accuracy, the latter significantly increased it. In this formalism, a bond was defined as completely frozen (*frozen* = 1.0) if the two atoms making this bond (atom 1 and atom 2) were within 0.5 Å of any protein atom van der Waals surface, each atom contributing 0.5 to *frozen*. If one of the atoms (i.e., atom 1) is not in close contact with the protein (atom 1 does not contribute to *frozen*), the atoms covalently bound to atom 1 are examined. If at least one of these connected atoms is within 0.5 Å of any protein atom van der Waals surface, atom 1 contributes 0.25 to *frozen*. If none of the atoms connected to atom 1 are in close contact with the protein, atom 1 does not contribute to *frozen*. In the example in Figure 4.3, atoms 1 and 3 have contacts with the protein, while atoms 2 and 4 do not. When considering the *frozen* value of bond 1-2, atom 1 would contribute 0.5, while atom 2 would contribute 0.25 (for being attached to atom 3, having a contact with the protein), yielding a *frozen* value of 0.75 for that bond. Analogously, bond 2-4 would have a *frozen* value of 0.25, stemming from the contribution of atom 2 only.



**Figure 4.3.** Frozen bond determination. Atoms 1 and 3 have van der Waals contacts (distance between van der Waals surfaces  $< 0.5 \text{ \AA}$ ) with the protein (purple surface), while atoms 2 and 4 do not. In this case, the bond between atoms 1 and 2 would have a *frozen* value of 0.75, while the bond between atoms 2 and 4 would have a value of 0.25.

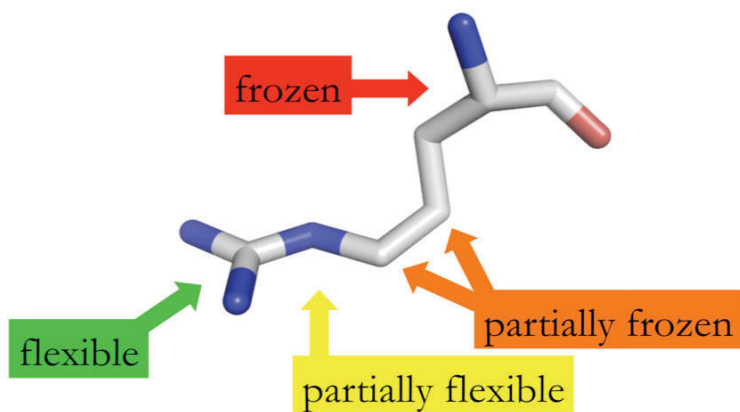
Conformational entropy loss upon ligand binding has two major components. First, the potential energy surface well in which a given conformation lies may get narrower. Second, some wells may disappear in the presence of the protein. Clearly, although strategies we investigated may lead to a more accurate description of the ligand entropy change than a function of  $N_{rot}$  alone, they evaluate the first component and not the second one, which would require a more exhaustive search of the ligand potential energy surface. This has indeed been proposed and implemented within AutoDock.<sup>19</sup>

The change in entropy of the protein upon binding (sixth term of Equation 4.2) was accounted for by reducing the interaction between ligand atoms and flexible side chains. This approach discussed previously<sup>21,22</sup> has been shown to increase the accuracy of the previous version of RankScore.<sup>9</sup> In the previous implementation, each side chain atom was assigned a scaling factor ( $\lambda$  in Equation 4.5) ranging from 0.6 to 1.0, which was next used to scale down the non-bonded interactions. In this

early implementation, the value of  $\lambda$  was residue-dependent (0.6 for arginine and 1 for alanine). However, interaction of the ligand with the beta carbon of an arginine does not lead to the freezing of the entire side chain and should not be penalized much while interactions with the guanidinium group of this same side-chain significantly reduce the mobility of the whole side chain. To account for this fact, the scaling factor  $\lambda$  is now dependent on the location of the atom on the side chain.

**Equation 4.5.** The entropic contribution to the binding from the protein is estimated by scaling down the interaction with the ligand ( $E_{\text{prot-lig}}$ ) calculated through a force field by a factor  $\lambda$ .

$$\Delta S_{\text{prot}} = \sum_{\text{non-bond pairs}} \lambda E_{\text{prot-lig}}^{\text{FF}}$$



**Figure 4.4.** Flexibility of side chain atoms. The further away from the peptide backbone an atom is, the more flexible it is considered, therefore the entropic penalty upon binding will be larger.

Finally, the solvation/desolvation contribution was evaluated using a generalized Born (GB) approach combined with an evaluation of the change in solvent accessible surface area (SASA) known to be proportional to the apolar change in solvation (Equation 4.6).<sup>20</sup> For this purpose, GB/SA has been fully implemented in FITTED.

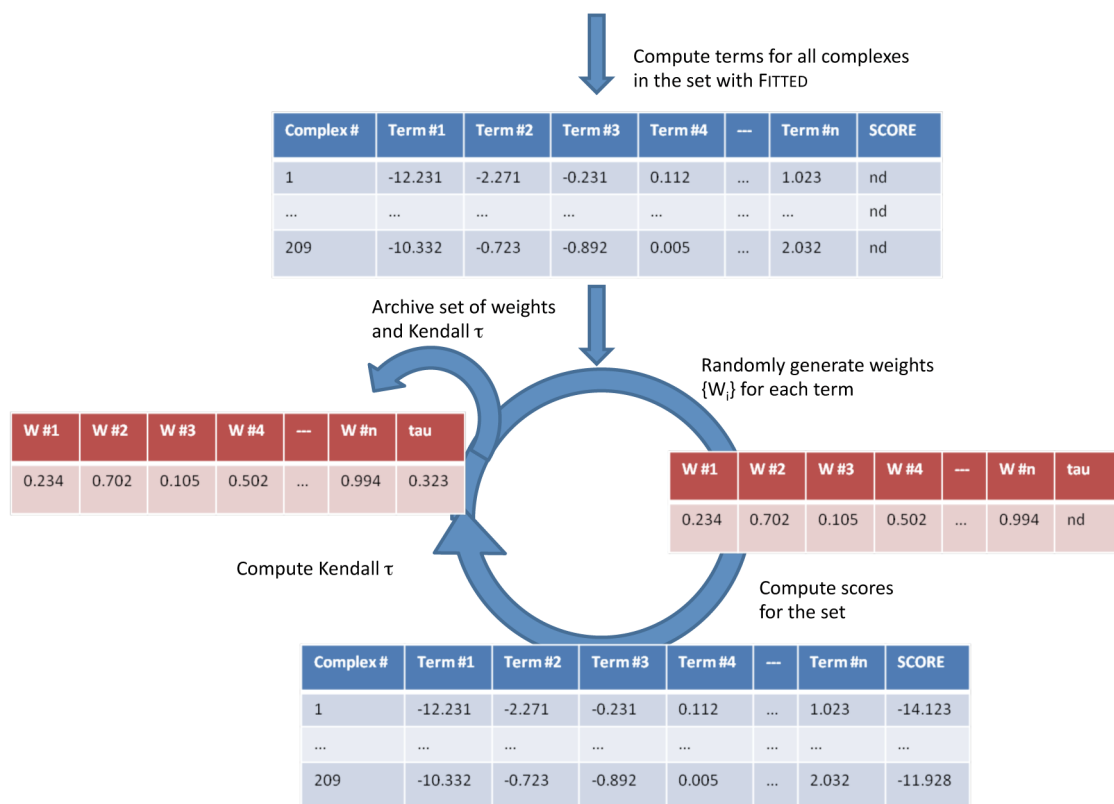
**Equation 4.6.** Free energy of solvation is calculated as a function of the solvent-accessible surface area (SASA) and the generalized Born approach.

$$\Delta G_{\text{solvation}} = f(\text{SASA}, \text{GB})$$

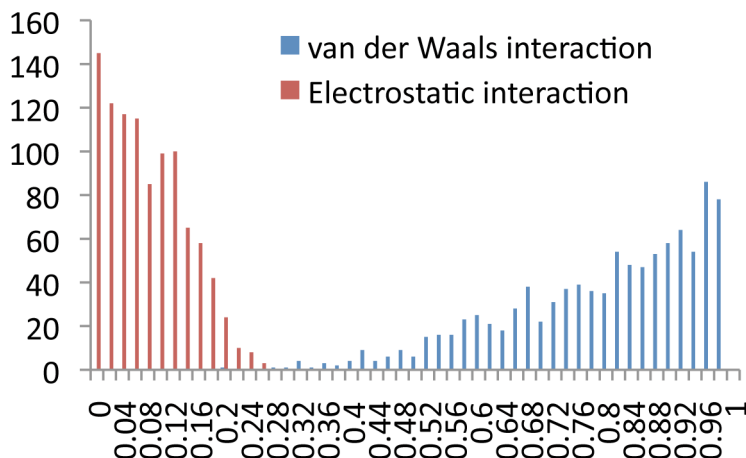
#### 4.2.8 Development of RankScore2 and RankScore3

The value for each of these terms was computed for the 209 complexes of the self-docking set 1 and for the nearly 1000 complexes of the cross-docking set 2.<sup>11</sup> Next,

random weights between 0 and 1 were generated for each term and the Kendall  $\tau$  measuring the correlation between the predicted and observed rankings was computed (Figure 4.5). This strategy will therefore lead to a scoring function optimized to predict the ranking of compounds and not to reproduce binding free energies. These steps were repeated 25,000 times and the corresponding table with 25,000 entries was sorted by decreasing  $\tau$ . In order to ensure a better transferability, we did not immediately select the best one but opted for an iterative approach. To do so, statistical analysis of the sets of weights leading to the top 4% (1,000 entries) correlation coefficients was carried out (Figure 4.6). Based on this information, the ranges for the different coefficients were constrained to the weights with the higher chance of leading to a better correlation with experimental binding affinities, and the protocol described in Figure 4.5 reiterated with the new ranges. For example, in a second iteration the coefficients for the van der Waals interaction were randomly generated between 0.00 and 0.40.



**Figure 4.5.** Procedure used to derive RankScore2 and RankScore3. In each case, 25,000 loops were performed, yielding an identical amount of sets of weights with their associated Kendall  $\tau$  value.



**Figure 4.6.** Distribution of scaling factors for van der Waals (blue) and electrostatic (red) interactions.

After five iterations of the entire protocol, RankScore2 was produced (Equation 4.7). This value is similar to those obtained previously with the same training set and the most accurate scoring functions assessed. The same procedure was carried out with the second set (cross-docked structures). Expectedly, as cross-docked structures are not as accurate as crystal structures (i.e., the amount of signal might be buried under the noise of the data), the trends (as the one shown in Figure 4.6) were not as marked and the function (referred to as RankScore3) derived from this set is not expected to be as accurate as RankScore2. In fact, most of the terms did not show any preferred range of values and none of the random set of weights led to Kendall  $\tau$  values as high as those observed with the previous set and RankScore2. Interestingly, these two functions RankScore2 and 3 were found to be very similar but RankScore3 will not be considered further.

**Equation 4.7.** Functional form of RankScore2. The different terms relate the different components ( $E^{vdW}$ ,  $E^{elec}$ ,  $E^{HB}$ ) of the intermolecular energy in the complex ( $E_{complex}$ ), the generalized Born polar solvation energy ( $\Delta G_{GB}$ ), the non-polar solvation energy proportional to the SASA ( $\Delta G_{SASA}$ ), the number of captured water molecules ( $N_{wat}$ ) and the number of rotatable bonds ( $N_{rot}$ ).

$$\begin{aligned}
 RankScore2 = & 0.680E_{complex}^{vdW} + 0.040E_{complex}^{elec} + 0.100E_{complex}^{HB} + \\
 & +0.000\Delta G_{GB} + 0.100\Delta G_{SASA} + 0.040N_{wat} + \\
 & +0.450\left(N_{rot} + 2 \cdot \sum_{bonds} f(N_{rot}, polarity, contact)\right)
 \end{aligned}$$

With  $N_{wat}$  being the number of captured water molecules,  $N_{Rot}$  being the number of rotatable bonds. Interestingly, the weight for the polar contribution to the solvation was found to be very low and setting it to zero reduces the computation time necessary to compute a score. This observation is consistent with our previous report.<sup>9</sup>

#### 4.2.9 Development of RankScore4

The last scoring function was developed using a slightly different approach. Six proteins (purine nucleoside phosphorylase, acetylcholinesterase, neuraminidase, oestrogen receptor, trypsin and P38 map kinase) and the corresponding decoys and active compounds were selected from the DUD set.<sup>21</sup> For each of the proteins, three or four conformations were considered. All the compounds were docked using FITTED in the flexible protein mode and the iterative protocol illustrated in Figure 4.5 applied. However, instead of computing  $t$  for the correlation between scores and binding affinities, we computed the area under a receiver operating characteristic (ROC) curve for the retrieval of known actives.

Once more the protocol was iterated five times with increasingly smaller ranges of coefficient values. At the end of this procedure RankScore4 was derived (Equation 10) with ROC values of 0.85 (PNP), 0.47 (AC), 0.73 (NA), 0.87 (ER), 0.95 (trypsin) and 0.67 (P38) with an average of 0.79. When comparing RankScore2 and RankScore4, it clearly appears that these two scoring functions are capturing very different information. While in RankScore 1, 2 and 3, the electrostatic interaction term was almost turned off, it became a major term in RankScore 4. In fact when RankScore2 was applied to these same 6 proteins and ligand sets, the ROC values were much lower (0.82, 0.40, 0.42, 0.67, 0.73, 0.47) with an average of 0.60. In addition, RankScore 2 provided rankings that were not better than random for AC, NA and P38 while only AC remained problematic after extensive training of the scoring function into RankScore4.

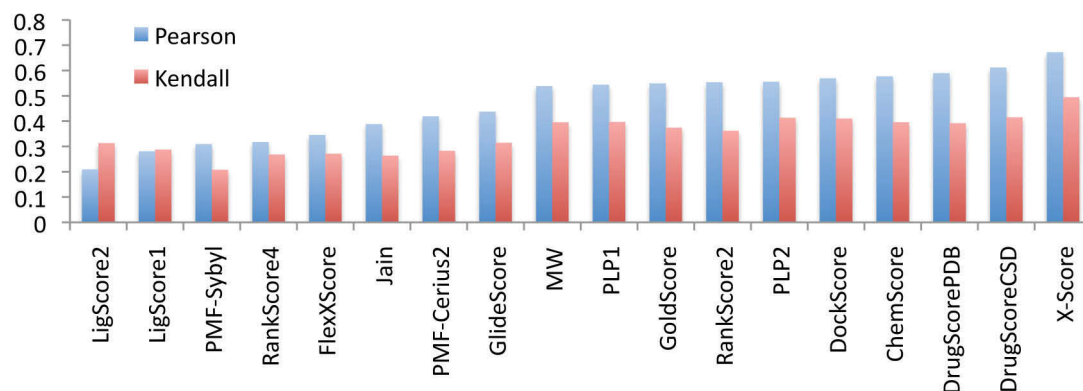
**Equation 4.8.** Functional form of RankScore4. See caption to Equation 4.7 for description of the terms.

$$\begin{aligned} \text{RankScore4} = & 0.184 \cdot E_{\text{complex}}^{\text{vdW}} + 0.746 \cdot E_{\text{complex}}^{\text{elec}} + 0.595 \cdot E_{\text{complex}}^{\text{HB}} + \\ & + 0.000 \cdot \Delta G_{\text{GB}} + 0.160 \cdot \Delta G_{\text{SASA}} + 0.050 \cdot N_{\text{wat}} + \\ & + 0.664 \cdot \sum_{\text{bonds}} f(N_{\text{rot}}, \text{polarity}, \text{contact}) \end{aligned}$$

#### 4.2.10 Application of the RankScore scoring functions to benchmark sets

In order to evaluate the predictive power and compare RankScore2 to other available functions, we applied it to the Wang's set of 100 protein/ligand complexes.<sup>22</sup> To our surprise, a close look at this set revealed some discrepancies and prompted us to curate it. Seven covalent inhibitors were removed and metal ions were added to metalloenzymes. With this set in hand, we applied 11 scoring functions including some that Wang and co-workers used. We observed that ChemScore was much better than previously reported with this set. In a previous report, we found that the accuracy of the scoring functions both Wang and co-workers and us looked at correlated well except for ChemScore.<sup>11</sup> With the set cleaned, accuracies obtained with both sets now correlate well. As shown on Figure 4.7 (see also Table A.7, Appendix), RankScore2 stands within the best scoring functions, behind X-Score and DrugScore, and within range of ChemScore, DockScore and PLP2. More interestingly, the correlation of the scores calculated with RankScore4 were among the least predictive of all the scoring functions considered.





**Figure 4.7.** Pearson correlation and Kendall  $\tau$  for a variety of scoring functions when applied to the testing set of 93 complexes.

As an additional validation, RankScore2 and 4 were also applied to the screening of libraries of known actives and decoys against thymidine kinase (TK), HIV-1 protease, thrombin, CDK2 and HIV reverse transcriptase. As can be seen in Table 4.1, RankScore4, developed for this specific purpose, is much more accurate than RankScore2 that was developed to reproduce binding affinities.

**Table 4.1.** AUC for the docking of libraries containing about 1000 ligands and decoys to 11 proteins (6 in the training set and 5 in the testing set)

SF	Training set						Testing set					Avg
	PNP	AC	NA	ER	TRP	P38	TK	HIVP	THR	CDK2	HIVRT	
RS2	0.82	0.40	0.42	0.67	0.73	0.47	0.56	0.63	0.63	0.65	0.61	0.60
RS4	0.85	0.47	0.73	0.87	0.95	0.67	0.88	0.87	0.90	0.83	0.65	0.79

This data indicates that distinct scoring functions for hit identification and lead optimisation should be developed.

### 4.3 Conclusions

It is well established that molecular mechanical force field energy on a single conformation is often not sufficient to provide a predictive tool for the fast estimation of the binding energy of ligands to proteins. However, the addition of other terms simulating other aspects of the energetics of binding to the equation

provides more predictive methods. In a first section we have shown that the intermolecular potential energies computed with many common force fields are highly correlated, showing that any force field would potentially perform as well in predicting ligand binding affinities. We next developed RankScore2 and RankScore3, built around the general Amber force field (GAFF), from an iterative process that optimized the scoring function weights in order to maximize the correlation of the calculated scores with experimental binding affinities. Validation of RankScore2 against a previously reported set of protein-ligand complexes indeed found it to perform better or as well as many commonly used scoring functions. We then trained RankScore4 to discriminate between active and inactive compounds in an analogous iterative fashion. Testing of RankScore2 and RankScore4 on other libraries of ligands and proteins revealed that RankScore4 perform significantly better than RankScore2, indicating that scoring function for VS should not be developed from active compounds only.

## **4.4 Experimental Section**

### **4.4.1 Preparation of the training set structures**

The preparation of the training set has been described in a previous report (see Chapter 3).<sup>11</sup> Succinctly, it involved: i. removal of all water molecules except the ones making bridging interactions (at least 3 hydrogen bonds) with both ligand and protein; ii. assignment of appropriate protonation states to both ligand and protein side chains; iii. constrained optimization by energy-minimization of the ligand with a force field.

### **4.4.2 Derivation of additional parameters for force fields**

Some of the force fields did not contain some parameters that were relevant to the calculations, so *ad hoc* parameters were derived for them. Some of the ligands featured moieties that were not included in the original force field parameterization, however they were deemed fairly rigid and starting from a crystalline structure, not much optimization would be necessary. Parameters (mostly bond stretch, and

torsions) for these moieties were derived by defining the new parameters so as to conserve the values observed in the crystal structure. For bond stretches, the equilibrium distance ( $r_0$ ) was defined as the average of the interatomic distance observed in all the molecules containing the particular moiety in the training set, with a stretch constant defined by analogy with another pair of atoms existing in the force field definition, or by default a large stretch constant to keep the bond stiff. The same strategy was used for bending and torsional parameters when needed. This study focuses on non-bonded interactions; hence the guessed parameters should not have much impact on the final result.

#### **4.4.3 Force field charges**

The forcefields included for use with Discover use a bond-charge increments system to assign partial charges, while Macromodel uses a bond dipole definition. When charge definitions were missing from the force field definition, semiempirical calculations using the AMPAC module in Insight II were performed on model molecules (e.g., for carbamates, *N*-methyl-methoxycarbamate was used) to determine appropriate bond increments. The bond increments were defined appropriately to reproduce the charge distribution observed by the semiempirical method.

#### **4.4.4 Development and validation of RankScore2 and RankScore4**

The sets used to develop RankScore2 and RankScore3 are those previously reported and were used with no further modifications. The set of protein-ligand complexes reported by Wang et al. used to validate RankScore2 was curated by a) removing covalent complexes (1a46, 1a5g, 1ba8, 1bb0, 1exw, 1yyy, 1zzz). In addition, missing metal atoms in metalloproteins (1af2, 1bzm, 1cbx, 1e96, 1mnc, 1tlp, 1tmn, 2ctc, 2tmn, 2xim, 2xis, 3cpa, 3tmn, 4tln, 4xia, 5p21, 5tln, 7tln, 8xia) were re-added from the original Protein Data Bank entry. Scores were re-calculated following the protocols recommended by the developers of the different scoring functions and described previously.<sup>11</sup> Proteins, active compounds and decoys used to derive and validate RankScore4 were retrieved from the DUD library.<sup>21</sup> A maximum of 1000

decoys were selected for each of the 11 proteins considered (AC: 1e66, 1gpn, 1h22, 1h23; CDK2: 1pxp, 1dm2, 1aq1, 1pxn; ER: 1sj0, 1err, 3ert; HIVP: 1b6l, 1hpo, 1hvp, 1pro; HIVRT: 1vrt, 1fk9, 1rt1, 1c1b; NA: 1f8d, 1f8e, 2qwe, 2qwf; P38: 1a9u, 1w7h, 1w82, 1w84; PNP: 1b8n, 1b8o, 1v48; THR: 1dwc, 1etr, 1tmt, 1ett; TK: 1e2k, 1ki3, 1of1, 2ki5; Trypsin: 1f0u, 1ghz, 1o2h, 1qb9).

The ligands from the various sets were docked using the FITTED2.6 suite and all the energy terms were output and tabulated. Python scripts were used to randomly generate the weights, compute the scores and either Kendall  $\tau$  (RankScore2) or area under ROC curves (RankScore4). Another Python script was used to analyze the generated data and define the range for the next cycle (see Figure 4.5).

## 4.5 References

1. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7-S26.
2. Aqvist, J.; Medina, C.; Samuelsson, J. E., A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385-391.
3. Hansson, T.; Marelus, J.; Åqvist, J., Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27.
4. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.
5. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411.
6. 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.

7. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
8. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289-307.
9. Moitessier, N.; Therrien, E.; Hanessian, S., A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic  $\beta$ -secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885-5894.
10. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of FITTED 1.5 to the Virtual Screening of Potential HCV Polymerase Inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 902-909.
11. 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.
12. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765-784.
13. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.
14. Böhm, H. J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243.
15. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
16. Jain, A. N., Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427-440.

17. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425-445.
18. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11-26.
19. Lee, J.; Seok, C., A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins: Struct., Funct., Genet.* **2008**, *70*, 1074-1083.
20. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C., The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005-3014.
21. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.
22. Wang, R.; Lu, Y.; Wang, S., Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287-2303.

[ This page was intentionally left blank ]

## **Chapter 5: Modeling of supramolecular compounds binding to guanine quadruplex DNA structures**

### ***5.1 Introduction***

#### **5.1.1 Cancer**

Cancer is a disease characterized by the invasive and uncontrolled growth of malignant tumour cells.<sup>1</sup> It is responsible for 13% of all human deaths (30% in developed countries), having claimed 7.4 million victims worldwide in 2004.<sup>2</sup> A malignant tumour can originate in many organs, and if left untreated it can metastasize to other parts of the organism eventually leading to death. The strong chemotherapeutics required for cancer treatment target different stages of the DNA transcription and replication pathways, and their high toxicity stems from their lack of specificity, which do not spare healthy cells.

#### **5.1.2 Telomeres**

The end of a eukaryotic DNA chromosome contains the telomere, a repetitive non-coding sequence of DNA that protects the genome from damage. Telomeres are around 10-15 kilobases in length and are mostly a duplex, except at the extreme 3' termini where they have a single strand.<sup>3-5</sup> A variety of proteins are associated to the DNA at the level of the telomeres, notably the shelterin complex, TRF1, TRF2 and POT1,<sup>6</sup> as well as the enzyme telomerase.<sup>7</sup> Due to the inefficient replication of the DNA the telomere becomes shorter after each replication round, losing 25-200 bases after each generation of cells and eventually reaching a critical length.<sup>8</sup> At this point, tumour suppressor mechanisms (e.g., p53, Rb) are activated forcing the cell into senescence, preventing further cell division. If the tumour suppressor mechanisms fail, cell division continues leading to the complete loss of the telomere, which in turn usually leads to apoptosis. Nevertheless, in some cases (a frequency of  $\sim 10^{-7}$ ) cells can prevent apoptosis by activating a mechanism to maintain telomere



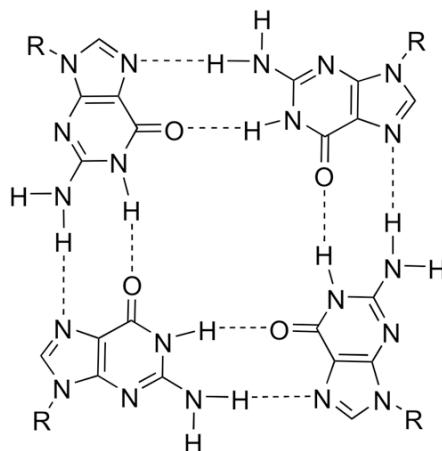
length, thus immortalizing the cell. In 85-90% of cancerous cells, this mechanism involves the enzyme telomerase.<sup>9</sup>

### **5.1.3 Telomerase**

Telomerase is a ribonucleoprotein, consisting of an RNA template and a reverse transcriptase domain (hTERT), with the ability of extending the genomic DNA by inserting tandem repeats of a sequence complementary to its template.<sup>10</sup> The repetitive sequence varies with the species; in the case of the human telomerase, this motif is GGGTTA.<sup>7</sup> The observation that telomerase is active in cancerous cells but not in normal somatic cells, makes it an attractive target for therapeutic intervention. Inhibition of telomerase activity has been attempted at different levels, namely: modulating the activity of the catalytic hTERT domain, targeting the RNA component, affecting other proteins binding to telomeres or interacting directly with telomeric DNA.<sup>11</sup>

### **5.1.4 Guanine quadruplexes**

The folding of guanine-rich single strands of DNA into quadruplexes through Hoogsteen base-pairing has received a lot of attention from the scientific community.<sup>12</sup> First identified in telomeric regions of the eukaryotic chromosome, G-quadruplexes (G4) have been since found in varied regions of the genome, notably in the promoter regions of the oncogenes c-kit and c-myc.<sup>13</sup> The structure of a guanine tetrad is shown in Figure 5.1. The structure of G4s has been studied by diverse biophysical techniques, including X-ray crystallography, NMR and circular dichroism (CD) spectroscopy.



**Figure 5.1.** Guanine quadruplex hydrogen-bonding structure showing Hoogsteen base-pairing.

The arrangement of single-stranded DNA to form a unimolecular G4 can lead to diverse foldings,<sup>14</sup> although few have been experimentally observed. In the case of the human telomeric sequence, 5'-d(GGG[TTAGGG]<sub>3</sub>)-3', Patel and co-workers found a basket-like folding for the human telomere in sodium buffer solution by NMR.<sup>15</sup> Later, Neidle and co-workers solved the crystal structure of a propeller-like folding in presence of potassium cations.<sup>16</sup> In addition to these two structures, a hybrid folding was observed by NMR in K<sup>+</sup>-containing solutions.<sup>17</sup>

### 5.1.5 G-quadruplex binders as telomerase inhibitors

Zahler *et al.* observed that telomerase activity could be inhibited by the formation of G-quadruplexes in telomeric DNA.<sup>18</sup> The stabilization of G-quadruplexes with small molecules emerged then as a potential strategy for telomerase inhibition, with the first success reported by Hurley and co-workers for a 2,6-diamidoanthraquinone-based ligand.<sup>19</sup> Since then, different organic molecules have been used for this purpose, such as telomestatin,<sup>20</sup> BRACO-19,<sup>21</sup> cationic porphyrins,<sup>22</sup> anthracene-9,10-diones<sup>23</sup> and bisquinolinium derivatives.<sup>24</sup>

### 5.1.6 Molecular modeling of G-quadruplex binders

Although the reports of modeling techniques applied to protein targets outnumber those applied to nucleic acids, most of the methods used for the former are applicable to the latter. Molecular dynamics simulations involving quadruplex DNA have been recently reviewed by Sponer and Spackova,<sup>25</sup> but little is mentioned

about systems including small molecules. Read *et al.* developed a method for calculating relative binding affinities from molecular dynamics time-averaged structures, which was successful at predicting the ranking of a family of compounds binding to the human telomere G-quadruplex.<sup>26</sup> Similar experiments using automated docking and simulated annealing refining were used to study the binding of peptides conjugated to acridines and acridones<sup>27</sup> and square planar nickel complexes.<sup>28</sup> When reporting the crystal structure of BSU6039 with the two-strand G-quadruplex from the *Oxytricha nova* telomeric sequence, a 2 ns molecular dynamics simulation was used to assess the stability of the pseudo-intercalating binding mode proposed.<sup>29</sup> While this work was in progress, the ICM docking program<sup>30</sup> was used for the automated docking of a library of compounds to the crystal structure of the human telomere G-quadruplex.<sup>31</sup>

#### **5.1.7 Summary of work presented in this chapter**

The present chapter describes the efforts performed to design and model platinum (II) complexes with a variety of heteroaromatic ligands as potential guanine quadruplex binders and telomerase inhibitors. In particular, we describe the development and application of a method to assess binding affinities of platinum complexes to G-quadruplex DNA. The first section describes the derivation of molecular mechanical terms compatible with the GAFF force field<sup>32</sup> for these inorganic complexes. In subsequent sections, we applied these parameters to the docking and molecular dynamics simulation of a platinum molecular square and different platinum complexes with extended  $\pi$ -surface ligands.

Part of this work has been published as a communication in the *Journal of The American Chemical Society*,<sup>33</sup> while other manuscripts are currently in preparation.

#### **5.1.8 Acknowledgments**

The work described in this chapter was done in collaboration with Roxanne Kieltyka (Department of Chemistry, McGill University) who performed the synthesis and biophysical studies of the platinum (II) complexes and Johans Fakhoury

(Experimental Medicine, McGill University), who performed the telomerase inhibition assays.

## **5.2 Results and discussion**

### **5.2.1 Development of molecular mechanical parameters for platinum (II) complexes with heteroaromatic ligands**

The modeling of platinum complexes poses the challenge of the lack of parameters describing square planar complexes within biomacromolecular force fields. Extension of the Amber force field is straightforward to implement, as new atom types and interactions are defined in an external force field modification file (*frcmod*). Development of a new parameter set requires the calculation of the energies of a set of distorted structures at a high level of theory. In the case of the Amber force field, the latest parameter set (*parm99*) was developed from calculations at the MP2/6-31G\* level of theory.<sup>34,35</sup> An extension of this parameter set to general organic molecules (*GAFF*, *Generalized Amber Force Field*), also developed from *ab initio* data at the MP2/6-31G\* level, allows for the simulation of most organic molecules with the Amber force fields, but it does not include organometallic compounds or transition metal complexes.<sup>32</sup> The choice of a basis set to treat platinum nuclei precludes the use of the 6-31G\* basis set used in the GAFF force field, which is not parameterized for transition metals, hence we decided to use the LACV3P\*\* basis set available in the Jaguar software, which includes parameters for all transition metals. This basis set is a triple-zeta contraction of the LACVP basis set, including the effective core potentials developed by Hay and Wadt for elements in the 4<sup>th</sup>, 5<sup>th</sup> and 6<sup>th</sup> periods of the periodic table;<sup>36-38</sup> light atoms are treated with 6-311G basis sets.

Development of force field parameters for each interaction requires the adjustment of a pair of parameters ( $K$  and  $x_0$ ) to fit a quadratic equation of the type  $E = K(x-x_0)$ , so that the relative energies of different conformers are accurately described by the molecular mechanical potential. To this effect, we constructed a series of structures distorted from the optimized conformation and calculated the energy at the

B3LYP/LACV3P\*\* level of theory, varying one internal coordinate (bond length, angle, torsion) at a time, sequentially covering all the parameters missing from the force field. We then calculated the energy using molecular mechanics with the set of available force field parameters using the GAFF force field. Figure 5.2 shows the format of the final *frcmod* file with the developed parameters.

```

MASS
Pt 195.08          2.000          Square planar platinum (II)
n5 14.01           0.530          idem n4
nz 14.01           0.530          idem na
DU 1.0             0.000          Dummy atom

BOND
Pt-n4 133.171      2.089          B3LYP/LACV3P**
Pt-n5 133.171      2.089          B3LYP/LACV3P**
Pt-na 128.571      2.054          B3LYP/LACV3P**
Pt-nz 128.571      2.054          B3LYP/LACV3P**
c3-n5 293.6        1.499          same as c3-n4
hn-n5 369.0        1.033          same as hn-n4
ca-nz 470.3        1.350          same as ca-na
Pt-DU 500.0        1.000          Dummy atom

ANGLE
Pt-n4-c3 26.906    119.664    B3LYP/LACV3P**
Pt-n5-c3 26.906    119.664    B3LYP/LACV3P**
Pt-n4-hn 19.277    106.877    B3LYP/LACV3P**
Pt-n5-hn 19.277    106.877    B3LYP/LACV3P**
Pt-na-ca 12.170    122.977    B3LYP/LACV3P**
Pt-nz-ca 12.170    122.977    B3LYP/LACV3P**
n4-Pt-n5 78.428    86.918    B3LYP/LACV3P**
na-Pt-nz 122.886    91.663    B3LYP/LACV3P**
n5-Pt-na 26.633    90.800    B3LYP/LACV3P**
n4-Pt-nz 26.633    90.800    B3LYP/LACV3P**
n4-Pt-na 0.000     180.000    B3LYP/LACV3P**
n5-Pt-nz 0.000     180.000    B3LYP/LACV3P**
na-ca-ha 51.200    112.420    same as ha-c2-na
nz-ca-ha 51.200    112.420    same as ha-c2-na
hn-n5-c3 46.200    110.110    same as hn-n4-c3
n5-c3-hx 49.000    107.910    same as n4-c3-hx
hn-n5-hn 40.500    108.110    same as hn-n4-hn
ca-ca-nz 70.2      118.34     same as ca-ca-na
ca-nz-ca 67.1      119.80     same as ca-na-ca
c3-c3-n5 66.0      108.93     same as c3-c3-n4
DU-Pt-nz 500.000    90.000     Dummy atom
DU-Pt-na 500.000    90.000     Dummy atom
DU-Pt-n4 500.000    90.000     Dummy atom
DU-Pt-n5 500.000    90.000     Dummy atom
DU-Pt-DU 0.000     180.000    Dummy atom

DIHE
n5-Pt-n4-c3 1      1.013      0.000      2.000      B3LYP/LACV3P****
n4-Pt-n5-c3 1      1.013      0.000      2.000      B3LYP/LACV3P****
n4-Pt-na-ca 1      0.755      0.000      1.000      B3LYP/LACV3P****

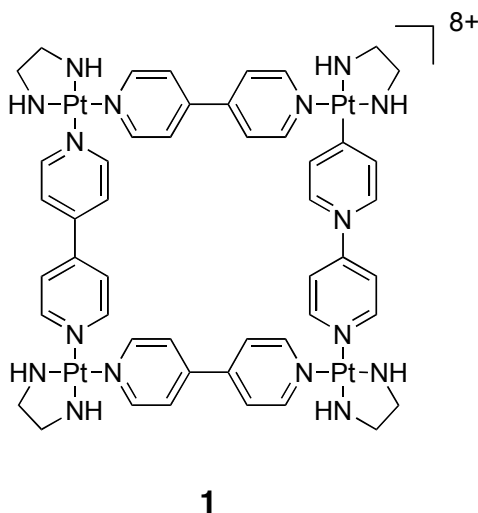
```

n5-Pt-nz-ca	1	0.755	0.000	1.000	B3LYP/LACV3P****
n4-Pt-nz-ca	1	0.000	0.000	0.000	B3LYP/LACV3P****
n5-Pt-na-ca	1	0.000	0.000	0.000	B3LYP/LACV3P****
na-Pt-nz-ca	1	0.000	0.000	0.000	B3LYP/LACV3P****
nz-Pt-na-ca	1	0.000	0.000	0.000	B3LYP/LACV3P****
c3-n4-Pt-na	1	0.000	0.000	0.000	B3LYP/LACV3P****
c3-n5-Pt-nz	1	0.000	0.000	0.000	B3LYP/LACV3P****
c3-n5-Pt-na	1	1.013	0.000	2.000	B3LYP/LACV3P****
c3-n4-Pt-nz	1	1.013	0.000	2.000	B3LYP/LACV3P****
n5-Pt-n4-hn	1	0.129	180.000	2.000	B3LYP/LACV3P****
n4-Pt-n5-hn	1	0.129	180.000	2.000	B3LYP/LACV3P****
hn-n4-Pt-na	1	0.000	0.000	0.000	B3LYP/LACV3P****
hn-n5-Pt-nz	1	0.000	0.000	0.000	B3LYP/LACV3P****
hn-n4-Pt-nz	1	0.129	180.000	2.000	B3LYP/LACV3P****
hn-n5-Pt-na	1	0.129	180.000	2.000	B3LYP/LACV3P****
X -c3-n5-X	9	1.400	0.000	3.000	same as X -c3-n4-X
X -ca-nz-X	4	1.200	180.000	2.000	same as X -ca-na-X
ca-nz-Pt-DU	1	0.000	0.000	0.000	Dummy atom
ca-na-Pt-DU	1	0.000	0.000	0.000	Dummy atom
hn-n4-Pt-DU	1	0.000	0.000	0.000	Dummy atom
hn-n5-Pt-DU	1	0.000	0.000	0.000	Dummy atom
c3-n5-Pt-DU	1	0.000	0.000	0.000	Dummy atom
c3-n4-Pt-DU	1	0.000	0.000	0.000	Dummy atom
IMPROPER					
ca-ca-na-Pt	1	7.526	180.000	2.000	B3LYP/LACV3P****
ca-ca-nz-Pt	1	7.526	180.000	2.000	B3LYP/LACV3P****
X -X -ca-ha	1	1.1	180.000	2.000	based on X-X-ca-hc
X -X -cp-cp	1	1.1	180.000	2.000	based on X-X-ca-hc
NONBON					
Pt		2.5400	0.3820		
n5		1.8240	0.1700	idem n4	
nz		1.8240	0.1700	idem na	
DU		0.0000	0.0000	Dummy atom	

**Figure 5.2.** Sample force field modification file (frcmod) for a platinum (II) square planar complex containing two aromatic nitrogen ligands and two aliphatic nitrogen ligands.

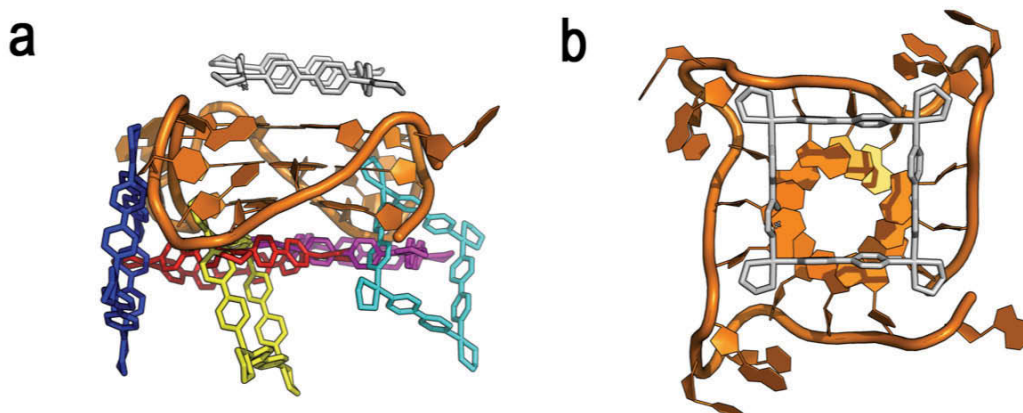
### 5.2.2 Platinum (II) square complex

The geometric arrangement of guanines in a G-quadruplex exposes a large flat  $\pi$ -surface with four aromatic groups equidistant from each other. As a model for complementarity, a molecular square was designed featuring four platinum (II) centers bridged by an aryl linker. The synthesis of these platinum squares is expeditive, relying on the self-assembly of the platinum ethylenediamine corners (as the nitrate) with the 4,4'-bipyridyl linkers.<sup>33</sup>



**Figure 5.3.** Structure of molecular square **1**.

Molecular modeling studies were conducted to understand the mode of binding and the fit of complex **1** within the G-quadruplex structure. For this, an approach that combines automated docking, molecular dynamics (MD) simulations, and evaluation of binding affinity was examined. We modified our recently developed docking program FITTED,<sup>39</sup> previously found to accurately predict binding modes of protein ligands,<sup>40</sup> to be able to dock ligands onto nucleic acids. With this tool, we docked **1** to the X-ray crystal structure of a G-quadruplex 22-mer;<sup>16</sup> a total of 100 docking runs were computed, with representative binding modes shown in Figure 5.4a. The predicted most favorable binding mode was one where the platinum square **1** is parallel to the plane of the terminal G-quartet (Figure 5.4b). In this mode, short Pt-P distances are consistent with electrostatic interaction of each of the Pt atoms of **1** with the backbone phosphates. Moreover, we observed that the NH<sub>2</sub> groups of each of the ethylenediamine ligands are hydrogen-bonded to the phosphate oxygens, as it was found previously for a monometallic platinum (II) complex interacting with an intermolecular G-quadruplex.<sup>41</sup> As well, one of the aromatic rings in each of the 4,4'-bipyridyl ligands interacts with a guanine base in a distorted T-shape geometry (see Figure 5.4b).

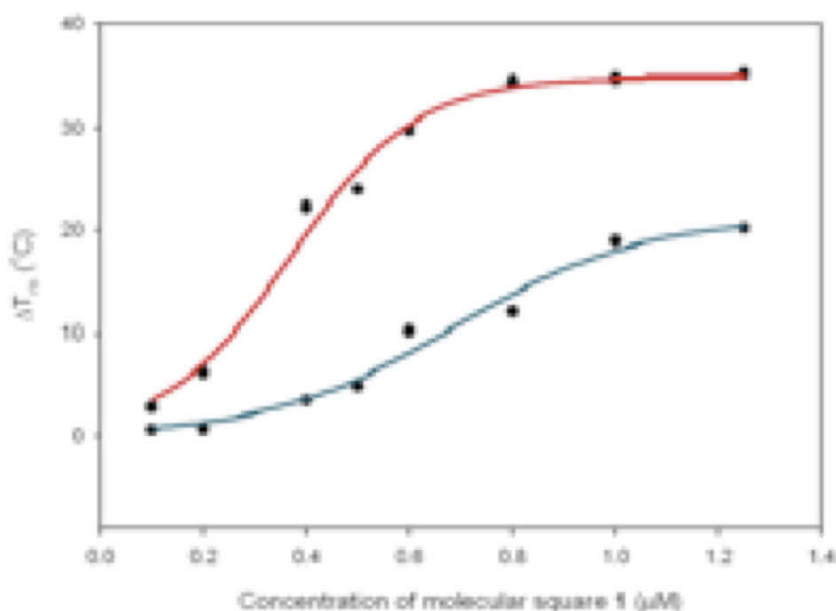


**Figure 5.4.** Docking of platinum square **1** to G-quadruplex structure 1kf1. (a) Representative docked conformations; (b) stacked conformation. This pictures were produced with PyMol.<sup>42</sup>

This binding mode, along with other representative docked complexes (Figure 5.4a), was further evaluated by running 4 ns MD simulations. Snapshots of these simulations were post-processed with the MM-PBSA formalism.<sup>43</sup> Relative free energies of binding show a stabilization of  $\sim 10$  kcal/mol for the parallel mode over all other structures, consistent with end-stacking of previously reported G-quadruplex binders. Thus, molecular modeling studies confirmed the excellent complementarity in size and interactions between square **1** and the quadruplex, in a parallel end-stacked mode.

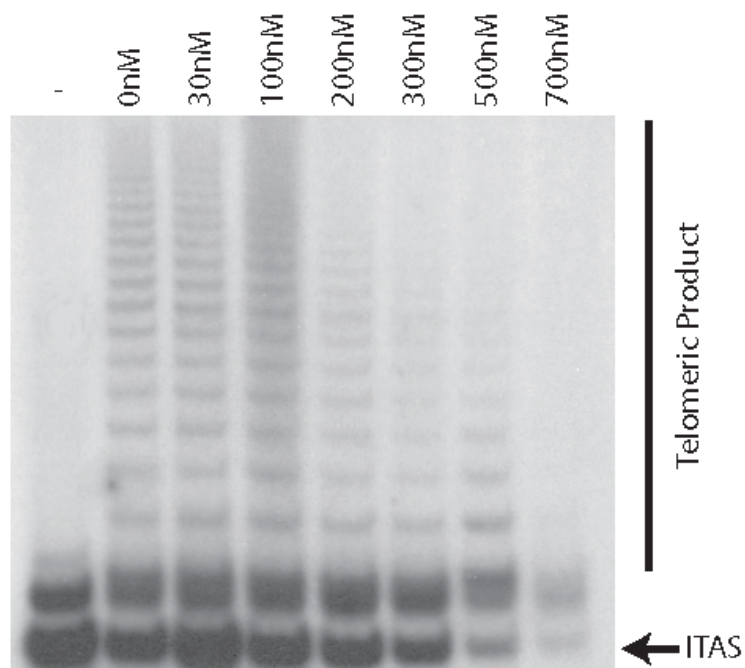
The binding ability of square **1** to the human telomere G-quadruplex was first evaluated using a FRET melting assay, which showed large stabilization of the G-quadruplex, with an increase of  $34.5$  °C in the thermal denaturation temperature with  $0.75$   $\mu\text{M}$  of **1**. This significant increase is competitive with many of the best reported quadruplex binders, such as telomestatin ( $30.3$  °C),<sup>44</sup> a nickel salen complex ( $33.2$  °C),<sup>28</sup> a macrocyclic oligoamide ( $33.8$  °C),<sup>44</sup> a bisquinolinium ( $29.7$  °C),<sup>24</sup> and BRACO-19 ( $27.5$  °C),<sup>28</sup> which required  $1$   $\mu\text{M}$  of the ligand to achieve these values. Thus, platinum square **1** is an excellent stabilizer of the G-quadruplex motif. A subsequent FRET assay was performed to evaluate the selectivity of **1** for G-quadruplex versus duplex structures (Figure 5.5). It was found that **1** indeed stabilizes duplex DNA, undoubtedly because of its high positive charge; however, far greater stabilization for G-quadruplex DNA was observed.





**Figure 5.5.** FRET stabilization curve of square 1 with quadruplex (red) and duplex DNA (blue).

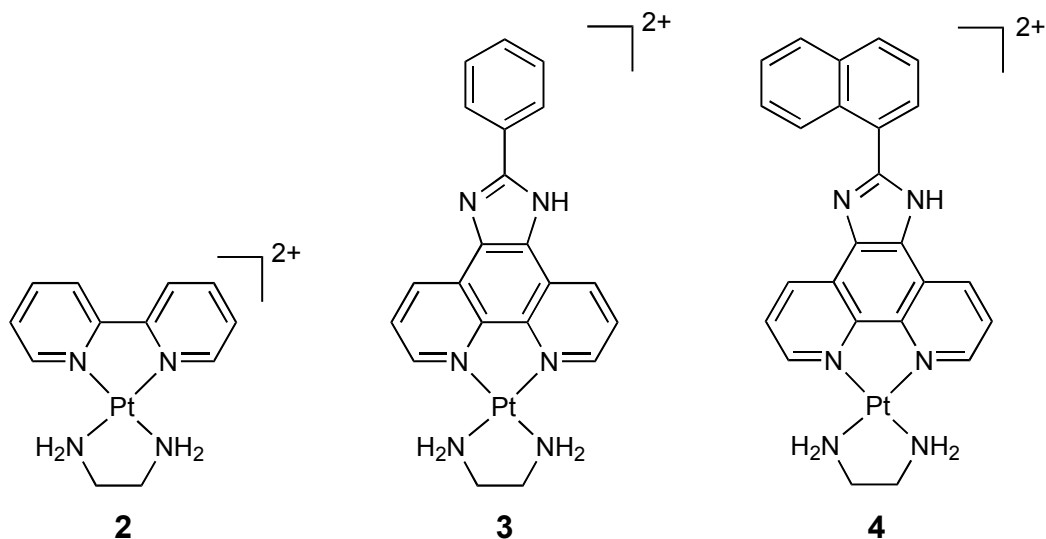
We were then interested in the potential of complex 1 to inhibit the enzyme telomerase. For this, a modified version of the telomeric repeat amplification protocol (TRAP) assay was performed (Figure 5.6). Inhibition of telomerase by complex 1 was found to be one of the strongest of reported G-quadruplex binders, with an IC<sub>50</sub> value of 0.197±0.056 μM.



**Figure 5.6.** TRAP assay of complex 1, showing ladders generated by the action of telomerase on a TS primer (PCR amplified). The lower band is an internal control primer (ITAS).

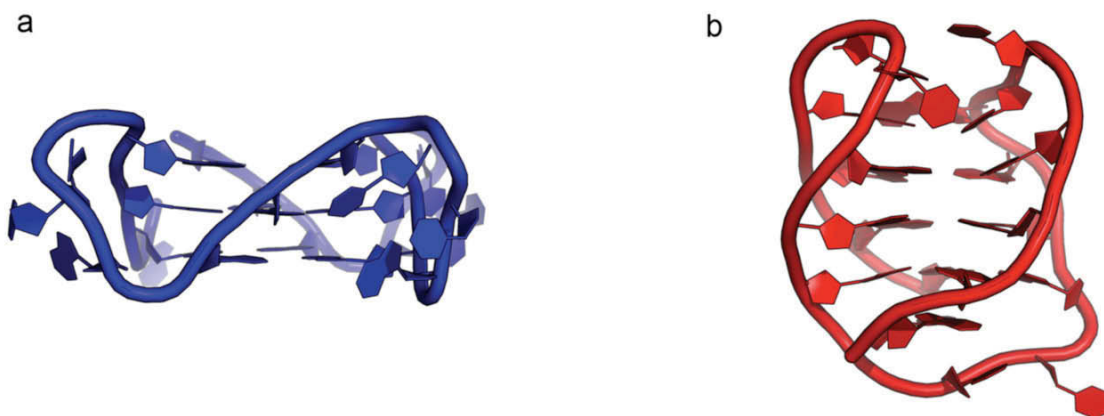
### 5.2.3 Extended $\pi$ -surface platinum (II) complexes as G-quadruplex binders

Platinum complexes having a single metal center can expose a large  $\pi$ -surface with the potential to complement the exposed surface of G-quadruplexes. Previously, complex **2** (see Figure 5.7) was reported as a duplex DNA binder with a constant of  $10^6 \text{ M}^{-1}$ ;<sup>45</sup> however, a contemporaneous study reported that larger complexes such as platinum ethylenediamine dipyridophenazine exhibit lower binding constants towards B-DNA, two orders of magnitude smaller.<sup>46</sup> This fact would point out a potential mismatch of sizes between both binding partners as the  $\pi$  surface increases, which could be exploited for binding to the larger surface available in G-quadruplexes. In a previous study, Kieltyka *et al.* showed that, indeed, extending the  $\pi$  surface of the aromatic ligand bound to the platinum center (i.e., moving from complex **2** to **3** and **4**, see Figure 5.7) increased the binding affinity of the platinum complex for an intermolecular G-quadruplex, while at the same time exhibiting a higher selectivity for quadruplex vs. duplex DNA.<sup>47</sup>



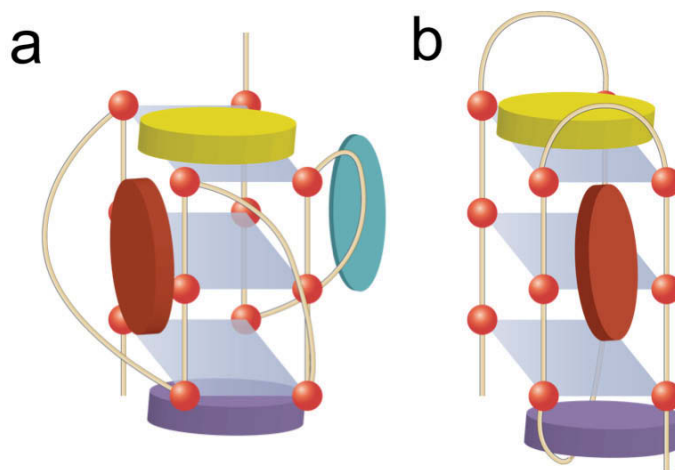
**Figure 5.7.** Structures of platinum (II) complexes: platinum ethylenediamine 2,2'-bipyridyl, Pt(en)bipy, **2**; platinum ethylenediamine phenylphenanthroimidazole, Pt(en)PIP, **3**; platinum ethylenediamine naphthylphenanthroimidazole, Pt(en)PIN, **4**.

The folding of the human telomeric sequence can lead to different topologies, among which we consider a propeller structure with all G strands oriented parallel to each other, observed in the crystal structure in presence of potassium cations;<sup>16</sup> and a basket structure, where one pair of strands flows in the opposite direction to the other pair in the quadruplex, observed in the NMR structure in Na<sup>+</sup>-containing medium (see Figure 5.8).<sup>15</sup>



**Figure 5.8.** Foldings of the human G-quadruplex considered. a: parallel structure (PDB code 1kf1); b: anti-parallel structure (PDB code 143d). This pictures were produced with PyMol.<sup>42</sup>

The goals of modelling in this project are two-fold: to determine the binding mode of the platinum (II) complexes to the G-quadruplex structure, and to explain the difference in activity of both complexes. To this effect, we applied a hybrid docking/molecular dynamics technique, previously developed to study the binding of a platinum molecular square to G-quadruplexes.<sup>48</sup> Briefly, different plausible binding modes were generated with the docking program FITTED,<sup>39,49,50</sup> and their binding affinity was then evaluated by analyzing snapshots of a MD simulation with the MM-PBSA formalism. We first considered the binding of the platinum complexes to the X-ray crystal structure of the G-quadruplex (PDB 1kf1, see Table 5.1), following a protocol previously reported.<sup>48</sup>



**Figure 5.9.** Schematic depiction of the binding modes considered as starting structures for complexes interacting with different foldings of a G-quadruplex. (a) Poses interacting with 1kf1; yellow: top; purple: bottom; teal: loop; red: groove. (b) Poses interacting with 143d; yellow: handles; purple: bottom; red: groove.

**Table 5.1.** Binding affinity of platinum (II) complexes to human telomere G-quadruplex parallel folding (1kf1) and antiparallel folding (143d) by the MM-PBSA method on 5 different initial conformations for each. Errors are one standard deviation from the average.

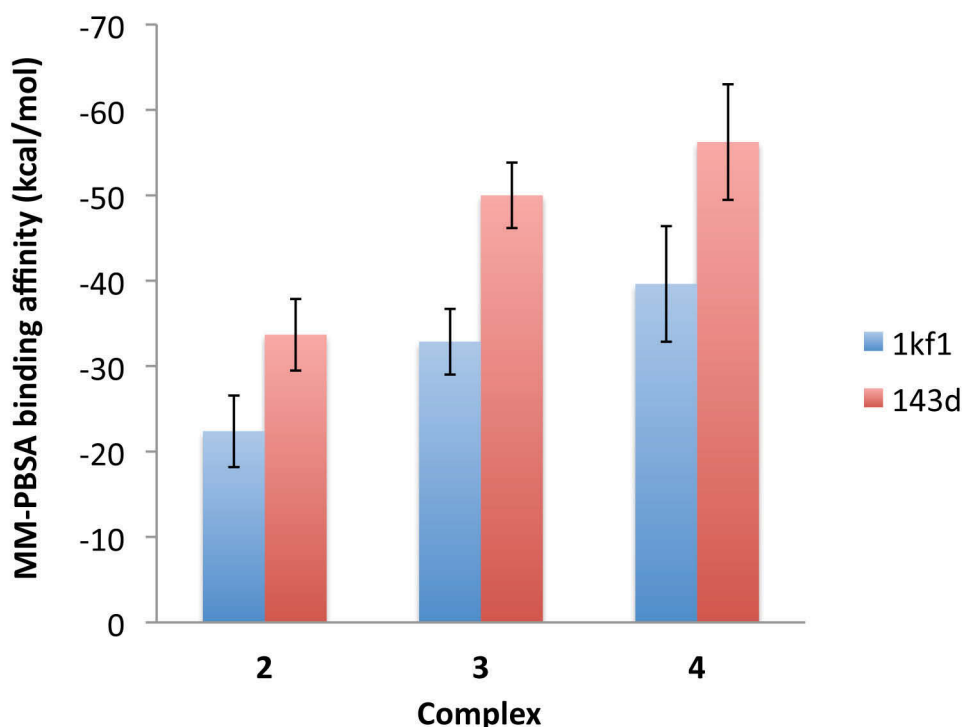
Complex	1kf1		143d	
	Binding mode <sup>a</sup>	E <sub>bind</sub> (kcal/mol)	Binding mode <sup>a</sup>	E <sub>bind</sub> (kcal/mol)
2	top	-22.37 ± 3.15	handles	-33.68 ± 4.19
	bottom	-22.18 ± 3.40	bottom	-20.79 ± 7.58
	loop	-17.29 ± 2.80	groove	-15.32 ± 3.50
	groove	-19.56 ± 4.20	groove	-14.18 ± 4.40
3	top	-32.86 ± 4.06	handles	-50.00 ± 3.84
	bottom	-32.79 ± 2.68	bottom	-44.85 ± 4.51
	loop	-22.02 ± 3.91	groove	-15.30 ± 4.89
	groove	-15.30 ± 3.29	groove	-32.51 ± 3.96
4	top	-39.62 ± 5.83	handles	-56.23 ± 6.77
	bottom	-36.59 ± 3.63	bottom	-40.07 ± 7.66
	loop	-36.46 ± 4.36	groove	-29.57 ± 5.62
	groove	-23.48 ± 2.40	groove	-17.16 ± 3.43

<sup>a</sup> see Figure 5.9

Following the experimental observation that an anti-parallel structure is favoured in presence of PIP, we turned our attention to the basket-like folding of the human telomeric sequence solved by NMR spectroscopy.<sup>15</sup> In this structure, the TTA loops are oriented *on top* of the guanine quadruplex stacks instead of *away* from them as in the all-parallel folding (see Figure 5.8); as a result, there are  $\pi$ - $\pi$  stacking interactions between the bases in the loops and the quadruplexes, and a binding molecule would be able to intercalate within this environment. To assess this possibility, we first docked the platinum complexes to a G-quadruplex structure missing the TTA loops. The binding modes obtained from these experiments were used as a starting structure for molecular dynamics simulations on the complete anti-parallel folded structure to assess the binding affinity through MM-PBSA (Table 5.1).

From these binding affinity values several conclusions can be drawn. First, stacking seems to be the preferred binding mode for all these platinum (II) complexes, as they show the best binding affinity among all other modes considered. Second, it is clear that the platinum (II) complexes with extended  $\pi$ -surface ligands have a much better binding affinity for the G-quadruplexes than the 4,4'-bipyridyl

complex, mainly stemming from a larger van der Waals contribution in the former structures. Finally, the binding affinities for the PIN and PIP complexes predicted by the MM-PBSA method appear to be very similar, as both values are within the standard deviations in both G-quadruplex foldings considered (see Figure 5.10). This correlates well with the small difference in IC<sub>50</sub> observed in the experimental assays (*vide infra*), where both complexes exhibit binding affinities within the same order of magnitude.

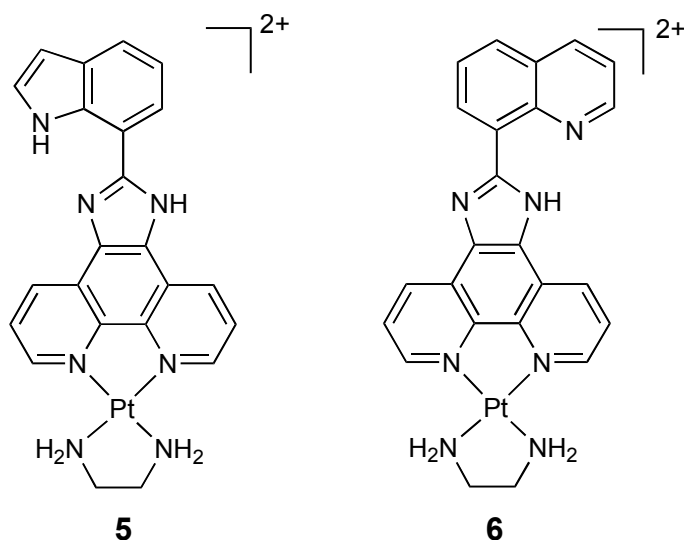


**Figure 5.10.** Binding affinities for stacking modes of complexes **2**, **3** and **4**. In blue, binding to parallel G-quadruplex folding (PDB code 1kf1); in red, binding to anti-parallel G-quadruplex folding (PDB code 143d). Error bars correspond to one standard deviation for the binding energy as calculated from the molecular dynamics snapshots.

#### 5.2.4 Hydrogen-bonded ligands for an even larger $\pi$ -surface

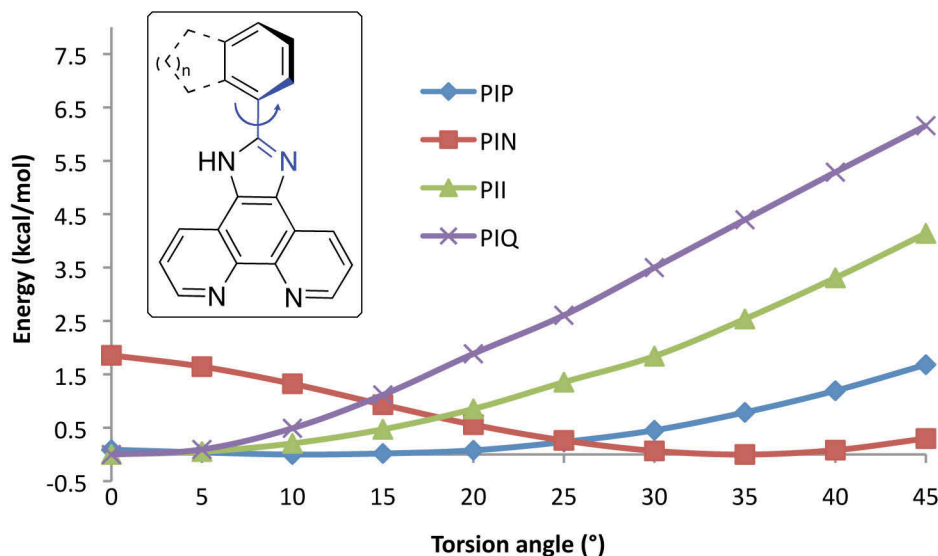
From the results obtained for the PIP and PIN ligands, it became evident that extending the  $\pi$ -surface of the ligands increased the binding affinity of the compounds for the G-quadruplex motif. However, the  $\pi$ -surface did not appear to be continuous, as the biphenyl-type bond (see Figure 5.12) between the imidazole

carbon and the aryl group forced the latter out of the plane of the rest of the ligand. We wished then to assess a new family of ligands featuring an internal hydrogen bond between one of the imidazole nitrogens and a new hydrogen bond donor or acceptor (Figure 5.11).



**Figure 5.11.** Structures of platinum (II) complexes: platinum ethylenediamine indolylphenanthroimidazole, Pt(en)PII, **5**; platinum ethylenediamine quinolylphenanthroimidazole, Pt(en)PIQ, **6**.

The additional contribution to the binding brought by the internal hydrogen bond in the PII and PIQ ligands would arise from two effects. First, the presence of an internal hydrogen bond would reduce the internal strain of the ligand and the entropic penalty upon binding, by locking the biphenyl torsion in place. Second, the increase in  $\pi$  surface area arising from stabilizing the conformation with a  $0^\circ$  torsional angle would lead to higher binding affinities when stacking on the G-quadruplex motif.



**Figure 5.12.** The biphenyl torsional angle is shown in blue.

The first effect was studied by analyzing the energy profile for the biphenyl torsion around the N-C-C-C bond (see Figure 5.12) with DFT. Calculations were performed at the B3LYP/LACV3P\*\* level of theory, scanning the torsional angle around the biphenyl moieties for PIP, PIN, PIQ and PII ligands at 5° increments. As expected, the ligands having an internal hydrogen bond exhibit a minimum in the potential energy surface for a torsion value of 0° in both cases, with a sharp increase in energy for deviations of more than 10-15°. The phenyl and naphthyl derivatives, on the other hand, have energy minimums at about 15° and 35° respectively. These energy minimums are shallower than the ones for the hydrogen bonded ligands, leading to the conformations with 0° torsional angles lying about 1.5-2.0 kcal/mol higher than the respective minimums.

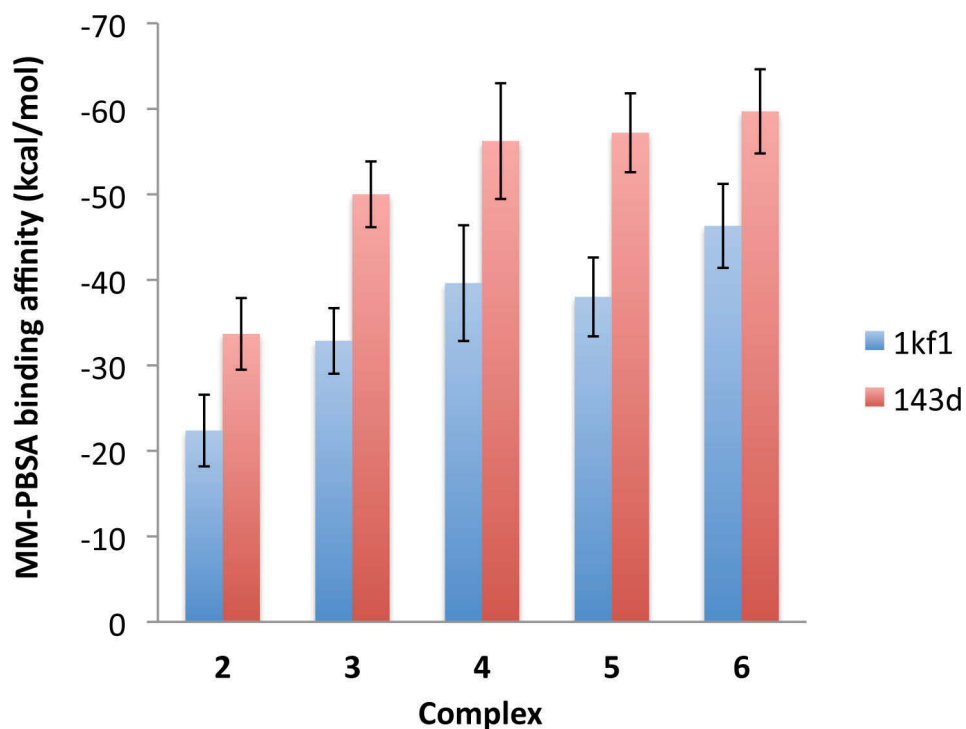
The second effect was investigated by estimating the binding affinity of the platinum (II) complexes to the G-quadruplexes, in an analogous way as performed with the previous ligands. The calculated binding affinity (see Table 5.2) was highest for Pt(PIQ)en; the value significantly higher than the one for Pt(PIP)en but at the same level as the one for Pt(PIN)en, while the PII complex seems to be equipotent with the PIN one (see Figure 5.13). The effect of the entropic contribution to the binding is unclear: a normal mode analysis of the snapshots did not exhibit significant



changes in the entropic contribution to the binding affinity for the different complexes.

**Table 5.2** MM-PBSA binding affinities for the different platinum(II) complexes to the human telomere G-quadruplex motif in a stacking binding mode. Energies in kcal/mol.

Compound #	Propeller	Basket
2	$-23.78 \pm 3.15$	$-33.68 \pm 4.19$
3	$-32.86 \pm 4.06$	$-50.00 \pm 3.84$
4	$-39.62 \pm 5.83$	$-56.23 \pm 6.77$
5	$-38.00 \pm 3.35$	$-57.20 \pm 4.61$
6	$-46.31 \pm 5.77$	$-59.70 \pm 4.92$



**Figure 5.13** Binding affinities for complexes 1-5 calculated with the MM-PBSA method. Error bars are one standard deviation.

### 5.3 Conclusions

A hybrid docking/molecular mechanics method was developed to study the binding of square planar transition metal complexes to DNA quadruplexes. As a necessity, molecular mechanical parameters compatible with the GAFF force field were

developed that enabled the FITTED docking program to place this type of compounds within the G-quadruplex and propose potential binding modes. Furthermore, the same molecular mechanical potentials allowed us to perform molecular dynamics simulations on the Pt complex/DNA systems to assess their stability.

This hybrid docking/DNA technique was used for the assessment of a platinum molecular square as a G-quadruplex binder, and found the stacking mode to be the most favoured one. When applied to a family of monometallic complexes with extended  $\pi$ -surfaces, the MM-PBSA method exposed the increase in binding affinity observed experimentally, within the error of the method.

This technique can potentially be applied to the virtual screening of libraries of complexes, thus reducing the synthetic effort required for the development of new G-quadruplex binders.

## **5.4 Methods**

### **5.4.1 Development of molecular mechanics parameters for Pt complexes**

A model of Pt(en)(bipy)<sub>2</sub> was built in the Maestro 8.0 interface (Schrödinger) and optimized at the B3LYP/LACV3P\*\* level of theory using Jaguar 7.0 (Schrödinger). Missing parameters (bond stretching, bending and torsions involving the Pt atom) from the GAFF force field were obtained running the *parmchk* module from Amber 10 on this model. Distorted structures around the equilibrium points were generated for each of the parameters that had to be defined, varying the bond distances by  $\pm 0.1$  Å at 0.01 Å intervals and angles by  $\pm 10^\circ$  at  $1^\circ$  intervals. The energy was calculated at the B3LYP/LACV3P\*\* level of theory for each distorted conformation. At the same time, RESP charges were assigned to the different conformations and their energy calculated with the available parameters within the GAFF force field. The difference between the DFT and the GAFF energies was fit to a quadratic term of the form  $K \cdot (x - x_0)$ . The resulting parameters were added to the GAFF force field as a *frcmod* file.

### 5.4.2 Docking

Models of the G-quadruplex structures were obtained from the Protein Data Bank (PDB IDs 1kf1, 143d). Hydrogens were added (at pH = 7.0), bond orders were fixed and structural cations were added if necessary (Na<sup>+</sup> in the case of 143d). In the case of 143d, a structure devoid of TTA loops was also constructed and prepared as follows. Models of the platinum(II) complexes were built in the Maestro 8.0 interface and optimized at the B3LYP/LACV3P\*\* level of theory using Jaguar 7.0. Two-stage RESP charges were assigned from the electrostatic potential (ESP) calculated by Jaguar at the B3LYP/LACV3P\*\* level of theory and the *antechamber*, *resp*gen and *resp* modules of Amber 10. The PROCESS and SMART modules of FITTED were used to prepare the G-quadruplexes and the ligands for docking, respectively. A total of 100 docking runs were performed for each ligand/G-quadruplex combination; the docked poses were clustered using XCluster (Schrödinger) and representative modes were selected by visual inspection.

### 5.4.3 Molecular dynamics simulations

The *leap* module from Amber 10 was used to assign *parm99*<sup>34,35</sup> and *parmbsc0* parameters<sup>49</sup> to the DNA G-quadruplex, and GAFF and the *ad hoc* parameters developed as described above to the platinum (II) complexes. The DNA:platinum (II) adducts were solvated by a 10 Å truncated octahedron of TIP3P water molecules and neutralized by the appropriate cation (K<sup>+</sup> for 1kf1, Na<sup>+</sup> for 143d). The solvent molecules were first relaxed by a conjugate gradient minimization, followed by a relaxation of the complete system. Heating of the system from 0K to 300K (20 ps, 1.6 fs time step) at constant volume and relaxation at constant pressure (1 atm, 100 ps, 1.6 fs time step) was followed by a 4 ns production run at constant pressure (1 atm, 1.6 fs time step).

### 5.4.4 Binding affinity calculations

Snapshots taken at 10 ps intervals from the production run processed to calculate binding affinities with the MM-PBSA method using the *mm\_pbsa.pl* scripts (relying on *sander*) included in Amber10.

## 5.5 References

1. Hanahan, D.; Weinberg, R. A., The hallmarks of cancer. *Cell* **2000**, *100*, 57-70.
2. World Health Organization Cancer. <http://www.who.int/mediacentre/factsheets/fs297/en/> (accessed 2009/04/01, 2009).
3. Wright, W. E.; Tesmer, V. M.; Huffman, K. E.; Levene, S. D.; Shay, J. W., Normal human chromosomes have long G-rich telomeric overhangs at one end. *Genes and Development* **1997**, *11*, 2801-2809.
4. McElligott, R.; Wellinger, R. J., The terminal DNA structure of mammalian chromosomes. *EMBO J.* **1997**, *16*, 3705-3714.
5. Makarov, V. L.; Hirose, Y.; Langmore, J. P., Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell* **1997**, *88*, 657-666.
6. De Lange, T., Shelterin: The protein complex that shapes and safeguards human telomeres. *Genes and Development* **2005**, *19*, 2100-2110.
7. Morin, G. B., The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. *Cell* **1989**, *59*, 521-529.
8. Shay, J. W.; Wright, W. E., Telomerase therapeutics for cancer: Challenges and new directions. *Nat. Rev. Drug Discovery* **2006**, *5*, 577-584.
9. Kim, N. W.; Piatyszek, M. A.; Prowse, K. R.; Harley, C. B.; West, M. D.; Ho, P. L. C.; Coviello, G. M.; Wright, W. E.; Weinrich, S. L.; Shay, J. W., Specific association of human telomerase activity with immortal cells and cancer. *Science* **1994**, *266*, 2011-2015.
10. Mergny, J. L.; Riou, J. F.; Mailliet, P.; Teulade-Fichou, M. P.; Gilson, E., Natural and pharmacological regulation of telomerase. *Nucleic Acids Res.* **2002**, *30*, 839-865.
11. Shin-ya, K.; Wierzba, K.; Matsuo, K.-i.; Ohtani, T.; Yamada, Y.; Furihata, K.; Hayakawa, Y.; Seto, H., Telomestatin, a Novel Telomerase Inhibitor from *Streptomyces anulatus*. *J. Am. Chem. Soc.* **2001**, *123*, 1262-1263.
12. Davis, J. T., G-Quartets 40 Years Later: From 5'-GMP to Molecular Biology and Supramolecular Chemistry. *Angew. Chem. Int. Ed.* **2004**, *43*, 668-698.

13. Simonsson, T., G-quadruplex DNA structures - Variations on a theme. *Biol. Chem.* **2001**, *382*, 621-628.
14. Webba da Silva, M., Geometric Formalism for DNA Quadruplex Folding. *Chem. Eur. J.* **2007**, *13*, 9738-9745.
15. Wang, Y.; Patel, D. J., Solution structure of the human telomeric repeat d[AG<sub>3</sub>(T<sub>2</sub>AG<sub>3</sub>)<sub>3</sub>] G-tetraplex. *Structure* **1993**, *1*, 263-282.
16. Parkinson, G. N.; Lee, M. P. H.; Neidle, S., Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **2002**, *417*, 876-880.
17. Phan, A. T.; Kuryavyi, V.; Luu, K. N.; Patel, D. J., Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K<sup>+</sup> solution. *Nucleic Acids Res.* **2007**, *35*, 6517-6525.
18. Zahler, A. M.; Williamson, J. R.; Cech, T. R.; Prescott, D. M., Inhibition of telomerase by G-quartet DNA structures. *Nature* **1991**, *350*, 718-720.
19. Sun, D.; Thompson, B.; Cathers, B. E.; Salazar, M.; Kerwin, S. M.; Trent, J. O.; Jenkins, T. C.; Neidle, S.; Hurley, L. H., Inhibition of Human Telomerase by a G-Quadruplex-Interactive Compound. *J. Med. Chem.* **1997**, *40*, 2113-2116.
20. Kim, M. Y.; Vankayalapati, H.; Shin-Ya, K.; Wierzba, K.; Hurley, L. H., Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular G-quadruplex. *J. Am. Chem. Soc.* **2002**, *124*, 2098-2099.
21. Harrison, R. J.; Cuesta, J.; Chessari, G.; Read, M. A.; Basra, S. K.; Reszka, A. P.; Morrell, J.; Gowan, S. M.; Incles, C. M.; Tanious, F. A.; Wilson, W. D.; Kelland, L. R.; Neidle, S., Trisubstituted acridine derivatives as potent and selective telomerase inhibitors. *J. Med. Chem.* **2003**, *46*, 4463-4476.
22. Haq, I.; Trent, J. O.; Chowdhry, B. Z.; Jenkins, T. C., Intercalative G-tetraplex stabilization of telomeric DNA by a cationic porphyrin. *J. Am. Chem. Soc.* **1999**, *121*, 1768-1779.
23. Perry, P. J.; Gowan, S. M.; Reszka, A. P.; Polucci, P.; Jenkins, T. C.; Kelland, L. R.; Neidle, S., 1,4- and 2,6-disubstituted amidoanthracene-9,10-dione derivatives as inhibitors of human telomerase. *J. Med. Chem.* **1998**, *41*, 3253-3260.

24. De Cian, A.; DeLemos, E.; Mergny, J. L.; Teulade-Fichou, M. P.; Monchaud, D., Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J. Am. Chem. Soc.* **2007**, *129*, 1856-1857.
25. Spöner, J.; Spackova, N., Molecular dynamics simulations and their application to four-stranded DNA. *Methods* **2007**, *43*, 278-290.
26. Read, M. A.; Wood, A. A.; Harrison, J. R.; Gowan, S. M.; Kelland, L. R.; Dosanjh, H. S.; Neidle, S., Molecular modeling studies on G-quadruplex complexes of telomerase inhibitors: Structure-activity relationships. *J. Med. Chem.* **1999**, *42*, 4538-4546.
27. Ladame, S.; Schouten, J. A.; Stuart, J.; Roldan, J.; Neidle, S.; Balasubramanian, S., Tetrapeptides induce selective recognition for G-quadruplexes when conjugated to a DNA-binding platform. *Organic and Biomolecular Chemistry* **2004**, *2*, 2925-2931.
28. Reed, J. E.; Arnal, A. A.; Neidle, S.; Vilar, R., Stabilization of G-quadruplex DNA and inhibition of telomerase activity by square-planar nickel(II) complexes. *J. Am. Chem. Soc.* **2006**, *128*, 5992-5993.
29. Haider, S. M.; Parkinson, G. N.; Neidle, S., Structure of a G-quadruplex-Ligand Complex. *J. Mol. Biol.* **2003**, *326*, 117-125.
30. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Bioinf.* **1997**, *29*, 215-220.
31. Ma, D.-L.; Lai, T.-S.; Chan, F.-Y.; Chung, W.-H.; Abagyan, R.; Leung, Y.-C.; Wong, K.-Y., Discovery of a Drug-Like G-Quadruplex Binding Ligand by High-Throughput Docking. *ChemMedChem* **2008**, *3*, 881-884.
32. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.
33. Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F., A platinum supramolecular square structure as a G-quadruplex interactive agent. *J. Am. Chem. Soc.* **2008**, *130*, 10040-10041.

34. Cheatham III, T. E.; Cieplak, P.; Kollman, P. A., A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics* **1999**, *16*, 845-862.
35. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
36. Hay, P. J.; Wadt, W. R., Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82*, 270-283.
37. Wadt, W. R.; Hay, P. J., Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *J. Chem. Phys.* **1985**, *82*, 284-298.
38. Hay, P. J.; Wadt, W. R., Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **1985**, *82*, 299-310.
39. 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.
40. Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N., Evaluation of Docking Programs for Predicting Binding of Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160-176.
41. Kieltyka, R.; Fakhoury, J.; Moitessier, N.; Sleiman, Hanadi F., Platinum Phenanthroimidazole Complexes as G-Quadruplex DNA Selective Binders. *Chem. Eur. J.* **2007**, *14*, 1145-1154.
42. Delano, W. L. *The PyMOL Molecular Graphics System*, Delano Scientific, LLC: San Carlos, CA, USA, 2002. <http://www.pymol.org>
43. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O., Calculating structures and free energies of

complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.

44. Shirude, P. S.; Gillies, E. R.; Ladame, S.; Godde, F.; Shin-ya, K.; Huc, I.; Balasubramanian, S., Macrocyclic and helical oligoamides as a new class of G-quadruplex ligands. *J. Am. Chem. Soc.* **2007**, *129*, 11890-11891.

45. Cusumano, M.; Di Pietro, M. L.; Giannetto, A., Stacking surface effect in the DNA intercalation of some polypyridine platinum(II) complexes. *Inorg. Chem.* **1999**, *38*, 1754-1758.

46. Che, C. M.; Yang, M.; Wong, K. H.; Chan, H. L.; Lam, W., Platinum(II) complexes of dipyridophenazine as metallointercalators for DNA and potent cytotoxic agents against carcinoma cell lines. *Chem. Eur. J.* **1999**, *5*, 3350-3356.

47. Kieltyka, R.; Fakhoury, J.; Moitessier, N.; Sleiman, Hanadi F., Platinum Phenanthroimidazole Complexes as G-Quadruplex DNA Selective Binders. *Chem. Eur. J.* **2008**, *14*, 1145-1154.

48. Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F., A Platinum Supramolecular Square as an Effective G-Quadruplex Binder and Telomerase Inhibitor. *Journal of the American Chemical Society* **2008**, *130*, 10040-10041.

49. Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Lughton, C. A.; Orozco, M., Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of  $\alpha/\gamma$  Conformers. *Biophys. J.* **2007**, *92*, 3817-3829.



[ This page was intentionally left blank ]

## **Chapter 6: Development of programs for the handling of ligands for virtual screening: SMART, REACTOR and SELECT**

### **6.1 Introduction**

#### **6.1.1 Ligand treatment for docking**

As was mentioned in Chapter 1, docking methods attempt to predict the binding mode of a small molecule within a biomacromolecular receptor, and provide insight on the relative binding affinities of a group of compounds. In order for a docking program to satisfactorily handle the ligands, the correct representation needs to be used in terms of atom types and charges. Additionally, the conformational sampling of ligands during the docking requires the definition of which torsional degrees of freedom are to be scanned, i.e., defining which bonds are *rotatable*.

#### **6.1.2 Molecular mechanical force fields**

The multiple molecular mechanical force fields described in the literature and in common use in structure-based drug design are well parameterized for biomacromolecules (proteins and nucleic acids),<sup>1,2</sup> or for organic molecules,<sup>3</sup> but few have been designed to be of general applicability for both types of molecules. For the docking of small molecules to proteins and nucleic acids the recently developed generalized Amber force field (GAFF)<sup>4</sup> is particularly attractive as an extension of the Amber parameters for biomolecules in use (with modifications) in the last 25 years,<sup>2,5,6</sup> and was therefore chosen for the description of ligands within the FITTED docking program.<sup>7</sup>

#### **6.1.3 Atomic charges**

The treatment of electrostatics in molecular mechanics requires the assignment of point charges on each atom. Charges can be assigned by different strategies, with variable computational requirements. The most CPU-intensive method (but also the golden standard), RESP or *Restrained*

*ElectroStatic Potential fitting*<sup>8</sup> requires an optimization of the structure at the HF/6-31G\* level of theory, followed by a calculation of the electrostatic potential (ESP) around the molecule and further fitting of atomic point charges to reproduce the ESP. This method is impracticable for large virtual libraries containing hundreds of thousands of compounds, due to the vast amount of CPU time that would be needed. An intermediate method, AM1-bcc,<sup>9</sup> has been reported to reproduce RESP charges with good fidelity,<sup>10,11</sup> requiring a semiempirical calculation with the AM1 set of parameters and a further adjustment. With a much lower computational requirement, methods relying on the equalization of atomic electronegativities have been established in the 1980's, and are widely applied to drug design due to their simplicity and speed.<sup>12</sup> On the fastest end of the spectrum, methods parameterized from the analysis of multiple *ab initio*-derived parameters can be used to quickly establish a set of charges, although in some cases they may disregard electronic effects that alter the charge distribution.<sup>13</sup>

#### **6.1.4 Virtual libraries of ligands**

As part of a structure-based virtual screening (VS) workflow, one needs to provide the computational method with 3D structures for the ligands under study. The 3D coordinates can be obtained either i) by manually drawing out the structures in a graphical interface, ii) by using 2D to 3D conversion tools<sup>14</sup> or iii) from ligand collections, either public repositories<sup>15-17</sup> or corporate databases. The three alternatives have their own drawbacks: i) can become prohibitively time-consuming for large libraries; ii) still requires the exact 2D representation (e.g., SMILES string) of each member of the library; iii) limits the library to available structures. Besides these considerations, a library containing only currently available compounds limits the coverage of chemical space being screened.

#### **6.1.5 Combinatorial libraries**

Especially for the case of lead optimization, there is interest in exploring Markush structures<sup>18</sup> where one or several regions of a molecule can be

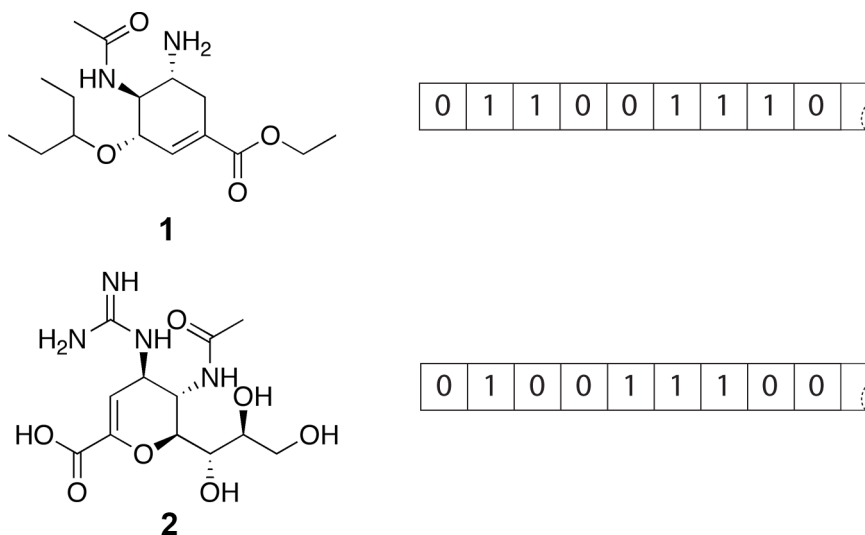
varied among different possible groups. Traditionally, this could be attempted by combinatorial synthesis, yielding libraries of related compounds that could be assayed for activity against a specific receptor. Analogously, a virtual library of compounds could be assayed by computer-aided drug design techniques; this library could be assembled from the combination of different fragment libraries.

#### **6.1.6 Virtual SAR (VSAR)**

Once a hit compound is found by screening techniques, it is common practice to probe plausible substitution sites in order to increase the activity of the compound, yielding a structure-activity relationship (SAR). Given a hit from a VS campaign, a virtual SAR could be obtained by applying computational methods to a library of derivatives of the hit. It has been observed that the scoring functions used in docking methods are not sensitive enough for this type of experiments;<sup>19</sup> however other higher-level techniques (e.g., LIE,<sup>20</sup> FEP<sup>21</sup>) for the prediction of binding affinity could be applied to this problem of hit-to-lead optimization.

#### **6.1.7 Fingerprints and similarity metrics**

When handling large libraries of ligands, it is necessary to be able to assess how similar their member compounds are. The inherent complexity of chemical structures requires the simplification of the molecular structure in order to make the search more efficient, especially when handling large libraries of molecules.<sup>22,23</sup> The most common simplification is the construction of *fingerprints* by analyzing the presence or absence of functional groups or substructures. This leads to a sequence of “0”s and “1”s (Figure 6.1) that globally can be used to evaluate the pairwise similarity of compounds by different metrics. Of the latter, the most commonly used for chemical similarity is the Tanimoto coefficient (Equation 6.1).



**Figure 6.1.** Molecular fingerprints. The two-dimensional structures of neuraminidase inhibitors oseltamivir **1** and zanamivir **2** are converted into a fingerprint based on their molecular features.

**Equation 6.1.** Tanimoto coefficient.  $T(A,B)$  is the Tanimoto similarity between A and B, that is, the ratio between the bits that are “on” in both A and B ( $A \cap B$ ) and the total number of bits ( $A \cup B$ ). In the third term,  $\chi_{iA}$  is the i-th bit of item A ( $\chi=0$  if off,  $\chi=1$  if on), and analogously for  $\chi_{iB}$ .

$$T(A,B) = \frac{A \cap B}{A \cup B} = \frac{\sum \chi_{iA} \chi_{iB}}{\sum \chi_{iA} + \sum \chi_{iB} + \sum \chi_{iA} \chi_{iB}}$$

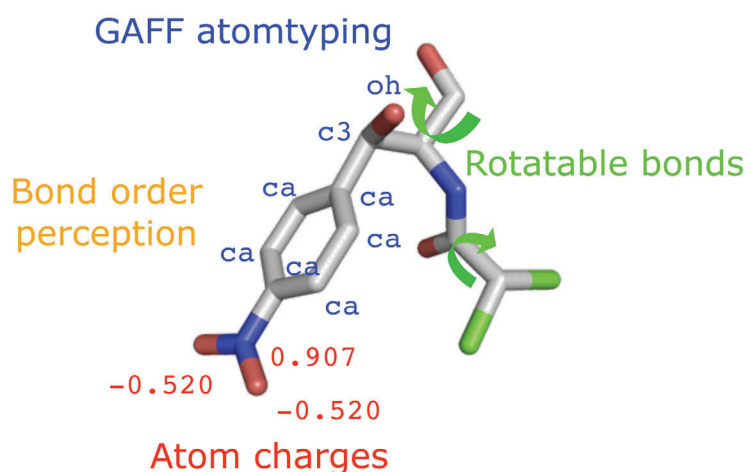
The values of the Tanimoto coefficient vary, therefore, between 0 (i.e., there are no similarities in the features considered between the two molecules) and 1 (the same features are present in both compounds, although they may still be not identical).

## 6.1.8 Goals

### 6.1.8.1 SMART

The goal of this project was to construct a module for the preparation of libraries of ligands to perform virtual screening on either FITTED (docking)<sup>7,24</sup> or ACE (asymmetric catalyst evaluation).<sup>24</sup> The resulting module was termed SMART (*Small Molecule Atomtyping and Rotatable Torsion assignment*) (see Figure 6.2). Separating the pre-processing of a library of ligands from the docking can save time in a VS campaign, as the operations are deterministic

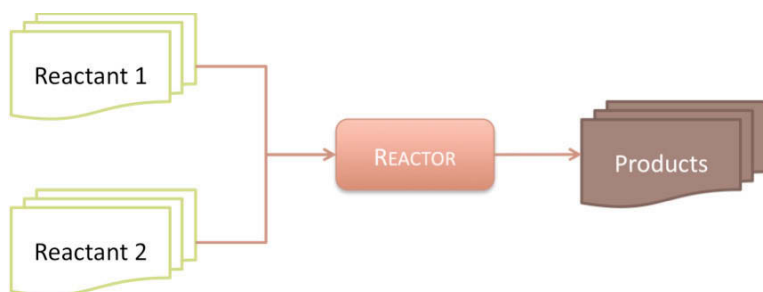
(i.e., with a unique and defined result, not dependent of a random number generator) and therefore can be performed just once, while the library could potentially be reused for different screens on the same or different targets. Additionally to the atom typing, a system to assign atomic point charges was implemented, based on the Merck Molecular Force Field (MMFF).<sup>3</sup>



**Figure 6.2.** Ligand properties assigned by SMART.

#### 6.1.8.2 REACTOR

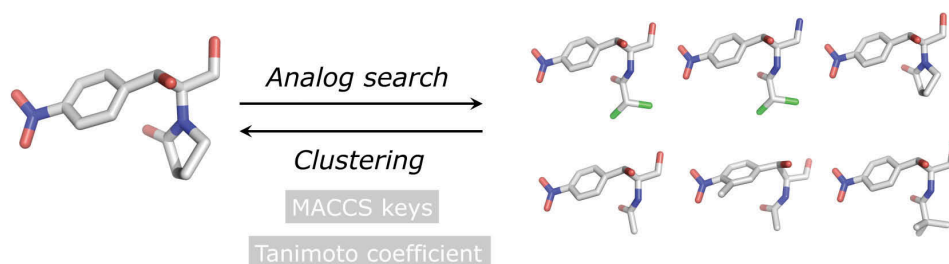
With the framework for the description of ligands laid down for SMART, we looked at constructing a program that would provide a virtual library of compounds when given a pair of virtual libraries of reactants and a set of rules for their conversion into products as input (Figure 6.3). With the application of structure-based methods in mind, the representation of input and output molecules was to be done in 3D coordinates, therefore yielding a library of compounds that would be ready for docking with FITTED or for further calculations with any other structure-based drug design tool.



**Figure 6.3** Information flow in REACTOR.

### 6.1.8.3 SELECT

We wanted to construct a program to handle virtual libraries of ligands in order to cluster ligands by similarity and to extract ligands resembling a query molecule (Figure 6.4, right-to-left and left-to-right respectively). Both tasks rely on the measurement of chemical similarity metrics, and their interest rely on the principle that similar molecules exhibit similar properties.<sup>22</sup> Based on this, one might argue that a screening performed on a subset of a library containing only one molecule from each similarity-derived cluster would yield as much information as screening the whole library, but at a fraction of the computational time. By the same token, if a given compound exhibits a certain biological activity, compounds similar to it are more likely than others to share that trait and also be active.



**Figure 6.4.** SELECT. Left to right: given a query molecule and a Tanimoto cutoff, the program can retrieve a subset of compounds from a library. Right to left: given a library and a Tanimoto cutoff, SELECT can extract a representative library of compounds with increased diversity.

## 6.2 Methods

The entire codes of SMART, REACTOR and SELECT were written in C++, and tested to work with *gcc* (Windows, Linux, Mac OS X) and Microsoft compilers. The following sections describe the different portions of the program.

### 6.2.1 SMART

#### 6.2.1.1 GAFF atom typing

The atom types defined in the generalized Amber force field<sup>4</sup> are specified in Table 6.1. The atom type assignment is performed on a per-element basis, and for each element the connectivity and chemical environment is defined so as to unequivocally assign an atom type to each atom. Some atom types defined in the original GAFF description are not used in the SMART assignment, namely the cc/cd and ce/cf types for conjugated systems (and their N and P counterparts), as well as the h1-h5 types for hydrogens on carbons with electron-withdrawing groups.

**Table 6.1.** GAFF atom types defined by SMART.

Element	Atom type	Description
C	c	sp <sup>2</sup> C bound to heteroatom (C=O, C=S)
	c1	sp C
	c2	aliphatic sp <sup>2</sup> C
	c3	sp <sup>3</sup> C
	ca	aromatic sp <sup>2</sup> C
	cp/cq	biphenyl bridging C
	cu	sp <sup>2</sup> C in three-membered rings
	cv	sp <sup>3</sup> C in three-membered rings
	cx	sp <sup>2</sup> C in four-membered rings
	cy	sp <sup>3</sup> C in four-membered rings
N	n	sp <sup>2</sup> N in amides
	n1	sp N
	n2	sp <sup>2</sup> N with 2 substituents
	n3	sp <sup>3</sup> N with 3 substituents
	n4	sp <sup>3</sup> N with 4 substituents
	na	sp <sup>2</sup> N with 3 substituents
	nh	N connected to aromatic ring (e.g., aniline N)
	no	N in nitro groups
	nb	aromatic N
O	o	sp <sup>2</sup> O in carbonyl, carboxylate, etc.
	oh	sp <sup>3</sup> O in hydroxyls



	os	sp <sup>3</sup> O in ethers and esters
S	s2	sp <sup>2</sup> S (e.g., thiocarbonyl)
	sh	sp <sup>3</sup> S in thiol groups
	ss	sp <sup>3</sup> S in thioether and disulfide groups
	s4	hypervalent S with 3 substituents
	s6	hypervalent S with 4 substituents
P	p3	sp <sup>3</sup> P with 3 substituents
	p4	hypervalent P with 3 substituents
	p5	hypervalent P with 4 substituents
H	hc	H on aliphatic C
	ha	H on aromatic C
	hn	H on N
	ho	H on O
	hs	H on S
	hp	H on P
F	f	fluorine
Cl	cl	chlorine
Br	br	bromine
I	i	iodine

#### 6.2.1.2 MMFF charge assignment

The assignment of point charges with the MMFF forcefield is a three-step process, requiring first the assignment of MMFF atom types (analogous to the other force fields described above),<sup>3</sup> followed by the assignment of bond-charge increments to each pair of atoms,<sup>13</sup> and finally the spreading of the formal charge arising from resonance structures.<sup>25</sup> Bond-charge increments are defined for select pairs of atom types bonded together,<sup>13</sup> while a general mechanism for deriving bond-charge increments relies on the combination of partial bond increments defined for an individual atom type.<sup>25</sup>

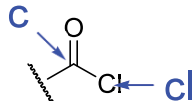
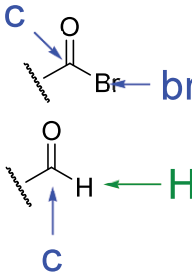
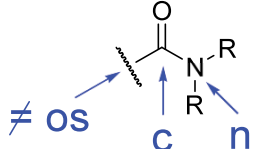
#### 6.2.1.3 Cycle perception

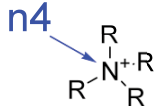
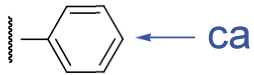
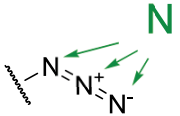
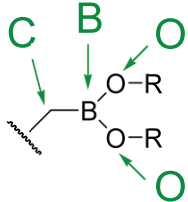
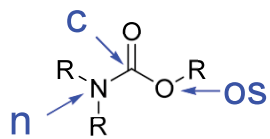
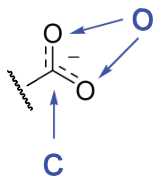
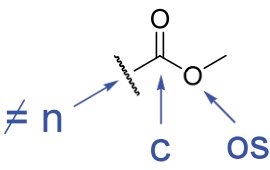
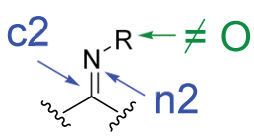
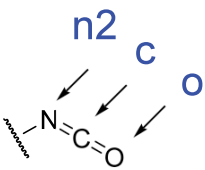
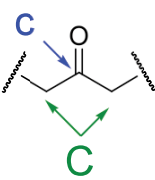
Cyclic groups have to be specially recognized, as the conformational sampling of rings requires a different algorithm for it to work.<sup>26</sup> To this effect, a graph of each molecule is constructed and scanned for connected paths via a depth-first search algorithm. Larger rings are then split into smaller rings if necessary, in order to retrieve the smallest set of smallest rings (SSSR) via an algorithm inspired by the one published by Fan *et al.*<sup>27</sup>

#### 6.2.1.4 Toxicophores and functional groups definitions

Compounds containing potentially toxic functionalities or structures that would not match a pharmacophore in the target should be filtered out as early as possible in a screening campaign.<sup>28</sup> To facilitate this, we implemented a mechanism to identify a series of functional groups as toxicophores, potentially not acceptable in a library for virtual screening. A bitstring identifying the presence of functional groups is appended to the description of a molecule, allowing for the filtering criteria to be defined by the user in a FITTED docking run. In addition to filtering by *presence* of a toxicophore, compounds can be filtered out if they do *not* contain features (i.e., filtering by *absence*) known to be necessary for binding to the target (or other desirable properties). For example, one might be interested in looking for compounds able to chelate a metal atom through a sulfonamide, but not having an aldehyde present. The groups defined as potential toxicophores are defined in Table 6.2.

**Table 6.2.** List of functional groups recognized by SMART. In the figures, blue lowercase indicates a GAFF atom type, while a green uppercase label denotes an element.

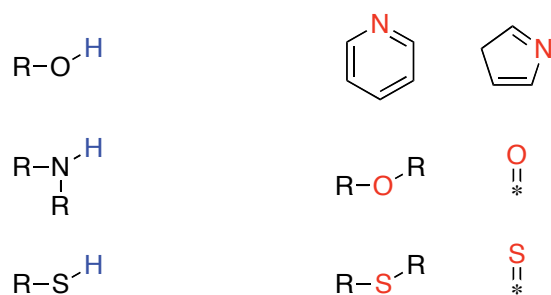
Functional group	Description
Acyl chloride	 <p>an atom of type c is bound to an atom of type cl or br</p>
Aldehyde	 <p>an atom of type c is bound to a hydrogen atom</p>
Amide	 <p>an atom of type c is bound to an atom with of type n, and not bound to an atom of type os; both c and n atoms are acyclic</p>

Ammonium		an atom with an n4 atom type is present in the molecule
Aromatic		an atom of type ca is present in the molecule
Azide		three acyclic nitrogen atoms in a linear arrangement are present in the molecule
Boronate		a boron atom is bound to a carbon and two oxygen atoms
Carbamate		an atom of type c is bound to an atom of type n and an atom of type os
Carboxylic acid		an atom of type c is bound to two atoms of type o
Ester		an atom of type c is bound to an atom of type os, but not bound to an atom of type n; both c and os atoms are acyclic
Imine		an atom of type c2 is bound to an atom of type n2, both acyclic; the nitrogen cannot be bound to oxygen
Isocyanate		an atom of type c is bound to one atom of type n2 and another of type o, where the c-n2 bond is acyclic.
Ketone		an atom of type c is bound to 2 carbon atoms

Lactame		an atom of type c is bound to an atom of type n, but not bound to an atom of type os; with both c and n atoms being cyclic
Lactone		an atom of type c is bound to an atom of type os, but not bound to an atom of type n; with both c and os atoms being involved in a ring
Michael acceptor		an atom of type c2 is bound to either 1) an atom with a c atom type which is not a carboxylate, or 2) a nitrile group; the bond between c2 and c/c1 must be acyclic
Nitrile		an atom of type c1 is bound to an atom of type n1
Nitro		an atom of type no is present in the molecule
Oxime		an atom of type c2 type is bound to an atom of type n2, which in turn is bound to an oxygen atom
Primary amine		an atom of type n3 is bound to two hydrogens
Secondary amine		an atom of type n3 is bound to a single hydrogen
Sulphonamide		an atom of type s6 is bound to an atom of type n

### 6.2.1.5 Lipinski's rules

In a seminal paper in the ADME-Tox field, Lipinski and co-workers analyzed a large number of marketed drugs and established a number of criteria a compound needs to meet in order to have good chances of being an orally available drug (i.e., to have good enough permeation and absorption properties).<sup>29</sup> These criteria are given in terms of molecular weight less than or equal to 500, number of hydrogen bond acceptors (HBA) less than or equal to ten, number of hydrogen bond donors (HBD) less than or equal to five, and calculated log P less than or equal to 5.0.<sup>29</sup> SMART calculates the molecular weight and counts the HBA/HBD groups. A HBA group is defined as an electronegative atom with a lone pair of electrons such as an  $sp^2$  oxygen (e.g., in a carbonyl or a sulfone group), a hydroxyl oxygen (e.g., an alcohol) or an  $sp^2$  nitrogen with a free lone pair (such as a N in pyridine). On the other hand, a HBD group requires a hydrogen atom attached to an electronegative atom such as oxygen, sulphur or nitrogen.

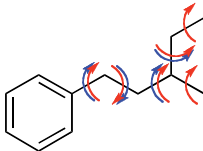


**Figure 6.5.** Hydrogen bond donor and acceptor groups. Left: hydrogen bond donor atoms shown in blue; right: hydrogen bond acceptor atoms coloured red. R groups can be hydrogen, alkyl, aryl or acyl, among other possibilities. Double-bonded O and S can be attached (\*) to C, S, N or P.

### 6.2.1.6 Assignment of rotatable bonds

The assignment of rotatable bonds is two-fold. On one hand, the bonds that will be considered rotatable for the conformational sampling during the docking need to be identified; on the other hand, the total number of rotatable bonds is a factor in the scoring of docked poses (see Chapters 3 and 4). The main difference between the two counts is that symmetric or

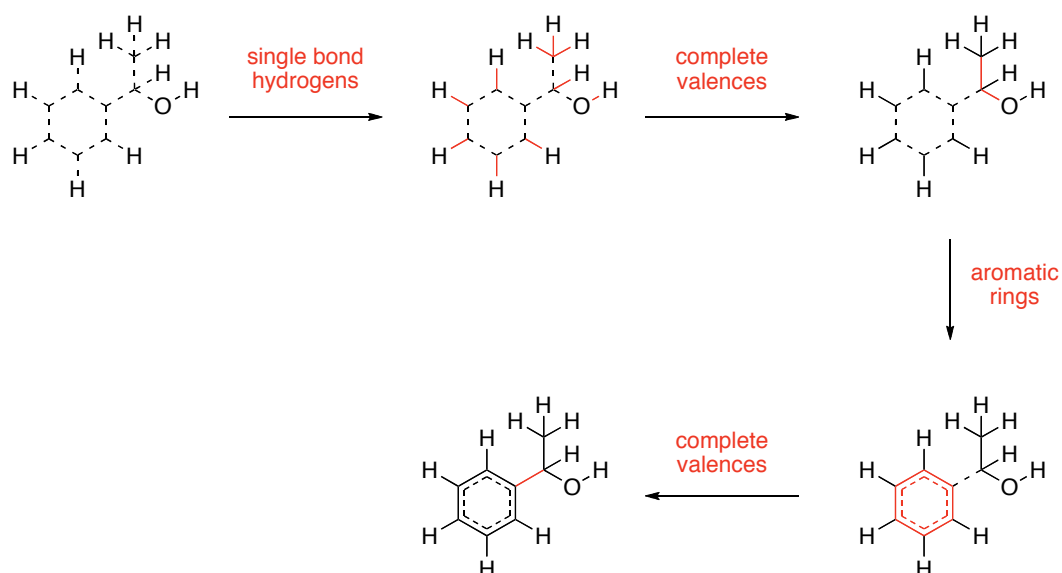
terminal groups need to be considered rotatable during the docking (for an optimal fit), however as a rotation around that bond would make an indistinguishable molecule, it is not counted for scoring purposes.



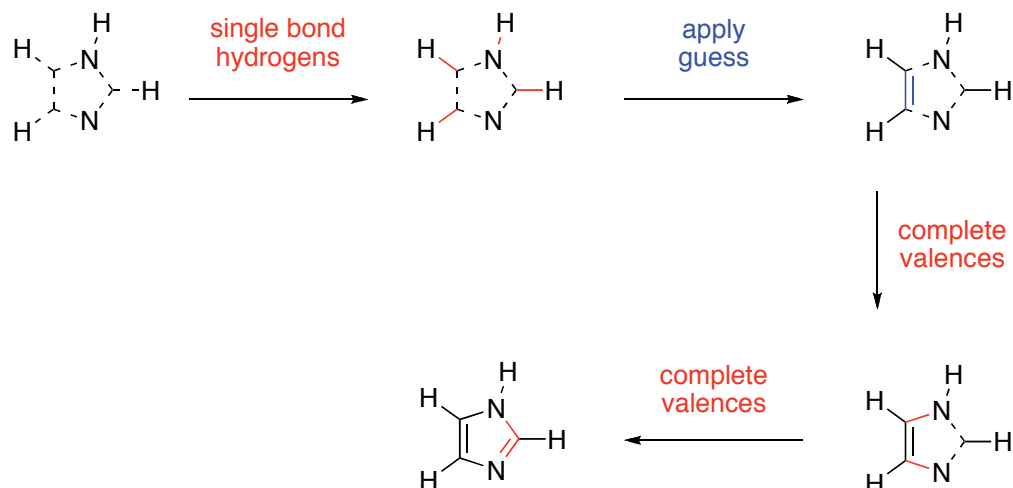
**Figure 6.6.** Rotatable bond assignment. Rotatable bonds assigned for docking are shown with red arrows; bonds considered rotatable for scoring are shown with blue arrows.

#### 6.2.1.7 Bond order assignment

One of the more challenging topics is the correct assignment of the bond orders of a ligand. This is relatively simple for a compound containing only carbon, hydrogen (or halogens) and oxygen atoms, as the valence (i.e., the sum of the bond orders) is unique for each element. However, problems arise when considering nitrogen atoms, as they can equally attain valences of 3 (in a neutral form, as in amines, amides, pyridines or imines) or 4 (in a cationic form, such as in ammoniums, guanidiniums or pyridiniums). A mechanism inspired by the one used in the *antechamber* program was implemented to assign bond orders (Figure 6.7).<sup>30</sup> A first round of assignment takes care of the obvious single bonds with elements having single valences (i.e., hydrogen, halogens). At this point, the order of bonds involving carbons with all bond orders but one defined can be unequivocally assigned; otherwise, pre-defined functional groups (e.g., carbonyl/carboxyl derivatives, aromatic cycles) are assigned and a new round is attempted. For some groups (especially heterocycles), this does not allow for an unequivocal assignment, therefore a bond order guess leading to multiple possible resonance structures are attempted in order to find a most reasonable one (Figure 6.8). In these cases, multiple attempts at defining a resonance structure are performed, and the most reasonable one is kept.



**Figure 6.7.** Bond order assignment in a molecule containing only C, H and O atoms. Bonds with (yet) unassigned order are pictured with dashed bonds, bonds assigned after each step are coloured red.



**Figure 6.8.** Bond order assignment in a heterocycle containing nitrogen atoms. The same colour convention as in previous figure applies, and in addition a bond order guess is coloured blue.

## 6.2.2 REACTOR

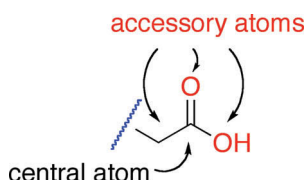
### 6.2.2.1 Encoding of chemical transformations

Chemical transformations can be described by different parameters. The connectivity of each atom in the reactant molecules can remain invariant (e.g., peptide bond, substitutions, metathesis), or it can change (e.g., additions, reductions/oxidations). The number of bonds formed upon the

transformation is usually one, but could also be two (e.g., Diels-Alder and other cycloadditions), or multiple (cascading or tandem reactions). The number of reactants involved in the transformation can be one (e.g., elimination, reduction, oxidation), two (e.g., substitutions, additions), or more (e.g., Ugi reaction, multi-component couplings). Given that most commonly combinatorial libraries are assembled piece-wise, with highly predictable addition or substitution reactions, we first focused our attention in reactions between two reactants, with no change in connectivity of the reacting centers with one new bond formed.

#### 6.2.2.2 Functional group recognition

For any transformation to take place, the site for the reaction needs to be identified. To this effect, a reaction center is defined as a *functional group*, consisting of a central atom bound to accessory atoms. This recognition can be done by GAFF atom type, by atom name or by element.



**Figure 6.9.** Functional group depiction: a central atom is bound to accessory atoms.

The definition of functional groups is stored in a text file under keyword/rules\_definitions.txt, and it is user-editable. Figure 6.10 shows an example of rules\_definitions.txt. The first column specifies the name of the functional group to be defined, which is then referred to in the definition of a transformation (*vide infra*). The second column defines the central atom; by default it expects GAFF atom types, but atom names can be specified preceding the name by an “@” sign, and elements by preceding them with an “&” character. The third column specifies the accessory atoms required, following the same rules as for the central atom. The fourth column specifies atoms that need not be present as accessory atoms (once again, following the conventions for the two previous columns); this feature is



useful for defining specific groups that otherwise fall to a default one, e.g., for aliphatic or aromatic versions of specific functional groups.

#funct_group	ctr_atom	acc_atoms	no_atoms
#-----	-----	-----	-----
carboxyl	c	o,oh	-
carboxylate	c	o,o	-
ester	c	o,os	-
acyl_chloride	c	o,cl	-
acyl_bromide	c	o,br	-
alkyl_chloride	c3	cl	-
alkyl_bromide	c3	br	-
alkyl_iodide	c3	i	-
aryl_chloride	ca	cl	-
aryl_bromide	ca	br	-
aryl_iodide	ca	i	-
anhydride	os	c,c	-
amine_1	n3	hn,hn,c3	-
ammonium_1	n4	hn,hn,hn,c3	-
amine_2	n3	hn,c3,c3	-
ammonium_2	n4	hn,hn,c3,c3	-
amine_3	n3	c3,c3,c3	-
ammonium_3	n4	hn,c3,c3,c3	-
ammonium_4	n4	c3,c3,c3,c3	-
aldehyde	c	o,hc	-
boron_acid_ar	b	ca,oh,oh	-
boron_ester_ar	b	ca,os,os	-
boron_acid_al	b	c3,oh,oh	-
boron_ester_al	b	c3,os,os	-
boron_acid_vi	b	c2,oh,oh	-
boron_ester_vi	b	c2,os,os	-
ketone	c	o,c3,c3	-
sulfonyl_cl	s6	o,o,cl	-
sulfonyl_br	s6	o,o,br	-
vinyl_chloride	c2	cl	-
vinyl_bromide	c2	br	-
vinyl_iodide	c2	i	-

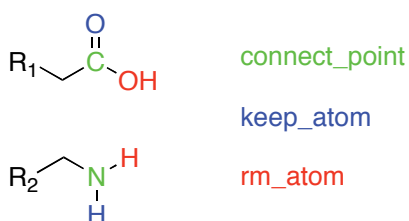
Figure 6.10. Sample rules\_definitions.txt file.

### 6.2.2.3 Specification of a transformation

A transformation is defined between a functional group in the first reagent and a corresponding one in the second reagent. A *connecting point* is defined in each functional group (which may or may not be the central atom as defined in the functional group specification), and the new bond is made

between them. Out of the atoms bound to the connecting point, atoms to be kept (*keep\_atom*) and to be removed (*rm\_atom*) are specified (Figure 6.11).

#rule	funct_group	cp	keep_atom	rm_atom	new_atom
#----	-----	--	-----	-----	-----
Rule1	carboxyl	c	o	oh	-
<b>Rule1</b>	<b>carboxylate</b>	<b>c</b>	<b>o</b>	<b>o</b>	-
Rule1	ester	c	o	os	-
<b>Rule2</b>	<b>amine_1</b>	<b>n3</b>	<b>hn,c3</b>	<b>hn</b>	-
Rule2	ammonium_1	n4	hn,c3	hn,hn	-
Rule2	amine_2	n3	hn,c3,c3	hn	-
Rule2	ammonium_2	n4	hn,c3,c3	hn,hn	-



**Figure 6.11.** Example of a rule for peptide bond formation. Top: content of rules.txt for a peptide bond formation (in bold, rules matched by bottom structures). Bottom: schematic representation of compounds matching Rule1 and Rule2, with atoms colour-coded for their function in the rule.

REACTOR includes a set of defined reactions that were used during the development and application of the program under keyword/rules.txt, although a different location for the rules can be specified. This becomes particularly useful for the generation of multi-step combinatorial libraries, where each step corresponds to a different rule, and each step can be encoded in a separate file (such as step1.txt, step2.txt, etc.).

In the case of transformations where dummy atoms are specified in each reactant molecule, it might not be practical to define a functional group for each case. For these cases, a quick definition of a functional group can be specified by preceding the name of the functional group by “!”; this instructs the program to not attempt to match the name of the functional group to a definition in rules\_definition.txt.

```

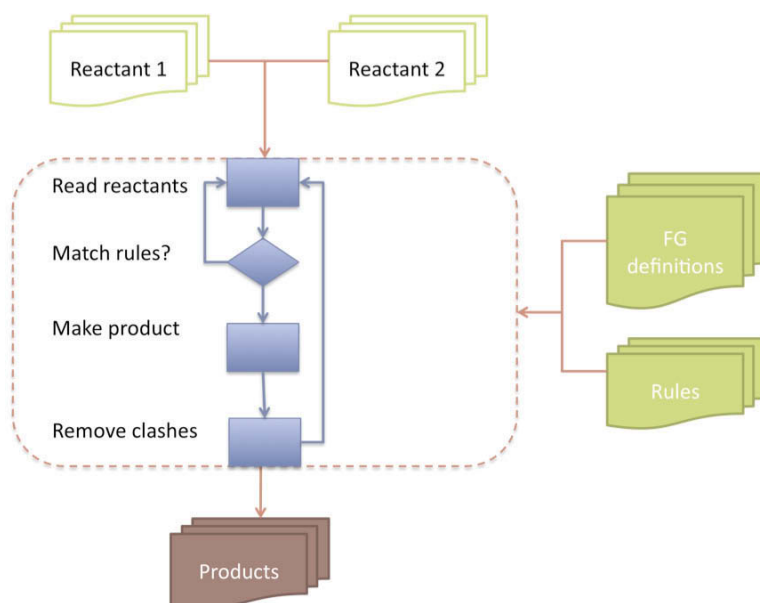
#rule funct_group      cp      keep_atom  rm_atom      new_atom
#-----
Rule1 !X1              *       -           @X1           -
Rule2 !X2              *       -           @X2           -

```

**Figure 6.12.** Example for a rule specified with dummy atoms. content of rules.txt for formation of a bond between fragments with X1 and X2 dummy atoms respectively. Bottom: schematic representation of compounds matching Rule1 and Rule2, with atoms colour-coded for their function in the rule; \* specifies any atom.

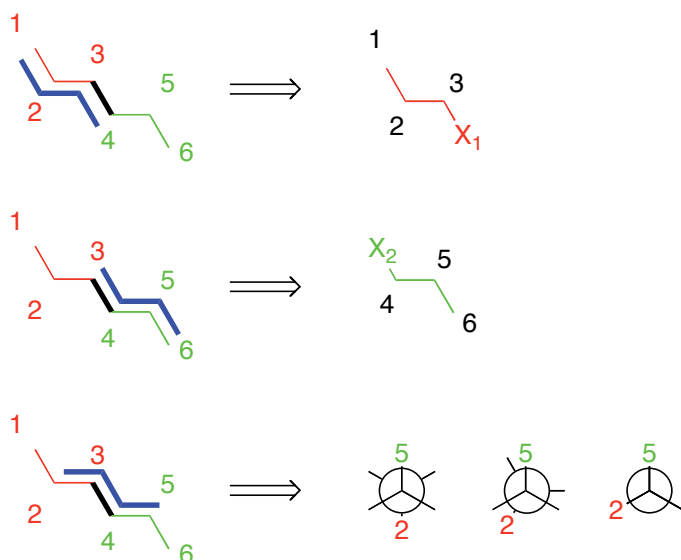
#### 6.2.2.4 Conversion of reactants into products

The cycle of the program is described in Figure 6.13. First, the program reads the description files (functional group definition and reaction rules), to then sequentially read the libraries provided for reactant 1 and reactant 2. For every reactant, a match of one of the specified rules is attempted; if the compound does not match any of the rules, the reactant is skipped. For every pair of successfully matched reactants, the transformation is performed: atoms to be removed are stripped and a new bond between the connecting points is made.



**Figure 6.13.** Schematic of the algorithm of REACTOR.

At this point, a rapid optimization of the torsions around the bond is done in order to remove the clashes that may arise as a consequence of the two groups coming together (Figure 6.14). The hybridization of both connecting points defines how this search is performed: connections between two  $sp^3$ -hybridized atoms require a search every  $30^\circ$ ; if one of the atoms is  $sp^2$ -hybridized the search happens every  $60^\circ$  degrees. These searches are performed in a way so as to ensure that the configuration around chiral centers remains unchanged. The final product is saved to an output file, and a new pair of reactants is read.



**Figure 6.14.** Assignment of new torsions upon product formation. Top and middle: the values for the torsions where the new bond is terminal (i.e., 1-2-3-4 and 3-4-5-6) are assigned the value of a torsion from the reactant (i.e., 1-2-3- $X_1$  and  $X_2$ -4-5-6 respectively). Bottom: the torsion around the newly formed bond (i.e., 2-3-4-5) is scanned to minimize clashes.

### 6.2.3 SELECT

#### 6.2.3.1 MACCS fingerprints

We implemented a subset of the MDL keys (also referred to as MACCS keys) as a 2D fingerprint,<sup>31</sup> that use molecular descriptors involving atom connectivity, element identity and the presence of functional groups. In total, there are 31 atom-based descriptors, 32 atomic environment descriptors and 351 atom-bond-atom combination descriptors.

### **6.2.3.2 Clustering**

Once the fingerprints are calculated for all the input structures, pairwise similarities can be computed through the Tanimoto coefficient. The library is then associated in as many clusters as specified by the user, each containing compounds with maximum similarity.

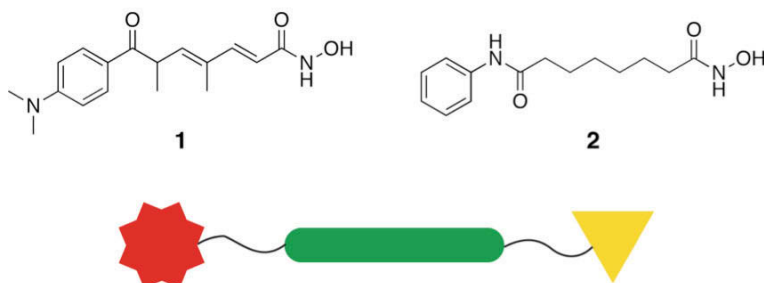
### **6.2.3.3 Analog search**

In the case of performing a search for analogs similar to a query molecule, the fingerprint of the query is compared to each of the members of the library. If the Tanimoto coefficient between both is higher than specified by the user, the library member is selected for output.

## **6.3 Application to the design of HDAC inhibitors**

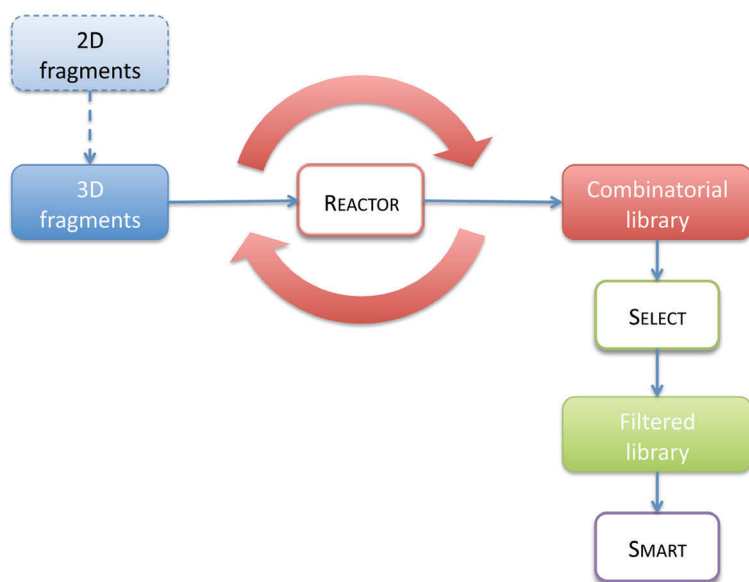
In collaboration with synthetic chemists from the Gleason research group (Dept. Chemistry, McGill University), there was interest in performing virtual screening of libraries of potential hybrid inhibitors of histone deacetylases (HDACs). HDACs are a family of zinc-containing enzymes that play a role in the remodeling of chromatin by deacetylating lysine residues on histone tails.<sup>32,33</sup> Histone deacetylation leads to transcriptional repression due to the formation of a condensed form of chromatin; inhibition of HDAC activity by small molecule inhibitors (HDAC inhibitors or HDACi) has been shown to be effective in the treatment of cancer.<sup>32,34-36</sup> Of the different families of HDACi, the hydroxamic acids have been among the most studied. For example, trichostatin A **1** (Figure 6.15, top) is an antifungal antibiotic exhibiting strong HDAC inhibition and preventing the growth of breast cancer tumour cell lines.<sup>37,38</sup> Suberoylanilide hydroxamic acid **2** (SAHA, vorinostat) is an HDACi approved by the FDA for the treatment of cutaneous T cell lymphoma.<sup>39,40</sup> These HDACis conform to a pharmacophore comprising a metal binding group linked to a hydrophobic chain and terminated by an aromatic group or a peptide;<sup>41</sup> we used this pharmacophore in the design of a virtual library of potential HDAC inhibitors (Figure 6.15, bottom). Zinc-binding groups included a variety of metal chelators such as hydroxamic acids, sulfonamides

and carboxylates. The hydrophobic linker was between five and nine atoms long, while the capping group was chosen from a diverse set of aromatic and aliphatic groups, as well as peptides.



**Figure 6.15.** Histone deacetylase inhibitors. Top: HDACis trichostatin A **1** and suberoylanilide hydroxamic acid (SAHA) **2**. Bottom: schematic representation of a library of HDAC inhibitors. In yellow, a capping group; in green, a hydrophobic linker of varying length; in yellow, a zinc-chelating moiety.

A protocol for the generation of a library of hybrid inhibitors is depicted in Figure 6.16. The first stage involves the enumeration of the library members for each of the fragments in the hybrid inhibitors; this can be done either by using a 3D interface (e.g., Maestro<sup>42</sup>) or a 2D drawing program (e.g., ChemDraw, ChemAxon) followed by a 3D coordinate generator (e.g., CORINA,<sup>43</sup> CONCORD,<sup>44</sup> OpenBabel<sup>45</sup>). With the fragment libraries in hand, the fragments are combined in as many REACTOR runs as necessary, resulting in a large combinatorial library. At this stage, redundancy in the library resulting from the combination of similar fragments can be achieved by filtering the library through SELECT, in an attempt to reduce the CPU time required for the virtual screening.



**Figure 6.16.** Workflow for the generation of a combinatorial library from libraries of 2D or 3D fragments. Libraries of three-dimensional fragments are fed to reactor, which outputs a combinatorial library with all possible products. The redundancy of this library can be reduced by removing compounds below a Tanimoto cutoff with SELECT. Finally, the filtered library can be fed to SMART and then docked with FITTED.

This virtual screening is ongoing, and its results will be reported in due course.

## 6.4 Conclusions

SMART, a module for the preparation of libraries of organic ligands for virtual screening was developed. It can be used to increase the throughput of virtual screening methods based on docking (FITTED) or in the evaluation of catalysts for asymmetric reactions (ACE). In addition to the assignment of suitable atom types for two different molecular mechanical force fields, the module is capable of assigning atomic charges based on the MMFF94 force field. We also implemented the ability of filtering compounds containing potentially toxic groups and/or not matching required features (such as the presence/absence of aromatic groups or hydrogen bond donors or acceptors), with a potential increase in VS throughput. Based on the basic components of SMART, we constructed a program to make combinatorial libraries of ligands from the combination of other libraries in a user-defined way. REACTOR is a simple yet powerful addition to the toolbox of the

computational medicinal chemist. Using data from publicly available databases containing three-dimensional structures from vendors, it is capable of providing a structure ready to be used in a structure-based drug design method such as docking. It is designed with the organic chemist in mind, allowing for the definition of reactions on chemical transformations in a simple language. It also allows for the formation of combinatorial libraries using Markush fragments. Multiple rounds of transformations can be used to build libraries from multiple components, as exemplified on the applications to VDR and HDAC.

### **6.5 Future developments and applications**

Future developments of these programs could include the incorporation of more powerful filtering techniques, such as skipping pairs of compounds that would make an unsuitable candidate (e.g., by using Lipinski's rules). One of the main present limitations of REACTOR is its inability to handle reactions where more than one bond is made at one time and where the connectivity of one of the reaction centers changes; however, these limitations are by application and not by design. The formation of multiple bonds in a single step can be attained by a redefinition of the rules specification, allowing for multiple cases to be matched. The circumvention of the requirement of fixed connectivity calls for the implementation of functions that would alter the geometry of a reaction center (e.g., changing from a 4-coordinate  $sp^3$  C to a 3-coordinate  $sp^2$  C), predicting the placement of the newly formed bonds upon a transformation.

### **6.6 References**

1. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187-217.
2. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A new force field for molecular mechanical



simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765-784.

3. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.

4. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

5. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A., An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230-252.

6. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

7. 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.

8. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **1993**, *97*, 10269-10280.

9. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132-146.

10. Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D., Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial charge models and optimization of nonpolar contributions. *J. Chem. Theor. Comp.* **2006**, *2*, 128-139.

11. Tsai, K. C.; Wang, S. H.; Hsiao, N. W.; Li, M.; Wang, B., The effect of different electrostatic potentials on docking accuracy: A case study using DOCK5.4. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3509-3512.

12. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity--a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-3228.
13. Halgren, T. A., Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520-552.
14. Sadowski, J.; Gasteiger, J.; Klebe, G., Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.
15. Irwin, J. J.; Shoichet, B. K., ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182.
16. Symyx Available Chemicals Directory.  
<http://www.symyx.com/products/databases/sourcing/acd/index.jsp>  
(accessed May 23, 2009).
17. National Cancer Institute Structural databases.  
[http://dtp.nci.nih.gov/docs/3D\\_database/structural\\_information/structural\\_data.html](http://dtp.nci.nih.gov/docs/3D_database/structural_information/structural_data.html) (accessed May 23, 2009).
18. Brown, L. J., The Markush Challenge. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 2-4.
19. Enyedy, I. J.; Egan, W. J., Can we use docking and scoring for hit-to-lead optimization? *J. Comput.-Aided Mol. Des.* **2008**, 1-8.
20. Aqvist, J.; Medina, C.; Samuelsson, J. E., A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385-391.
21. Kollman, P., Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395-2417.
22. Willett, P., Chemoinformatics - Similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **2000**, *11*, 85-88.
23. Willett, P.; Barnard, J. M.; Downs, G. M., Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.

24. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development and application of a rapid and accurate program based on organic principles. *Angew. Chem. Int. Ed.* **2008**, *47*, 2635-2638.
25. Halgren, T. A., Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17*, 616-641.
26. Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997-1009.
27. Fan, B. T.; Panaye, A.; Doucet, J. P.; Barbu, A., Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 657-662.
28. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580-594.
29. 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 Delivery Rev.* **1997**, *23*, 3-25.
30. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247-260.
31. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273-1280.
32. Johnstone, R. W., Histone-deacetylase inhibitors: Novel drugs for the treatment of cancer. *Nat. Rev. Drug Discovery* **2002**, *1*, 287-299.

33. De Ruijter, A. J. M.; Van Gennip, A. H.; Caron, H. N.; Kemp, S.; Van Kuilenburg, A. B. P., Histone deacetylases (HDACs): Characterization of the classical HDAC family. *Biochem. J.* **2003**, *370*, 737-749.
34. Marks, P. A.; Rifkind, R. A.; Richon, V. M.; Breslow, R.; Miller, T.; Kelly, W. K., Histone deacetylases and cancer: Causes and therapies. *Nat. Rev. Cancer* **2001**, *1*, 194-202.
35. Vigushin, D. M.; Coombes, R. C., Histone deacetylase inhibitors in cancer treatment. *Anti-Cancer Drugs* **2002**, *13*, 1-13.
36. Xu, W. S.; Parmigiani, R. B.; Marks, P. A., Histone deacetylase inhibitors: Molecular mechanisms of action. *Oncogene* **2007**, *26*, 5541-5552.
37. Yoshida, M.; Kijima, M.; Akita, M.; Beppu, T., Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. *J. Biol. Chem.* **1990**, *265*, 17174-17179.
38. Vigushin, D. M.; Ali, S.; Pace, P. E.; Mirsaidi, N.; Ito, K.; Adcock, I.; Coombes, R. C., Trichostatin A is a histone deacetylase inhibitor with potent antitumor activity against breast cancer in vivo. *Clin. Cancer Res.* **2001**, *7*, 971-976.
39. Kelly, W. K.; Marks, P. A., Drug insight: Histone deacetylase inhibitors - Development of the new targeted anticancer agent suberoylanilide hydroxamic acid. *Nature Clinical Practice Oncology* **2005**, *2*, 150-157.
40. Marks, P. A.; Breslow, R., Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. *Nat. Biotechnol.* **2007**, *25*, 84-90.
41. Finnin, M. S.; Donigian, J. R.; Cohen, A.; Richon, V. M.; Rifkind, R. A.; Marks, P. A.; Breslow, R.; Pavletich, N. P., Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **1999**, *401*, 188.
42. *Maestro*, 8.5; Schrödinger, LLC: Portland, OR, 2008. <http://www.schrodinger.com>
43. Gasteiger, J.; Rudolph, C.; Sadowski, J., Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Meth.* **1990**, *3*, 537-547.

44. *CONCORD*, Tripos, Inc.  
[http://www.optive.com/index.php?family=modules,SimplePage,,,&page=sybyl\\_concord](http://www.optive.com/index.php?family=modules,SimplePage,,,&page=sybyl_concord)
45. *The OpenBabel Package*, version 2.2.1, 2009.  
<http://www.openbabel.org>

## **Chapter 7: Contributions to knowledge**

### **7.1 Conclusions**

#### **7.1.1 Evaluation of docking programs for Golgi $\alpha$ -mannosidase**

We have assessed scoring functions used in molecular docking in the context of the application to the design of  $\alpha$ -mannosidase inhibitors. We found that most scoring functions assessed had problems at reproducing the binding modes of the ligands in the binding site, although in most cases they successfully predicted the binding mode around the metal centre. Despite these shortcomings, a virtual screening application of Glide was successful at retrieving seeded known actives from a library of decoys.

#### **7.1.2 Assessment of scoring functions for flexible docking**

In an attempt at deepening our assessment of scoring function performance, we have developed a challenging set of protein-ligand complexes for the assessment of structure-based CADD methods. Additionally, we prepared a set of cross-docked structures that can be used for the consideration of protein flexibility. The evaluation of the ability of several scoring functions to predict binding affinity showed that XScore, the eHiTS scoring function, DrugScore, GoldScore and ChemScore were the most reliable scoring functions overall. In particular, eHiTS and XScore were the least sensitive to the use of conformational ensembles instead of protein native structures as targets.

#### **7.1.3 Development of scoring functions for protein-ligand interactions**

We used the general Amber force field (GAFF) as the basis to develop new scoring functions, for its parameterization allows for broad applicability to organic molecules. The entropic energy of the ligand was best modeled by a count of rotatable bonds modified by a function that accounted for the polarity and buriedness of the bonds within the binding site. Entropic effects on the protein were considered by scaling down of the interactions with side chain atoms. Each term of

the scoring function was tuned in an iterative way, optimizing an objective function within increasingly focused ranges. The objective function was either the correlation between the ranked lists of binding affinities and predicted scores, or the area under the curve of a receiver-operating characteristic for the retrieval of actives in a VS application. Validation of these scoring functions with standard benchmark sets showed that the first developed scoring function (RankScore2) was among the best scoring functions at binding affinity prediction, while the second one (RankScore4) was successful at retrieving known actives from libraries of decoys in a different set of targets.

#### **7.1.4 Modelling of platinum complexes as G-quadruplex binders**

We have applied a hybrid docking/molecular dynamics/MM-PBSA scoring technique for the development of platinum square-planar complexes as DNA G-quadruplex binders. With this technique, we were able to gain evidence for the parallel stacking mode these compounds exhibit, as well as explain differences in activity in congeneric series of compounds. Furthermore, when applying the technique to a pair of alternate foldings of the human G-quadruplex motif, the preference of different Pt complexes for each folding could be explained.

#### **7.1.5 Development of programs for the handling of ligands for virtual screening: SMART, REACTOR and SELECT**

Finally, we described the development of a set of programs to handle one part of the input for docking programs: the ligands. The SMART module of FITTED automates the setting up of the ligands prior to a docking run, assigning GAFF atom types, and MMFF charges and setting up rotatable bonds. An implementation of the MACCS fingerprint led to the development of SELECT, which exploits 2D similarity search to cluster libraries of ligands and to extract analogs to a query from a library. A third program, REACTOR, is able to construct combinatorial libraries of ligands with user-defined transformations, either with a defined chemistry or with a target-oriented behaviour. These programs are currently being used for the preparation of libraries to be used in the virtual screening of histone deacetylase inhibitors.

## **7.2 Papers and conference presentations**

### **7.2.1 Papers published**

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.

Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F. A platinum supramolecular square structure as a G-quadruplex interactive agent. *J. Am. Chem. Soc.* **2008**, *130*, 10040.

Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of FITTED 1.5 to the Virtual Screening of Potential HCV Polymerase Inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 902.

Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7.

Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N. Evaluation of Docking Programs for Predicting Binding of Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160.

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*, 43.

Moitessier, N.; Englebienne, P.; Chapleur, Y. Directing-protecting groups for carbohydrates. Design, conformational study, synthesis and application to regioselective functionalization. *Tetrahedron* **2005**, *61*, 6839.



### 7.2.2 Book chapters

Kieltyka, R.; Englebienne, P.; Moitessier, N.; Sleiman, H. F. Quantifying interactions between G-quadruplex DNA and transition metal complexes. *Methods in Molecular Biology*, Special Issue on G-quadruplexes, submitted.

### 7.2.3 Forthcoming papers

Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 5. Force Field-Based Prediction of Binding Affinities of Ligands To Proteins.

Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F. Interaction of Platinum Phenanthroimidazole Complexes with the Human G-Quadruplex Sequence.

Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Langille, A.; Autexier, C.; Moitessier, N.; Sleiman, H. F. Increasing Planarity of Platinum Phenanthroimidazole G-Quadruplex Binders Through Hydrogen Bonding.

### 7.2.4 Conference presentations

Kieltyka, R.\*; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F. Platinum complexes with variable  $\pi$ -surfaces and their interactions with the guanine quadruplex. 236<sup>th</sup> ACS National Meeting, Philadelphia, PA, Aug/2008. (*oral presentation*)

Kieltyka, R.\*; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F. Platinum-based G-quadruplex binders as potential telomerase inhibitors. 6<sup>th</sup> Canadian Symposium on Telomeres and Telomerase, Lake Winnipeg, MB, May/2008. (*oral presentation*)

Englebienne, P.\*; Corbeil, C. R.; Moitessier, N. FORECASTER: A new platform for drug discovery. 8<sup>th</sup> CERMM Symposium, Montreal, QC, Apr/2008.

Englebienne, P.\*; Corbeil, C. R.; Moitessier, N. FORECASTER: A new platform for drug discovery. 235<sup>th</sup> ACS National Meeting, New Orleans, LA, Apr/2008.

Englebienne, P.\*; Moitessier, N. RankScore2: A novel scoring function for ligand-protein binding affinities. 235<sup>th</sup> ACS National Meeting, New Orleans, LA, Apr/2008.

Kieltyka, R.\*; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F. Platinum phenanthroimidazole complexes as G-quadruplex binders. 235<sup>th</sup> ACS National Meeting, New Orleans, LA, Apr/2008. (*oral presentation*)

Englebienne, P.\*; Moitessier, N. Development of a docking scoring function from a challenging training set. Computer-Aided Drug Design Gordon Research Conference, Tilton, NH, Jul/2007.

Moitessier, N.\*; Corbeil, C. R.; Englebienne, P. FITTED 1.0, docking to flexible and solvated macromolecules. 6<sup>th</sup> Canadian Computational Chemistry Conference, Vancouver, BC, Jul/2006.

Englebienne, P.\*; Moitessier, N. Prédiction *in silico* de l'affinité ligand/protéine par fonctions de score. 74<sup>th</sup> ACFAS, Montreal, QC, May/2006. (*oral presentation*)

Englebienne, P.\*; Moitessier, N. Predictive tools for structure-based drug design: A comparative study of force field-based scoring functions. 230<sup>th</sup> ACS National Meeting, Washington, DC, Aug/2005.

Englebienne, P.\*; Moitessier, N. Development of Novel Software for Docking Ligands to Flexible Proteins. 88<sup>th</sup> CSC National Conference & Exhibition, Saskatoon, SK, May/2005.

Corbeil, C.R.\*; Englebienne, P.\*; Moitessier, N. Development of a novel software to predict the binding affinity of ligands to flexible proteins. 5<sup>th</sup> CERMM Symposium, Montreal, QC, Feb/2005.

Englebienne, P.\*; Moitessier, N.; Chapleur, Y. Carbohydrate Protecting-Directing Groups. 15<sup>th</sup> QOMSBQC, Ottawa, ON, Dec/2004. (*oral presentation*)

[ This page was intentionally left blank ]

## Appendix

### *A.1 Supplementary information for Chapter 3*

**Table A.1.** Description of the 209 complexes included in Set 1.

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1a30	HIV-1 protease	-5.86	375.29
1a85	MMP-8	-6.16	393.45
1adl	Trp RNA-binding attenuation protein	-7.30	303.47
1af6	Maltoporin	-2.48	342.30
1afk	Ribonuclease a	-9.01	504.16
1aid	HIV-1 protease	-6.57	453.07
1ajv	HIV-1 protease	-10.51	574.70
1apw	Penicillopepsin	-10.89	506.64
1b8n	Purine nucleoside phosphorylase	-14.33	282.28
1b8o	Purine nucleoside phosphorylase	-14.48	267.27
1bdl	HIV-1 protease	-7.97	520.68
1biw	MMP-3	-9.51	415.53
1bma	Elastase	-6.25	521.61
1bmh	Thrombin	-11.49	531.66
1bn4	Carbonic anhydrase	-12.67	362.45
1bnn	Carbonic anhydrase	-13.61	374.46
1br5	Ricin	-3.68	253.22
1br6	Ricin	-4.38	311.28
1bxo	Penicillopepsin	-13.61	638.70
1c3x	Pentosyltransferase	-5.01	277.02
1c4u	Thrombin	-14.12	582.51
1c4v	Thrombin	-14.70	531.68
1c5c	chimeric decarboxylase antibody 21d8	-9.48	343.34

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1c5q	Trypsin	-8.66	303.15
1cet	Lactate dehydrogenase	-3.93	320.89
1ciz	MMP-3	-10.13	502.62
1d3p	Thrombin	-10.06	545.75
1d3t	Thrombin	-8.90	543.73
1d4p	Thrombin	-8.58	361.47
1d8m	MMP-3	-11.58	419.46
1db1	Vitamin d nuclear receptor	-12.61	416.65
1dmp	HIV-1 protease	-13.00	536.68
1dvz	Transthyretin	-9.68	280.23
1dy4	Cellobiohydrolase I	-5.94	258.34
1e1x	Cyclin dependent kinase 2	-8.02	251.29
1e2k	Thymidine kinase	-6.73	252.27
1e4h	Transthyretin	-11.45	488.62
1e5a	Transthyretin	-10.40	330.82
1e66	Acetylcholinesterase	-13.46	298.82
1eb1	Thrombin	-14.20	439.58
1ecv	PTP 1b	-6.61	333.04
1efy	ADP-ribose polymerase	-11.19	267.29
1ejn	Urokinase-type plasminogen activator	-7.65	342.47
1elc	Elastase	-9.07	507.58
1erb	Retinoic acid receptor rxr-alpha	-9.60	327.51
1evh	Mena evh1 domain	-4.38	594.69
1ezq	Factor xa	-12.32	460.58
1f0u	Trypsin	-9.75	460.58
1f4e	Thymidylate synthase	-4.03	268.31
1f4g	Thymidylate synthase	-8.82	496.48
1f73	n-Acetyl neuraminate lyase	-3.25	310.28

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1f8d	Neuraminidase	-4.63	290.28
1f8e	Neuraminidase	-6.57	290.30
1fcx	Retinoic acid receptor gamma-1	-9.79	387.50
1fcy	Retinoic acid receptor gamma-1	-11.60	385.49
1fcz	Retinoic acid receptor gamma-1	-12.55	361.47
1fd0	Retinoic acid receptor gamma-1	-11.44	400.50
1fh7	Xylanase beta-1	-7.13	266.27
1fh8	Xylanase beta-1	-9.38	250.27
1fh9	Xylanase beta-1	-8.75	294.26
1fhd	Xylanase beta-1	-9.28	302.29
1fjs	Factor xa	-13.56	526.50
1fkh	FK506 binding protein	-11.10	455.64
1fmb	EIAV protease	-13.61	566.77
1g2l	Factor xa	-9.86	526.62
1g30	Thrombin	-9.33	526.62
1g4o	Carbonic anhydrase	-11.23	290.34
1ghv	Thrombin	-5.92	254.27
1ghw	Thrombin	-5.72	253.29
1ghz	Trypsin	-6.53	254.27
1gi1	Trypsin	-6.49	253.29
1gi6	Trypsin	-8.47	252.30
1gi8	Urokinase-type plasminogen activator	-6.87	253.29
1gj4	Thrombin	-5.75	362.84
1gj5	Thrombin	-7.16	329.38
1gj7	Urokinase-type plasminogen activator	-10.74	362.84
1gj9	Urokinase-type plasminogen activator	-10.18	382.46
1gja	Urokinase-type plasminogen activator	-7.38	256.29
1gpn	Acetylcholinesterase	-8.82	257.36

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1h0a	Epsin	-7.41	417.07
1h1s	Cyclin dependent kinase 2	-11.19	402.48
1h22	Acetylcholinesterase	-12.38	468.69
1h23	Acetylcholinesterase	-11.36	496.74
1h46	Exoglucanase I	-4.86	257.34
1hpo	HIV-1 protease	-12.55	504.61
1hpv	HIV-1 protease	-12.55	505.64
1htf	HIV-1 protease	-9.30	575.76
1hwr	HIV-1 protease	-11.34	406.53
1hy7	MMP-3	-7.34	402.45
1i73	MMP-8	-7.48	449.47
1igj	Precursor of periplasmic sugar receptor	-13.61	520.67
1ikt	Estradiol 17 beta-dehydrogenase 4	-4.63	352.52
1j8v	Beta-d-glucan glucohydrolase isoenzyme exo1	-4.91	358.37
1jan	MMP-8	-6.43	301.37
1jao	MMP-8	-8.06	323.42
1jaq	MMP-8	-6.10	302.33
1jj9	MMP-8	-7.85	332.38
1jwt	Thrombin	-10.69	493.61
1k22	Thrombin	-11.43	430.53
1k4g	tRNA-guanine transglycosylase	-7.96	288.33
1k9s	Purine nucleoside phosphorylase	-8.88	282.28
1kbc	MMP-8	-10.48	399.53
1kdk	Sex hormone-binding globulin	-12.32	290.45
1kel	28b4 fab	-9.91	345.27
1koj	Glucose-6-phosphate isomerase	-9.12	260.12
1ksn	Factor xa	-12.80	447.52

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1kts	Thrombin	-10.93	500.59
1kzk	HIV-1 protease	-14.14	575.73
1l2s	Lactamase beta	-6.25	316.76
1l8b	Eukaryotic translation initiation factor 4e	-9.33	535.20
1lee	Plasmepsin II	-10.54	531.70
1lf2	Plasmepsin II	-10.24	531.70
1lhw	Sex hormone-binding globulin	-11.11	302.42
1lnm	Diga16	-11.84	374.53
1loq	Orotidine 5'-monophosphate decarboxylase	-5.04	323.18
1lor	Orotidine 5'-monophosphate decarboxylase	-15.06	339.18
1m0n	2,2-Dialkylglycine decarboxylase	-3.02	392.24
1m0o	2,2-Dialkylglycine decarboxylase	-3.14	380.23
1m0q	2,2-Dialkylglycine decarboxylase	-4.03	352.18
1m48	Interleukin-2	-6.93	447.56
1mai	Phospholipase C delta-1	-9.09	417.07
1mfa	14-3-3-like protein c	-6.86	486.47
1mfl	Erb-b2 interacting protein	-5.30	540.64
1mmb	MMP-8	-11.30	477.65
1mnc	MMP-8	-12.25	349.43
1moq	Glucosamine 6-phosphate synthase	-4.71	259.15
1mq6	Factor xa	-15.18	568.87
1mrw	Pol polyprotein	-13.21	527.69
1mu6	Thrombin	-11.41	432.41
1mu8	Thrombin	-12.25	446.44
1n3i	Purine nucleoside phosphorylase	-12.10	265.29
1n46	Thyroid hormone receptor beta-1	-14.32	367.41
1n4k	Inositol 1	-13.68	417.07
1nfw	Factor xa	-12.20	436.94



<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1nfy	Factor xa	-12.10	463.99
1nje	Thymidylate synthase	-5.17	306.19
1njj	Ornithine decarboxylase	-2.86	500.59
1njs	Phosphoribosylglycinamide formyltransferase	-10.65	543.46
1nl9	PTP 1b	-8.11	531.57
1nm6	Thrombin	-13.68	511.07
1no6	PTP 1b	-10.09	333.30
1nq7	Nuclear receptor ror-beta	-9.26	339.50
1nvq	Serine/threonine-protein kinase chk1	-11.23	481.54
1nvr	Serine/threonine-protein kinase chk1	-11.04	467.55
1nw4	Uridine phosphorylase	-12.35	267.27
1nwl	PTP 1b	-3.25	468.58
1o0m	Ribonuclease a	-7.01	323.18
1o0n	Ribonuclease a	-5.57	323.18
1o0o	Ribonuclease a	-6.94	425.19
1o2h	Trypsin	-8.37	336.42
1o2p	Trypsin	-6.60	359.84
1o36	Trypsin	-8.11	488.44
1o3d	Trypsin	-9.71	329.38
1o3f	Trypsin	-10.84	328.40
1ocq	Endoglucanase 5a	-7.07	310.33
1oe8	Glutathione s-transferase	-7.51	306.32
1okl	Carbonic anhydrase	-8.21	250.32
1ony	PTP 1b	-9.22	588.64
1osv	Bile acid receptor	-4.74	419.63
1ow4	Pheromone binding protein	-7.73	299.35
1p57	Serine protease hepsin	-5.99	253.29

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1pkx	Bifunctional purine biosynthesis protein purh	-9.42	363.20
1pme	Erk2 map kinase	-12.80	377.44
1pr5	Purine nucleoside phosphorylase	-5.34	266.26
1pvn	Inosine-5'-monophosphate dehydrogenase	-13.37	338.19
1pzi	Heat-labile enterotoxin b subunit	-5.74	331.28
1pzp	Beta-lactamase tem	-4.51	305.34
1q1g	Uridine phosphorylase putative	-11.67	297.36
1q66	Queuine trna-ribosyltransferase	-7.86	313.40
1qb9	Trypsin	-10.13	490.63
1qbv	Thrombin	-7.33	382.49
1qk4	Hypoxanthine phosphoribosyltransferase	-5.73	347.20
1rbo	Ribulose bisphosphate carboxylase/oxygenase (rubisco)	-17.32	353.09
1sl3	Thrombin	-16.13	552.33
1sqa	Urokinase-type plasminogen activator	-12.54	413.49
1sqn	Progesterone receptor	-12.80	300.44
1sr7	Progesterone receptor	-13.75	523.46
1sri	Streptavidin	-8.28	269.28
1stc	Cyclin dependent kinase 2	-11.03	467.55
1ta6	Thrombin	-12.43	480.03
1tom	Thrombin	-11.30	388.56
1ttm	Carbonic anhydrase	-10.00	309.34
1tyr	Transthyretin	-9.53	303.47
1ur8	Chitinase b	-5.92	422.39
1urg	Maltose-binding protein	-7.92	342.30
1usn	MMP-3	-10.54	427.38
1v2k	Trypsin	-8.43	499.04

<b>PDB</b>	<b>Protein</b>	<b>Ebind (kcal mol<sup>-1</sup>)</b>	<b>MW</b>
1v2o	Trypsin	-6.44	433.51
1v48	Purine nucleoside phosphorylase	-10.61	335.21
1vpo	Anti-testosterone (light chain)	-12.96	288.43
1vyf	14 kda fatty acid binding protein	-10.96	281.46
1w1d	3-phosphoinositide dependent protein kinase-1	-8.88	496.05
1xkk	Epidermal growth factor receptor	-11.60	582.08
1yej	Ig antibody d2.3 (light chain)	-10.16	400.33
2bpv	HIV-1 protease	-10.45	622.84
2gss	Glutathione s-transferase	-6.73	302.14
2qwe	Neuraminidase	-10.19	332.32
2qwf	Neuraminidase	-7.71	341.37
2rkm	Oligo-peptide binding protein	-5.31	276.38
2std	Scytalone dehydratase	-13.41	334.68
3gst	Glutathione S-transferase	-9.15	500.55
3std	Scytalone dehydratase	-15.12	364.45
4sga	Proteinase a	-9.94	471.54
4std	Scytalone dehydratase	-14.06	338.18
5std	Scytalone dehydratase	-14.28	375.42
6cpa	Carboxypeptidase	-15.69	476.43
6std	Scytalone dehydratase	-11.76	413.17
7std	Scytalone dehydratase	-14.59	334.68
8cpa	Carboxypeptidase	-12.45	462.40

**Table A.2.** Listing of 87 complexes included in Set 2.

<b>PDB code</b>	<b>Proteins</b>
1bn4	Carbonic anhydrase
1bnn	Carbonic anhydrase
1okl	Carbonic anhydrase
1ttm	Carbonic anhydrase
1ezq	Factor xa
1fjs	Factor xa
1g2l	Factor xa
1ksn	Factor xa
1mq6	Factor xa
1nfw	Factor xa
1nfy	Factor xa
1a30	HIV-1 protease
1aid	HIV-1 protease
1ajv	HIV-1 protease
1bdl	HIV-1 protease
1dmp	HIV-1 protease
1hpo	HIV-1 protease
1hpy	HIV-1 protease
1htf	HIV-1 protease
1hwr	HIV-1 protease
1kzk	HIV-1 protease
2bpv	HIV-1 protease
1biw	MMP-3
1ciz	MMP-3
1d8m	MMP-3
1usn	MMP-3
1a85	MMP-8
1i73	MMP-8

---

1jan	MMP-8
1jao	MMP-8
1jaq	MMP-8
1jj9	MMP-8
1kbc	MMP-8
1mmb	MMP-8
1mnc	MMP-8
1ecv	PTP 1b
1nl9	PTP 1b
1no6	PTP 1b
1nwl	PTP 1b
1ony	PTP 1b
2std	Scytalone dehydratase
3std	Scytalone dehydratase
4std	Scytalone dehydratase
5std	Scytalone dehydratase
6std	Scytalone dehydratase
7std	Scytalone dehydratase
1bmn	Thrombin
1c4u	Thrombin
1c4v	Thrombin
1d3p	Thrombin
1d3t	Thrombin
1d4p	Thrombin
1eb1	Thrombin
1g30	Thrombin
1ghv	Thrombin
1ghw	Thrombin
1gj4	Thrombin
1gj5	Thrombin

---

---

1jwv	Thrombin
1k22	Thrombin
1kts	Thrombin
1mu6	Thrombin
1mu8	Thrombin
1nm6	Thrombin
1qbv	Thrombin
1sl3	Thrombin
1ta6	Thrombin
1tom	Thrombin
1c5q	Trypsin
1f0u	Trypsin
1ghz	Trypsin
1gi1	Trypsin
1gi6	Trypsin
1o2h	Trypsin
1o2p	Trypsin
1o36	Trypsin
1o3d	Trypsin
1o3f	Trypsin
1qb9	Trypsin
1v2k	Trypsin
1v2o	Trypsin
1ejn	Urokinase-type plasminogen activator
1gi8	Urokinase-type plasminogen activator
1gj7	Urokinase-type plasminogen activator
1gj9	Urokinase-type plasminogen activator
1gja	Urokinase-type plasminogen activator
1sqa	Urokinase-type plasminogen activator

---

**Table A.3.** Accuracy of the scoring functions on 5 subsets of set 1. Each cell presents the ranked correlation coefficient  $\tau$  for each scoring function within a subset of proteins.

	<b>Thrombin</b>	<b>HIV-1P</b>	<b>Trypsin</b>	<b>MMP-3/8</b>	<b>trypsin/thrombin/FXa</b>
	<b>(22)</b>	<b>(11)</b>	<b>(13)</b>	<b>(14)</b>	<b>(42)</b>
Min $E_{\text{bind}}$	-16.1	-14.1	-10.8	-12.2	-16.1
Max $E_{\text{bind}}$	-5.7	-5.8	-6.4	-6.1	-5.7
ChemScore (Sybyl)	0.056	0.624	0.564	0.341	0.266
DockScore (Sybyl)	0.368	0.624	0.385	0.297	0.459
DrugScore <sup>CSD</sup>	0.42	0.55	0.462	0.077	0.443
DrugScore <sup>PDB</sup>	0.558	0.477	0.179	0.077	0.556
eHiTS SF	0.455	0.367	0.077	0.209	0.388
FlexXScore (Sybyl)	-0.022	0.661	0.385	0.209	0.075
GlideScore	0.307	0.514	0.308	0.165	0.333
GoldScore (Sybyl)	0.55	0.44	0.179	0.077	0.503
Hammerhead	0.304	0.477	0.487	0.253	0.313
(Cerius2)					
LigScore1 (Cerius2)	0.429	0.44	0.179	0.088	0.369
LigScore2 (Cerius2)	0.451	0.514	0.245	0.165	0.419
PLP1 (Cerius2)	0.481	0.587	0.308	0.099	0.487
PLP2 (Cerius2)	0.394	0.587	0.308	0.143	0.447
PMF (Sybyl)	0.29	-0.183	-0.026	-0.209	0.303

PMF (Cerius2)	0.247	0.55	0.359	-0.209	0.354
RankScore	0.68	0.55	0.359	0.099	0.605
Surflex SF	0.177	0.477	0.308	0.231	0.298
XScore	0.312	0.734	0.436	0.199	0.414
MW	0.489	0.257	0.128	0.341	0.459



**Table A.4.** Accuracy ( $\tau$ ) of the scoring functions on three subsets of Set 1.

Scoring function	Hydrophobic proteins (85 complexes)	Hydrophilic proteins (93 complexes)	Metalloproteins (23 complexes, all hydrophobic)
ChemScore (Sybyl)	0.301	0.277	0.043
DockScore (Sybyl)	0.230	0.282	0.083
DrugScore <sup>CSD</sup>	0.218	0.357	-0.036
DrugScore <sup>PDB</sup>	0.176	0.346	-0.107
eHiTS SF	0.227	0.365	0.02
FlexXScore	0.223	0.211	0.083
GlideScore	0.288	0.240	0.04
GoldScore (Sybyl)	0.278	0.315	0.02
Hammerhead (Cerius2)	0.292	0.221	0.043
LigScore1 (Cerius2)	0.109	0.109	0.083
LigScore2 (Cerius2)	0.190	0.317	0.051
PMF (Sybyl)	0.245	0.380	-0.051
PMF (Cerius2)	0.269	0.352	-0.043
PLP1 (Cerius2)	0.109	-0.064	0.075
PLP2 (Cerius2)	0.255	0.004	0.051
RankScore	0.223	0.402	-0.012

Surflex SF	0.282	0.238	0.067
XScore	0.307	0.361	0.1
MW	0.153	0.289	0.249

**Table A.5.** Accuracy of the scoring functions ( $\tau$ ) on set 2 when waters are kept for docking and scoring, kept for docking and removed for scoring, removed for both docking and scoring, or made displaceable.

Scoring function	Native	Non-native protein conformation <sup>a</sup>				
	protein conformation	Waters kept	Wet/wet	Wet/dry	Dry/dry	Displaceable
ChemScore (Sybyl)	0.338	0.307	0.307	0.328	0.330	
DockScore (Sybyl)	0.280	0.273	0.274	0.267	0.281	
DrugScore <sup>CSD</sup>	0.351	0.295	0.296	0.320	0.326	
DrugScore <sup>PDB</sup>	0.333	0.336	0.340	0.337	0.342	
eHiTS SF	0.366	0.359	0.380	0.380	0.386	
FlexXScore (Sybyl)	0.164	0.081	0.048	0.069	0.100	
GlideScore	0.307	0.071	0.076	0.126	0.133	
GoldScore (Sybyl)	0.259	0.274	0.274	0.272	0.286	
Hammerhead SF	0.176	0.116	-0.009	0.001	0.116	
(Cerius2)						
LigScore1	0.114	-0.017	-0.026	0.007	0.015	
(Cerius2)						
LigScore2	0.266	0.076	0.064	0.101	0.135	
(Cerius2)						
PLP1 (Cerius2)	0.289	0.261	0.253	0.276	0.280	
PLP2 (Cerius2)	0.305	0.260	0.241	0.261	0.270	
PMF (Sybyl)	0.230	0.209	0.209	0.203	0.212	
PMF (Cerius2)	0.315	0.297	0.288	0.292	0.296	
RankScore	0.295	0.242	0.243	0.286	0.286	
Surflex SF	0.218	0.050	0.055	0.056	0.073	
XScore	0.413	0.403	0.405	0.412	0.412	
MW	0.276					

<sup>a</sup> random sampling of cross-docked structures, 1 protein per ligand; maximum and minimum correlation values displayed for 10,000 trials.

**Table A.6.** Accuracy of the scoring functions ( $\tau$ ) on the complete set 2 when protein conformational ensembles were considered.

Scoring function	Self-docking	Protein conformational ensemble <sup>a</sup>			
	Wet/wet	Wet/wet	Wet/dry	Dry/dry	displaceable
ChemScore (Sybyl)	0.338	0.317	0.317	0.290	0.305
DockScore (Sybyl)	0.280	0.260	0.260	0.235	0.261
DrugScore <sup>CSD</sup>	0.351	0.296	0.297	0.311	0.315
DrugScore <sup>PDB</sup>	0.333	0.329	0.333	0.320	0.336
eHiTS SF	0.366	0.451	0.441	0.441	0.451
FlexXScore (Sybyl)	0.164	0.102	0.052	0.058	0.098
GlideScore	0.307	0.228	0.219	0.267	0.276
GoldScore (Sybyl)	0.259	0.287	0.287	0.251	0.282
Hammerhead SF (Cerius2)	0.176	0.135	0.007	0.002	0.135
LigScore1 (Cerius2)	0.114	0.072	0.046	0.048	0.057
LigScore2 (Cerius2)	0.266	0.170	0.113	0.129	0.177
PLP1 (Cerius2)	0.289	0.294	0.280	0.281	0.290
PLP2 (Cerius2)	0.305	0.298	0.279	0.280	0.298
PMF (Sybyl)	0.230	0.215	0.215	0.215	0.219
PMF (Cerius2)	0.315	0.281	0.280	0.272	0.277
RankScore	0.295	0.302	0.302	0.305	0.314
Surflex SF	0.218	0.105	0.075	0.075	0.112
XScore	0.413	0.391	0.391	0.390	0.390
MW			0.277		

<sup>a</sup> best scored complex for each ligand.

## A.2 *Supplementary information for Chapter 4*

**Table A.7.** Correlations between predicted score and experimental binding affinity for the test set of 93 complexes.

Scoring function	Pearson	Spearman	Kendall
LigScore2	0.209	0.449	0.313
LigScore1	0.281	0.414	0.287
PMF-Sybyl	0.309	0.295	0.208
RankScore4	0.317	0.395	0.268
FlexXScore	0.345	0.417	0.271
Jain	0.388	0.401	0.264
PMF-Cerius2	0.419	0.382	0.282
GlideScore	0.437	0.452	0.314
MW	0.538	0.552	0.395
PLP1	0.544	0.560	0.397
GoldScore	0.549	0.530	0.374
RankScore2	0.553	0.524	0.362
PLP2	0.555	0.579	0.413
DockScore	0.569	0.574	0.410
ChemScore	0.577	0.562	0.396
DrugScorePDB	0.590	0.577	0.392
DrugScoreCSD	0.612	0.585	0.415
X-Score	0.672	0.682	0.494

### ***A.3 Supplementary information for Chapter 5***

#### **A.3.1 Fluorescence resonance energy transfer assay (FRET) experiment details**

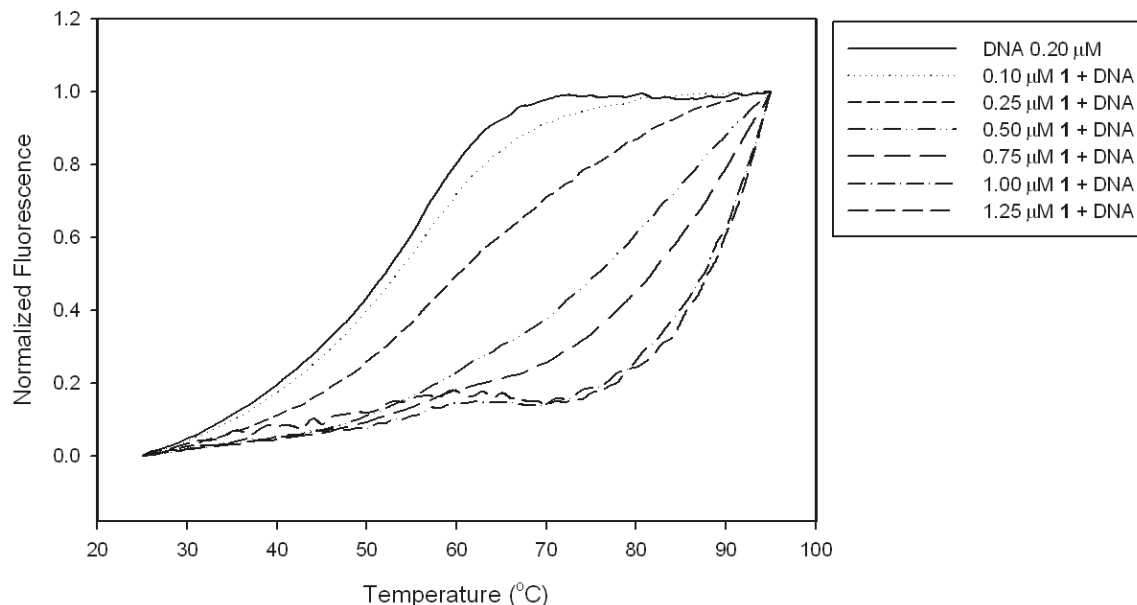
Complex **1** is highly soluble in water and as a result, a stock solution of 10 mM was prepared in deionized water. Both duplex (5'-FAM-TATAGCTATA-HEG-TATAGCTATA-TAMRA-3' where FAM = fluorescein, TAMRA=tetracarboxylrhodamine, and HEG = hexaethyleneglycol) and a quadruplex (F21T = 5'-FAM-GGG(TTAGGG)3-TAMRA-3') forming FRET oligonucleotides (SigmaGenosys) were dissolved in deionized water and their stock solutions were quantified by UV/Vis spectroscopy. Any dilutions past this point were performed in a 10 mM sodium cacodylate buffer with 100 mM LiCl (pH 7.4).

A 400 nM solution of each oligonucleotide was prepared in the aforementioned cacodylate buffer and heated to 90°C in the UV/Vis instrument for 5 minutes and then allowed to cool within the instrument for 2-3 hours. In the meantime, several solutions of varying concentration of the molecular square, **1**, were prepared 2x as concentrated as to be used in the assay, from 0.1  $\mu$ M to 2  $\mu$ M.

The Fluorescence Resonance Energy Transfer (FRET) assay was performed as a high throughput screen in 256-well format with F21T and duplex DNA. Fluorescence measurements were recorded in an Applied Biosystems Real-Time PCR (ABI HT 7900). Solutions of quadruplex or duplex DNA and complex **1** (20  $\mu$ L of each component) were pipetted into the wells to give a total reaction volume of 40  $\mu$ L. The emission of FAM was followed with its excitation at 494 nm and emission at 522 nm. Samples were first equilibrated within the instrument at 25°C prior to heating to 95°C in 71 cycles at 1°C/min. Fluorescence readings were recorded every 0.5°C/min.

$\Delta T_{1/2}$  values were obtained by normalizing the FRET data from 0 to 1 using a concentration of less than 1  $\mu$ M in platinum square.  $\Delta T_{1/2}$  was taken as the

temperature at which the normalized emission is equal to 0.5. Each concentration value was collected as a triplicate and experiments were repeated 2-3 times.

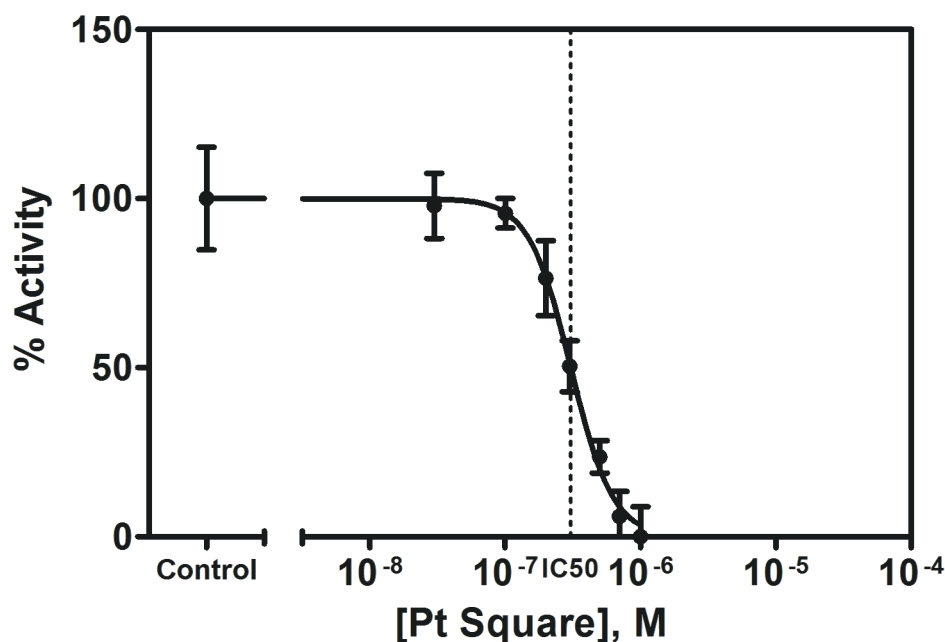


**Figure A.1.** Titration of complex **1** (0.1 to 1.25  $\mu\text{M}$ ) with G-quadruplex DNA.

### A.3.2 TRAP assay protocol

Inhibition of telomerase activity was detected by a modified, two-step version of the telomeric repeat amplification protocol (TRAP).<sup>3</sup> Telomerase extract was prepared from 1  $\mu\text{g}$  of exponentially growing HeLa cells and was used for every reaction in a final volume of 50  $\mu\text{L}$  consisting of TRAP buffer (final concentration of 20 mM Tris-HCl pH 8.3, 1.5 mM  $\text{MgCl}_2$ , 63 mM KCl, 1 mM EGTA pH 8.0, 0.01% Tween-20, 100 ng/ $\mu\text{L}$  BSA), 2.5mM dNTP mix, 40 pmol TS primer, 20 pmol NT primer,  $1 \times 10^{-13}$  M TSNT internal control primer. Each reaction was incubated with increasing concentration or without complex **1** for 30 minutes at 30°C. To each reaction, 5  $\mu\text{Ci}$  of [ $^{32}\text{P}$ ]dGTP, 2 units of *Taq* polymerase, and 20 pmol of ACX reverse primer were added. Reactions were amplified by the polymerase chain reaction (PCR) for 30 cycles with the following conditions: 30 seconds at 94°C, 30 seconds at 60°C, and 30 seconds at 72°C. DNA products were resolved on a 10% acrylamide gel that was dried and exposed to an autoradiography film (GE Healthcare, Canada).

Quantification of inhibition was performed by using Imagequant software (GE Healthcare, Canada). Telomeric ladder products were normalized to the internal control, and the ratio obtained for inhibited telomerase reactions were compared the ratio of uninhibited telomerase. We determined the  $IC_{50}$  obtained when the ligand is added after the telomerase extension step. As a control experiment, we find that the  $IC_{50}$  obtained when the ligand is added after the telomerase extension step to be  $0.304\ \mu\text{M}$ , higher than when added during the telomerase extension step ( $IC_{50} = 0.197\ \mu\text{M}$ ). These results are consistent with results obtained with other G-quadruplex ligands such as telomestatin that have been established as specific telomerase inhibitors.<sup>4</sup>



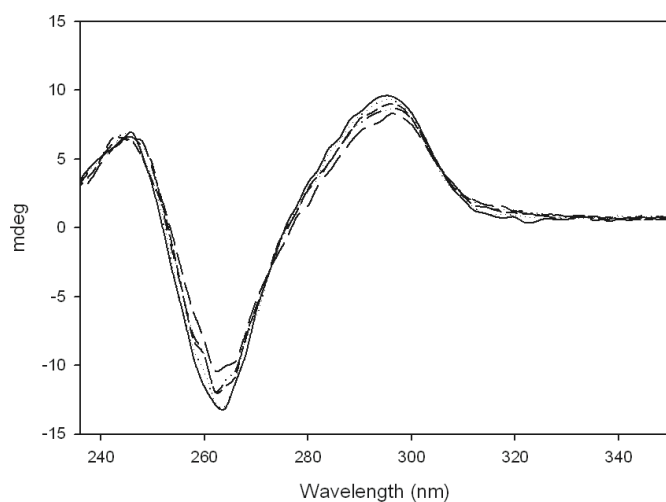
**Figure A.2.** Percent activity of telomerase upon addition of increasing amounts of complex **1**.

#### **A.4 Circular dichroism (CD) study**

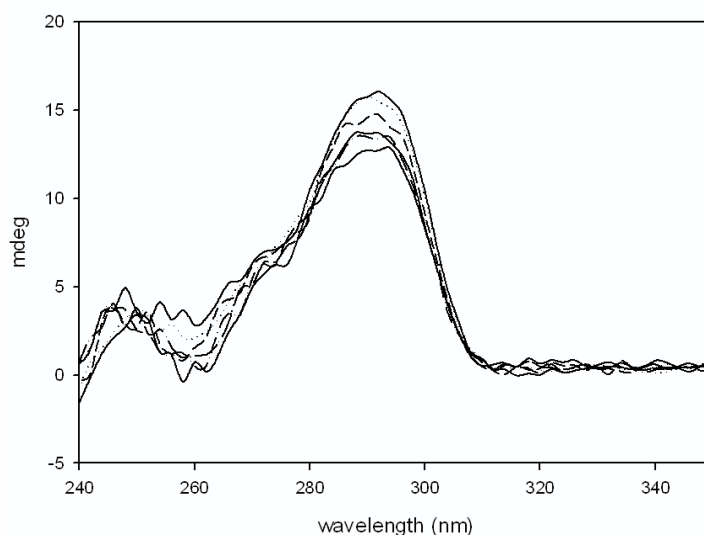
A circular dichroism study was conducted to observe the effect of complex **1** on the structure of the G-quadruplex. CD studies were performed on a JASCO J-810 spectrophotometer with a 1 cm path length cuvette. Scans were taken from 350-235 nm at a scan speed of 200 nm/min with 5 acquisitions. A solution of complex **1**



was prepared as a 400  $\mu\text{M}$  solution in water. The human telomere sequence, 5'-AGGGTTAGGGTTAGGGTTAGGGT-3' (SigmaGenosys), was dissolved in a 10 mM sodium phosphate buffer with 100 mM NaCl (pH 7.2) and heated to 90°C for 5 minutes and then cooled to 25°C over 1 hour in a UV/Vis spectrophotometer to obtain the intramolecular quadruplex structure. The 400  $\mu\text{M}$  solution of complex **1** was then titrated, 3  $\mu\text{L}$  at a time, to a 3  $\mu\text{M}$  solution of the G-quadruplex in the abovementioned buffer. A cuvette filled with the sodium phosphate buffer solution was used as a blank and subtracted from previously recorded spectra. This experiment was also repeated in  $\text{K}^+$  based buffer with similar results.

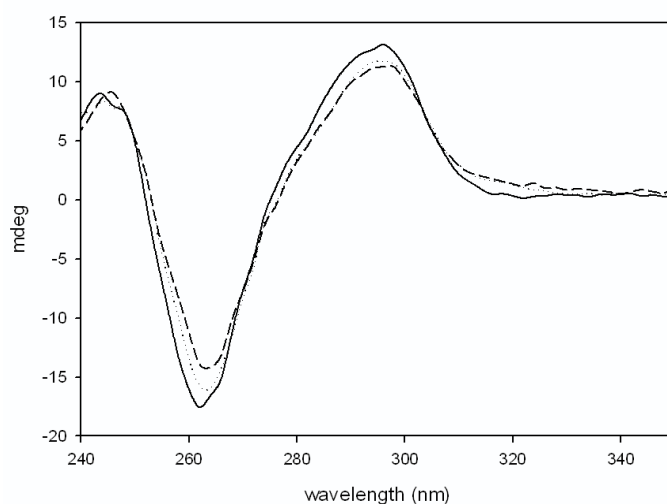


**Figure A.3.** Titration of complex **1** with a 3  $\mu\text{M}$  solution of G-quadruplex in a sodium phosphate buffer. The solid black line represents the G-quadruplex with no complex added. There is a slight decrease in the peak near 300 nm and an increase in the signal around 250 nm.



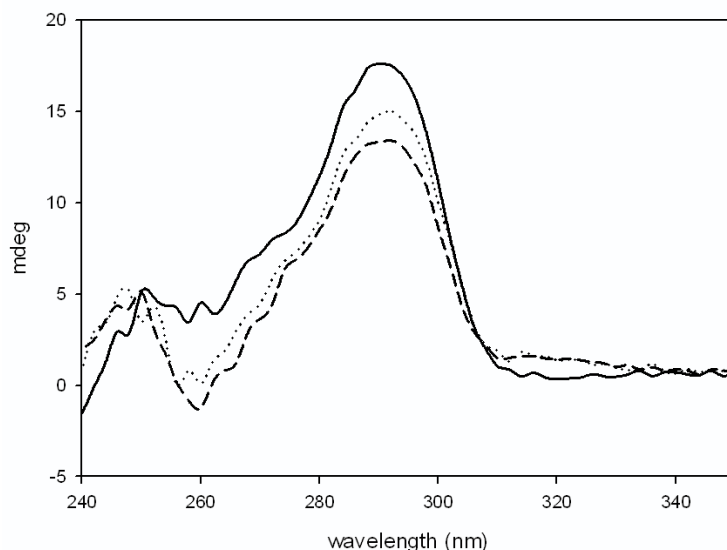
**Figure A.4.** Titration of complex **1** with a 3  $\mu\text{M}$  solution of G-quadruplex in potassium phosphate buffer. Solid black line is the G-quadruplex without complex **1**. Successive additions of complex **1** lead to a slight decrease in the peak at 295 nm with each aliquot.

In a subsequent experiment, complex **1** (3  $\mu\text{M}$ ) was added to the G-quadruplex (3  $\mu\text{M}$ ) after its formation and a CD spectrum was recorded. Then, this solution was heated to 90°C and allowed to cool to room temperature, where another CD spectrum was recorded. This experiment was conducted with aforementioned sodium and potassium based buffers with different results shown below.



**Figure A.5.** Heating and cooling of complex **1** and the G-quadruplex in a sodium phosphate buffer. The solid black line represents G-quadruplex after annealing. The dotted line is the result of the

addition of **1** to the G-quadruplex. The dashed line arises from the heating and cooling of the mixture of the G-quadruplex and **1**. The signal for the intramolecular G-quadruplex is retained.



**Figure A.6.** Heating and cooling of platinum square **1** and G-quadruplex in a potassium phosphate buffer. Black solid line represents the annealed G-quadruplex. The dotted line shows the effect of addition of the platinum molecular square **1** to the annealed G-quadruplex. The dashed line is the result after heating the platinum molecular square with the G-quadruplex and cooling the mixture to room temperature. There appears to be a slight change in the conformational preference from the original quadruplex structure upon the addition of the platinum square complex **1**. However, after heating and cooling this signal is also maintained.

## A.5 References

1. Fujita, M.; Yazaki, J.; Ogura, K., Spectroscopic Observation Of Self-Assembly Of A Macrocyclic Tetranuclear Complex Composed Of  $Pt^{2+}$  And 4,4'-Bipyridine. *Chem. Lett.* **1991**, 1031-1032.
2. Basolo, F.; Bailar, J. C.; Tarr, B. R., The Stereochemistry of Complex Inorganic Compounds. X. The Stereoisomers of Dichlorobis-(ethylenediamine)-platinum (IV) Chloride. *J. Am. Chem. Soc.* **1950**, 72, 2433-2438.
3. Autexier, C.; Pruzan, R.; Funk, W. D.; Greider, C. W., Reconstitution of human telomerase activity and identification of a minimal functional region of the human telomerase RNA. *EMBO Journal* **1996**, 15, 5928-5935.
4. De Cian, A.; Cristofari, G.; Reichenbach, P.; De Lemos, E.; Monchaud, D.; Teulade-Fichou, M. P.; Shin-ya, K.; Lacroix, L.; Lingner, J.; Mergny, J. L., Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, 104, 17347-17352.