LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

TRINITY COLLEGE DUBLIN

Development and application of novel virtual High Throughput Screening (vHTS) Technologies

Giorgio Carta

PhD Thesis

TRINITY COLLEGE
05 AUG 2008
LIBRARY DUBLIN
TH081S
8535

DECLARATION

This thesis has not been submitted as an exercise for a degree at any other university. Except where stated, the work described therein was carried out by me alone.

I give permission for the Library to lend or copy this thesis upon request.

Signed:

Giorgio Carta

"All truths are easy to understand

once they are discovered;

the point is to discover them"

(G.Galilei)

To my parents.

Summary

The development of new pharmaceutical drugs is an expensive, time-consuming task with an often low success rate. Traditionally, once a therapeutic target and its role in disease is identified, the process of identification of new drugs was performed through combinatorial chemistry. This is a technique for rapidly and systematically assembling a variety of molecular entities, or building blocks, in many different combinations, to create tens of thousands of diverse compounds that can be tested or screened with *in vitro* assays to identify potential valuable candidates. Chapter 1 gives a general overview of vHTS techniques and algorithms with particular emphasis to those methods that will be used later in the following chapter.

Chapter 2, in the context of virtual screening, investigates and evaluates different tools designed for the generation of molecular 3D structures. The chapter investigates the effect that different SMILES (Simplified Molecular Input Line Entry System) strings permutations and other classical 2D structural formats have on chemical databases to be screened. In particular, the effect on the generation of conformers and their effect within the context of vHTS is explored.

Chapter 3 describes the development of parameters and techniques needed to carry out a successful virtual screen for ligands that bind to Peroxisome Proliferator-Activated Receptor (PPARγ), a nuclear receptor of particular interest due to its role played in varies diseases including Type II Diabetes. A biological introduction of PPAR is included at the beginning of the chapter. The outcome of this chapter is the identification of a novel putative agonist for the PPARγ receptor and a Structural Activity Relationship (SAR) of a class of known agonists. A biological *in vitro* assay was used to determine the activity of the novel compound identified is presented.

Whilst Chapter 2 and 3 were focused on pre-processing techniques, Chapter 4 explores how ligand bias can contribute significantly to the number of false positives observed in a virtual screening campaign. A post-processing normalisation

procedure to reduce the dependence of certain scoring functions on molecular weight was developed.

Chapter 5 finally introduces a new statistical protocol, combined with a virtual combinatorial enumeration, to post-process, optimize and decorate the ranked output list obtained from an enriched set of data such as the one produced by a scoring function. The aim is to create a solid post-processing method to enhance the fit of novel molecules within a receptor pocket.

Acknowledgments

I wish to express my honest gratitude to my supervisor Dave for giving me the amazing opportunity to carry out my research within his fantastic group!

A special thanks goes to the people that have been orbiting around my desk for the last three years, in no particular order!: Darren, Andy, Georgia, Mr. Yang, Tim, Nathalia, Gemma, Paul and the folks just arrived in the Lab: Martin, Nick, Miriam, Trevor… this would not be happened without the support and help of all of you.

I would like to express a special thanks to Vale for her constant support, friendship and honesty. She has been a lead and an inspiration not only those last three years, but since we know each other (more than 10 years!!!). Huge thanks!!!

A maffiss thanks goes to the all Department of Biochemistry… and a name came up in my mind most than any other: Thanks Mary Pat! You are wonderful!

To my wonderful girlfriend Aga, I thank you for your love, support, patient and understanding!

Thanks to the lads in Italy: Simone, il Nonno, Perseu… *siete amici unici!*

To my mother, my father, my brother and sister I dedicate this work. Long those three years in Ireland I missed all of you! Thanks for supporting and believing in my decision!

List of publications

- ***Considerations in Compound Database Preparation-"Hidden" Impact on Virtual Screening.***
  Results.Andrew J. S. Knox, Mary J. Meegan, Giorgio Carta and David G. Lloyd .
  J. Chem. Inf. Model., 45 (6), 1908 -1919, 2005. 10.1021/ci050185z S1549-9596(05)00185-3

- ***Permuting input for more effective sampling of 3D conformer space***
  Giorgio Carta, Valeria Onnis, Andrew J. S. Knox, Darren Fayne and David G. Lloyd
  Comput Aided Mol Des (2006) 20:179–190
  DOI 10.1007/s10822-006-9044-4

- ***On the Effects of Permuted Input on Conformational Sampling of Drug-like Molecules: an Evaluation of Stochastic Proximity Embedding.***
  Dimitris K. Agrafiotis, Deepak Bandyopadhyay, Giorgio Carta, Andrew J. S. Knox and David G. Lloyd
  Chem Biol Drug Des 2007; 70: 123–133
  DOI: 10.1111/j.1747-0285.2007.00538.

- ***Unbiasing Scoring Functions: A New Normalization and Rescoring Strategy***
  Giorgio Carta, Andrew J. S. Knox, and David G. Lloyd
  J. Chem. Inf. Model. 2007, 47, 1564-1571

Abbreviations

3D: three dimension

AA: Amino Acid

AD: Alzheimer's Disease

AMBER: Assisted Model Building with Energy Refinement

AP: Adipocyte specific lipid binding Protein

AR: Androgen Receptor

AUC: Area Under Curve

BAT: Brown Adipose Tissue

CADD: Computer-Aided Drug Design

CAP: c-Cbl-associated protein

CBP: CREB Binding Protein

CNS: Central Nervous System diseases

DBD: DNA Binding Domain

DC: Divide and Conquer

DFS: Depth-First Search

DR: Direct Repeated sequences

EPA: Eicosapentaenoic Acid

ER: Estrogen Receptor

FDA: Food and Drug Administration

FFA: Free Fatty Acid

FN: False Negative

FP: False Positive - Fluorescent Polarization

FRED: Fast Exhaustive Rigid Docking

FRODA: Framework Rigidity Optimized Dynamic Algorithm

GLUT: Glucose Transporters

GPCR: G-Protein Coupled Receptor

GR: Glucocorticoid Receptor

GyK: *Glycerol Kinase*

HB: Hydrogen Bond

IND: Investigational New Drug

L: Ligand

LBD: Ligand Binding Domain

LBVS: ligand based virtual screening

logBB:Log of Brain/Blood partition coefficient

LPC: Ligand Protein Contact

MC: Monte Carlo

MDS: Multi Dimensional Scaling

MF: Molecular Forcefield

MMFF: Merck Molecular Force Field

MR: Mineral-corticoid Receptor

MS: Multiple Scleroris

MW: Molecular Weight

NASH: Non Alcoholic Steatohepatitis

NLM: Non-Linear Mapping

NMR: Nuclear magnetic resonance

NR: Nuclear Receptor

PCA: Principal Component Analysis

PD: Parkinson's Disease

PDB: Proteing Data Bank

PEPCK: phosphoenolpyruvate dehydrogenase

PL: Protein Ligand

PMF: Potential of Mean Force

PP: Pipeline Pilot

PPAR: Peroxisome Proliferator Activated Receptor

PR: Progesterone Receptor

QSAR: Quantitative Structure Activity Relationship

R: Receptor

RAR: Retinoic Acid Receptor

RD: Repressor Domain

RID: Nuclear Receptor Interaction domain

RL: Receptor-Ligand

RMSD: Root Mean Square Deviation

ROC: Receiver Operating Characteristic

RXR: Retinoic X Receptor

SBVS: Structure Based Virtual Screening

sdf: Structural Data File

SIFt: Structural Interaction FingerprinT

SMARTS: SMiles ARbitrary Target Specification

SMILES: Simplified Molecular Input Line Entry System

SPE: *Stochastic Proximity Embedding*

SR: Steroid Receptor

Tc: Tanimoto Coefficient

tdt: Thor Data Tree

Tdt: Thor Data Tree

TN: True Negative

TP: True Positive

TR: Thyroid Hormone Receptor

TS: Triangle Smoothing

TZD: Thiazolidinediones

UCP: UnCoupling Protein

VDR: Vitamin D Receptor

vdW: Van der Walls

vHTS: Virtual High Throughput screening

VS: Virtual Screening

WAT: White Adipose Tissue

WDI: World Drug Index

WOMBAT: *WOrld of Molecular BioAcTivity*

ZINC: ZINC Is Not Commercial

Table of contents

Summary	i
Acknowledgements	iii
List of Publication	iv
Abbreviations	v
Table of contents	ix
Table of Figures	xiv
Table of Tables	xix

Chapter 1

VIRTUAL HIGH THROUGHPUT SCREENING (VHTS): AN OVERVIEW 1

1.1 PART I: DRUG DISCOVERY AND THE ADVENT OF VHTS 3
1.1.1 OVERVIEW OF THE DRUG DISCOVERY PROCESS 3
1.1.2 DOCKING 5
1.1.3 SCORING 7
1.2 PART II: HIGH-THROUGHPUT SCREENING PROTOCOL 9
1.2.1 SOURCE OF 3D STRUCTURES OF RECEPTOR-LIGAND COMPLEXES 11
1.2.2 RECEPTOR PREPARATION 12
1.2.3 DATABASE PREPARATION: PRE-FILTERING AND PRE-PROCESSING 13
1.2.4 DOCKING ALGORITHM AND TOOLS 22
1.2.5 SCORING PROCEDURES 26
1.2.6 SCORING FUNCTIONS DETAILS 28
1.2.7 POST-PROCESSING METHODS 36
1.2.8 VALIDATION METHODS 40
1.3 REFERENCES 48

Chapter 2

PERMUTING INPUT STRINGS FOR MORE EFFECTIVE GENERATION OF 3D CONFORMERS. 57

2.1 ABSTRACT 59
2.2 INTRODUCTION 60
2.3 METHODS 64
2.3.1 PREPARATION OF LIGANDS 65
2.3.2 CONFORMER GENERATION 66
2.3.3 FINGERPRINT ANALYSIS 69
2.3.4 PRINCIPAL COMPONENT ANALYSIS 69
2.3.5 ALTERNATIVE 2D REPRESENTATIONS: CONNECTION TABLES. 70
2.4 RESULTS AND DISCUSSION 72
2.4.1 FINGERPRINT ANALYSIS 81

2.4.2 CONNECTION TABLE ANALYSIS. 83
2.5 CONCLUSION 84
2.6 FURTHER APPLICATIONS 86
2.6.1 METHODS FOR SPE EVALUATION 86
2.6.2 SPE CONFORMERS GENERATION 87
2.6.3 DISCUSSION AND RESULTS 89
2.7 REFERENCE 92

Appendix to Chapter 2

SCRIPTS 95
DISTANCE GEOMETRY 102
GENERATION OF INTERATOMIC DISTANCE 103
ASSIGNMENT OF A RANDOM VALUES AND EMBEDDING 106
REFINEMENT 109

Chapter 3

APPLICATION OF VHTS TECHNOLOGY: DISCOVERY OF NOVEL PPAR GAMMA LIGANDS 113

3.1 INTRODUCTION 115
3.2 NUCLEAR RECEPTOR 116
3.3 PPARGAMMA 118
3.4 BIOLOGICAL RELEVANCE OF PPAR GAMMA 120
3.4.1 ADIPOGENESIS, DYSLIPIDEMIA AND PPAR GAMMA 120
3.4.2 DIABETES AND PPARGAMMA 123
3.4.3 CANCER AND PPAR GAMMA 124
3.4.4 NEUROPROTECTION AND PPAR GAMMA. 125
3.5 PPARGAMMA LIGANDS 126
3.5.1 SYNTHETIC LIGANDS 126
3.5.2 DUAL AGONISTS 130
3.5.3 PARTIAL AGONISTS 132

3.5.4 NATURAL LIGANDS 133
3.6 PPARGAMMA LBD 135
3.7 COMPUTATIONAL METHODS 141
3.7.1 VALIDATION SET AND SCREENING SET 141
3.7.2 PRE-PROCESSING 142
3.7.3 GENERATION OF CONFORMERS 143
3.7.4 FRED: RIGID DOCKING 144
3.7.5 CONNOLLY AND GAUSSIAN SURFACES 144
3.8 BIOLOGICAL METHODS 145
3.8.1 METHODOLOGY 145
3.8.2 FP COMPETITIVE BINDING ASSAY 145
3.9 RESULTS 147
3.9.1 VALIDATION – LIGAND RE-INSERTION 147
3.9.2 SCREENING 150
3.10 DISCUSSION 153
3.11 FUTURE WORK 160
3.12 **REFERENCE 161**

Appendix to Chapter 3

LIST OF PPARGAMMA AGONISTS FROM WOMBAT 167
OMEGA 2.0 LIST OF PARAMETERS 174
FRED LIST OF PARAMETERS 176
COMPETITIVE BINDING EXPERIMENTS AND FP 182

Chapter 4

CHAPTER 4: UNBIASING SCORING FUNCTIONS: A NEW RE-SCORING AND NORMALISATION STRATEGY 189

4.1 ABSTRACT 191
4.2 INTRODUCTION 192
4.3 COMPUTATIONAL METHODS 195
4.3.1 MULTI-CONFORMERS DATABASE PREPARATION 195

4.3.2 ANALYSIS OF THE RESULTS 201
4.4 RESULTS 202
4.5 DISCUSSION AND CONCLUSIONS 214
4.6 REFERENCE 216

Chapter 5

POST-PROCESSING OPTIMISATION AND COMBINATORIAL DECORATION BASED ON SCORING FREQUENCY 219

5.1 ABSTRACT 221
5.2 INTRODUCTION 222
5.3 COMPUTATIONAL METHODS 228
5.3.1 DOCKING/SCORING METHOD. 228
5.3.2 MOLECULAR FRAGMENTATION AND FREQUENCY SELECTION. 229
5.4 RESULTS 233
5.5 DISCUSSION 238
5.6 FUTURE WORKS 239
5.7 REFERENCES 240

Table of Figures

Chapter 1

Figure 1-1. In rigid docking a library of ligands is pre-generated (B) and then the best fit for each conformer is found within the receptor (C). 6

Figure 1-2. In Flexible docking a fragment of the molecule (seeding phase) is first placed within the receptor (A) and further extended by additional blocks (B and C). 6

Figure 1-3. Outline of a SBVS protocol (see the text for more detail). 10

Figure 1-4. General maximum dissimilarity algorithm. Database represents the collection of molecules prior the clusterisation. 16

Figure 1-5. MW distribution for a set of known active compounds for the 7 receptors in the box (pdb Id entry) 18

Figure 1-6. MW distribution of ZINC before and after filtering against the MW distribution of PPARγ agonists. 19

Figure 1-7. PCA results: the red sphere represents the chemical space occupied by the whole SPECS database. Blue represents the SPECS after filtering and a substructure search. 20

Figure 1-8. Different systematic docking algorithms. 1. Library of pre-computed conformers can be docked in a rigid fashion in the binding pocket. 2. In the incremental algorithm a fragment is first seeded in the binding pockets and incremental re-construction of the ligand follow. 3. In the hammerhead fragments are placed with regards to their affinity and in a second stage "chained"-back together. (Sketch adapted from Böhm[95]). 24

Figure 1-9. On the left the ligand is free to undergo translational motion - a high entropy situation. After binding to the receptor the system is highly ordered - a low entropy complex. 27

Figure 1-10 (A-B-C). Representation of the main contributions to a force-filed based scoring function and their relative graphs schematic demonstration. 32

Figure 1-11 (D-E). Representation of the main contributions to a force field based scoring function and their graphical schematics. 33

Figure 1-12. Example ROC curve. See text for details (from http://gim.unmc.edu). 42

Figure 1-13. SIFt analysis for the Chemgauss scoring function. 45

Figure 1-14. Scheme of FP competitive binding assay. (Figure from www.invitrogen.com). 46

Chapter 2

Figure 2-1. Classification of conversion programs applied and their associated SMILES strings Generated. 62

Figure 2-2. Protocol for the evaluation of SMILES string permutations. 65

Figure 2-3. Potential energy graph for conformer set A, generated using omega (280 members). Black line illustrates result for permuted input string, grey illustrates result for canonical. 77

Figure 2-4. PCA analysis of conformers generated by Omega (set A): in violet the 280 conformers produced from the canonical string of 4DFR are shown. In green the conformers generated from 8 permutation of the same string. In red the crystal structure of the ligand of 4DFR.78

Figure 2-5. The Potential energy graph for conformer set B, generated using OMEGA (ca. 2,200 members). The black line illustrates the result for the permuted input string, the grey illustrates result for the canonical. 80

Figure 2-6. PCA of conformers generated by OMEGA (set B). In violet the conformers (about 1200) generated setting an energy window equal of 60 and a RMSD cut off of 0.5 Å from the canonical string of 4DFR. In green the conformers generated from the 8 permutations of 4DFR keeping all the parameters in OMEGA as default (with the exception of –max_conf, that was set to 310 to generate a comparable number of final conformers with the canonical set. In red the PCA coordinate calculated for the crystal structure is shown..... 80

Figure 2-7. Fingerprint analysis of conformers generated by Omega, Catalyst and Rubicon. The height of the 3 bar graphs is proportional to the total amount of conformers generated. Cansmi: Canonical SMILES string. 82

Figure 2-8. Fingerprint analysis of conformers generated by Corina..... 82

Figure 2-9. Number of unique conformations (within 1 Å RMSD) discovered at the end of each SPE trial for the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (F) 1GLQ. (E) Five independent runs of the first SDF permuted input of 4DFR, each starting from a different random seed..... 90

Appendix to Chapter 2

Figure 2-A. Algorithm used to permutate the sdf connection table in elaborate_sdf.pl..... 100

Figure 2-B. Scheme illustrating how the sdf file is divided in blocks by elaborate_sdf.pl 101

Figure 2-C. Five-carbon fragment. The angle between each bond is assumed to be 120º and the bond C-C to have a length of 1.3 Å (hydrogens are not displayed). 102

Figure 2-D. Distance d1,3 depends by length of bonds l1,2 and l2,3 and the angle α. 103

Figure 2-E . The length of distance between atoms 1 and 4 will depends on the lengths between atoms 1,2,3,4 (ln) and angle θ 104

Figure 2-F. Distance between atoms 1,5 cannot be lower than the sum of the vdW radii of the atoms 1,5 (in red). 104

Figure 2-G. The maximum distance between atoms 1,5 (labeled A and C) or the upper limits (1,5) is equal to the sum of AC and BC. 106

Figure 2-H. Conformation of the five-carbon fragment generated by distance geometry 108

Figure 2-I. The stereochemistry about tetrahedral atoms can be maintained with an appropriate chiral constraint. 1,2,3,4 represent vectors v1 ,v2 ,v3 ,v4. 112

Chapter 3

Figure 3-1. PPARγ role in the activation of different genes. 122

Figure 3-2. PPARγ LBD bound with Rosiglitazone (1FM6). TYR-473, HIS-323 and HIS-449 are labelled. 137

Figure 3-3. ROCs curve obtained from docking in PDB 1NYX, ranked using chemgauss2 scoring function. 148

Figure 3-4. Superimposition of the original Ragaglitazar as in the PDB entry 1NYX (red) and the original re-docked ligand using FRED (cyan) within the binding pocket. The blue box represents the area in which FRED performed the docking. 149

Figure 3-5. Superimposition of the original Ragaglitazar as in the PDB entry 1NYX (red), the re-docked using FRED (cyan) and the closest conformers created and subsequently docked (gray). 149

Figure 3-6. Docked structure of compound 79 into PPARγ LBD. 151

Figure 3-7. FP binding assay curve of compound 79. An IC_{50} of 7 μM was obtained. Test was run twice in duplicate.(MP: millipolarization units) 152

Figure 3-8. Comparison between Compound 79, Rivoglitazone (a selective agonist), JTT-501 (dual agonist PPARγ and PPARα.) and Ragaglitazar (ligand in PDB 1NYX). 154

Figure 3-9. Ligand- Receptor Interactions for the Ragaglitazar - 1NYX crystal structure (top). Putative Compound 79- Receptor Interaction of the binding mode predicted by FRED (bottom) 156

Figure 3-10. Gaussian surface calculated on ligand 79 and residue HIS323of PPARγ LBD.Green surface: hydrophobic area; purple: HBs; blue: polar area. The distance between HIS323 and the sulphur atom in compound 79 is 3.19Å. 157

Figure 3-11. Site P1 highlighted with a Gaussian surface; left Ragaglitazar (ligand in PDB 1NYX). Right: compound 79 as predicted by the docking. 158

Figure 3-12. Site P2 highlighted with a Gaussian surface; left: Ragaglitazar (ligand in PDB 1NYX); Right: compound 79 as predicted by the docking. 159

Figure 3-13. Site P4 highlighted with a Gaussian surface; left Ragaglitazar (ligand in PDB 1NYX). Right: compound 79 as predicted by the docking. 159

Appendix to Chapter 3

Figure 3-A. Example of Scatchard Plot. See text for details. 184

Figure 3-B. Example of a time dependent saturation experiment 185

Figure 3-C. The IC_{50}, the concentration of substance that provides 50% inhibition of a certain reaction 186

Chapter 4

Figure 4-1. Descriptor (y axis) plotted against ranked position (x axis) 194

Figure 4-2. Correlation matrix of 12 descriptors. For example the correlation between molecular weight (labelled as MW) and a count of rotatable bond (opr_nrot) is highlighted in orange. a_IC : Atom information content (total); rings: number of rings; a_heavy: number of heavy atoms; a_nC: number of carbon atoms; Zagreb: Zagreb index; opr_brigid: number of rigid bonds; opr_nrot: number of rotatable bonds; Kier1: first kappa shape index; apol: sum of the atom polarizablilities; SMR: Molecular refractivity; vdw_vol: van der Waals surface. Complete documentation about the descriptors used above may be found in the MOE reference manual. 199

Figure 4-3. Plot of the function $y=(1-x)^{0.25}$ 200

Figure 4-4. Effect of the re-scoring procedure on the molecular-weight distribution plotted versus the rank position for the Shapegauss scored set. Top part: before the normalisation; bottom part: after normalisation. Molecules with a molecular weight above 500 are highlighted in both the plots. 202

Figure 4-5. The enrichments curves before and after re-scoring 204

Figure 4-6. Molecular weight versus the rank position plotted for all scoring functions. 206

Figure 4-7. Score distributions for all the scoring functions considered. Set A: before normalisation, Set B: after normalisation. 207
Figure 4-8. (Continuous). 208
Figure 4-9. (Continuous). 209
Figure 4-10. SIFt analysis results for all the scoring functions 211
Figure 4-11. (Continuous). 212

Chapter 5

Figure 5-1. Fragmentation scheme. 225
Figure 5-2. Fragments are classified as scaffolds or lateral chains according to the rules described in the Methods section. 226
Figure 5-3. In our example, the ensemble of scaffolds labeled A2 are more frequent and receive a better score than chains labeled A1. Scaffold A2 are significant with respect of the total score distribution (in red) and therefore will be selected as a privileged scaffold. Chains A1 will be discarded as they appear to be insignificant. Chains B1 represents another hypothetic structure selected to be privileged by the scoring function. 226
Figure 5-4. Statistically selected chains and scaffolds are recombined together. 227
Figure 5-5. Outline of the protocol 227
Figure 5-6. An example of fragmentation. Group A represents the fragments identified as chain and the group B those fragments identified as scaffolds. "As" represents a dummy atom. 231
Figure 5-7. Score distribution generated from calculation A. 233
Figure 5-8. Score distribution generated from calculation B. 235
Figure 5-9. ROC curve computed for the original set in calculation B. An accuracy of 0.752 (1 represents the max value) in identifying the seeds is calculated. 236
Figure 5-10. ROC curve computed for the combinatorial set in calculation B. An accuracy of 0.604 (1 represents the max value) in identifying the seeds is calculated. 236
Figure 5-11. Known actives successfully reassembled by the post-processing method in calculation B. 237
Figure 5-12. Top two molecules of the decoy set successfully reassembled by the post-processing method in calculation B. 238

Table of Tables

Chapter 1

Table 1-1. Report from a pre-filtering protocol: PDB code: RCSB database ID of the protein; Receptors: common name of the protein; Retrieved: known active retrieved. Selectivity Filter: selective exclusively towards that receptor; Bioeffect Filter: agonists/antagonists depending on the receptor considered. Lipinski Filter: compounds passing the Lipinski drug-like filter.18

Table 1-2. Scoring Functions commonly used in Virtual Screening. 29

Table 1-3. In a false positive a test wrongly shows an effect or condition to be absent. Conversely, in a false positive a test wrongly shows an effect or condition to be present. In true positive or a false positive the test is exact in the prediction. 41

Table 1-4. Hypothetical screening results considering 10% (100) of a database of 1000 molecules. 42

Chapter 2

Table 2-1. List of protein ligands complexes extracted from the PDB and their associated resolutions.[A] Specifies the number of permutations generated by 'permsmi' (Daylight Inc.).[B] Indicates the number of strings selected to carry out the analysis, including the canonical SMILES string, as generated by Mol2smi (daylight Inc). 62

Table 2-2. RMSD Standard Deviation for Omega.[a] can = canonical SMILES string..... 74

Table 2-3. RMSD Standard Deviation for Rubicon[a] can = canonical SMILES string 74

Table 2-4. RMSD Standard Deviation for Catalyst.[a] can = canonical SMILES string..... 74

Table 2-5. Minimum RMSD for Omega. [a] can = canonical smile string 76

Table 2-6. Minimum RMSD for Rubicon. [a] can = canonical smile string..... 76

Table 2-7. Minimum RMSD for Catalys.[a] can = canonical smile string 76

Table 2-8. Minimum RMSD for Corina. [a] can = canonical smile string 76

Table 2-9. RMSD minimum, maximum, average and standard deviation calculation for the pdb entry 4DFR4DFR: the table shows the set where 280 conformers were produced..... 77

Table 2-11. RMSD minimum, maximum, average and standard deviation calculation for 4DFR: the table shows the set where 280 conformers were produced using one Connection Table ('One CT' column) and 10 permutation of a single connection table (Permuted column). Unit are expressed in Å..... 83

Table 2-12. PDB codes, structures, number of rotatable bonds, number of SPE trials and number of boosts for each of the molecules in the data set. 86

Table 2-13. ANOVA of the conformational properties of the ensembles produced by the permuted inputs for each molecule under investigation (when F is less than Fcrit, the differences are considered to be statistically insignificant) 91

Appendix to Chapter 2

Table 2-A. Initial bounds matrix. Cyan cells indicate the lower limits. Orange cells indicate the upper limits. (all the units are in Å). 103

Table 2-B. Resulting distance matrix after TS of the matrix in Table 1. (all the units are in Å). 106

Table 2-C Randomly assigned distances between the upper and the lower bounds indicated in Table 2-B. (all the units are in Å) 107

Table 2-D. XYZ coordinates obtained after the embedding of matrix in table 3 107

Table 2-E. Interatomic distances of conformation in Figure 2-H (all the units are in Å). 108

Table 2-F. Updated matrix of table 2 with the actual distance found after embedding indicated in table 5 111

Chapter 3

Table 3-1. TZD derivatives. 127

Table 3-2. Cinnamate derivatives. 129

Table 3-3. Tyrosine derivatives 130

Table 3-4. Dual Agonists. 131

Table 3-5. GW-0072. 133

Table 3-6. Natural Ligands for PPARγ. 134

Table 3-7. PPARγ crystal structure available on the PDB databank[59] 136

Table 3-8. List of hot spots recognisable in PPARγ LBD identified by solvent mapping technique (taken from reference [62]). 138

Table 3-9. Putative Hydrogen bonds and their length (Å) predicted by LPC. In the columns the crystal structures are depicted and in the rows the main residue interactions identified as essential in each complex. 140

Table 3-10. ROCs values calculated for each scoring function available in FRED against eight different PDB entries for PPARγ. In dark yellow is highlighted the best combination. 147

Table 3-11. Neighbouring amino acids to Ragaglitazar and Compound 79 158

Chapter 4

Table 4-1. Structure of the 19 estrogens selected from the literature. 197

Table 4-2. Consensus enrichment obtained with the original scoring function versus the one obtained after normalisation. 213

Chapter 5

Table 5-1. Statistical details associated with the graph in Figure 8 - calculation **A** (alpha 0.05).234

Table 5-2. Statistical details associated with graph in figure 5B - Calculation **B** (alpha 0.05).... 235

1 Virtual High Throughput Screening (vHTS): an overview

1.1 Part I: Drug discovery and the advent of VHTS

The development of new pharmaceutical drugs is an expensive, time-consuming task with an often low success rate. According to recent surveys the average pre-approval development cost is estimated at $800 million per compound[1] with a seven year average time from initial Investigational New Drug (IND) filing until approval[2,3]. Traditionally, once a therapeutic target and its role in disease is identified the process of identification of new drugs was performed through combinatorial chemistry. This is a technique for rapidly and systematically assembling a variety of molecular entities, or building blocks, in many different combinations, to create tens of thousands of diverse compounds that can be tested or screened with *in vitro* assays to identify potential valuable candidates. Virtually every product of human genes and every pathogen can be a potential pharmacological target resulting in a combinational explosion if multiplied by the number of chemicals that can be synthesised. It is thus clear that a fully combinatorial approach is not feasible and a rational approach to pharmaceutical chemistry is required to lead and drive the drug discovery process. Hence, in recent years, there has been major investment in techniques to improve the discovery and development process with benefits in reduced cost and time.

1.1.1 Overview of the drug discovery Process

As with many disciplines, biochemistry and drug discovery have undergone an evolution due to the impact of computer science. In parallel with the increased performance and functionality of computer hardware and software, advances in protein crystallography and NMR (Nuclear Magnetic Resonance) have laid the foundation for the development of Computer-Aided Drug Design (CADD)[4-6]. The identification of the three-dimensional (3D) coordinates of protein-ligand complexes led to the development of a wide series of CADD techniques aiming to predict the interactions taking place in those complexes. A main area of the drug discovery process that has benefited by the introduction of computational techniques has been the initial screening of potential hit molecules to elucidate those that exhibit some activity against relevant pharmacological

targets through Virtual High Throughput Screening (vHTS). vHTS takes advantage of the great calculation power offered by modern computers and the 3D model of receptors provided by crystallography and NMR techniques. It is possible to replace or integrate combinatorial chemistry by simulating the synthesis of random molecules using informatics tools with "chemical" intelligence. Basically, by applying bio- and cheminformatics techniques the "hits" identification stage is shifted from *in vitro* to *in silico* testing. This is accomplished by simulating and modelling the biochemistry through the conversion of chemical atoms to binary sequences of zeros and ones. The final aim of *in silico* screening is the prediction and estimation of the ability, or affinity, of a set of molecules to bind to a target (e.g. a receptor). Subsequently, this simulation allows one to focus the *in vitro* screening towards just a small subset of compounds selected by the vHTS, in order to increase hit rates in the drug discovery process.[7-9]

vHTS can be divided into two categories: ligand-based (indirect methods) or structure-based (direct methods)[10] drug design. In ligand-based virtual screening (LBVS), properties and features of known active ligands against a target are determined. Based on the assumption that active molecules should have similar physical and chemical characteristics, these properties are used as a filter[5] or as guide towards the identification of similar compounds. LBVS can be divided into further classifications depending on the categories of properties calculated. There are be one-dimensional, such as number of hydrogen bond (HB) donors or acceptors and molecular weight (MW)[11,12], two-dimensional (substructure searching, identification of maximum common substrate)[13-15] and three dimensional properties (pharmacophore, 3D similarity, 3D fingerprints)[6,16-19]. The prevailing trait that gathers together these methodologies is that the 3D coordinates of the target protein is not required but rather a set of known active molecules is required, from which specific properties are derived.

In contrast, for structure-based virtual screening (SBVS) a more detailed knowledge of the target protein (usually derived from crystallographic studies or NMR data)[6,20] is required. A pre-requisite for this approach is an understanding of the rationale of molecular recognition in protein-ligand complexes. When the

3D structure of a given protein is known, this information can be directly used for the retrieval or design of new active molecules considering the receptor's coordinates. Analysis of the 3D complexes reveals the binding mode between a specific ligand and the ligand binding domain (LBD): non-covalent bonding interactions, ligand conformations, shape complementarities, secondary and tertiary structure changes upon binding of a ligand etc. A wide number of computational tools have been designed to help in this process through the visualisation and the calculation of the ligand-protein interactions. The knowledge gathered in this phase can be subsequently used as a query for the screening of large libraries of compounds[5]. For example, common methodologies used in SBVS include receptor based pharmacophore, molecular dynamics calculations or virtual protein-ligand docking. As docking is the predominant method employed in this work it shall now be discussed in further detail.

1.1.2 Docking

Docking techniques are extensively used today for drug discovery of new leads or, to a lesser extent, lead optimisation. Docking is defined as the prediction of the structure of receptor-ligand complexes, where the receptor is usually a protein or a protein oligomer and the ligand is either a small molecule or another protein, and the estimation of their binding affinity. Virtual protein-ligand docking is always a two-step methodology. In the first step, a conformational search of the ligand is performed and subsequently or simultaneously poses are generated within the binding site of the receptor[4]. The second step involves an evaluation of those poses and the estimation of the binding affinities; the so called scoring phase. With regards to the conformational search of degrees of freedom of the ligand, *rigid docking* and *flexible docking* methods can be distinguished.

In **Rigid docking** the search algorithm explores different positions for the ligand within the receptor. A pre-computed set of conformers are virtually individually introduced in the binding site, translated and rotated until a rational pose is generated. In other words, the conformer is generated prior to virtual

docking in the receptor, and only in a second step each conformer generated is fitted within the LBD, which is also one of the main disadvantages of rigid docking. On the other hand, the main advantage is the high speed of the algorithms utilised, which allows the screening of very large databases of compounds in a relatively short period of time. This method is outlined in Figure 1-1.

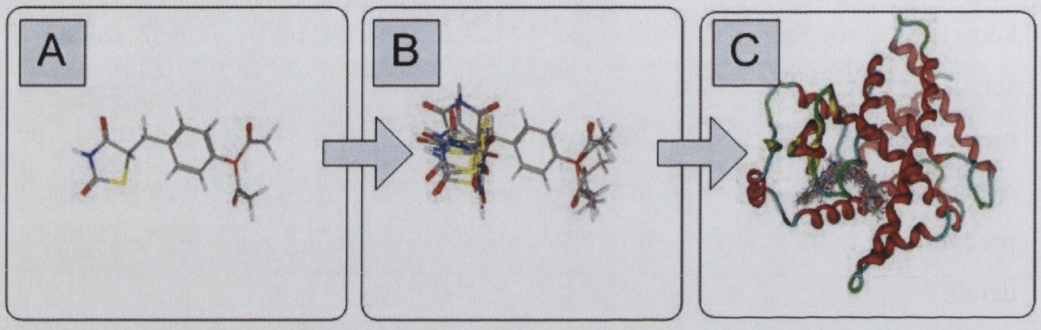


Figure 1-1 In rigid docking a library of ligands is pre-generated (B) and then the best fit for each conformer is found within the receptor (C).

Flexible docking adds exploration of ligand torsional degrees of freedom into the docking process within the confines of the binding pocket[21-24]. In a typical flexible procedure the ligand is first fragmented into rigid blocks. The fragments are then placed in the pocket and the molecule rebuilt exploring different conformational solutions. The process is depicted in Figure 1-2.

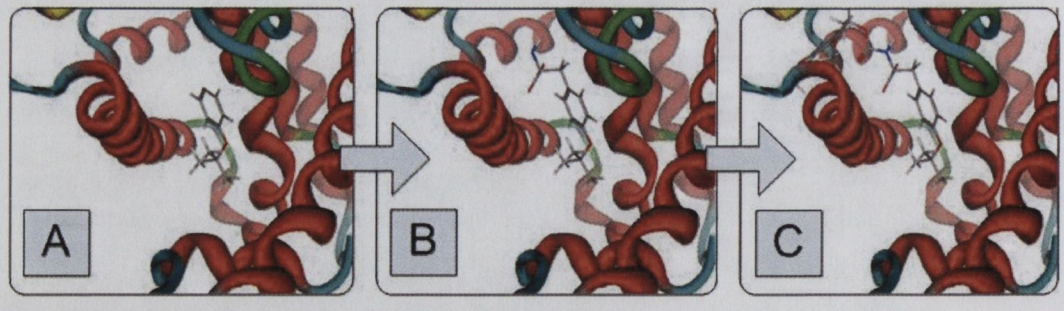


Figure 1-2 In Flexible docking a fragment of the molecule (seeding phase) is first placed within the receptor (A) and further extended by additional blocks (B and C).

Thus, with rigid docking the conformers are pre-computed "*outside*" the receptor, whereas in flexible docking the conformers are computed "*inside*" the receptor, taking into account the features of the receptor itself.

Flexible-flexible docking. As a key objective of docking is the prediction of the structure of protein-ligand complexes, methods that account for both ligand and receptor flexibility (flexible-flexible) have been investigated. Receptor flexibility can be considered during docking, either accounting for different conformations of the protein (from crystallographic data or from *in-silico* simulations) or using the so called *"induced fit"* method developed by Schrodinger[25]. Using the former method the ligand conformers are docked (rigidly or not) in different conformations of the receptor. Conversely, the latter is a method for the prediction of ligand induced conformational changes in receptor active sites by combining docking with protein structure prediction techniques[26]. The Methodology simulates the modification of the receptor tertiary structure by the docked pose of the ligand thus producing the *induced fit*.

1.1.3 Scoring

In the second so called "scoring" phase, the poses generated during the docking (either rigid or flexible) are evaluated and for each one a "score" is calculated, which represents an estimation of the binding affinity. In this step a mathematical algorithm, the "scoring function", is applied to recognise which poses yield the most stable interactions with the protein. Based on this estimation of the binding affinity the compounds are then ranked in an output list. A large number of scoring functions are available based on different methodologies, algorithms and developed using different tools. It becomes evident in the following paragraphs and in Chapter 4, that this may lead to completely different approaches and, even often, contradictory results[27-30]. Moreover, evidence exists that there is a strong correlation between the type of interactions and the scoring prioritisation[31]. Thus, a scoring function may perform impeccably for one receptor and at the same time poorly for another. More details about docking and scoring algorithms will be explored in the following sections.

In the second part of this introduction an overview of a typical vHTS protocol according with the established procedures[32] is outlined. The aim is to

provide examples of the different tools that are applied to virtual screening (VS) for novel pharmaceuticals.

1.2 Part II: High-throughput screening protocol

In this work, we have applied a virtual screening protocol to a validated target of therapeutic importance, peroxisome proliferator-activated receptor gamma (PPARγ), where several different pre-processing techniques were used to narrow down the number of molecules considered before the docking step. A typical protocol for VS is depicted in Figure 1-3.

A target protein is selected and 3D coordinates are taken from crystal structures available in the RCSB Protein Data Bank[33] or NMR results.

- A library of drug-like molecules is then selected and/or filtered.
- Conformers for each molecule in the dataset are produced and fitted within the receptor in the "docking" step.
- Poses are subsequently evaluated and their binding affinity estimated.
- In order to refine the result from the scoring, some post-processing steps are usually required.
- Finally, a number of hits are selected and tested with a biological assay *in vitro*.

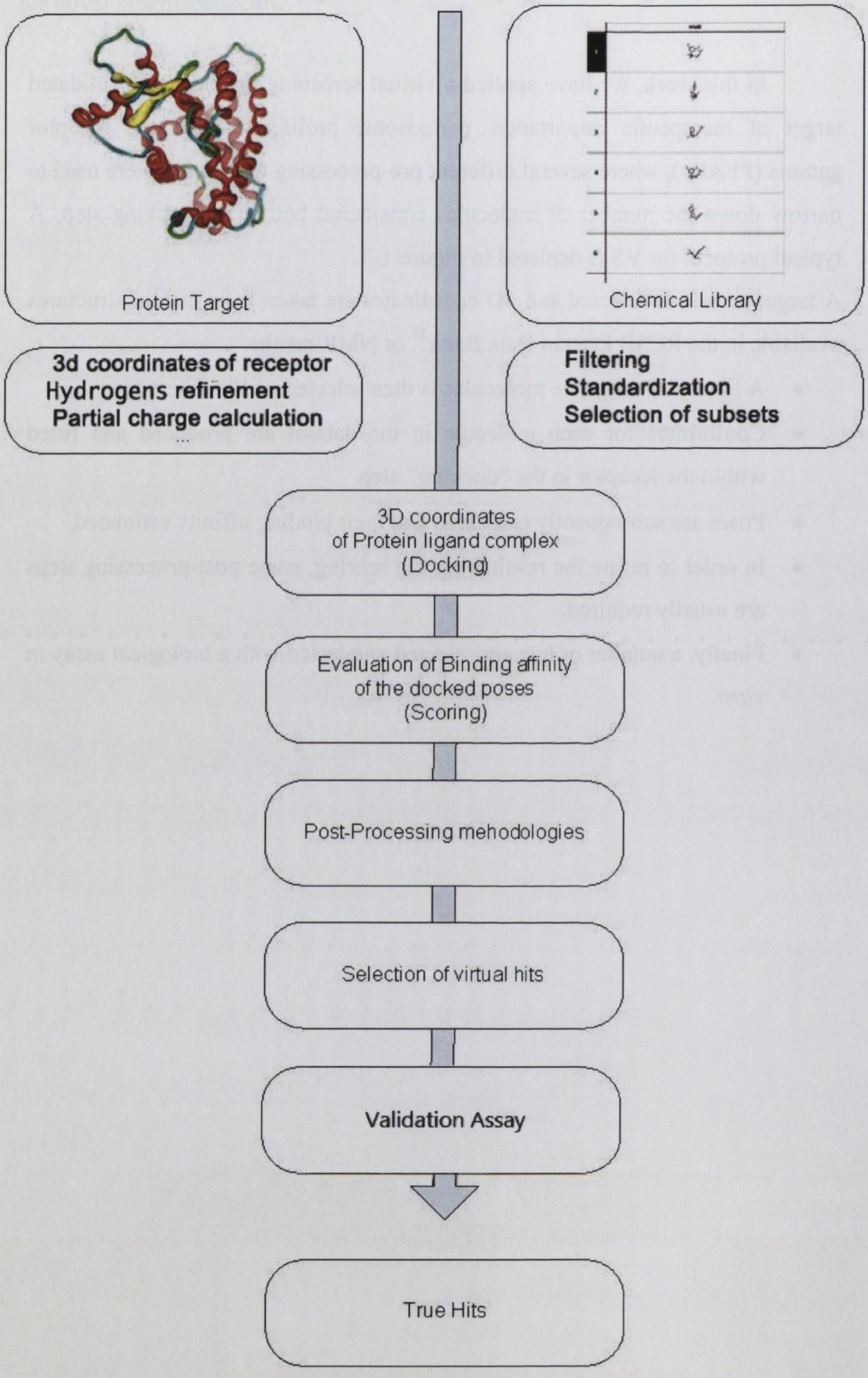


Figure 1-3 Outline of a SBVS protocol (see the text for more detail).

1.2.1 Source of 3D structures of receptor-ligand complexes

Major advances in protein expression and purification, sample preparation, microcrystallisation[34], use of synchrotron light[35] and automation[36,37] have led to an explosion in bio-macromolecular structure determination. Subsequently, in the past few years the protein data bank[33] (PDB) grew dramatically and presently over 47,625 protein structures are available, both in the unbound form and complexed with a variety of small ligands, other proteins or nucleic acids[6,36]. The PDB repository represents the backbone of SBVS as it is the main source of information on the atomic Cartesian coordinates of drug interactions with proteins. To perform vHTS, the first step is target selection. Proteins and their binding pockets must have some "druggable" properties suitable for drug design. For example, it is preferable a receptor with a small and well defined by crystallographic study binding site, which can increase the rate of success in SBVS.[38] Having a *snapshot* of a target protein, allows for the identification of cavities and putative binding pockets. If a static 3D representation of a receptor-ligand (RL) complex is available it can be used as a reference structure for ligand docking. The following questions can be posed: do the ensemble of conformations generated, during the docking, contain a structure similar to the bound form of the ligand as found in the crystallographic data[39] or, in other words, is possible to reproduce the crystal structure binding mode?

Additionally, estimations of protein flexibility can be extracted from the crystallographic data. The B-factors of protein crystal structures reflect the fluctuation of atoms about their average positions and provide important information about protein dynamics[40]. Furthermore, we can predict the flexibility of the main chains in the proteins. FRODA (Framework Rigidity Optimized Dynamic Algorithm), a software developed by Thorpe *et al.*, identifies first the rigid regions in the protein, and then replaced them by ghost templates which are used to guide the movements of the atoms in the protein[41].

Examining macro-molecular flexibility is critical for a thorough understanding of the principles that control ligand binding to proteins. Structural

changes are commonly induced in the receptor upon ligand binding[42]. Therefore by ignoring this effect it may have a strong impact on vHTS. A very common approach to incorporate receptor flexibility is the use of multiple receptor conformations derived from crystallographic data or as generated by such tools as FRODA. With regards to the first option, sources of different receptor conformations can usually be retrieved from the PDB. Different PDB entries of proteins complexed with ligands offer structural information about the mechanism underlying either the activation of the receptor or the antagonistic effect. For example, in nuclear receptor-like structures, a macromolecular rearrangement of the receptor upon the binding of a ligand can be detected by comparing the agonist bound and the antagonist bound PDB entry.

1.2.2 Receptor Preparation

As X-ray crystallography is not able to locate 3D coordinates of hydrogen atoms, they must be added manually to the protein (numerous software tools are available to carry out this task: MOE[43,44], SYBYL[44] are just a few examples). Furthermore, the position of hydrogen atoms must have the correct spatial orientation, according with the geometry of the root atom and their non-covalent interactions. Hence, a molecular force field minimisation step is also required to correct and refine the positions of hydrogen atoms in the receptor. Different molecular force fields are available: widely used are MMFF94 (Merck Molecular Force Field[45-49]) or AMBER99[50] (Assisted Model Building with Energy Refinement). AMBER is specifically parameterised for proteins and nucleic acids. Partial atomic charges are also added during this step to each atom of the protein since the PDB file does not include such information. The rational behind the partial charge calculation is that valence in molecules is not fixed to any one particular atom, but moves around the molecule as a whole. Since the electrons are likely to spend more time in proximity of more electronegative atoms, this distribution is not uniform in some parts of the receptors. Partial charges affect the scoring process, being part of columbic term used in evaluating the binding affinity of the ligand (see sections below).

1.2.3 Database Preparation: Pre-filtering and pre-processing

SBVS consists of the screening of large databases of drug-like molecules to determine potential hits able to bind to a target protein. A wide number of commercial, non-commercial and academic compound databases are available today (e.g. ZINC, SPECS, WOMBAT [51-53]). Furthermore, datasets of ligands are also available pre-focused on particular chemical-physical properties, such as those exploring particular regions of chemical space or having drug-like characteristics (e.g. natural compounds, G-Protein Coupled Receptor – GPCR libraries).

From a biological point of view, an ideal screening library should contain low toxicity, orally bioavailable compounds. This excludes peptides or proteins as they are usually broken down by peptidases during digestion. Depending on the tissue localisation of the target, one might be interested in compounds with a high tropism for a specific organ. Also, from a chemical prospective, the compounds in a virtual library must also be synthetically accessible or available for purchase.

Generally, those collections of compounds are already oriented and downsized by clustering or filtering. The former is needed to ensure coverage of the largest possible expanse of chemical space in the search for bioactive molecules without greatly increasing the size of the libraries. This is formulated by the "Similar Property Principle": if structurally similar compounds are likely to exhibit similar activity, then maximum coverage of the activity space should be achieved by selecting a structurally diverse set of compounds. To obtain chemical diversity one needs first to quantify the chemical diversity and then, based on the dissimilarity, a group of molecules can be divided into clusters where objects in the cluster are similar, but objects in other clusters are dissimilar[54-56]. This sort of calculation first requires a definition of *diversity*, or *similarity* which is the inverse of the diversity (typically, similarity lies in a range between 0 and 1, thus a similarity of 0.8 corresponds to a diversity of 0.2)[57].

Dissimilarity/Similarity searching involves the comparison of a set of descriptors that define or characterise a dataset of one or more molecules with the corresponding set of descriptors defining each structure in the database.

The two most important types of descriptors described are the "fragment substructures" and "physical properties". In the former case, a molecule is checked for the presence of various atom- or bond- centred fragment substructures and their presence encoded in a Boolean bit-string vector, or fingerprint. A common example is represented by the *daylight fingerprint*[58]. In detail the Daylight fingerprint algorithm examines the molecules and generates the following:

- A pattern for each atom
- A pattern representing each atom and its nearest neighbours
- A pattern representing each group of atoms and bonds connected by path up to two bonds long
- A pattern representing each group of atoms and bonds connected by path up to three, four, five, six and seven bonds long

The similarity between the two molecules is calculated by a comparison of the fingerprint. The "physical" descriptors depict inherent properties of the molecules, such as number of HB donors, logP, topological features, van der Waals (vdW) area or surface, polarisability and so on. Calculation of the similarity between two objects, each characterised by a common set of descriptors, can be achieved by using different coefficients.
As an example, a molecular object A can be defined by means of a vector X_A of n attributes such that:

$$X_A = \{x_{1A}, x_{2A}, x_{3A} \dots x_{jA} \dots x_{nA}\} \qquad 1\text{-}1$$

Where X_{jA} is the values of the j^{th} attribute of object A. The values of attributes may be a real number or bit. In the former we have a "physical properties" type of descriptors: values can any numerical value (i.e. MW). In the latter, a bit indicates the absence or the presence of a particular property or feature in the object and it can have value of 1 (feature is present) or 0 (feature is absent).

In the case of a molecular object the bits may also represent topological features of the molecules (Daylight fingerprint, extended connectivity fingerprint), physical properties or a combination of both of them[57]. Where the

attribute values are restricted to 0 or 1 (binary bits), the expressions used for the calculation of the similarity and distance between molecular objects can be substantially simplified[57]. For objects A and B characterised by vectors X_a and X_b containing n binary values (such as fingerprints) we can write:

$$a = \sum_{j=1}^{j=n} x_{jA} \qquad \textit{Number of bits "on" in A.} \qquad \textit{1-2}$$

$$b = \sum_{j=1}^{j=n} x_{jB} \qquad \textit{Number of bits "on" in B.} \qquad \textit{1-3}$$

$$c = \sum_{j=1}^{j=n} x_{jA} x_{jB} \qquad \textit{Number of bits "on" in both A and B.} \qquad \textit{1-4}$$

$$d = \sum_{j=1}^{j=n} \left(1 - x_{jA} - x_{jB} + x_{jA} x_{jB}\right) \qquad \textit{Number of bits "off" in both A and B.} \qquad \textit{1-5}$$

And hence:

$$n = a + b - c + d \qquad \textit{1-6}$$

Thus to measure the distance between object A and B, the Tanimoto coefficient (Tc), one the most commonly used coefficients, can be used (Equation. 1-7)

$$Tc = \frac{a}{(a + b - c)} \qquad \textit{1-7}$$

Tanimoto coefficient is most widely used for binary fingerprints. Other similar methods are the Dice coefficient, Cosine similarity or the Euclidean Distance.
In order to perform the clustering based on dissimilarity, various algorithms are available. As an example, the general steps of one of the most widely used algorithm called general maximum dissimilarity (Figure 1-4) are as follows.

1. Initialise subset by transferring a compound from *Database* to it
2. Calculate the dissimilarity between each remaining compound in *Database* and the compounds in *Subset.*
3. Select that compound from *Database* that is most dissimilar to *Subset* and transfer it to *Subset.*
4. Return to Step 2 if there are less than n compounds in *Subset.*

Figure 1-4 General maximum dissimilarity algorithm. Database represents the collection of molecules prior the clusterisation.

The initial molecule selected in the first step can be (i) randomly selected, (ii) the most dissimilar one from all the other or (iii) the nearest to the centre of the Database relatively to the other molecules. The algorithm will loop until either, the pre-defined number of clusters or an average number of molecules per cluster is reached.

Filtering is used to remove non drug-like compounds from a collection and to decrease the number of compounds in the docking process thus reducing the computational time required[59]. Even though the virtual vendor libraries are often already filtered, it is common to restrict chemical space to the area of interest by applying more filtering or pre-processing techniques. This is done to decrease the number of *in silico* false positives brought forward to biological tests *in vitro* and speed up the process by further downsizing the collections. Watson et al.[60] also indicated how the inappropriate selection of libraries may produce artificial results. They highlighted the necessity of consistency between the libraries and the "*target*" chemical space, usually identified through known active molecules. It is important to screen a library of compounds tailored towards compounds that display similar characteristics to known actives or ligands of a particular target[32] as shown previously. A large variety of filters are available, implemented in several software applications (FILTER[61], Pipeline Pilot[62]) and are widely customisable. Molecules containing specific functional groups, known to be toxic, can be discarded early in the screening protocol (alkyl-bromides, metals etc). Next, there are functional groups that are not strictly forbidden, but are not allowed in quantities of more the one. For instance, para-fluorobenzene, or trifluoromethyl have specific purposes, but heavily fluorinated molecules are usually eliminated.

Beyond simple functional group filtering, one can consider both simple and complex physical properties which can be used to filter compound libraries.

These filters attempt to consider "drug-likeness" properties, such as bioavailability, solubility, toxicity and synthetic accessibility before the primary vHTS. The best known of the physical property filters is Lipinski's "rule-of-five"[63], which focuses on bioavailability, where the molecules should not have:

- Molecular mass greater than 500 Da
- High lipophilicity (expressed as cLogP greater than 5)
- More than 5 HB donors
- More than 10 HB acceptors

However, many other physical properties, such as solubility, atomic content, ring structures and surface area ratios can also be considered.

Finally, compounds that can be problematic at later stages should be eliminated at this point. For instance, Shoichet's aggregating compounds often produce false positives in *in-vitro* assay[64]. It can save enormous resources if they were identified by virtual or high-throughput screening. Similarly (and sometimes concomitantly), dyes can appear to be inhibitors by interfering with colorimetric/fluorometric assays or binding non-specifically to the target protein[65]. Moreover, pre-filtering the database by focussing in on regions of chemical space containing known actives on one's protein of interest increases the likelihood of finding novel molecules that also target that protein[59].

We shall now explain how knowledge of known actives can be exploited (Table 1-1). WOMBAT is a compounds collection containing 154,236 entries, totalling over 307,700 biological activities on 1,320 unique targets. Known active ligands against a set of proteins for which the crystal structures were available[66-72], have been extracted from the WOMBAT[53] database ensuring the removal of duplicates. As described in Table 1-1 the original list was narrowed down from ~3000 (known active against the targets illustrated in Table 1-1) to 569 molecules: general drug-like filters were applied using components implemented in 'Pipeline Pilot'[62] (PP), a multi purpose data management and manipulation software platform. This selection of compounds will be also used in chapter 4.

	PDB code	Receptors	Retrieved	Selectivity	Bioeffect	Lipinski
1	1GLQ	GSTase	9	9	9	9
2	1AQ1	CDK2	1158	300	300	259
3	1DQ	PTPase	29	29	29	29
4	4DFR	DHFR	219	219	219	139
5	1I7I	PPARγ	587	149	79	58
6	1P93	GR	151	151	142	39
7	3ERT	Erα	903	167	59	30
TOT			3056	1024	837	569

Table 1-1 Report from a pre-filtering protocol: PDB code: RCSB database ID of the protein; Receptors: common name of the protein; Retrieved: known active retrieved. Selectivity Filter: selective exclusively towards that receptor; Bioeffect Filter: agonists/antagonists depending on the receptor considered. Lipinski Filter: compounds passing the Lipinski drug-like filter.

The MW distribution of the active molecules versus each target has been calculated and plotted as depicted in Figure 1-5. Each class of actives shows a particular MW profile. Therefore, when designing a screening dataset for vHTS, ideally the same profile should be maintained. In other words, if PPARγ agonists have a MW in the range between approx 320 and 620, we should explore and use a screening database in this MW range to increase the chance of finding hits.

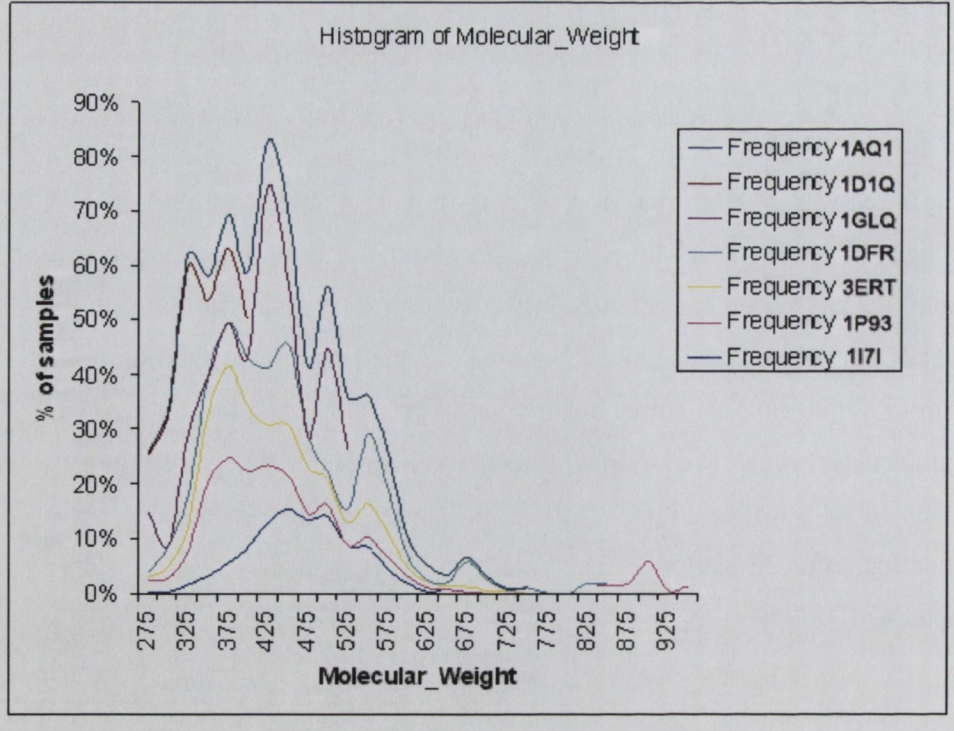


Figure 1-5 MW distribution for a set of known active compounds for the 7 receptors in the box (pdb Id entry).

On larger scale, ZINC (Zinc Is Not Commercial)[51] is a freely available database, was selected as the screening database. Each molecule in the ZINC library contains vendor and purchasing information and is ready for docking using a number of popular docking programs. This database is available for download (http://zinc.docking.org) in several common file formats including SMILES[73], mol2[74], 3D SDF[75]. ZINC contains about 2,700,000 drug-like molecules. Within certain limits, the molecules are prepared in multiple protonation states and multiple tautomeric forms. After applying various filters and clusterisation mentioned above the number of compounds was narrowed down to 11,000 for the same targets in Table 1-1.

Figure 1-6 compares the MW distribution of ZINC before and after filtering against the MW distribution of PPARγ agonists. As depicted in Figure 1-6, the distribution of MW in the case of the PPARγ receptor after filtering still coincides with the MW distribution of known actives giving an appropriate coverage for further analysis (distribution of known actives molecules against PPAR target is considered).

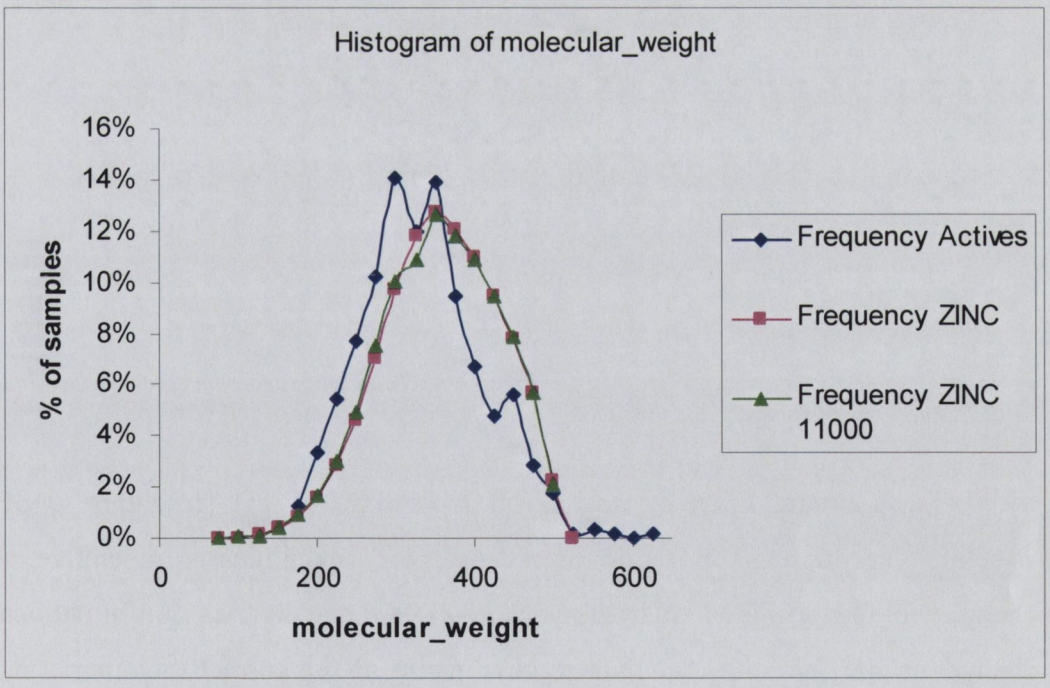


Figure 1-6 MW distribution of ZINC before and after filtering against the MW distribution of PPARγ agonists.

Another way to visualise the result is using Principal Components Analysis (PCA), (Figure 1-7) on the physicochemical descriptors of the molecule

dataset. Applied to molecular modelling, PCA is a method to visually represent the chemical space "explored" or covered by libraries of compounds. In this example the red sphere depicts a chemical space occupied by the commercial database SPECS[52]. The blue sphere represents the filtered database using focused filtering. Finally, the yellow solid sphere shows the space occupied by a set of known PPARγ agonists. It is clear how pre-filtering the database narrows down the library in a chemical space area where PPARγ ligands are located.

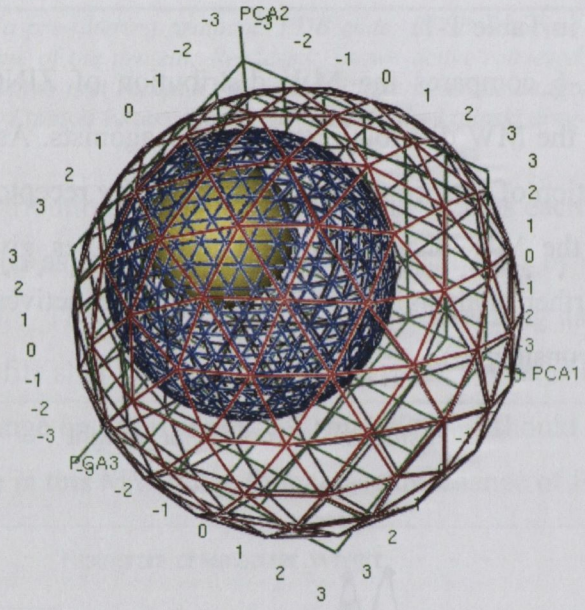


Figure 1-7 PCA results: the red sphere represents the chemical space occupied by the whole SPECS database. Blue represents the SPECS after filtering and a substructure search.

Furthermore, upon expansion of a database of 2D structures to 3D structures, accounting for HB donor and acceptor capabilities is imperative, as changes in the positions of hydrogens in 2D format denotes conformational changes in 3D format[76]. Accurate representation of the correct tautomeric and protonated forms of a compound, depending on the ultimate physiological environment of a compound, is also extremely important in this case. It is computationally expensive and difficult to assign the most probable state of a compound, thus all physiological relevant states are typically enumerated as a

representation. Similarly, it is necessary to generate all physiologically relevant stereoisomers of a compound arising from ambiguous stereo center descriptions.

Conformational changes (see section 1.2.4) must also be accounted for as ligands rarely bind to a receptor in their lowest energy form and usually experience some strain upon binding that induces an increase in the energy of the ligand[77]. It has been previously noted that in order to reproduce the binding mode of a ligand in a receptor it is beneficial for the docking algorithm to provide a number of docked poses on the basis of RMSD (see section 1.2.8.3)[78].

In a recent publication[79] we provided an insight into the implications of cheminformatic pre-processing of a validation database of compounds where multiple protonated, tautomeric, stereochemical and conformational states have been enumerated. Several commonly used methods for the generation of ligand conformations and conformational ensembles are examined, paired with an exhaustive rigid-body algorithm for the docking of different 'multimeric' compound representations to the ligand binding site of the human estrogen receptor alpha.

Different scoring function (see section 1.2.6) with intrinsic chemical knowledge was combined with a consensus-scoring scheme (section 1.2.7.1) to rank output from the docking protocol, and enrichment rates (section 1.2.8.1) calculated for each screen. The overheads of CPU consumption and effect on relative database size (disk requirement) for each of the protocols employed were considered. Assessment of these parameters indicated that SBVS enrichments are highly dependent on the initial cheminformatic treatment(s) used in database construction. The interplay of SMILES representations (Chapter 2), stereochemical information, protonation state enumeration and ligand conformation ensembles were identified as critical in achieving optimum enrichment rates in such screening.

1.2.4 Docking algorithm and tools

Docking and scoring allow the user to search for structures that have the correct geometric and electrostatic features to fit in the protein binding site. Conformational searching during docking is essential as usually it is not known which conformation of a ligand interacts most favourably with a receptor. Algorithms able to simulate flexibility of a molecule and generate conformers have been classified by Kuntz[4] into three main groups: systematic, stochastic and deterministic algorithms[5,6].

Systematic search algorithms are based on a grid of values for each formal degree of freedom, and each of these grid values is explored in a combinatorial fashion during the search. As the number of the grid values increases, the number of evaluations needed increases exponentially. Examples of systematic algorithms are represented by the anchor and grow (incremental construction) or the hammerhead algorithm[80].

The anchor and grow method usually divides the ligand into flexible fragments and rigid fragments. A particular fragment is docked, and the full ligand is built up either incrementally or separately docked fragments are joined together[81]. In the anchor and grow algorithm one or more rigid fragments, or anchors, are placed within the binding pocket considering either sphere matching[82-84] shape complementarities (DOCK4.0, LIGIN[85]), chemical interaction (FlexX[86,87]) or HB interactions as used by Leach[88].

In the *hammerhead algorithm*[80] method fragments are determined automatically by disassembling the ligands, aligning the fragments within the binding pocket using probing dummy atoms and finally retaining just the fragments showing the best interactions. A process called "chaining" links together the different fragments by considering bond angles. Therefore, instead of systematically searching torsion angles, the ligands are optimised after each fragment is added using energy minimisation. The algorithm was specifically developed for database screening. SURFLEX[23,24], is a fully automatic flexible molecular docking that uses the hammerhead docking system.

The systematic algorithm methods are usually applied in docking for the generation of pre-computed 3D conformer libraries[89]. In these datasets, each compound is represented by an ensemble of conformations whose coordinates are saved. Molecular library conformations are calculated once, the conformers are rotated and translated within the binding pocket and the best poses are outputted (e.g. FRED: fast exhaustive rigid docking[90]).

Pre-generated conformational libraries are an efficient way of considering the flexibility of the ligands as the cost of generating multiple conformers per molecule is only incurred once and the internal energy of the conformers can be assessed prior to docking. A large variety of software can be employed for the generation of multiple conformers (OMEGA[91], CORINA[92], RUBICON[93], MOE[94]). Although efficient, the disadvantage is that receptor interactions are not taken into account during the conformer generation. Furthermore as the procedure necessitates the generation of a library of conformers, a significant amount of storage space is also required. A scheme depicting the different approaches of systematic algorithms is shown in Figure 1-8.

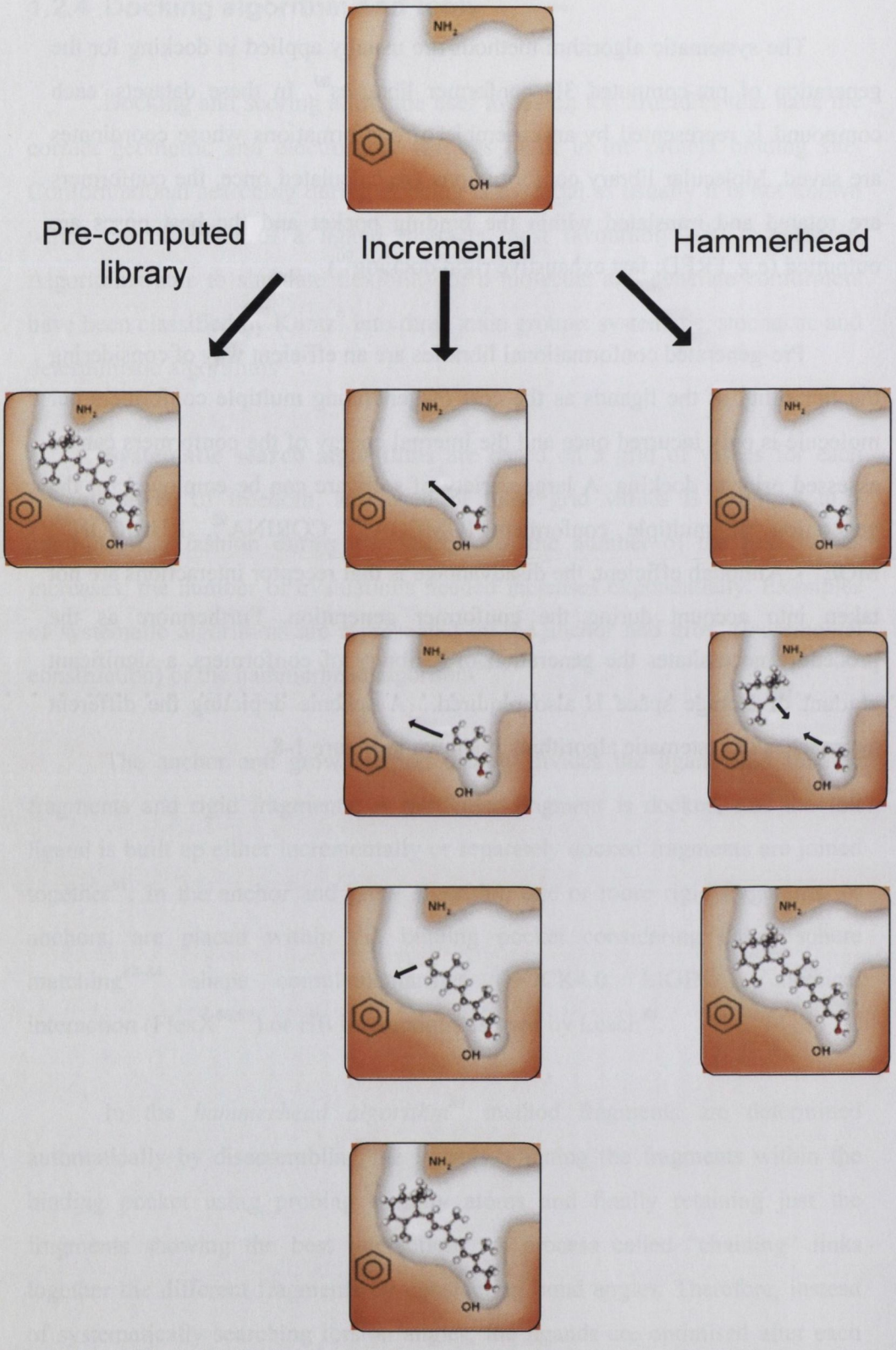


Figure 1-8 Different systematic docking algorithms. 1. Library of pre-computed conformers can be docked in a rigid fashion in the binding pocket. 2. In the incremental algorithm a fragment is first seeded in the binding pockets and incremental re-construction of the ligand follow. 3. In the hammerhead fragments are placed with regards to their affinity and in a second stage "chained"-back together. (Sketch adapted from Böhm[95]).

In **deterministic methods** or searches, a generic initial state determines the move that can be made to generate the next state, which generally has to be equal to or lower in energy than the initial state. Deterministic searches performed on exactly the same system with the same initial parameters will produce the same final state[4]. Examples of deterministic algorithm are represented by energy minimisation and molecular dynamics simulation. The former is currently used in many docking applications as a post-processing technique in order to optimise the potential energy associated with the ligand. The latter is not used for vHTS due to the computational time required per complex.

Stochastic search algorithms make random changes, usually varying one degree of freedom of the system at a time. With stochastic searches there are convergence uncertainties. Different runs, with the same input and parameters may produce different results. Examples of stochastic searches include *Monte Carlo* (MC) methods or the *evolutionary* algorithm. In MC implementations, the ligand is considered as a whole, and random changes are made in order to translate and rotate the ligand, as well as to modify the torsion angles. After each move, the structure is minimised and the energy of the new structure is determined. In this procedure the ligand is usually placed in a random position in the binding site and to increase the chance of reaching the global minima the simulation consists of numerous cycles. The first cycle is performed simulating "high temperature" (high translational movement is allowed) and later cycles are done at decreasingly lower temperatures (simulated annealing MC). Usually each cycle starts with the lowest energy from the precedent. AutoDOCK[91,96,97] was the first program to be implemented with MC based searches.

Additionally, independent from the method used to fit and optimise a ligand within the binding site, different methods to represents the binding pockets can be considered. The protein can be represented using its Connolly surface,[98] which consists of all the points of the vdW surface at which a solvent sphere can touch it. These methods attempt to align points on the surface (ligand-protein or protein-protein).

Other computational methods use energy grids[99] pre-computed prior the docking. The idea is to store information about the binding site and receptor's potential

energy contribution on grid points so that it only needs to be read during the ligand evaluation. Typically the grid points can be electrostatic or vdW. The Electrostatic potential is represented by a pairwise summation of Columbic forces (Equation 8) and the vdW term is usually modelled using a Lennard-Jones 12-6 (Equation 9).

$$E_{coul}(r)\sum_{i=1}^{Na}\sum_{j=1}^{Nb}\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad 1\text{-}8$$

Equation 8 Columbic potential: N represents the number of atoms in molecules a and b and q the charge on each atoms. r is the distance between atoms i and j, respectively in molecules a and b.

$$E_{vdW}(r)\sum_{i=1}^{Na}\sum_{j=1}^{Nb}4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad 1\text{-}9$$

Equation 9 vdW potential: N represents the number of atoms in molecules A and B and q the charge on each atom. r is the distance between atom i and j. σ represents the collision diameter of atom i in molecule a with atom j in molecule b.

1.2.5 Scoring procedures

The majority of published scoring functions have been developed in association with docking tools. However, the scoring functions used to rank compounds post docking do not have to be linked to the originating docking method but rather are independent and should be able to estimate the binding affinities of the protein-ligand complexes generated by any structure-based approach.

As mentioned in the above sections, a scoring function estimates the binding affinity between RL or Protein-Protein complexes. The calculation is not an easy task as there are many thermodynamic variables that should be taken into account. Firstly, one should consider that before binding, both ligand and protein are surrounded by molecules of solvent. The desolvation occurring during binding leads, therefore, to an increase in entropy of the system, driving the equilibrium toward the products. The increase in solvent entropy upon binding to

non-polar surfaces in the receptor is the known as the hydrophobic effect[100]. The desolvation phenomenon can also be seen from an enthalpy point of view. The desolvation occurring during the binding implies that a number of interactions with the solvent must be lost and will only partially be replaced with the novel contacts within the binding site[4]. An entropy loss should also be considered, arising from the fact that the PL complex is a highly organised (low-entropy) entity compared to P + L in solution. Furthermore, the entry of the ligand into the binding site brings the reacting groups of the ligand and coordinating residues of the protein together. The substrate will fit in the proper position for binding, with a net loss of entropy (Figure 1-9).

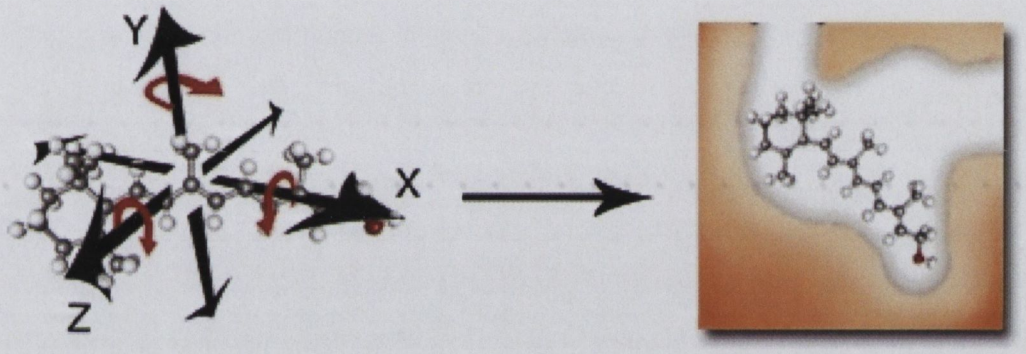


Figure 1-9 On the left the ligand is free to undergo translational motion - a high entropy situation. After binding to the receptor the system is highly ordered - a low entropy complex.

Finally, in order to estimate the binding affinity the internal, or strain energy, of the ligand and protein must be taken into account. This energy can be higher in the complex than in both the un-complexed states[4].

A common approach to approximate the binding affinity is the thermodynamic analysis of the interaction occurring within a RL complex. The stability of a particular complex can be measured by determining the equilibrium binding constant K_{eq}, which is directly related to the Gibbs free energy (equations 10-12):

$$[P]+[L]\xrightarrow{Kon}[PL] \qquad 1\text{-}10$$

$$K_{eq} = \frac{Kon}{Koff} \qquad 1\text{-}11$$

$$\Delta G_{bind} = -RT\ln K_{eq} = -RT\left(\frac{K_{on}}{K_{off}}\right) = \Delta H - T\Delta S \qquad 1\text{-}12$$

Equation 10-12: P: protein; L: Ligand; PL: Complex Protein-Ligand; K_{eq}: equilibrium binding constant; G_{bind}: Gibbs free energy; K_{on} and K_{off} : on-rate and off-rate of the reaction; H-TS: decomposition of Gibbs free energy in enthalpy (H) and entropy contribution (S). T: temperature

As the Gibbs free energy is theoretically calculable using thermodynamic equations, the binding constant, K_{eq}, can be derived easily using eq. 12. As solving the theoretical thermodynamic equation accounting for enthalpy and entropy of the system is complex and time-consuming, an approximation is required for vHTS.

To estimate ΔG_{bind} for vHTS the calculation is simplified using a scoring function, which compromises between the accuracy in the prediction of the binding affinity and calculation speed. In many cases, the scoring functions can consistently predict the binding poses of ligands with nanomolar affinity; however, they perform poorly at predicting lower affinity. As explained in the following paragraph in more detail, different methods have been developed in order to estimate ΔG_{bind}.

1.2.6 Scoring functions details

Two aspects of the CADD primarily determine the success of the process: the generation of reasonable ligand-binding poses through sampling the conformational space of both ligand and receptor (the "docking problem") and the identification of those binding poses that correspond best to the experimental given situation based on reasonable estimates of binding affinities, namely the scoring problem[8,101]. Scoring functions can be divided into three classes depending on how the binding affinity of a molecular complex is determined.

1) Physics (Molecular simulation) based approach[102,103]

2) Empirical/force field additivity based approach[4,104]

3) Knowledge approaches[105-109]

The basic idea of molecular simulation approaches derives from statistical mechanics and molecular dynamics. Explicit atomic level consideration of solvent molecules, ions and flexibility of both ligand and protein makes these approaches computationally intensive and limits the usage of simulation strategies in screening large numbers of ligands against a protein target. Force field methods mainly account for the binding energy between ligand and protein by summing up non-bonded interactions such as vdW and Columbic potentials. Knowledge based scoring function are based on statistical analysis of a large number of PL complexes available in the structural repositories[101]. Table 1-2 lists some of the different scoring functions available along with the class they belong to.

Type of Scoring Function	***Function Name***
Physics based	DOCK[110]
	CHARMm[111]
	AutoDock[112]
Empirical	ChemScore[92]
	ScreenScore[113]
	ShapeGauss[114]
	PLP[115]
	FlexX[116]
	LUDI[95]
	LigScore[117]
Knowledge	BLEEP[118]
	DrugScore[107,108]
	PMF[119,120]
	SMoG[121]

Table 1-2 Scoring Functions commonly used in Virtual Screening.

1.2.6.1 First principle method

The first principle method, or molecular mechanics force-field method, replace all the electrons in the system by empirical parameters and simple analytical expressions.

These methods represent a straightforward approach to scoring by using the non-bonded energies of well-established molecular mechanical force fields and usually quantify the sum of two energies, the RL interaction energy and internal energy of the ligand (as steric strain induced by binding). Interactions between RL are often described using vdW terms and electrostatic energy terms. In doing so, estimates of the free energy of binding in solution are substituted by estimates of the gas-phase enthalpy of binding. A typical paradigm of molecular force field calculation is showed in eq. 13.

$$E_{total} = \sum E_{bond} + E_{angle} + E_{dihedral} + E_{vdW} + E_{Electrostatic} \qquad 1\text{-}13$$

The E_{total} represents an estimation of the potential energy of the system. The constituent terms consist of an energy penalty associated with the deviation of bonds and angles away from their 'reference' values (E_{bonds} and E_{angle}), there is a function that describes how the energy changes as bonds are rotated ($E_{dihedral}$) and finally a terms which consider the non-bonded interaction in the system (E_{vdW}_and $E_{Electrostatic}$). The various contributions are depicted in Figure 1-10 and Figure 1-11.

Examples of classical force field models include G-Score[122] (based on the Tripos Force Field[122]), and AutoDock[112] (based on AMBER[50] force field).

As an example (eq. 14), in AMBER[50] the overall docking energy of a given ligand molecule is expressed as the sum of intermolecular interactions between the complex and the internal energy of the ligand.

$$V(r^N) = \sum_{Bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 +$$

$$+ \sum_{Torsion} \frac{V_i}{2}(1 + \cos(n\omega - \gamma)) +$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \qquad 1\text{-}14$$

$V(r^N)$ represents the potential energy of the system. The first term accounts covalently bonded atoms using a harmonic (ideal spring) function (Figure 1-10-A) which is a good approximation near equilibrium bond lengths, but becomes increasingly poor as atoms separate. The second term (Figure 1-10-B), harmonic as the former, represents the internal energy contribution derived from the summation over all valence angles in the molecules. For these terms in the equation the potential energy will increase as the actual length or angle (l_i and θ_i) of the molecules diverge from the *reference* length or angle ($l_{i,0}$ and $\theta_{i,0}$). The summing over torsion in the molecule is represented by the third term (Figure 1-11C).

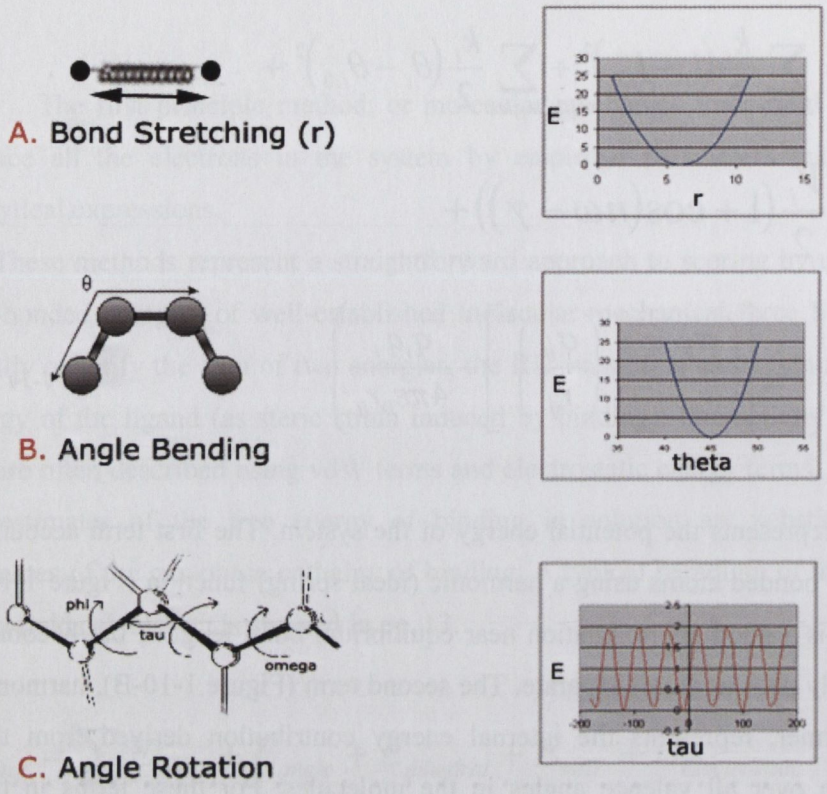


Figure 1-10 (A-B-C): Representation of the main contributions to a force-filed based scoring function and their relative graphs schematic demonstration.

The last two terms accounts for the non-bonded interactions, vdW and electrostatic (Figure 1-11).

The vdW interactions are represented with a 12-6 Lennard-Jones potential in the gas-phase. The Lennard-Jones potential makes atoms repulsive as the distance between the atoms approaches the sum of their vdW radii (Pauli Exclusion Principle), and is slightly attractive or makes a negligible contribution to the potential energy as the distance between the atoms increases. In the formula, σ_{ij} are coefficients dependent on the atoms i and j, and r_{ij} represents the distance between the atoms i and j. The parameters of the Lennard-Jones potential vary depending on the desired 'hardness" of the potential. Higher terms (12-6) result in increasingly repulsive potentials and will be less forgiving of close contacts between receptor and ligand atoms. Lower terms (i.e. 8-4 Lennard-Jones) make the potential softer. In those scoring functions, the vdW term is often responsible for penalising docking solutions with respect to overlap

between receptor and ligand atoms. Distance-dependent dielectric constants (ε_{ij}) are usually employed to approximate the long-range electrostatic interactions.

Standard force-field methods have major limitations because they were originally computed for gas-phase contributions to structure and energetic terms, therefore do not include solvation effects and entropic terms.

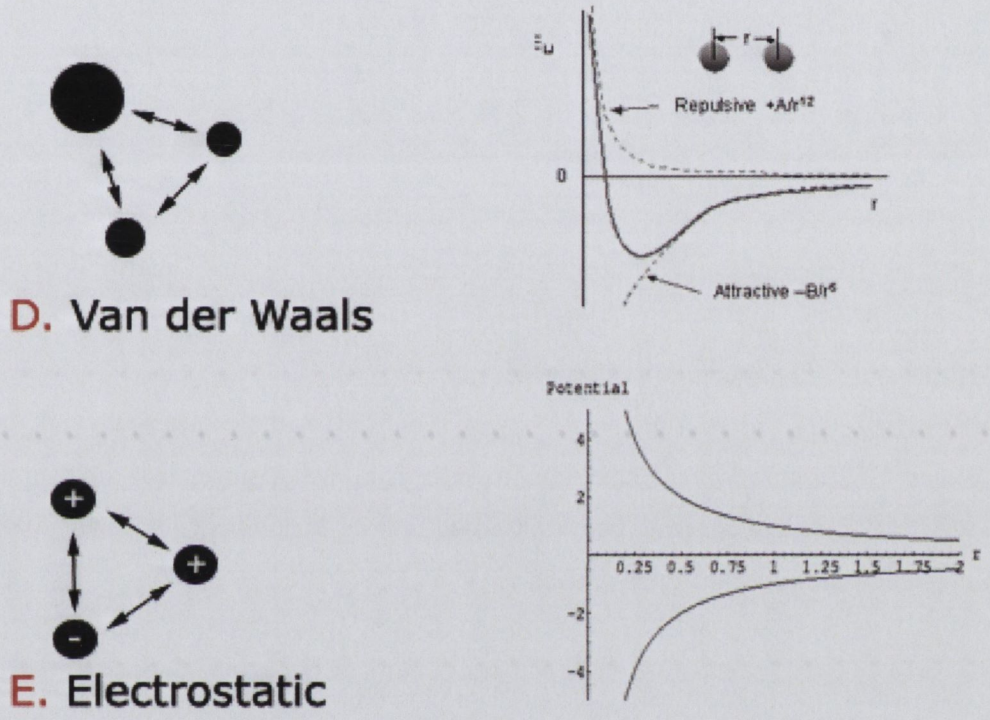


Figure 1-11 (D-E): Representation of the main contributions to a force field based scoring function and their graphical schematics.

1.2.6.2 Empirical method

Empirical scoring functions are an alternative to the pure force-field method for scoring. The underlying idea is that the binding free energy of a non-covalent receptor-ligand complex can be decomposed into a sum of localised, chemically intuitive interactions[92]. As pointed by Kuntz and Brooijmans[4], empirical scoring functions can be considered as a logical extension of structure-activity relationships first developed by Hansch in 1960s. Hansch correlated the octanol-water partition of solute with some biological activities of the molecules. This deduction was therefore based on the physical-properties of the ligands only. This work led to the development of a new field of investigation, Quantitative

Structure Activity Relationship (QSAR), broadly used today in the absence knowledge of the target structure. Empirical scoring functions are based on receptor-ligand properties rather than on ligand property alone as in the QSAR study.

In other words, empirical scoring function are fit to reproduce experimental data, such as binding energies/or conformations, as a sum of several parameterised functions[89]. The binding affinity ΔG_{bind} is thus estimated as a sum of interactions ΔG_i multiplied by weighting factors f_i :

$$\Delta G_{binding} \approx \sum \Delta G_i f_i \qquad 1\text{-}15$$

The coefficient factors f_i are obtained from regressional analysis using experimentally determined binding energy, free energy perturbation method (FEP) and X-ray structural information.

The first empirical scoring function was proposed in the pioneering work of Böhm[95,123] (Figure 1-8 and equation 16). This approach assumes that the overall RL binding free energy can be decomposed into basic components, each one of them accounting for different terms.

$$\begin{aligned} \Delta G_{binding} = \Delta G_0 &+ \Delta G_{h-bonds} \sum\nolimits_{h-bonds} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{ionic} \sum\nolimits_{ionic} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{lipo} \left| A_{lipo} \right| + \Delta G_{rot} NROT \end{aligned} \qquad 1\text{-}16$$

Terms represents HB interactions, electrostatics, lipophilic and entropic interactions. This scoring function was first implemented in LUDI[123], now implemented in FlexX[10,124], and is derived using experimental binding free energies of 45 complexes.

The first two terms account for the hydrogen bonds formed between ligand and protein, where neutral H-bonds and ionic interactions are treated individually. $\Delta G_{h\text{-}bonds}$ and ΔG_{ionic} describe the contributions respectively from an

ideal hydrogen bond and an unperturbed ionic interaction. The contribution of the first two terms (*h-bonds* and *ionic*) is scaled by a distance and angle-dependent function ($f(\Delta R, \Delta\alpha)$) in order to penalise the deviations from an ideal geometry. The third term accounts for lipophilic interactions ΔG_{lipo} assumed to be proportional to the lipophilic contact surface A_{lipo}. The fourth term accounts for all the rotatable single bonds (*NROT*) in the ligands, since this relates to the loss in torsional entropy ΔG_{rot} of the ligand upon binding within the receptor pocket. The last term is a regression constant. ΔG_0 is a contribution to the binding energy that does not directly depend on any specific interactions with the protein but is associated with the overall loss of translational and rotational entropy of the ligand.

In empirical scoring functions, terms accounting for non-bonded interactions can be implemented in rather different fashions. For example, ChemScore[92] does not differentiate between ionic and hydrogen bond interaction.

The main disadvantage of the empirical scoring functions is the strong dependence on the training data set used to perform the regressional analysis. This often yields different weighting factors for the various terms. Consequently, terms from different scoring function can not be combined together. Another disadvantage is that pH, salt concentration and temperature can influence the experimental measurement of binding constants. These different conditions are generally ignored when calculating binding free energies from experimental binding constants and this limits both the training of empirical scoring function and their accuracy[4].

1.2.6.3 Knowledge based potentials

Knowledge based scoring functions[118,125] are designed to reproduce experimental structures rather than binding energies. RL complexes are modelled using relatively simple atomic interaction-pair potentials. A number of atom-type interactions are defined and their frequencies observed. If these frequencies are converted into free energies using the Boltzmann distributions, the potentials are generally called *potential of mean force* (PMF). Therefore, while empirical scoring functions require experimental binding data in order to fit the equation,

knowledge based methods requires a training set of structural data, which has the advantages of easily devising large training sets.

In more detail, the Helmholtz free energy of binding (ΔA_{bind}) is calculated by summing over all protein-ligand atom pairs interactions as show in equation 1-17 and 1-18:

$$\Delta A_{bind} = \sum A_{ij}(r) \qquad 1\text{-}17$$

$$A_{ij}(r) = -kt \ln \frac{g_{ij}(r)}{g(r)} \qquad 1\text{-}18$$

Where A_{ij} is the PMF between atom types i and j, with the particular kind of interaction *ij*. The sum is over all protein-ligand atom pair interactions between atom *i* and *j*. The PMF is derived from equation 18, where ***k*** is the Boltzmann's constant and g_{ij} (r) is the normalised probability of atom pairs *ij* to be in contact at distance r, while g(r) is the normalised reference probability for the reference state where the protein and ligand do not interact with each other.

1.2.7 Post-processing methods

As shown by Watson et al[60] an adequate post processing procedure avoids artificial enrichment (see section 1.2.8.1) and decreases the chance of obtaining false positives. Various post-processing methods can be used to achieve this result. In the following section some of these methods relevant to this work are covered.

Docking constraints: Requirements such as HB to specific residues of the active site that are known to be important in the ligand binding process can be used as a filter to select possible candidates. Specific computational tools, such as LPC (Ligand Protein Contacts)[126], can rapidly estimate the presence or absence of specific interactions. Therefore, compounds that do not show a

specific contact within the binding pocket can be discarded and the final hit list reduced.

Guided Docking: During the docking procedure, different software applications help defining some constraints to guide the docking itself ("guided docking"). In other words, an area within the binding pocket can be selected to allow just a defined pattern (usually coded as a SMART string, a language for describing molecular patterns[127]). For example, certain polar groups can be preferentially placed in particular polar pockets, helping therefore to achieve a more accurate docking. For example PPARγ agonists have a polar functional group usually placed in proximity to the AF-2 domain. A pattern can be described to instruct the docking software to not consider those ligand poses which lack polar atoms in that region. Hindle et al.[128] incorporated a pharmacophore matching tool in the docking program and showed an increase in the enrichment (see below for enrichment).

Another common way to post process results is applying *consensus scoring*, *similarity* and *sub-structure searching*, as described in the following sections.

1.2.7.1 Consensus scoring function

As described above scoring functions have primarily two main goals: to rapidly determine the most likely binding mode of each ligand and to return an estimation of its binding affinity for the target protein. Different scoring functions use distinct atom types, different atomic partial charges, and different molecular force fields and have been trained with different protein – ligand complexes.

As a result of these different approaches in estimating the binding affinity, different scoring functions may produce considerably varying results[27,30]. One way to get around such differences is to combine the results from different scoring functions into a single consensus one[129]. This method was first introduced by Charifson et al[77] where they docked ligands with their in-house docking tool "*Gambler*", re-scored the generated docked ligand conformations with an ensemble of scoring functions and then took the intersection of the top *N%* of each of the sorted ranking lists. In their work they

demonstrated that the number of false positives is significantly reduced by calculating the consensus scoring.

Watson et al.[60] also exemplified different consensus procedures. The simplest way to combine different scoring functions is to sum the individual scores (rank by score) after normalisation. Another approach is to combine the relative positions and compute an average. Thus, if compound A is classified 3rd in scoring function X and 8th in scoring function Y, the resulting consensus score will be represented by the average, 5.5 (rank by rank). Finally in rank-by-vote, compounds have to rank in the top x% of the ranked lists of each of the individual scoring-functions. Both methods show good efficiency compared to the scoring functions individually although "rank by rank" and "rank by score" perform better then "rank by vote".

1.2.7.2 Similarity searching

In many cases, even after applying virtual screening, there are still too many compounds to be tested in biological assay. In these cases, a common tactic is to select hit compounds which present some degree of similarity with known actives. Similarity is a numerical estimation of the difference between two objects, for our purposes, molecules. Usually it ranges from 0 to 1. A similarity of 0 means the two molecules are completely dissimilar; a similarity of 1 means they are exactly the same.

A significant number of similarity algorithms are available: examples include 2D, 3D, fingerprint based and shape similarity. Furthermore, different mathematical approaches exist to calculate the similarity between two entities. As an example the *Tanimoto similarity* score is a method of calculating the similarity between two ligand fingerprints. As shown in the equation 19 where $\boldsymbol{T}$ is the Tanimoto Score, N_a and N_b are the number of bits set to 1 in the fingerprint of ligand *a* and *b* respectively and N_c is the total number of bits set to 1 found in fingerprints of both ligands *a* and *b*.

$$T = \frac{N_c}{(N_a + N_b - N_c)} \qquad 1\text{-}19$$

A Tanimoto score (T) of 0.7 or higher is representative of two molecules having high structural similarity which taken as a good indication of possible similar biological activity.

1.2.7.3 Re-scoring procedure

A common strategy to improve the result from a scoring function is to re-score the result by applying a coefficient before either the final score or in front of singular terms in the equation. This may be done to enhance or to emphasise singular contributions known to be essential for the binding. If HBs are essential for the binding of a particular set of ligands, the term in the scoring function accounting for them can be weighted over the other terms.

Furthermore, scoring functions are known to be affected by the MW of the ligands. This is due to terms such as the vdW contribution where attractive potentials are favourable to the binding mode. Due to the fact that this calculation is the sum of pair wise interactions, the more atoms involved, the higher the MW and hence the bigger this contribution will be. In chapter 4 a normalisation-rescoring strategy is applied to reduce the dependency between MW and score in order to gather more consistent results.

1.2.8 Validation methods

1.2.8.1 Enrichment

The potential of docking/scoring programs to identify a set of known active ligands from a set of drug-like 'decoys' is measured using the metric of Enrichment. Enrichment is a measure of the proportion of hits retrieved in a subset of compounds compared to the proportion of hits expected from a random sample of compounds,

$$Enrichment = \frac{\frac{Hits_{Sampled}}{N_{Sampled}}}{\frac{Hits_{Total}}{N_{Total}}} \qquad 1\text{-}20$$

Where

$Hits_{sampled}$ = the actual number of hits

$Hits_{Total}$: the total number of hits

$N_{sampled}$: the actual number of compounds sampled

N_{Total}: the total number of compounds

1.2.8.2 Area under ROC (Receiver Operating Characteristic) curve

ROC curves represent an alternative evaluation method for the validation of a vHTS. By generating a ROC curve and by calculating the Area under the Curve (AUC), a single value representing an estimation of the quality of the vHTS is obtained. In order to understand a ROC curve, the concepts of false positive (FP), false negative (FN), sensitivity and specificity must be given (Table 1-3).

As outlined in Table 1-3, a FP is an inactive molecule predicted by the vHTS to be active. On the contrary, a FN is a true active predicted to be inactive the by the vHTS.

	True Active	*True inactive*	
Test Positive	True positives	False positive	Total positives
Test Negative	False negative	True negatives	Total negatives
	Total actives	Total inactives	Total

Table 1-3 In a false positive a test wrongly shows an effect or condition to be absent. Conversely, in a false positive a test wrongly shows an effect or condition to be present. In true positive or a false positive the test is exact in the prediction.

Sensitivity is thus defined as the proportion of true positives (TP) predicted to be active over the total number of actives.

$$Sensitivity = TP / (TP+FN) \qquad 1\text{-}21$$

Specificity is the proportion of true negatives (TN) predicted to be inactive over the total number of actives.

$$Specificity = TN / (TN + FP) \qquad 1\text{-}22$$

In hypothetical screening, we define N equal to the total number of molecules in the database (1000). Applying the concept of FPs, FNs, TPs and TNs and considering just the top 10% (100 molecules), a typical scenario can be described as in Table 1-4.

	True Active	*True inactive*	
Test Positive	10	30	40
Test Negative	20	40	60
	30	70	100

Table 1-4 Hypothetical screening results considering 10% (100) of a database of 1000 molecules.

Thus

$$Sensitivity = TP / (TP+FN) = 10/(10+20) = 10/30 = 0.33$$

and

$$Specificity = TN / (TN + FP) = 40/(40+30) = 40/70 = 0.57$$

The ROC curve is therefore defined as the plot of the true positive rate (*sensitivity*) against the false positive rate (1 - *specificity*) for the different possible cut-off points of a test (for example, 10%, 20% or 30% of the database screened). An example is given in Figure 1-12.

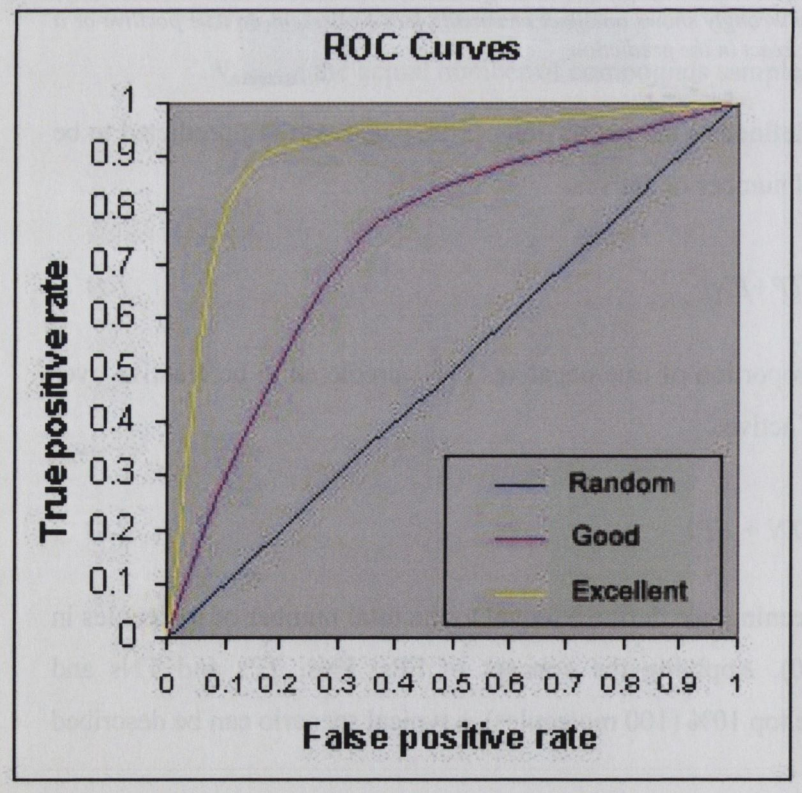


Figure 1-12 Example ROC curve. See text for details (from http://gim.unmc.edu).

The graph in Figure 1-12 shows three ROC curves representing excellent, good and random distribution tests plotted on the same graph. The accuracy of the test depends on how well the vHTS separates the molecules being tested into those with and without experimentally determined activity. *Accuracy* is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0.5 represents a poor test. A rough guide for classifying the accuracy of a ROC curve is the following point system:

0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)

ROCs Curves will be used in Chapter 3 in order to validate a vHTS.

1.2.8.3 RMSD (Root Mean Square Deviation)

RMSD is a measure of the difference between two conformations of the same molecule. RMSD is used to compare two conformations of a given set of points, typically atoms. Given a list of paired points, it gives a measure of the distance between these points. Normally a rigid superposition which minimises the RMSD is performed, and this minimum is returned. Given two sets of N points v_n and w_n, the RMSD is defined as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2} \qquad 1\text{-}23$$

Where

$$\delta = v_i - w_i \qquad 1\text{-}24$$

An RMS distance value between molecular objects is expressed in Angstroms (Å). RMSD analyses are usually applied to test the ability of conformer generator software to replicate the ligand conformation in a co-crystal structure. The use of RMSD calculations will be further discussed in Chapter 2.

1.2.8.4 Scoring frequencies

Recent publications[130,131] highlighted the calculation of score distributions in order to assess the discriminatory power of scoring functions and on the other hand that neither enrichment or lead identification alone are sufficient for optimising an vHTS protocol. Cole et al.[132] suggested using a statistical analysis of variance (ANOVA) of the score distribution in order to establish a significant variation when analysing various PL docking programs. These methods will be discussed and applied in Chapter 4 and Chapter 5.

1.2.8.5 SIFt fingerprints

Structural Interaction Fingerprint (SIFt) is a 1D binary fingerprint representation of intra-molecular interactions in a 3D structure[133,134], like protein-ligand complex. SIFt can be used to compare the poses generated and ranked in the docking/scoring process with the one retrieved in the crystal structure. A MOE[43] implementations of SIFt can be used to calculate the fingerprints and Tanimoto distance to evaluate the similarity between the crystal structure and the scored poses. Tanimoto similarity for each scoring functions before and after re-scoring, can be calculated allowing an increasing overlap over each calculation (65%, 70%, 75%, 80%, 85%) (See chapter 4). In Figure 1-13 the result histogram from an analysis carried out with SIFt is shown.

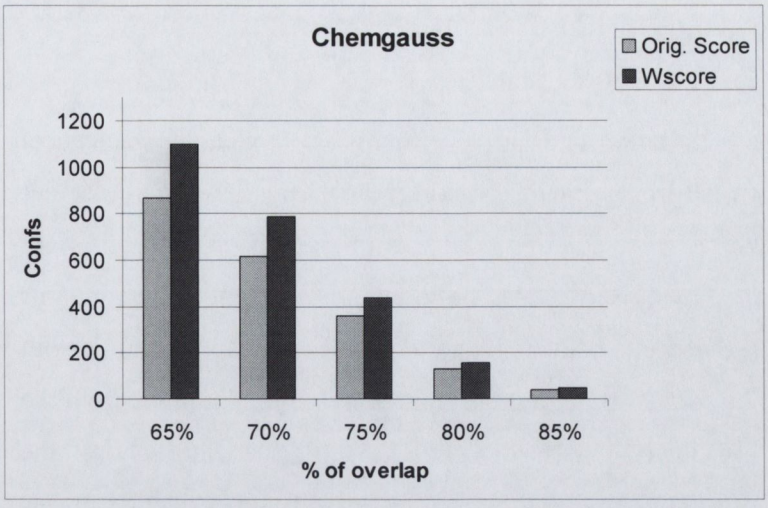


Figure 1-13 SIFt analysis for the Chemgauss scoring function.

1.2.8.6 Correlation Matrices and Spearman coefficient

The Spearman formula is a measure indicating how much two ordered lists are correlated with each other. It is usually applied in two different fashions. It can be used to test a scoring function and its ability in predicting the binding affinity (the predicted binding affinity and experimentally determined one can be plotted and the correlation calculated). It also can be used to estimate the degree of agreement between different scoring functions. This method will be investigated thoroughly in Chapter 4. A Correlation matrix is calculated using MOE[43] tools by applying the Spearman formula:

$$R_s = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad 1\text{-}25$$

$$R_s = 1 - \frac{\sum}{N(N^2-1)}$$

D is the difference between each rank of corresponding values and N represents the number of pairs of values.

1.2.8.7 In vitro screening: Fluorescent Polarisation

Fluorescence polarisation (FP) measurements is a valuable biophysical research tool for investigating ligand–protein interactions at the molecular level[135]. When a fluorescent molecule is excited with plane polarised light, light is emitted in the same polarised plane, provided that the molecule remains stationary throughout the excited state. If the molecule rotates and tumbles out of this plane during the excited state, light is emitted in a different plane from the excitation light. If vertically polarised light is exciting a fluorophore, the intensity of the emitted light can be monitored in vertical and horizontal planes (the degree of movement of emission intensity from vertical to horizontal plane is related to the mobility of the fluorescently labelled molecule). If a molecule is very large or bound to a receptor, little movement occurs during excitation and the emitted light remains highly polarised. If a molecule is small, rotation and tumbling is faster and the emitted light is depolarised relative to the excitation plane.

In other words, the binding of a fluorophore labelled ligand to a protein results in a decrease of ligand rotation and hence a measurable increase in polarisation upon excitation with polarised light.

The FP technique used in Chapter 3 for the detection of PPARγ agonists employs a high-affinity steroidal compound containing a covalently linked fluorophore. As illustrated in Figure 1-14, the displacement of the fluorescent steroid by competitor compounds serves as the basis for high throughput screening assays. By displacing the fluorophore, the competitor generates a fall in the light polarisation emitted (low polarisation) (Figure 1-14).

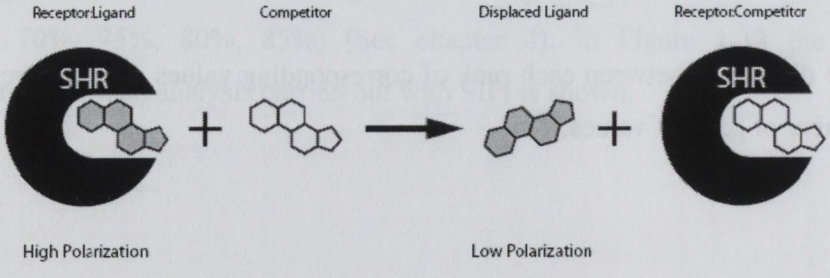


Figure 1-14 Scheme of FP competitive binding assay. (Figure from www.invitrogen.com).

Competitive binding experiments are used to analyse binding affinities of non-fluorescent ligands. It is therefore possible to compare the relative binding affinities (IC_{50}) of two or more unlabeled ligands or, when the K_d of the labelled ligand is known, calculate the K_i of the unlabeled ligands (see appendix of Chapter 3 for more details about RL thoery and binding assay experiments).

1.3 References

1. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ 2003;22(2):151-185.
2. Reichert JM. Trends in development and approval times for new therapeutics in the United States. Nat Rev Drug Discov 2003;2(9):695-702.
3. Verkman AS. Drug discovery in academia. Am J Physiol Cell Physiol 2004;286(3):C465-474.
4. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 2003;32:335-373.
5. Klebe G. Recent developments in structure-based drug design. J Mol Med 2000;78(5):269-281.
6. Gubernator K, Böhm H-J. Structure-based ligand design. Weinheim ; Chichester: Wiley-VCH; 1998. xiv,153 p.
7. Lyne PD. Structure-based virtual screening: an overview. Drug Discov Today 2002;7(20):1047-1055.
8. Jorgensen WL. The many roles of computation in drug discovery. Science 2004;303(5665):1813-1818.
9. Kubinyi H. Strategies and Techniques for Identification of Novel Bioactive Compounds - CHI's Second Annual Conference. IDrugs 1998;1(8):883-886.
10. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43(25):4759-4767.
11. Blin N, Federici C, Koscielniak T, Strosberg AD. Predictive quantitative structure-activity relationships (QSAR) analysis of beta 3-adrenergic ligands. Drug Des Discov 1995;12(4):297-311.
12. Hansch C MP, Fujita T, Muir RM. Correlation of Biological Activities of Phenoxyacetic Acids with Hamme Substituent Constants and Partition Coefficents. 1962;194:178-180.
13. Hoffman BT, Kopajtic T, Katz JL, Newman AH. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn Z descriptors. J Med Chem 2000;43(22):4151-4159.
14. Netzeva T, Doytchinova I, Natcheva R. 2D and 3D QSAR analysis of some valproic acid metabolites and analogues as anticonvulsant agents. Pharm Res 2000;17(6):727-732.
15. Hunt PA. QSAR using 2D descriptors and TRIPOS' SIMCA. J Comput Aided Mol Des 1999;13(5):453-467.
16. Klebe G, Abraham U. Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput Aided Mol Des 1999;13(1):1-10.
17. Good AC, Peterson SJ, Richards WG. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. J Med Chem 1993;36(20):2929-2937.
18. Makhija MT, Kulkarni VM. 3D-QSAR and molecular modeling of HIV-1 integrase inhibitors. J Comput Aided Mol Des 2002;16(3):181-200.

19. Liao C, Xie A, Zhou J, Shi L, Li Z, Lu XP. 3D QSAR studies on peroxisome proliferator-activated receptor gamma agonists using CoMFA and CoMSIA. J Mol Model (Online) 2004;10(3):165-177.
20. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol 2006.
21. Mizutani MY, Itai A. Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors. J Med Chem 2004;47(20):4818-4828.
22. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47(7):1739-1749.
23. Jain AN. Ligand-based structural hypotheses for virtual screening. J Med Chem 2004;47(4):947-961.
24. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 2003;46(4):499-511.
25. Schrodinger. Schrodinger, Inc.:http://www.schrodinger.com.
26. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 2006;49(2):534-553.
27. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46(12):2287-2303.
28. Bursulaya BD, Totrov M, Abagyan R, Brooks CL, 3rd. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17(11):755-763.
29. Xing L, Hodgkin E, Liu Q, Sedlock D. Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 2004;18(5):333-344.
30. Marsden PM, Puvanendrampillai D, Mitchell JB, Glen RC. Predicting protein-ligand binding affinities: a low scoring game? Org Biomol Chem 2004;2(22):3267-3273.
31. Gasteiger J. Handbook of chemoinformatics - The docking Problem (chapter 4.9 - vol 4 - Klebe et Al). Weinheim ; [Chichester]: Wiley-VCH; 2003. 1732-1763 p.
32. Knox AJ MM, Lloyd DG. Estrogen receptors: molecular interactions, virtual screening and future prospects. Curr Top Med Chem 2006;6(3):217-243.
33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
34. Abola E, Kuhn P, Earnest T, Stevens RC. Automation of X-ray crystallography. Nat Struct Biol 2000;7 Suppl:973-977.
35. Sorensen TL, McAuley KE, Flaig R, Duke EM. New light for science: synchrotron radiation in structural medicine. Trends Biotechnol 2006;24(11):500-508.
36. Marco E, Gago F. Overcoming the Inadequacies or Limitations of Experimental Structures as Drug Targets by Using Computational Modeling Tools and Molecular Dynamics Simulations. ChemMedChem 2007.

37. Blundell TL, Jhoti H, Abell C. High-throughput crystallography for lead discovery in drug design. Nat Rev Drug Discov 2002;1(1):45-54.
38. Brive L, Abagyan R. Computational structural proteomics. Ernst Schering Res Found Workshop 2002(38):149-166.
39. Lorber DM, Udo MK, Shoichet BK. Protein-protein docking with multiple residue conformations and residue substitutions. Protein Sci 2002;11(6):1393-1408.
40. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58(4):905-912.
41. Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2005;2(4):S127-136.
42. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. J Am Chem Soc 2005;127(26):9632-9640.
43. MOE. Molecular Operating Environment a software by Chemical Computing Group.www.chemcomp.com.
44. SYBIL. Molecular Modelling a sofware by TRIPOS.www.tripos.com.
45. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 1996;17:490–519.
46. Halgren TA. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 1996;17:520–552.
47. Halgren TA. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem 1996;17:553–586.
48. Halgren TA. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem 1996;17:616-641.
49. Halgren TA, Nachbar RB. Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem 1996;17:587–615.
50. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem 2004;25(9):1157-1174.
51. Shoichet Ia. J Chem Inf Model 2005;45(1):177-182
52. SPECS. Chemistry Solutions for drug discovery.http://www.secs.net.
53. WOMBAT. WOrld of Molecular BioAcTivity - Sunset Molecular Discovery LLC.http://sunsetmolecular.com.
54. Gillet VJ, Wild DJ, Willett P, Bradshaw J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. The Computer Journal 1998;41(8):547-558.
55. Snarey M, Terrett NK, Willett P, Wilton DJ. Comparison of algorithms for dissimilarity-based compound selection. J Mol Graph Model 1997;15(6):372-385.
56. Willet P. Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds. Journal of Computational Biology 1999;6(3/4):447-457.
57. Willet P. Chemical Similarity Searching. J Chem Inf Comput Sci 1998;38:983-996.

58. Butina D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. J Chem Inf Model 1999;39(4):747-750.
59. Oprea TI. Property distribution of drug-related chemical databases. J Comput Aided Mol Des 2000;14(3):251-264.
60. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44(3):793-806.
61. FILTER. a software by Openeye Scientific Software.http://eyesopen.com.
62. PipelinePilot. a software by SciTegic http://scitegic.com.
63. Lipinsky CA LF, Dominy BW Feeney PJ. Experimental and computational approached to estimate solubility and permeability in drug discovery and develepment settings. Adv Drug Deliv Rev 1997;23:3-25.
64. McGovern SL. A specific mechanism of nonspecific inhibition". J Med Chem 2003;46:4265-4272.
65. McGovern S. J Med Chem, 2003;46:4265-4272.
66. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate. J Biol Chem 1982;257(22):13650-13662.
67. Cronet P, Petersen JF, Folmer R, Blomberg N, Sjoblom K, Karlsson U, Lindstedt EL, Bamberg K. Structure of the PPARalpha and -gamma ligand binding domain in complex with AZ 242; ligand selectivity and agonist activation in the PPAR family. Structure 2001;9(8):699-706.
68. Garcia-Saez I, Parraga A, Phillips MF, Mantle TJ, Coll M. Molecular structure at 1.8 A of mouse liver class pi glutathione S-transferase complexed with S-(p-nitrobenzyl)glutathione and other inhibitors. J Mol Biol 1994;237(3):298-314.
69. Kauppi B, Jakob C, Farnegardh M, Yang J, Ahola H, Alarcon M, Calles K, Engstrom O, Harlan J, Muchmore S, Ramqvist AK, Thorell S, Ohman L, Greer J, Gustafsson JA, Carlstedt-Duke J, Carlquist M. The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. J Biol Chem 2003;278(25):22748-22754.
70. Lawrie AM, Noble ME, Tunnah P, Brown NR, Johnson LN, Endicott JA. Protein kinase inhibition by staurosporine revealed in details of the molecular interaction with CDK2. Nat Struct Biol 1997;4(10):796-801.
71. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell 1998;95(7):927-937.
72. Wang S, Tabernero L, Zhang M, Harms E, Van Etten RL, Stauffacher CV. Crystal structures of a low-molecular weight protein tyrosine phosphatase from Saccharomyces cerevisiae and its complex with the substrate p-nitrophenyl phosphate. Biochemistry 2000;39(8):1903-1914.
73. Weininger D. SMILES: a chemical language and information system. 1 Introduction to methodology and encoding rules J Chem Inf Comput Sci 1999;2:31-36.

74. TRIPOS. Tripos Mol2 File Format. www.tripos.com.
75. MDL. MDL SDF File Format - www.mdl.com.
76. Miller MA. Chemical database techniques in drug discovery. Nat Rev Drug Discov 2002;1(3):220-227.
77. Perola E, Charifson PS. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem 2004;47(10):2499-2510.
78. Kallblad P, Mancera RL, Todorov NP. Assessment of multiple binding modes in ligand-protein docking. J Med Chem 2004;47(13):3334-3337.
79. Knox AJ, Meegan MJ, Carta G, Lloyd DG. Considerations in compound database preparation--"hidden" impact on virtual screening results. J Chem Inf Model 2005;45(6):1908-1919.
80. Welch W, Ruppert J, Jain AN. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 1996;3(6):449-462.
81. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. Protein Sci 1998;7(4):938-950.
82. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161(2):269-288.
83. DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem 1988;31(4):722-729.
84. DesJarlais RL, Sheridan RP, Dixon JS, Kuntz ID, Venkataraghavan R. Docking flexible ligands to macromolecular receptors by molecular shape. J Med Chem 1986;29(11):2149-2153.
85. V. Sobolev RW, G Vriend, M Edelman. Molecular docking using surface complementarity. Proteins 1996;25:120-219.
86. Rarey M, Wefing S, Lengauer T. Placement of medium-sized molecular fragments into active sites of proteins. J Comput Aided Mol Des 1996;10(1):41-54.
87. Kramer B, Rarey M, Lengauer T. CASP2 experiences with docking flexible ligands using FlexX. Proteins 1997;Suppl 1:221-225.
88. Leach AR, Kuntz ID. Conformational Analysis of Flexible Ligands in Macromolecular Receptor Sites. Journal of Computational Chemistry 1991;13(6):730-748.
89. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3(11):935-949.
90. FRED. Fast Exhaustive Rigid Docking: a software by Openeye Scientific Software.http://eyesopen.com.
91. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ. Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. J Comput Aided Mol Des 2003;17(8):525-536.
92. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11(5):425-445.

93. RUBICON. distributed by Daylight Chemical Informations Systems Inc.www.daylight.com.
94. MOE: Molecular Operating Enviroment: Chemical Computer Group.
95. Bohm HJ. New Computational Approaches to Predict Ligand-Protein Interactions. Kubinyi H, editor: Wiley-VCH; 1998.
96. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 1996;9(1):1-5.
97. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 2002;46(1):34-40.
98. Connolly ML. Molecular surface Triangulation. Journal of Applied Crystallography 1985;18(6):499-505.
99. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28(7):849-857.
100. Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem 1959;14:1-63.
101. Jain T, Jayaram B. An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. FEBS Lett 2005;579(29):6659-6666.
102. Brandsdal BO, Smalas AO, Aqvist J. Free energy calculations show that acidic P1 variants undergo large pKa shifts upon binding to trypsin. Proteins 2006;64(3):740-748.
103. Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 2001;30:211-243.
104. Novotny J, Bruccoleri RE, Saul FA. On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. Biochemistry 1989;28(11):4735-4749.
105. Sotriffer CA, Gohlke H, Klebe G. Docking into knowledge-based potential fields: a comparative evaluation of DrugScore. J Med Chem 2002;45(10):1967-1970.
106. Gohlke H, Klebe G. Statistical potentials and scoring functions applied to protein-ligand binding. Curr Opin Struct Biol 2001;11(2):231-235.
107. Gohlke H, Klebe G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. J Med Chem 2002;45(19):4153-4170.
108. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295(2):337-356.
109. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL, 3rd. Assessing scoring functions for protein-ligand interactions. J Med Chem 2004;47(12):3032-3047.
110. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15(5):411-428.
111. Momany FA, Rone R. Validation of the general purpose QUANTA 3.2/CHARMm force field J Comput Chem 1992 13 (7):888-900
112. Morris G, Goodsell D, Halliday R, Huey R, Hart W, Belew R, Olson A. Automated docking using a Lamarckian genetic algorithm and an

empirical binding free energy function. Journal of Computational Chemistry 1999;19(14):1639-1662.
113. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44(7):1035-1042.
114. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. Biopolymers 2003;68(1):76-90.
115. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 2000;14(8):731-751.
116. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261(3):470-489.
117. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model 2005;23(5):395-407.
118. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 2005;48(7):2325-2335.
119. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 1999;42(5):791-804.
120. Muegge I, Martin YC, Hajduk PJ, Fesik SW. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. J Med Chem 1999;42(14):2498-2503.
121. Shimada J, Ishchenko AV, Shakhnovich EI. Analysis of knowledge-based protein-ligand potentials using a self-consistent method. Protein Sci 2000;9(4):765-775.
122. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 1999;37(2):228-241.
123. Bohm HJ. 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(3):243-256.
124. Muryshev AE, Tarasov DN, Butygin AV, Butygina OY, Aleksandrov AB, Nikitin SM. A novel scoring function for molecular docking. J Comput Aided Mol Des 2003;17(9):597-605.
125. Velec HF, Gohlke H, Klebe G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem 2005;48(20):6296-6303.
126. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15(4):327-332.
127. SMARTS. Smiles Extension - developed by Daylight (Chemical Information System, Inc).www.daylight.com.
128. Hindle SA, Rarey M, Buning C, Lengaue T. Flexible docking under pharmacophore type constraints. J Comput Aided Mol Des 2002;16(2):129-149.

129. Yang JM, Chen YF, Shen TW, Kristal BS, Hsu DF. Consensus scoring criteria for improving enrichment in virtual screening. J Chem Inf Model 2005;45(4):1134-1146.
130. Seifert MH. Assessing the discriminatory power of scoring functions for virtual screening. J Chem Inf Model 2006;46(3):1456-1465.
131. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49(20):5912-5931.
132. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R. Comparing protein-ligand docking programs is difficult. Proteins 2005;60(3):325-332.
133. Singh J, Deng Z, Narale G, Chuaqui C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. Chem Biol Drug Des 2006;67(1):5-12.
134. Deng Z, Chuaqui C, Singh J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem 2004;47(2):337-344.
135. Jameson DM, Seifried SE. Quantification of protein-protein interactions using fluorescence polarization. Methods 1999;19(2):222-233.

2. Permuting input strings for more effective generation of 3D conformers.

2.1 Abstract

SMILES (Simplified Molecular Input Line Entry System) strings and other classical 2D structural formats offer a convenient way to represent molecules as a simplistic connection table, with the inherent advantages of ease of handling and storage. In the context of virtual screening, chemical databases to be screened are often initially represented by canonicalised SMILES strings that can be filtered and pre-processed in a number of ways, resulting in molecules that occupy similar regions of chemical space to active compounds of a therapeutic target. A wide variety of software exists to convert molecules into SMILES format, namely, Mol2smi (Daylight Inc.), MOE (Chemical Computing Group) and Babel (Openeye Scientific Software). Depending on the algorithm employed, the atoms of a SMILES string defining a molecule can be ordered differently. Upon conversion to three dimensional coordinates they result in the production of ostensibly the same molecule. In this work, we show how different permutations of a SMILES string can affect conformer generation, affecting the reliability and the repeatability of the results. Furthermore, we propose a novel procedure for the generation of conformers, taking advantage of the permutation of the input strings – both SMILES and other 2D formats, leading to more effective sampling of conformation space in the output. A fingerprint and principal component analyses (PCA) step is implemented to post process and visualise the results. Finally, in collaboration with Johnson&Johnson the permutation of the input string was tested for Stochastic Proximity Embedding (SPE), a distance geometry method for the generation of conformers.

Published as:

- *Permuting input for more effective sampling of 3D conformer space*
  Giorgio Carta, Valeria Onnis, Andrew J. S. Knox, Darren Fayne and David G. Lloyd
  Comput Aided Mol Des (2006) 20:179–190
  DOI 10.1007/s10822-006-9044-4

- *On the Effects of Permuted Input on Conformational Sampling of Drug-like Molecules: an Evaluation of Stochastic Proximity Embedding.*
  Dimitris K. Agrafiotis, Deepak Bandyopadhyay, Giorgio Carta, Andrew J. S. Knox and David G. Lloyd
  Chem Biol Drug Des 2007; 70: 123–133
  DOI: 10.1111/j.1747-0285.2007.00538.

2.2 Introduction

The advent of drug design, concurrent with the accelerated evolution of computational power, allow us to manipulate and store large amounts of data in different formats, for pre-processing, substructure searching, 3D-matching and for the creation of a combinatorial library[1-3]. A widely used method to handle compound structures is to represent them in Simplified Molecular Input Line Entry Specification (SMILES) format. SMILES, proposed by Weininger in 1988[4] is a simple linear chemical language for specifying molecules or molecular fragments. SMILES is broadly used in computational chemistry and is supported by most major software tools in the field, reducing the volume of data exchange due to its compact nature. SMILES strings have been extensively used in both pharmaceutical and academic settings to uniquely code a 2D structure of a molecule with the correct stereochemical and protonated states represented.

However, a molecule although uniquely represented, can be described by differently ordered SMILES strings. SMILES denote a molecular structure as a graph with an optional chiral specification and SMILES describing only the labeled molecular graph (i.e. atoms and bonds, but no chiral or isotopic information) are known as generic SMILES. There are usually a large number of valid generic SMILES which represent a given structure. A canonicalisation algorithm exists to generate one unique generic SMILES among all valid possibilities[5].

Rules can be applied that specify the generation of SMILES but, different software packages produce different SMILES strings for the same encoded molecule. If the SMILES strings are used for representing the 2D structure of the molecules, possible permutations of the string should not affect the generation of the 3D structure. However, when the SMILES strings are used as input for the generation of conformers, it becomes clear that different permutations of the same string can generate diverse sets of conformers. This has a large impact for all applications and validations where a dataset was initially represented by SMILES strings. Our group has previously shown the impact that different SMILES strings representations have on the docking process in terms of Enrichment and *FP* rate calculations[6].

In that work a subset of the Derwent World Drug Index (WDI)[7] was selected using Lipinski's rules[8], by removing compounds with intrinsically non-drug like

properties using an MCL script implemented in the Daylight toolkit[5]. Additional compound filtering was carried out with FILTER[9]. The dataset was retained in SMILES format. A set comprising 40 active ligands (needles) for the estrogen receptor (ER) alpha was selected from literature where binding and/or anti-proliferative data were experimentally determined, with activities ranging from nanomolar to low micromolar potency. This active ligand set was then added to the validation set to make the total number of compounds up to 1000.

The Daylight toolkit[5] was used to export a dataset of actives in tab format, allowing retention of the correct assignment of isomeric where specified and 'ambiguous' SMILES where stereochemistry was not defined in the compound records. All structures were stored as SMILES format using two methods. MOL2SMI (Daylight Toolkit) and CONVERT (Molecular Networks GmbH)[10] were used to produce two alternate SMILES representations. Different software were compared for the generation of 3D structure (CORINA[11], OMEGA[12], RUBICON[13] and CATALYST[14]) and FRED2.01[15] was utilised in this study to dock all pre-processed compound sets. If alternate representations of SMILES strings were used as input 2-D structures, such as those produced by MOL2SMI and CONVERT, radically different effects on the enrichment rates were achieved when the same pre-processing protocols were utlilised in advance of rigid docking experiments.

To examine this concept, this chapter explores the effect the permutation of SMILES strings on the conformers generation using data from a set of 15 crystal structures extracted from a recent study by Vigers et al.[16] and also two X-ray structures of interest within our group [17,18] (Table 2-1). Our aim was to quantitatively show the influence that each permutation of a SMILES string, generated by several software programs, has on the production of conformers. We also show how this phenomenon can be used, in some cases, to increase the rate of production of a diverse set of conformers.

	PDB code	Ligand-Protein Complexes	Resolution	Perm[A]	P. Sel.[B]
1	1AQ1	CDK	2.2 Å	35	6
2	1D1Q	Tyrosine Phosphatase	1.7 Å	14	3
3	1GLQ	Glutathione-S-transferase	1.8 Å	30	6
4	1ABE	Arabinose-binding protein	1.7 Å	10	2
5	1AZM	Carboic Anidrase	2.0 Å	14	3
6	1CBX	Carboxypeptidase	2.0 Å	15	3
7	DFR	DHFR (dihydrofolate reductase)	1.7 Å	33	7
8	1EBP	Retinoic acid binding protein	2.2 Å	22	4
9	1HYT	Thermolysin	1.7 Å	15	3
10	1MRK	Ribosome inactivating protein	1.6 Å	19	4
11	1P93	Glucorticoid	2.7 Å	28	5
12	1PHF	Cytochrome p450-cam	1.6 Å	11	3
13	4PHV	HIV-protease	2.1 Å	46	9
14	1I7I	PPAR gamma	2.35 Å	28	6
15	1POC	Phospholipase A2	2.0 Å	31	6
16	1SRJ	Streptavidin	1.8 Å	22	4
17	1TPP	Trypsin	1.4 Å	15	3

Table 2-1: List of protein ligands complexes extracted from the PDB and their associated resolutions[A] Specifies the number of permutations generated by 'permsmi' (Daylight Inc.).[B] Indicates the number of strings selected to carry out the analysis, including the canonical SMILES string, as generated by Mol2smi (daylight Inc).

Figure 2-1 illustrates the process whereby a single PPARγ agonist 3D structure was reduced to SMILES format using Mol2smi (Daylight Chemical Information Systems Inc.)[5], MOE (Chemical Computing Group)[19], Babel (Openeye Scientific Software)[20] and ChemSketch (Advanced Chemistry Development)[21]. The SMILES string produced by each software package, as shown in Figure 2-1, are different for the same molecule.

As SMILES notation represents a linear form of a connection table, we also wanted to investigate if the concept was extendible to a more elaborate connection table, as found in formats such as sdf (Structures Data File)[22]

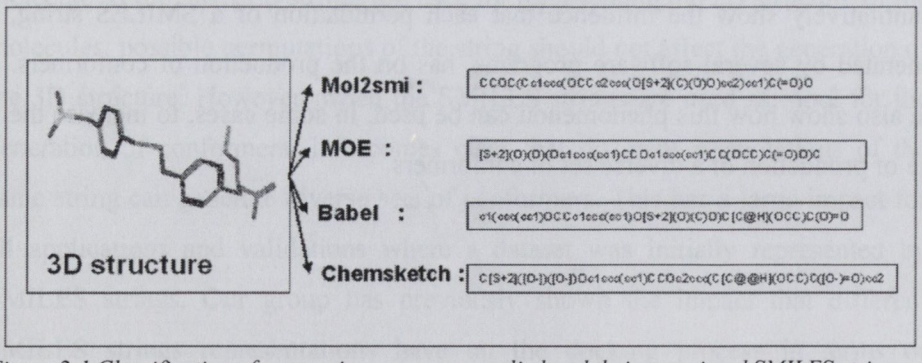


Figure 2-1 Classification of conversion programs applied and their associated SMILES strings Generated.

Permuted SMILES strings were generated for ligands extracted from 17 crystal structures in the PDB and then used as input for different conformer

generating software packages (Omega v1.8.1, Corina v3.6, Rubicon and Catalyst v4.9.1; also in SPE in the *Further Application* section).

The conformational space explored by the same output quantity of conformers arising from a permuted set of a SMILES, in comparison to conformational space sampled when using a single unique canonical SMILES string were investigated using PCA of several 3D descriptors and using the potential energy.

To further explore the effect of the input connection table and SMILES permutations on the different algorithms for the generation of conformers a later study in collaboration with Johnson&Johnson (Dimitris K. Agrafiotis) was carried out and included in the section *Further application*. In this study a validation and examination of the performance of SPE, using a smaller set of five X-ray crystal structures was used (1CBX, 1ETT, 1HVR, 4DFR, 1GLQ).

SPE is a distance geometry method aiming at generating conformations that satisfy a set of interatomic distance constraints derived from the molecular connectivity table and defined in the form of lower and upper distance bounds$\{l_{ij}\}$ and $\{u_{ij}\}$ for each atom pairs *i* and *j* in the system. (See Appendix to Chapter 2).

The results obtained on this distance geometry method establish that the conformational ensemble produced by SPE using different permuted inputs is statistically indistinguishable and within the range of variability that would be expected from the stochastic nature of the method itself. This, along with the results of a more comprehensive comparative study (Agrafiotis et al., J. Chem. Info. Model, 2007, in press), provides evidence that SPE is a robust conformational search method.

2.3 Methods

Seventeen crystal structures from the PDB were selected. Most of the structures were selected from the list used to validate FlexX [23]/Gold [24] and also MASC [16]. All structures were solved at a resolution ≤ 2.5 Å, except for 1P93 (2.7 Å) which is the glucocorticoid complex. The ligands represent a wide variety of different structural characteristics such as size, number of rotatable bonds and features. The list of compounds studied is shown in Table 2-1

2.3.1 Preparation of Ligands

The structures of complexes were downloaded and visually inspected with MOE [19]. Ligands were extracted from the complexes and all hydrogen and coordination atoms deleted. Mol2smi by Daylight was utilised to reduce the 3D representation to an initial canonical SMILES version.

Permsmi[5] was subsequently used to produce permutations of the SMILES string, starting with each of the atoms in the input SMILES in turn. The set of permutations of the molecule are written out in pseudo-canonical order (i.e. ring closures are renumbered). Thus, the longer the string, the higher the numbers of permutations 'Permsmi' is able to generate. An average of 22.8 permutated smiles strings were created for each molecule. We chose to use every fifth permutated SMILES string of a molecule as we deemed it to be representative of the set produced by Permsmi, making sure to include the canonical SMILES strings initially produced by mol2smi. This procedure narrows down to an average of 5.5 permutation strings per molecule, which were used as input for generating 3D conformers as illustrated in Figure 2-2.

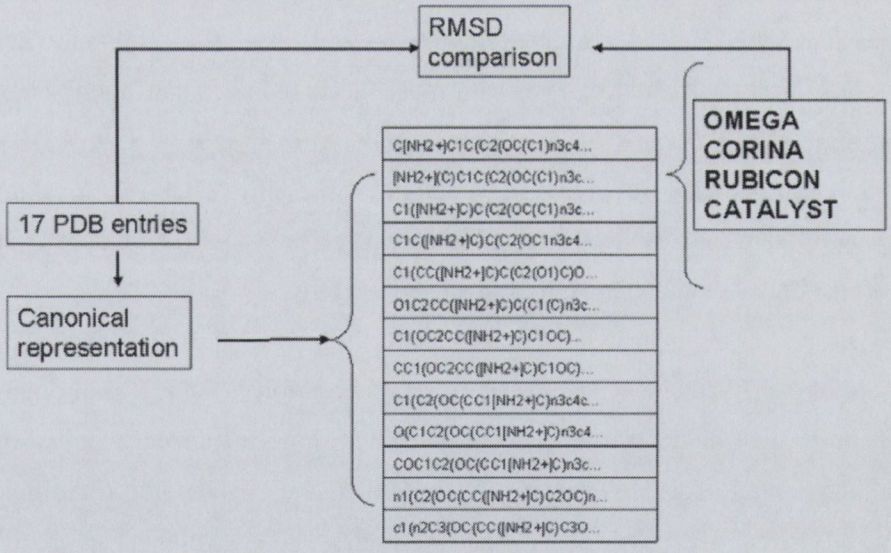


Figure 2-2 Protocol for the evaluation of SMILES string permutations.

2.3.2 Conformer Generation

Corina: Corina v3.6 [11] produces 3D conformations by combining monocentric fragments of standard bond lengths, appropriate angles and dihedral angles. To incorporate appropriate torsion angles, Corina uses a set of rules derived from statistical analysis of the Cambridge Structural Database[25]. Furthermore, Corina is able to generate only a single conformer per SMILES string. Corina, retaining default settings, was used to generate conformers of the permuted list and the output was saved in mol2 format. One conformer per SMILES string was created to a total of **94** conformers (permutation average generated * 17 PDB ligands)

Omega: Omega v1.8.1 [12] generates conformers in three separate steps. First, a generic 3D structure is built. This primary conformation can be minimised with the MMFF force field. In the second stage, in order to generate different 3D conformations, Omega applies a depth-first search (DFS) divide and conquer algorithm (DC). DFS is an algorithm for traversing a searching tree, tree structure or graph. DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it had not finished exploring.

DC is an important algorithm design paradigm. It works by recursively breaking down a problem into two or more sub-problems of the same (or related) type, until these become simple enough to be solved directly.

More simply, Omega disassembles the molecules into small fragments, and calculates the fragments energy using a Dreiding force field[26], and finally reassembles the fragments in order. The generated conformations are limited by their total energy (e-windows parameter in OMEGA - the Dreiding force field has no electrostatic interaction) and those molecules whose RMSD values are too similar (RMSD value cut-off) are discarded. The third and final step offers the possibility to refine conformations with standard MMFF[27-31] force minimisation to avoid the presence of high energy unrealistic conformations.

Omega was set to generate 50 conformers per SMILES string. The output file was set to save as mol2 format and the number of rotatable bonds increased

to 30, to allow all molecules to pass the Omega filter. All other remaining parameters were kept as default. 1azm was discarded by the Omega filter due to the presence of a zinc atom. A total of **2,406** conformers were generated.

Catalyst: Catalyst [32] uses a poling algorithm for the generation of conformers with the CHARMM force field[33], which does not include electrostatic terms. There are two integrated algorithmic methods of conformation creation to choose from: the BEST and FAST algorithms, of which FAST were the method of choice for this study. Its conformational search algorithm is based on a grid search of torsion values in combination with energy minimisation to resolve high-energy configurations, such as overlaps of vdW radii.

A maximum number of 50 conformers per each SMILES string were created with the output saved in sdf format. Again, 1azm was discarded, and a total of **3,623** conformers were generated.

Rubicon: Employing Rubicon[13], conformers are initially generated with a distance-geometry method that samples random conformations – this is by default a non-stochastic process, yielding reproducible output for a given input string in separate runs. In other words RUBICON uses a pseudo-random generator which provides identical behaviour on all platforms. The parameter RUBE_SEED sets the seed for this number generator, which must be in the range 0 to 9,000,000,000. Given otherwise identical input, Rubicon produces identical results for a given seed on all platforms. These three dimensional structures are then minimised for distance and volume bounds violations. Only one minimiser algorithm is provided with version 4.3 of Rubicon [34]. SMILES strings were first converted into tdt (Thor Data Tree) format by smi2tdt [5] since Rubicon does not read the SMILES format and the output was set to pdb format. A total number of **3,845** conformers were generated. The default settings were used, except where noted, when implementing the conformer generation software packages.

All conformations were superimposed by matching non-hydrogen atoms, to the corresponding unmodified X-ray ligand structure. To remove artificial differences resulting from RMSD deviations attributed to automorphisms (an artificial RMSD difference caused by symmetry in the structure)[35,36], the

program RMSD (Openeye Scientific Software) was used for analysis. The minimum RMSD values (the closest conformers to the crystal structure) and the standard deviations were then calculated for each set.

As sampling conformer space with an ensemble size of fifty output molecules is not sufficient to allow the software to produce a reasonable and diverse ensemble of conformers, two further experiments were performed. Thus, a subsequent RMSD analysis was carried out allowing the generation of a larger number of conformers (sets A and B described below).

Set A: From the original set of crystal structures we selected a molecular "probe" characterised by an average number of rotatable bonds (4DFR, 10 RBs) – representative of the typical drug-like compounds encountered in commercial screening datasets used in virtual screening. The canonical string (cansmi) and subsequently eight permutations (permsi) were generated as described above. In this study, the conformers were built using OMEGA. The 4DFR canonical SMILES string was first used and the highest number of conformers with the default setting was obtained (a total 280 conformers with *max_confs* flag set to 500). To produce a comparable set of conformers the same number of output structures were replicated for the permuted string setting the *max_conf* flag to 35 (35 confs by 8 permutations). The results were then imported in MOE and 3D potential energy descriptor (MMFF94), PCA analysis were then calculated and plotted.

Set B: The same canonical SMILES strings and the eight permutations were then used for a subsequent run. OMEGA was used to generate conformers from the canonical string with a *RMSD (-rms* flag) cut off of 0.5 Å (default = 1) and an increased *energy window* (*ewindow* flag) to reflect the energy window span demonstrated by the conformers previously generated in A in terms of potential energy. (about 60 kcal/mol calculate with MMFF94 force field). *Max_confs* flag was increased to 1,000,000 to remove any restriction in the number of output structures. This produced an upper ensemble of conformers from the single input string. As in experiment ***A,*** similar final number of conformers were produced from the permuted strings through setting *max_confs* to 310 and leaving the *energy window* and the *RMSD* cut off at the default

settings. Again, the sets generated were imported to MOE and 3D potential energy descriptors were generated and PCA analysis were performed and plotted in graphs.

2.3.3 Fingerprint Analysis

Conformers generated from the ensemble of permutations of a single molecule were then considered. For example, a set was represented as a collection of conformers produced from the six permutations of the PPARγ ligand SMILES string. To measure the number of duplicate conformations presents in the set we imported the conformational file into MOE and an eigenvalue spectrum shape fingerprint ESshape3D [19] analysis was undertaken. Each fingerprint is of fixed length and allows for a comparison of 3D molecular shapes considering the heavy atoms of a molecule. Duplicate conformers appeared in the database with identical fingerprints, allowing us to calculate the percentage of unique conformers.

2.3.4 Principal Component Analysis

In order to investigate the chemical space explored set **A** and **B** were analysed with a Principal Component Analysis. A total of six MOE 3D descriptors (from all 3D descriptors, x-y-z dipole moment, and Principal moment of inertia x-y-z) were computed and selected for all the conformers generated to reflect the different topological characteristics of the 3D molecules in the output datasets.

The aim of PCA is to reduce the dimensionality of a set of descriptors by linearly transforming the data. In our analysis six vectors have been transformed into a manageable 3D space described by three principal component vectors, where each of the 3D vectors is a combination of the selected descriptors.

To achieve this result the number of principal components were limited to three and all the other options in the MOE PCA module were run as default settings, without applying weightings to any of the thirty descriptors selected.
The three principal components were then plotted as XYZ coordinates respectively for each compound. At the end of the process each molecule in the

database is described by a set of coordinates, and may be plotted relative to all other members in the three dimensional space.

2.3.5 Alternative 2D representations: connection tables.

The connection table is a common way to simply represent a chemical structure. A typical file extension that contains a connection table describing the organisation of the atoms within a molecule is the SDF file format[37]. Typically an SDF file includes two essential main sections: the first describing the 3D coordinates of the individual atoms in the molecules, the second depicting the connections of each atom with the others and the bond order present (1 = single bond, 2 = double bond)[37].

```
Section 1 (3D coordinates)
-------------------------------------------------------------

 54 57  0  0  0  0  0  0  0  0  0
   -0.3250     0.1625     0.0000 C   0  0  0  0  0
    2.6042     1.8667     0.0000 C   0  0  0  0  0
   -0.3250    -0.9917     0.0000 C   0  0  0  0  0
    0.6458     0.7083     0.0000 C   0  0  0  0  0
    2.6042     0.7083     0.0000 C   0  0  0  0  0
    1.6250     0.1625     0.0000 C   0  0  0  0  0
     ....
     ....
-------------------------------------------------------------
Section 2 (Connection table)
-------------------------------------------------------------

  3  1  1  0  0  0
  4  1  1  0  0  0
 12  1  1  0  0  0
  1 22  1  0  0  0
  2  9  1  0  0  0
  8  2  1  0  0  0
  2 23  1  0  0  0
     ....
     ....
-------------------------------------------------------------
```

For this part of our study a PERL script (see appendix for code and details) script utility was written to produce random permutation of the connection table. In detail, the utility randomises the position of the first section of the file and, while keeping track of atom movement, it rewrites a new

connection table underneath. To avoid the production of duplicates an adjacency matrix is first calculated from the connection table and subsequently its connectivity stack[38]. The connectivity stack of a matrix is a sequence of the adjacency matrix values $s_k = a_{ij}$:

$$S = (s_1, s_2, s_3 \dots s_{1k} \dots s_n) \qquad 2\text{-}1$$

where the position of the k_{th} element is given by:

$$k = i + \frac{(j-1)(j-2)}{2} \quad with\ (i < j) \qquad 2\text{-}2$$

Identical stack matrixes correspond to identical adjacency matrix and, thus, identical connection tables.

To investigate if permutation of the connection tables resulted in the same effect observed for the permuted SMILES strings on conformer generation, we first generated an SDF from the ligand co-crystallised in PDB entry 4DFR[39]. OMEGA was then used to generate 280 conformers *(-max_confs* set to 500 and remaining parameters as default).

As described above, the perl script was used to produce ten permutations of the crystal structure sdf file. Finally, to reproduce the same amount of conformers as above, *–max_conf* was set to 10 to furnish 280 conformers from the list of permuted connection tables of 4DFR.

A RMSD calculation was then carried out between the resulting conformers and the co-crystal ligand in 4DFR using OpenEye's RMSD[40].

2.4 Results and Discussion

Two categories of data are reported in the results: a quantitative result taking into account the number of conformers generated by each software package and a qualitative result, where we considered the characteristics of the conformers generated.

From a quantitative point of view, some software tools show interdependence between string permutations and the numbers of conformers generated. In others words, a different number of conformers were generated when starting from a different permutation of the same SMILES string. In particular, Omega and Rubicon exhibited this behaviour, as although the maximum numbers of conformers to be generated was set to 50, this number was not always constructed. For example, Rubicon generated 27 conformers from the initial canonical SMILES string and up to 32 for one of its permutations for the ligand extracted from 4DFR. Omega produced only 9 conformers for the canonical string and 14 for a permutation of it in the case of the X-ray ligand taken from 1MRK. Less affected was Catalyst, which in the worst case (1CBX) generated 39 conformers for the canonical and 36 for most other permutations. The full set of results is shown in tables 2-8. Results in tables 2, 3 and 4 report the RMS deviation values calculated for canonical SMILES strings and their permutations. Cells labelled with 'D' were molecules discarded by the software in the case of both Omega and Rubicon. Cells labelled 'A' indicates that standard deviation was not computable because just one conformer per string was produced. Empty cells indicate that no permutation was produced for that string.

It is clear from Table 2-2 that each set of the permutated SMILES string generated molecules exhibits a different standard deviation (SD) of the RMSD values for each set of conformers compared to the X-ray structure. That suggests that each set includes a different pool of RMSD values when compared with their respective crystal structures. Similar results as those generated by Omega are also observed in the case of Rubicon, with different RMSD values also being produced (Table 2-3). Table 2-4, however, shows that Catalyst is not as affected as Omega or Rubicon, in the conformers being generated from the permutations of the canonical SMILES string. Approximately, the same RMSD values appear

for the set indicating that identical conformers were created by each permutation. In the case of Corina, no RMSD calculations could be carried out as only one conformer per SMILES string could be produced.

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	A	0.21	0.407	0.28	D	0.353	0.291	0.369	0.36	0.426	0.12	0	0.717	0.434	0.417	0.417	0.516
P1[b]	0.007	0.021	0.398	0.249	D	0.258	0.53	0.355	0.269	0.394	0.184	0.304	0.741	0.666	0.342	0.698	0.516
P2[b]	A	0.2	0.456	0.224	D	0.344	0.316	0.308	0.135	0.445	0.163	0	0.568	0.888	0.324	0.85	0.6
P3[b]	A	0.177	0.461		D	0.338	0.365	0.349	0.337	0.316	0.099	0.304	0.602	0.88	0.345	0.253	0.554
P4[b]	A		0.477				0.461	0.332		0.443	0.163		0.612	0.767	0.476	0.864	
P5[b]	A		0.476				0.346				0.177		0.496	0.649	0.425		
P6[b]	A		0.487				0.437						0.532	0.426	0.433		
P7[b]							0.504						0.695				
P8[b]													0.495				
P9[b]													0.496				

Table 2-2 RMSD Standard Deviation for Omega.[a] can = canonical SMILES string
b 'P1, P2, P3…' = different permutations of the canonical SMILES string *(all units in Å).*

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
can[a]	0.447	0.067	0.348	0.217	D	0.303	0.422	0.455	0.369	0.337	0.288	0.149	0.642	0.385	0.264	0.834	0.300
P1[b]	0.337	0.047	0.296	0.215	D	0.267	0.357	0.493	0.327	0.333	0.330	0.182	0.607	0.482	0.249	0.754	0.252
P2[b]	0.401	0.063	0.405	0.193	D	0.340	0.325	0.45	0.411	0.351	0.345	0.154	0.548	0.405	0.206	0.781	0.261
P3[b]	0.424	0.046	0.294		D	0.367	0.35	0.519	0.350	0.307	0.321	0.181	0.533	0.348	0.290	0.853	0.253
P4[b]	0.428		0.33				0.251	0.471		0.363	0.321		0.503	0.329	0.282	0.761	
P5[b]	0.382		0.327				0.322				0.352		0.628	0.411	0.235		
P6[b]	0.404		0.338				0.316						0.565	0.34	0.215		
P7[b]							0.338						0.477				
P8[b]													0.544				
P9[b]													0.628				

Table 2-3 RMSD Standard Deviation for Rubicon[a] can = canonical SMILES string
[b] 'P1, P2, P3…' = different permutations of the canonical SMILES string (all units in Å).

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
can[a]	0.445	0.179	0.596	0.285	0.346	0.413	0.519	0.476	0.382	0.419	0.209	A	0.670	0.806	0.377	0.768	0.278
P1[b]	0.487	0.177	0.586	0.264	0.346	0.413	0.519	0.476	0.382	0.470	0.209	A	0.598	0.806	0.377	0.768	0.278
P2[b]	0.340	0.177	0.586	0.198	0.346	0.302	0.519	0.476	0.384	0.470	0.209	A	0.598	0.764	0.319	0.768	0.278
P3[b]	0.483	0.177	0.626		0.346	0.413	0.519	0.476	0.382	0.439	0.209	A	0.598	0.764	0.319	0.768	0.278
P4[b]	0.483		0.626				0.519	0.476		0.477	0.209		0.670	0.764	0.319	0.768	
P5[b]	0.483		0.622				0.519				0.209		0.670	0.806	0.319		
P6[b]	0.483		0.622				0.426						0.598	0.806	0.377		
P7[b]							0.519						0.598				
P8[b]													0.670				
P9[b]													0.670				

Table 2-4 RMSD Standard Deviation for Catalyst.[a] can = canonical SMILES string
[b] 'P1, P2, P3…' = different permutations of the canonical SMILES string (all unit sin Å).

We next sought to assess the quality of the conformers produced by each software package with respect to the different permutations of a SMILES string. Table 2-5, Table 2-6, Table 2-7 and Table 2-8 show the minimum RMSD values retrieved from each set. Label D and empty spaces have the same significance as in the previous tables.

Table 2-5 depicts that the minimum RMSD values (the closest conformers to the crystal structure) are derived from a permutated string rather than the canonical one which would be normally used in a virtual screening process. Importantly, in some cases, the lowest RMSD value is nearly half that observed with the canonical SMILES string (eg.1POC, 4DFR, 1ABE). As before, we see that the same holds for Rubicon as in the case of Omega, where the minimum RMSD values retrieved in the permuted sets are lower than those of the canonical string (Table 2-8). A less dramatic difference is observed when using Catalyst or Corina.

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	1.82	0.3	1.52	0.61	D	0.42	1.7	0.53	0.68	0.8	1.3	0.24	2.41	3.41	1.06	0.39	0.3
P1[b]	1.26	0.52	1.73	0.68	D	0.68	0.76	0.74	0.45	0.95	1.3	0.24	1.72	2.1	1.18	0.39	0.3
P2	1.77	0.32	1.44	0.37	D	0.49	1.98	1.19	1	0.71	1.64	0.67	2.38	1.98	1.26	0.18	0.3
P3	0.86	0.29	1.72		D	0.68	1.96	0.61	0.45	1.06	1.26	0.24	2.28	1.97	1.04	0.46	0.3
P4	1.27		1.17				1.17	0.66		0.83	1.61		1.99	1.73	0.78	0.19	
P5	1.51		1.27				1.98				1.25		2.57	2.49	0.95		
P6	1.43		1.66				1.15						2.41	3.44	1.1		
P7							1.09						1.86				
P8													2.45				
P9													2.57				

Table 2-5 Minimum RMSD for Omega. [a] *can = canonical smile string*
[b] *'P1, P2, P3... ' = different permutations of the canonical SMILES string (all units in Å).*

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can	0.45	0.31	1.53	0.38	D	0.49	1.67	0.57	0.24	0.91	1.17	0.1	2.02	2.46	1.26	0.17	0.39
P1	0.38	0.31	1.28	0.11	D	0.61	1.73	0.89	0.23	0.94	0.99	0.09	1.72	2.12	1.46	0.19	0.46
P2	0.45	0.31	1.49	0.47	D	0.46	2.07	0.45	0.24	0.8	0.66	0.11	2.09	2.54	1.76	0.18	0.39
P3	0.47	0.31	1.35		D	0.49	1.84	0.76	0.25	0.85	1.12	0.09	2.24	2.57	1.3	0.18	0.47
P4	0.35		1.29				2.37	0.61		0.81	1.12		2.53	2.50	1.25	0.17	
P5	0.37		1.42				1.93				1.12		2.02	2.31	1.53		
P6	0.43		1.32				2.22						2.29	2.43	1.68		
P7							2.05						2.2				
P8													2.29				
P9													2.02				

Table 2-6. Minimum RMSD for Rubicon. [a] *can = canonical smile string*
[b] *'P1, P2, P3... ' = different permutations of the canonical SMILES string (all unit in Å).*

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
can	0.53	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	1.91	2.02	1.27	0.17	0.43
p1	0.26	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.02	1.27	0.17	0.43
p2	0.26	0.42	0.99	0.38	0.32	0.58	1.16	1.22	0.29	0.82	1.53	0.25	2.18	2.08	1.48	0.17	0.43
p3	0.26	0.42	1.35		0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.08	1.48	0.17	1.31
p4	0.26		1.35				1.16	1.22		0.8	1.53		1.91	2.08	1.48	0.17	
p5	0.26		1.35				1.16				1.53		1.91	2.02	1.48		
p6	0.26		1.35				1.3						2.18	2.02	1.27		
p7							1.16						2.18				
p8													1.91				
p9													1.91				

Table 2-7. Minimum RMSD for Catalys.[a] *can = canonical smile string*
[b] *'P1, P2, P3... ' = different permutations of the canonical SMILES string (all units in Å).*

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
can	1.14	0.76	2.38	0.63	1.31	0.56	2.57	1.35	0.91	1.09	1.47	0.46	3.78	1.82	4.15	0.16	0.42
p1	1.85	0.76	2.38	0.78	1.35	0.56	2.57	1.34	0.91	1.71	1.44	0.46	2.74	1.86	4.15	0.16	0.42
p2	1.82	0.76	2.41	0.73	1.35	1.52	2.58	1.46	1.49	1.71	1.46	0.46	2.75	2.24	4.15	0.16	0.42
p3	1.83	0.76	2.38		1.35	0.56	2.57	1.29	0.91	1.94	1.81	0.46	2.58	2.22	4.9	0.16	0.42
p4	1.83		2.38				1.98	1.02		1.96	1.22		2.05	2.24	4.9	0.16	
p5	1.82		2.41				1.97				0.56		2.59	2.19	4.15		
p6	1.81		2.41				1.75						2.75	1.82	4.15		
p7							1.97						2.48				
p8													2.05				
p9													2.05				

Table 2-8. Minimum RMSD for Corina. [a] *can = canonical smile string*
[b] *'P1, P2, P3... ' = different permutations of the canonical SMILES string (all units in Å).*

Table 2-9 (set A) and Table 2-10 (set B), show the RMSD values obtained comparing the crystal structure of 4DFR and different sets of conformers.

Results for set A: In detail, Table 2-9 shows that operating with default parameters and allowing OMEGA to produce as many conformers as it can, the output from the treatment of permuted input strings delivers conformers closer to the bioactive conformation when compared to those generated from the canonical string (280 final unique conformers were produced both from the canonical and the permuted string of 4DFR)

Table A	Can	Perm
min :	1.29	1
max :	3.44	3.44
Average :	2.52921429	2.492357143
Dev St.	0.39585861	0.478173962

Table 2-9 RMSD minimum, maximum, average and standard deviation calculation for the pdb entry 4DFR4DFR: the table shows the set where 280 conformers were produced.

The graph in Figure 2-3 shows a comparison between the potential energy of the conformers generated from the canonical and the permuted strings. It is quite clear that the potential energy span produced by the conformers from the permuted input is wider then that originating from the canonical input.

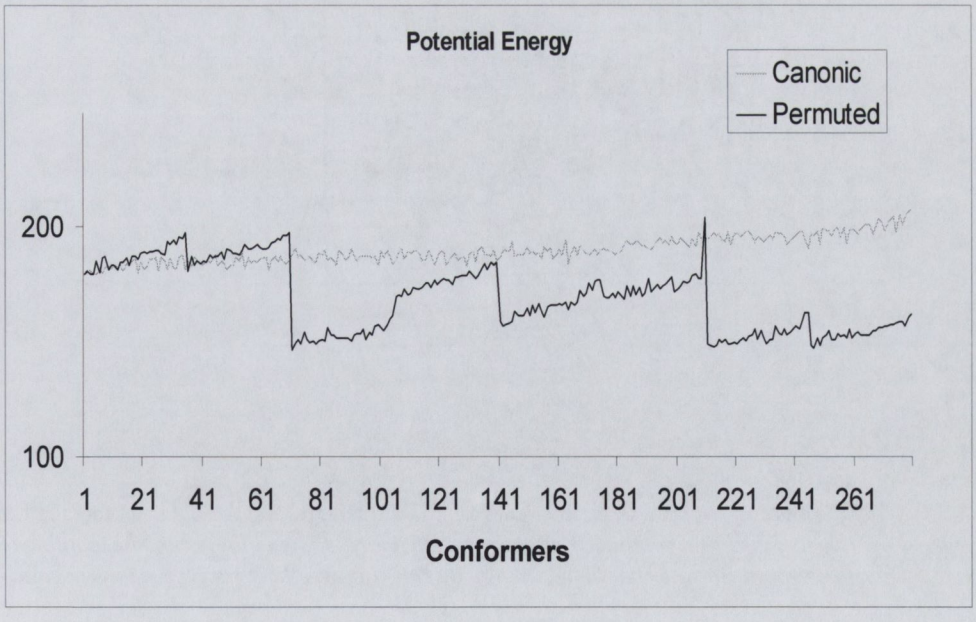


Figure 2-3 Potential energy graph for conformer set A, generated using omega (280 members). Black line illustrates result for permuted input string, grey illustrates result for canonical.

To further examine this phenomenon we undertook to represent the conformational space explored by the two output sets using PCA analysis. The data is shown in Figure 2-4.

In red is depicted the conformational space location of the co-crystal ligand (the bioactive conformer) from 4DFR. Interestingly, we note that the immediate conformational neighbourhood around the bioactive conformer predominately originates from the permuted string output. Moreover, the conformational space explored by the conformers from the canonical structure appears to be smaller and less diverse then that from the permuted strings. Hypothetically, by exhaustively rotating every single rotatable bond of 4DFR, no matter the original format (canonical or permuted), by small steps (i.e. 1 degree * 360), we would have obtained a full coverage of the conformational space in Figure 2-4, with some blank areas due to not sterically or stereochemically-allowed conformations.

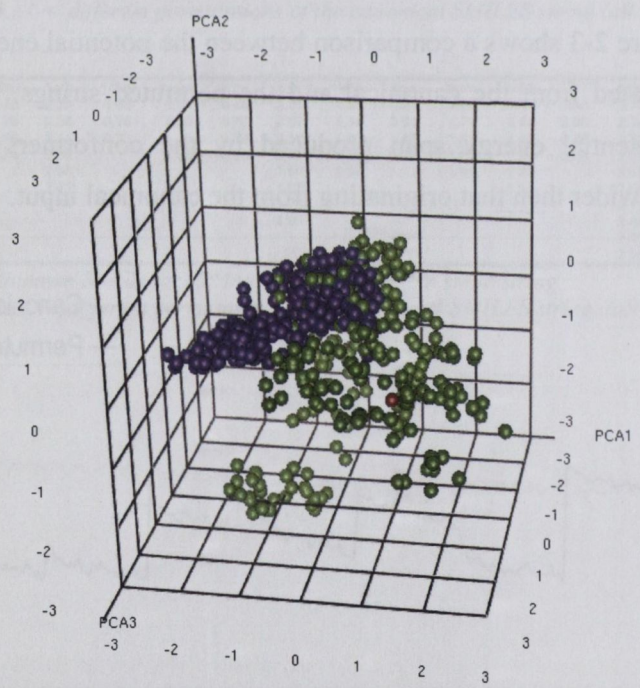


Figure 2-4 PCA analysis of conformers generated by Omega (set A): in violet the 280 conformers produced from the canonical string of 4DFR are shown. In green the conformers generated from 8 permutation of the same string. In red the crystal structure of the ligand of 4DFR.

Results set B: In Table 2-10 the unique conformers produced from the permuted strings are once again closer to the bioactive conformer (RMSD 0.76 Å calculated for permuted strings against 1.23 Å from canonical), when comparing

ensembles of similar size. From the canonical input, OMEGA produced 2,239 conformers. The set originating from the permuted input comprised of 2,326 conformers.

Table B	Can	Perm
min :	1.23	0.76
max :	3.47	3.44
Average :	2.53093345	2.398963887
Dev St.	0.38276039	0.476240746

Table 2-10 RMSD minimum, maximum, average and standard deviation calculation for the pdb entry 4DFR: the table shows the set where 2,239 and 2,326 conformers were produced for the canonical and the permuted SMILES strings respectively.

As with set A, the potential energy has been calculated as previously described. In this case, the potential energy span originating from the permuted strings input is larger, members of the ensemble are energetically much closer to the energy calculated for the bioactive conformation than those members of the ensemble originating from the canonical input (Figure 2-5).

Figure 2-6 shows the plotted result of the PCA for set B. It depicts clearly how the conformational space explored by the conformers originating from permutated input is diverse and closely surrounds the bioactive space.

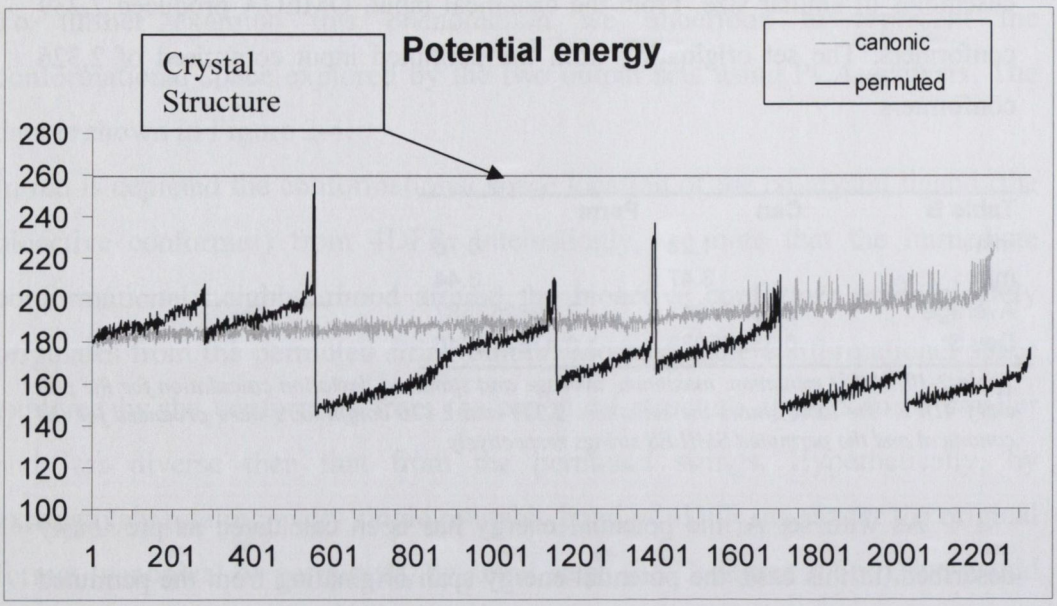


Figure 2-5 The Potential energy graph for conformer set B, generated using OMEGA (ca. 2,200 members). The black line illustrates the result for the permuted input string, the grey illustrates result for the canonical.

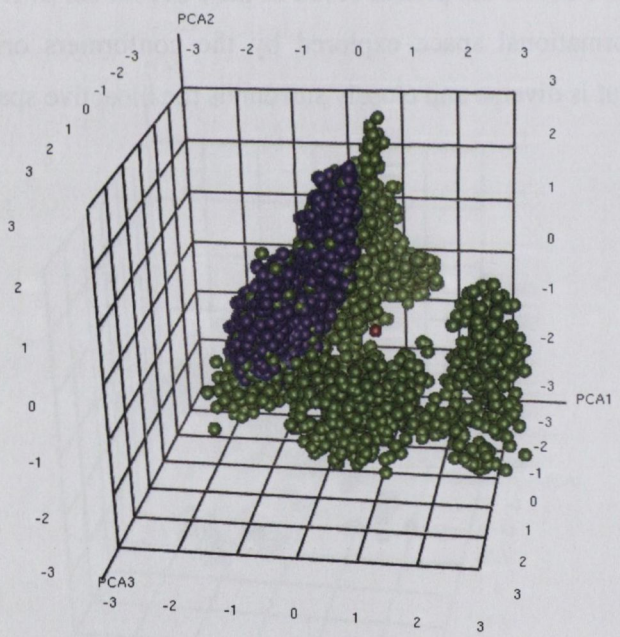


Figure 2-6 PCA of conformers generated by OMEGA (set B). In violet the conformers (about 1200) generated setting an energy window equal of 60 and a RMSD cut off of 0.5 Å from the canonical string of 4DFR. In green the conformers generated from the 8 permutations of 4DFR keeping all the parameters in OMEGA as default (with the exception of –max_conf, that was set to 310 to generate a comparable number of final conformers with the canonical set. In red the PCA coordinate calculated for the crystal structure is shown.

2.4.1 Fingerprint Analysis

A critical method for incorporating ligand flexibility in a vHTS process is to generate multiple conformers of each molecule. One of the key factors to take into account at this stage in the process is the balance between the number of conformers generated and the diversity of each with regard to conformational space sampling. Generally, less conformers with a wider degree of RMSD separation between each is beneficial.

To investigate the uniqueness of the conformers generated from each permutated SMILES string and the software used, all the conformers were read into MOE and fingerprint analysis carried out using Esshape3D. By exploring the percentage of conformers that were identical in each set for a molecule it was possible to determine whether this was due to the orientation of each conformer. For example, it was possible that the ensemble of conformers derived from different SMILES permutations of 1AQ1, contained duplicates.

Figure 2-7 shows the result from the fingerprint analysis for OMEGA, Catalyst and Rubicon. The numbers of conformers produced by each software package were normalised to the one with the highest value. Thus, Rubicon (3,845) was 100%, Catalyst (3,623) 94.22% and OMEGA (2,406) 62.57%. The percentage of unique conformers generated from the permuted SMILES strings by Omega was 99.38% of the total. If the conformers were generated from only the canonical SMILES string, the percentage of the total would be just 11.33%. It is apparent at this stage that generating conformers using Omega with default settings is highly efficient when combined with permuted SMILES strings. In the case of Rubicon, 87.73% of the total numbers of conformer were unique in conformational space. Again starting from a single SMILES string, only 18.11% would be unique. To a lesser extent, but still evident, was Catalyst with the percentage of unique conformers being 23.98% of the total. Beginning from a single SMILES string would allow 17.06% unique conformers to be produced.

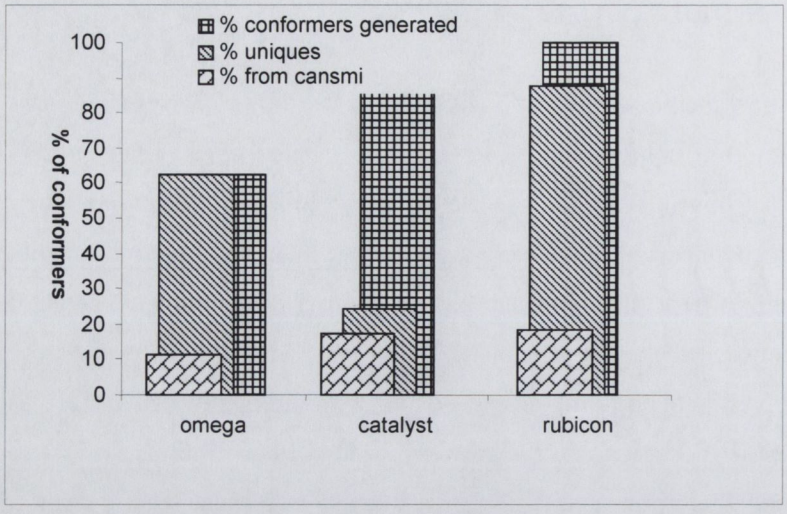


Figure 2-7 Fingerprint analysis of conformers generated by Omega, Catalyst and Rubicon. The height of the 3 bar graphs is proportional to the total amount of conformers generated. Cansmi: Canonical SMILES string.

Since Corina produced just one conformer per string (94 total amount = 100%) it is represented on its own in Figure 2-8 so as to keep it to scale. The percentage of unique conformations using Corina was 71.28%. With only the canonical SMILES string it would have been 18.18% of the total.

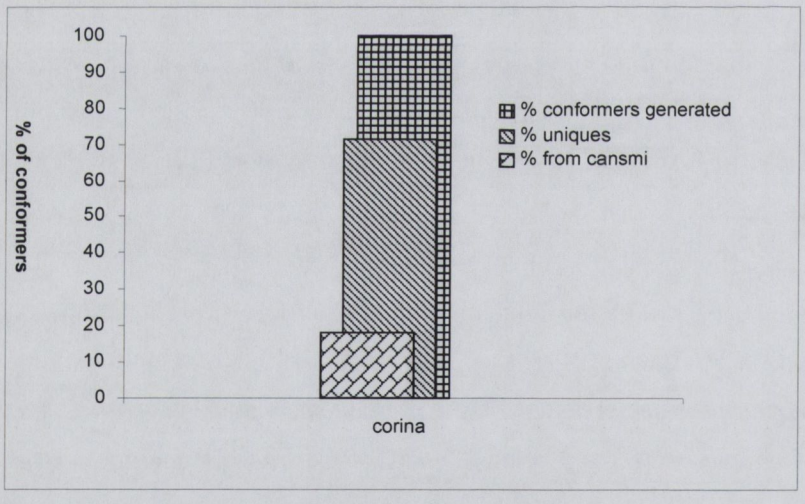


Figure 2-8 Fingerprint analysis of conformers generated by Corina.

2.4.2 Connection Table analysis.

Table 2-11 depicts the RMSD difference between the crystal structure and the conformers produced from the crystal structure itself. Subsequently RMSD analysis was also carried out between the conformers produced from the permutation crystal structure connection table and the co-crystal structure itself.

As the values suggest the same effect obtained above for the permutation of the SMILES string is reproduced for the permutation of a connection table. As with the SMILES strings, different conformers were generated for the sets resulting in different standard deviations.

Moreover, beginning from ten permutations, the 280 conformers generated are closer to the crystal structure (Average: 1.78 Å, Min: 0.37 Å) than the 280 conformers originating from a single connection table (Average: 2.37 Å, Min: 1.14 Å).

4DFR	One CT	Permuted
min :	1.14	0.37736
max :	3.26	3.26
Average :	2.3736	1.7877
Dev St.	0.3773	1.2805

Table 2-11 RMSD minimum, maximum, average and standard deviation calculation for 4DFR: the table shows the set where 280 conformers were produced using one Connection Table ('One CT' column) and 10 permutation of a single connection table (Permuted column). Unit are expressed in Å.

2.5 Conclusion

In this study, we have shown that permutation of input strings leads to the generation of different numbers of conformers and that some software programs are affected more than others by this depending on their innate algorithm. Depending on the software used the conformers generated from permuted SMILES strings are more diverse than if they were generated from a single representation. The differing amounts of conformations generated are not an inherent fault of the software packages used, rather it is an outcome of the input SMILES string and how the software deals with that SMILES string. When a SMILES string is permutated this possibly has an incidental effect on the order of the torsions which are subsequently used to generate the conformations.

Conformers produced from permuted SMILES strings exhibit a smaller RMSD value, when compared to the X-ray structure, than do conformers generated from the canonical SMILES string. Other work carried out by our group has shown the dramatic effect that this phenomenon has on Enrichment and False Positive rates. Thus, many validations of rigid docking/scoring combination platforms where the same dataset was used, but converted to SMILES strings using different software, may show better or worse performance as a direct result of this phenomenon and not due to the platforms utilised. We suggest that all validations of rigid docking/scoring platforms account for this effect prior to the docking step. Increasing the number of conformers generated, may possibly reduce this effect.

Importantly, all of the software tested compares the generated conformers and discards those that are judged to be too closely related. The ultimate measure of this is an RMSD assessment. For example, one of the comparison steps in Omega involves applying a conformation RMSD exclusion threshold of 0.8 Å. If this value is reduced an increased amount of conformations will be accumulated but they will be more conformationally similar. Conversely, when using permutation of the same SMILES string those conformers would not be discarded and thus some of the conformers obtained might have similar RMSD values.

Finally, we suggest a novel way of generating conformers when using Omega, Rubicon or Corina with their default settings. Permutated input strings,

in conjunction with the canonical SMILES string, could be utilised to more rapidly and comprehensively sample conformational space of a molecule. This method extends to connection table representations of 2D molecular structure such as SDF. PCA analyses clearly illustrate superior conformational sampling arising from use of permuted input. To ensure that the final result does not include any duplicates a quick fingerprint analysis can be carried out as we have shown. We also propose that future versions of software employed in the generation of conformers from SMILES strings could include a switchable flag allowing the canonicalisation of the SMILES strings before conformer generation. The same flag, when off, would allow the users to generate multiple molecular conformations starting from permutations of the same SMILES string.

The effect is also observed using connection tables rather than SMILES strings. To this end, the same method described above may be used in the generation of a diverse unique set of conformers. Clearly, SMILES presents a far more limited set of possible atom/bond orderings unless the use of dot disconnections closure bonds are employed. Thus, SD or any format incorporating a connection table may be used to generate (#atoms!)*(#bonds!) possible orderings.

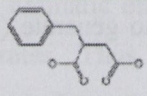
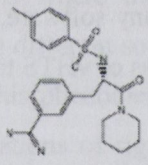
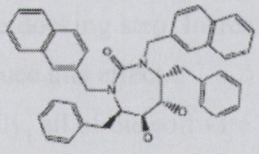
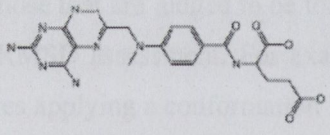
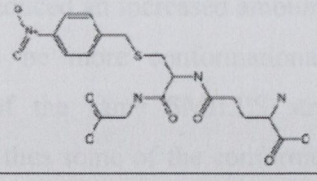
Finally, although not currently implemented in any software, canonicalisation of CTs would also be useful in the validation process of vHTS tools.

2.6 Further applications

Differing from the algorithms taken in consideration previously in this Chapter, in this section we investigate SPE which uses an alternative approach for the generation of conformers known as distance geometry[41,42] (see appendix for a description of the algorithm).

2.6.1 Methods for SPE evaluation

The evaluation was based on a subset of the set listed in Table 2-12, along with the number of SPE trials and the number of conformational boosts employed in each trial. Each ligand was neutralised and the proper chirality was assigned to each chiral centre, as inferred from the PDB 3D co-ordinates. For 1CBX, 4DFR and 1GLQ, these chiral flags were subsequently cleared to study the performance of SPE in the presence and absence of stereochemistry. For 1ETT and 1HVR, the chiral flags remained as shown in Table 2-12.

PDB code	Structure	Rotatable bonds	Trials	Boosts[a]
1CBX		5	1000	0
1ETT		7	5000	1
1HVR		8	5000	1
4DFR		9	10 000	1
1GLQ		13	10 000	1

[a]Number of boosts per trial. For example, the 5000 conformations for 1ETT were obtained by generating 2500 conformations using standard distance bounds, and boosting each of these conformations once to obtain 2500 additional conformations.

Table 2-12 PDB codes, structures, number of rotatable bonds, number of SPE trials and number of boosts for each of the molecules in the data set.

For each ligand, five different SMILES and five different SDF variants were generated, using the Daylight utility permsmi[40] for the former and an in-house program for the latter (see Appendix for more detail). Each variant was used as input to SPE, which was instructed to produce between 1,000 and 10,000 conformations for each one, depending on the size of the molecule (1,000 conformations for each 1CBX variant, 5,000 for each 1ETT and 1HVR variant and 10,000 for each 4DFR and 1GLQ variant). The SPE conformer generator reads both SMILES and SD files and does not necessitate explicit file conversion.

2.6.2 SPE Conformers Generation

SPE is a recently developed technique for finding molecular geometries that satisfy distance constraints in an efficient manner[41,43,44]. The method starts from random initial atomic positions and gradually refines them by repeatedly selecting a pair of atoms at random and updating their respective co-ordinates toward satisfying the corresponding distance boundaries. This procedure is performed repeatedly until a reasonable conformation is obtained.

The actual algorithm involves stochastic refinement of all geometric constraints, both distance and volume. The co-ordinate adjustments differ for each type of constraint, for example, a volume constraint is minimised by moving the positions of the four atoms associated with that constraint in the direction opposite to the gradient of an error function that measures the deviation from the target value. The algorithm also uses a number of parameters such as learning rates and sampling frequencies to control the self-organisation process (see Appendix for more details).

While SPE was originally shown to provide a good sampling of conformational space[43], it was observed that 'extreme' conformations located near the periphery of conformational space were not as likely to be visited and in principle some important conformations could be missed. To remedy this problem, a boosting heuristic was introduced that can be used in conjunction with SPE to bias the search toward more extended or more compact geometries[45]. The method generates increasingly extended (or compact) conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed,

generating a chemically sensible conformation c1. The lower bounds of all atom pairs$\{l_{ij}\}$ are then replaced by the actual interatomic distances$\{d_{ij}\}$in conformation c1 and used along with the unchanged upper bounds$\{u_{ij}\}$to perform a second embedding to generate another conformation, c2. This process is repeated for a prescribed number of iterations, yielding additional conformations. The lower bounds are then restored to their original default values and a new sequence of embeddings is performed using a different random number seed. This process will never yield a set of distance constraints that are impossible to satisfy, because there exists at least one conformation (i.e. the one generated in the preceding iteration) that satisfies them[46].

All stochastic proximity embeddings were carried out on hydrogen-depleted graphs in collaboration with Johnson & Johnson (Dimitris Agrafiotis) and the generation of conformers was also performed entirely by the same using the following parameters:

- $\lambda d = \lambda v = 1$
- $C = 50$
- $\delta\lambda_d = \delta\lambda_v = 0.9/(C - 1)$
- $S = 50 \cdot N$

$$v = \min\left(0.5, 8.0 \times \frac{N_v}{\frac{N(N+1)}{2+8N_v}}\right) \qquad 2\text{-}3$$

where λd and λv are the initial distance and volume learning rates, C is the number of cycles, S is the number of steps, $\delta\lambda d$ and $\delta\lambda v$ are the decrements of the distance and volume learning rates in each cycle. v (equation 3) is the ratio of the volume to distance refinements, N is the number of atoms in the molecule and NV is the total number of volume constraints [44].

2.6.3 Discussion and Results

Being a stochastic method, SPE is not guaranteed to produce the same results when run under different initial conditions. Diverse runs, starting from different random number seeds, are expected to produce slightly different sets of conformations, but these should converge as sampling becomes more exhaustive and the search space becomes saturated. This is illustrated in Figure 2-9, which shows the number of unique conformations discovered at the end of each trial for each SMILES (black) and SDF (red) permuted input for each of the five molecules. When contrasted with multiple independent random runs of only one of the permuted inputs (Figure 2-9E for 4DFR), it becomes clear that the input representation does not have any discernible impact and that the variation in the sampling profiles is within the range expected by the stochastic nature of the method.

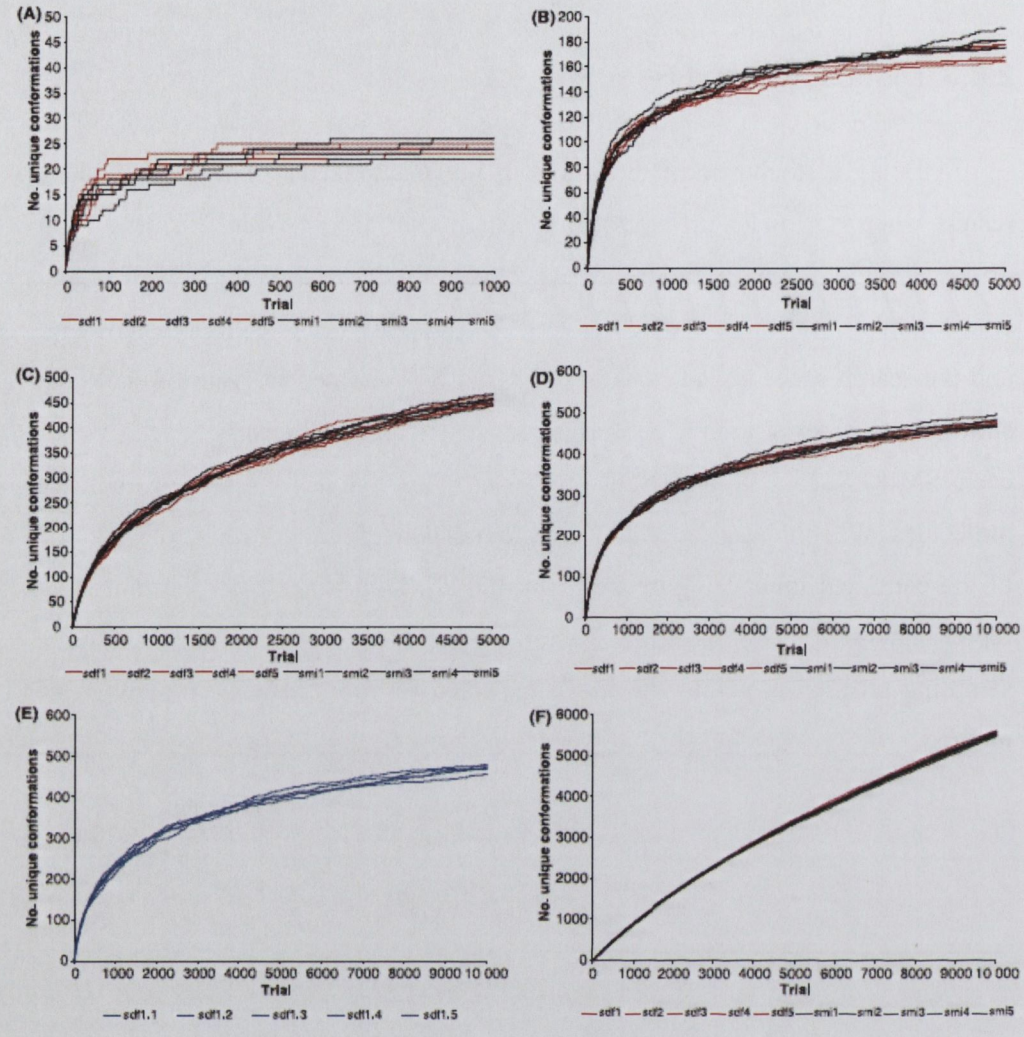


Figure 2-9 Number of unique conformations (within 1 Å RMSD) discovered at the end of each SPE trial for the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (F) 1GLQ. (E) Five independent runs of the first SDF permuted input of 4DFR, each starting from a different random seed.

The ensembles produced by each permuted input were further compared using three different conformational properties:

- the RMSD from the X-ray crystal structure ligand
- the potential energy computed by MMFF94s
- the radius of gyration. The radius of gyration describes the overall spread of the molecule and is defined as the RMSD of the collection of atoms from their common centre of gravity. In other words it measures a conformation's degree of compactness (or extendedness)

In all cases, the conformational ensembles seemed to be drawn from the same probability distribution and the variation appears to be random and not dependent on the input representation.

To obtain a more robust statistical estimate, we compared these distributions using single-factor ANOVA. ANOVA is a well-established technique that attempts to test the null hypothesis that two or more mean values are equal, by comparing the variation observed within each class against the one observed across the entire sample. The key parameters of an ANOVA study include the F-statistic which is compared to the critical value of the F distribution, F_{crit}, and the p-value which represents the smallest level of significance for which the observed sample information becomes significant, provided the null hypothesis is true. The results, which are summarised in Table 2-13, show that the differences in the conformational property distributions of the permuted inputs are statistically insignificant, except for the energy distribution of the conformations of 1GLQ.

	Energy			RMSD			Radius of gyration		
PDB code	F	F_{crit}	p-Value	F	F_{crit}	p-Value	F	F_{crit}	p-Value
1C8X	0.269	1.924	0.982	0.027	1.924	0.999	0.078	1.924	0.999
1ETT	0.124	1.885	0.999	0.214	1.885	0.992	0.189	1.885	0.995
1HVR	0.281	1.883	0.980	0.189	1.883	0.995	0.289	1.883	0.980
4UHK	0.834	1.893	0.532	0.316	1.893	0.952	0.185	1.893	0.999
1GLQ	13.337	1.880	1.45E-21	0.346	1.880	0.960	0.499	1.880	0.876

Table 2-13: ANOVA of the conformational properties of the ensembles produced by the permuted inputs for each molecule under investigation (when F is less than Fcrit, the differences are considered to be statistically insignificant)

Previously in this Chapter we showed that some systematic conformational search methods, such as Omega, can be seriously impacted by minor rearrangements of the connection table and recommended the use of multiple permuted inputs to improve their performance and enhance their sampling capacity. While this approach only requires a way to generate permuted connection tables and therefore can be used with any conformational search program to circumvent its intrinsic bias, it remains symptomatic. It does not guarantee that the additional permutations will lead to an exhaustive sampling of conformational space and that important regions of conformational space will not be missed. The preceding analysis demonstrates that stochastic proximity embedding does not exhibit such bias and that its sampling capacity is limited solely by the geometric constraints encoded in the rule base.

2.7 Reference

1. Hou T, Xu X. Recent development and application of virtual screening in drug discovery: an overview. Curr Pharm Des 2004;10(9):1011-1033.
2. Liao C, Liu B, Shi L, Zhou J, Lu XP. Construction of a virtual combinatorial library using SMILES strings to discover potential structure-diverse PPAR modulators. Eur J Med Chem 2005;40(7):632-640.
3. Bringmann B. KA. Frequent SMILES. Lernen, Wissensentdeckung und Adaptivität 2004;Workshop GI Fachgruppe Maschinelles Lernen, part of LWA 2004.
4. Weininger D. SMILES: A Chemical Language and Information System J Chem Inf Comput 1988;28:31-36.
5. Daylight Chemical Informations Systems Inc. (URL: http://www.daylight.com).
6. Andrew J.S. Knox MJM, Giorgio Carta, David G. Lloyd2. Considerations in compound database preparation - 'hidden' impact on virtual screening results. J Chem Inf Model 2005;ASAP Web Release Date: 12-Oct-2005.
7. Derwent World Drug Index. (URL: http://thomsonderwent.com/products/lr/wdi).
8. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46(1-3):3-26.
9. FILTER. distributed by Openeye Scientific Software
10. CONVERT. developed and distributed by Molecular Networks GmbH (URL: http://www.mol-net.de).
11. CORINA 3.6. distributed by Molecular Networks GmbH
12. OMEGA 1.8.1. distributed by Openeye Scientific Software
13. RUBICON. distributed by Daylight Chemical Informations Systems Inc.
14. CATALYST. developed and distributed by Accelrys, 9685 North Scanton Road, San Diego, CA 92121, USA (URL:http://wwwaccelryscom).
15. FRED (version 2.0.1). developed and distributed by Openeye Scientific Software.(URL:http://www.eyesopen.com).
16. Vigers GP, Rizzi JP. Multiple active site corrections for docking and virtual screening. J Med Chem 2004;47(1):80-89.
17. Kauppi B, Jakob C, Farnegardh M, Yang J, Ahola H, Alarcon M, Calles K, Engstrom O, Harlan J, Muchmore S, Ramqvist AK, Thorell S, Ohman L, Greer J, Gustafsson JA, Carlstedt-Duke J, Carlquist M. The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. J Biol Chem 2003;278(25):22748-22754.
18. Cronet P, Petersen JF, Folmer R, Blomberg N, Sjoblom K, Karlsson U, Lindstedt EL, Bamberg K. Structure of the PPARalpha and -gamma ligand binding domain in complex with AZ 242; ligand selectivity and agonist activation in the PPAR family. Structure (Camb) 2001;9(8):699-706.

19. Molecular Operating Environment (MOE). developed and distributed by Chemical Computing Group,. (http://www.chemcomp.com).
20. Babel v2.0A3. distributed by Openeye Scientific Software
21. Chemsketch. v8.17. wwwacdlabscom.
22. Dalby A. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. J Chem Inf Comput Sci 1992;32:244-255.
23. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261(3):470-489.
24. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267(3):727-748.
25. Cambridge Structural Database.http://www.ccdc.cam.ac.uk/.
26. SL Mayo BO, WA Goddard III. DREIDING: A Generic force field for molecular simulation. JPhysChem 1990;94:8897-8909.
27. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 1996;17:490–519.
28. Halgren TA. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 1996;17:520–552.
29. Halgren TA. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem 1996;17:553–586.
30. Halgren TA. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem 1996;17:616-641.
31. Halgren TA, Nachbar RB. Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem 1996;17:587–615.
32. Catalyst v4.9.1. www.accelrys.com.
33. AD MacKerell BN, N Foloppe. Development and current status lf the CHARMM force field for nucleic acids. Byopoymers 2000;56(4):257-265.
34. D. F. Shanno, Phua KH. ACM Transactions on Mathematical Software 1980(6):618-622
35. Bostrom J, Greenwood JR, Gottfries J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model 2003;21(5):449-462.
36. Bostrom J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comput Aided Mol Des 2001;15(12):1137-1152.
37. Ivanciuc O. Graph Theory in Chemistry. Handbook of Chemoinformatic Edited by J Gasteiger 2003;Vol. 1:103-138.
38. Kudo Y, Sasaki S. The Connectivity Stack, a New Format for Representation of Organic Chemical Structures. Journal of Chemical Documentation 1974;14(4):200-202.
39. Babel 1.100.2. distributed by Openeye Scientific Software.
40. oechem, rmsd. distributed by Openeye Scientific Software.
41. Agrafiotis DK, Xu H. A self-organizing principle for learning nonlinear manifolds. Proc Natl Acad Sci U S A 2002;99(25):15869-15872.

42. Spellmeyer DC, Wong AK, Bower MJ, Blaney JM. Conformational analysis using distance geometry methods. J Mol Graph Model 1997;15(1):18-36.
43. Agrafiotis DK. Stochastic proximity embedding. J Comput Chem 2003;24(10):1215-1221.
44. Xu H, Izrailev S, Agrafiotis DK. Conformational sampling by self-organization. J Chem Inf Comput Sci 2003;43(4):1186-1191.
45. Izrailev S, Zhu F, Agrafiotis DK. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. J Comput Chem 2006;27(16):1962-1969.
46. Agrafiotis DK, Gibbs AC, Zhu F, Izrailev S, Martin E. Conformational sampling of bioactive molecules: a comparative study. J Chem Inf Model 2007;47(3):1067-1086.

2-Ap. Appendix to chapter 2

Scripts

Below is listed the code of *elaborate_sdf.pl*. This script was used in Chapter 2 for the generation of permutation of the connection table from an sdf file. Figure 2-A and Figure 2-B depicted the algorithms used in the script. The comments and details of the script are highlighted red.

elaborate_sdf.pl

```
#!/usr/bin/perl
#
# © Copyright Trinity College Dublin.
#
# usage : perl elaborate_sdf.pl infile.sdf
# Description:
# it generates permutation in the connection table of an sdf file
# Author Giorgio Carta, 9/02/2006

# include "strict" library

use strict;

# open zero argument in the command line

open(INFILE,$ARGV[0]) or die "cannot open $! \n";

# set environment variables

my $intro;
my @nodes;
my @edges;
my $ending;
my $count_mol;
my $flag_intro;
my $flag_edges = "true";
my $counter_intro = 1;
my $counter_nodes = 0;

# start loop in the file

foreach(<INFILE>)
{

        # save first 4 lines of sdf file

        if ($counter_intro <= 4) { $intro = $intro.$_; $counter_intro++;next; }

        # split and isolates sdf fields

        my ($trash,$xa,$ya,$za,$ele,$trash,$trash,$trash,$trash,$trash) = split(/\s+/,$_);

        if (($ele =~ /\S+/)&&(!($ele =~ /\d+/))) {        # identifies atom element and skip what is not
```

```
                    my $string = $_;
                    chomp($string);              # deletes new line character

    my @line = ($string,++$counter_nodes);  # gets the nodes and add a ranking number
                    push(@nodes,\@line);                    # creates the array of address
                    next;                                   # bo back to loop, do not go forward
                    }

                                                            # end when molecule ends

    if ($_ =~ /END/) { $flag_edges = "false"; $ending = $ending.$_; next;}

    if ($flag_edges eq "true") {
                    my $string = $_;
                    chomp($string);              # deletes new line character

                                                            #split the connection string 1 2->1 & 2

                    my @splitted = split(/\s+/,$_);
                    shift(@splitted);            # the first element is skipped!!!
                    push(@edges,\@splitted);     # creates the array of addresses
                    next;}

    $ending = $ending.$_;

                                        # everything is packed now: send it to subroutines!

    if ($_ =~ /\$\$\$\$/) {

    print "Molecule n. ".(++$count_mol)."\n";

                            # calls for routine engine (intro file, nodes, edges, ending)


    my $confs = engine($intro,\@nodes,\@edges,$ending);

                            # routine returns number of permutation generated


    print "\nPermutations generated : $confs \n";     # output number of
                                                                    # permutation generated


                                                                    # reset the counter and
                                                                    # flag to default

    $counter_intro = 1;
    $flag_edges = "true";
    $intro = ""; $ending = "";
    $counter_nodes = 0;
    @nodes = ();
    @edges = ();
    }

}


sub engine()                                # sub engine($intro,\@nodes,\@edges,$ending);
{


my $intro = shift();
my $nodes = shift();
my $conn_table_old = shift();
my $endin = shift();
```

```
open (OUT, ">>outfile.sdf");                          # open for append the OUT outfile

# local variables

my $tag = "true";
my $max_count = 0;
my %hash_id;
my $iteration;                                        # counter for the iteration
my $max_id = 10;                                      # max number of permutation defined by the user

                                                      # begin the permutaion process

while ($tag == "true") {
my @permutation = random_perm(@$nodes);               # permutate @nodes, which contains the
                                                      # original connection table
my $size = @nodes;
my @conn_table_new = create_new_table($conn_table_old,\@permutation);

my $id = create_matrix($size,@conn_table_new);        # create a matrix with the new c.t and
                                                      # return "Kudo e Sasaki" stack

                                                      # user max of permutation per connection table
if (!(defined($hash_id{$id}))) {                      # check if the id had already been defined in the
                                                      # %hash_id

          #display(@nodes);                           # verbose mode : originary structure
          #display(@conn_table_old);                  # verbose mode

          print OUT $intro;                           # saving mode : molecule introduction
          display(@permutation);                      # saving mode : permutated structure
          display1(@conn_table_new);                  # saving mode : connection table
          print OUT $ending;                          # saving mode : end of the molecule
          #print "id : $id \n";                       # debug       : displai id in the STDOUT
          print ".";
          $hash_id{$id} = defined;                    # difenes id in the %hash_id
          $max_count++;                               # count the permutation generated
          $iteration = 0;                             # reset to zero when finds a new valid permutation
          if ($max_count == $max_id)
                    {$tag = "false";
                     last;}                           # ends if goes over the
                                                      # thershold user set by the user
          }
else { $iteration++;
          if ($iteration >= 100) { last;}             # after 100 times it doesn`t find anything
                                                      # different assumes that no more
                                                      # permuts are available
          next;}                                      # skip to next if the id already defined


          }                                           # close while ($tag = "true")

return $max_count;                                    # return the number of permutation generated

}                                                     # close engine sub routine



sub random_perm{                                      # generates random permutation of an array
                                                      # receives as argoument an array. Here the array
my @r;                                                # is an array of pointers.
push(@r,splice(@_,int(rand()*@_),1)) while(@_);       # Engine.
return @r;                                            # Return a permutation of the array reveived
}
```

```
sub display()                                          # save in OUT the nodes
{
foreach (@_) {
          print OUT $$_[0];
          print OUT "\n";
}


}

sub display1()                                         # save in OUT the connection table formatted
{
foreach (@_) {

          my  $pipi  =  sprintf  "%3d%3d%3d%3d%3d%3d%3d",$$_[0],$$_[1],$$_[2],$$_[3],$$_[4],$$_[5],$$_[0];  #  format
output

          print OUT $pipi;
          print OUT "\n";
}


}


sub create_new_table()                                  # create the new connection table
{

my $address_ct = shift();                               # address of address of the connection table;
my $address_nodes = shift();                            # address of address of nodes;
my %hash;
my $count2 = 1;
my @conn_table_new;                                     # array of address for the new connection table

foreach (@$address_nodes)
          {
                    #print "key :$$_[1] >> value $count2 \n"; # debug : print STOUT the
                                                        # correspondance between old and
                                                        # new c.t.
                    $hash{$$_[1]} = $count2;            # old value : key --- new value :
                                                        #  values
                    $count2++;
          }


foreach (@$address_ct)
          {


          my @line_address = sort {$a <=> $b} ($hash{$$_[0]},$hash{$$_[1]}); # sort numerically
                    my @temp = @$_;
                    splice(@temp,0,2);                  # splice the first two element in the array
                                                        # containing the old connection table
                    push (@line_address,@temp);         # Push in the new array what is left
                                                        # from the old connection table
                    push (@conn_table_new,\@line_address);  # ex. double bond, etc.
          }

          return(@conn_table_new);
}


sub create_matrix                                       # create half matrix for the calculation of the ID
of                                                      # the conn. table
{
```

```
        my $size = shift();                          # pick the size of the matrix equal to the number
        my @matrix;                                  # nodes it s going to contain half the matrix itself
        my $mio;
        my $i;
        my $id;
        my $o;
        my $counter;

        foreach(@_)
        {
        $matrix[$$_[0]][$$_[1]] = 1;                 # assign the 1 values to index i j in the matrix
        }



        foreach(2 ..$size){
                $mio = $_;
                for($i = 0; $i <= $counter; $i++)    # calculate the ID from half the matrix
                        {
                        $o = $i + 1;
                                                     # kudo e Sasaki method

                        if (!($matrix[$o][$mio] == 1))  {$matrix[$o][$mio] = 0; }

                        $id = $id.$matrix[$o][$mio];
                        #print $matrix[$o][$mio];    # debug : print the 2D matrix :
                                                     # activate this line
                        }
                        #print "\n";                 # and this one too.
                $counter++;
        }


return ($id);                                        # identification of the
                                                     #  connectivity stack


}                                                    # close create_matrix
```

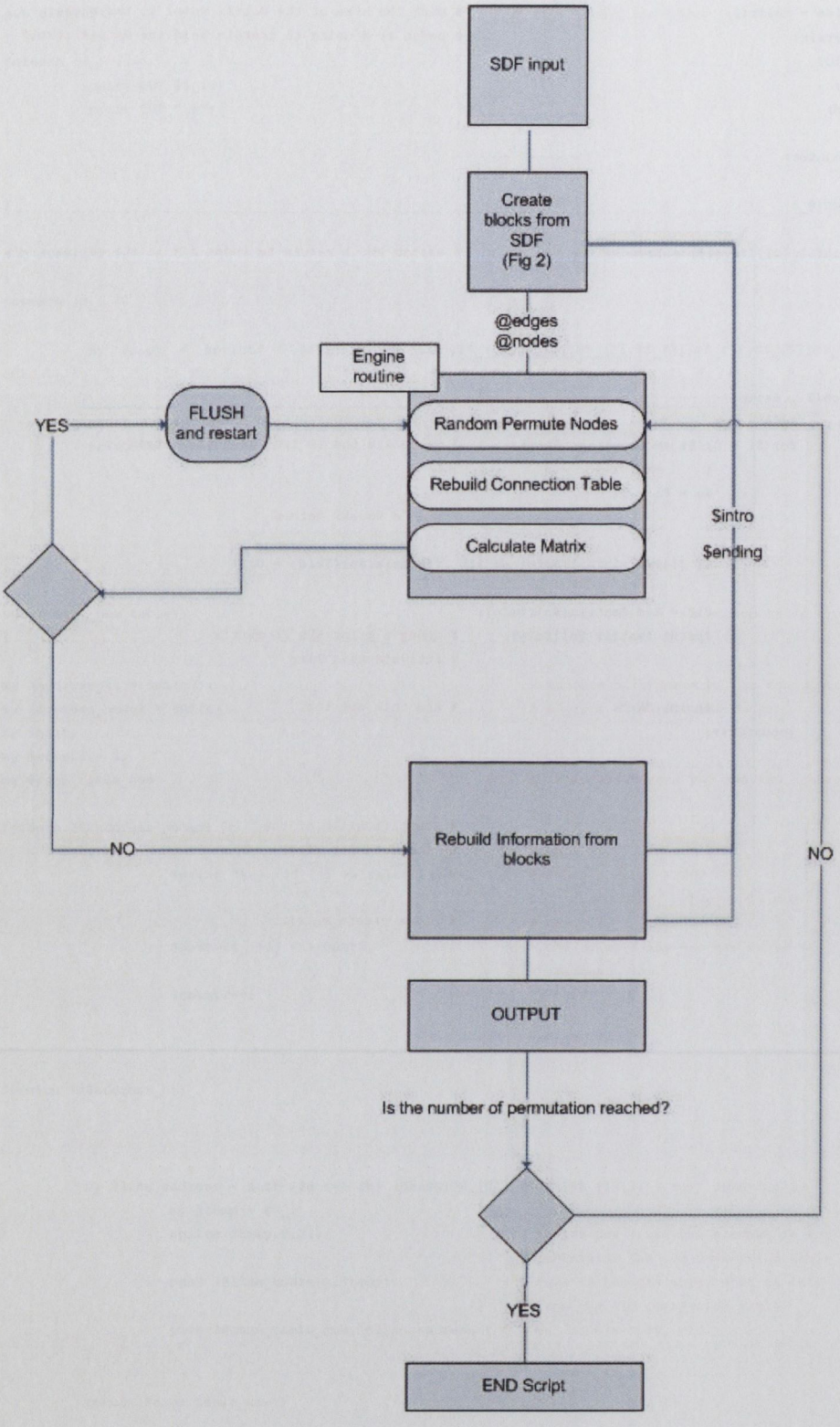


Figure 2-A:Algorithm used to permutate the sdf connection table in elaborate_sdf.pl

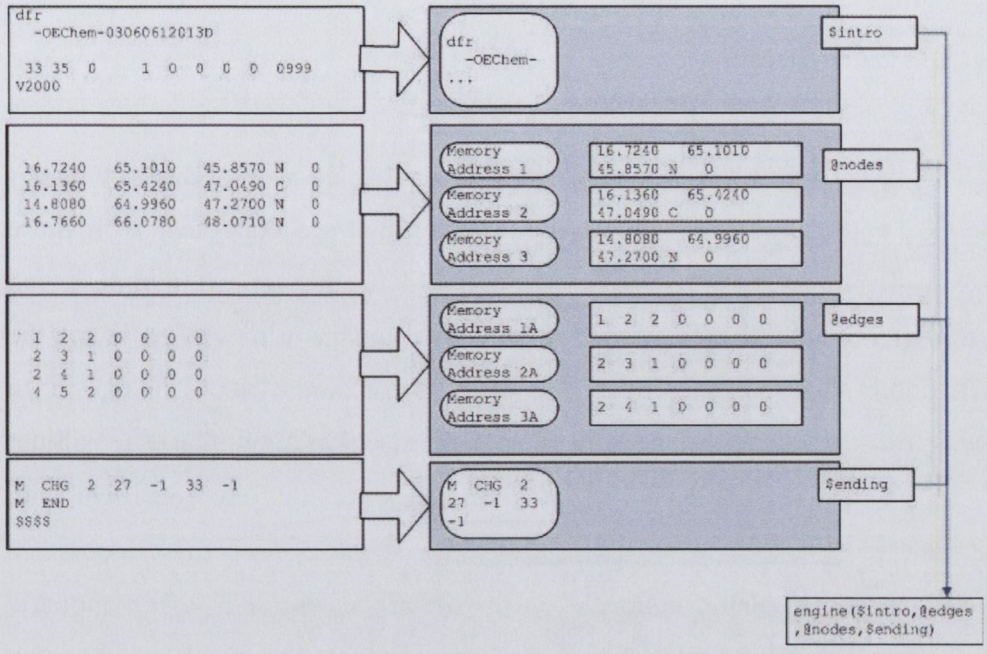


Figure 2-B: Scheme illustrating how the sdf file is divided in blocks by elaborate_sdf.pl

Distance Geometry

Distance geometry is a general method for converting a set of distance ranges (or "bounds") into a set of Cartesian coordinates that are consistent with those bounds. Any molecular system can be described as the set of minimum and maximum inter-atomic distances between all pairs of atoms in the molecule and the complete conformational space of the molecule is contained within this space. In distance geometry, a matrix is defined as the set of minimum and maximum distances, and then used to create a series of conformers that are consistent with those distances.

A three dimensional generic molecule can be described with a N × N symmetric matrix of inter atomic distances whose diagonal elements are equal to 0. In such matrix, the elements (i,j) contains the distance between i and atoms j. For clarity we define as $u_{i,j}$ and $l_{i,j}$ respectively the upper and lower distance of the bond between atom i and atom j or in other words, the maximum and the minimum distance between these atoms.

Distance geometry methods involve four basic steps: (i) generation of inter-atomic distance bounds, (ii) assignment of a random value to each distance within the respective bounds, (iii) conversion of the resulting distance matrix into a starting set of Cartesian co-ordinates (embedding), and (iv) refinement of the co-ordinates through minimization of an error function measuring the violation of the input constraints. In order to illustrate how the sampling of conformational space is achieved by distance geometric methods, a simple five–carbon fragment molecule will be considered throughout appendix B (Figure 2-C)

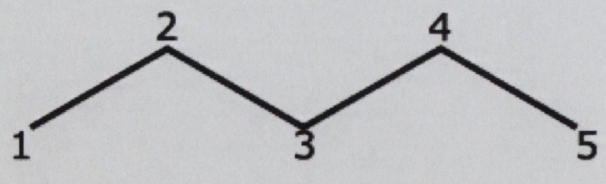


Figure 2-C Five-carbon fragment. The angle between each bond is assumed to be 120° and the bond C-C to have a length of 1.3 Å (hydrogens are not displayed).

Generation of Interatomic distance

Assuming the structure in Figure 2-C as the starting structure we can generate the upper and lower distance Table 2-A. The lower bounds occupy the left-bottom half of the matrix and the upper bounds the upper-right half.

	Atom				
Atom	1	2	3	4	5
1	0	1.3	2.2517	3.4395	99.0
2	1.3	0	1.3	2.2517	3.4395
3	2.2517	1.3	0	1.3	2.2517
4	2.6	2.2517	1.3	0	1.3
5	2.8	2.6	2.2517	1.3	0

Table 2-A Initial bounds matrix. Cyan cells indicate the lower limits. Orange cells indicate the upper limits. (all the units are in Å).

As is shown in Table 2-A the covalent bond length is assumed equal to 1.3 Å, which is a standard C-C bond length. The distance of the non-bonded relationship between carbon 1 and 3 is restricted by geometrical constraints and it depends on the central angle α and the length of the two bonds (Figure 2-D).

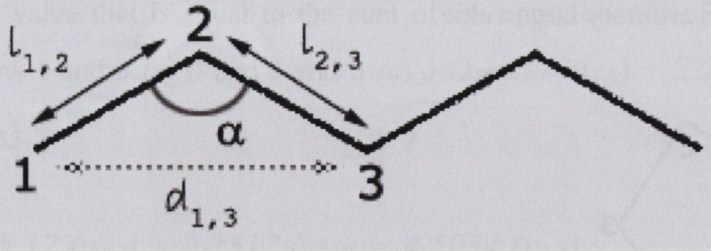


Figure 2-D Distance d1,3 depends by length of bonds l1,2 and l2,3 and the angle α.

Given the bond length is equal to 1.3 Å and the angle equal to 120°, with a simple trigonometric function we obtain the distance between atoms 1,3 equal to 2.2517 Å. Furthermore the distance between two atoms with a 1,4 relationship is restrained as

well as the 1,3 relationship and it is correlated to the dihedral angle of the central bond (from 0° to 180°) and the lengths of the internal bonds (Figure 2-E).

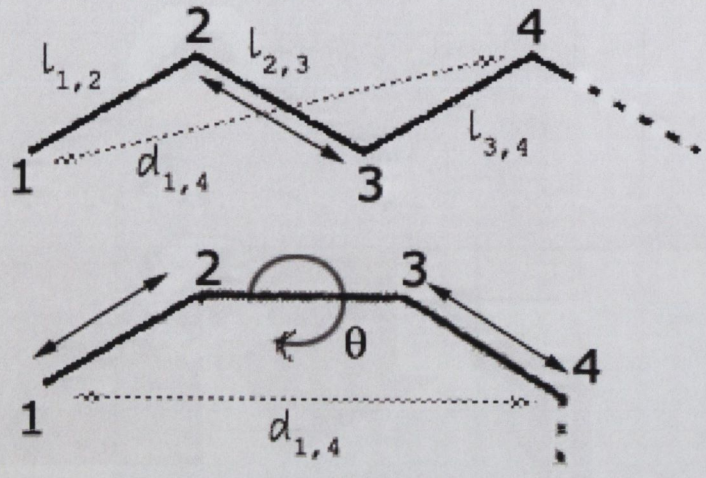


Figure 2-E The length of distance between atoms 1 and 4 will depends on the lengths between atoms 1,2,3,4 (ln) and angle θ.

On the basis of geometrical arguments the distance values in a 1,4 relationship can therefore vary between 2.6 Å (the lower limit) and 3.4395 Å (the upper limit). It is not easy to determine limits on other inter atomic distances (between atoms in a 1,n relationship where n > 4) but it is usual to require that the lower limit never approach the sum of the vdW radii of the two atoms (Figure 2-F). The upper bond is then usually assigned an arbitrary large value.

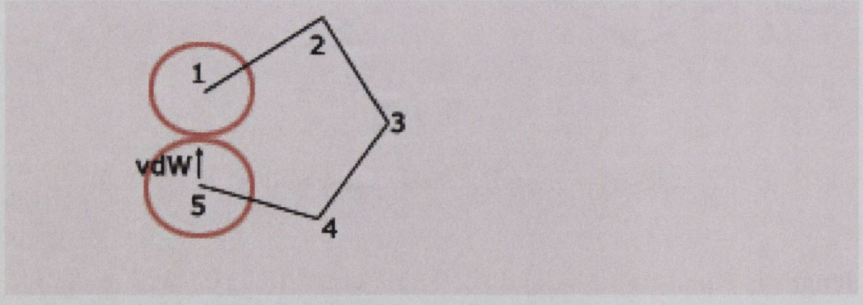


Figure 2-F Distance between atoms 1,5 cannot be lower than the sum of the vdW radii of the atoms 1,5 (in red).

As we can see in matrix 1, the lower distance is thus set to 2.6 Å (vdW radius 1.3 × 2) and the upper distance to 99.0 Å. During this phase it is also possible to include different constraints in order to restrict the inter-atomic distance of double or triple bonds, cyclic structure etc.

A procedure called *triangle smoothing* (TS) is then subsequently applied to refine the initial distance matrix. TS uses two intuitive trigonometric restrictions on groups of three atoms. Considering a three atoms bonded system with atom A, B and C, the first restriction postulates that the upper distance between two atoms A and C cannot be greater than the sum of the maximum values of the distances AB and BC (equation 1).

$$u_{AC} \leq u_{AB} + u_{BC} \qquad \text{2-A}$$

The second restriction is that the minimum value of the AC distance cannot be less than the difference between the lower bound AB and the upper bound BC.

$$l_{AC} \leq l_{AB} - u_{BC} \qquad \textit{2-B}$$

where u and l represent again respectively the upper and the lower distances. Applying those restrictions to the matrix in Table 2-A only one distance is changed to a value that is equal to the sum of the upper bounds of the distances between atoms 1 and 3 $(u_{1,3})$ and 3 and 5 $(u_{3,5})$ (2.25.17 $(u_{1,3})$ + 2.2517 $(u_{3,5})$ = 4,5034 $(u_{1,5})$ *2-C*).

$$2.25.17\ (u_{1,3}) + 2.2517\ (u_{3,5}) = 4{,}5034\ (u_{1,5}) \qquad \text{2-C}$$

The resulting smoothed matrix is show in Table 2-B.

	Atom				
Atom	1	2	3	4	5
1	0	1.3	2.2517	3.4395	4.5033
2	1.3	0	1.3	2.2517	3.4395
3	2.2517	1.3	0	1.3	2.2517
4	2.6	2.2517	1.3	0	1.3
5	2.8	2.6	2.2517	1.3	0

Table 2-B Resulting distance matrix after TS of the matrix in Table 1. (all the units are in Å).

The geometrical explanation of triangle smoothing applied to our five atom molecule can be depicted as in Figure 2-G.

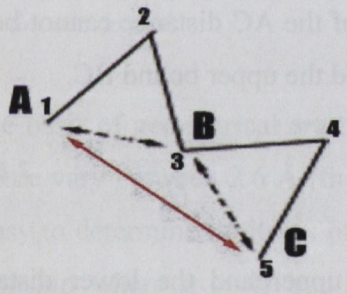


Figure 2-G The maximum distance between atoms 1,5 (labeled A and C) or the upper limits (1,5) is equal to the sum of AC and BC.

Assignment of a random values and embedding

In order to generate new valid distances that at the end will represent a new conformer of the molecule, random distances are assigned between the lower and the upper bounds to give the following distance matrix (Table 2-C).

	Atom				
Atom	1	2	3	4	5
1	0	1.3	2.2517	3.11	3.42
2		0	1.3	2.2517	2.85
3			0	1.3	2.2517
4				0	1.3
5					0

Table 2-C Randomly assigned distances between the upper and the lower bounds indicated in Table 2-B. (all the units are in Å).

It is now possible to proceed to the generation of conformers by calculating three dimensional coordinates that satisfy the distances indicated in the matrix in Table 2-C.

In more details embedding is the process of arranging a set of objects in a space with a particular number of dimensions so as to reproduce the observed distances between them. This problem is also known as multidimensional scaling (MDS) or nonlinear mapping (NLM) and has two main applications: (i) reducing dimensionality of high-dimensional observations and (ii) producing coordinate vectors from data supplied in the form of proximities.

The conversion into Cartesian coordinates is performed through a series of matrix operation. A metric matrix is first calculated and its *eigenvalues* and *eigenvectors* used to obtain a close 3D representation to the proximities indicated in the distance matrix in Table 2-C. An example of the resulting coordinates and distances after embedding is given in Table 2-D, Table 2-E and Figure 2-H.

Atom	X coordinate	Y coordinate	Z coordinate
1	1.77	0.601	-0.218
2	1.014	-0.244	0.410
3	-0.001	-0.757	-0188
4	-1.166	-0.379	-0.079
5	-1.623	0.799	0.075

Table 2-D XYZ coordinates obtained after the embedding of matrix in table 3.

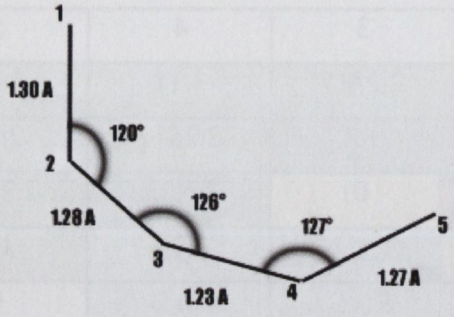


Figure 2-H Conformation of the five-carbon fragment generated by distance geometry

	Atom				
Atom	1	2	3	4	5
1	0	1.29	2.24	3.10	3.42
2		0	1.29	2.24	2.85
3			0	1.23	2.25
4				0	1.25
5					0

Table 2-E Interatomic distances of conformation in Figure 2-H (all the units are in Å).

It is important to note that the distance in this conformation does not equal the original randomly chosen distance, nor do they all lie between the upper and lower values indicated in the matrix in Table 2-B.

This is because it may be necessary to use more than just three dimensions to find a conformation that satisfies the initial distance matrix. Moreover, as the triangle smoothing is applied to the bounds matrix of upper and lower bounds, it is also possible at this stage to have conformers which do not satisfy the triangular inequalities.

Refinement

To improve the embedding a refinement step, is required so that the conformation better satisfies the initial distance bounds. The quality of the projection can be determined using a sum-of-square error function (Kruskal's stress):

$$S = \sqrt{\frac{\sum_{i<j}\left(d_{ij} - r_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}} \qquad \text{2-D}$$

Where S is the Kruskal's stress, d_{ij} represents the Euclidean distance between atom i and j and r_{ij} the proximity of the same atoms indicated in a distance matrix.

In the refinement step the actual embedding is carried out in an iterative fashion by: (i) generating an initial set of coordinates $\boldsymbol{x_i}$,(ii) computing the distance $d_{i,j}$, (iii) finding a new set of coordinates $\boldsymbol{x_i}$ using a steepest descent algorithm and (iv) repeating steps two and three until the stress function (equation 2-D) falls below some predefined threshold. Stochastic proximity embedding (SPE) uses a self-organizing scheme that attempts to bring each individual stress ($d_{i,j} - r_{i,j}$) rapidly to zero. This method starts with an initial configuration and iteratively refines it by repeatedly selecting two points at random, and adjusting their coordinates so that their Euclidean distance on the map $d_{i,j}$ matches more closely their corresponding distance or proximity in the distance matrix $r_{i,j}$.

The correction is proportional to the disparity $\lambda = \frac{\left|r_{ij} - d_{ij}\right|}{d_{ij}}$, where λ is a learning rate parameter that decreases during the course of the iteration to avoid oscillatory behavior. The detailed algorithm is as follows:

1. Initialize the coordinate $\boldsymbol{x_i}$. Select an initial learning rate λ.
2. Select a pair of point, i and j at random and compute their distance $d_{i,j} = ||\boldsymbol{x_i} - \boldsymbol{x_j}||$. If $d_{i,j} \neq r_{i,j}$ update the coordinates $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ by:

$$x_i \leftarrow x_i + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} \left(x_i - x_j \right) \qquad 2\text{-}E$$

and

$$x_j \leftarrow x_j + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} \left(x_j - x_i \right) \qquad 2\text{-}F$$

where ε is a small number to avoid division by zero.

3. Repeat (2) for a prescribed number of steps S
4. Decrease the learning rate λ by a prescribed decrement δλ.
5. Repeat steps from (2) to (4) for a prescribed number of cycles *C*.

The SPE conformational algorithm generates conformations that satisfies a set of distance constraints as specified by the lower and upper bounds $l_{i,j}$ and $u_{i,j}$.

With a heuristic method it is then possible to bias the extendedness of the generated conformations. For example, by increasing the values of the lower bounds, one could enforce an increase of the inter-atomic distances, thereby generating more extended conformations. Such artificial modification, however, may be problematic, because when the distance bounds are tightened, the existence of a geometry satisfying these constraints is not always guaranteed.

The method generates increasingly extended (or compact) conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed, generating a chemically sensible conformation c1. The lower bounds of all atom pairs$\{l_{ij}\}$are then replaced by the actual inter-atomic distances$\{d_{ij}\}$in conformation c1, and used along with the unchanged upper bounds$\{u_{ij}\}$to perform a second embedding to generate another conformation, c2. This process is repeated for a prescribed number of iterations, yielding additional conformations. The lower bounds are then restored to their original default values, and a new sequence of embeddings is performed using a different random number seed. This process will never yield a set of distance

constraints that are impossible to satisfy, because there exist at least one conformation (i.e. the one generated in the preceding iteration) that satisfies them. An analogous procedure can be used to generate increasingly compact conformations.

As an example, the resulting upper and lower distance matrix obtained by substituting the lower limit of matrix in Table 2-B with the actual values in Table 2-E is showed in Table 2-F.

	Atom				
Atom	1	2	3	4	5
1	0	1.3	2.2517	3.4395	4.5033
2	1.2	0	1.3	2.2517	3.4395
3	2.24	1.29	0	1.3	2.2517
4	3.1	2.24	1.23	0	1.3
5	3.42	2.85	2.25	1.25	0

Table 2-F Updated matrix of table 2 with the actual distance found after embedding indicated in table 5

As it shows Table 2-F, the lower limit is moving toward the upper limit, in other words the random values selected between the lower and upper bounds will be biased toward the upper limits, atoms are therefore diverging from each other and an expanded conformation is thus generated.

During the optimization of the structure against the distance constraints is usual to incorporate chiral constraints. These are used to ensure that the final conformation is the desired stereoisomers. Chiral constraints are usually incorporated as a chiral volume, calculated as a scalar triple product. For example, to maintain the correct stereochemistry about the tetrahedral atom number 4 in Figure 2-I, the following scalar triple produce in equation 2-G must be positive:

$$(v_1 - v_4) \cdot [(v_2 - v_4) \times (v_3 - v_4)] > 0 \qquad 2\text{-}G$$

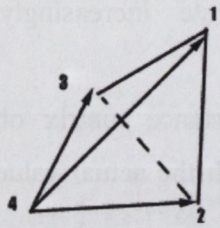


Figure 2-I The stereochemistry about tetrahedral atoms can be maintained with an appropriate chiral constraint. 1,2,3,4 represent vectors v1 ,v2 ,v3 ,v4.

Finally, many other enhancements can be included such as the chiral constraints to improve the quality of the conformers generated. Few of them are described in this chapter. Others could be included to avoid eclipsed conformations by setting a specific values of lower limit of 1,4 interactions or aromatic, amide and ketones groups can be forced to apply a planar structure by an appropriate choice of distance bounds and chiral constraints.

3. Application of vHTS technology: discovery of novel PPARγ ligands

3.1 Introduction

This Chapter details the development of parameters and techniques to carry out a successful virtual screen for ligands that bind to PPARγ. The protocol is fully integrated within Pipeline Pilot (PP), a common platform for the processing of information in vHTS. Details are given about the protocols and the rationale behind the screening itself. The general details and the advantages of using VS for the discovery of novel actives molecules was already described in Chapter 1 and some of the techniques used in this chapter were discussed in Chapter 2.

A biological introduction to PPARγ and its superfamily, the nuclear receptors, is also given to offer a biological background on the target as an ideal link between the rational discovery of a novel molecule and its potential use in a therapy.

The outcome of this chapter is the identification of a novel putative agonist for PPARγ receptor and a structural activity relationship within similar classes of known agonists. A biological *in vitro* assay was performed to determine the activity of the novel compound identified.

Identification of "decoration" sites around the novel active for lead-optimisation is also illustrated at the end of the discussion.

3.2 Nuclear Receptor

Nuclear Receptors (NRs) are transcription factors essential in embryonic development, maintenance of differentiated cellular phenotypes, metabolism and cell death[1]. This set of receptors act as master switches to initiate, regulate and terminate the expression of specific genes or genetic programs[1-3]. Dysfunction of nuclear receptor signalling leads to proliferative, reproductive and metabolic diseases such as cancer, infertility, obesity and diabetes[4]. NRs, such as estrogen and retinoic acid receptors (ER and RAR), form a super family of ligand-inducible transcription factors directly interacting with DNA[5]. They integrate the signals of a number of different and independent pathways and, hence, they are involved in the regulation of many fundamental processes[3].

To date, 49 NRs have been identified[6] throughout the animal kingdom, ranking from nematodes to humans (24 NRs[1]). The NR superfamily contains many "liganded" receptors, but also "orphan" receptors for which no ligand has yet been identified.

Almost two decades have gone by since the cloning of the first NR for a steroid hormone. Since then, other nuclear hormone receptors were rapidly cloned and their target sequences of DNA identified. Our knowledge of the regulation of gene expression by NR has grown enormously over the last few years, mainly due to the realisation that not only the interaction of the receptors with DNA was important for transcriptional responses, but also that many co-regulators were crucial in transmitting the hormonal signal to the transcriptional machinery. Additionally, crystal structures of the ligand binding domain of NRs have been solved, and this allowed the definition of the structural basis of their transcriptional functions.

Endogenous ligands of NRs origin from biosynthetic pathways that originate from isoprenoid moieties and involve enzymes of the P450 family. One key intermediate is all-*trans*-retinal for retinoic acid derivatives, which are the ligands of RAR and Retinoid X Receptor (RXR). Other key intermediates are squalene, cholesterol and pregnenolone for vitamin D (Vitamin D receptor: VDR), corticoids (glucocorticoids and mineralcorticoids receptor: GR and MR), and reproductive

steroids for estrogen, progesterone and androgen receptors (PR and AR). Thyroid Hormones (Thyroid hormone Receptor TR), leukotriene, and prostaglandins that bind and activate the Peroxisome PPAR represent exceptions to the present status of knowledge not being originated from the same biosynthetic pathway.

A large number of synthetic drugs, agonists and antagonists, are also currently used for targeting different disorders: tamoxifen (antagonist for ER) is used in the therapy of breast cancer, thiazolidinediones (TZD) are used for type II diabetes or dexomethasone (agonist for GR) for the treatment of inflammatory diseases.

3.3 PPAR γ

The world obesity drug market has been predicted to reach $3.7 billion by 2008[7-9]. Dietary fat intake is an environmental factor that affects many aspects of human life today[10]. A fat rich, sedentary lifestyle represents an increased risk for the development of diseases such as obesity, diabetes type II, and cardiovascular complication or nonalcoholic steatohepatitis (NASH). All these syndromes aggregated together are named generically syndrome X, which is the principal cause of morbidity in the developed world.

The Diabetes Federation Ireland estimates a prevalence of diabetes as high as 6% of the population. Figures indicate that up to 200,000 people have been diagnosed in Ireland. Another 200,000 are unaware that they have the condition, with type II diabetes accounting for almost 90% of all cases. There has been a significant increase in overweight 13 years old children of both sexes in Ireland[11]. The prevalence of diabetes type II in this country is rapidly increasing due to factors such as, an aging population, a rise in sedentary lifestyle, obesity and inappropriate diet. The number of Irish people suffering from diabetes is expected to double by 2010[11,12]. As well as lifestyle, other factors may play an essential role. Genes are also as determinant and important as the environment factors in originating obesity and syndrome X.

The determination to develop new treatments for obesity is therefore immense. Many of the problems related to obesity are due to the development of insulin resistance and diabetes. The development of agents able to both improve insulin resistance and reduce body weight may be particularly beneficial. The PPAR NR family has received particular interest and as a result the TZDs have emerged as a therapeutic class for the treatment of diabetes. One adverse effect associated with some PPARγ receptor agonists however is weight gain. Given the co-morbidity of type II diabetes and obesity, this can limit the usefulness of such molecules. Quite unexpectedly, a moderate reduction of PPARγ activity observed in heterozygous PPARγ-deficient mice was shown to prevent insulin resistance and obesity induced by a high-fat diet[9]. Consequently PPARγ antagonists are now being developed.

Originally belonging to the family of orphan receptors, since no natural ligands were identified, nowadays polyunsaturated fatty acid and their metabolites have been characterised as endogenous ligands for this class of proteins. As members of hormone activated nuclear receptors, the binding with an agonist lead to a transcriptional cascade resulting in the regulation mostly of lipid metabolism[13].

To date three subclasses of PPARs have been identified, namely PPARα, γ and β (also called delta or NUC-1). Each sub form has different tissue distributions. PPARα is abundant in the liver, kidney and muscle and regulates fatty acid metabolism and inflammatory responses. PPARβ is ubiquitously expressed and its physiological functions are not fully understood[9]. Finally, PPARγ is expressed mainly in adipose tissue, muscle and heart. PPARγ is also the most extensively investigated among the PPAR subtypes. The protein has been cloned from a variety of species showing a high level of conservation across all the species. The human PPARγ presents 95% identity with the murine one. This high level of conservation may reflect the pivotal role that PPARγ plays as a regulator of glucose and lipid homeostasis[14].

PPARγ is expressed in human tissues as three isoforms, PPARγ 1, PPARγ 2 and PPARγ 3, each of them differing at the 5' end and each under control of their own specific promoter. In detail, PPARγ 1 and 3 are expressed highest in adipocytes and in the intestine. In contrast, PPARγ 2 appears to be limited to adipose tissue and liver[15]. PPARγ is also found to be expressed in a number of human tumours, such as neuroglioblastoma[16], liposarcoma[17], colon cancer[18], lung[19,20], prostate[21] and breast[22]. PPARγ shows the most restricted fatty acid binding profile of all the PPARs, interacting mostly with monounsaturated fatty acids.

3.4 Biological relevance of PPAR γ

3.4.1 Adipogenesis, Dyslipidemia and PPAR γ

Adipocytes have recently been identified as a principal component and mediator in many physiological and pathological processes. Much literature has highlighted the role of adipose tissue in obesity, cardiovascular disease and non-insulin dependent diabetes mellitus. In this regard, increasing interest has been applied in the past ten years to adipocyte differentiation and the role that transcription factors and various receptors may have in this process[8,23,24].

The classic view of fat cells identified them as a depot for the storage of energy, with lipids being added or released under the influence of the appropriate hormones. An active role for the adipocyte in energy metabolism emerged recently with the leptin discovery[25], a hormone secreted uniquely by fat cells mediating different signalling to the hypothalamus such as satiety feeling.

Moreover, other pleitropic functions have been conferred to adipocytes: the release of cytokines and lipids leads to a decrease in glucose utilisation in skeletal muscle and enhanced glucose production by the liver, both of which contribute to high levels of glucose in the blood stream (as occurs in non-insulin dependent diabetes disease).

Cytokines also play a role in activating various inflammatory responses that are considered to be an important mediator of cardiovascular disease. Besides fat cells, skeletal and heart muscle, are the only tissues known to express the glucose transporters Glut4. Adipocytes have been shown to control endocrine function by secreting angiotensin, TNFα and interleukine-6. Secretion of glycerol and fatty acids from the adipocyte also plays an important role in hepatic and peripheral metabolism. Leptin itself has a wide spectrum of biological activity, independent of satiety, including effects on fertility, reproduction and hematopoiesis[26]. Thus, the view of the adipocyte has shifted from the notion of an inert storage vessel to that of a central integrator of the metabolic program.

The comprehension of these processes and the need for finding new drugs able to modulate these metabolic paths emerged due to the extraordinary increasing

incidence in obesity, atherosclerosis, type II diabetes and other related diseases. Genetic mutations able to alter the release of leptin or decrease its interaction with the receptors are known causes of obesity in mice and humans.

The release of cytokines and lipids by fat cells lead to a decrease of glucose usage in skeletal muscle and enhancement of glucose production in liver, both of which contribute to high glycaemia in peripheral circulation, a typical sign of diabetes non insulin-dependent (type II).

Knowledge concerning the sequence of transcriptional events during adipogenesis provides a basis for understanding the physiological processes associated with adipose tissue as well as for the development of therapeutic intervention in these adipocyte-related diseases.

A high number of the proteins involved in these metabolic pathways have been enumerated and investigated in recent publications. Within all the factors involved in the cascade of transcriptional events leading to the adipocyte differentiation a prominent role has emerged to be played by PPARγ.

The first stage of differentiation of fat cells is the growth arrests usually determined by contact inhibition. In cultured cells, initial growth arrest is induced by addition of a pro-differentiative hormonal regime and is followed by one or two additional rounds of cell division known as clonal expansion. This process ceases coincident with the expression of the transcription factors PPARγ and C/EBPα (CCAT enhancer binding protein). The induction of these two proteins leads to a fully differentiated phenotype. This is followed by de novo or enhanced expression of most or all the genes that characterise the adipocyte phenotype, along with triglyceride accumulation. The products of the genes include glycerophosphate dehydrogenase, fatty acid synthase, acetyl CoA carboxylase, malic enzyme, Glut 4, the insulin receptor and AP2 (adipocyte-specific lipid binding protein). The analysis of the aP2 5' flanking region led to the discovery of one of the first enhancer elements for fat cells, the PPARγ receptor.

This gene regulatory sequence, located 5 Kb upstream of the initiation point, has been identified as both essential and sufficient for the expression of the gene. Differentiation-dependent, tissue specific transcription is directed by this adipocyte-specific enhancer, which includes several regulatory DNA elements. In detail, a *cis*-

acting element (ARA6 and 7) has been identified, as well as a *trans*-acting factor, ARF6, that binds to these sites.

PPARγ exists in two different isoforms that are produced by alternative splicing: PPARγ 1 and 2. In detail, PPARγ 2 contains 30 AA more then the isoform 1. While many tissues express low levels of PPARγ 1, PPARγ 2 is present only in fat cells. Thus, PPARγ 2, with the RXR is definitely required for the function for the expression of aP2. Moreover, PPARγ's crucial role has been identified in the function of many others fat-cell specific genes, such as PEPCK (phosphoenol pyruvate carboxykinase) and GyK (glycerol kinase), FATP and CD36 (fatty acid transport protein) (Figure 3-1).

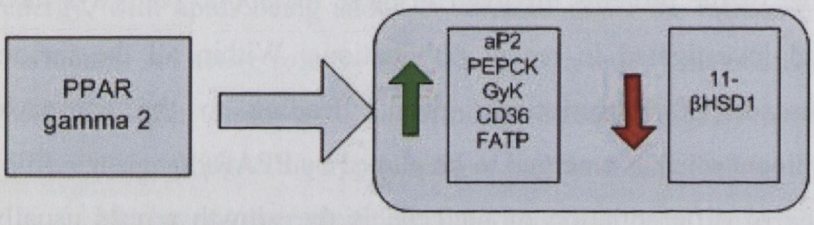


Figure 3-1 PPARγ role in the activation of different genes.

Thus, PPARγ agonist-mediated differentiation includes lipid accumulation and the expression of many endogenous genes characteristic of adipocyte. Interestingly, the ability of PPARγ to promote adipogenesis is not limited to fibroblastic cells. Myoblastic cell lines can also be converted to adipocytes, particularly when the cells co-express C/EBP alpha in addition to PPARγ

The brown adipocyte tissue (BAT) is another important tissue where PPARγ is highly expressed. Conversely to the white adipocyte tissue (WAT), the primary function of BAT is to dissipate energy in form of heat. Although brown fat cells have virtually the same enzymatic machinery as white fat cells, the accumulated lipid in BAT serves mainly for a source of fatty acid to be oxidised in mitochondria. Recent data showed that BAT in humans is present not just in the neonatal period, but also in a small island within the ordinary WAT. Although no transcription factors have been identified that are specifically expressed in brown fat cells, PPARγ and C/EBPα are induced during brown adipogenesis, in a similar way to white cell differentiation. The connection between PPARγ and differentiation of WAT and BAT led to the discovery of an important PPARγ co-activator: PGC-1.

PGC-1 is a protein of 90 Kd that is expressed in skeletal muscle, heart, kidney, brain and brown fat but not in white fat. Upon cold exposure the level PGC-1 increases mostly in muscle and BAT, the two major sites of adaptive thermo-genesis.

Furthermore, several observations indicate that of PGC-1 can preferentially direct pre-adipocyte to a brown phenotype. It is important to underline that PCG-1 is able on the one hand to bind and activate directly PPARγ, and on the other hand to increase the level of UCP-1, UCP-2 and stimulate mitochondrial biogenesis.

The prominent role of PPARγ in adipocyte metabolism and the understanding of the regulatory mechanism described above made this protein a plausible target for different diseases. One could imagine using PPARγ antagonist to reduce fat accumulation in WAT. Naturally, any attempt to manipulate adipogenesis without addressing the deregulated energy balance of the organism would mean that the excess of calories could be deposited elsewhere with potentially deleterious effects (i.e. liver leading to hepatomegaly).

It is also possible to imagine that increasing adipogenesis selectively in BAT could be metabolically useful. It could potentially have an anti-obesity effect by increasing the energy expenditure.

3.4.2 Diabetes and PPARγ

One of the approved pharmacological uses of PPARγ agonists (troglitazone, pioglitazone and rosiglitazone) is for the treatment of insulin independent diabetes type II (currently marked agents for diabetes by the FDA). In humans with this disease, activators of PPARγ, such as TZD, significantly improved insulin sensitivity.

PPARγ agonists are proven to reduce plasma glucose, lipids and insulin levels. However, the detailed mechanism of these drugs is currently not clear or completely understood. Considering that PPARγ is mostly localised in the adipose tissue, low expression in liver and very low expression in muscle, it is not clear how it can affect the muscle uptake of glucose, which is the major tissue of glucose disposal. A plausible mechanism of action emerges after evaluating the set of transcriptional changes originating from the activation of PPARγ. These include

genes involved in glucose uptake (c-Cbl-associated protein/CAP and GLUT 4), lipid uptake and storage (CD36, aP2, LPL, FATP and acyl-CoA syntase) and energy expenditure (glycerol kinase/GyK, uncoupling protein UCP-2 and UCP-3). However, the effect on CAP and GLUT4 cannot solely be responsible for the efficacy of TZD, since adipose tissue is accountable for the 5% of the total glucose disposal.

Sequestering lipids into fat stores reduces the metabolic burden on the liver and muscle as well as promoting glucose use. Free fatty acid (FFA), in particular, causes insulin resistance in muscle, thus lowering these molecules has some beneficial effects. Similarly, UCPs allow protons to cross the mitochondrial membrane bypassing the ATPase, thus diverting potential energy into heat instead of ATP formation. Increased energy expenditure should be therapeutically beneficial in diabetic patients, especially in those associated with obesity. PPARγ ligands also appear to increase glucose oxidation in the muscle and decrease gluco-neo-genesis in the liver in part by down-regulating pyruvate dehydrogenase 4 (PDK40) and PEPCK. Finally, by repressing the expression of molecules such as TNF alpha and leptin in the adipose tissue, they are able to influence insulin resistance.

Thus, rather then a singular mechanism accountable for the effect of PPARγ agonists in type II diabetes, an ensemble of actions has emerged that might be synergistically responsible for the lower level of glucose and the improved insulin sensitivity.

3.4.3 Cancer and PPAR γ

The discovery of PPARγ and its capacity to promote the differentiation and stop the proliferation in adipocytes, suggested that this receptor may be capable of reducing malignant cancer behaviour. This was first highlighted in human liposarcoma, a malignant cancer of adipose tissues. Many sarcoma have been found to over express PPARγ and be more sensitive to administration of PPARγ agonists. Exposure of the sarcoma to PPARγ ligands caused differentiation, lipid accumulation, cessation of growth and expression of mRNAs characteristic of mature adipocyte. PPARγ is also expressed in different epithelial tissues, important

in human cancer, such has breast, prostate and colon. This evidence suggests the importance of PPARγ agonists as a potential target for cancer.

3.4.4 Neuroprotection and PPAR γ.

Pioglitazone (see section 1.5), as with other PPARγ agonists, have been discovered to possess the ability to suppress T-cell activation suggesting that it might be useful in the treatment of inflammatory diseases. Moreover, activation of PPARγ in microglia (as well as macrophages) results in decreased activation of these cells[27,28]. This implies that PPARγ may be useful in a number of central nervous system diseases (CNS), such as Parkinson Disease (PD), Alzheimer's Disease (AD) and Multiple Sclerosis (MS). It is now evident that PPARγ plays a major role in the regulation of the inflammation process[29,30].

3.5 PPARγ ligands

3.5.1 Synthetic Ligands

With some exceptions, PPARγ agonists can be classified into three groups of molecules: **TZD ,Dihydro-cinnamatate derivates** (such as AZ-242) and **tyrosine derivates** (GI 262570).

3.5.1.1 TZD derivatives

The first agonists with high affinity for PPARγ identified were a class of anti-diabetic drugs known as TZD[31] (Table 3-1). TZD are generally selective for PPARγ over the other subtypes of PPARs although exceptions are known in the literature[10,32]. TZDs have been developed over a period of 20 years. The structure of Rosiglitazone in complex with PPARγ has been reported in the presence (PDB entry 1FM6)[33] and in the absence (PDB entry 2PRG)[34] of the hetero-dimer RXR. TZDs generally show an anti-diabetic effect lowering the blood level of glucose, although they often produce weight gain and do not improve the lypidemia in patients. Furthermore they exhibit side effects, such as increased adiposity, edema, significant cardiac hypertrophy and the risk of heart failure.

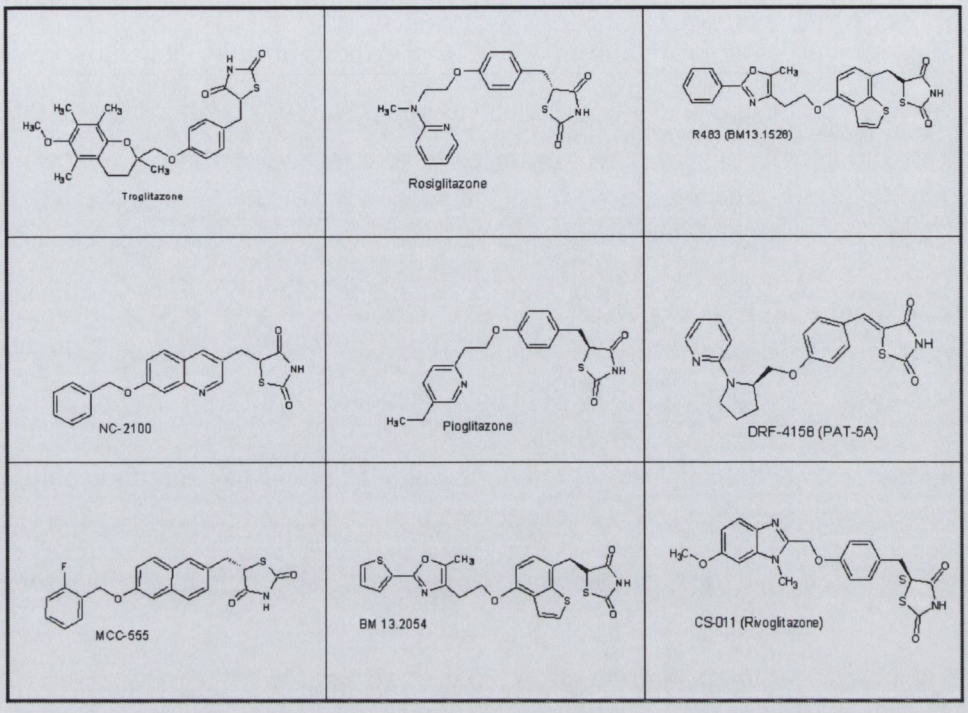


Table 3-1 TZD derivatives.

Trogliatazone[35] was approved in the US in 1997 and withdrawn in 2000 for liver damage occurring in around 1 in 100,000 patients[36]. Side effects apart, troglitazone lowered the level of blood glucose in patients and reduced insulin resistance.

Rosiglitazone (BRL 49,653) is the current leader of the insulin sensitiser market. It was launched by Avandia in the US in 1999 and in Europe in 2000. The properties of rosiglitazone are also extensively reported in literature[37,38]. The drug is a very successful treatment for type II diabetes. It does not appear to cause liver damage, as troglitazone, and effectively lowers blood glucose. In contrast, rosiglitazone has negative effects on body weight and lipid profiles[36].

pioglitazone[39,40] (patented by Actos) was approved in the US in 1999. The literature reports similar anti-diabetic effect as rosiglitazone and it does not appear to produce any liver damage. In contrast with rosiglitazone, pioglitazone has a favourable lipid profile and appears to have anti-atherosclerotic effect[36].

MCC-555[41] (isogliatazone) was reported to be more efficacious then pioglitazone and troglitazone in lowering the plasma levels of glucose, insulin and triglyceride. The activity of MCC-555 appears to be context specific on PPARγ, as the activity varies from full agonist to partial agonist depending on the target tissue.

R-483 (BM 13.1528) and the similar **BM13.2054** are other examples of TZD developed in recent years[42] and their acute and chronic effect on body weight and glucose levels has been investigated. In addition to PPARγ activation, evidence shows that these compounds affect muscle glucose metabolism[36].

NC-2100 is a TZD with a similar structure to MCC-555 and, like that molecule, it seems to have a specific activity on PPARγ depending on the context. The EC_{50} of NC-2100 is 30-50 fold less active then rosiglitazone, although it has the same capability to reduce plasma levels of glucose in mice. It also causes lower body weight increases, probably due to the high induction of UCP1 and UCP2 (Uncoupling Protein).

CS-011 (rivoglitazone) is a TZD under development for the treatment of type II diabetes. CSS-011 has been shown to have superior bioavailability, efficacy and a longer half-life in rats when compared with rosiglitazone. Body weight increases were also observed at therapeutic doses. CSS-011 had greater hypoglycaemic potency compared with rosiglitazone in mice and was more potent than rosiglitazone in adipocyte differentiation assay[36].

DRF-4158 (PAT-5A) was reported as having comparable anti-hyperglycemic activity to rosiglitazone and troglitazone with a better lipid profile[43]. The TZDs contain a stereogenic centre at C-5 of the heterocyclic head group but have been developed as a racemate since they undergo racemisation under physiological conditions[36]. A PPARγ binding assay revealed that only the (S)-enantiomer of the TZDs bind to the receptor with high affinity[44]. This result suggests that only 50% of the drug substance in the currently approved TZDs binds to the target receptor, while 50% of the drug substance is inactive[10].

3.5.1.2 Dihydro-cinnamate derivatives

To overcome the racemisation problem Buckle et al. have synthesised acyclic head group compounds that are less prone to racemisation: **SB 213068** and its (S)-isomer, **SB 236636**, are representatives of a series of R-alkoxy-phenylpropanoic acids[45,46]. From a formal point of view these compounds can be classified as dihydro-cinnamate derivatives (Table 3-2).

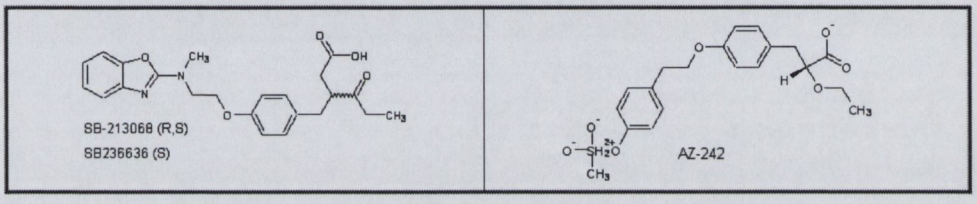


Table 3-2 Cinnamate derivatives.

AZ-242 (AR-H039242 - Astra Zeneca) is a PPARα/γ co-agonist with a cinnamate head group. Its structure in complex with both subtypes of PPARs has been solved (1I7G and 1I7I)[47]. It has been reported to have enantiomeric stability, which distinguishes it from classic TZDs. AZ-242 exhibits the ability to lower the blood glucose and a positive lipid profile.

3.5.1.3 Tyrosine derivatives

Farglitazar (GI262570) was one of the first compounds to be optimised based on its activity at human PPARγ. It represents the first of a series of molecules that contains some of the most potent agonists reported.

In addition, these compounds showed >1000 fold selectivity on PPARγ over PPARα and β[34,48]. Chemically farglitazar belongs to the class of N-(2-benzoylphenyl)-L-tyrosine derivative. This molecule formally represents a hybridisation between the new tyrosine based scaffolding with "left-hand" sides from TZDs. **GW-1929**, **GW-7872** and **GW-7845**[48] also belong to the series of

tyrosine derivatives and are similar to Farglitazar for potency and selectivity over PPARγ over the other subtypes (Table 3-3).

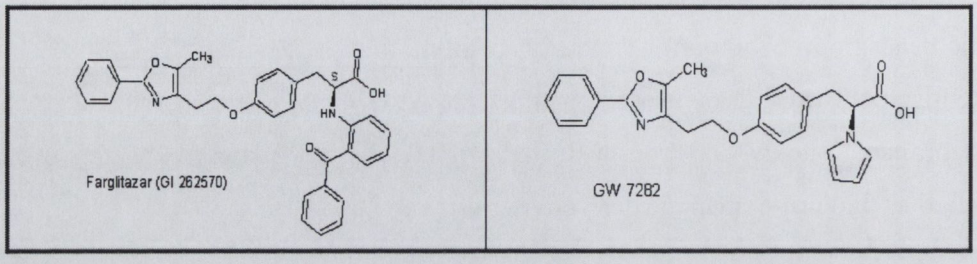


Table 3-3 Tyrosine derivatives.

3.5.2 Dual Agonists

Diabetic patients are also at a high risk for atherosclerosis and myocardial infarction, and some risk can probably be attributed to diabetic dyslipidemia, which is characterised by high triglyceride and low HDL levels. The PPARγ selective agonists, such as rosiglitazone and pioglitazone are currently used in therapy for lowering glucose levels in diabetes although they do not significantly improve lipid levels in patients. In both diabetic and non-diabetic patients, hyper tryglyceridemia is often treated with *fibrates*, which are PPARα agonists. Fibrates can lower triglycerides levels 30-50% and also increase HDL up to 20% but do not improve glucose levels. Therefore, the discovery of dual agonists (Table 3-4) active on PPARγ and α may correct hyper-tryglyceridemia and diabetic dyslipidemia as well as lower the plasma level of glucose.

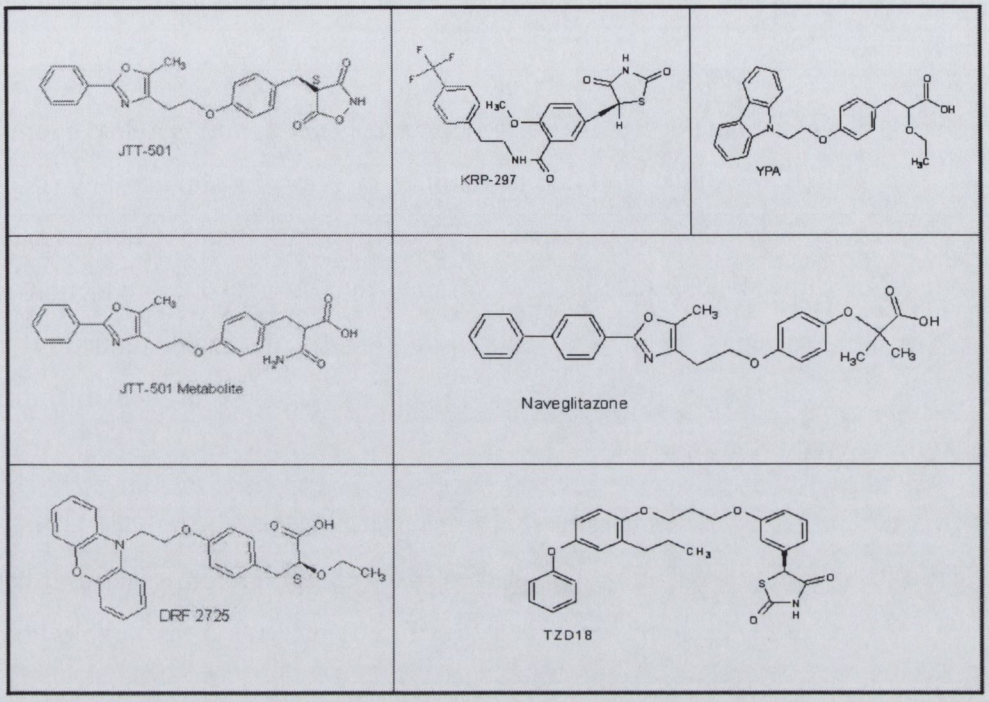


Table 3-4 Dual Agonists.

TZD **KRP-297** (discovered by Kyorin Pharmaceutical) has been reported to have dual PPARα/γ activity in a trans-activation assay. KRP binds directly with K_d values of 228 and 326 nM/L, respectively for the subunit α and γ[49]. Unlike rosiglitazone, KRP-297 has a meta-substituted side chain across the central phenyl rings. The meta-substituted side chain may allow an improved fit to the PPARα protein[50].

LY519818[32] (Lilly 2001, naveglitazone) represents a new class of non-TZD showing dual activity for PPARγ/α. The series of compounds, for which LY19818 represents the leader, were prepared based on the rational design of a dual PPAR agonist agent using the putative pharmacophores of known PPARγ and α selective compounds. Chemically it can be classified as a cinnamate derivative.

JTT-501 has a non-TZD scaffold with a dual agonist activity, although the isoxazolidinedione is structurally related to TZDs. It was reported to activate the α subunit 10-fold higher then for activation of PPARγ. Its activity is likely to be

mediated through a malonic amide metabolite that is generated by hydrolysis of the heterocyclic ring.

DRF-2725[51] is a cinnamate derivative (also known as NN622) and is a dual agonist PPARγ/α whose binding mode has been determined by X-ray crystallography (PDB entry: 1NYX). In diabetic, insulin resistant, hyper-lipidemic mice, treatment with this compound showed 56% reduction of plasma glucose and 62% reduction in triglyceride as compared to 33% reduction in glucose and 16% reduction in triglyceride by rosiglitazone. It has also shown extremely good oral bioavailability and pharmacokinetic characteristics.

TZD-18[52] is a TZD derivative that displayed robust anti-diabetic activity in mice and Zucker fatty rats and was more potent then rosiglitazone in both species. It has the capability to lower the lipids level and therefore it presents a favourable lipid profile for the treatment of type II diabetes or syndrome X.

YPA[53] is a cinnamate derivative that has been co-crystallised with PPARγ (PDB entry 1KNU). It has shown both anti-diabetic and hypolipidemic effects.

3.5.3 Partial Agonists

Several alternative approaches have been taken to seek classes of PPAR ligands with improved therapeutic profiles thus minimising the undesirable side effects.

As an example **GW-0072** (GSK) (Table 3-5) is a compound that does not classify easily into any of the above chemical categories. GW-0072 was identified from a combinatorial library of non-TZD thiazolidine acetamides. It has a profile as a partial agonist in trans-activation assay with only 15%-20% of the efficacy of rosiglitazone. It antagonises rosiglitazone in this assay to the level of its own partial agonist activity. GW-0072 behaves as a full antagonist of rosiglitazone-induced adipocyte differentiation. Thus, this drug is a PPARγ modulator that can inhibit the adipogenic effect of rosiglitazone, but not of insulin, since it appeared to function as an insulin sensitiser.

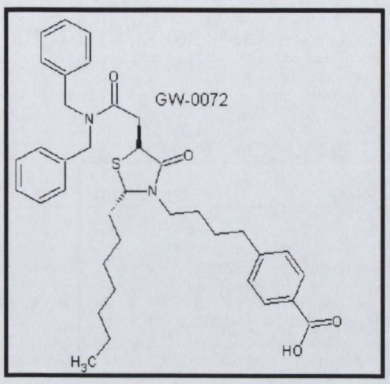


Table 3-5 GW-0072.

Furthermore, GW-0072, as shown in the following paragraphs, shows an atypical binding mode within the LBD of PPARγ, without any direct interaction with the region AF-2, believed to be essential for the activity for PPAR γ agonists.

3.5.4 Natural Ligands

The search for the natural PPARγ ligands led to the discovery of a number of fatty acids and eicosanoid derivatives that bind and activate the receptor[10] (Table 3-6).

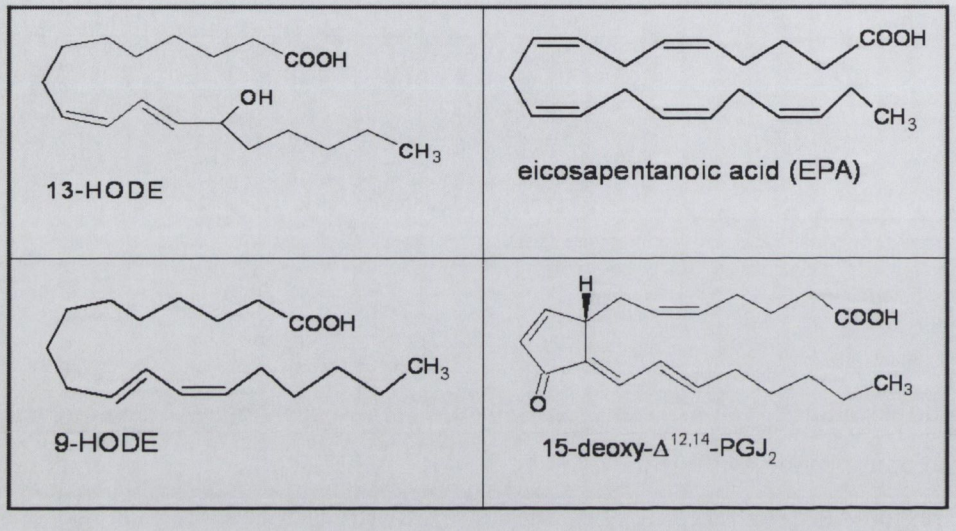


Table 3-6 Natural Ligands for PPARγ.

The essential fatty acids linoleic acid, linolenic acid, arachidonic acid and eicosapentaenoic acid (EPA) have been shown to bind PPARγ at micromolar concentrations found to be present in human serum[54]. It is unclear if these concentrations can be present inside cells. A possible mechanism of action may involve their metabolites. The 15-lipoxygenase metabolites of linoleic acid, 9-HODE and 13-HODE, have been shown to function as micromolar PPARγ agonists[55]. Thus, metabolic conversion of polyunsaturated fatty acids inside cells could supply an extra level of hormonal regulation of PPARγ. Additionally, the *J*-series of prostaglandins derived from PGD2 have also been identified as PPARγ ligands. The terminal metabolite 15-deoxy-12,14-prostaglandin J2 (15d-PGJ2) activates PPARγ in low micromolar concentrations and also induces adipocyte differentiation[31,56]. This prostaglandin has become the most widely used natural PPARγ ligand. However, it is important to note that 15d-PGJ2 mediates some of its effects in cells through PPARγ's independent signalling pathways, and caution must be exercised in attributing the biological effects of 15d-PGJ2 solely to activation of

PPARγ[10]. Furthermore, the molecular recognition of fatty acids by PPARγ has been resolved with X-ray crystallography (PDB entry: 3GWX)[57].

3.6 PPAR γ LBD

Several structures of PPARγ have been crystallised either with natural ligands[57], with synthetic ligands[58] or the apo-structure without any ligand[34]. Some of the structures were solved as homo-dimers, others as hetero-dimer with RXR. Furthermore PPARγ has been also crystallised with some peptidic co-activators where it is easy to identify interacting features as the LxxLL motif (L:Leucine).

A list of the LBD crystal structures available on the PDB database is shown in Table 3-7 with the relative codes and resolutions.

PDB ID	Res Å	Year	Authors
1GWX	2.50	2000	Xu, H.E., Lambert, M.H., Montana, V.G., Park, D.J., Blanchard, S., Brown, P., Sternbach, D., Lehmann, J., Bruce, G.W., Willson, T.M., Kliewer, S.A., Milburn, M.V.
1I7I	2.35	2002	Petersen, J.F.W., Cronet, P., Folmer, R., Blomberg, N., Sjoblom, K., Karlsson, U., Lindstedt, E.-L., Bamberg, K.
1KNU	2.50	2002	Svensson, L.A., Mortensen, S.B., Fleckner, J., Woeldike, H.F.
1PRG	2.20	2001	Nolte, R.T., Wisely, G.B., Milburn, M.V.
1RDT	2.40	2004	Haffner, C.D., Lenhard, J.M., Miller, A.B., McDougald, D.L., Dwornik, K., Ittoop, O.R., Gampe Jr., R.T., Xu, H.E., Blanchard, S., Montana, V.G.
1WM0	2.90	2004	Ostberg, T., Svensson, S., Selen, G., Uppenberg, J., Thor, M., Sundbom, M., Sydow-Backman, M., Gustavsson, A.L., Jendeberg, L.
1ZEO	2.50	2006	Shi, G.Q., Dropinski, J.F., McKeever, B.M., Adams, A.D., MacNaul, K.L., Elbrecht, A., Berger, J.P., Zhou, G., Doebber, T.W.
2ATH	2.28	2006	Mahindroo, N., Huang, C.-F., Wu, S.-Y., Hsieh, H.-P.
2F4B	2.07	2006	Lu, I.L., Peng, Y.H., Mahindroo, N., Hsieh, H.P., Wu, S.Y.
2FVJ	1.99	2006	Benz, J., Burgermeister, E., Flament, A., Schnoebelen, A., Stihle, M., Gsell, B., Rufer, A., Ruf, A., Kuhn, B., Maerki, H.P., Mizrahi, J., Sebokova, E., Niesor, E., Meyer, M.
2GTK	2.10	2006	Kuhn, B., Hilpert, H., Benz, J., Binggeli, A., Grether, U., Humm, R., Maerki, H.-P., Meyer, M., Mohr, P.
2HFP	2.00	2006	Pokross, M.E., Evdokimov, A.G., Walter, R.L., Mekel, M.J., Hopkins, C.R.
2I4J	2.10	2007	Pochetti, G., Mazza, F.
2I4P	2.10	2007	Pochetti, G., Mazza, F.
2I4Z	2.25	2007	Pochetti, G., Mazza, F.
2OM9	2.80	2007	Ambrosio, A.L.B., Garratt, R.C.
2PRG	2.30	1999	Nolte, R.T., Wisely, G.B., Milburn, M.V.
2Q59	2.20	2007	Bruning, J.B., Nettles, K.W.
2Q5P	2.30	2007	Bruning, J.B., Nettles, K.W.
2Q5S	2.05	2007	Bruning, J.B., Nettles, K.W.
2Q61	2.20	2007	Bruning, J.B., Nettles, K.W.
2Q6R	2.41	2007	Bruning, J.B., Nettles, K.W.
2Q6S	2.40	2007	Bruning, J.B., Nettles, K.W.
2REW	2.35	2007	Muckelbauer, J.
4PRG	2.90	1999	Milburn, M.V.
1FM6	2.1	2000	Gampe Jr., R.T., Montana, V.G., Lambert, M.H., Miller, A.B., Bledsoe, R.K., Milburn, M.V., Kliewer, S.A., Willson, T.M., Xu, H.E.
1FM9	2.1	2000	Gampe Jr., R.T., Montana, V.G., Lambert, M.H., Miller, A.B., Bledsoe, R.K., Milburn, M.V., Kliewer, S.A., Willson, T.M., Xu, H.E.
1K74	2.3	2001	Xu, H.E., Lambert, M.H., Montana, V.G., Moore, L.B., Collins, J.L., Oplinger, J.A., Kliewer, S.A., Gampe Jr., R.T., McKee, D.D., Moore, J.T., Willson, T.M.
1NYX	2.65	2003	Ebdrup, S., Pettersson, I., Rasmussen, H.B., Deussen, H.-J., Frost Jensen, A., Mortensen, S.B., Fleckner, J., Pridal, L., Nygaard, L., Sauerberg, P.
1KNU	2.5	2001	Svensson, L.A., Mortensen, S.B., Fleckner, J., Woeldike, H.F.

Table 3-7 PPARγ crystal structure available on the PDB databank[59].

PPARγ crystal structures revealed a large (1,300 $Å^3$) Y-shaped binding site inside which different ligands may be accommodated in different regions of the

LBD itself[47,57]. Most of the ligands bind the PPARγ LBD in a U shape conformation, and usually occupy a small percentage of the cavity itself (Rosiglitazone 40%)[60]. In detail, the structure of Rosiglitazone in complex with PPARγ (Figure 3-2) has been reported both in the presence (1FM6) and in the absence (2PRG) of RXR. Both complexes are quite similar, showing fundamentally the same essential ligand-protein interactions. An inspection of the crystal structure shows how the polar TZD head forms hydrogen bonds with essential residues located in the proximity of the Helix 12 (in the AF-2 region): HIS-323, TYR-473 and HIS-449 (Figure 3-2).

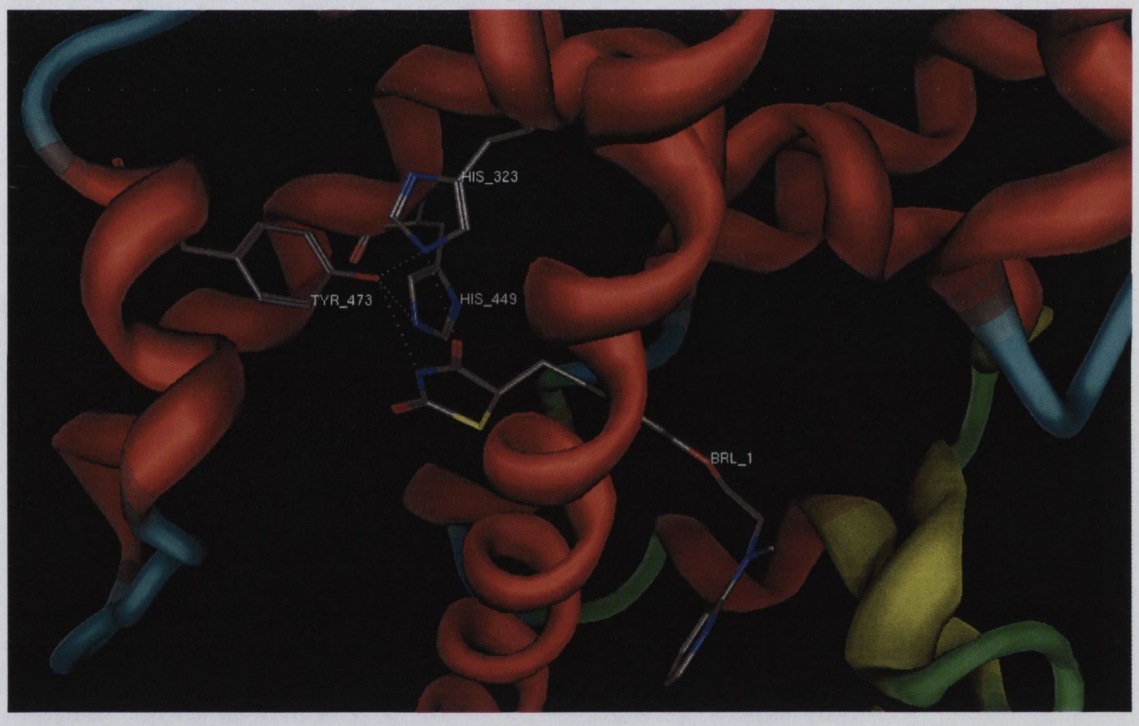


Figure 3-2 PPARγ LBD bound with Rosiglitazone (1FM6). TYR-473, HIS-323 and HIS-449 are labelled.

Similar interactions are shown in the co-crystallised structure of AZ-242 bound to PPARγ. The polar head of the ligand (Dihydro-cinnamatate group) is located again close to the residues mentioned above.

A PPARγ/α dual agonist (Farglitazar) has been co-crystallised in the hetero-dimeric complex with RXR (1FM9). Again, farglitazar binds the receptor in a similar fashion to that depicted above. Its polar head (dihydro-cinnamatate

derivates) binds identically to AZ242, with the carboxylic function interacting with the three residues.

The tails of the two ligands form vdW contacts in the same pocket although the details of the binding are different. Finally, a notable exception to this rule is represented by ligand GW-0072 (partial agonist) which shows no interaction with the AF-2 portion of the LBD.

By analysing the binding mode, the essential role of those residues emerges. Located in proximity to the AF-2 region, they mediate the activity of the agonists. It is likely, as mentioned above, that these hydrogen bond interactions stabilise helix-12 in a conformation allowing the co-factor to bind the receptor with its LxxLL motif. The interaction with the co-factor would then trigger the cascade of events leading to the activation of the transcriptional machinery and the effect of PPAR agonists[61]. Apart from those residues, many more interactions, mostly hydrophobic, and cluster of AA have been identified in recent publications with solvent mapping techniques[62]. The list of features, or hot spots, is shown in Table 3-8.

Site	description surrounding	residues
P1	binding the TZD and carboxyl groups of agonists and interacting with helix H12	F282, C285, Q286, S289, H323, Y327, F363, H449, L469, Y473
P2	buried between H3 and H5 and overlapping the middle of the agonists	C285, R288, S289, I326, Y327, L330, F363, M364
P3	upper distal end of the binding site between H2 and H3 and reached only by the partial agonist GW0072	R288, E291, A292, E295, M329, E343
P4	hydrophobic pocket close to the entrance and close to the distal-end group of agonists	S255, E259, F264, H266, V277, A278, R280, I281, G284, F287, I341, S342, M348
B	surface pocket in the back between H7 and H10/H11 and overlapping the dimerisation region	P366, E369, F370, K373, D441, Q444, I445, E448
F	surface pocket between H3 and H12	Q283, Q286, F287, S464, H466, L469
C1	surface pocket between H10 and H12 and possibly contributing to cofactor binding	V450, L453, Q454, K457, M463, L465, Q470, Y473, K474, D475
C2	between H5 and H12 and overlapping with the binding site of the coactivator peptide SRC-1	V293, T297, V315, L318, K319, T447, L468, E471, I472
E1	pocket defined by the lower ends of helices H3, H7, and H10	E276, I279, R357, F360, M364, E460, D462
E2	putative ligand entrance between H2' and the β sheet	K244, F247, G258, E259, K261, I262, F264, K265, I341, S342, Q345, G346

Table 3-8 List of hot spots recognisable in PPARγ LBD identified by solvent mapping technique (taken from reference [62]).

The binding site identified as P1 contains the important residues SER-289, HIS-323, HIS-449 and TYR-473. This cluster is present in co-crystallised structures with strong agonists suggesting that these compounds are able to produce this protein conformation. In contrast, site P2 is formed in the ligand-free PPARγ (structure PDBentry 4PRG). Moreover, PPARγ bound with the partial agonist GW0072[63] has been found forming interactions with the site P2 rather than the site P1. The area P2 is located between helix 3 and 5 and it can be depicted as a box that buries the central link of a classic TZD ligand and is highly hydrophobic.

The site P3 is also rather important. This is a pocket on the right arm of the Y-shaped binding site, which is out of reach of the known agonists. The only agonist able to interact with this portion of the protein is the partial agonist GW0072[63].

Site P4 is highly hydrophobic and it is close to the lower end of the ligand-binding site, defined by H3 from the left, and by the β sheet and helix H2' from the right.

A potential spot for the dimerisation region has been identified (site B) as well as the interaction site with co-activators (site C1 and C2). Site B is located in the back of the LBD, between helices 7 and 10/11. Site C is located around H12 where, in a similar fashion with the other nuclear receptors, a hydrophobic box is defined and recognised by the co-activator LxxLL motif.

Finally a putative entrance site has been identified close to site E1 and E2, (located between the helix H2' and the β-sheet). This site is only present in the agonist-bound crystal structure, with the exception of 1I7I. In this structure the distal methylsulfoxy group of AZ 242 reaches further down into pocket P4 than in the other agonist structures.

The main interactions between a protein and a ligand can be deduced using LPC (Ligand Protein Contacts)[64]. This software computes distances between the ligand and the protein present in the file coding for 3D coordinate of the crystal structure. A list of putative hydrogen bonds calculated by LPC is shown in Table 3-9.

Residue	1NYX	1KNU	1K74	2PRG	1FM9	1I71	1GWX	1WMO
HIS449	2.8	..9	2.7	2.8	2.8	2.7	2.4	
HIS449	3.2	3.2	3.6		3.5	3.2	3	
HIS323	2.6	3	2.7	2.8	2.5	2.8	2.7	
SER289		2.8	2.9	2.9	2.9	2.8		
TYR473	3.5	2.9	3.2	2.8	3.5	2.7	2.5	
TYR473	2.8		2.5		2.4			
GLY284			5	4.5	4.6			
LEU340	4.3							
SER342			5.1	4.7	5.4	4.8		3.3
CYS285					4.6		4.5	
THR288							3.9	
ALA342							5.6	
GLU343								4.9
HIS266								3.2
HIS266								3.3
ARG280								5.5

Table 3-9 Putative Hydrogen bonds and their length (Å) predicted by LPC. In the columns the crystal structures are depicted and in the rows the main residue interactions identified as essential in each complex.

Table 3-9 shows how the most frequent interactions in almost every crystal structure solved are the contacts between the ligand and residues in the AF-2 region (HIS-449, HIS-323 and TYR-473).

3.7 Computational Methods

The computational protocol used for screening was entirely integrated into PP[65]: A protocol was developed in order to fully integrate third party software (e.g. OMEGA, CORINA, FRED and in-house scripts). Furthermore the integration scripts were specifically developed to run in a multiprocessor environment reducing therefore the overall computational time required for the virtual screen (Sun Grid Engine running of a cluster of 16 Intel Xeon CPU 3.06 GHz).

3.7.1 Validation set and screening Set

35 PPARγ agonists were initially retrieved from the WOMBAT database[66], selecting those molecules with associated biological activity data. The full list of PPARγ selected structures with references is depicted in the appendix of this chapter. All of the 35 molecules passed vHTS filters and organic filters implemented in PP. The protocol was first validated and optimised using a validation set of decoy molecules seeded with the 35 PPARγ known actives. The quality of the protocol was evaluated considering the ROCs results after implementing post-processing techniques. Screening and decoy validation sets were obtained following essentially an identical protocol. The only difference between the sets was the final number of molecules, adjusted using a clustering step: for the validation set the clustering restricted the total number of structure to 1,000. For the screening set the output cluster centres were 10,000. The source for the validation set was SPECS[67] and for the screening set Maybridge[68], SPECS and Asinex Platinum were utilised. The mol2 file format has been utilised throughout the validation and the screening.

3.7.2 Pre-processing

Both validation and screening sets were initially filtered using drug-like filters implemented in PP (Organic Filter, HTS filter and Lipinski Filter), ensuring duplicates were removed.

An ADMET prediction model implemented in PP pilot was used to rule out compounds predicted to poorly cross the Blood Brain Barrier[69]. This component can calculate two properties:

ADMET_BBB: Log of Brain/Blood partition coefficient (LogBB).

ADMET_BBB_Level: Ranking of the LogBB values into one of the following levels: 0 : Very High, 1 : High, 2 : Medium, 3 : Low, 4 : Undefined (Molecule is outside the confidence area of the regression model used to calculate LogBB). Molecules with level equal 3 were discarded. Undefined molecules were kept. `(ADMET_BBB_Level <= 2 or ADMET_BBB_Level == 4).`

The ADMET filter was applied in order to increase the chance of identifying PPARγ ligands with a favourable Blood-Brain Barrier pharmacodynamics and the ability to enter the CNS.

Clustering was carried out using the component implemented in PP. This component clusters the input molecules based on molecular descriptors and adds a cluster membership property to each record. Clustering is based on Tanimoto similarities between the user-selected descriptor sets. The clustering method is a relocation method based on maximal dissimilarity partitioning. Details of the algorithm were discussed in Chapter 1. The cluster centres and closest to centre were included in the output `(DistanceToClosest >= 0.7 or ClusterCenter != 0).`

A bad valence filter was applied to leave out molecules containing one or more atoms with bad valence states: input molecules are examined for bad valence states (such as quaternary nitrogen atoms with no charge on the N).

A Bayesian classifier was trained on the set of actives to ensure that similar ligand properties for both the validation and screening sets was used.

Finally, prior to the generation of conformers, stereocentres and charges were standardised and ionisation states were enumerated between a pH of 7 and 8.

Using the PP **Standardize Stereo** component atoms that are perceived as true stereo atoms but which have no stereochemical markings are set to UnknownAtomStereo. Atoms that have stereochemical markings but which are not true stereoatoms are set to `NoAtomStereo`. Similarly, bonds that are perceived as true stereo double bonds, but which have no stereochemical markings are set to `UnknownBondStereo`. Bonds that have stereochemical markings but which are not true stereo bonds are set to NoBondStereo.

For the **Charges Standardization** component the following formal charges for common functional groups are applied:

```
Quaternary Nitrogen: +1
Nitro groups: N+(=O)O-
Diazonium N: +1
Oxygen with three attachments: +1
X=O-C Oxygen: +1
Sulfur with three attachments: +1
S=O-C Sulfur: +1
F, Cl, Br, I with no attachments: -1
```

3.7.3 Generation of conformers

Following the approach of Chapter 2, for each entry in the validation and screening set in sdf format nine different permutations were produced using the script described in the Appendices to Chapter 2. Subsequently CORINA3.4[70] and OMEGA2.0[71] (described in Chapters 2 and 4) were used for the generation of three dimensional structures. CORINA was employed with default settings. A record of the parameters used for OMEGA2.0 is presented in the Appendix of this chapter. This list of parameters was obtained after numerous trials following the broad approach detailed in Chapter 2.

Following conformer generation with CORINA and then OMEGA, each structure was checked for the presence of overlapping atoms using an algorithm implemented within PP. The coordinates of each input molecule were checked to

determine whether or not there were pairs of atoms too close to each other. Molecules must have either 2D or 3D coordinates. This component is useful to assess the quality of 2D molecular depictions or of 3D conformations. For this purpose the default vdW scale factor of 0.7 was used. This value specifies the scaling factor to apply to each atom's vdW radius for bump checks. This factor is used to multiply the vdW radius of each heavy atom in the molecule.

3.7.4 FRED: Rigid Docking

After numerous tests consisting of docking into different crystal structure coordinates available from the PDB repository, the coordinates of the crystal structure identified with PDB code 1NYX was selected as the frame reference to be used to carry out the docking with FRED[72] (see section 1.9.1).

1NYX is the Ligand binding domain of the human PPARγ in complex with a dual-acting PPARγ/α agonist Ragaglitazar. The coordinates of the receptor were pre-processed as explained in Chapter 1, including partial charge optimisation (AMBER99[73]), hydrogen addition and subsequent minimisation (AMBER99[73]). The ligand and apo-protein were then saved as different files and used respectively as the box and pocket reference. All the scoring functions available in FRED were evaluated.

A complete list of FRED parameters is given in the Appendix to this chapter.

3.7.5 Connolly and Gaussian Surfaces

The Connolly Surface, consists of the boundary of the volume of space defined by removing all points inside the van der Waals surface of spherical probe (that represents the solvent) located at all positions for which the probe does not contact the van der Waals volume of a molecule.

The *Gaussian Surface* is a smooth approximation of the Connolly Surface and is constructed from a sum-of-Gaussians density derived from the atomic coordinates of the molecule. The particular approximation has found use in determining atomic volumes and dielectric regions in electrostatics calculations[74].

3.8 Biological methods

Competitive binding affinity experiments were carried out using purified human PPARγ LBD and applying Fluoromone, a fluorescein-labeled PPARγ ligand. Both were contained in the PPARγ Green Assay Kits obtained from Invitrogen Corporation. HrPPARγ was stored at -80°, and not subjected to any vortexing.

3.8.1 Methodology

Upon binding of a fluorescent molecule to the HrPPARγ and formation of the Fluoromone-receptor complex, it tumbles slowly resulting in a high polarisation value. Polarisation, measured in mP units, is directly related to the molecular volume of the tumbling molecule. Thus, if a competing compound displaces the Fluorormone, it causes the molecule to tumble rapidly in solution and results in a low polarisation value being obtained. This change in mP value allows the determination of the binding affinity of the compound. The concentration of the competitor that results in a half-maximum shift in polarisation is equivalent to the IC_{50} value, and a curve can be plotted using different concentrations of the test compound.

3.8.2 FP Competitive Binding assay

The FP assay was performed as per the manufacturer instructions. Briefly, complete PPARγ Green Screening buffer was made up by addition of DTT (di-thiothreitol) to a final concentration of 5 mM in the buffer provided. PPARγ LBD and Fluoromone were combined in complete PPARγ Green Screen buffer to 0.001 mg/ml and 5 nM respectively to give 2X PPARγ LBD/Fluoromone green complex.

The vial with the complex LBD/Fluoromone was gently inverted, ensuring no vortexing of the mixture so as to avoid denaturation.

Competing compounds were prepared at a standard concentration of 4 mM in DMSO. Compounds were diluted to 200 μM in Complete Screening Buffer.

The 2X PPARγ LBD/Fluoromone green complex (20 μL) was plated in 384 well flat bottom black plates and 20 μL of competing compounds were subsequently added to each well. Each compound was repeated in triplicate.

Wells for positive control (Rosiglitazone), vehicle (DMSO) and blank were plated in triplicate as well.

Following two hours incubation, the sample underwent FP measurement using a 535 nm excitation. More details about the competitive binding assay theory are given in the appendix of this chapter.

3.9 Results

3.9.1 Validation – Ligand re-insertion

The validation set of molecules was used to refine the VS protocol and to prove its utility before the screening of a larger library of compounds.

Various trial experiments were run in order to determine the optimal parameters to be used for the generation of conformers taking into account the quality of (i) conformers and (ii) computational time spent for their generation. Furthermore, different PPARγ co-crystal structures were tested against different scoring functions implemented in FRED. As depicted in Table 3-10, the most efficacious scoring function was Chemgauss2 with the co-crystal structure 1NYX. However, even when considering just the averages of the ROCs score for each scoring function in the different PDB entries, 1NYX remained to be best structure and Chemgauss2 the best scoring function.

	Chemgauss2	*Chemgauss*	*Chemscore*	*PLP*	*Screenscore*	*Shapegauss*	*Average*
1fm6	0.843	0.87	0.53	0.697	0.726	0.857	0.754
1k74	0.884	0.79	0.574	0.692	0.758	0.785	0.747
1knu	0.907	0.778	0.646	0.789	0.772	0.768	0.777
1nyx	0.953	0.832	0.661	0.843	0.828	0.815	0.822
1rdt	0.858	0.769	0.544	0.704	0.764	0.779	0.736
2fb4	0.921	0.763	0.654	0.794	0.779	0.75	0.777
2prg	0.85	0.796	0.596	0.673	0.708	0.777	0.733
4prg	0.858	0.752	0.677	0.806	0.803	0.734	0.772
Average	0.884	0.794	0.610	0.750	0.767	0.783	0.765
StDev	0.039	0.039	0.057	0.065	0.039	0.038	0.029

Table 3-10 ROCs values calculated for each scoring function available in FRED against eight different PDB entries for PPARγ. In dark yellow is highlighted the best combination.

Figure 3-3 shows the ROCs curve obtained. Furthermore, the combination of CORINA, OMEGA for the generation of the conformers and the rigid docking with FRED was capable of reproducing the orientation of Ragaglitazar (ligand in 1NYX PDB entry) with an RMSD of 0.6 Å (RMSD by Openeye was used for every RMSD calculation) compared to the crystallographic ligand pose. Figure 3-4 illustrates the

superimposition of the ligand in the original structure 1NYX (red) and the orientation reproduced by FRED (cyan). Figure 3-5 also shows how the best conformer generated using the method described in the previous section, resembled the coordinates and the orientation in the binding pocket of the co-crystal structure 1NYX. An RMSD of 1.6Å of the ligand fixed in the receptor was calculated.

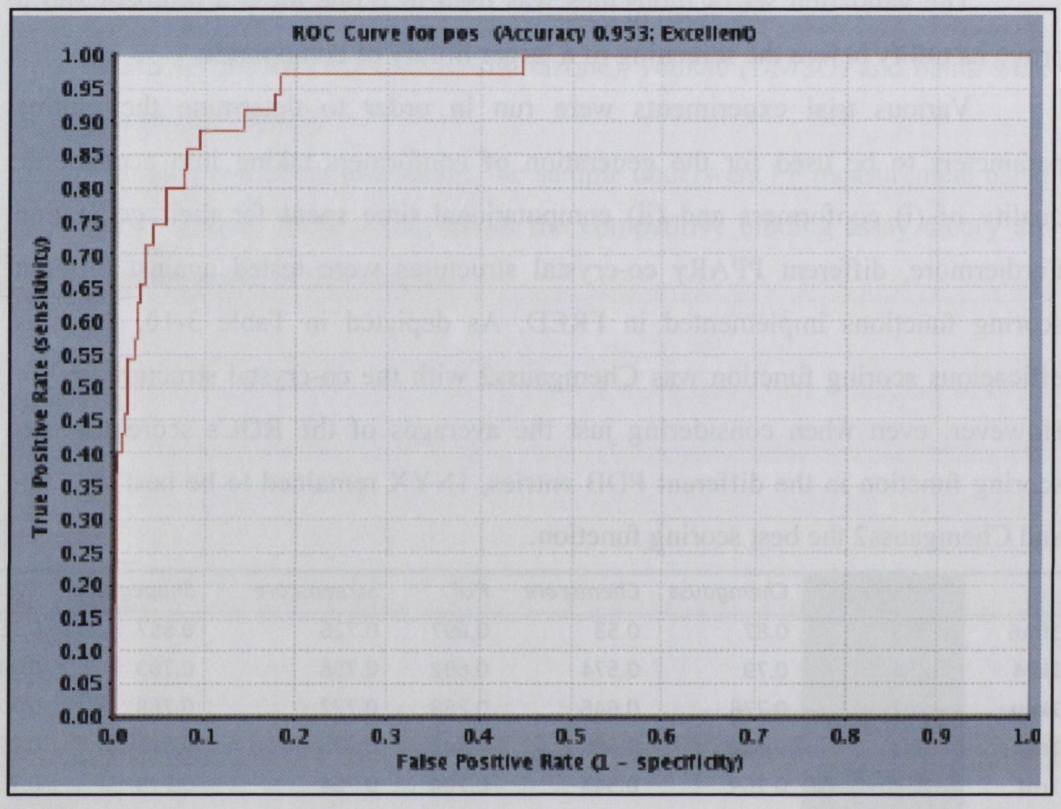


Figure 3-3 ROCs curve obtained from docking in PDB 1NYX, ranked using chemgauss2 scoring function.

Figure 3-4 Superimposition of the original Ragaglitazar as in the PDB entry 1NYX (red) and the original re-docked ligand using FRED (cyan) within the binding pocket. The blue box represents the area in which FRED performed the docking.

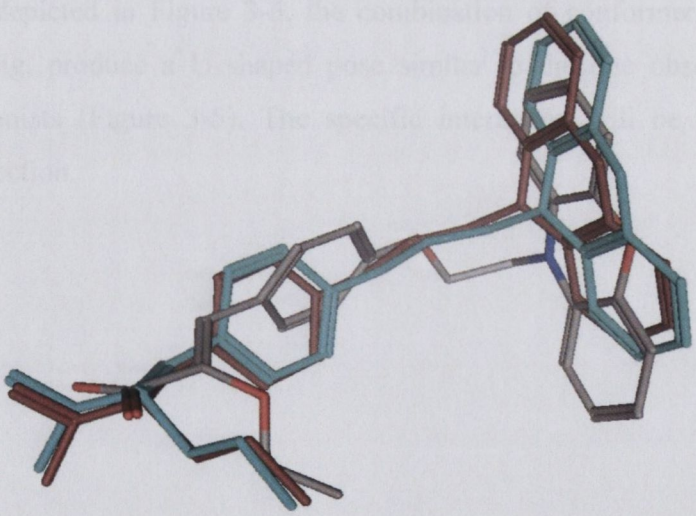


Figure 3-5 Superimposition of the original Ragaglitazar as in the PDB entry 1NYX (red), the re-docked using FRED (cyan) and the closest conformers created and subsequently docked (gray).

3.9.2 Screening

SPECS, Maybridge and Asinex Platinum commercial collections were virtually screened for potential PPARγ agonists following the approach described in the computational method section. The highest molecules ranked were ordered and tested for *in vitro* activity using the FP assay (20 from Specs, 10 from Maybridge and Asinex for a total of 40 molecules).

Initial *in vitro* screening was performed using a single point concentration (100 μM final compound concentration in the well). Compounds exhibiting activity at that concentration were taken forward on for further examination with a six point concentration FP assay.

Out of 40 compounds, two structures exhibited single point activity and six point concentration analysis revealed for one of them (compound internal ID 79 – Maybridge ID: HTS10669) an IC_{50} of about 7 μM. Figure 3-7 shows the graph obtained from the FP assay.

Figure 3-6 shows the binding pose predicted by the docking of compound 79 into the PPARγ LBD.

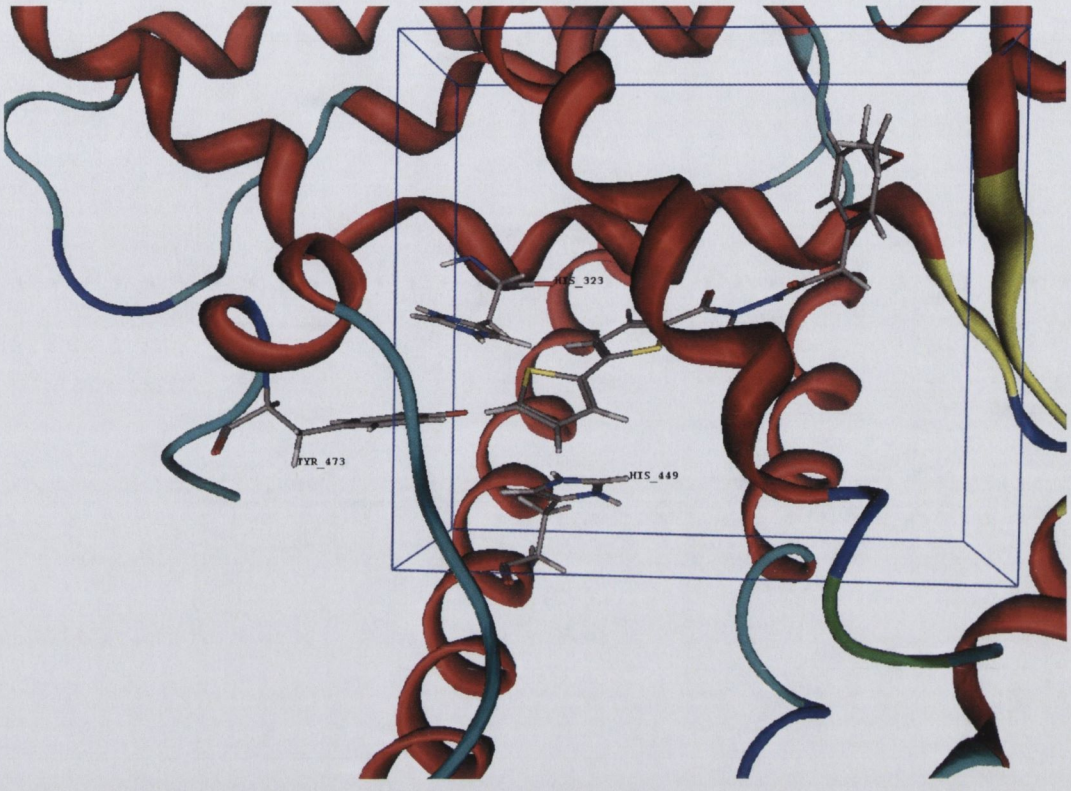
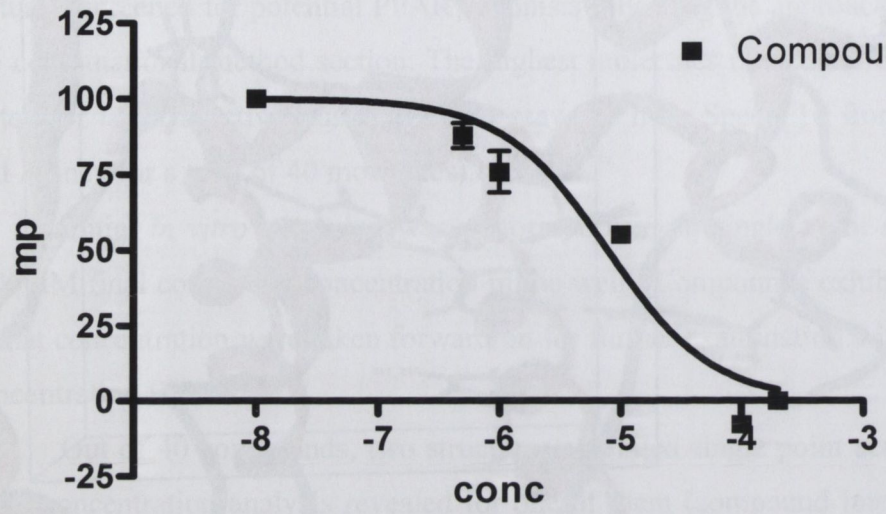


Figure 3-6 Docked structure of compound 79 into PPARγ LBD.

As depicted in Figure 3-6, the combination of conformers generation and rigid docking, produce a U shaped pose similar to the one observed for known PPARγ agonists (Figure 3-5). The specific interaction will be discussed in the following section.



IC_{50}: 7.8017e-006 µM

IC_{50} Standard Error: 4.7803e-006 to 1.2733e-005 µM

R^2 0.9311

Type:	**PPARγ Polarisation Assay**
Cmp ID	79
SupplierName	Maybridge
Positive Control	1µM Rosigitazone
Vehicle	DMSO
Concentration	Six concentration duplicate
Activity (% vehicle)	~7 µM (IC_{50})
R^2	<1
Rosiglitazone IC_{50}	1.6 nM

Figure 3-7 FP binding assay curve of compound 79. An IC_{50} of 7 µM was obtained. Test was run twice in duplicate.(MP: millipolarization units)

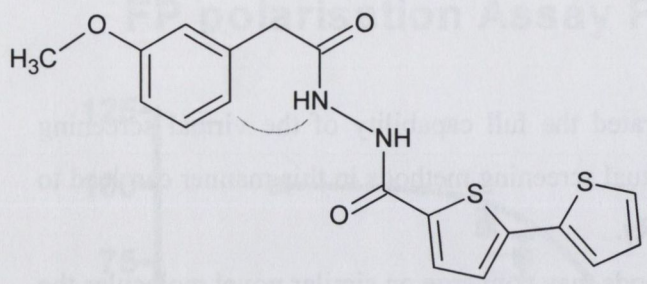
3.10 Discussion

In this chapter we demonstrated the full capability of the virtual screening process. It is clear that utilising virtual screening methods in this manner can lead to the retrieval of a novel active PPARγ.

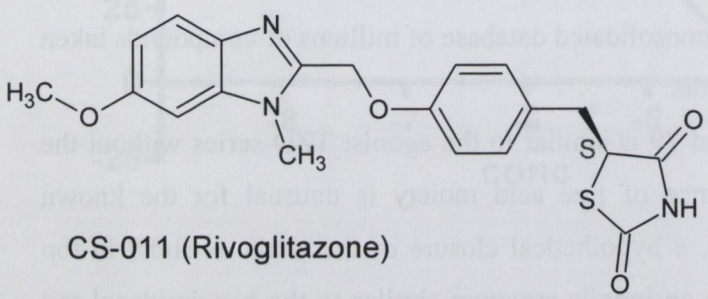
Although current vHTS methods may converge on similar novel molecules the VS protocol permits retrieval of these compounds in a fraction of the time.

In particular, in this work a novel potent ligand with an IC_{50} around 7 μM was successfully identified from a consolidated database of millions of compounds taken from number of database vendors.

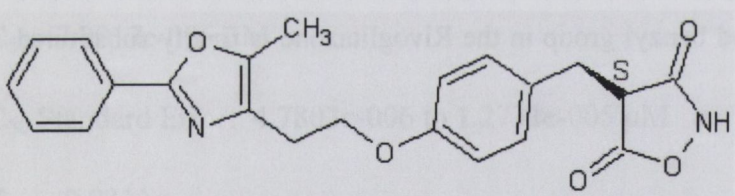
The "head" of compound 79 is similar to the agonist TZD series without the decoration although the absence of free acid moiety is unusual for the known PPARγ agonists. Furthermore, a hypothetical closure of the chain as show in top part Figure 3-8 would lead to an indolic structure similar to the benzimidazol tail present in the Rivoglitazone (Figure 3-8). The link in case of compound 79 is represented by a secondary amine, in the Rivoglitazone the related link is an ether group. The para-substituted benzyl group in the Rivoglitazone is finally substituted with a thiophene group.



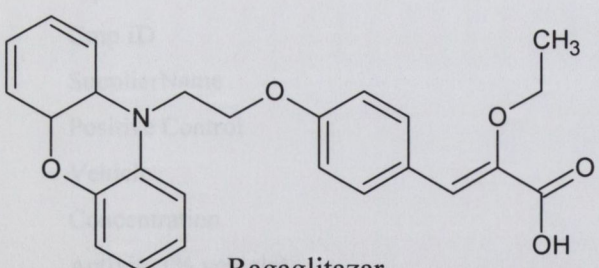
Compound 79



CS-011 (Rivoglitazone)



JTT-501



Ragaglitazar

Figure 3-8 Comparison between Compound 79, Rivoglitazone (a selective agonist), JTT-501 (dual agonist PPARγ and PPARα.) and Ragaglitazar (ligand in PDB 1NYX).

As for Rivoglitazone, similar features are exhibit by the dual agonist JTT-501 with a further modification of the tail section and a methyl group in the azo-oxy ring which allegedly confer the dual agonism to the molecule as discussed previously in the ligand section of this chapter.

Those structural similarities lead us to predict that compound 79 should act as an agonist against PPARγ, although the different "linkers" could lead to some activity in the PPARα or to a dual activity α/γ.

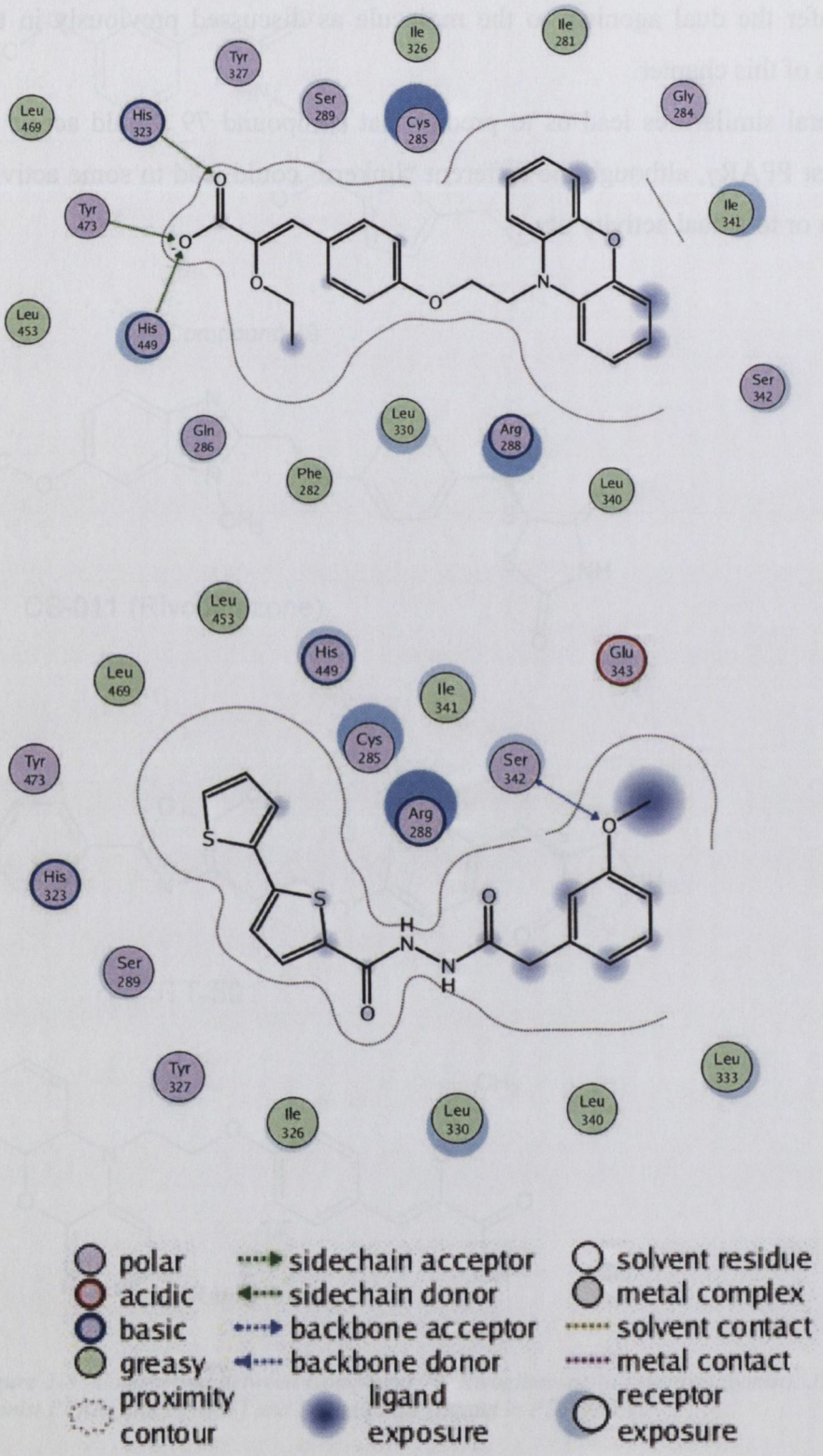


Figure 3-9 Ligand- Receptor Interactions for the Ragaglitazar - 1NYX crystal structure (top). Putative Compound 79- Receptor Interaction of the binding mode predicted by FRED (bottom)

Table 3-11 and Figure 3-9 describe the interactions that compound 79 and Ragaglitazar share with the receptor. The binding mode is similar to the one adopted by Ragaglitazar, with the distal thiophene head of compound 79 interacting with the AF-2 region of the LBD.

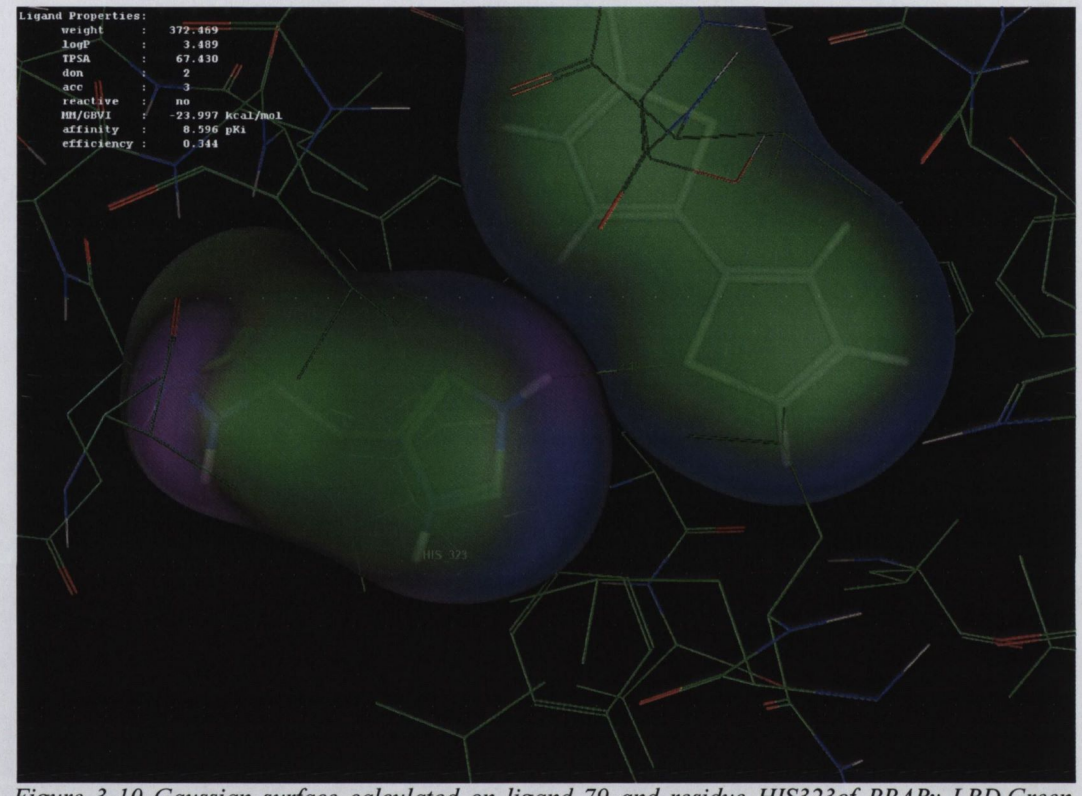


Figure 3-10 Gaussian surface calculated on ligand 79 and residue HIS323of PPARγ LBD.Green surface: hydrophobic area; purple: HBs; blue: polar area. The distance between HIS323 and the sulphur atom in compound 79 is 3.19Å.

As shown in Figure 3-10, polar interactions calculated with MOE, occur between the sulphur of the distal thiophene ring and HIS323, even though the pose predicted by FRED does not place those atoms close enough (3.19Å) and with the correct geometry. As shown in the picture, the polar surface (in blue) of compound 79 matches the polar surface calculated over the HIS323.

The absence of hydrogen bonds between compound 79 and the receptor in this hydrophilic pocket, leads us to the assumption that this compound has been prioritised by ChemGauss2 mainly due to ligand-pocket shape complementarily.

Interaction Site	Residue	Index
Site P1	TYR	473
	HIS	323
	SER	289
	TYR	327
	HIS	449
Site P2	ARG	288
	ILE	326
	LEU	330
	CYS	285
Site P4	LEU	340
	SER	342
	ILE	341
Site C1	LEU	453
Site F	LEU	469

Table 3-11 Neighbouring amino acids to Ragaglitazar and Compound 79

Table 3-11 lists the common residues in the PPARγ LBD which compound 79 and Ragaglitazar are buried within.

As described in Table 3-11, three different "hot spots" can be recognised. Site P1 represents the interaction with helix 12 made by the *head* of the ligand. Decorations of the distal thiophene of compound 79 with carboxylic or hydroxylic functions should strengthen the polar interactions with those residues and increase activity (Figure 3-11).

Figure 3-11 Site P1 highlighted with a Gaussian surface; left Ragaglitazar (ligand in PDB 1NYX). Right: compound 79 as predicted by the docking.

The molecule *linker* in the actives is buried between helix H3 and H5 (site P2) and in both cases it again exhibits interactions with similar AA (Figure 3-12).

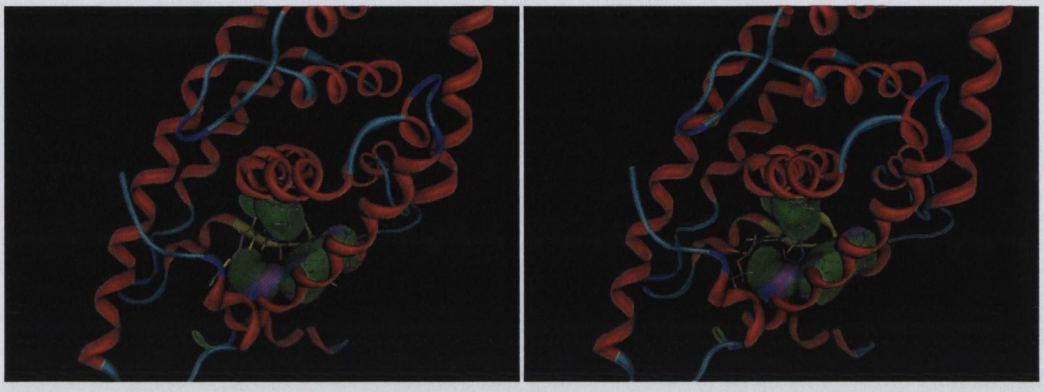


Figure 3-12 Site P2 highlighted with a Gaussian surface; left: Ragaglitazar (ligand in PDB 1NYX); Right: compound 79 as predicted by the docking.

Site P4 clusters together the hydrophobic AAs close to the pocket entrance and close to the distal-end group of the agonists. Hydrophobic contacts are probably established between the benzyl group of compounds 79 and those AA (Figure 3-13).

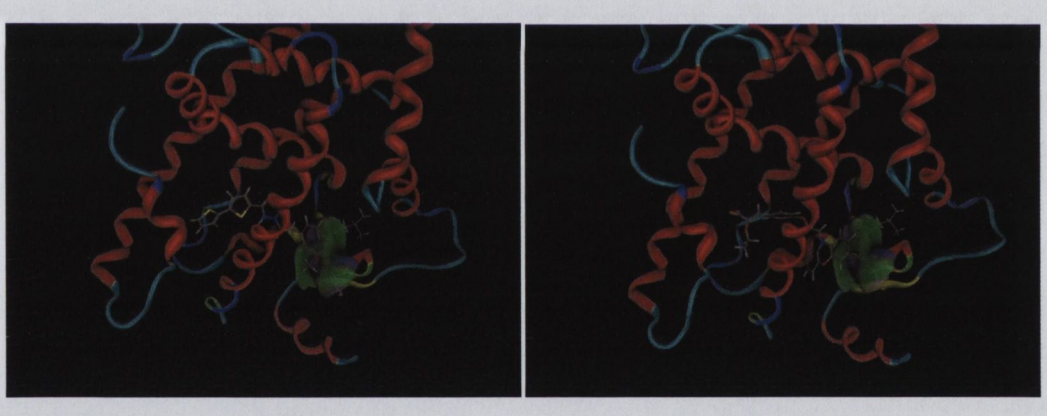


Figure 3-13 Site P4 highlighted with a Gaussian surface; left Ragaglitazar (ligand in PDB 1NYX). Right: compound 79 as predicted by the docking.

To establish if compound 79 was selective for only PPARγ in the context of NR family, binding affinity was also established for the GR, Androgen Receptor and Estrogen Receptor ($ER_{\alpha/\beta}$). Single points concentration (80 μM) showed activity against GR and a multiple points assay identified an IC_{50} for ER_{α} and ER_{β} of 5.8 μM and 5.7 μM respectively.

3.11 Future Work

Compound 79 exhibited potent on target binding to PPARγ, however it is only a first step towards developing a novel therapeutic. As it was identified through the screening of a 3rd party vendor database it is devoid of lead optimisation. Putative interactions with the receptor offer much room for optimisation through the introduction of substituents on the thiophene group in order to increase the interaction with the residues located in proximity to the AF-2 helix. Outside the scope of this work, vendor databases should be screened using similarity and substructure searches in order to find commercially available analogues. SAR data generated through this work will aid in focussing the design of synthetically feasible analogues.

Based on the results from the Chemgauss2 scoring function a potential lead optimisation stage should also be performed using the protocol described in Chapter 5. Further *in vitro* assays will be performed on the LNCap cell line that expresses PPARγ which would give an indication as to the agonistic nature of compound 79. Compound 79 has also already been submitted to the National Cancer Institute to investigate the potential chemotherapeutic effect of compounds 79 in 60 different cell lines.

In vivo tests in mice should also be performed on the optimised 'hit' in order to assess the ADME properties of compound 79 (e.g. real plasma/BBB partition coefficient).

Dual activity against both PPARγ and PPARα must also be investigated as the dual activity may be a key factor for the future pharmacological development of compound 79.

3.12 References

1. Gronemeyer H, Gustafsson JA, Laudet V. Principles for modulation of the nuclear receptor superfamily. Nat Rev Drug Discov 2004;3(11):950-964.
2. Aranda A, Pascual A. Nuclear hormone receptors and gene expression. Physiol Rev 2001;81(3):1269-1304.
3. Steinmetz AC, Renaud JP, Moras D. Binding of ligands and activation of transcription by nuclear receptors. Annu Rev Biophys Biomol Struct 2001;30:329-359.
4. Laudet V, Gronemeyer H. The Nuclear Receptors Factbooks. Academic, San Diego 2001.
5. Gronemeyer H, Laudet V. Transcription factors 3: nuclear receptors. Protein Profile 1995;2(11):1173-1308.
6. Robinson-Rechavi M, Carpentier AS, Duffraisse M, Laudet V. How many nuclear hormone receptors are there in the human genome? Trends Genet 2001;17(10):554-556.
7. Wu Z, Xie Y, Morrison RF, Bucher NL, Farmer SR. PPARgamma induces the insulin-dependent glucose transporter GLUT4 in the absence of C/EBPalpha during the conversion of 3T3 fibroblasts into adipocytes. J Clin Invest 1998;101(1):22-32.
8. Morrison RF, Farmer SR. Hormonal signaling and transcriptional control of adipocyte differentiation. J Nutr 2000;130(12):3116S-3121S.
9. Wang YX, Lee CH, Tiep S, Yu RT, Ham J, Kang H, Evans RM. Peroxisome-proliferator-activated receptor delta activates fat metabolism to prevent obesity. Cell 2003;113(2):159-170.
10. Timothy M. Willson PJB, Daniel D. Sternbach, and Brad R. Henke. The PPARs: From Orphan Receptors to Drug Discovery†. Journal of Medicinal Chemistry, 2000;43:527-550.
11. Lissau I. Overweight and obesity epidemic among children. Answer from European countries. Int J Obes Relat Metab Disord 2004;28 Suppl 3:S10-15.
12. Diabetes: The Policy Puzzle: Towards Benchmarking in the EU25. International Diabetes Federation European Region - Federation of European Nurses in Europe.
13. Lazar MA. PPAR gamma, 10 years later. Biochimie 2005;87(1):9-13.
14. Hoekstra BGSaWJ. Peroxisome Proliferator-Activated Receptors (PPARs): Choreographers of Metabolic Gene Transcription. SIMGA RBI.
15. Bishop-Bailey D, Wray J. Peroxisome proliferator-activated receptors: a critical review on endogenous pathways for ligand generation. Prostaglandins Other Lipid Mediat 2003;71(1-2):1-22.
16. Shouwei Han RKW, and Neil Sidell. Differentiation of Human Neuroblastoma by Phenylacetate Is Mediated by Peroxisome Proliferator-activated Receptor g1. CANCER RESEARCH 2001;61:3998–4002,.
17. Tontonoz P, Singer S, Forman BM, Sarraf P, Fletcher JA, Fletcher CD, Brun RP, Mueller E, Altiok S, Oppenheim H, Evans RM, Spiegelman BM. Terminal differentiation of human liposarcoma cells induced by ligands for

peroxisome proliferator-activated receptor gamma and the retinoid X receptor. Proc Natl Acad Sci U S A 1997;94(1):237-241.

18. Sarraf P, Mueller E, Jones D, King FJ, DeAngelo DJ, Partridge JB, Holden SA, Chen LB, Singer S, Fletcher C, Spiegelman BM. Differentiation and reversal of malignant changes in colon cancer through PPARgamma. Nat Med 1998;4(9):1046-1052.
19. Tsubouchi Y, Sano H, Kawahito Y, Mukai S, Yamada R, Kohno M, Inoue K, Hla T, Kondo M. Inhibition of human lung cancer cell growth by the peroxisome proliferator-activated receptor-gamma agonists through induction of apoptosis. Biochem Biophys Res Commun 2000;270(2):400-405.
20. Inoue K, Kawahito Y, Tsubouchi Y, Yamada R, Kohno M, Hosokawa Y, Katoh D, Bishop-Bailey D, Hla T, Sano H. Expression of peroxisome proliferator-activated receptor (PPAR)-gamma in human lung cancer. Anticancer Res 2001;21(4A):2471-2476.
21. Butler R, Mitchell SH, Tindall DJ, Young CY. Nonapoptotic cell death associated with S-phase arrest of prostate cancer cells via the peroxisome proliferator-activated receptor gamma ligand, 15-deoxy-delta12,14-prostaglandin J2. Cell Growth Differ 2000;11(1):49-61.
22. Mueller E, Sarraf P, Tontonoz P, Evans RM, Martin KJ, Zhang M, Fletcher C, Singer S, Spiegelman BM. Terminal differentiation of human breast cancer through PPAR gamma. Mol Cell 1998;1(3):465-470.
23. Lee CH, Olson P, Evans RM. Minireview: lipid metabolism, metabolic diseases, and peroxisome proliferator-activated receptors. Endocrinology 2003;144(6):2201-2207.
24. Rosen ED, Spiegelman BM. Molecular regulation of adipogenesis. Annu Rev Cell Dev Biol 2000;16:145-171.
25. Flier JS. Clinical review 94: What's in a name? In search of leptin's physiologic role. J Clin Endocrinol Metab 1998;83(5):1407-1413.
26. Gregoire FM, Smas CM, Sul HS. Understanding adipocyte differentiation. Physiol Rev 1998;78(3):783-809.
27. Mrak RE, Landreth GE. PPARgamma, neuroinflammation, and disease. J Neuroinflammation 2004;1(1):5.
28. Sundararajan S, Landreth GE. Antiinflammatory properties of PPARgamma agonists following ischemia. Drug News Perspect 2004;17(4):229-236.
29. Schulz JB, Falkenburger BH. Neuronal pathology in Parkinson's disease. Cell Tissue Res 2004;318(1):135-147.
30. Teismann P, Schulz JB. Cellular pathology of Parkinson's disease: astrocytes, microglia and inflammation. Cell Tissue Res 2004;318(1):149-161.
31. Kliewer SA, Lenhard JM, Willson TM, Patel I, Morris DC, Lehmann JM. A prostaglandin J2 metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation. Cell 1995;83(5):813-819.
32. Brooks DA, Etgen GJ, Rito CJ, Shuker AJ, Dominianni SJ, Warshawsky AM, Ardecky R, Paterniti JR, Tyhonas J, Karanewsky DS, Kauffman RF, Broderick CL, Oldham BA, Montrose-Rafizadeh C, Winneroski LL, Faul MM, McCarthy JR. Design and synthesis of 2-methyl-2-[4-(2-[5-methyl-2-

aryloxazol-4-yl]ethoxy)phenoxy]propionic acids: a new class of dual PPARalpha/gamma agonists. J Med Chem 2001;44(13):2061-2064.

33. Gampe RT, Jr., Montana VG, Lambert MH, Miller AB, Bledsoe RK, Milburn MV, Kliewer SA, Willson TM, Xu HE. Asymmetry in the PPARgamma/RXRalpha crystal structure reveals the molecular basis of heterodimerization among nuclear receptors. Mol Cell 2000;5(3):545-555.
34. Nolte RT, Wisely GB, Westin S, Cobb JE, Lambert MH, Kurokawa R, Rosenfeld MG, Willson TM, Glass CK, Milburn MV. Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. Nature 1998;395(6698):137-143.
35. T. Y. Jpn Kokai Tokkyo Koho 1985;C07D417-12:29 pp.
36. Leff T. RJE. The Antidiabetic PPAR gamma Ligands: An Update on Compounds in Development. Curr Med Chem - Imun, Endoc & Metab Agents 2002;2:33-47.
37. Balfour JA, Plosker GL. Rosiglitazone. Drugs 1999;57(6):921-930; discussion 931-922.
38. Wolffenbuttel BH, Sels JP, Huijberts MS. Rosiglitazone. Expert Opin Pharmacother 2001;2(3):467-478.
39. Sohda T, Momose Y, Meguro K, Kawamatsu Y, Sugiyama Y, Ikeda H. Studies on antidiabetic agents. Synthesis and hypoglycemic activity of 5-[4-(pyridylalkoxy)benzyl]-2,4-thiazolidinediones. Arzneimittelforschung 1990;40(1):37-42.
40. Ikeda H, Taketomi S, Sugiyama Y, Shimura Y, Sohda T, Meguro K, Fujita T. Effects of pioglitazone on glucose and lipid metabolism in normal and insulin resistant animals. Arzneimittelforschung 1990;40(2 Pt 1):156-162.
41. Tontonoz P, Hu E, Spiegelman BM. Stimulation of adipogenesis in fibroblasts by PPAR gamma 2, a lipid-activated transcription factor. Cell 1994;79(7):1147-1156.
42. Furnsinn C, Brunmair B, Meyer M, Neschen S, Furtmuller R, Roden M, Kuhnle HF, Nowotny P, Schneider B, Waldhausl W. Chronic and acute effects of thiazolidinediones BM13.1258 and BM15.2054 on rat skeletal muscle glucose metabolism. Br J Pharmacol 1999;128(6):1141-1148.
43. Vikramadithyan RK, Chakrabarti R, Misra P, Premkumar M, Kumar SK, Rao CS, Ghosh A, Reddy KN, Uma C, Rajagopalan R. Euglycemic and hypolipidemic activity of PAT5A: a unique thiazolidinedione with weak peroxisome proliferator activated receptor gamma activity. Metabolism 2000;49(11):1417-1423.
44. Parks DJ, Tomkinson NC, Villeneuve MS, Blanchard SG, Willson TM. Differential activity of rosiglitazone enantiomers at PPAR gamma. Bioorg Med Chem Lett 1998;8(24):3657-3658.
45. Buckle DR, Cantello BCC, Cawthorne MA, Coyle PJ, Dean DK, Faller A, Haigh D, Hindley RM, Jefcott LJ, Lister CA, Buckle DR, Cantello BCC, Cawthorne MA, Coyle PJ, Dean DK, Faller A, Haigh D, Hindley RM, Jefcott LJ, Lister CA. Non thiazolidendione antihyerglycemic agents. 2: alpha carbon substituted beta-phenilpropanoic acids. Bioorg Med Chem Lett 1996;6(17):2127-2130.

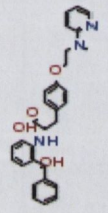
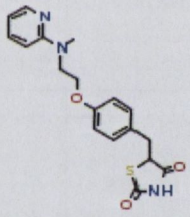
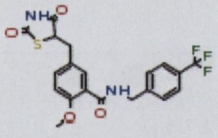
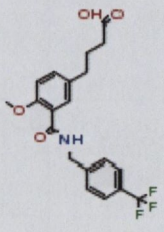
46. Buckle DR, Cantello BCC, Cawthorne MA, Coyle PJ, Dean DK, Faller A, Haigh D, Hindley RM, Jefcott LJ, Lister CA, Buckle DR, Cantello BCC, Cawthorne MA, Coyle PJ, Dean DK, Faller A, Haigh D, Hindley RM, Jefcott LJ, Lister CA. Non thiazolidendione antihyerglycemic agents. 1: alpha heteroatom substituted beta-phenilpropanoic acids. Bioorg Med Chem Lett 1996;6(17):2121-2126.
47. Cronet P, Petersen JF, Folmer R, Blomberg N, Sjoblom K, Karlsson U, Lindstedt EL, Bamberg K. Structure of the PPARalpha and -gamma ligand binding domain in complex with AZ 242; ligand selectivity and agonist activation in the PPAR family. Structure (Camb) 2001;9(8):699-706.
48. Collins JL, Blanchard SG, Boswell GE, Charifson PS, Cobb JE, Henke BR, Hull-Ryde EA, Kazmierski WM, Lake DH, Leesnitzer LM, Lehmann J, Lenhard JM, Orband-Miller LA, Gray-Nunez Y, Parks DJ, Plunkett KD, Tong WQ. N-(2-Benzoylphenyl)-L-tyrosine PPARgamma agonists. 2. Structure-activity relationship and optimization of the phenyl alkyl ether moiety. J Med Chem 1998;41(25):5037-5054.
49. Nomura M, Tanase T, Ide T, Tsunoda M, Suzuki M, Uchiki H, Murakami K, Miyachi H. Design, synthesis, and evaluation of substituted phenylpropanoic acid derivatives as human peroxisome proliferator activated receptor activators. Discovery of potent and human peroxisome proliferator activated receptor alpha subtype-selective activators. J Med Chem 2003;46(17):3581-3599.
50. Xu HE, Lambert MH, Montana VG, Plunket KD, Moore LB, Collins JL, Oplinger JA, Kliewer SA, Gampe RT, Jr., McKee DD, Moore JT, Willson TM. Structural determinants of ligand binding selectivity between the peroxisome proliferator-activated receptors. Proc Natl Acad Sci U S A 2001;98(24):13919-13924.
51. Lohray BB, Lohray VB, Bajji AC, Kalchar S, Poondra RR, Padakanti S, Chakrabarti R, Vikramadithyan RK, Misra P, Juluri S, Mamidi NV, Rajagopalan R. (-)3-[4-[2-(Phenoxazin-10-yl)ethoxy]phenyl]-2-ethoxypropanoic acid [(-)DRF 2725]: a dual PPAR agonist with potent antihyperglycemic and lipid modulating activity. J Med Chem 2001;44(16):2675-2678.
52. Guo Q, Sahoo SP, Wang PR, Milot DP, Ippolito MC, Wu MS, Baffic J, Biswas C, Hernandez M, Lam MH, Sharma N, Han W, Kelly LJ, MacNaul KL, Zhou G, Desai R, Heck JV, Doebber TW, Berger JP, Moller DE, Sparrow CP, Chao YS, Wright SD. A novel peroxisome proliferator-activated receptor alpha/gamma dual agonist demonstrates favorable effects on lipid homeostasis. Endocrinology 2004;145(4):1640-1648.
53. Sauerberg P, Pettersson I, Jeppesen L, Bury PS, Mogensen JP, Wassermann K, Brand CL, Sturis J, Woldike HF, Fleckner J, Andersen AS, Mortensen SB, Svensson LA, Rasmussen HB, Lehmann SV, Polivka Z, Sindelar K, Panajotova V, Ynddal L, Wulff EM. Novel tricyclic-alpha-alkyloxyphenylpropionic acids: dual PPARalpha/gamma agonists with hypolipidemic and antidiabetic activity. J Med Chem 2002;45(4):789-804.

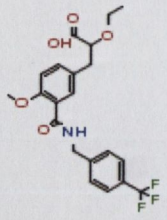
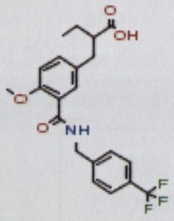
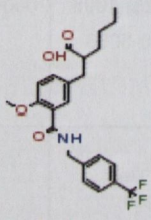
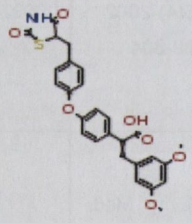
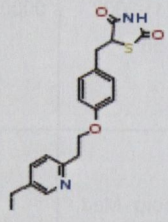
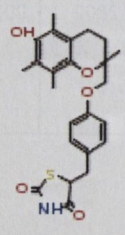
54. Jungling E, Kammermeier H. A one-vial method for routine extraction and quantification of free fatty acids in blood and tissue by HPLC. Anal Biochem 1988;171(1):150-157.
55. Nagy L, Tontonoz P, Alvarez JG, Chen H, Evans RM. Oxidized LDL regulates macrophage gene expression through ligand activation of PPARgamma. Cell 1998;93(2):229-240.
56. Forman BM, Tontonoz P, Chen J, Brun RP, Spiegelman BM, Evans RM. 15-Deoxy-delta 12, 14-prostaglandin J2 is a ligand for the adipocyte determination factor PPAR gamma. Cell 1995;83(5):803-812.
57. Xu HE, Lambert MH, Montana VG, Parks DJ, Blanchard SG, Brown PJ, Sternbach DD, Lehmann JM, Wisely GB, Willson TM, Kliewer SA, Milburn MV. Molecular recognition of fatty acids by peroxisome proliferator-activated receptors. Mol Cell 1999;3(3):397-403.
58. Jonas Uppenberg CS, Maria Jaki, Go¨ran Bertilssoni, Lena Jendeberg, and Anders Berkenstam. Crystal Structure of the Ligand Binding Domain of the Human Nuclear Receptor PPARg. THE JOURNAL OF BIOLOGICAL CHEMISTRY 1998;273(47).
59. H.M. Berman JW, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Research 2000;28 235-242
60. Sheu SH, Kaya T, Waxman DJ, Vajda S. Exploring the binding site structure of the PPAR gamma ligand-binding domain by computational solvent mapping. Biochemistry 2005;44(4):1193-1209.
61. Yu C, Chen L, Luo H, Chen J, Cheng F, Gui C, Zhang R, Shen J, Chen K, Jiang H, Shen X. Binding analyses between Human PPARgamma-LBD and ligands. Eur J Biochem 2004;271(2):386-397.
62. Shu-Hsien Sheu TK, David J. Waxman, and Sandor Vajda. Exploring the Binding Site Structure of the PPARgamma Ligand-Binding Domain by Computational Solvent Mapping. Biochemistry 2005;44(4):1193-1209.
63. Lu IL, Huang CF, Peng YH, Lin YT, Hsieh HP, Chen CT, Lien TW, Lee HJ, Mahindroo N, Prakash E, Yueh A, Chen HY, Goparaju CM, Chen X, Liao CC, Chao YS, Hsu JT, Wu SY. Structure-based drug design of a novel family of PPARgamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. J Med Chem 2006;49(9):2703-2712.
64. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15(4):327-332.
65. Pipeline. Pilot - a software by scigetic -;www.scitegic.com.
66. WOMBAT. WOrld of Molecular BioAcTivity - Sunset Molecular Discovery LLC.http://sunsetmolecular.com.
67. SPECS. Chemistry Solutions for drug discovery.http://www.secs.net.
68. Maybridge. Thermo Fisher Scientific.http://www.maybridge.com.
69. Egan WJ, Merz KM, Jr., Baldwin JJ. Prediction of drug absorption using multivariate statistics. J Med Chem 2000;43(21):3867-3877.
70. CORINA 3.6. distributed by Molecular Networks GmbH
71. OMEGA 1.8.1. distributed by Openeye Scientific Software

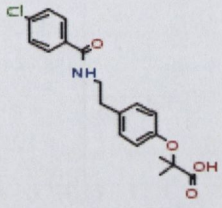
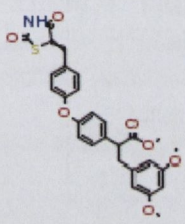
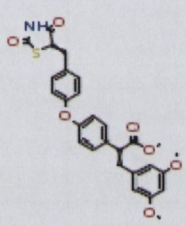
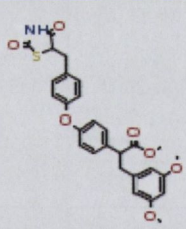
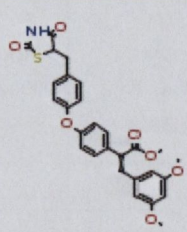
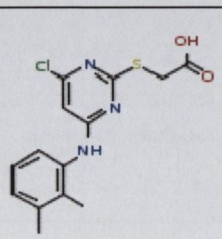
72. FRED (version 2.0.1). developed and distributed by Openeye Scientific Software.(URL:http://www.eyesopen.com).
73. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem 2004;25(9):1157-1174.
74. J.A. Grant BTP, A Nicholls. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. J Comput Chem 2001;22(6):608-640.

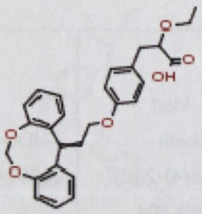
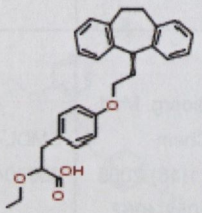
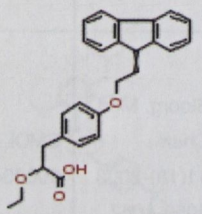
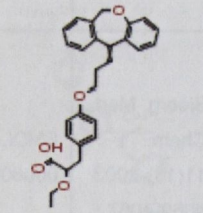
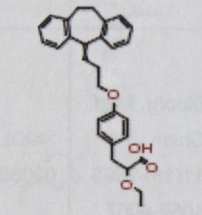
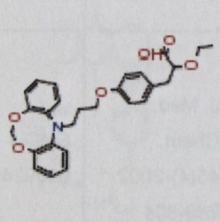
3-Ap. Appendix to Chapter 3

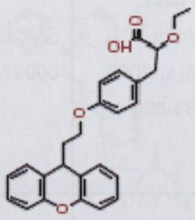
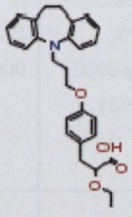
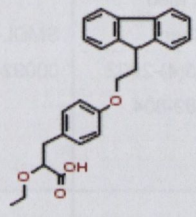
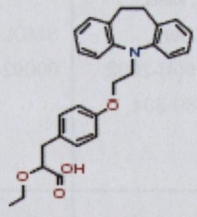
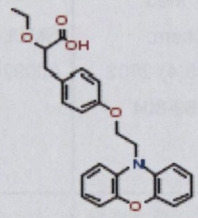
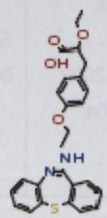
List of PPARγ agonists from WOMBAT

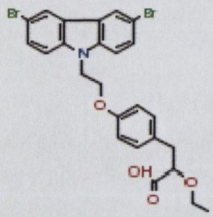
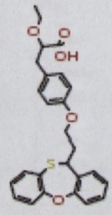
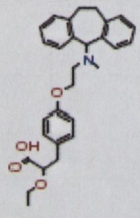
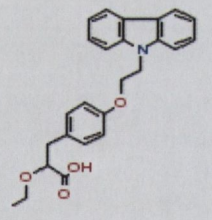
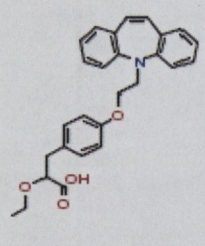
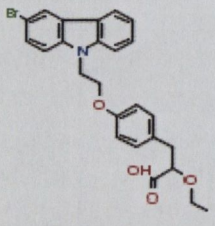
Molecule	Activity_type	Activity_value	Activity_species	References	Name
	EC50	6.3500	human	J. Med. Chem. 41(25)-1998 5020-5036	SMDL-00048608
	EC50	5.3872	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092465
	EC50	6.0000	human	Bioorg. Med. Chem. Lett. 12(1)-2002 77-80	SMDL-00000198
	EC50	5.6576	human	Bioorg. Med. Chem. Lett. 12(1)-2002 77-80	SMDL-00000200

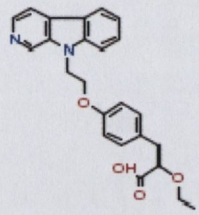
	EC50	5.7959	human	Bioorg. Med. Chem. Lett. 12(1)-2002 77-80	SMDL-00000207
	EC50	7.3979	human	Bioorg. Med. Chem. Lett. 12(1)-2002 77-80	SMDL-00000202
	EC50	6.0000	human	Bioorg. Med. Chem. Lett. 12(1)-2002 77-80	SMDL-00000205
	EC50	6.1612	human	Bioorg. Med. Chem. 11(18)-2003 4059-4067	SMDL-00080484
	EC50	5.1752	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092463
	EC50	6.0088	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092464

	EC50	4.4473	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092462
	EC50	4.6216	human	Bioorg. Med. Chem. 11(18)-2003 4059-4067	SMDL-00080485
	EC50	5.9446	human	Bioorg. Med. Chem. 11(18)-2003 4059-4067	SMDL-00080482
	EC50	4.2388	human	Bioorg. Med. Chem. 11(18)-2003 4059-4067	SMDL-00080486
	EC50	5.4334	human	Bioorg. Med. Chem. 11(18)-2003 4059-4067	SMDL-00080487
	EC50	4.8996	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092461

	EC50	5.4685	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092447
	EC50	4.8996	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092444
	EC50	4.9031	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092449
	EC50	4.9066	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092445
	EC50	4.7773	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092443
	EC50	5.9208	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092451

	EC50	5.0177	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092448
	EC50	6.1192	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092452
	EC50	4.9318	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092450
	EC50	6.1249	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092453
	EC50	5.4935	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092460
	EC50	5.1249	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092442

	EC50	6.6576	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092459
	EC50	5.0458	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092446
	EC50	5.7696	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092441
	EC50	6.4437	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092457
	EC50	5.0223	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092454
	EC50	5.6198	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092458

	EC50	6.8539	human	J. Med. Chem. 45(4)-2002 789-804	SMDL-00092456

OMEGA2.0 list of parameters

```
      #Interface settings
      #-pvmconf (Not set, no default)

      #File Options :
          -commentEnergy  true (default)
          #-in ../my_validation_set.sdf
          #-log (Not set, no default)
          -out  omega_results5.sdf
          #-param (Not set, no default)
          -prefix  ppar_ (default)
          -rotorOffsetCompress  false (default)
          -sdEnergy  true (default)
          -verbose  false (default)
          -warts  true (default)

      #3D Construction Parameters :
          -buildff  mmff94s_NoEstat (default)
          -canonOrder  true (default)
          -deleteFixHydrogens  true (default)
          #-fixfile (Not set, no default)
          -fixrms  0.150000 (default)
          -fraglib
/project/software/openeye/omega2.0/openeye/data/omega2/fraglib.oeb.
gz
          -fromCT  true (default)
          -maxmatch  10 (default)
          -umatch  true (default)

      #Structure Enumeration :
          -enumNitrogen  true (default)
          -enumRing  true (default)

      #Torsion Driving Parameters :
      #   -erange
          -ewindow  30.000000 (default)
```

```
#     -rangeIncrement  2 (default)
      -rms  0.800000 (default)
      #-rmsrange (Not set, no default)
      -searchff  mmff94s_NoEstat (default)
      -tordrive  true (default)
      #-torlib (Not set, no default)

#PVM Parameters :
      -pvmdebug  false (default)
      -pvmpass  10 (default)
```

FRED list of parameters

```
#Interface settings
# -param
/project/software/scitegic/public/users/cartag/parameters/param_fre
d.txt
#-pvmconf (Not set, no default)

#Input Ligands :
    -dbase  /project/scitegic/web/jobs/cartag/{ppcB6E64-0C88-4A4A-
B6DB-29C65490EDAA}/temp/temp_in4.mol2
    -conftest  none
    -assign_ligand_charges  true
    #-molnames (Not set, no default)

#Active Site :
    -pro
/project/software/scitegic/public/users/cartag/apostructures/1NYX_a
po.mol2
    -box
/project/software/scitegic/public/users/cartag/ligands/1NYX_ligand.
mol2
    -addbox  2.500000
    #-pharm (Not set, no default)
    -assign_protein_charges  false

#Exhaustive Docking :
    -no_dock  false

    #Exhaustive :
        -exhaustive_scoring  chemgauss2
        -rstep  1.500000
        -tstep  1.000000

    #Negative_Image :
        -clash_scale  0.750000
        -neg_img_size  normal

    #Number of poses :
        -num_poses  100
        -sqrt_poses  false
```

```
#Optimization :
    -optimization  plp

#Consensus Pose Selection :

    #Scoring Functions :
        -pose_select_weight_shapegauss  1
        -pose_select_weight_plp  1
        -pose_select_weight_chemgauss  0
        -pose_select_weight_chemgauss2  1
        -pose_select_weight_chemscore  1
        -pose_select_weight_screenscore  1

    #Consensus Method :
        -consensus_struct_by_product  false
        -consensus_struct_by_score  true

    #Alternate Poses :
        -num_alt_poses  0

#Force Field Refinement :
    -discard_refined_poses_with_constraint_violations  true
    -refine  no_refinement

#Scoring :

    #Standard Scoring Functions :
        -shapegauss  true
        -plp  true
        -chemgauss  true
        -chemgauss2  true
        -chemscore  true
        -screenscore  true
        -zapbind  false

    #MASC Corrected Scoring Functions :
        -shapegauss_masc  false
        -plp_masc  false
        -chemgauss_masc  false
        -chemgauss2_masc  false
        -chemscore_masc  false
        -screenscore_masc  false
```

```
        -zapbind_masc  false

#Scoring Function Modifications :

    #Shapegauss :
        -set_shapegauss_gamma  4.500000
        -set_shapegauss_kappa  0.700000

    #PLP :
        -scale_plp_clash  1.000000
        -scale_plp_hb  1.000000
        -scale_plp_metal  1.000000
        -scale_plp_steric  1.000000
        -scale_plp_sulphur  1.000000
        -include_plp_desolvation_penalty  false

    #Chemgauss2 :
        -scale_chemgauss2_aromatic  1.000000
        -scale_chemgauss2_hb  1.000000
        -scale_chemgauss2_metal  1.000000
        -scale_chemgauss2_steric  1.000000

    #Chemscore :
        -scale_chemscore_hb  1.000000
        -scale_chemscore_lipophilic  1.000000
        -scale_chemscore_metal  1.000000
        -scale_chemscore_rb  1.000000
        -set_chemscore_use_oelib_rad  false
        -set_chemscore_use_oelib_typing  false

    #Screenscore :
        -scale_screenscore_ambiguous  1.000000
        -scale_screenscore_aromatic  1.000000
        -scale_screenscore_clash  1.000000
        -scale_screenscore_hb  1.000000
        -scale_screenscore_lipo  1.000000
        -scale_screenscore_metal  1.000000
        -scale_screenscore_plp  1.000000
        -scale_screenscore_rb  1.000000

#Output :
    -prefix  fred
    -oformat  sdf
    -output_scores  true
```

```
-output_structs  true
-serial  false
-sort_by_top_consensus_pose  false

#Scoring Function Hitlists :

    #Shapegauss :
        -shapegauss_size  1000000
        #-shapegauss_cut (Not set, no default)
        -shapegauss_consensus_score_weight  1

    #PLP :
        -plp_size  1000000
        #-plp_cut (Not set, no default)
        -plp_consensus_score_weight  1

    #Chemgauss :
        -chemgauss_size  1000000
        #-chemgauss_cut (Not set, no default)
        -chemgauss_consensus_score_weight  1

    #Chemgauss2 :
        -chemgauss2_size  1000000
        #-chemgauss2_cut (Not set, no default)
        -chemgauss2_consensus_score_weight  1

    #Chemscore :
        -chemscore_size  1000000
        #-chemscore_cut (Not set, no default)
        -chemscore_consensus_score_weight  1

    #Screenscore :
        -screenscore_size  1000000
        #-screenscore_cut (Not set, no default)
        -screenscore_consensus_score_weight  1

    #Zapbind :
        -zapbind_size  1000000
        #-zapbind_cut (Not set, no default)
        -zapbind_consensus_score_weight  1

    #MASC Shapegauss :
```

```
        -shapegauss_masc_size  1000000
        #-shapegauss_masc_cut (Not set, no default)
        -shapegauss_masc_consensus_score_weight  1

    #MASC PLP :
        -plp_masc_consensus_score_weight  1
        -plp_masc_size  1000
        #-plp_masc_cut (Not set, no default)

    #MASC Chemgauss :
        -chemgauss_masc_size  1000000
        #-chemgauss_masc_cut (Not set, no default)
        -chemgauss_masc_consensus_score_weight  1

    #MASC Chemgauss2 :
        -chemgauss2_masc_size  1000000
        #-chemgauss2_masc_cut (Not set, no default)
        -chemgauss2_masc_consensus_score_weight  1

    #MASC Chemscore :
        -chemscore_masc_size  1000000
        #-chemscore_masc_cut (Not set, no default)
        -chemscore_masc_consensus_score_weight  1

    #MASC Screenscore :
        -screenscore_masc_size  1000000
        #-screenscore_masc_cut (Not set, no default)
        -screenscore_masc_consensus_score_weight  1

    #MASC Zapbind :
        -zapbind_masc_size  1000000
        #-zapbind_masc_cut (Not set, no default)
        -zapbind_masc_consensus_score_weight  1

#Consensus Scoring Hitlists :
    -consensus  true
    -consensus_size  1000000
    -consensus_pool_size  1000000
```

```
        #MASC Chemscore :
            -chemscore_masc_size  1000000
            #-chemscore_masc_cut (Not set, no default)
            -chemscore_masc_consensus_score_weight  1

        #MASC Screenscore :
            -screenscore_masc_size  1000000
            #-screenscore_masc_cut (Not set, no default)
            -screenscore_masc_consensus_score_weight  1

        #MASC Zapbind :
            -zapbind_masc_size  1000000
            #-zapbind_masc_cut (Not set, no default)
            -zapbind_masc_consensus_score_weight  1

    #Consensus Scoring Hitlists :
        -consensus  true
        -consensus_size  1000000
        -consensus_pool_size  1000000
```

Competitive binding experiments and FP

Competitive binding assays are experiments in which a biological binding agent competes against labelled or unlabeled compounds for a specific target. The labeling may consist of a fluorescent tracer (as in the fluorescence polarization assay), or a radioactively labelled tracer etc. In order to quantify a physical constant such as the affinity of a ligand for a receptor, various assumptions and simplifications are typically made:

1. The interaction between the receptor (R) and the ligand (L) is reversible.
2. All the receptors molecules are considered as independent and equivalent.
3. The interactions are all measured at equilibrium.
4. Neither the ligand nor the receptor undergoes any degradation during the measured reaction.

Thus, at equilibrium the reaction between L and R can be described as:

$$Ligand_{free} + Receptor_{free} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} Ligand{:}Receptor \qquad 3\text{-}A$$

where k1 and k-1 are respectively the association and the dissociation kinetic constants. As the mass action equation postulates, at equilibrium one can write the following equivalence:

$$K_d = \frac{[Ligand_{free}][Receptor_{free}]}{[Receptor-Ligand]} = \frac{[L_f][R_f]}{[B]} = \frac{k_{-1}}{k_1} \qquad 3\text{-}B$$

which can be rewritten as:

$$K_d = \frac{L_f(R_t - B)}{B} \qquad 3\text{-}C$$

where L_f is the ligand concentration in its unbound state (free), R_t is the total concentration of the receptor in either bound or unbound state and B represents the receptor in its bound states (receptor-ligand). Rearranging Equation 3-C:

$$\frac{L_f R_t - L_f B}{B} = K_d$$

$$\Rightarrow \quad BK_d + BL_f = L_f R_t$$

$$\Rightarrow \quad B\frac{(K_d + L_f)}{(K_d + L_f)} = \frac{L_f R_t}{(K_d + L_f)}$$

$$\Rightarrow B = \frac{L_f R_t}{K_d + L_f}$$

$$\Rightarrow \frac{B}{R_t} = \frac{L_f}{(K_d + L_f)} \qquad 3\text{-}D$$

K_d can be also depicted as the concentration of free ligand at which 50% of the receptor sites are occupied as shown in the equation below, obtained by setting B/R_t equal to 0.5 (1/2):

$$\frac{1}{2} = \frac{L_f}{K_d + L_f}$$

$$\Rightarrow \mathbf{0} = \frac{2L_f}{K_d + L_f}$$

$$\Rightarrow K_d + L_f = 2L_f$$

$$\Rightarrow K_d = L_f \qquad 3\text{-}E$$

$$\Rightarrow \frac{B}{R_t} = \frac{L_f}{(K_d + L_f)} \qquad 3\text{-}D$$

is an equation for a rectangular hyperbola with a horizontal asymptote at high L_f corresponding to 100% of saturation of the receptor. The ratio B/R_t is also called the fractional occupancy and it varies in the range between 0 and 1 (from fully saturated receptor to completely unbound receptor). Rearrangement of the previous equation yields to the following:

$$\frac{B}{L_f} = \frac{-B}{K_d} + \frac{R_t}{K_d} \qquad 3\text{-}F$$

which is the equation of a line ($y = mx - b$; plotting B/L_f vs. B) where $m = 1/K_d$ and $b = R_t/K_d$. Thus R_t/K_d is the y-intercept, R_t is the x-intercept and $-1/k_d$ is the slope. This equation is represented in the so called scatchard-plot below.

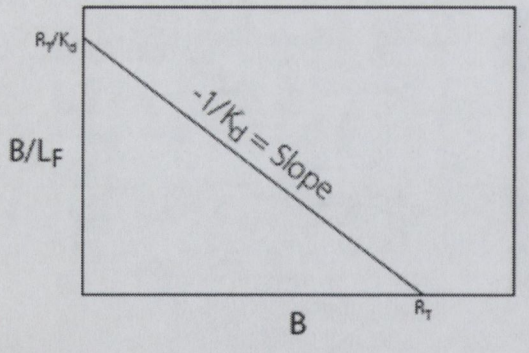


Figure 3-A: Example of Scatchard Plot. See text for details.

The slope of the straight line in a Scatchard plot is related to the dissociation constant of the ligand-receptor complex while the intercept on the abscissa reveals

the number of available receptors. Straight lines indicate that a single affinity class of binding sites is present. Deviations from straight lines are more difficult to interpret and cannot be resolved simply into two or more linear components; nonlinear curves can show the existence of more than one class of binding sites. Another advantage of the scatchard plot is that the R_t and K_d values can be determined by linear regression of Scatchard transformed data.

Analysis of affinity constants in binding experiments assumes that the measurement is performed at equilibrium. This is usually accomplished by incubating a low concentration of ligand with the receptor and following the amount of ligand bound over time (see picture below). The incubation time for most assays is the time required for 90% of the ligand to bind to the receptor (Figure 3-B).

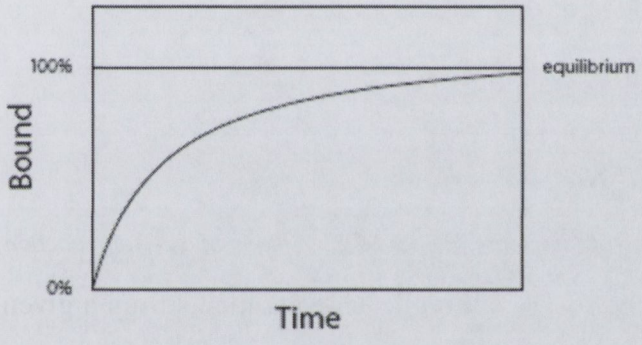


Figure 3-B Example of a time dependent saturation experiment

In most traditional binding experiments, the receptor concentration is kept constant and the ligand concentration is varied. It is equally correct to keep the ligand concentration constant and vary the receptor concentration. In FP, on the other hand, it is the small, labeled ligand that is held constant in addition to the receptor at low concentrations and a much higher concentrated, unlabeled molecule is titrated against it. Thus, a FP competitive binding experiment measures the binding of a single concentration of labeled ligand in the presence of various concentrations of unlabeled ligand. Ideally, the concentration of unlabeled ligand is varied over at least six orders of magnitude. The experiments are done this way in order to vary the labeled ligand from the completely free state (and lowest polarization value) to the

completely bound state (highest polarization value), thus maximizing the dynamic range of the experiment (details about FP principles are given in Chapter 1).

The analysis of the FP results yields a measurement of the IC_{50}, the concentration of substance that provides 50% inhibition of a certain reaction. In this case the binding of the labeled ligand to the receptor. The IC_{50} is the parameter most commonly used to indicate the inhibitory effect of compounds in competitive binding and functional antagonist assays. In similar terms, an EC_{50} parameter is used to describe the effect of compounds in agonist/stimulator assays. The EC_{50} shows the concentration giving 50% of that compounds maximal response.

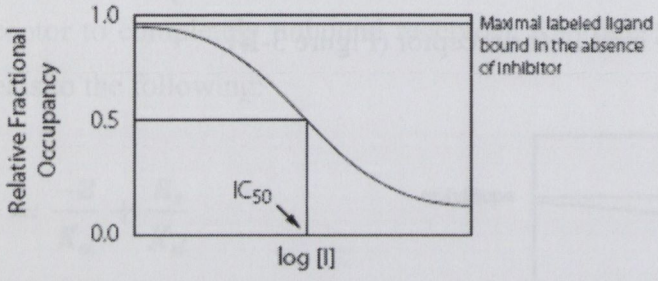


Figure 3-C: The IC_{50}, the concentration of substance that provides 50% inhibition of a certain reaction

Before fitting the curve shown in Figure 3-C, all replicate experiments from a given run (including multiple plates per run) for a given concentration are averaged. Thus only one point per concentration per run is used to fit the dose-response equation to the data. The four parameter logistic model, also called the Hill-Slope model, is the most common equation fitted to *in vitro* dose-response data. One form of the equation is

$$y = bot + \frac{(top - bot)}{1 + (x/IC_{50})^{slope}} \qquad 3\text{-}G$$

where *top* and *bot* are in that order the higher and the lower fractional occupancy, y is the percentage activity and x is the corresponding concentration. The fitted IC_{50} parameter is the relative IC_{50}, and is defined as the concentration giving a response half way between the fitted top and bottom of the curve. One can compare the

relative binding affinities (IC_{50}) of two or more unlabeled ligands or, when the K_d of the labeled ligand is known, calculate the K_i of the unlabeled ligands. K_i is the binding affinity of the inhibitor. Whereas the IC_{50} value for a compound may vary between experiments depending on labeled ligand concentration, the K_i is an absolute value. K_i is the inhibition constant for a drug; it represents the concentration of competing ligand in a competition assay which would occupy 50% of the receptors if no labeled ligand were present.

Obtaining IC_{50} values with Equation 3-G under identical conditions is an excellent approach if the aim of the study is to compare the relative affinities of a series of ligands. Otherwise, in order to compare affinities from different experiments, the K_i must be derived. This can be achieved using the Cheng and Prussof correction:

$$f = \frac{I}{I+K_i(1+L_t/K_d)} \qquad 3\text{-}H$$

Where f is fractional inhibition, L_t is the total concentration of the labeled Ligand, I is the concentration of the unlabeled ligand, K_d is the dissociation constant for the labeled ligand and K_i is the dissociation constant for the unlabeled ligand.

When f=1/2 (50% of occupancy of the receptor), I is equal to the IC_{50} , Equation 3-H can be rearranged in order to calculate the K_i of the unlabeled ligand:

$$\frac{1}{2} = \frac{IC_{50}}{IC_{50} + K_i\left(1 + {}^{L_t}/_{K_d}\right)}$$

$$\Rightarrow K_i = \frac{IC_{50}}{1 + {}^{L_t}/_{K_d}} \qquad 3\text{-}I$$

The great benefit of competitive binding experiments is their ability to determine binding affinities without requiring labeling of the ligand under study. In a typical competition experiment, the tracer is incubated with receptor and various concentrations of the unlabeled competing ligand. As the concentration of unlabeled

ligand increases, it competes with the tracer for receptor binding. As the fraction of tracer that is bound to the receptor drops, there is a corresponding drop in the polarization value. Competitive experiments performed using FP require relatively high receptor concentrations in order to bind significant amounts of the tagged ligand. FP competitive binding experiments should be designed such that the R/K_d ratio is at least 1, so that the polarization value in the absence of the unlabelled compound will represent at least 50% of the maximal shift.

In the case of fluorescence polarization for PPARγ discussed in chapter 3, polarization values were plotted against the concentration of the test compound. Competitive binding experiments were used to analyze the binding affinities of untagged ligands. The concentration of the test compound that results in a shift toward higher polarization values is a measure of the relative affinity of the test compound for the PPARγ ligand binding domain. The fitting of the curve shown in Chapter 3 was performed using the Prism software from GraphPad™ Software, Inc.

4. Unbiasing Scoring Functions: a new re-scoring and normalisation strategy

4.1 Abstract

Ligand bias can contribute significantly to the number of false positives (FP) observed in a virtual screening campaign. Utilising a receptor-based docking approach against a well established target of therapeutic importanceERα, coupled with several common scoring functions taken both individually and as a consensus, we sought to examine the characteristics of molecules retrieved by each. A set of 19 actives and approximately 10,000 decoys trained by a Bayesian classifier to ensure similar ligand properties were docked to PDB entry 3ERT and ranked using ChemGuass, ChemGauss2, ChemScore, ScreenScore, ShapeGauss and PLP. 2D descriptors were calculated for all compounds and plotted versus the rank position of the compounds. In particular, several descriptors displayed a strong correlation using the Spearman coefficient with the ranked position of a compound. All of those descriptors also exhibited a positive correlation with molecular weight and to counteract this we have introduced a new normalisation procedure to reduce the dependence of certain scoring functions on molecular weight. It is also shown that scoring frequency analysis rather than enrichment calculation is a better method for gaining a more complete picture of a docking protocol. Finally, SIFt fingerprint analysis was also employed to highlight the advantages of introducing these normalisation procedures.

Published as:
Unbiasing Scoring Functions: A New Normalization and Rescoring Strategy
Giorgio Carta, Andrew J. S. Knox, and David G. Lloyd
J. Chem. Inf. Model. 2007, 47, 1564-1571

4.2 Introduction

Improvements in crystallography, genomics and proteomics have permitted a deeper understanding of the interaction between proteins and their endogenous ligands. More specifically, in the realm of computational chemistry, development of algorithms and new mathematical approaches to describe the binding between protein and ligand has allowed characterisation of many important interactions such as hydrogen bonding, hydrophobic contacts and vdW interactions. This knowledge is now commonly utilised in processes such as SBVS, whereby the goal is to be able to simulate and accurately predict the docked position of a molecule inside a binding pocket allowing identification of 'binders' from 'non-binders'.

Subsequently each docked compound receives a score equivalent to the free energy of binding, generated from the predicted binding pose. The general assumption is that molecules that bind tightly to the receptor should be given a higher score. Once scored, molecules can be ranked, sorted and the top x% is generally retained for biological evaluation.

A large variety of methods have been described in the literature to calculate and predict *in silico* the binding affinity between a receptor and ligand, however, it is also well known that most fail to accurately predict it[1]. Nowadays, scoring functions can be classified into three main categories: Force-Field, Empirical and Knowledge-Based methods[2]. Empirical and Knowledge-based methods are the most commonly used scoring functions in a VS campaign due to their inherent speed and ease of calculation.

In this work we focused our attention on the empirical method, such as that described originally by Bohm[3,4], who dissected the overall Gibbs free energy of binding into single contributions:

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta\alpha) + \\ + \Delta G_{ionic} \sum_{ionic_int} f(\Delta R, \Delta\alpha) + \Delta G |A_{lipo}| + \\ + \Delta G_{rot} N_{rot} \qquad 4\text{-}1$$

where, ΔG_{bind} is a summation of hydrogen bonds interactions, ionic interactions, hydrophobic contacts and minus an entropy penalty depending on the number of rotatable bonds. A wide variety of empirical scoring functions now exist as a result of Bohm's pioneering work several of which we have utilised for the purpose of this study (ChemGauss[5], ChemGauss2[5], ChemScore[6], ScreenScore[7], ShapeGauss[5] and PLP[8]) to rank a set of docked molecules generated from the FRED docking algorithm. Empirical scoring functions by nature are additive and generally exhibit a dependence on molecular weight resulting in a bias towards heavier molecules (false positives) in the top of a ranked list of molecules post-scoring. It has also been widely shown in the literature how the different scoring functions produce drastically different results when applied to same target[1,9-12]. Combining two or more of these individual scoring functions into a consensus score can be more effective in the retrieval of active molecules and reduce this effect.

In this study, we propose a method of unbiasing individual scoring functions and suggest a normalisation procedure to refine the result and improve performance of the functions. The impact of this on combining the new scoring functions in a consensus scheme is also investigated.

A set of 19 actives and approximately 10,000 decoys, selected by a Bayesian classifier trained on the set of actives to ensure similar ligand properties, were docked against the ERα (PDB ID: 3ERT) using FRED2.11[5] as shown in Chapter 3. Enrichments were calculated and the results from each scoring function compared by a correlation matrix.

The 19 actives were obtained by processing the 40 estrogens antagonists used in our recent publication[13] through FILTER by Openeye. FILTER is a screening and selection tool that uses a combination of physical-properties calculation and functional-group knowledge to assess drug-likeness of compounds.

A variety of 2D descriptors (available in MOE.2005.06[8]) were calculated for all molecules and plotted against the ranked position produced by each scoring function for each molecule (Figure 4-1).

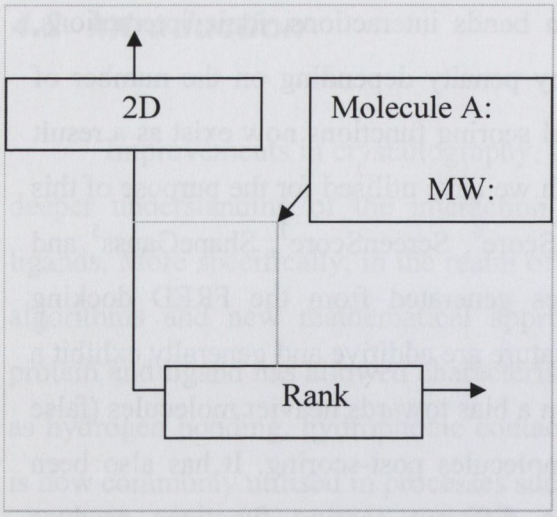


Figure 4-1 Descriptor (y axis) plotted against ranked position (x axis)

For each plot a trend line was calculated and overlayed for each scoring function. The descriptors which exhibited a trend towards the best ranking molecules were selected and the correlation among them was explored. A normalisation procedure was applied to each of the scores depending on the strength of the trend exhibited by selected descriptors.

Comparison of enrichment rates before and after the rescoring procedure was carried out to illustrate the benefits of this normalisation procedure in reducing the number of false positives ranked highly due to increasing molecular weight. Enrichment is a commonly utilised measure of VS performance but it fails to depict the effect that each adjustment to a VS protocol has on the whole of the database, not just the actives. In line with this, Seifert[14] has observed that enrichment alone is not sufficient for optimising a VS protocol[2]. Thus, we propose that scoring frequency, i.e. the distribution of scores observed for a set of ranked ligands, is a better method for assessing the discriminatory power of these scoring functions.

To analyse the binding modes of the ligands used in the docking studies, **S**tructural **I**nteraction **F**ingerprint**s** (SIFt[15]) were employed as a means of determining whether ligands ranked at the top of a scored database after application of our normalisation procedure now exhibit better interactions with key residues in the active site of ERα.

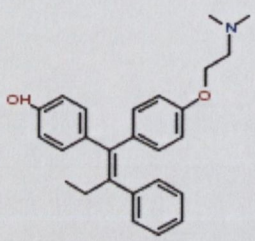
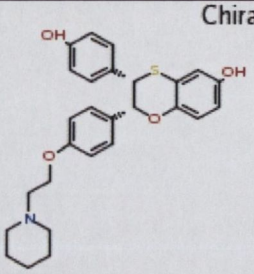
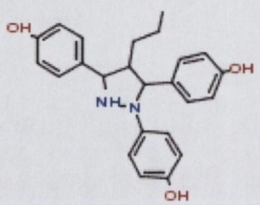
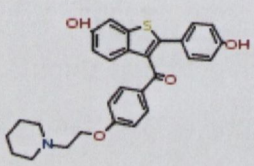
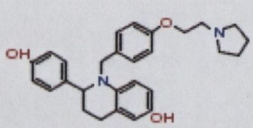
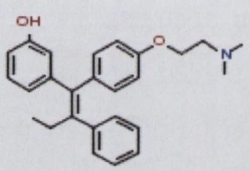
4.3 Computational Methods

4.3.1 Multi-conformers database preparation

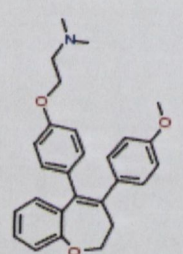
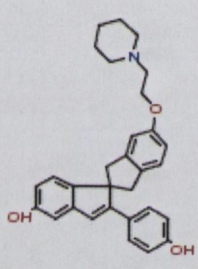
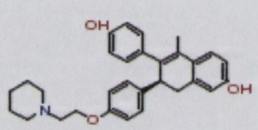
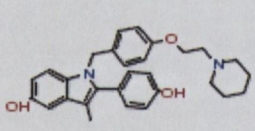
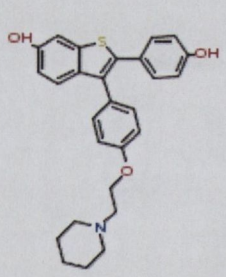
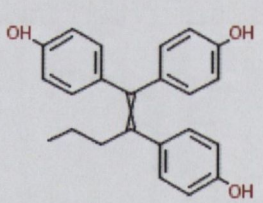
The following protocol was fully automated through PP[16]. Maybridge[17], Bionet and IBS2004 databases were filtered using a Bayesian algorithm trained for ERα. All molecules were standardised (e.g. keep largest fragment, check valences) and ionised at pH 7.4.

Nineteen estrogen antagonist structures were retrieved from literature (Table 4-1) and parsed as above. The active and decoy set were merged and duplicates removed ensuring the final database size was 10,000. Omega 1.81[18] was employed for the generation of conformers. Default settings were kept with the exception of `maxconfs=10`, `finalopt=true` and `output= mol2` to retain charges.

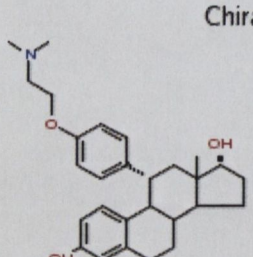
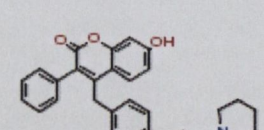
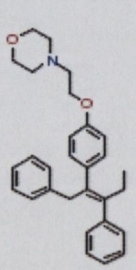
Chiral

Chiral

Chiral

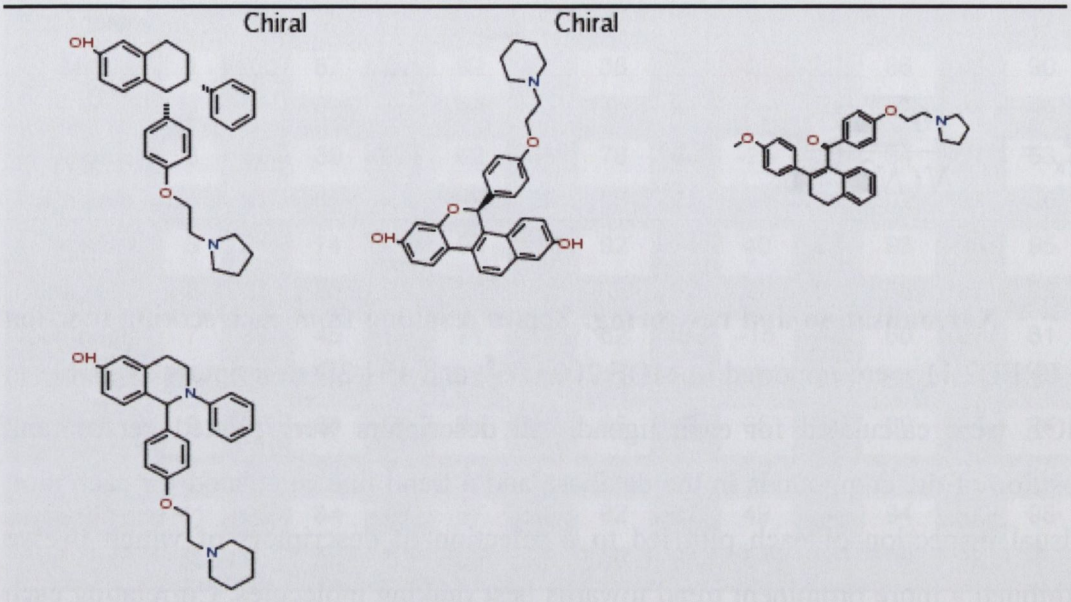


Table 4-1 Structure of the 19 estrogens selected from the literature.

Docking: The coordinates of the estrogen receptor alpha co-crystal structure 3ERT[19] with antagonist 4-hydroxytamoxifen were downloaded from PDB[20]. Water molecules were removed and MOE.2005.06[8] was used to add hydrogens to the protein and a minimisation protocol using MMFF94 force-field was implemented to adjust the positions of hydrogen while keeping the heavy atoms fixed at their coordinates. The ligand was extracted and used to define the 5 Å search box for docking. FRED2.11[5] was then used to generate poses inside the binding pocket using the standard default settings. The same protocol for docking was validated and is has been used in previous work by our lab [13,21].

Scoring: Scoring of the docked poses for each ligand was achieved using the scoring functions implemented in FRED2.1.1 (ChemGauss2, ChemGauss, ChemScore, PLP, ScreenScore, ShapeGauss). Consensus scoring (*testconf* = none) was also calculated. In a previous publication we already showed that FRED2.01 in conjunction with ChemGauss and PLP provides excellent enrichment with ERα validating our choice of docking algorithm and set of scoring functions.

Correlation Matrices: Correlation matrices were calculated using MOE.2005.06 applying Spearman's formula (Equation 4-2). Spearman Rank correlation measures the correlation between two sequences of values. The two sequences are ranked separately and the differences in rank are calculated at

each position (D). Spearman Correlation is from -1 to 1. *N* represents the total number of element in the lists.

$$R_s = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad 4\text{-}2$$

Normalisation and re-scoring: Scores resulting from each scoring function in FRED2.11 were imported in MOE.2005.06[8], and 194 2D descriptors available in MOE were calculated for each ligand. All descriptors were plotted versus rank position of the compounds in the database and a trend line calculated for each plot. Visual inspection of each plot led to a selection of descriptors of which twelve exhibited a more prominent trend towards best ranking molecules. Correlating each the descriptors as below in Figure 4-2, it was observed that each of the 12 correlated well with *MW*.

Descriptors		1	2	3	4	5	6	7	8	9	10	11	12
MW	1	100	87	52	92	81	88	59	46	90	86	93	90
a_IC	2	87	100	39	86	74	80	45	58	88	89	84	89
rings	3	52	39	100	62	65	78	96	-25	36	54	57	53
a_heavy	4	92	86	62	100	95	96	71	46	95	92	97	96
a_nC	5	81	74	65	95	100	92	74	40	88	93	95	95
zagreb	6	88	80	78	96	92	100	82	23	85	88	92	90
opr_brigid	7	59	45	96	71	74	82	100	-15	48	60	67	61
opr_nrot	8	46	58	-25	46	40	23	-15	100	65	50	49	54
Kier1	9	90	88	36	95	88	85	48	65	100	89	94	94
apol	10	86	89	54	92	93	88	60	50	89	100	94	98
SMR	11	93	84	57	97	95	92	67	49	94	94	100	98
vdw_vol	12	90	89	53	96	95	90	61	54	94	98	98	100

Figure 4-2 Correlation matrix of 12 descriptors. For example the correlation between molecular weight (labelled as MW) and a count of rotatable bond (opr_nrot) is highlighted in orange. a_IC : Atom information content (total); rings: number of rings; a_heavy: number of heavy atoms; a_nC: number of carbon atoms; Zagreb: Zagreb index; opr_brigid: number of rigid bonds; opr_nrot: number of rotatable bonds; Kier1: first kappa shape index; apol: sum of the atom polarizablilities; SMR: Molecular refractivity; vdw_vol: van der Waals surface. Complete documentation about the descriptors used above may be found in the MOE reference manual.

Since the 12 descriptors correlation with each other, only the *MW* was taken into account for further calculations. To negate the trend towards higher MW compounds being prioritised at the top of the database, a normalisation procedure was introduced as follows. Firstly, the scores resulting from each scoring function and the corresponding MW of each were normalised from 0 to 1 according to:

$$normValue = (x - \min)/(\max - \min) \qquad 4\text{-}3$$

where ***x*** represent either the MW or the actual Score for each compound and min, max are the minimum and maximum values of the ranked set.

The ranked list was then re-scored in order to reduce the dependence of the scoring function on molecular weight. A new formula was introduced to penalise only those molecules which exhibit high *MW*, and leave the remainder relatively unchanged.

$$WScore = normScore(1 - normMW)^{0.25} \qquad 4\text{-}4$$

A visual representation of the formula above is depicted in the Figure 4-3.

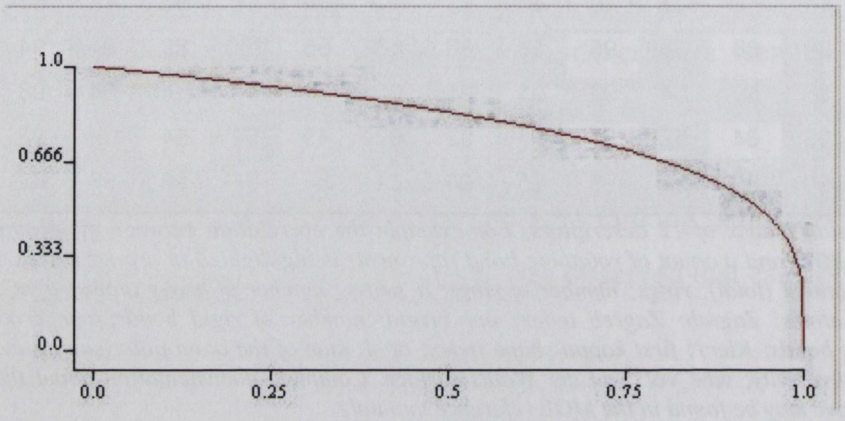


Figure 4-3 Plot of the function $y=(1-x)^{0.25}$.

A linear scoring will award molecules with low MW and penalise molecules with higher MW. The function showed in Figure 4-3 is intended to only penalise molecules with higher MW and not affect those with low MW.

4.3.2 Analysis of the results

Enrichment: The potential of the scoring functions to identifying a set of known active ligands from a set of drug-likes 'decoys' is typically measured using the metric enrichment . Enrichment is a measure of the proportion of hits retrieved in a subset of compounds compared with the proportion of hits expected from a random sample of compounds.

$$E = \frac{Hits_{sampled} / N_{sampled}}{Hits_{total} / N_{total}} \qquad 4\text{-}5$$

Enrichments were calculated over all the scoring functions, before and after re-scoring procedure.

Scoring frequencies: Scoring functions results can be analysed plotting the score frequencies. In other words, how many molecules (%) receive a certain score. Scoring frequencies were normalised for ease of plotting, and results prior to and post re-scoring were analysed by this method.

SIFt fingerprints: **S**tructural **I**nteraction **F**ingerprint**s** (SIFt) are used in this study to capture the interactions occurring between each docked pose of a compound and the residues of the active site of ERα. VS results for a set of protein-kinase inhibitors have previously been shown to produce superior enrichment using SIFt compared with traditional scoring functions. It has also been used to design target-focused libraries due to its inherent sensitivity to the target. In this light we have employed a SVL implementation of SIFt available in MOE.2005.06 to calculate the 1D binary string. Tanimoto coefficient was utilised to discern the similarity between the string generated for the PDB entry 3ERT and the ranked poses. The top scoring 200 molecules for each scoring function were assessed for similarity to the crystal structure with an overlap of 65%, 70%, 75%, 80%, and 85%.

4.4 Results

The graphics depicted in Figure 4-4 show the effect of the re-scoring procedure on the MW distribution plotted versus the rank position for the Shapegauss scored set.

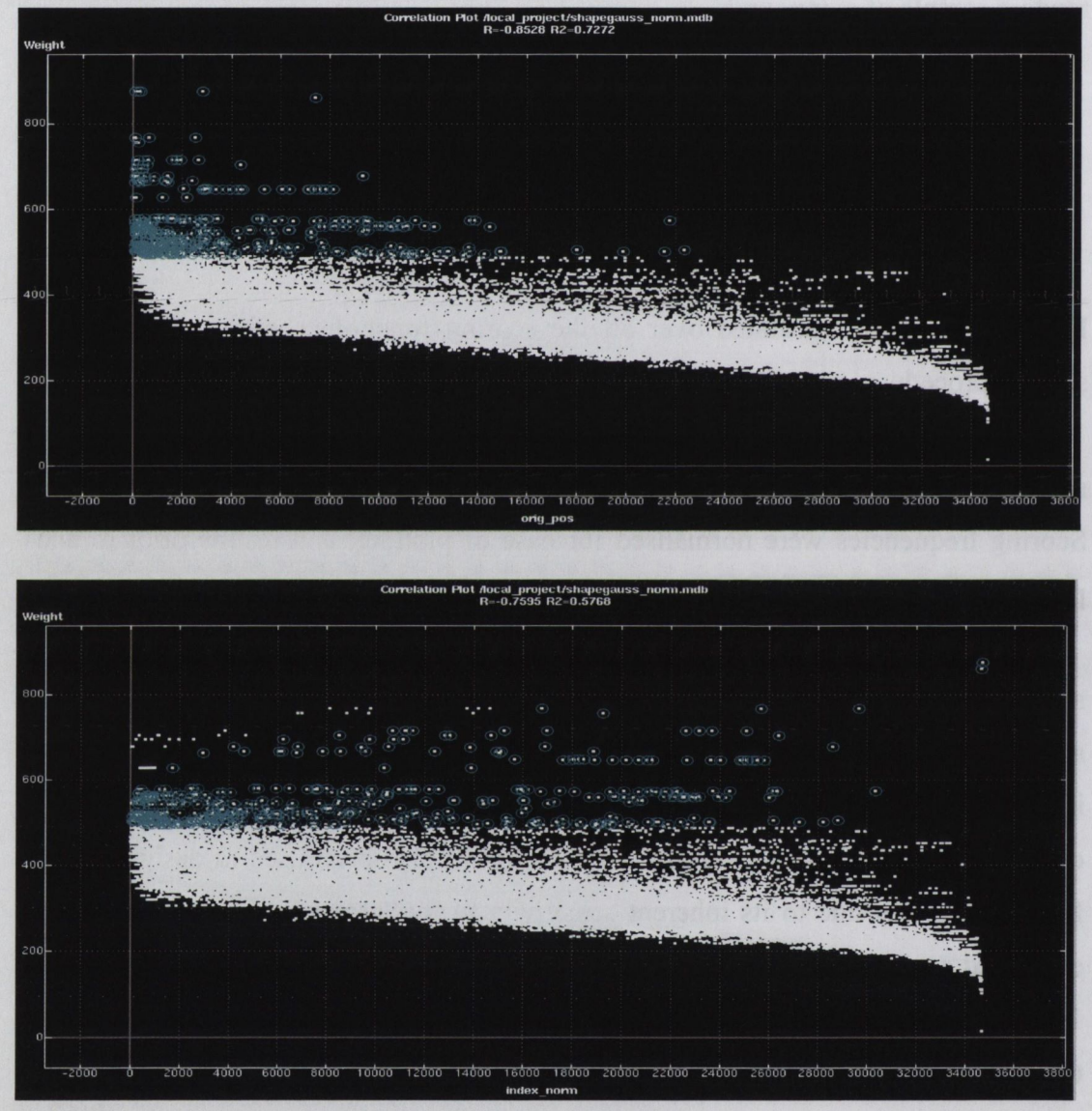


Figure 4-4: Effect of the re-scoring procedure on the molecular-weight distribution plotted versus the rank position for the Shapegauss scored set. Top part: before the normalisation; bottom part: after normalisation. Molecules with a molecular weight above 500 are highlighted in both the plots.

Molecules with a MW above 500 are highlighted, and moving from the first plot to the second we can observe these molecules are more evenly distributed

throughout the ranked set. A clear trend however, is still observed towards the rank position of higher MW molecules.

The enrichment curves retrieved for the scoring functions as implemented in FRED2.1.1 before rescoring and after re-scoring, are showed in Figure 4-5.

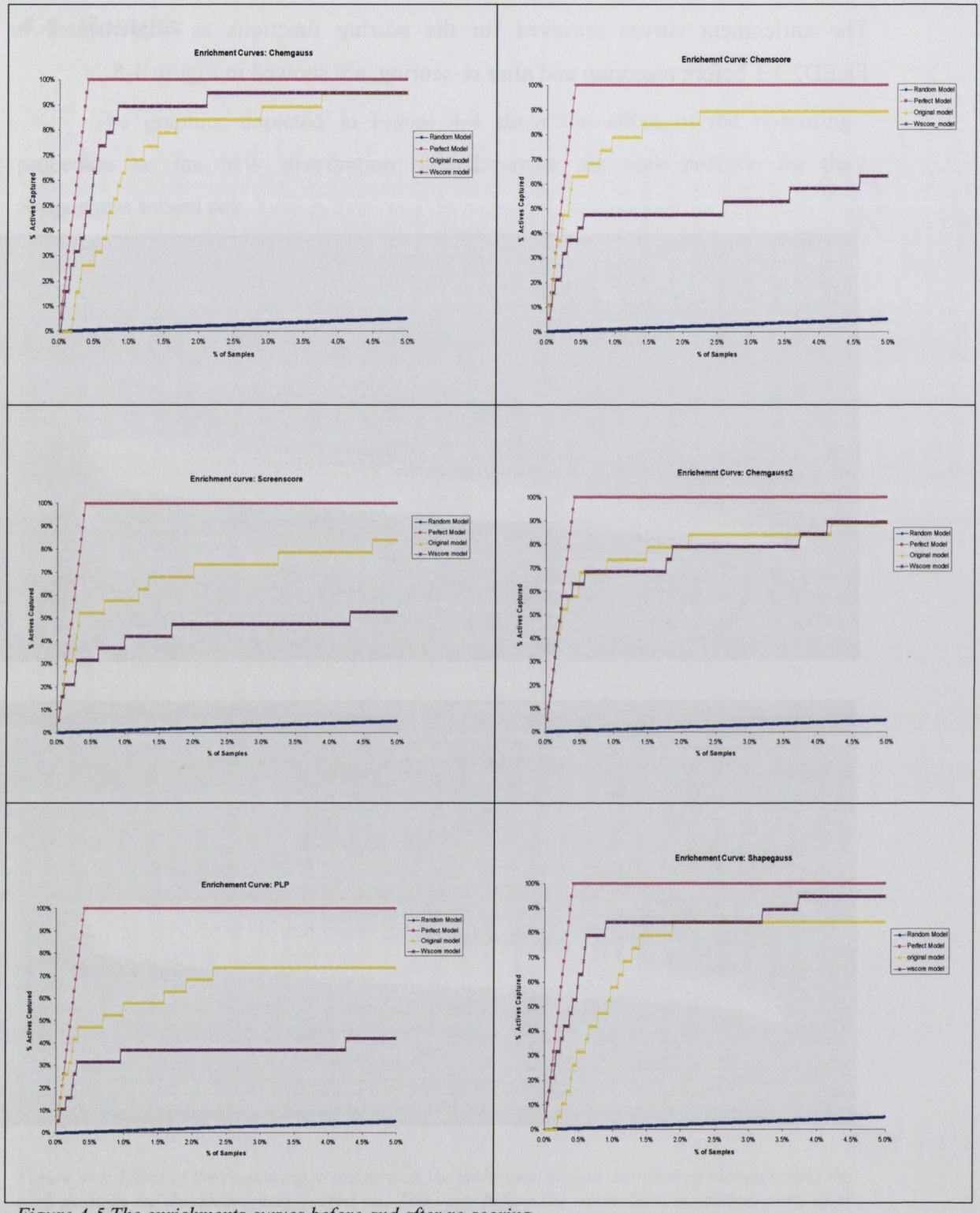


Figure 4-5 The enrichments curves before and after re-scoring.

Shapegauss and Chemgauss re-scored results (Wscore) show higher enrichment values when compared with the original scoring functions. Chemgauss2

exhibits neither improvement nor a reduction of enrichment. PLP, Chemscore and Screenscore original score enrichments outperform the re-scored enrichment results.

To further investigate the increase in enrichment observed for Shapegauss and Chemgauss, but reduction for PLP, Chemscore, Screenscore, molecular weight versus the rank position was plotted for all scoring functions. The upper and lower MW for the active set lies between ~340-480 and from Figure 4-6 it is clear that prior to the re-scoring procedure, both Shapegauss and Chemgauss retrieve molecules with a MW outside this range. After implementing our re-scoring method, the molecules retrieved by Shapegauss and Chemgauss are now in the same MW range as the active set, thus producing a higher enrichment. It is also shown that Chemgauss2 remains unaffected, correlating well with no change observed in enrichment.

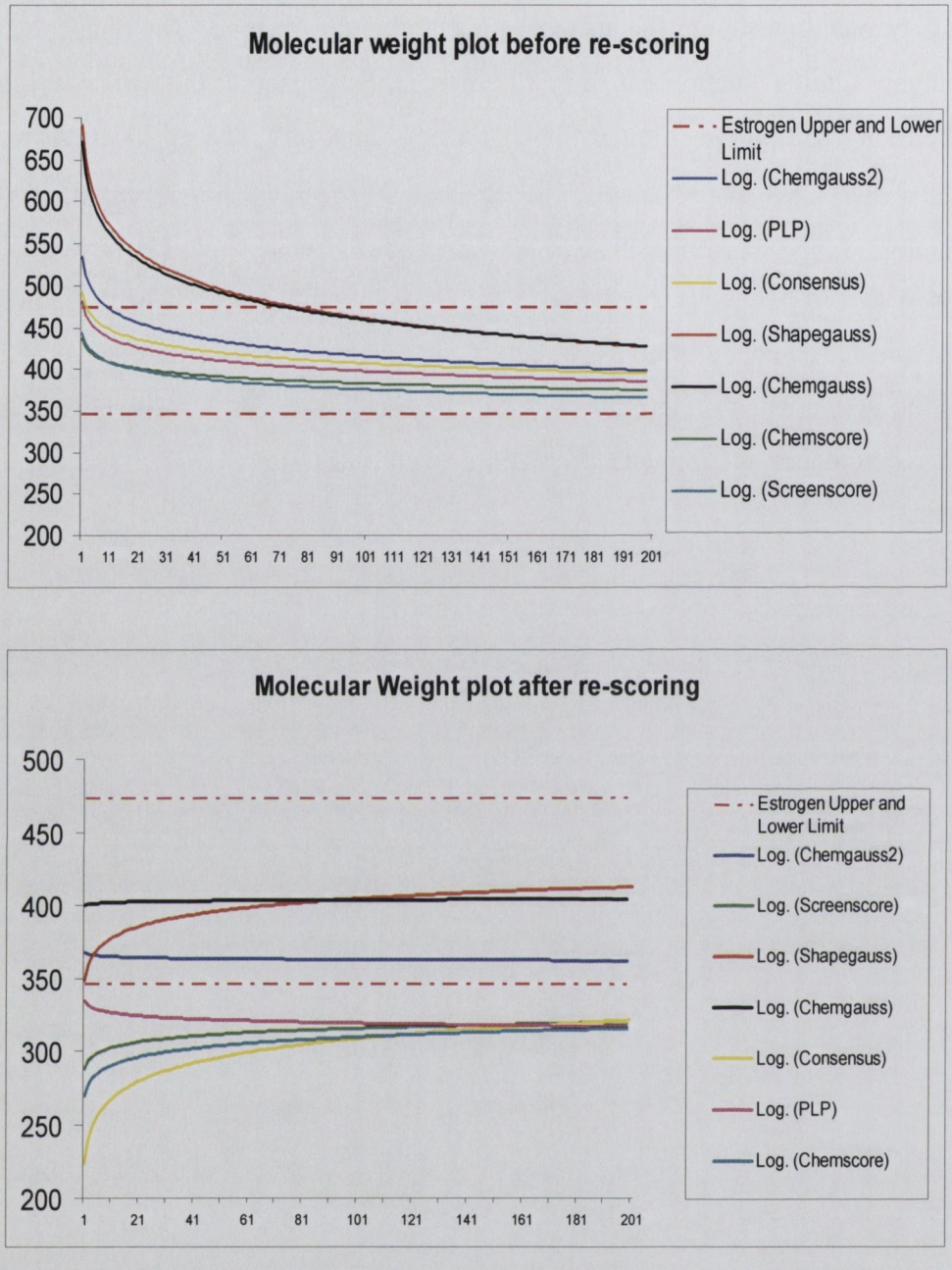


Figure 4-6 Molecular weight versus the rank position plotted for all scoring functions.

However, there is a significant shift towards the retrieval of molecules that have a lower MW than the active set. This would obviously result in an enrichment reduction for those particular scoring functions (Chemscore, PLP, Screenscore).

To examine the full effect of our re-scoring procedure on both the active and decoy sets, it is necessary to look at the score distribution of each molecule for all the scoring functions (Figure 4-7).

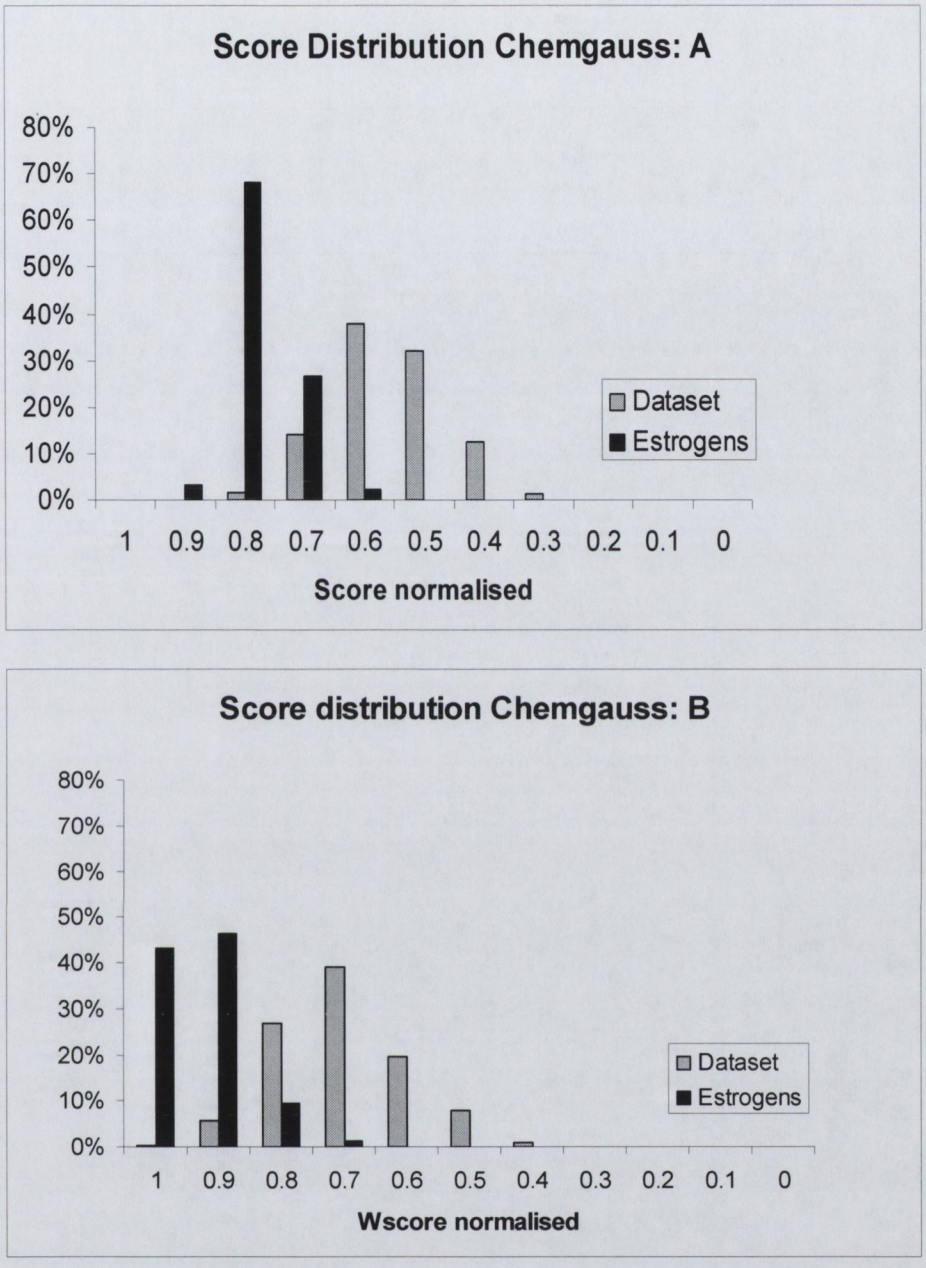
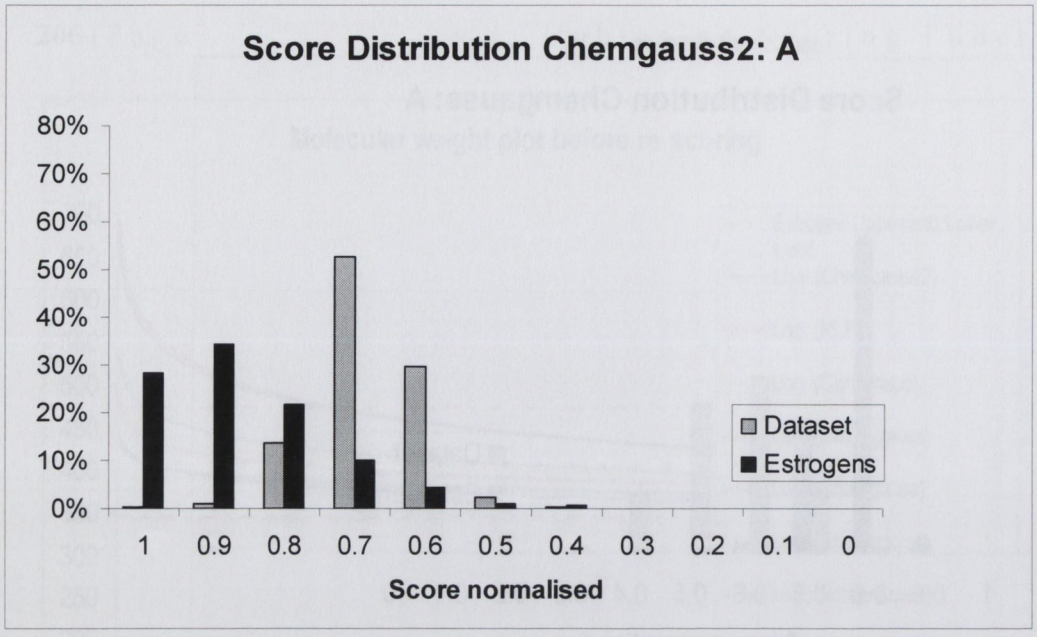


Figure 4-7 Score distributions for all the scoring functions considered. Set A: before normalisation, Set B: after normalisation.



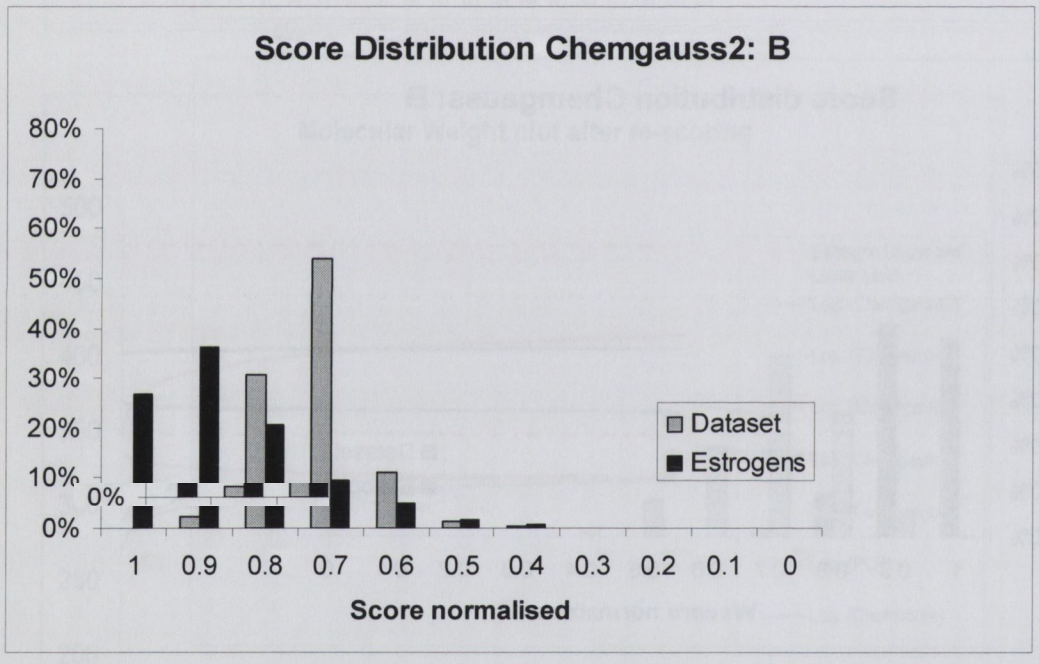
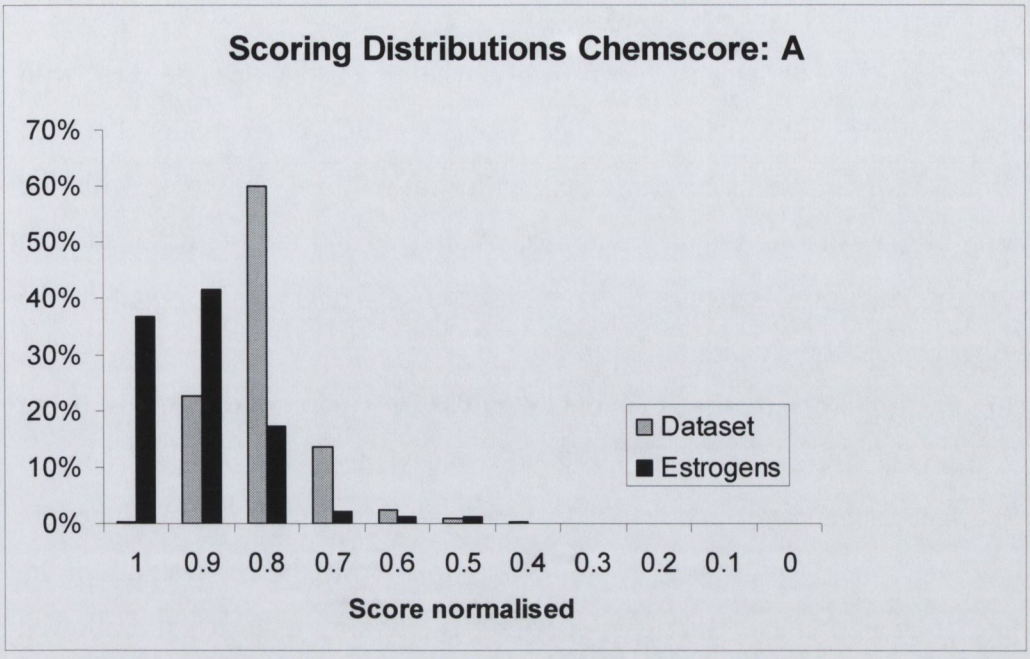


Figure 4-8 (Continuous).



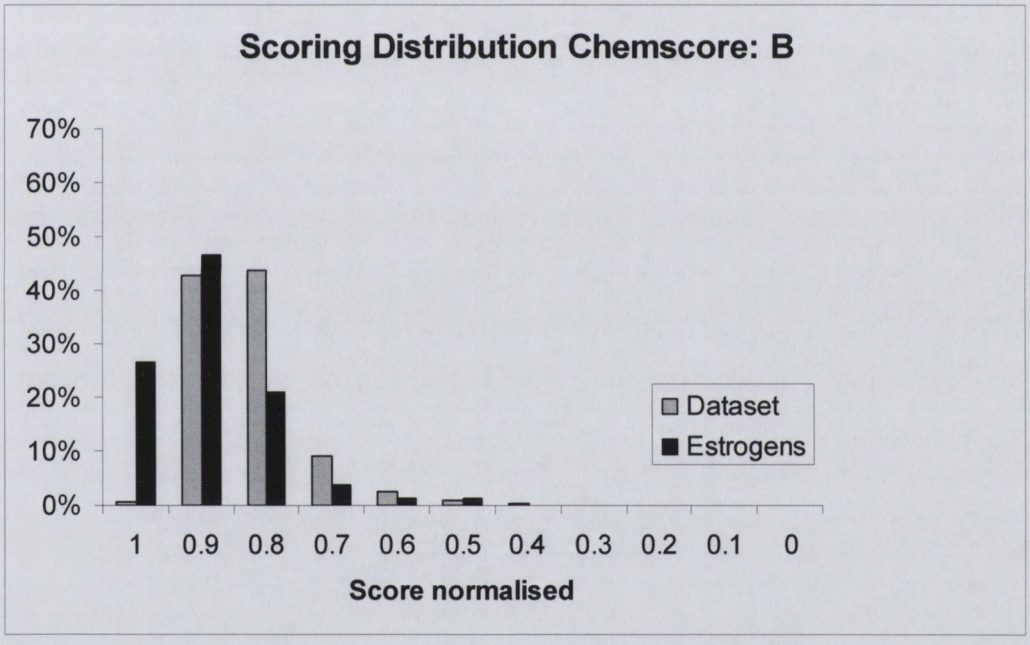


Figure 4-9 (Continuous).

The graphs in Figure 4-7, Figure 4-8 and Figure 4-9 represent the score distributions for Chemgauss, Chemscore and Chemgauss2. For conciseness, the results for just three scoring functions are depicted, being representative of the three different scenarios obtained from the enrichment results:

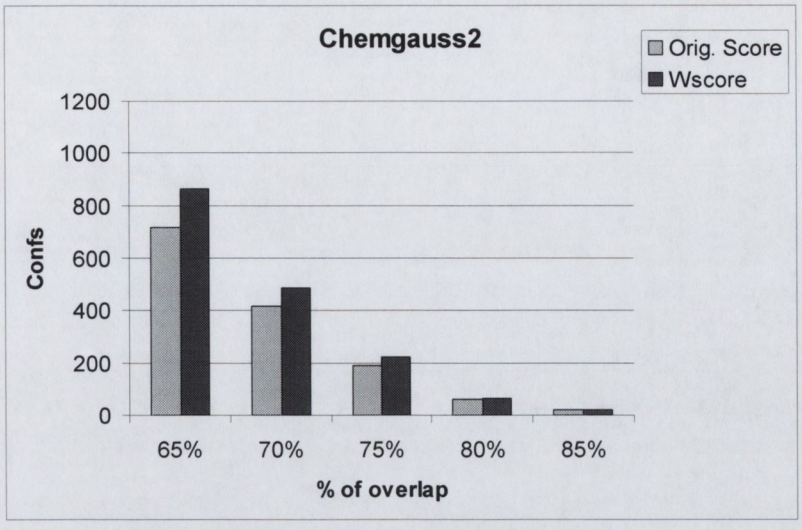
- Re-scoring outperforms original score (Chemgauss)
- The original score outperforms re-scoring method (Chemscore)

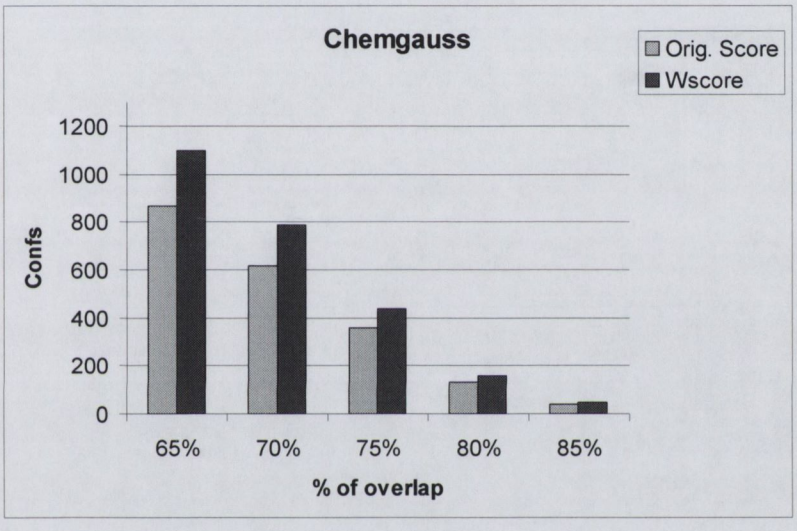
- No appreciable change (Chemgauss2)

In the case of Chemgauss, a shift for the whole ranked set can be seen, with greater separation of the actives from decoys also. The discrimination between actives and decoys appears to decrease significantly for Chemscore, with the distribution of the estrogen remaining almost unchanged and the decoy set shifting higher. In the case of Chemgauss2, no significant difference is observed in the scoring distribution of the dataset.

Finally, to critically analyse and classify the type of compounds being retrieved from each screen, we introduced SIFt analysis to the procedure. The goal was to estimate if the compounds ranked in the top 200 for each scoring function, after re-scoring with our procedure, could better interact with the key residues of the active site compared than prior to re-scoring. The x axis indicates the allowed overlap and in the y axis the numbers of molecules with that percentage overlap. Thus the higher the number of molecules retrieved for a given overlap translates into closer protein-ligand equivalent to the ones observed in the crystal structure now being made for the top molecules in the ranked dataset.

The graphs below (Figure 4-10) show that after the re-scoring procedure the molecules recovered with an overlap of 65%-70% increase for all the scoring functions. For the scoring functions, Shapegauss, Chemgauss, Chemgauss2 improved interactions are observed for all the higher ranked molecules with better overlapping at 70%-85% also.





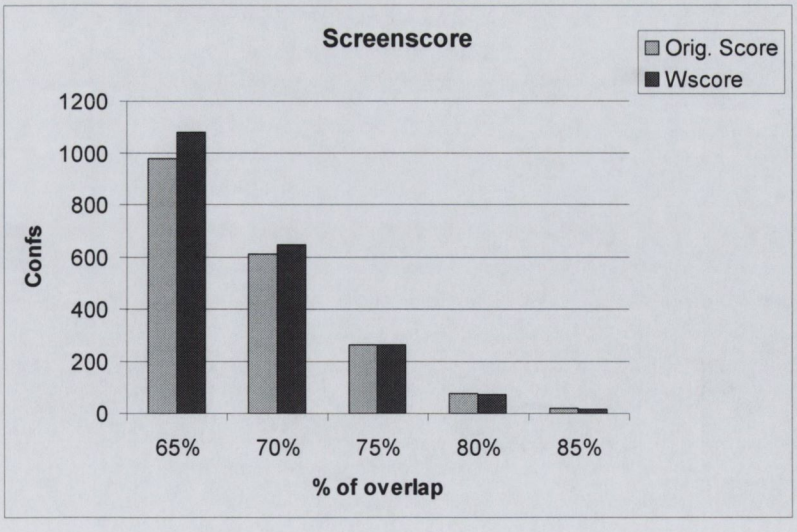
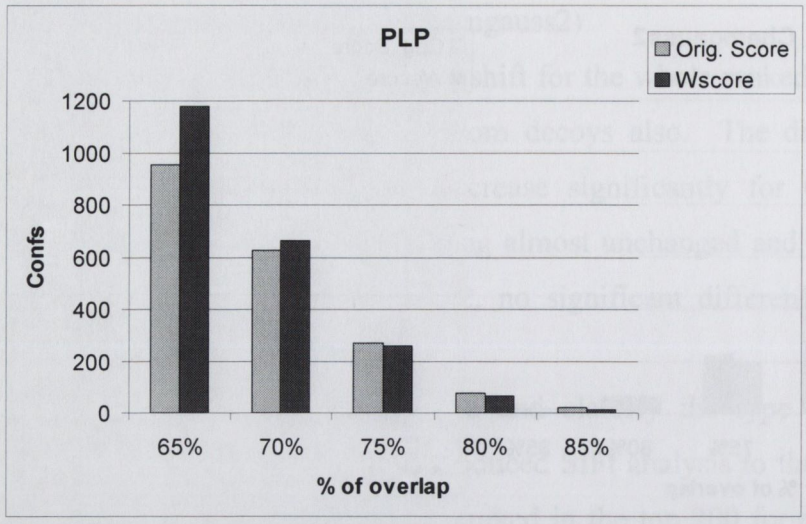
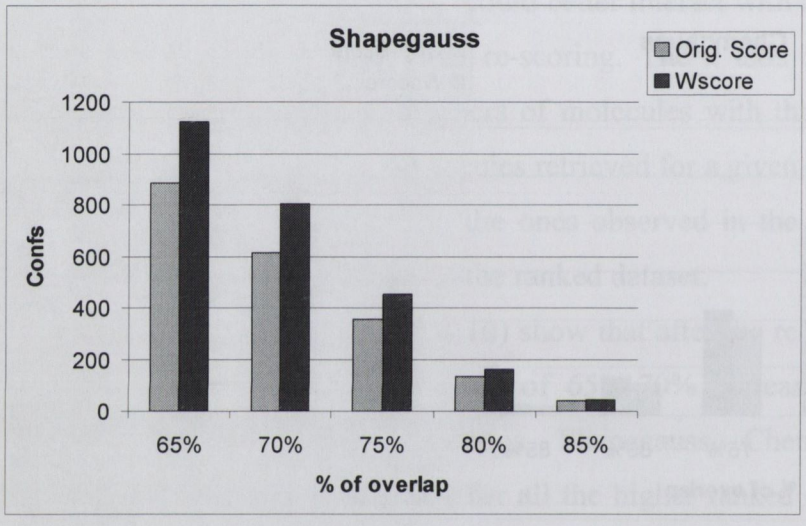


Figure 4-10 SIFt analysis results for all the scoring functions.





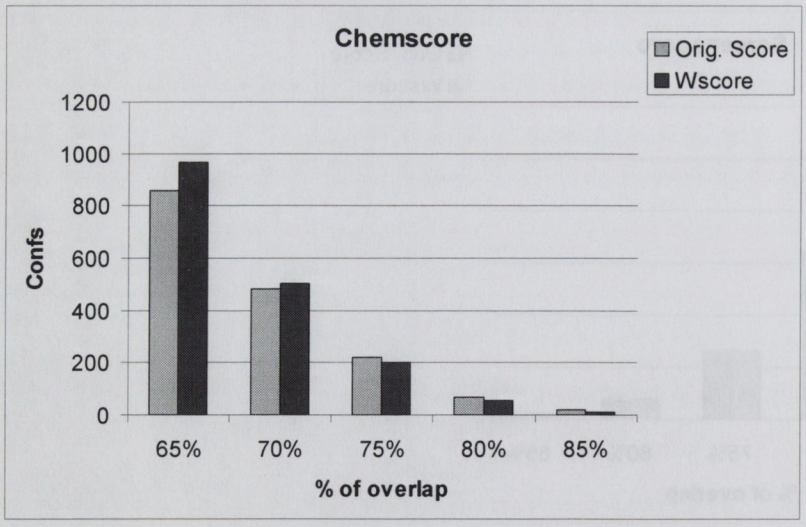


Figure 4-11 (Continuous).

As we have seen, after re-scoring there are improvements in the abilities of Chemgauss and Shapegauss to discriminate between actives and decoys in the database used. To further illustrate the value of unbiasing the scoring functions with respect to MW, we formed a consensus score consisting of adding the rank by rank values from the original positions of PLP, Chemscore, Screenscore, Chemgauss2 and the re-scored ranks of Chemgauss and Shapegauss.

%	0.20%	0.50%	1%	1.50%	2%	3%	5%
Ligands No.	9	23	47	70	94	142	236
Enrichment_original	193	140	79	57	22	28	18
Enrichment_New_Wscore	220	140	79	57	42	28	18
Theoretical Maximum	*248*	*204*	*100*	*67*	*50*	*33*	*20*

Table 4-2 Consensus enrichment obtained with the original scoring function versus the one obtained after normalisation.

The new consensus enrichment (Table 4-2) obtained higher values in the first 0.20% of the database, whilst the values in the rest of the database remained the same. The benefit of this procedure is evident considering the formula can be tuned by altering the exponential value to penalise more heavily compounds with high MW or low MW. We have not attempted to tune the function to its maximum potential as the purpose of the study was not to achieve high enrichment values, but to prevent high MW compounds from ranking highly in a database screen.

4.5 Discussion and Conclusions

In this work we propose a procedure to un-bias scoring functions with regard to MW. It has been noted previously by Jacobsson *et al.*[22] that ligand bias is a source of false positives in structure-based virtual screening. Importantly, scoring functions are generally correlated with molecular size due to their inherently additive nature. Pan *et al.* proposed size normalisation by dividing by the square or cubic root of the Heavy Atom Count (HAC) to negate the effect of this dependence. We have introduced a function that can be tailored or tuned to lightly or heavily impact this MW bias. This formula penalises exponentially higher molecular weight molecules and leaves other molecules untouched.

Of great importance in this respect is that it is calculated post-scoring and thus can be re-evaluated easily. The ranked set can also be easily, heavily biased towards the retrieval of more lead-like MW compounds without loss in their ability to interact correctly with key residues within the active site by increasing the value of the exponential in our function.

Plotting the ranked position of compounds versus their associated descriptors is a fast way to inspect the characteristics of the compounds which scoring functions rank better than others. The formula we propose can be applied to any descriptor where un-biasing is required. We propose that the Spearman coefficient be employed with the aim of exploring possible relations between descriptors values and ranking position.

Although improvements in enrichments were observed for Chemgauss and Shapegauss after the application of our formula, a reduction was also seen for the remaining scoring functions with the exception of Chemgauss2. However, using SIFt analysis, it was shown that although there is a fall in enrichment for some of the scoring functions that there is conversely an improvement in the quality of possible false positives in the ranked dataset.

A better understanding of the full impact of applying our formula can be derived from the score distributions, since it brings more information about how scoring functions rank all the compounds in the database and not just the positive control. Important feedback results from analysis of the scoring frequencies. Taking

Chemscore as an example, it was clear that it outperformed most of the scoring functions in discriminating between actives and decoys. The re-scoring procedure clearly decreases the enrichment value; however analysis of the score distribution shows that on average the score of the positive controls remain more or less unchanged. Thus, the re-scoring procedure brings up more molecules from the decoy set among the higher rankings. It is possible that some of these molecules that are now ranked highly with equivalent scores to those of the antiestrogens are actually actives, as the database chosen was one compiled from commercially available compounds.

Noteworthy, the Spearman coefficients obtained when comparing Shapegauss and Chemgauss after the normalisation, versus all of the other original scoring functions show slightly higher r^2 values. This can be interpreted as a better convergence of the scoring functions towards similar results. Furthermore, the scoring functions involved in this study are empirical and therefore an approximation of binding affinity. Thus they should converge to approximately similar results, even when composed by different terms.

Finally we demonstrate the benefit of adjusting the bias of individual scoring functions for incorporation in a consensus scheme. Utilising the ranks obtained for re-scored Shapegauss and Chemgauss in conjunction with the ranks obtained for the original versions of Chemgauss2, PLP, Screenscore and Chemscore, higher enrichment values were generated.

These methods can be readily applied to any descriptors (QSAR etc.) to tune the scoring functions towards retrieval of molecules with defined properties and un-bias particular parameters within a scoring function if the user has prior knowledge of the type of ligands needed to be selected.

4.6 Reference

1. Marsden PM, Puvanendrampillai D, Mitchell JB, Glen RC. Predicting protein-ligand binding affinities: a low scoring game? Org Biomol Chem 2004;2(22):3267-3273.
2. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49(20):5912-5931.
3. Bohm HJ. 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(3):243-256.
4. Bohm HJ. New Computational Approaches to Predict Ligand-Protein Interactions. Kubinyi H, editor: Wiley-VCH; 1998.
5. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. Biopolymers 2003;68(1):76-90.
6. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11(5):425-445.
7. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44(7):1035-1042.
8. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 2000;14(8):731-751.
9. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46(12):2287-2303.
10. Bursulaya BD, Totrov M, Abagyan R, Brooks CL, 3rd. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17(11):755-763.
11. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43(25):4759-4767.
12. Xing L, Hodgkin E, Liu Q, Sedlock D. Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 2004;18(5):333-344.
13. Knox AJ, Meegan MJ, Carta G, Lloyd DG. Considerations in compound database preparation--"hidden" impact on virtual screening results. J Chem Inf Model 2005;45(6):1908-1919.
14. Seifert MH. Assessing the discriminatory power of scoring functions for virtual screening. J Chem Inf Model 2006;46(3):1456-1465.
15. Singh J, Deng Z, Narale G, Chuaqui C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. Chem Biol Drug Des 2006;67(1):5-12.
16. Hassan M, Brown RD, Varma-O'brien S, Rogers D. Cheminformatics analysis and learning in a data pipelining environment. Mol Divers 2006;10(3):283-299.

17. Maybridge.www.maybridge.com.
18. OMEGA 1.8.1. distributed by Openeye Scientific Software http://www.eyesopen.com.
19. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell 1998;95(7):927-937.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
21. Knox AJ, Meegan MJ, Sobolev V, Frost D, Zisterer DM, Williams DC, Lloyd DG. Target specific virtual screening: optimization of an estrogen receptor screening platform. J Med Chem 2007;50(22):5301-5310.
22. Jacobsson M, Karlen A. Ligand bias of scoring functions in structure-based virtual screening. J Chem Inf Model 2006;46(3):1334-1343.

5. Post-processing optimisation and combinatorial decoration based on scoring frequency

5.1 Abstract

As described in Chapter 4, scoring frequency analysis can be a useful way to inspect measures of molecular fitness, such as those provided by computational docking algorithms or activity data such as IC_{50}'s. This Chapter introduces a statistical selection protocol, combined with virtual combinatorial enumeration, to post-process, optimise and decorate a ranked output list obtained from an enriched set of data such as the one produced by a scoring function. Assuming the ability of the scoring function to discern correct binding modes from false positives as seen in Chapter 3, this protocol creates a second generation of virtual compounds by recombination of those molecular fragments assessed to be significant in mediating receptor binding. The protocol is exemplified by application to a screening library docked into a nuclear receptor target, PPARγ, which is therapeutically relevant for the treatment of various diseases such as Type II diabetes (Chapter 3). The second generation set of compounds shows a significant shift towards better scores, as compared to the original population, and a novel group of molecules that can be either synthesised or matched within vendor databases was created.

5.2 Introduction

Increasing understanding of protein-ligand interactions and the development of vHTS tools such as ligand docking in recent years has allowed the identification of many pharmacologically active molecules. An example of such capability is presented in Chapter 3 where by application of VS techniques novel PPARγ agonists were identified.

As shown in the previous chapters, the search for new active compounds against a specific target requires the screening of a large and diverse dataset of in-house and/or vendor compounds. Drug-like molecules are virtually docked into a receptor active site and a scoring step is applied where all the poses produced by the docking are evaluated and an estimation of the binding affinity is obtained. Scoring functions are essential tools in this process as the final hits are usually selected on the basis of their scores. However, numerous publications provide evidence that the hits generated by a virtual screening protocol typically contain a high percentage of both false positives and false negatives[1,2]. In Chapter 3, although 40 molecules that scored highly were tested *in vitro*, just one true active compound was retrieved.

The inability to screen out false positives and negatives is in part due to the inherent inaccuracy of the whole simulation in vHTS and, as shown in the preceding chapters, to the many parameters and factors that can eventually affect the screening process[1,3,4]. Docking itself presents some degree of imprecision due either to incomplete sampling of conformation space or scoring function approximation and inaccuracy[5,6]. An imperfect scoring function can mislead by predicting incorrect ligand geometries or by selecting non-binding molecules over true ligands. Artificial results can also be traced to the decoy set used for the screening and the pre-processing applied to it. The choice of the decoy set used during the screening may bias or affect the results[7]. As also shown in Chapters 1 and 2 the results may drastically depend on the pre-processing of a dataset. As described in Chapter 1, in order to reduce the impact of false positives or false negatives many different post-processing and protocol optimisation strategies

have been suggested and are usually required to correctly prioritise a collection of compounds.

Post-processing methods are usually applied in order to optimise the results of a scoring function or to improve the performance and the discriminating capability of the screening protocol. In Chapter 4, for example, we showed how a re-scoring strategy can un-bias the scoring phase, thus leading to improved results. Another typical strategy to reduce the number of false positives is to merge the hit lists generated by different scoring functions, generating a so called *consensus score*[2,8-12]. This method proved in many cases to enhance the performance of the docking process in discerning active molecules from a decoy set of inactives.

Making an overview of the commonly used processing and post-processing methods described in the precedent chapters and in literature, we can see that the majority adapt the virtual docking/scoring routine to the target and/or the dataset. This is done until a good enrichment and low false positive rate validates the method itself and it can therefore be used for vHTS.

In Chapter 3 the generation of conformers, the selection of a receptor frame reference and the parameters of the docking procedure were adjusted until a good level of confidence with a validation set were achieved. The scoring function itself can also be adjusted in its singular components in order to "tune" it up with the target. As an example, FRED, mentioned in the earlier chapters, allows the user to weight different components within the individual scoring functions. Furthermore, different weights can be given to the different scoring functions when a consensus strategy is chosen. On the other hand, as described in Chapter 1, these scoring functions offer an estimation of binding affinity by summing interaction terms derived from weighted parameters. These parameters typically are determined by fitting the scoring function itself to real experimental binding results. In other words, the scoring functions are trained to properly weight different types of interactions. The shortcoming of these methods can be sometime traced to the bias of the scoring function due to the original training set.

Therefore, the general process of optimisation of a scoring function, implicitly engineered with certain bias, involves iterative assessment of the

results, until a reasonable efficiency in the detection of known actives is achieved.

This Chapter suggests an alternative approach. In our protocol, optimisation is accomplished by the adaptation of the hit list to the scoring function and not *vice-versa*. In other words, instead of training the scoring function to recognise the needles (known actives) in the haystack (decoy set), this chapter proposes a method to build a set of molecules capable of capturing the scoring function sensitivity. In order to *cast* the dataset over the scoring function, a first necessary assumption is that the scoring function itself must be able to give an acceptable estimation of the binding affinity. As scoring functions have been used for many years, we can comfortably assume some efficacy in vHTS.

The suggested technique rationally builds and decorates new chemical structures with the aim of obtaining a better score than the original set with that particular screening protocol.

This approach effectively can be placed among the post-processing methods and also lead optimisation, even though it is operated on a list of molecules rather than on a single hit. Usually, a number of considerations are relevant to the process of optimisation[13,14]:

1. Improvement of conformational properties: where the docked structure is strained or has a high degree of flexibility, it may be possible to introduce changes that make the molecule less strained or more rigid.
2. Improvement of bioavailability: it may be necessary to modify physicochemical properties such as the octanol/water partition coefficient and/or a poor metabolic profile in order to improve bioavailability.
3. Improvement of complementarily to the binding site: improvements in complementarily based on small changes can usually be proposed by simply looking at the docked complex on a high-resolution graphics workstation.

The method we propose automatically and rationally post-optimises output from vHTS in a manner that could easily be extended to address the third of these consideration.

In more detail, even though we consider a scoring function only "roughly" able to distinguish between what is active and what is inactive, an overall trend

that highlights certain molecules or substructures will be present when examining a population of docked molecules as a whole. The classical output of a scoring function, a ranked list of molecules with an associated score when overviewed with a statistical methodology, can give an indication of which scaffolds and side chains are predicted to make the strongest interactions with the target protein. A second generation of virtual compounds can then be created through the recombination of those molecular fragments assessed to be significant in mediating protein binding.

This is accomplished using a statistical protocol, combined with a virtual combinatorial enumeration that leads to a target-driven decoration of promising scaffolds. This protocol derives its roots in precedent works that showed how the computation and analysis of scoring frequencies (Chapter 4) is a potent tool capable of quantifying and modulating the discriminatory ability of a scoring function.

In a three step procedure, the full output set from an empirical scoring function (Chemgauss2 implemented in FRED by Openeye[15]) is firstly fragmented to a specific set of disconnection rules. The score associated with each parent molecule is inherited by all of its child fragments (Figure 5-1).

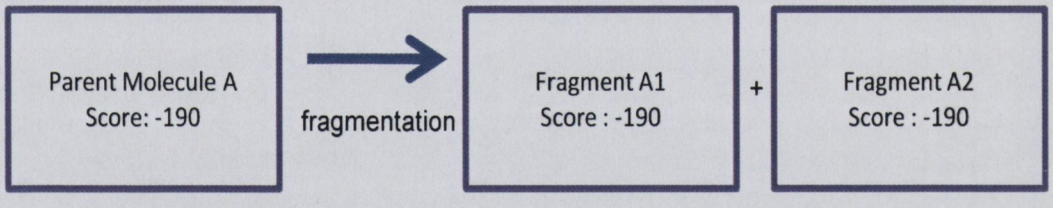


Figure 5-1 Fragmentation scheme.

Based on their topological features (see the Methods section), the fragments are then classified as either *scaffolds* or *chain* (Figure 5-2). In Figure 5-2 Fragment A1 is, as an example, classified as Chain A1 and Fragment A2 classified as a Scaffold A2.

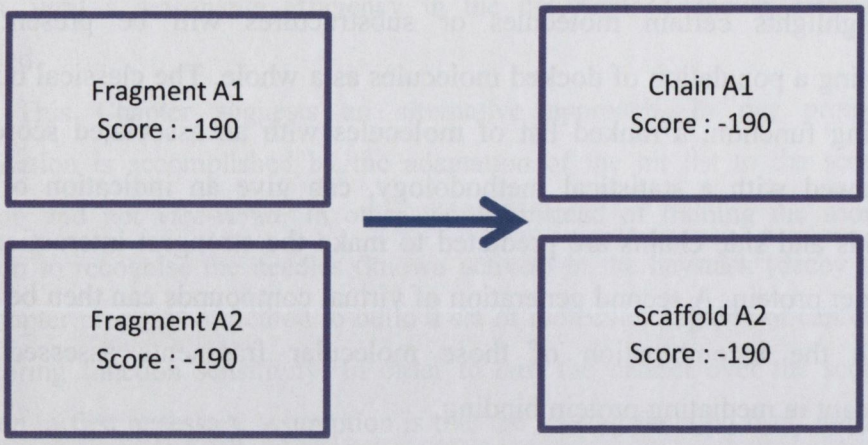


Figure 5-2 Fragments are classified as scaffolds or lateral chains according to the rules described in the Methods section.

Both Chain and Scaffold inherit the score of the parent molecule.

Score distributions are then calculated for each chemically distinct scaffold and chain generated by the fragmentation process. In the second step, a statistical analysis is carried out to discern those fragments where the associated score distribution is significantly better than the distribution of scores obtained for the population of successfully docked molecules as a whole (Figure 5-3).

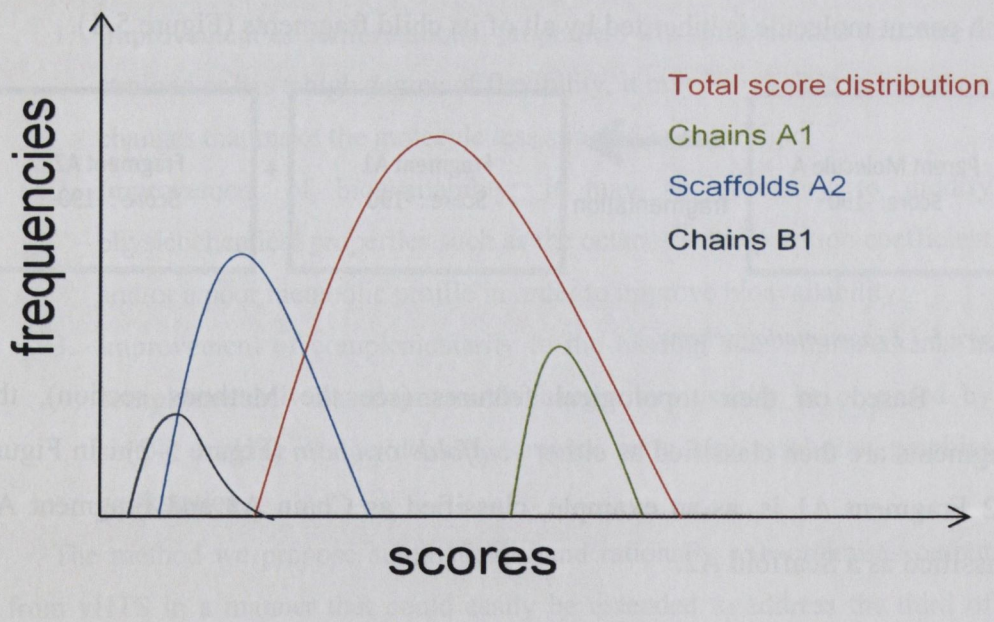


Figure 5-3 In our example, the ensemble of scaffolds labeled A2 are more frequent and receive a better score than chains labeled A1. Scaffold A2 are significant with respect of the total score distribution (in red) and therefore will be selected as a privileged scaffold. Chains A1 will be discarded as they appear to be insignificant. Chains B1 represents another hypothetic structure selected to be privileged by the scoring function.

In the third step the selected fragments are then combined together and synthetic filters are applied to avoid chemically unfeasible molecules. In our

example (Figure 5-4), Scaffold A2 (cyan line in Figure 5-3) and Chain B1 (blue line), are selected as significant, within a certain level of confidence with respect of total score distribution (red line in Figure 5-3), whilst Chain A1 (green line) is not considered to be significant. As a result, Chain B1 and Scaffold A2 are combined together to build a new second generation molecule.

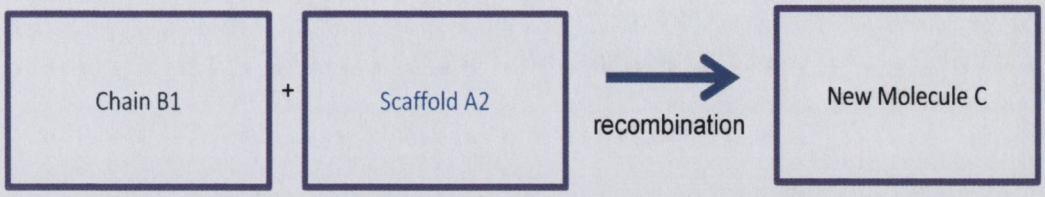


Figure 5-4 Statistically selected chains and scaffolds are recombined together.

Applying this technique the underlying trend that the scoring function exhibits, which might be otherwise be hidden within the data and left unexploited, is instead considered and used to build a focused combinatorial library. The library so obtained is subsequently docked and the new set of scores calculated. An overview of the process is depicted in Figure 5-5.

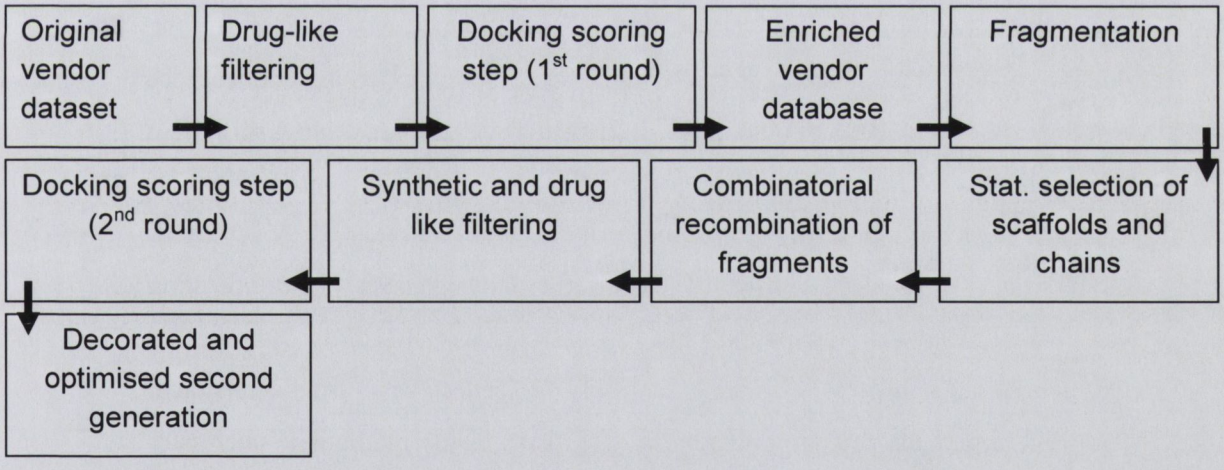


Figure 5-5 Outline of the protocol.

For the purpose of this work, the target identified in Chapter 3 was used. A docking/scoring procedure has been used to screen for potential agonists of the PPARγ, following a similar protocol to the one used in Chapter 3.

5.3 Computational Methods

The computational protocol was developed using PP[16]; to integrate various and commercial software and in-house bespoke scripts: Omega1.8[17] and Corina[18] were used for the generation of conformers, FRED[19] was used to carry out the rigid docking and FILTER[20] was optionally used to select a subset of the initial population based on the drug-likeness of the structures.

The process can be broken up into three main segments: docking/scoring, fragmentation, and generation of the virtual combinatorial library.

5.3.1 Docking/Scoring Method.

One of the numerous publicly available co-crystal structures of PPARγ was selected (PDB identifier 1NYX), and the 3D coordinates retrieved from the PDB. Greater details are described in Chapter 3. The Maybridge database[21] of commercially available compounds (about 60,000 structures) was selected as a source of screening molecules. To reduce the computational demands of the docking stage, this database was clustered using extended-connectivity fingerprints (Scitegic functional class fingerprints with a maximum path length of 6). The cluster centres, plus those members whose distance from their cluster centre exceeded 0.7 in terms of Tanimoto similarity, were retained. A variety of other filters were also applied to ensure the lead-likeness of the docking candidates. Lead-like compounds are typically characterised by sub-optimal target binding affinity[22,23]. Usually, lead-like molecules have simple chemical features, amenable to combinatorial chemistry and medicinal chemistry optimisation[22,23]. A set of 35 known PPARγ gamma agonists retrieved from the WOMBAT database[24] was added to the surviving set of Maybridge compounds. Corina (Molecular Network) was used to create an initial 3D conformation for each potential ligand, using default parameters. Subsequently, Omega[17] (Openeye) was used for the generation of the conformational library. Ten conformers per molecule were generated using the standard energy window and RMSD cut-offs. The rigid docking step was carried out using FRED[19], using the 3D coordinates and reference ligand structure 1NYX to define the binding site.

The Chemgauss2 scoring function was used for the evaluation of the poses and a ROC plot calculated.

Two different calculations were carried out; the first without applying FILTER (indicated by the letter **A** in the result section) and the second applying FILTER as a pre-processing method before the docking (indicated by the letter **B**).

5.3.2 Molecular Fragmentation and Frequency Selection.

The score for the best-docked conformation was retained for each molecule. Those molecules where no conformer could be successfully docked were discarded at this point. The structures were then fragmented, and the most statistically significant fragments across the whole input set identified.

Fragmentation: Single, acyclic bonds between non-hydrogen atoms were considered to be potential cleavage sites. We excluded from this list any bond involving an sp3 hybridised carbon atom not adjacent to a ring, in order to avoid excessive fragmentation of comparatively uninteresting aliphatic chains. This exclusion was achieved by preserving the integrity of aliphatic chains matching the following SMARTS pattern:

```
[CX4;!R;!$(C-[*:R])]
```

For each molecule we created a unique set of fragments generated by all possible combinations of the remaining bond cleavages. The fragmentation was specified in MDL RXN format:

```
$RXN
Break acyclic bond

Comment line
 1 2
$MOL


  2  1  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  1
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  2
  1  2  1  0  0  2  4
M  END
$MOL


  2  1  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  1
    0.0000    0.0000    0.0000 As  0  0
  1  2  1  0  0  0  4
M  END
$MOL


  2  1  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  2
    0.0000    0.0000    0.0000 As  0  0
  1  2  1  0  0  0  4
M  END
```

This process is also illustrated in Figure 5-6. We excluded from the output set any un-fragmented molecules as well as those that contained in excess of 9 cleavage points, to avoid excessive fragmentation and computational expense. In processing the whole input set we derived a list of fragments, each with associated set of scores for the molecule that contained those fragments. The fragments were subsequently split into two different sets on the basis of molecular weight and the presence or absence of rings (number of rings > 0 and molecular weight >= 140). Dummy atoms were used to map the sites where cleavage had occurred.

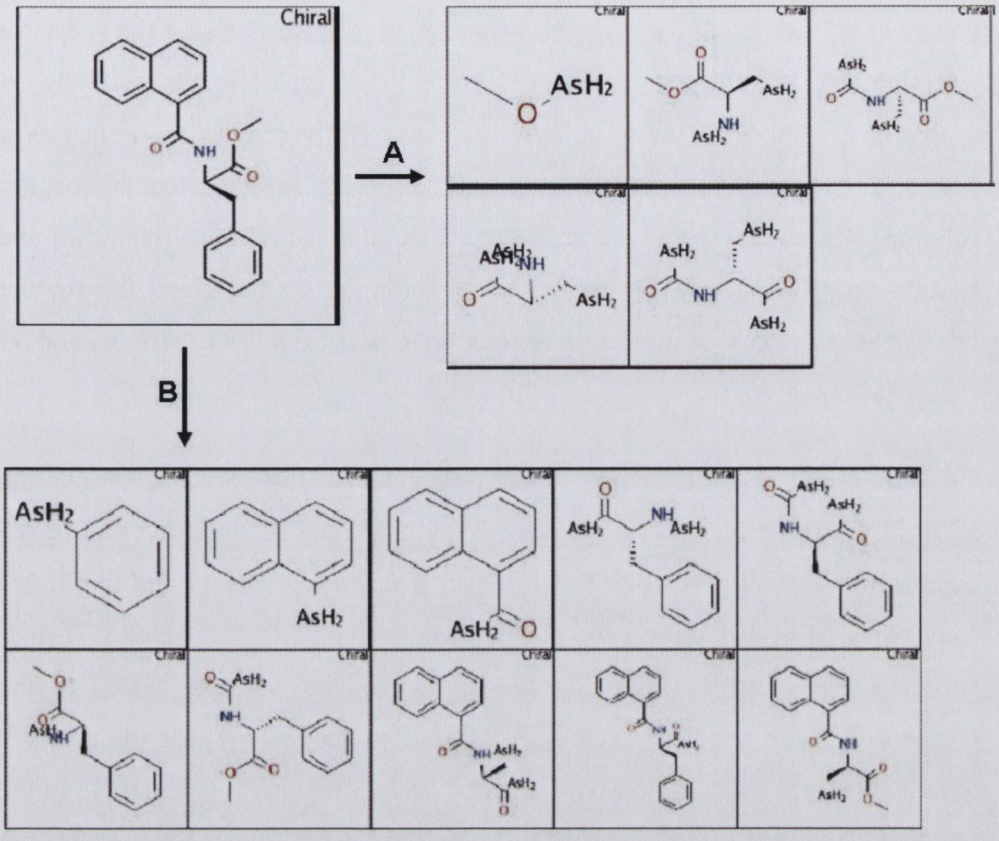


Figure 5-6 An example of fragmentation. Group A represents the fragments identified as chain and the group B those fragments identified as scaffolds. "As" represents a dummy atom.

Selection of interesting fragments: A t-test statistical analysis, assuming unequal variance, was used to discern significant scaffolds and chains. The two-sample t-test analysis tests for equality of the population means underlying each sample. This t-test assumes that the two datasets came from distributions with unequal variances. The t-test is equivalent to an Anova single factor analysis with a sample of two. The following formula is used to determine the statistical value *t*.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad 5\text{-}1$$

where $\bar{x}_1$ $\bar{X}_1$and $\bar{x}_2$ $\bar{X}_2$are the averages of the two samples and s_1 and s_2 are the variances of the populations of size n_1 and n_2 respectively. Using the t-test we therefore compared the set of scores associated with each fragment with the set of scores obtained for the population of successfully docked molecules as a whole, looking for fragments that appear preferentially in the better docking solutions. In order to obtain a moderate number of output fragments the threshold significance level was set to 0.01 for the scaffolds and 0.5 for the lateral chains. Applying these settings led to the selection of 87 scaffolds and 45 chains for input into the next stage of process.

Combinatorial recombination: A protocol that exhaustively recombines the scaffolds with the chains was constructed within Pipeline Pilot. To reduce the combinatorial explosion, HTS and Lipinski filters were applied and duplicates were removed before, during and after the generation of new molecules. A filter to assess the synthetic feasibility of the compounds was also added to the protocol with the intent of excluding chimeric molecules because the protocol does not use any chemical knowledge when performing the fragmentation. In the recombination procedure side chains with a single attachment point are combined with scaffolds fragments presenting an available insertion point in a combinatorial fashion. This procedure is repeated until no further attachment points are available.

5.4 Results

Docking/Scoring

Figure 5-7 and Figure 9 show the Chemgauss2 score distribution of the filtered (set A) and not filtered Maybridge database (set B) after docking/scoring. Also shown is the distribution of score for the combinatorial set obtained after the post-processing. It can be seen that the variance is lower for the combinatorial set than for the original set. A significant shift of the combinatorial set curve towards a lower score with respect to the original set is also observed. Table 5-1 presents the statistical data associated with Figure 5-7. Despite the overall shift in population, the best individual molecules in this second generation have worse scores than their counterparts from the original set. This was caused by the filtering applied only on the recombination step. In other words, the set shown in Figure 5-7 are not fairly comparable as the combinatorial library underwent tight filtering, while the original was not filtered at all. Thus we make the assumption that better scored molecules of the combinatorial set were not included as they may have been filtered out. Thus the procedure was changed and set B was used instead.

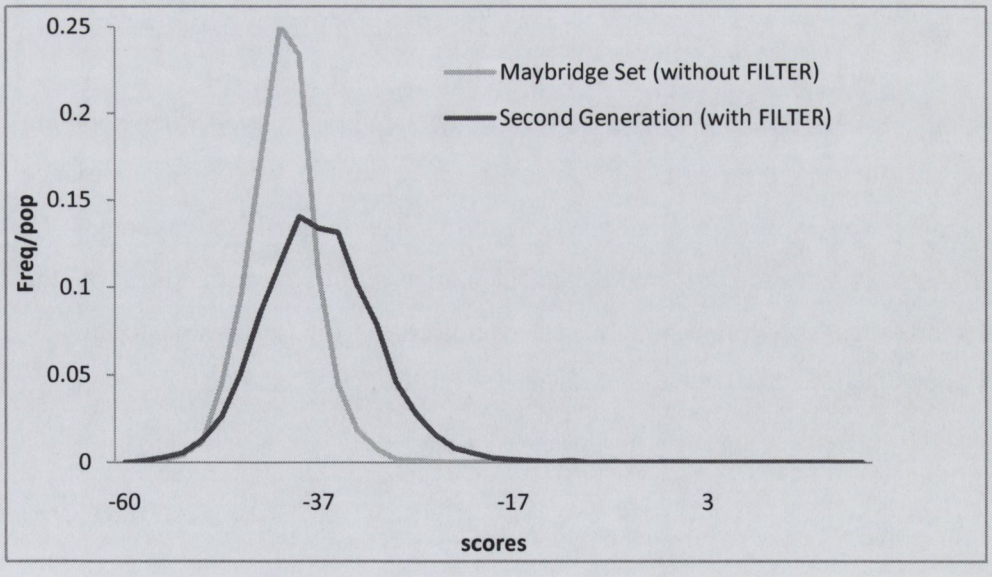


Figure 5-7 Score distribution generated from calculation A.

t-Test: Two-Sample Assuming Unequal Variances		
	Original Set	*Combinatorial Set*
Mean	-35.8087	-39.69432665
Variance	37.28	12.60
Observations	6335	5331
Hypothesized Mean Difference	0	
Df	10443	
t Stat	42.77819	
P(T<=t) one-tail	0	
T Critical one-tail	1.645	
P(T<=t) two-tail	0	
T Critical two-tail	1.960191	

Table 5-1 Statistical details associated with the graph in Figure 8 - calculation A (alpha 0.05).

The results of calculation **B** are shown in

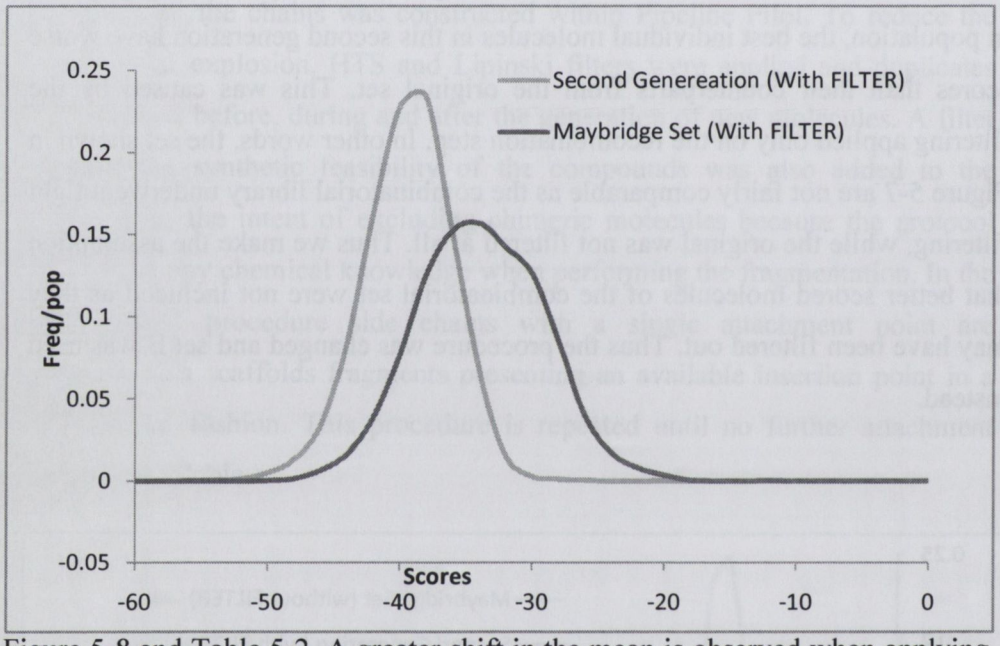


Figure 5-8 and Table 5-2. A greater shift in the mean is observed when applying the extra filtering to the original Maybridge set before the docking. The variance is still lower for the combinatorial than for the original set, although the difference is reduced compared to the calculation A. In contrast with run **A**, the best individual score are now achieved by molecules from the combinatorial.

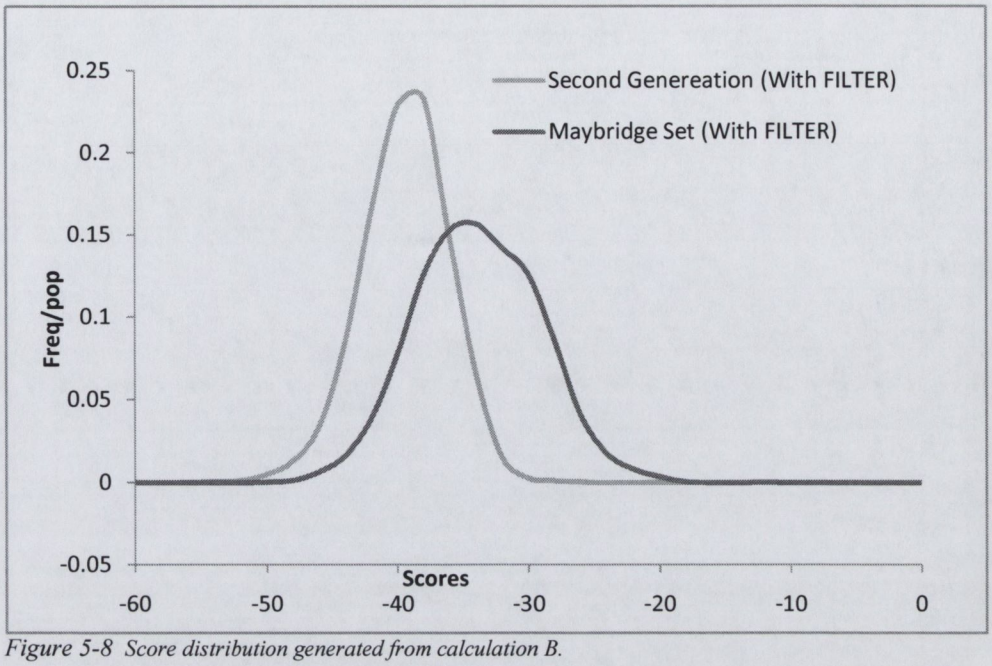


Figure 5-8 Score distribution generated from calculation B.

t-Test: Two-Sample Assuming Unequal Variances		
	Original Set	*Combinatorial set*
Mean	-34.8845	-40.42241884
Variance	22.79227	11.82876191
Observations	5172	8786
Hypothesized Mean Difference	0	
Df	8354	
t Stat	73.01211	
P(T<=t) one-tail	0	
t Critical one-tail	1.645036	
P(T<=t) two-tail	0	
t Critical two-tail	1.960248	

*Table 5-2 Statistical details associated with graph in figure 5B - Calculation **B** (alpha 0.05).*

Figure 5-9 and Figure 5-10 show the ROC plot evaluations for the original and recombined population in calculation B. As shown in Figure 5-9 the model was able to discern known actives from the original decoy set with an accuracy of 0.752, with 1 being the maximum. The ROC evaluation after the docking/scoring of the combinatorial set (Figure 5-10) shows a decreased accuracy of 0.604. After visual inspection we determined that this was due to the fact that some new second generation molecules receive better scores than the known actives.

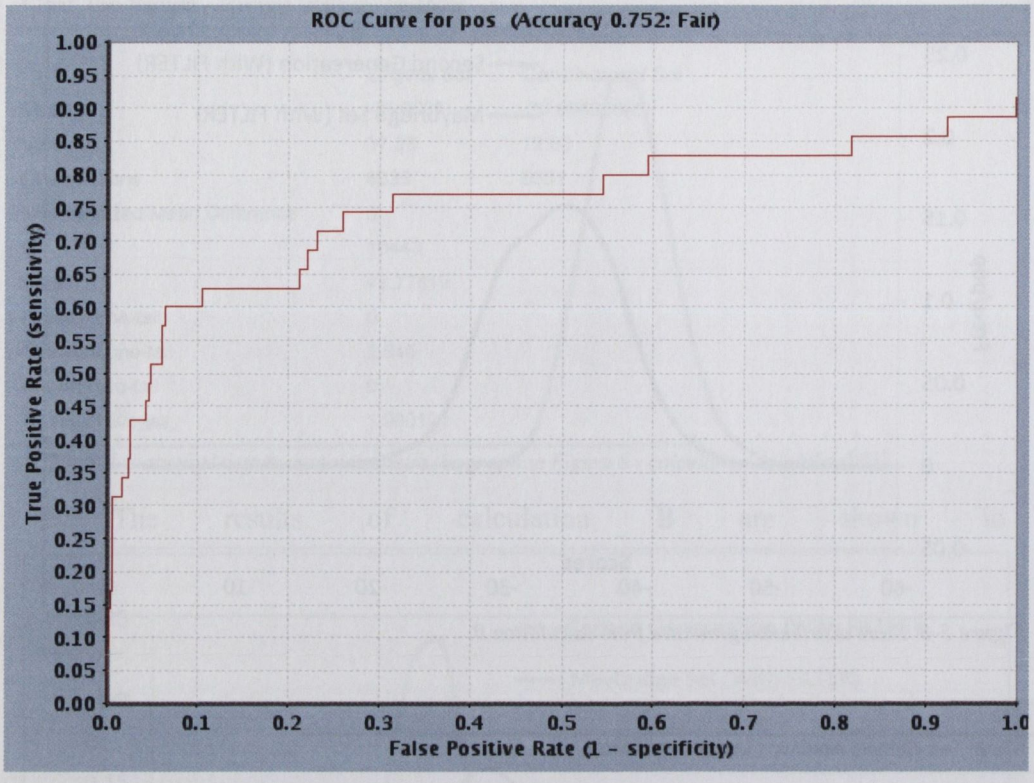


Figure 5-9 ROC curve computed for the original set in calculation B. An accuracy of 0.752 (1 represents the max value) in identifying the seeds is calculated.

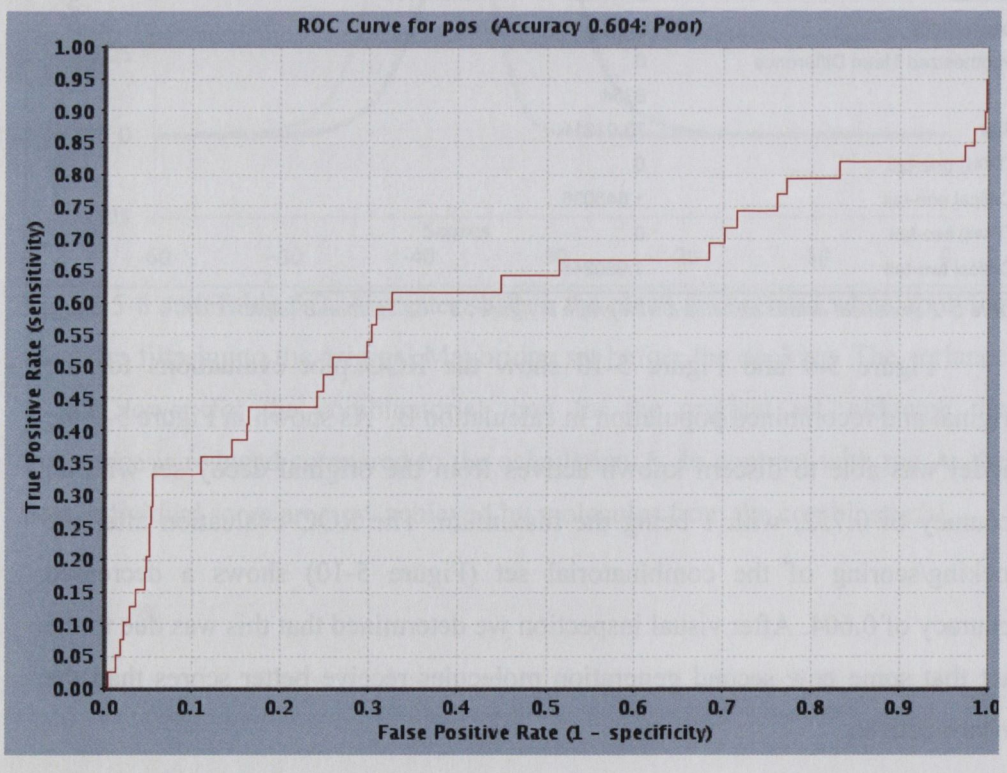


Figure 5-10 ROC curve computed for the combinatorial set in calculation B. An accuracy of 0.604 (1 represents the max value) in identifying the seeds is calculated.

Figure 5-11 and Figure 5-12 show some of the real structures successfully reassembled by the post-processing method in calculation B which

were also present in the original set. In total three known PPARγ agonist and 41 members decoy set were properly re-assembled (Figure 5-12 shows the best scoring three).

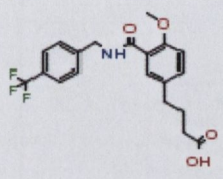
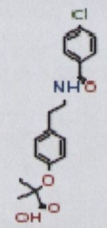
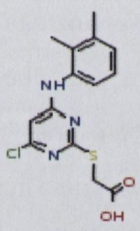
Molecule	Name	Chemgauss2	Index_comb	Frequency
	SMDL-00000200_7	-44.495300	982	2
	SMDL-00092462_1	-38.473854	6295	2
	SMDL-00092461_9	-37.270287	7306	2

Figure 5-11 Known actives successfully reassembled by the post-processing method in calculation B.

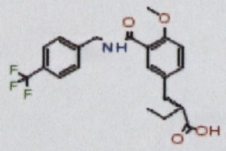
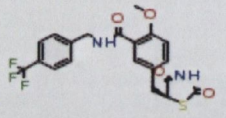
Molecule	Name	Chemgauss2	Index_orig	Index_comb	Frequency
Chiral 	SMDL-00000202_9 SMDL-00000202_9	-50.854374 -50.854374	1	34	2
Chiral 	SMDL-00000198_6 SMDL-00000198_6	-48.802837 -48.802837	7	117	2

Figure 5-12 Top two molecules of the decoy set successfully reassembled by the post-processing method in calculation B.

5.5 Discussion

Post-processing the hit list of a scoring function is a powerful technique to decrease false positive rates and focus on a more relevant region of chemical space. By applying our post-processing technique, it has been possible to automatically generate structures adapted to fit a scoring functions specification. In both calculations (A and B), the score distribution is significantly improved. In calculation A, the best scoring individual molecules from the original set received higher scores than those in the combinatorial set. A visual inspection of the results lead to the conclusion that was this caused by the filtering applied during the combinatorial phase applied in order to reduce the combinatorial explosion. We therefore then decided to apply the same filtering in a pre-processing fashion prior to the docking of the original set, in an attempt to make the challenge equal for both combinatorial and original sets (calculation **B**). The resulting data revealed a more significant difference between the original set and the scoring set. A better score average and lower individual scores are achieved comparing to the original set (calculation **A** and **B)** and to the combinatorial set from calculation **A.** The lower variances observed for both calculations when scoring the combinatorial sets indicates a narrowed and more focused chemical space exploration. Moreover, because the method is led by the scoring function itself, the exploration of the chemical space is target-driven.

The lower ROC score observed in Figure 5-10 compared with that in Figure 5-9 is simply due to the fact that novel molecules produced by the combinatorial approach are receiving better scores than the known actives. As the known actives always receive the same scores, shifting the overall population to improved values results in a lower ROC score. This is an indication that the molecules have been successfully engineered in order to maximise the scoring function results.

By retrieving the higher scored re-assembled structures from available vendor databases, using similarity or seeking exact structural matches, or by carrying out *de novo* chemical synthesis of the new structures the chances of finding active molecules should be increased.

5.6 Future Works

The synthetic feasibility of the output structures could be improved either by incorporating more detailed synthetic screening rules or by integrating the protocol with software for the estimation of the synthetic feasibility of compounds (eg. SYLVA[25]) to prioritise the structures according to their synthetic complexity.

A more restricted fragmentation/recombination could also reduce the reliance on Openeye's FILTER, which is very restrictive in some chemical areas. Another improvement could also be made by including in the statistical calculation the 3D conformations of the fragments therefore implicitly including in the method information about the binding pocket.

This method can be used to optimise against actual experimental data (IC_{50}), should a large enough dataset be available. Finally, fragmental consensus scores produced by different scoring functions could be adopted within this method.

5.7 References

1. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44(3):793-806.
2. Xing L, Hodgkin E, Liu Q, Sedlock D. Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 2004;18(5):333-344.
3. Carta G, Onnis V, Knox AJ, Fayne D, Lloyd DG. Permuting input for more effective sampling of 3D conformer space. J Comput Aided Mol Des 2006;20(3):179-190.
4. Knox AJ, Meegan MJ, Carta G, Lloyd DG. Considerations in compound database preparation--"hidden" impact on virtual screening results. J Chem Inf Model 2005;45(6):1908-1919.
5. Carta G, Knox AJ, Lloyd DG. Unbiasing scoring functions: a new normalization and rescoring strategy. J Chem Inf Model 2007;47(4):1564-1571.
6. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 2003;32:335-373.
7. Graves AP, Brenk R, Shoichet BK. Decoys for docking. J Med Chem 2005;48(11):3714-3728.
8. Baber JC, Shirley WA, Gao Y, Feher M. The use of consensus scoring in ligand-based virtual screening. J Chem Inf Model 2006;46(1):277-288.
9. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42(25):5100-5109.
10. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20(4):281-295.
11. Oda A, Tsuchida K, Takakura T, Yamaotsu N, Hirono S. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. J Chem Inf Model 2006;46(1):380-391.
12. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46(12):2287-2303.
13. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 1996;16(1):3-50.
14. Bohm HJ. A novel computational tool for automated structure-based drug design. J Mol Recognit 1993;6(3):131-137.
15. FRED. Fast Exhaustive Rigid Docking. http://eyesopencom;a software by Openeye Scientific Software.
16. Pipeline. Pilot - a software by Scitegic.www.scitegic.com.
17. OMEGA 1.8.1. distributed by Openeye Scientific Software
18. CORINA 3.6. distributed by Molecular Networks GmbH
19. FRED (version 2.0.1). developed and distributed by Openeye Scientific Software. (URL:http://wwweyesopencom).

20. FILTER. a software by Openeye Scientific Software.http://eyesopen.com.
21. Maybridge. Thermo Fisher Scientific. http://wwwmaybridgecom.
22. Oprea TI, Allu TK, Fara DC, Rad RF, Ostopovici L, Bologa CG. Lead-like, drug-like or "Pub-like": how different are they? J Comput Aided Mol Des 2007;21(1-3):113-119.
23. Oprea TI, Davis AM, Teague SJ, Leeson PD. Is there a difference between leads and drugs? A historical perspective. J Chem Inf Comput Sci 2001;41(5):1308-1315.
24. WOMBAT. WOrld of Molecular BioAcTivity - Sunset Molecular Discovery LLC.http://sunsetmolecular.com.
25. SYLVA. A software by Molecular Networks -;http://www.molecular-networks.com/.

Accepted Publications

Considerations in Compound Database Preparation—"Hidden" Impact on Virtual Screening Results

Andrew J. S. Knox,† Mary J. Meegan,† Giorgio Carta,‡ and David G. Lloyd*,‡

School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland, and Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland

Received May 5, 2005

Structure-based virtual screening (SBVS) utilizing docking algorithms has become an essential tool in the drug discovery process, and significant progress has been made in successfully applying the technique to a wide range of receptor targets. In silico validation of virtual screening protocols before application to a receptor target using a corporate or commercially available compound collection is key to establishing a successful process. Ultimately, retrieval of a set of active compounds from a database of inactives is required, and the metric of enrichment (E) is habitually used to discern the quality of separation of the two. Numerous reports have addressed the performance of docking algorithms with regard to the quality of binding mode prediction and the issue of postprocessing "hit lists" of docked ligands. However, the impact of ligand database preprocessing has yet to be examined in the context of virtual screening and prioritization of compounds for biological evaluation. We provide an insight into the implications of cheminformatic preprocessing of a validation database of compounds where multiple protonated, tautomeric, stereochemical, and conformational states have been enumerated. Several commonly used methods for the generation of ligand conformations and conformational ensembles are examined, paired with an exhaustive rigid-body algorithm for the docking of different "multimeric" compound representations to the ligand binding site of the human estrogen receptor alpha. Chemgauss, a shapegaussian scoring function with intrinsic chemical knowledge, was combined with PLP as a consensus-scoring scheme to rank output from the docking protocol and enrichment rates calculated for each screen. The overheads of CPU consumption and the effect on relative database size (disk requirement) for each of the protocols employed are considered. Assessment of these parameters indicates that SBVS enrichments are highly dependent on the initial cheminformatic treatment(s) used in database construction. The interplay of SMILES representations, stereochemical information, protonation state enumeration, and ligand conformation ensembles are critical in achieving optimum enrichment rates in such screening.

INTRODUCTION

A typical drug discovery research program involves the testing of every available compound in a corporate or commercial compound library using high-throughput biological screening techniques (HTS). Such an approach inevitably leads to high cost and large time scales.[1] One of the major problems emerging in the pharmaceutical industry is that biologically assaying many thousands of compounds for receptor affinity is unlikely to yield the potential number of druglike molecules that realistically exist in a random set, because the screening set is not enriched with compounds that have previously been categorized according to specific "druglike" parameters.

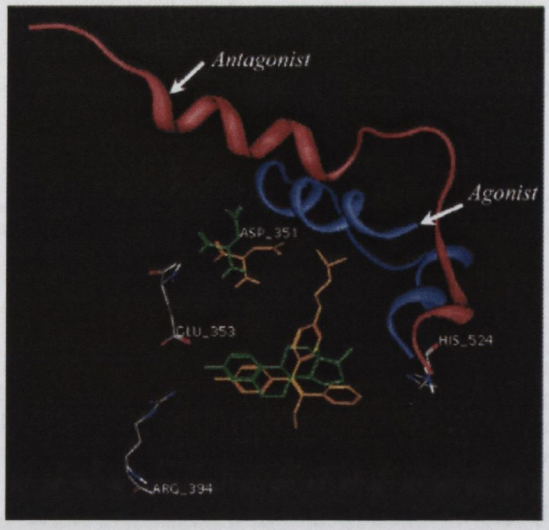
Evidence for this becomes apparent when we see that R&D production of new drugs has remained constant over the past number of years, with major pharmaceutical companies each launching roughly one new drug per year.[2] A recent report also suggested that HTS has generated no lead compounds when used as a sole technique by a large number of major R&D companies.[3]

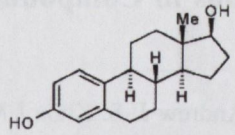
An increase in the number of highly resolved X-ray crystal structures of pharmaceutically relevant biological targets[4] has prompted the use and development of computational techniques to predict ligand affinity to such macromolecules.[5,6] Structure-based virtual screening (SBVS), the most prominent computational technique employed, involves the use of a docking program to generate ligand poses in the active site of a receptor and identification of the optimally docked pose using a scoring function.[7] The scoring function should reflect the complementarity or binding affinity of the ligand for the receptor.[8,9] In the context of a virtual screen, this method should discriminate compounds that bind to the receptor from nonbinders and, where possible, subsequently rank the binders according to their potency.

Virtual screening of compound libraries against therapeutic targets for a particular disease state has become integrated in the drug discovery process and provides a low-cost, rapid, and effective method of enriching a random compound library with the possibility of identifying active species directly. This technique has been applied successfully to compound libraries against a number of targets using various docking programs.[10−12] It is still at an early stage in its development, and the improvement of sampling methods and scoring functions will undoubtedly advance both the reli-

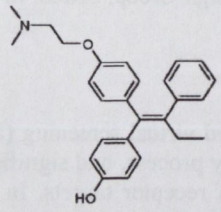
* To whom correspondence should be addressed. Phone: +353-1-6082904. Fax: +353-1-677240. E-mail lloyddg@tcd.ie.
† School of Pharmacy and Pharmaceutical Sciences.
‡ Molecular Design Group, School of Biochemistry and Immunology.

10.1021/ci050185z CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/12/2005





Agonist (Estradiol)



Antagonist (4 - Hydroxytamoxifen)

Figure 1. Binding of agonist (green) and antagonist (orange) induces different helix-12 conformations.

ability and the efficacy of the technique. The significance of these factors will manifest themselves in the form of increased differentiation between active and inactive compounds in a corporate compound collection. The impact of ligand database preprocessing prior to SBVS on enrichment (*E*) has yet to be examined.

In this study, we seek to examine the effects of preprocessing on the prioritization of known active ligands from a database containing both known actives and inactives, where protonation, tautomeric, stereochemical, and conformational states are represented. Research in our group is focused on the identification of novel modulators of human nuclear hormone receptors.[13−17] In this light, we have applied a virtual screening protocol to a validation target of therapeutic importance, estrogen receptor (ER) alpha, where several different preprocessing techniques were used to generate the database of ligands to be screened.

The ER alpha is a nuclear hormone receptor[18] with a buried lipophilic binding site where liganding is highly dependent on hydrogen bonds as well as lipophilic contacts. The binding site is enclosed upon ligand binding, and liganding results in reorganization of the receptor. In particular, helix-12 encapsulates the receptor if an agonist (estradiol) is bound but is prevented from attaining this orientation when an antagonist (4-hydroxytamoxifen) is bound,[19] as in Figure 1. The main differences between the antagonist 4-hydroxytamoxifen and the endogenous agonist ligand (estradiol) are that the antagonist lacks a second hydroxyl group, which prevents hydrogen bonding with His524, but has an extended side chain to accommodate additional interaction with Asp351. The large amount of crystallographic data available and our understanding of the mechanism of action make the ER a viable and therapeutically important target for virtual screening. More specifically, estrogens are mitogenic for ER-positive breast cancer cells, and as 50% of primary breast cancers contain ER alpha,[20] we deemed this to be an important target for application to the optimization of virtual screening approaches.

This study is split into three main stages, where stage 1 involved the assessment of the effect on the prioritization of the actives from the compound collection where protonation, tautomeric, stereochemical, and conformational states are enumerated. To quantify these effects, we report *E* values for each level of preprocessing where a compound database consisting of 1000 compounds (haystack) seeded with 40 ER known actives (needles) was utilized as input. Figure 2 depicts an indicative portion of several of the ER-alpha-active ligands that were included in the needle set.

The second stage consisted of a more rigorous test using 10 000 compounds (claimed inactive, with disclosed biological data) seeded with a single known active, in this case, a previously disclosed potent flexible antagonist,[21] illustrated in Figure 2. Using the same preprocessing protocols, the impact on ranking of the known active was assessed. Application of these processes using a training database allows one to effectively calibrate the docking and scoring protocol prior to deployment on a large corporate or commercial compound library. We also determined the quality of the binding poses produced under each protocol compared with those of the crystal structure 3ERT.[22]

Finally, the third stage divulged an analysis of the same 10 000 compounds seeded with the set of 40 actives used in stage 1. The impact of preprocessing here was assessed through the calculation of false positive (FP) rates for 50% of the true positives.

Why is it necessary to preprocess a database prior to virtual screening? Upon expansion of a database of 2D structures to one of 3D structures, accounting for hydrogen-bond donor and acceptor capabilities is imperative, as changes in the positions of hydrogens in the 2D format denotes conformational changes in the 3D format.[23] Accurate representation of the correct tautomeric and protonated forms of a compound, depending on the ultimate physiological environment of a compound, is extremely important in this case. It is computationally expensive and difficult to assign the most probable state of a compound; thus, all relevant states are typically enumerated as a representation. Similarly, it is necessary to generate all physically relevant stereoisomers of a compound arising from ambiguous stereocenter descriptions.

Conformational changes must also be accounted for as ligands rarely bind to a receptor in their lowest energy form and usually experience some strain upon binding that induces an increase in the energy of the ligand.[24] Computational

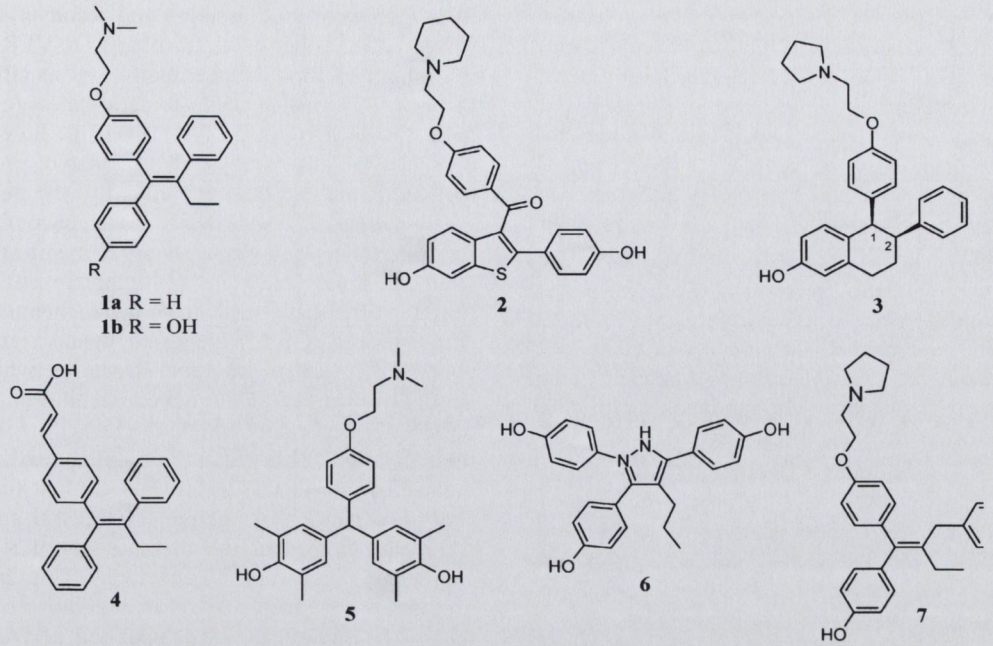


Figure 2. (**1a**) Tamoxifen, (**1b**) 4-hydroxytamoxifen, (**2**) raloxifene, (**3**) lasofoxifene, (**4**) GW5638, (**5**) sumimoto biphenol, (**6**) pyrazole antagonist, and (**7**) flexible antagonist.

methodologies that account for ligand flexibility can be divided into two types. First, generation of conformational ensembles prior to a docking experiment, or "on the fly",[25] can account for conformational changes upon ligand binding. Second, and less widely used, an evolutionary algorithmic approach is taken where core fragments of molecules are "grown" in the binding site of a receptor.[26] The former method of ensemble generation is addressed in this study. Conformer ensemble generation is also shown to be imperative in achieving optimum enrichment for the ER.

For the docking algorithm, this is CPU costly, as multiple forms of the same ligand are docked; however, this overhead is usually only a fraction of the time needed to predict the individual most dominant and relevant state of a molecule using software modules currently available. To further reduce docking CPU time, we chose to use a docking tool with extremely rapid docking times.[27] FRED (Fast Rigid Exhaustive Docking)[28] from OpenEye Scientific was used in combination with a consensus scoring scheme, Chemgauss and PLP, to prioritize the compounds in our docking studies in the ER. The use of FRED integrated with several scoring functions has been previously reviewed in screening experiments for seven targets with different active site characteristics, for example, lipophilic buried cavities, intermediate polarity, and very polar solvent-exposed binding sites. Chemscore was found to be the most applicable general scoring function for SBVS.[29] We take this scoring a step further by implementing a consensus scoring scheme to rank the database postdocking, where a shape-based function with terms accounting for chemical potentials, most significantly hydrogen-bonding interactions, is combined with PLP, an empirical scoring function shown to correlate well with protein−ligand binding affinities. Preliminary docking and scoring studies in our laboratory have highlighted the efficacy of this method combined with FRED over other scoring methods when the ER is used as a target. This study also highlights the unanticipated impact on enrichment rate. Very different in silico enrichments can be achieved depending on the initial SMILES string representation used.

COMPUTATIONAL METHODS

Preparation of Estrogen Receptor (ER) Alpha. Protein target coordinates were extracted from the PDB entry corresponding to the crystal structure of the estrogen receptor complexed with 4-hydroxytamoxifen (3ERT).[22] Structural waters were removed from the monomer, and Macromodel 6.5[30] was used to subsequently re-establish the correct connections in the PDB file. MOE[31] was used to add hydrogens to the protein, and a minimization protocol using a MMFF94 force field was implemented to adjust the positions of hydrogens and keep the heavy atoms fixed at their respective crystallographic positions. The complexed ligand was extracted and utilized as the search box for docking in the receptor. Protein and ligand files were saved in the mol2 format using MOE.

Preparation of Validation Set (1000). Verdonk et al. recently described "virtual enrichments" achieved using decoy sets that are dissimilar from the active set.[32] Here, we implement the same strategy of a "focused" decoy set with similar properties to those of the active set using a validation set (haystack) of 1000 compounds. The haystack was built as follows.

A subset of the Derwent World Drug Index (WDI)[33] was selected using Lipinski's rules,[34] by removing compounds with intrinsically nondruglike properties such as those with molecular weights < 200 or > 550, the number of hydrogen bond donors $0 < x < 6$ and acceptors $0 < x < 10$, and calculated log P values < 7, using an MCL script implemented in the Daylight toolkit.[35] Additional compound filtering was carried out with FILTER.[36] A Perl script was used to select a random subset of compounds from this filtered data set. Over half of all known marketed drugs contain chiral centers, and it was deemed of importance to

represent this in the data set. To this end, 500 molecules with their respective specific active chiral and isomeric data were taken from the WDI using the Daylight toolkit. Subsequently, 460 molecules whose active chirality and isomers were unspecified or "ambiguous" were also selected and added to the data set. This prevented any sources of imbalanced results where the decoy compounds are not representative of active species. The data set was retained in SMILES format.

A set comprising 40 ligands (needles) active for ER alpha was selected from the literature where binding and antiproliferative data were experimentally determined, with activities ranging from nanomolar to low micromolar potency. This active ligand set was then added to the validation set to bring the total number of compounds up to 1000.

Preparation of Stages 2 and 3 Decoy Sets (10 000). To further, more comprehensively test the protocols, a decoy database of 9999 inactive compounds was built from the WDI and CHEMBANK[37] and seeded with a single potent flexible estrogen alpha antagonist[21] (shown in Figure 2) to bring the total database size to 10 000 ligands for stage 2. In the generation of this dataset, all known estrogen actives were excluded, and as with the preparation of the 1000 ligand dataset, all compounds with nondruglike properties were removed using MCL scripting in Daylight and the application of FILTER. As before, a Perl script was used to randomly select the final 9999 compounds from a larger filtered population before seeding with the active ligand. The data set was stored in SMILES format.

Also, a database of 10 000 compounds combining the decoy set from stage 2 and the 40 active ligands (needles) for ER alpha used in stage 1 was prepared for stage 3. Similarly, the procedure was as above for stage 2. Stage 3 was carried out to make sure that the results obtained from stage 2 were not biased because of the incorporation of only a single active antiestrogen.

Preprocessing of Validation Set. *Generation of SMILES.* The Daylight toolkit[35] was used to export a data set of actives in tab format, allowing retention of the correct assignment of isomeric where specified and "ambiguous" SMILES where stereochemistry was not defined in the compound records. All structures were stored in SMILES format using two methods. MOL2SMI (Daylight Toolkit) and CONVERT (Molecular Networks GmbH)[38] were used to produce two alternate SMILES representations, A and B, respectively.

Generation of Tautomeric and Protonated States. The utility TAUTOMER (Molecular Networks GmBH)[39] was used to generate relevant tautomeric states of each molecule in the database. Conversion of SMILES strings to SDF format using UNITY[40] was necessary as strings lose their stereochemical information through canonicalization. To preserve the effects of input SMILES formatting differences, tautomerically processed SDF files were reconverted to SMILES strings using either MOL2SMI or CONVERT as required. This procedure was repeated using TAUTOMER (Openeye Scientific Software)[41] to facilitate a direct comparison of two commercially available and widely used tautomer generators.

The computational utility QUACPAC[42] was used to enumerate physiologically relevant protonation states of the validation set. Again, to preserve stereochemical information, SDF files were used as input and reconverted to SMILES strings using either MOL2SMI or CONVERT as required.

Options to limit enumeration to a specified maximum number of protonated and tautomeric states are possible using both QUACPAC and TAUTOMER; however, for the purposes of this study, all calculable protonation and tautomeric states were enumerated in the pH range 2−14.

Generation of Stereoisomers. Different conformations exist for enantiomers, and it is necessary to manifest this molecular conformational space in a virtual screen as compound libraries often have inadequate stereochemical information denoted. STERGEN[43] identified ligand stereocenters and generated a set of isomeric structures where none was explicitly specified. This step enumerates only the multiple possible stereocenters for "ambiguous" SMILES strings in the validation set as the actives and approximately 50% of the selected haystack strings had explicit (correct) stereochemistry defined. As before, MOL2SMI and CONVERT were used to produce the alternate sets of SMILES strings following stereoisomer generation. This procedure was repeated using FLIPPER[44] as an alternative stereochemistry tool to facilitate a direct comparison of two widely used stereochemical generators.

Conformer Generation. To account for the fact that a molecule can adopt several 3D conformations by rotation about single and acyclic bonds, four 2D to 3D conformer generators were considered in this study, CORINA,[45] OMEGA,[46] RUBICON,[47] and CATALYST.[48]

In all cases, with the exception of CORINA, a single-conformer database and a multiple-conformer database (10 conformers) were generated. In the case of CORINA, generation of only one true conformer (when one discounts ring-flipping variants) is possible and, thus, OMEGA was subsequently used to expand the data to 10 conformers from the original input conformer passed by CORINA.

CORINA uses monocentric fragments with standard bond lengths, angles, and dihedral angles to form a 3D representation of a molecule. Sadowski has shown that CORINA reproduced the correct conformation of bound ligands for almost half of a data set of 639 X-ray structures.[49] OMEGA uses a torsion-driving beam rule-based method to generate conformational ensembles. A SMILES string is reduced to fragments with rotatable bonds, and rules are then applied to regenerate the ensembles. Application of the MMFF force field to refine input geometries allows any high-energy constructs to be minimized. RUBICON uses distance-geometry methods to randomly sample conformations. A rule-based method for establishing geometric constraints based on SMARTS[50] is utilized. CATALYST employs two methods of conformer sampling using a poling algorithm, FAST and BEST. CPU time is a contributing factor to the choice of preprocessing protocol in SBVS, so for the purpose of this study, the FAST option was chosen.

Structure-Based Virtual Screening Protocol. FRED 2.01[28] was utilized in this study to dock all preprocessed compound sets. FRED 2.01 uses a systematic, nonstochastic algorithm to ensure that reproducible results are attained. FRED rigidly and exhaustively examines all poses in an active site, filters by shape complementarity, then ranks by "fitness" prior to scoring using Gaussian functions which have chemical awareness incorporated (e.g., Chemgauss and Shapegauss). The final poses can be scored simultaneously

Table 1. Classification of Database Preprocessing Protocols Applied

level	preprocessing protocol
LEVEL1_X_A/B	SMILES − 1 conformer
LEVEL2_X_A/B	SMILES − 10 conformers
LEVEL3_X_A/B	SMILES − protonation − 1 conformer
LEVEL4_X_A/B	SMILES − protonation − 10 conformers
LEVEL5_X_A/B	SMILES − stereoisomers − 1 conformer
LEVEL6_X_A/B	SMILES − stereoisomers −10 conformers
LEVEL7_X_A/B	SMILES − tautomers −1 conformer
LEVEL8_X_A/B	SMILES − tautomers −10 conformer

utilizing a number of scoring functions such as Shapegauss, PLP, Chemgauss, Chemscore, Screenscore, and Zapbind.

For this study, default operational values were applied and the docking of separately generated input conformers was enabled. Following rigid-body optimization of the ligands in the docking, ranking of the ligand poses using several scoring functions is possible. In internal validation studies using rigid-body docking algorithms and scoring with several scoring functions, either separately or as a consensus[51] scoring function, we have found Chemgauss and PLP[52] used as a consensus score to be the most efficacious scoring method for ranking the docked poses of a lipophilic binding site such as that of ER alpha.[17] The Chemgauss scoring function accounts most significantly for hydrogen-bond interactions. PLP scoring accounts for both simple and steric hydrogen-bond interactions. A recent report reviewed a set of screening experiments for seven targets with different active site characteristics, for example, lipophilic buried cavities, intermediate polarity, and very polar solvent-exposed binding sites. Chemscore emerged as the most applicable general scoring function for SBVS.[27] In previous studies using Chemscore as implemented with FRED 1.1, we have produced results in agreement with this report. However, in this work, we have chosen to utilize the most recent code release—FRED 2.01—where shape docking with a chemical knowledge function (Chemgauss) in combination with PLP delivers superior enrichment results using the same data sets in comparative trials.

Computational Overheads—CPU Time Consumption and Database Size. Despite steady drops in the cost of computational equipment, in parallel with increases in processing power, all computational SBVS experiments have associated overheads in terms of the time required for processing and the resultant physical database size produced. We therefore examine the time involved in producing the various "multimeric" databases examined, and their relative sizes.

Stage 1. Impact of Preprocessing Levels on Enrichment Rate (1000 Compounds). The two different representations of SMILES strings generated according to section 1 of the Preprocessing of Validation Set section above were used as the input for all subsequent enumerations of protonation, stereochemical, and tautomeric states shown as A (MOL2SMI) or B (CONVERT). In the presentation of the data, a qualifier "X" denotes which of four 2D−3D toolkits (CORINA, OMEGA, RUBICON, and CATALYST) was used for the validation set. Table 1 outlines the various preprocessing levels considered in this study. Each level was repeated for each of the SMILES generated; that is, LEVEL1−8_X_A/B is run for both MOL2SMI and CONVERT SMILES string representations, to furnish 64 individual protocols in total, when all four conversion tools are employed.

Table 2. Classification of 10 000 Compound Database Preprocessing Protocols Applied

level	protocol
9	single conformer
10	10 conformers
11	100 conformers
12	protonation + single conformer
13	protonation + 10 conformers
14	stereoisomers + single conformer
15	stereoisomers + 10 conformers
16	tautomers + single conformer
17	tautomers + 10 conformers

Following SBVS in each of the preprocessed databases outlined, enrichment rates for the first 0.5%, 1%, 1.5%, 2%, and 4% of the screen population were calculated. E indicates the ratio of the yield of actives in the hit list (post-screen ranked database population) relative to the random yield of actives as distributed throughout the unranked database and is calculated as follows:

$$E = \left(\frac{\frac{H_a}{H_t}}{\frac{A}{D}} \right) = \frac{H_a}{H_t} \times \frac{D}{A}$$

where H_t is the total number of compounds in the hit list, H_a is the number of known actives in the hit list, A is the number of active compounds in the database, and D is the number of compounds in the database. Thus, if only five actives (H_a) are retrieved in the first 1% of the ranked database (H_t), $D = 1000$, and $A = 40$ actives, the enrichment would be 12.5.

Stage 2. Ranking of a Single Potent ER-Alpha Antagonist in a 10 000-Decoy Compound Set. This set of 10 000 compounds was created to more stringently test the docking and scoring procedure utilizing the optimum levels of preprocessing identified from the above protocols as determined by their respective enrichment values in the 1000-ligand validation set. The efficacy of each protocol was determined according to the ability of each to prioritize the single active ligand contained in the data set. In each case, the protocols that achieved the highest enrichment rates for each level in experiments using the database of 1000 structures were employed and designated as levels 9−17. For example, if, in level 1 processing, SMILES generated using MOL2SMI and subsequent 3D conformers produced the optimum enrichment post docking, this protocol would then be used for single-conformer generation (denoted level 9) for the set of 10 000 compounds. A level of conformer generation producing 100 conformers of each compound in the data set (level 11) was also added to assess the effect of increased conformer generation. The classification of each of the database preprocessing protocols is outlined in Table 2.

Stage 3. Ranking of a Diverse Set of ER-Alpha Antagonists in a 10 000-Decoy Compound Set. To ensure that the results obtained from stage 2 reveal the full potential for variation of E rates, a decoy dataset of 9960 compounds "spiked" with the diverse set of 40 estrogen antagonists was utilized. This would account for any discrepancies that may be observed through diversity, as some of the antagonists

Table 3. Classification of 10 000 Compound Database Preprocessing Protocols Applied

level	protocol
18	single conformer
19	10 conformers
20	100 conformers
21	protonation + single conformer
22	protonation + 10 conformers
23	stereoisomers + single conformer
24	stereoisomers + 10 conformers
25	tautomers + single conformer
26	tautomers + 10 conformers

may not intrinsically have different protonation, tautomeric, and stereochemical states. The classification of each of the database preprocessing protocols is outlined as per Table 3. The efficacy of each protocol was measured by assessing false positive (FP) rates for 50% of the true positives.

RESULTS AND DISCUSSION

Computational Overheads. The dependence of 2D to 3D conversion rates on the nature of input SMILES strings passed to the conversion tools is illustrated in parts a and b of Figure 3. An overview of the CPU time used and the conversion rate achieved using each conformer generation program where a single conformer is constructed from the SMILES strings in the validation set is provided. All programs were run on 32-bit Linux (Fedora) architecture with a 3.00 GHz Intel Xeon CPU with 2 GB of RAM.

Upon examination of the data presented in Figure 3a, we find that CORINA converts the data set with a wall-clock time of <0.05 s/molecule, expending the least CPU time when processing MOL2SMI SMILES representations. However, ~1.6% of the data set remained unprocessed through conversion errors. The significance of this is apparent when

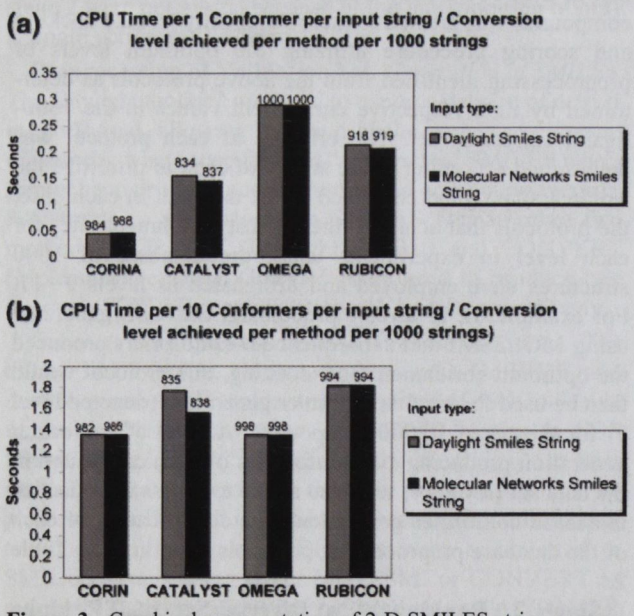


Figure 3. Graphical representation of 2-D SMILES string conversion from MOL2SMI (Daylight) SMILES string and CONVERT (Molecular Networks GmbH) SMILES string to 3-D molecules. Data labels over each column are equivalent to the number of molecules converted in the validation set. (a) When four conformer generation methods were used, one conformer was generated per molecule. (b) Conformational ensembles of 10 conformers were generated per molecule, for each of the four methods.

the data set is scaled to a more representative 1 million compounds—a database size employed for typical discovery screening. In such a set, up to 16 000 molecules could remain unprocessed, leading to potential "hits" being overlooked through this initial attrition. Similarly, CATALYST and RUBICON exhibit string-parsing errors, which could cumulatively impact on the quality of the database produced. OMEGA converts 100% of the data set in <0.29 s/molecule when input SMILES were generated using either MOL2SMI or CONVERT strings.

Good and Cheney have previously demonstrated that increasing sampling increases the chance of producing conformers closer to the bioactive conformer in a crystal structure.[53] Figure 3b shows conformational ensemble generation where 10 conformers of each molecule are produced. For CORINA and OMEGA, the conversion rates remain relatively unchanged, scaling as expected. With RUBICON, increased conformer sampling increases the performance of both conversion time and rate. CATALYST is seen to produce a single conformer per molecule in ~0.15 s from a SMILES string, but when the production of 10 conformers per molecule is undertaken, we observe an increase in processing time to 1.75 s per molecule. This is in line with all of the conversion tool rates. However, the low conversion rate observed when using CATALYST could mitigate against its incorporation as a large-scale preprocessing tool for database construction in SBVS.

Stage 1. Effect of Preprocessing Levels on Enrichment Rate. Table 4 depicts LEVEL1−8 of preprocessing using MOL2SMI and CONVERT SMILES string representations of the validation set of compounds. These data illustrate clearly the impact on the enrichment rate when protonated, tautomeric, stereochemical, and multiple conformations of the validation set are docked and scored, and they also illustrate the clear significance of using alternate SMILES strings as input for such SBVS protocols.

LEVEL1 and LEVEL2 depict "entry level" preprocessing, where consideration of either one (LEVEL1) or 10 (LEVEL2) conformers per ligand in the SBVS protocol is applied, in this instance, using both MOL2SMI- (X_A) and CONVERT-generated (X_B) SMILES input. In all cases, the enrichment is calculated using the number of molecules converted by each 2D−3D conversion program as input. We observe LEVEL1_OMEGA_A and LEVEL1_OMEGA_B outperforming other levels in LEVEL1. A large difference in enrichment rates can also be seen between the two alternate representations of SMILES strings. Screening a database prepared with LEVEL1_OMEGA_A, for instance, achieves the maximum possible enrichment until 1.5% and then reduces. LEVEL1_OMEGA_B, however, exhibits poorer enrichment rates in the same section of the database. The same trend is equally observed using RUBICON. Both RUBICON and CATALYST failed to convert ~17% and ~8.2% of the database, respectively, with lowered enrichment rates resulting. RUBICON conformer generation results in the lowest enrichment rate in the first 1% of the data set when only a single conformer is presented in the SBVS, but increasing the sampling rate, as with OMEGA and CORINA, considerably improves the outcome. Also, the conversion rates increase dramatically with RUBICON, converting 99.4% of the database. LEVEL2_CORINA_A and LEVEL2_OMEGA_A perform almost equally well

Table 4. Enrichment Results Obtained for Each Level (LEVEL1−8) of Preprocessing[a]

level	subset size % 0.5	1	1.5	2	4	database size
LEVEL1_CATALYST_A	20.85	20.85	19.46	16.68	9.38	834
LEVEL1_CATALYST_B	20.93	20.93	19.46	16.68	9.94	837
LEVEL1_CORINA_A	19.68	22.14	21.32	19.68	17.22	984
LEVEL1_CORINA_B	19.76	22.23	21.40	19.76	16.67	988
LEVEL1_OMEGA_A	**25.00**	**25.00**	**25.00**	**21.25**	**15.63**	**1000**
LEVEL1_OMEGA_B	25.00	22.50	21.66	18.75	11.88	1000
LEVEL1_RUBICON_A	18.36	18.36	19.89	17.21	11.48	918
LEVEL1_RUBICON_B	13.79	18.38	15.31	16.08	13.21	919
LEVEL2_CATALYST_A	20.88	20.88	19.48	18.78	14.09	835
LEVEL2_CATALYST_B	20.95	20.95	19.55	17.80	14.14	838
LEVEL2_CORINA_A	24.55	24.55	24.55	24.55	17.19	982
LEVEL2_CORINA_B	24.65	22.09	23.00	23.42	16.64	986
LEVEL2_OMEGA_A	**24.95**	**24.95**	**24.95**	**24.95**	**19.34**	**998**
LEVEL2_OMEGA_B	24.95	24.95	24.95	23.70	17.47	998
LEVEL2_RUBICON_A	24.85	22.37	21.54	21.12	16.77	994
LEVEL2_RUBICON_B	24.85	22.37	21.54	21.13	16.77	994
LEVEL3_CATALYST_A	20.65	18.59	15.14	14.46	9.80	826
LEVEL3_CATALYST_B	19.66	22.12	18.02	14.75	11.06	983
LEVEL3_CORINA_A	24.60	22.14	22.96	22.14	18.45	984
LEVEL3_CORINA_B	24.48	22.03	22.84	20.80	17.74	979
LEVEL3_OMEGA_A	**25.00**	**25.00**	**25.00**	**22.50**	**18.13**	**1000**
LEVEL3_OMEGA_B	25.00	25.00	23.33	23.75	14.38	1000
LEVEL3_RUBICON_A	19.40	19.40	21.02	20.61	15.76	970
LEVEL3_RUBICON_B	23.85	23.85	22.26	20.27	16.10	954
LEVEL4_CATALYST_A	20.65	20.65	19.27	17.55	13.42	826
LEVEL4_CATALYST_B	24.33	24.33	22.70	19.46	16.42	973
LEVEL4_CORINA_A	24.53	24.53	24.53	24.53	14.10	981
LEVEL4_CORINA_B	24.40	24.40	24.40	24.40	16.47	976
LEVEL4_OMEGA_A	24.93	24.93	24.93	23.68	15.58	997
LEVEL4_OMEGA_B	**24.93**	**24.93**	**24.93**	**23.68**	**16.20**	**997**
LEVEL4_RUBICON_A	24.85	22.37	23.19	22.37	16.77	994
LEVEL4_RUBICON_B	24.85	24.85	24.85	22.37	17.40	994
LEVEL5_CATALYST_A	20.65	20.65	19.27	17.55	11.36	826
LEVEL5_CATALYST_B	24.33	21.89	21.08	18.24	13.99	973
LEVEL5_CORINA_A	24.45	22.00	21.19	20.78	16.50	978
LEVEL5_CORINA_B	24.43	24.43	21.17	20.76	15.27	977
LEVEL5_OMEGA_A	**25.00**	**22.50**	**21.66**	**16.25**	**13.75**	**1000**
LEVEL5_OMEGA_B	20.00	22.50	21.66	16.25	13.75	1000
LEVEL5_RUBICON_A	18.44	18.44	18.44	16.14	13.25	922
LEVEL5_RUBICON_B	23.00	23.00	23.00	19.55	13.80	920
LEVEL6_CATALYST_A	20.65	20.65	20.65	19.62	12.90	826
LEVEL6_CATALYST_B	24.33	24.33	24.33	23.10	13.98	973
LEVEL6_CORINA_A	24.38	24.38	24.38	21.94	17.06	975
LEVEL6_CORINA_B	24.33	21.89	21.08	20.67	15.20	973
LEVEL6_OMEGA_A	24.95	24.95	23.29	23.70	17.47	998
LEVEL6_OMEGA_B	**24.95**	**24.95**	**24.95**	**23.70**	**16.84**	**998**
LEVEL6A_OMEGA_B	24.93	24.93	23.26	23.68	14.33	997
LEVEL6_RUBICON_A	24.60	22.14	22.96	20.91	15.38	984
LEVEL6_RUBICON_B	24.58	24.58	24.58	20.88	15.36	983
LEVEL7_CATALYST_A	22.08	22.08	19.13	16.56	10.49	883
LEVEL7_CATALYST_B	19.78	22.25	21.43	19.78	12.36	989
LEVEL7_CORINA_A	24.73	22.25	23.07	22.25	16.07	989
LEVEL7_CORINA_B	24.68	22.20	21.39	20.97	14.81	987
LEVEL7_OMEGA_A	**25.00**	**25.00**	**25.00**	**22.50**	**15.00**	**1000**
LEVEL7A_OMEGA_A	25.00	25.00	25.00	23.75	17.50	1000
LEVEL7_OMEGA_B	25.00	25.00	25.00	21.25	14.38	1000
LEVEL7_RUBICON_A	19.00	20.58	20.58	17.81	12.47	950
LEVEL7_RUBICON_B	23.40	21.06	18.72	16.38	10.53	936
LEVEL8_CATALYST_A	22.10	19.89	20.67	18.79	13.26	884
LEVEL8_CATALYST_B	24.75	24.75	24.75	21.04	14.23	990
LEVEL8_CORINA_A	24.95	24.95	24.95	23.70	15.59	998
LEVEL8_CORINA_B	24.95	24.95	24.95	22.46	14.35	998
LEVEL8_OMEGA_A	24.95	24.95	24.95	23.70	15.59	998
LEVEL8_OMEGA_B	**24.98**	**24.98**	**24.98**	**22.48**	**14.36**	**999**
LEVEL8_RUBICON_A	24.90	22.41	21.58	19.92	14.32	996
LEVEL8_RUBICON_B	24.93	24.93	23.26	21.19	13.71	997
theoretical optimal value	**25.00**	**25.00**	**25.00**	**25.00**	**25.00**	**1000**

[a] Optimum enrichments of each level obtained are highlighted in bold.

with enrichment rates up to 2% of 24.55 and 24.95, respectively.

LEVEL3 illustrates the impact on virtual screening arising from the introduction of a step to incorporate protonation states in the processing of the screening database. Single conformers generated by OMEGA contribute to achieving higher enrichment values and are enhanced by the addition of protonation, as seen in Table 4. Initial observation clearly shows that LEVEL3_OMEGA_A achieves a superior enrichment rate over the 4% of the database when compared with the others. LEVEL3_RUBICON_A produces the lowest *E*. Both CATALYST and RUBICON exhibit a large deviation in *E* between using two alternate SMILES representations as input. RUBICON generates higher *E* values using CONVERT SMILES strings.

A remarkable increase is observed using RUBICON where multiple conformers with a treatment of protonation are considered. LEVEL4_RUBICON_B accomplishes an *E* of 24.85, just slightly lower than that of LEVEL4_OMEGA_B. CORINA performs well also with MOL2SMI input giving an *E* of 24.53 across the first 2% of the ranked database. However, the opposite is seen with CATALYST at this stage in the process. CATALYST converts 97.3% of the database from CONVERT SMILES strings; however, only 82.6% is converted using MOL2SMI. Again, the effect of alternating SMILES string representations is highlighted here.

In general, protonation has a considerable influence on the orientation of a docked ligand, as the conformers produced by each 2D−3D tool will vary according to the positions of the hydrogens on a molecule. OMEGA has been previously shown to achieve significantly better results when the input structure is supplied from CORINA.[54] Additional refinement and minimization of the CORINA input structure under the MMFF force field before ensemble generation using OMEGA also enhances the performance. These concepts were integrated in the current version of OMEGA used in this study, and so, in general, the performance using OMEGA appears to be superior.

In the absence of specification of the chemical structure, the addition of arbitrary stereochemical information may introduce stereoisomers and regioisomers that may not actually exist in reality and impact on the enrichment rate achieved. STERGEN was used in the context of preserving SMILES strings with assigned stereochemical information and to assign multiple stereochemical representations to those ligands with partial or ambiguous information. The value of this is immediately apparent as typical commercial compound libraries often contain incomplete or no specific stereochemical information for a percentage of compound entries. LEVEL5 denotes how the consideration of stereochemical information assists Chemgauss and PLP in prioritizing the actives in the validation set only in the case of LEVEL5_RUBICON_B. Interestingly, this effect was more pronounced using FLIPPER (LEVEL6A_OMEGA_B), which we used to enumerate *all* possible stereoisomers for *all* ligands—without preserving the information of those with defined chirality. Accordingly, a slightly lower *E* is achieved as decoys were introduced to the docking and scoring procedure. There is clearly a fine balance which needs to be achieved in generating realistic stereochemical information for "ambiguous" structures in a data set and preserving the

known structural information down through the processing stages.

The introduction of a tautomer treatment to database preprocessing demonstrated that, in combination with a single conformer consideration, processed using LEVEL7_OMEGA_A/B, a higher enrichment could be achieved than when processing multiple conformers (LEVEL8_OMEGA_A/B). However, for all other tools, a benefit is observed on addition of multiple conformers to the database treatment. All possible tautomeric forms were enumerated, not solely the one considered to be most prevalent in solution at physiological pHs. The addition of tautomeric representations, for consideration in the SBVS protocol, makes the scoring function work harder but helps in more finely discriminating actives from inactives, as is demonstrated in LEVEL7_OMEGA_A/B. To facilitate a direct comparison of tools, LEVEL7A_OMEGA_A made use of an alternate processing utility, TAUTOMER (Openeye Scientific Software). A slight enrichment increase is observed over the 4% of the ranked data set using this tool. Tautomeric variation of a ligand also impacts on conformer orientation and, thus, the poses generated in a binding site during docking and prior to scoring. Although an additive effect on enrichment is observed by the addition of tautomers as with protonation, it is difficult to immediately establish if the highest scoring tautomer ranked is actually representative of the most prevalent species in vivo.

Finally, to ensure that the values in Table 4 for each level were statistically significant, we conducted an ANOVA two-way analysis of variance using Minitab 14.20. At a confidence level of 99%, a statistical difference was observed between programs within each level ($p < 0.0001$). A Friedman two-way analysis of variance was also used to decide the optimum levels of preprocessing using the sum of ranks from enrichments observed in each level. We chose LEVEL5_OMEGA_A rather than LEVEL5_OMEGA_B as being superior because enrichment in the first 0.5% was 25 compared with 20, respectively. This is more important as when searching a larger database it is preferable to search the smallest ranked hit list possible (i.e., 0.5−1%).

The Impact of Alternate SMILES Representations. If alternate representations of SMILES strings are used as input 2D structures, such as those produced by MOL2SMI and CONVERT, radically different effects on the enrichment rates are achieved when the same preprocessing protocols are utilized in advance of rigid docking experiments. For example, LEVEL1_RUBICON_A and LEVEL1_RUBICON_B use the same preprocessing protocol of generating a single conformer, yet enrichment rates in the first 0.5% of the database were 18.36 and 13.79, respectively. Major enrichment differences can be seen throughout Table 4 because of variation in initial SMILES depictions. This effect is not restricted to any individual 2D−3D conformer generator studied, as is evident from LEVEL5_OMEGA_A/B, where enrichments are 25 and 20, respectively. These differences are possibly caused by the way in which each conformer generator parses a SMILES string. For example, programs that use a library of SMARTS strings to generate segments of a compound from a SMILES string would produce different conformers depending on the initial representation.

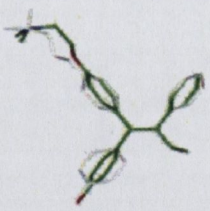
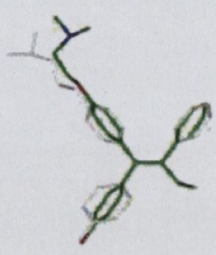
To emphasize the significance of this observation, we have endeavored to test six alternate SMILES string representations of the known ER-active modulator hydroxytanoxifen and compare the generated conformers with those of a set of six conformers produced from a *single* SMILES string representation of the ligand. The root mean square difference (RMSD) of each conformer generated was compared with the cocrystal structure of bound hydroxytamoxifen in the ER-active site and each conformer was also docked using FRED in the binding site of 3ERT to determine the best ranking conformer score. Table 5 shows the results.

It is clear that different SMILES representations of the same molecule produce different conformers, and as we have shown in Table 3, this has a clear impact on enrichment values obtained also. Alternating the initial SMILES string produces a range of conformers, the best of which exhibits an RMSD of 0.75 Å when compared to the crystal structure of hydroxytamoxifen, and results in the highest scoring docked pose (−55.90) using FRED 2.01. The immediate conclusion to draw from these data is that, nominally, while each individual molecule is constant, not all SMILES representations are equal, and a hitherto unexplored link exists between this simplest cheminformatic treatment of molecular representations and the results obtained in 3D molecular recognition studies involving SBVS of this nature.

Computational Overheads—How Big Is Too Big? As previously discussed, each level of "multimeric' ligand treatment increases the physical size of the database under consideration. This size limit begins to impact on the feasibility and practicality of the SBVS when the base number of compounds is large. Figure 4 illustrates how the size of the data set using each protocol differs. This poses a significant quandary with respect to virtual screening. Database size must be reduced as much as possible to keep disk space at an affordable minimum, and also to prune the overall number of molecules to be screened from a performance cost/benefit perspective. If we consider the relative scaling for LEVEL8, 1000 compounds (a 312k SMILES file) expands to a 3D SDF file containing 27 212 multimeric forms (73 Mb). Scaling this up to a database of 1 000 000 screening compounds will accordingly generate 25−30 million ligands for consideration in both the preprocessing steps and the actual docking procedure. While the ability of newer software to utilize and read compressed data files may alleviate this difficulty, the issues of how large a data set one can afford to use, and correspondingly which preprocessing method one will employ, will ultimately vary according to available resources.

Stage 2. Ranking of a Single Potent ER-Alpha Antagonist in a 10 000-Molecule Compound Set. Tautomerism and protonation states have little synthetic consequence, although their consideration may be crucial in enhancing the enrichment level of a virtual screening protocol. A successful virtual screening protocol should not only prioritize compounds for biological testing but also provide useful information about the stereochemistry of the compound should synthesis be required. We therefore sought to utilze from the previous steps protocols that provide the highest possible enrichment but also give the most information about the exact 3D or stereochemical nature of the compounds identified. The importance of this with respect to the estrogen receptor is clear when one considers that the *E* isomer of the ligand

Table 5. Comparison of RMSD of Alternate SMILES-Generated Conformers versus Conformer Generation Taken from a Single SMILES String[a]

6 SMILES permutations	RMSD	TOP SCORE PLP = -55.90
CC\C(c1ccccc1)=C(/c2ccc(O)cc2)c3ccc(OCCN(C)C)cc3	0.75	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3	0.76	
CC\C(=C(/c1ccc(cc1)O)c1ccc(cc1)OCCN(C)C)c1ccccc1	1.07	
CN(C)CCO(c1ccc(cc1)C(\c2ccc(O)cc2)=C(c3ccccc3)\CC)	1.25	
CN(C)CCO(c1ccc(cc1)C(=C(c2ccccc2)\CC)\c3ccc(O)cc3)	0.77	
CN(C)CCOc(ccc(c1)C(\c(ccc(c2)O)c2)=C(c(cccc2)c2)\CC)c1	1.23	
6 SMILES of equivalent permutation	**RMSD**	**TOP SCORE PLP = -54.91**
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3	0.6	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc4	0.76	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc5	0.99	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc6	1.25	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc7	1.13	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc8	0.89	

[a] RMSD values were calculated using the OpenEye OEChem RMSD calculator, which fully accounts for automorphisms (self-symmetry of the molecules being compared). The crystal structure of hydroxytamoxifen (white) is superimposed with the conformer of hydroxytamoxifen (colored by atom) generated from a SMILES string with the lowest RMSD value. The PLP values shown correspond to the lowest RMSD structure docked in the active site of 3ERT.

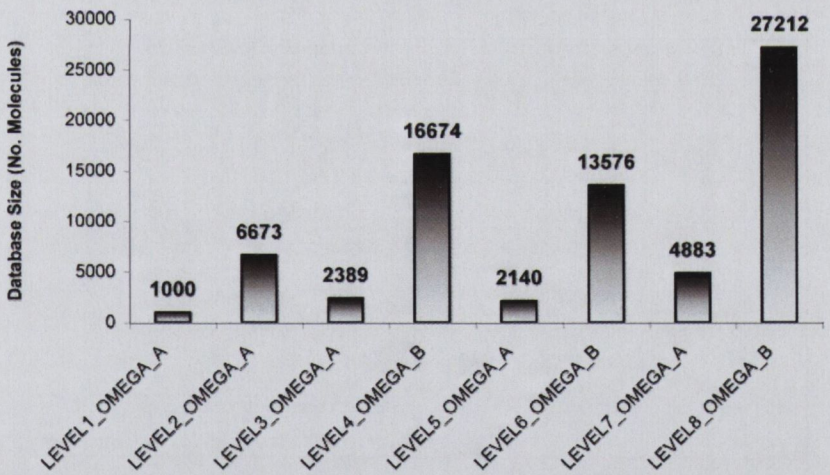


Figure 4. Column data labels show the number of molecules produced from 1000 input SMILES strings using the optimum preprocessing techniques as previously evaluated.

tamoxifen exhibits estrogenic activity, while the *Z* isomer exhibits antiestrogenic potency. From our initial individual screening runs, it appears that the addition of stereochemical information to ensemble conformer generation provides the optimum benefit in pre- and post-screening, where the compound collection lacks defined stereochemical information.

To assess the utility of the various screening levels in a "real-world" application, our dataset of 10 000 ligands containing only a single known active was used. The impact of the various protocols on the ability of SBVS to identify the active is given in Table 6. Optimal ranking is achieved using only 10 conformers, with no additional benefit seen when expanded to a treatment of 100 conformers per ligand. In these cases, all levels involved generation of conformers using OMEGA. Ligand conformer sampling is highly important in SBVS when the active or binding site is deemed to be flexible. The estrogen receptor exhibits a relatively rigid binding site upon examination. Nonetheless, slight variations in binding site residue positions occur, and a greater treatment of ligand flexibility is expected to improve the enrichment rate. Bostrom et al.[54] deemed 1000 conformations per molecules to be adequate sampling in the context of a virtual screen; while we agree that such a treatment is highly

Table 6. Ranking of a Single Active by FRED 2.01 from a Screening Database Totaling 10 000 Druglike Compounds Using Preprocessing Protocol Levels 9−17.

level	rank	protocol details
9	146	single conformer
10	**1**	10 conformers
11	**1**	100 conformers
12	399	protonation + single conformer
13	**6**	protonation + 10 conformers
14	254	stereoisomers + single conformer
15	**6**	stereoisomers + 10 conformers
16	163	tautomers + single conformer
17	**3**	tautomers + 10 conformers

important when the target is flexible, no apparent benefit is seen for the estrogen receptor above 10 conformers. This emphasizes the requirement for SBVS protocol optimization on a target-by-target basis.

This comparative study highlights the wide impact on enrichment and variance possible when moving from single- to multiple-conformer treatments (level 9→10) and where protonation (level 9→12, level 10→13, and level 12→13), chirality (level 9→14, level 10→15, and level 14→15), and tautomer treatments (level 9→16, level 10→17, and level 16→17) are explored in SBVS.

As a final comparison of the impact of preprocessing, we sought to elucidate the RMSD between docked solutions for hydroxytamoxifen-generated (colored by atom) conformers generated using each of the above preprocessing protocols and the ligand cocrystal structure pose (white) taken from 3ERT. All of the docked structures are shown to be close to the crystal structure of hydroxytamoxifen (Figure 5). The docked structure of the single potent antiestrogen used to seed the database of 10 000 structures is also overlayed (colored by atom) to show its equivalent docked solutions for each processing level. Interestingly, although a minimal difference is observed in the RMSD values, a large difference is seen in the relative rankings of the seed antiestrogen.

Stage 3. FP Rate of 40 ER-Alpha Antagonists in a 10 000-Molecule Compound Set. False positive rates for the ER-alpha antagonist set are outlined in Table 7. It is evident that the use of 10 conformers again gives the lowest

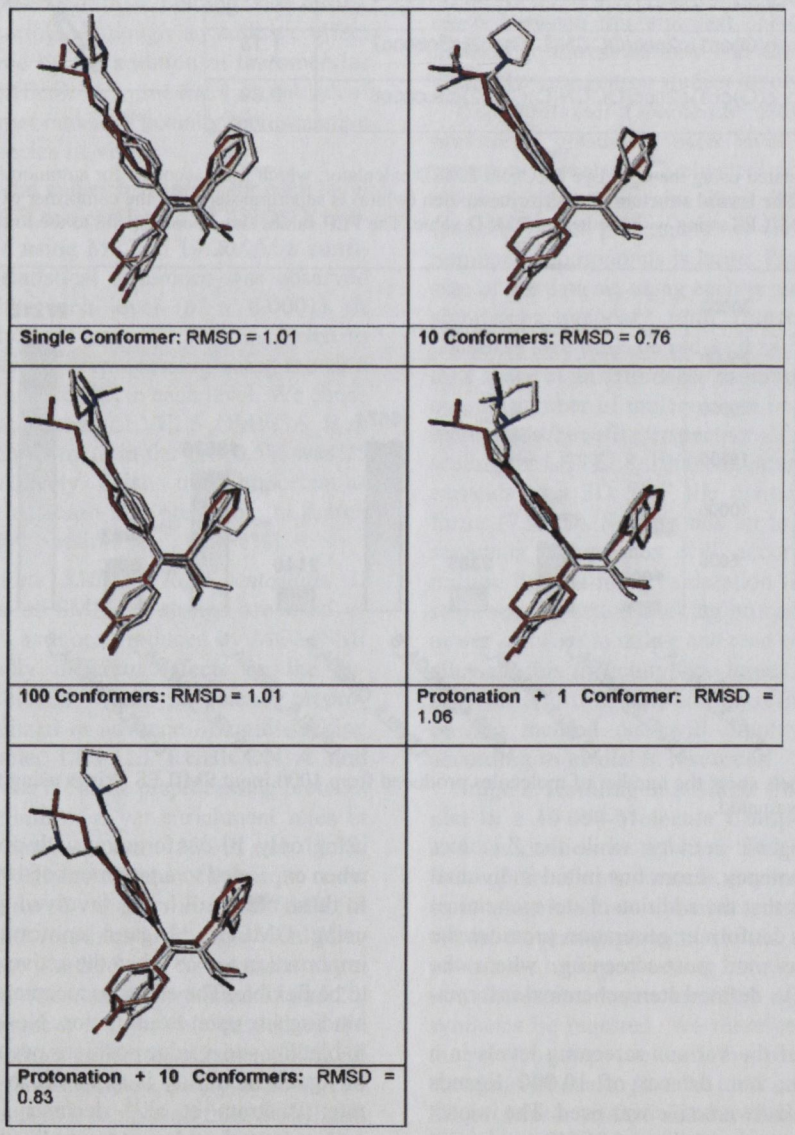


Figure 5. RMSD difference between docked conformers generated using each of the above preprocessing protocols and the ligand crystal structure taken from 3ERT.

Table 7. False Positive Rates for Recovery of 50% of the True Positives

level	FP 50%	protocol details
LEVEL18	0.44	single conformer
LEVEL19	0.35	10 conformers
LEVEL20	1.10	100 conformers
LEVEL21	0.58	protonation + single conformer
LEVEL22	2.30	protonation + 10 conformers
LEVEL23	0.65	stereoisomers + single conformer
LEVEL24	1.44	stereoisomers + 10 conformers
LEVEL25	1.87	tautomers + single conformers
LEVEL26	1.50	tautomers + 10 conformers

FP rate at 0.35% for a true positive rate of 50%, corroborating results from the previous section. The use of a more broad and diverse set of antagonists spiked within a large decoy set of 9960 enables us to definitively show the influence of each level of preprocessing on the database. We observe that the database conversion rate is also optimal using multiple conformers only. The addition of multiple conformers with protonation or stereochemical generation results in a higher FP rate and, thus, an increase in the number of random molecules ranked among the actives. The addition of tautomers also increases the FP rate.

CONCLUSIONS

We have endeavored to elucidate the optimal preprocessing protocols and determine a method generic for the optimizing of enrichments in SBVS. Establishing a successful virtual screening process also requires that the techniques used must be CPU and resource "friendly", while keeping database physical size to a minimum. This study is a snapshot of some of the most popular available techniques used in database preprocessing. All of the programs used were utilized with their default settings, and the adjustment of certain parameters may further enhance the enrichment rates achieved. The results presented here have important implications for those embarking on structure-based virtual screening experimentation. With increasing amounts of commercial and academic code available to researchers, the choice of preprocessing technique used to expand and represent a screening compound collection can and will have a significant impact on the performance of the virtual screen. Compound libraries are often represented in 2D format as SMILES strings and are converted to 3D format for the purpose of pharmacophore searching or structure-based virtual screening. SMILES strings can be constructed in a number of ways, and we have demonstrated clearly in this study that different representations have markedly different effects, not only on virtual screening enrichment rates achieved but also on the quality of the docked structures. If speed is a concern, then different tools are available with associated performance benefits, but not all methods will convert all input. A clear pattern is observed when using Daylight's MOL2SMI SMILES strings, where optimum enrichment rates are achieved over those found using CONVERT string representations when compounds are enumerated as single conformations. Protonation, tautomerization, and the assignment of correct stereochemistry appears to have little benefit in this test case, but with respect to single conformers, an optimal enrichment can be achieved starting from a MOL2SMI string. SMILES strings generated from CONVERT require enumeration of multiple conformations to produce a high enrichment regardless of whether they are protonated, tautomerized, or the stereochemistry is assigned.

Importantly, enrichment rates observed when using a smaller data set of 1000 compounds seeded with 40 actives show a marked difference to the ranking of a single active in 10 000. The best enrichment is achieved using OMEGA in combination with the propagation of 10 conformers per compound. No additional benefit is observed when using 100 conformers per compound with this particular receptor (ER), but this may not be the case when working with more flexible target systems. To allow a sufficient amount of synthetic information to be retrieved about each compound, a slight decrease in the ranked position of the active must be accepted where stereochemical information is added. However, a balance between the introduction of false positive results and the exact structural information needs to be achieved, and a cost/benefit assessment must be made for each target studied. A marked difference in the ranking ability of SBVS imbued by alternate preprocessing protocols is observed when using the larger test set. We therefore suggest the adoption of larger ("real-scale") validation and training data sets as being more beneficial to those involved in training a docking procedure for the identification of active species in virtual screening. We are currently applying these findings to other virtual screening protocols optimized for the identification of novel modulators of the human estrogen receptors.

ACKNOWLEDGMENT

We acknowledge the funding support from the Irish Health Research Board (HRB), the Higher Education Authority's Program for Research in Third Level (PRTLI) Institute for Information Technology and Advanced Computation (IITAC), and Science Foundation Ireland (SFI). We thank the software vendors for their continuing support of such academic research efforts, in particular, the contributions of Daylight Corporation and OpenEye Scientific, which significantly underpinned the work performed in this study.

REFERENCES AND NOTES

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *22* (2), 151−85.

(2) Smith, A. Screening for drug discovery: The leading question. *Nature* **2002**, *418*, 453−459.

(3) Lahana, R. How many leads from HTS? *Drug Discovery Today* **1999**, *4* (10), 447−448.

(4) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, *7* (Suppl.), 957−9.

(5) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813−8.

(6) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862−5.

(7) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7* (20), 1047−55.

(8) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43* (25), 4759−67.

(9) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44* (7), 1035−42.

(10) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213−21.

(11) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46* (14), 3045−59.

(12) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43* (3), 401−8.
(13) Lloyd, D. G.; Hughes, R. B.; Zisterer, D. M.; Williams, D. C.; Fattorusso, C.; Catalanotti, B.; Campiani, G.; Meegan, M. J. Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J. Med. Chem.* **2004**, *47* (23), 5612−5.
(14) Lloyd, D. G.; Buenemann, C. L.; Todorov, N. P.; Manallack, D. T.; Dean, P. M. Scaffold hopping in de novo design. Ligand generation in the absence of receptor information. *J. Med. Chem.* **2004**, *47* (3), 493−6.
(15) Meegan, M. J.; Lloyd, D. G. Advances in the science of estrogen receptor modulation. *Curr. Med. Chem.* **2003**, *10* (3), 181−210.
(16) Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M. Ethyl side-chain modifications in novel flexible antiestrogens−design, synthesis and biological efficacy in assay against the MCF-7 breast tumor cell line. *Anti-Cancer Drug Des.* **2001**, *16* (1), 57−69.
(17) Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M. Flexible estrogen receptor modulators: design, synthesis, and antagonistic effects in human MCF-7 breast cancer cells. *J. Med. Chem.* **2001**, *44* (7), 1072−84.
(18) MacGregor, J. I.; Jordan, V. C. Basic guide to the mechanisms of antiestrogen action. *Pharmacol. Rev.* **1998**, *50* (2), 151−96.
(19) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389* (6652), 753−8.
(20) Scott, J.; McGuire, W. L. New Molecular Markers of Prognosis in Breast Cancer. In *Endocrine-Dependent Tumors*; Voigt, K. D., Knabe, C., Eds.; Raven Press: New York, 1991; pp 179−196.
(21) Lloyd, D.; Smith, H.; O'Sullivan, T. P.; Zisterer, D. M.; Meegan, M. *J. Med. Chem.* **2005**, in press.
(22) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95* (7), 927−37.
(23) Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1* (3), 220−7.
(24) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47* (10), 2499−510.
(25) Bostrom, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137−52.
(26) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470−89.
(27) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57* (2), 225−42.
(28) *FRED*, version 2.0.1; Openeye Scientific Software. http://www.eyesopen.com.
(29) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9* (1), 47−57.
(30) *MACROMODEL*, version 6.5; Schrodinger Inc. http://www.schrodinger.com.
(31) *Molecular Operating Environment*; Chemical Computing Group. http://www.chemcomp.com.
(32) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein−ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793−806.
(33) Derwent World Drug Index. http://thomsonderwent.com/poducts/lr/wdi.
(34) 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.* **2001**, *46* (1−3), 3−26.
(35) Daylight Chemical Information Systems Inc. http://www.daylight.com.
(36) *FILTER*; Openeye Scientific Software. http://www.eyesopen.com.
(37) Strausberg, R. L.; Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **2003**, *300* (5617), 294−5.
(38) *CONVERT*; Molecular Networks GmbH. http://www.mol-net.de.
(39) *TAUTOMER*; Molecular Networks GmbH. http://www.md-net.de.
(40) *UNITY*; Tripos Inc. http://www.tripos.com.
(41) *TAUTOMER*; Openeye Scientific Software. http://www.eyesopen.com.
(42) *QUACPAC*, version 1.1; Openeye Scientific Software. http://www.eyesopen.com.
(43) *STERGEN*; Molecular Networks GmbH. http://www.mol-net.de.
(44) *FLIPPER*; Openeye Scientific Software. http://www.eyesopen.com.
(45) *CORINA*, version 3.6; Molecular Networks GmbH. http://www.mol-net.de.
(46) *OMEGA*, version 1.8.1; Openeye Scientific Software. http://www.eyesopen.com.
(47) *RUBICON*; Daylight Chemical Information Systems Inc. http://www.daylight.com.
(48) *CATALYST*; Accelrys: San Diego, CA. http://www.accelrys.com.
(49) Sadowski, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000.
(50) *SMARTS*; Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
(51) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20* (4), 281−95.
(52) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14* (8), 731−51.
(53) Good, A. C.; Cheney, D. L. Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *J. Mol. Graph. Model.* **2003**, *22* (1), 23−30.
(54) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **2003**, *21* (5), 449−62.

CI050185Z

J Comput Aided Mol Des (2006) 20:179–190
DOI 10.1007/s10822-006-9044-4

Permuting input for more effective sampling of 3D conformer space

Giorgio Carta · Valeria Onnis · Andrew J. S. Knox · Darren Fayne · David G. Lloyd

Received: 21 December 2005 / Accepted: 18 March 2006 / Published online: 14 July 2006
© Springer Science+Business Media B.V. 2006

Abstract SMILES strings and other classic 2D structural formats offer a convenient way to represent molecules as a simplistic connection table, with the inherent advantages of ease of handling and storage. In the context of virtual screening, chemical databases to be screened are often initially represented by canonicalised SMILES strings that can be filtered and preprocessed in a number of ways, resulting in molecules that occupy similar regions of chemical space to active compounds of a therapeutic target. A wide variety of software exists to convert molecules into SMILES format, namely, Mol2smi (Daylight Inc.), MOE (Chemical Computing Group) and Babel (Openeye Scientific Software). Depending on the algorithm employed, the atoms of a SMILES string defining a molecule can be ordered differently. Upon conversion to 3D coordinates they result in the production of ostensibly the same molecule.In this work we show how different permutations of a SMILES string can affect conformer generation, affecting reliability and repeatability of the results. Furthermore, we propose a novel procedure for the generation of conformers, taking advantage of the permutation of the input strings—both SMILES and other 2D formats, leading to more effective sampling of conformation space in output, and also implementing fingerprint and principal component analyses step to post process and visualise the results.

Keywords Conformers · Docking · Drug discovery · Fingerprints · Scoring · SMILES · Virtual screening

G. Carta · V. Onnis · D. Fayne · D. G. Lloyd (✉)
Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland
e-mail: lloyddg@tcd.ie

A. J. S. Knox
School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland

Introduction

The advent of drug design, concurrent with the accelerated evolution of computational power, allows us to manipulate and store large amounts of data in different formats, for pre-processing, substructure searching, 3D-matching and finally for creation of a combinatorial library [1–3]. A widely used method to handle compound structures is to represent them in SMILES format. Simplified Molecular Input Line Entry Specification (SMILES), proposed by Weininger in 1988 [4] is a simple linear chemical language for specifying molecules or molecular fragments. Compared with most file formats, SMILES is broadly used in computational chemistry and is supported by most major software tools in the field, reducing data exchange due to its compact nature. SMILES strings have been extensively used in both pharmaceutical and academic settings to uniquely code a 2D structure of a molecule with correct stereochemical and protonated states represented.

However, a molecule although uniquely represented, can be described by differently ordered SMILES strings. Rules can be applied that specify the generation of SMILES but, intriguingly, we show that different software packages produce different SMILES

strings for the same encoded molecule. If the SMILES strings are used for representing the 2D structure of the molecules, possible permutations of the string should not affect the generation of the 3D structure.

However, when the SMILES strings are used as input for the generation of conformers, it becomes clear that different permutations of the same string can generate diverse sets of conformers. This has a large impact for all applications and validations where a dataset was initially represented by SMILES strings. Our group has previously shown the impact that different SMILES strings representations have on the docking process in terms of enrichment (E) and false positive (FP) rate calculations [5].

To examine this concept we carried out a study on the permutation of SMILES strings using data from a set of 15 crystal structures extracted from a recent study by Vigers et al. [6] and also 2 X-ray structures of interest within our group [7,8] (Table 1). Our aim was to quantitively show the influence that each permutation of a SMILES string, generated by several software programs, has on the production of conformers. We also show how this phenomenon can be used, in some cases, to increase the rate of production of a diverse set of conformers.

Figure 1 illustrates the process whereby a single PPAR agonist 3D structure was reduced to SMILES format using Mol2smi (Daylight Chemical Information Systems, Inc.) [9], MOE (Chemical Computing Group) [10], Babel (Openeye Scientific Software) [11] and ChemSketch (Advanced Chemistry Development) [12]. The SMILES string produced by each software package, as shown in Fig. 1, are different for the same molecule.

As SMILES notation represents a linear form of a connection table, we also wanted to investigate if the concept was extendible to more elaborate connection tables, as those found in formats such as sdf.

Permuted SMILES strings were generated for ligands extracted from 17 crystal structures in the Protein Data Bank (PDB) and then used as input for different conformer generating software packages (Omega v1.8.1, Corina v3.6, Rubicon and Catalyst v4.9.1).

Principal component analysis (PCA) of several 3D descriptors and potential energy was employed to investigate and compare the conformational space explored by the same output quantity of conformers arising from a permuted set of a SMILES, in comparison to conformational space sampled when using a single unique canonical SMILES string.

Methods

Seventeen crystal structures from the PDB were selected. Most of the structures were selected from the list used to validate FlexX [13]/Gold [14] and also MASC [6]. All structures were solved at a resolution ≤ 2.5 Å, except for 1P93 (2.7 Å) the glucocorticoid complex) and they represent a wide variety of different structural characteristics such as size, number of

Table 1 List of ligands extracted from the PDB and their associated resolutions

	PDB code	Ligand–protein complexes	Resolution (Å)	Perm[a]	P. Sel.[b]
1	1AQ1	CDK	2.2	35	6
2	1D1Q	Tyrosine phosphatase	1.7	14	3
3	1GLQ	Glutathione-*S*-transferase	1.8	30	6
4	1ABE	Arabinose-binding protein	1.7	10	2
5	1AZM	Carboic anidrase	2.0	14	3
6	1CBX	Carboxypeptidase	2.0	15	3
7	4DFR	DHFR	1.7	33	7
8	1EBP	Retinoic acid binding protein	2.2	22	4
9	1HYT	Thermolysin	1.7	15	3
10	1MRK	Ribosome inactivating protein	1.6	19	4
11	1P93	Glucorticoid	2.7	28	5
12	1PHF	Cytochrome p450-cam	1.6	11	3
13	4PHV	HIV-protease	2.1	46	9
14	1I7I	PPAR gamma	2.35	28	6
15	1POC	Phospholipase A2	2.0	31	6
16	1SRJ	Streptavidin	1.8	22	4
17	1TPP	Trypsin	1.4	15	3

[a]Specifies the number of permutations generated by 'Permsmi' (Daylight Inc.)

[b]Indicates the number of strings selected to carry out the analysis, including the canonical SMILES string, as generated by Mol2smi (daylight)

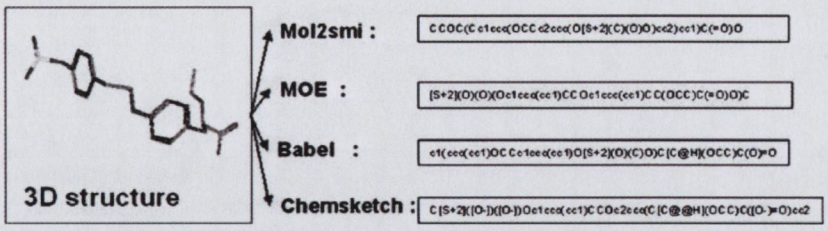


Fig. 1 Classification of conversion programs applied and their associated SMILES strings generated

rotatable bonds and features. The list of compounds studied is shown in Table 1.

Preparation of ligands

Compounds were downloaded and visually inspected within MOE [10]. Ligands were extracted from the complex and all hydrogen and coordination atoms deleted. Mol2smi was utilised to reduce the 3D representation to an initial canonical SMILES version.

Permsmi [9] was subsequently used to produce permutations of the SMILES string, starting with each of the atoms in the input SMILES in turn. The set of permutations of the molecule are written out in pseudo-canonical order (i.e. ring closures are renumbered). Thus, the longer the string, the higher the numbers of permutations 'Permsmi' is able to generate. An average of 22.8 permuted SMILES strings were created for each molecule. We chose to use every fifth permuted SMILES string of a molecule as we deemed it to be representative of the set of five making sure to include the canonical SMILES strings initially produced by mol2smi. This procedure narrows down to an average of 5.5 permutation strings per molecule, which were used as input for generating 3D conformers as illustrated in Fig. 2.

Conformer generation

Corina

Corina v3.6 [15] produces 3D conformations by combining monocentric fragments of standard bond lengths, appropriate angles and dihedral angles. To incorporate appropriate torsion angles, Corina uses a set of rules derived from statistical analysis of the Cambridge Structural Database (CSD) [16]. Furthermore, Corina is able to generate only a single conformer per SMILES string.

Corina, retaining default settings, was used to generate conformers of the permuted list and the output was saved in mol2 format. One conformer per SMILES string was created for a total of 94 conformers.

Omega

Omega v1.8.1 [17] generates conformers in three separate steps. First, a generic 3D structure is built. This primary conformation can be minimized with the MMFF force field. In the second stage, in order to generate different 3D conformations, Omega applies a depth-first divide and conquer algorithm. More simply, Omega disassembles the molecules into small

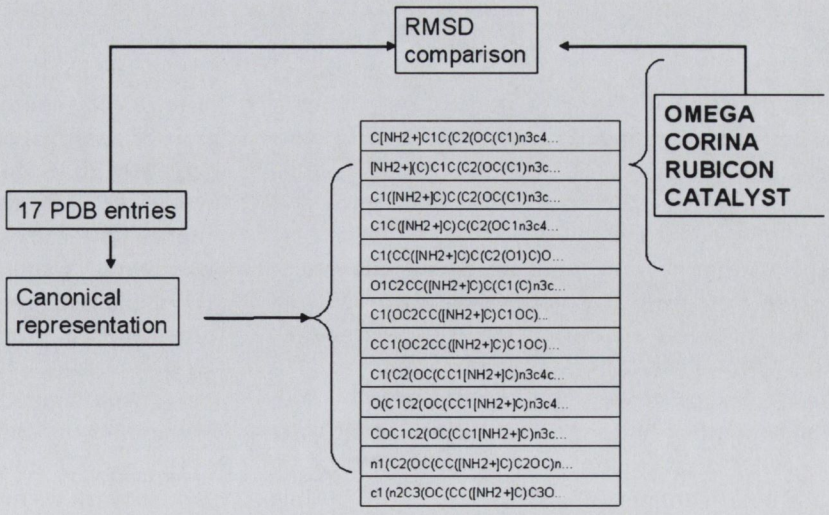


Fig. 2 Protocol for the evaluation of SMILES string permutations

fragments, and calculates the fragments energy using a Deidral FF, and finally reassembles the fragments in order. The generated conformations are limited by their total energy (e-windows) (Deidral force field, no electrostatic interaction) and molecules whose RMSD values are too similar (RMSD value cut off) are discarded. The third and final step offers the possibility to refine conformations with standard MMFF to avoid the presence of high energy unrealistic conformations.

Omega was set to generate 50 conformers per SMILES string. The output file was set to save as mol2 format and the number of rotatable bonds increased to 30, to allow all molecules to pass the Omega filter. All other remaining parameters were kept as default. 1 azm was discarded by Omega filter due to the presence of a zinc atom. A total of 2,406 conformers were generated.

Catalyst (fast)

Catalyst [18] uses a poling algorithm for the generation of conformers with the CHARMM force field, which does not include electrostatic terms. There are two integrated algorithmic methods of conformation creation to choose from: the BEST and FAST algorithm, of which FAST was the method of choice for this study. Its conformational search algorithm is based on a grid search of torsion values in combination with some energy minimisation to resolve high-energy configurations, such as overlaps of van der Waals radii.

A maximum number of 50 conformers per each SMILES string were created with the output saved in sdf format. Again, 1 azm was discarded, and a total of 3,623 conformers were generated.

Rubicon

Employing Rubicon [19], conformers are initially generated with a distance-geometry method that samples random conformations—this is by default a non-stochastic process, yielding reproducible output for a given input string in separate runs. These 3D structures are then minimized for distance and volume bounds violations. Only one minimizer algorithm is provided with version 4.3 of Rubicon [20].

SMILES strings were first converted into tdt format by smi2tdt [9] since Rubicon does not read the SMILES format and the output set to pdb format. A total number of 3,845 conformers were generated.

The default settings were used, except where noted, when implementing the conformer generation software packages. All conformations were superimposed by matching non-hydrogen atoms, to the corresponding unmodified X-ray ligand structure. To remove artificial differences resulting from RMSD deviations attributed to automorphisms [21,22], the program RMSD (Openeye Scientific Software) was used for analysis. The minimum RMSD values (the closest conformers to the crystal structure) and the standard deviations were then calculated for each set.

As it is clear that sampling conformer space with an ensemble size of 50 output molecules might be not sufficient to allow the software to produce a reasonable and diverse ensemble of conformers, two further experiments were performed. Thus, a subsequent RMSD analysis was carried out allowing the generation of a larger number of conformers (sets A and B described below).

Set A: From the original set of crystal structures we selected a molecular "probe" characterized by an average number drug like rotable bonds (1DFR, 10 RBs)—representative of the typical drug-like compounds encountered in commercial screening datasets used in virtual screening. The canonical string (cansmi) and subsequently eight permutations (Permsi) were generated as described above. In this study case the conformers were built using OMEGA. The 1DFR canonical SMILES string was first used and highest number of conformers with the default setting was obtained (a total 280 conformers with *max_confs* flag set to 500). To produce a comparable set of conformers the same number of output structures were replicated for the permuted string setting the *max_conf* flag to 35 (35 confs by 8 permutations). The results were then imported in MOE and 3D potential energy descriptor (MMFF94), PCA analysis were then calculated and plotted.

Set B: The same canonic SMILES string and its eight permutations were then used for a subsequently run. OMEGA was used to generate conformers from the canonic string with a *RMSD (rms* flag) cut off of 0.5 (default = 1) and an increased *energy window* (*ewindow* flag) to reflect the energy window span demonstrated by the conformers previously generated in A in terms of potential energy. (about 60 kcal/mol calculate with MMFF94 force field). *Max_confs* flag was increased to 1,000,000 to remove any restriction in the number of output structures. This produced an upper ensemble of **X** conformers from the single input string. As in experiment *A* a similar final number of conformers were produced from the permuted strings through setting *max_confs* to 310 and leaving the *energy window* and the *RMSD* cut off at the default settings. Again, the sets generated were imported in MOE and 3D potential energy descriptor and PCA analysis were carried out and plotted in graphs.

Fingerprint analysis

Conformers generated from the ensemble of permutations of a single molecule were then considered. For example, a set was represented as a collection of conformers produced from the six permutations of the PPAR ligand SMILES string. To measure the number of duplicate conformations presents in the set we imported the conformational file into MOE and an eigenvalue spectrum shape fingerprint ESshape3D 10 analysis was undertaken. Each fingerprint is of fixed length and allows for a comparison of 3D molecular shapes considering the heavy atoms of a molecule. Duplicate conformers appeared in the database with identical fingerprints, allowing us to calculate the percentage of unique conformers.

PCA analysis

In order to investigate the chemical space explored by the set generated for the set A and B previously described was analysed with a principal component analysis. A total of six MOE 3D descriptors (all x3D descriptors, *x–y–z* dipole moment, and principal moment of inertia *x–y–z*) were selected and computed for all the conformers generated to reflect the different topological characteristics of the 3D molecules in the output datasets. The aim of PCA is to reduce the dimensionality of a set of descriptors by linearly transforming the data. In detail, in our analysis six vectors have been transformed into a manageable 3D space described by three principal component vectors, where each of the 3D vectors is a combination of the selected descriptors.

To achieve this result the number of principal components were limited to three and all the other options in the MOE PCA module were run as default settings, without applying specific weighting to any of the 30 descriptors selected. The three principal components were then plotted as *XYZ* coordinates respectively for each compound. At the end of the process each molecule in the database is described by a set of coordinates, and may be plotted relative to all other members in the three-dimensional space.

Alternative 2D representations—connection tables

The connection table is a common and predominant way to simply represent a chemical structure—in its simplest incarnation we encounter SMILES strings. A typical file extension that contains a more complex connection table describing the organization of the atoms within a molecule is the SDF file format. Typically an SDF file includes two essential main sections: the first describing the 3D coordinates of the individual atoms in the molecules, the second depicting the connections of each atom with the others and the bond order present (1 = single bond, 2 = double bond) [23].

Section 1 (3D coordinates)

54	57	0	0	0	0	0	0	0	0	0						
	−0.3250			0.1625			0.0000			C	0	0	0	0	0	
	2.6042			1.8667			0.0000			C	0	0	0	0	0	
	−0.3250			−0.9917			0.0000			C	0	0	0	0	0	
	0.6458			0.7083			0.0000			C	0	0	0	0	0	
	2.6042			0.7083			0.0000			C	0	0	0	0	0	
	1.6250			0.1625			0.0000			C	0	0	0	0	0	
	....															
	....															

Section 2 (Connection table)

3	1	1	0	0	0
4	1	1	0	0	0
12	1	1	0	0	0
1	22	1	0	0	0
2	9	1	0	0	0
8	2	1	0	0	0
2	23	1	0	0	0

For this part of our study a bespoke perl utility was written to produce random permutation of the connection table. In detail, the utility randomises the position of the first section of the file and, while keeping trace of movement, it rewrites a new different connection table underneath. To avoid the production of duplicates an adjacency matrix is first calculated from the connection table and subsequently its connectivity stack [24].

The connectivity stack of a matrix is a sequence of the adjacency matrix values $s_k = a_{ij}$:

$$S = (s_1, s_2, s_3 \ldots s_k \ldots s_n)$$

where the position of the kth element is given by

$$k = \frac{i + (j\ 1)(j-2)}{2} \quad \text{with } (i<j)$$

Identical stack matrixes correspond to identical adjacency matrix and, thus, identical connection tables.

To investigate if permutation of connection tables in more complex formats such as SDF resulted with the same effect observed for the permuted SMILES strings on conformer generation, we first generated an SDF file from the ligand co-crystallised in PDB entry 1DFR [25]. OMEGA was then used to generate 280 conformers (*max_confs* set to 500 and remaining parameters as default). As described above, the perl script was used to produce 10 permutations of the crystal structure sdf file.

Finally, to reproduce the same amount of conformers as above, *max_conf* was set to 10 to furnish 280 conformers from the list of permuted connection tables of 1DFR. RMSD calculation was then carried out between the resulting conformers and the co-crystal ligand in 1DFR using OpenEye's RMSD [26].

Results and discussion

Two categories of data are reported in the results: a quantitative result taking into account the number of conformers generated by each software package and a qualitative result, where we considered the characteristics of the conformers generated. From a quantitative point of view, some software tools show interdependence between string permutations and the numbers of conformers generated. In others words, a different number of conformers were generated when starting from a different permutation of the same SMILES string.

In particular, Omega and Rubicon exhibited this behaviour, as although the maximum numbers of conformers to be generated was set to 50, this number was not always constructed. For example, Rubicon generated 27 conformers from the initial canonical SMILES string and up to 32 for one of its permutations for the ligand extracted from 1DFR. Omega produced only 9 conformers for the canonical string and 14 for a permutation of it in the case of the X-ray ligand taken from 1MRK. Less affected was Catalyst, which in the worst case (1CBX) generated 39 conformers for the canonical and 36 for most other permutations. The full set of results is shown in Tables 2–8. Results in Tables 2, 3, 4 report the RMS standard deviation values calculated for canonical SMILES strings and their permutations. Cells labelled with 'D' were molecules discarded by the software in the cases of both Omega and Rubicon. Label 'A' indicates that standard deviation was not computable because just one conformer per string was produced. Empty cells indicate that no permutation was produced for that string.

It is clear from Table 2 that each set of the permuted SMILES string generated molecules exhibits a different SD of the RMSD values for each set of conformers compared to the X-ray structure. That suggests that each set includes a different pool of RMSD values when compared with their respective crystal structures. Similar results generated by Omega are also observed in the case of Rubicon, with different RMSD values also being produced (Table 3). Table 4, however, shows that Catalyst is not as affected as Omega or Rubicon, in the conformers being generated from the permutations of the canonical SMILES string. Approximately the same SD values appear for the set indicating that identical conformers were created by each permutation. In the case of Corina, no SD calculations could be carried out as only one conformer per SMILES string could be produced.

We next sought to assess the quality of the conformers produced by each software package with respect to the different permutations of a SMILES string. Tables 5, 6, 7, 8 show the minimum RMSD values retrieved from each sets. Label D and empty spaces have the same significance as in the previous tables.

Table 2 depicts that minimum RMSD values (the closest conformers to the crystal structure) are derived from a permuted string rather than the canonical one which would be normally used in a virtual screening process. Importantly, in some cases, the lowest RMSD value is nearly half that observed with the canonical SMILES string (eg.1POC, 4DFR, 1ABE). As before, we see that the same holds for Rubicon as in the case of Omega, where the minimum values retrieved in the permuted sets are lower than those of the canonical string (Table 6). A less

Table 2 RMSD standard deviation for Omega

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	A	0.21	0.407	0.28	D	0.353	0.291	0.369	0.36	0.426	0.12	0	0.717	0.434	0.417	0.417	0.516
P1[b]	0.007	0.021	0.398	0.249	D	0.258	0.53	0.355	0.269	0.394	0.184	0.304	0.741	0.666	0.342	0.698	0.516
P2	A	0.2	0.456	0.224	D	0.344	0.316	0.308	0.135	0.445	0.163	0	0.568	0.888	0.324	0.85	0.6
P3	A	0.177	0.461		D	0.338	0.365	0.349	0.337	0.316	0.099	0.304	0.602	0.88	0.345	0.253	0.0554
P4	A		0.477				0.461	0.332		0.443	0.163		0.612	0.767	0.476	0.864	
P5	A		0.476				0.346				0.177		0.496	0.649	0.425		
P6	A		0.487				0.437						0.532	0.426	0.433		
P7							0.504						0.695				
P8													0.495				
P9													0.496				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 3 Minimum RMSD for Rubicon

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.45	0.31	1.53	0.38	D	0.49	1.67	0.57	0.24	0.91	1.17	0.1	2.02	2.46	1.26	0.17	0.39
P1[b]	0.38	0.31	1.28	0.11	D	0.61	1.73	0.89	0.23	0.94	0.99	0.09	1.72	2.12	1.46	0.19	0.46
P2	0.45	0.31	1.49	0.47	D	0.46	2.07	0.45	0.24	0.8	0.66	0.11	2.09	2.54	1.76	0.18	0.39
P3	0.47	0.31	1.35		D	0.49	1.84	0.76	0.25	0.85	1.12	0.09	2.24	2.57	1.3	0.18	0.47
P4	0.35		1.29				2.37	0.61		0.81	1.12		2.53	2.50	1.25	0.17	
P5	0.37		1.42				1.93				1.12		2.02	2.31	1.53		
P6	0.43		1.32				2.22						2.29	2.43	1.68		
P7							2.05						2.2				
P8													2.29				
P9													2.02				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 4 RMSD standard deviation for Rubicon

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.447	0.067	0.348	0.217	D	0.303	0.422	0.455	0.369	0.337	0.288	0.149	0.642	0.385	0.264	0.834	0.300
P1[b]	0.337	0.047	0.296	0.215	D	0.267	0.357	0.493	0.327	0.333	0.330	0.182	0.607	0.482	0.249	0.754	0.252
P2	0.401	0.063	0.405	0.193	D	0.340	0.325	0.45	0.411	0.351	0.345	0.154	0.548	0.405	0.206	0.781	0.261
P3	0.424	0.046	0.294		D	0.367	0.35	0.519	0.350	0.307	0.321	0.181	0.533	0.348	0.290	0.853	0.253
P4	0.428		0.33				0.251	0.471		0.363	0.321		0.503	0.329	0.282	0.761	
P5	0.382		0.327				0.322				0.352		0.628	0.411	0.235		
P6	0.404		0.338				0.316						0.565	0.34	0.215		
P7							0.338						0.477				
P8													0.544				
P9													0.628				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 5 Minimum RMSD for Omega

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	1.82	0.3	1.52	0.61	D	0.42	1.7	0.53	0.68	0.8	1.3	0.24	2.41	3.41	1.06	0.39	0.3
P1[b]	1.26	0.52	1.73	0.68	D	0.68	0.76	0.74	0.45	0.95	1.3	0.24	1.72	2.1	1.18	0.39	0.3
P2	1.77	0.32	1.44	0.37	D	0.49	1.98	1.19	1	0.71	1.64	0.67	2.38	1.98	1.26	0.18	0.3
P3	0.86	0.29	1.72		D	0.68	1.96	0.61	0.45	1.06	1.26	0.24	2.28	1.97	1.04	0.46	0.3
P4	1.27		1.17				1.17	0.66		0.83	1.61		1.99	1.73	0.78	0.19	
P5	1.51		1.27				1.98				1.25		2.57	2.49	0.95		
P6	1.43		1.66				1.15						2.41	3.44	1.1		
P7							1.09						1.86				
P8													2.45				
P9													2.57				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

dramatic difference is observed when using Catalyst or Corina.

Table 9, A and B, show the RMSD values obtained comparing the crystal structure of 1DFR and different sets of conformers.

In detail, Table 8A shows that operating with default parameters and allowing OMEGA to produce as many conformers as it can, the output from the treatment of permuted input strings delivers conformers closer to the bioactive conformation when compared to those generated from the canonical string (280 final unique conformers were produced both form the canonical and the permuted string of 1DFR).

The graph in Fig. 3 shows a comparison between the potential energy of the conformers generated from the canonical and the permuted strings. It is quite clear

Table 6 RMSD standard deviation for Catalyst

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.445	0.179	0.596	0.285	0.346	0.413	0.519	0.476	0.382	0.419	0.209	A	0.670	0.806	0.377	0.76	0.278
P1[b]	0.487	0.177	0.586	0.264	0.346	0.413	0.519	0.476	0.382	0.470	0.209	A	0.598	0.806	0.377	0.76	0.278
P2	0.340	0.177	0.586	0.198	0.346	0.302	0.519	0.476	0.384	0.470	0.209	A	0.598	0.764	0.319	0.76	0.278
P3	0.483	0.177	0.626		0.346	0.413	0.519	0.476	0.382	0.439	0.209	A	0.598	0.764	0.319	0.76	0.278
P4	0.483		0.626				0.519	0.476		0.477	0.209		0.670	0.764	0.319	0.76	
P5	0.483		0.622				0.519				0.209		0.670	0.806	0.319		
P6	0.483		0.622				0.426						0.598	0.806	0.377		
P7							0.519						0.598				
P8													0.670				
P9													0.670				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 7 Minimum RMSD for Catalyst

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1sr	1tpp
Can[a]	0.53	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	1.91	2.02	1.27	0.17	0.43
P1[b]	0.26	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.02	1.27	0.17	0.43
P2	0.26	0.42	0.99	0.38	0.32	0.58	1.16	1.22	0.29	0.82	1.53	0.25	2.18	2.08	1.48	0.17	0.43
P3	0.26	0.42	1.35		0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.08	1.48	0.17	1.31
P4	0.26		1.35				1.16	1.22		0.8	1.53		1.91	2.08	1.48	0.17	
P5	0.26		1.35				1.16				1.53		1.91	2.02	1.48		
P6	0.26		1.35				1.3						2.18	2.02	1.27		
P7							1.16						2.18				
P8													1.91				
P9													1.91				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 8 Minimum RMSD for Corina

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1sr	1tpp
Can	1.14	0.76	2.38	0.63	1.31	0.56	2.57	1.35	0.91	1.09	1.47	0.46	3.78	1.82	4.15	0.15	0.42
P1	1.85	0.76	2.38	0.78	1.35	0.56	2.57	1.34	0.91	1.71	1.44	0.46	2.74	1.86	4.15	0.15	0.42
P2	1.82	0.76	2.41	0.73	1.35	1.52	2.58	1.46	1.49	1.71	1.46	0.46	2.75	2.24	4.15	0.15	0.42
P3	1.83	0.76	2.38		1.35	0.56	2.57	1.29	0.91	1.94	1.81	0.46	2.58	2.22	4.9	0.15	0.42
P4	1.83		2.38				1.98	1.02		1.96	1.22		2.05	2.24	4.9	0.15	
P5	1.82		2.41				1.97				0.56		2.59	2.19	4.15		
P6	1.81		2.41				1.75						2.75	1.82	4.15		
P7							1.97						2.48				
P8													2.05				
P9													2.05				

that the potential energy span produced by the conformers from the permuted input is wider then that originated from the canonical input.

To greater examine this phenomenon we undertook to represent the conformational space explored by the two output sets using PCA analysis. These data are shown in Fig. 4.

In red is depicted the conformational space location of the co-crystal ligand (the bioactive conformer) from 1DFR. Interesting, we note that the immediate conformational neighbourhood around the bioactive conformer predominately originates from the permuted string output. Moreover, the conformational space explored by the conformers from the canonic appears to be smaller and less diverse then that from permuted strings.

In Table 9 the unique conformers produced from the permuted strings are once again closer to the bio-

Table 9 RMSD minimum, maximum, average and standard deviation calculation for the ligand contained in PDB entry 1DFR: the top panel A shows the set where 280 conformers were produced; the panel below B shows the set where 2,239 and 2,326 conformers were produced for respectively the canonic (Can) and the permuted (Perm) SMILES strings

	Can	Perm
A		
Min	1.29	1
Max	3.44	3.44
Average	2.52921429	2.492357143
Dev st.	0.39585861	0.478173962
B		
Min	1.23	0.76
Max	3.47	3.44
Average	2.53093345	2.398963887
Dev st.	0.38276039	0.476240746

active conformer (RMSD 0.76 calculated for permuted strings against 1.23 from canonical), when comparing ensembles of similar size. From the canonical input, OMEGA produced 2,239 conformers. The set originating from the permuted input comprised 2,326 conformers.

As for set A, the potential energy has been calculated as previously described. In this case, the potential energy span originating from the permuted strings input is larger. Members of the ensemble are energetically much closer to the energy calculated for bioactive conformation then those members of the ensemble originating from the canonical input (Fig. 5).

Figure 6 shows the plotted result of the PCA. It depicts clearly how the conformational space explored by the conformers originating from permuted

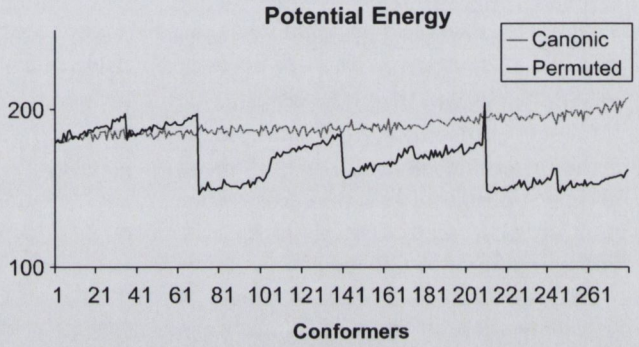


Fig. 3 Potential energy graph for conformer set A, generated using omega (280 members). Black line illustrates result for permuted input string, grey illustrates result for canonical

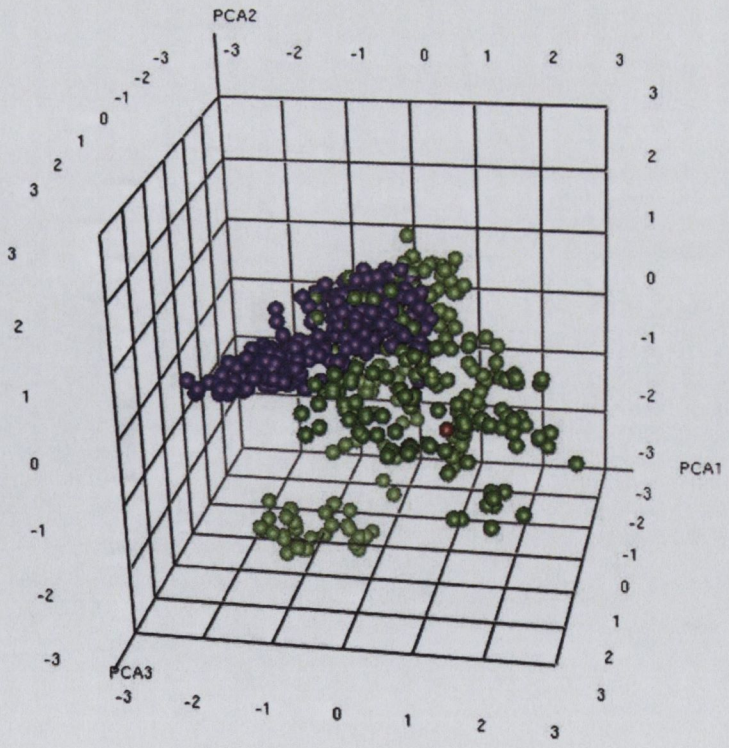


Fig. 4 PCA analysis of conformers generated by Omega: in violet are showed the 280 conformers produced form the canonical string of 1DFR. In green the conformers generated from 8 permutation of the same string. In red the crystal structure of 1DFR

input is diverse and closely surrounds the bioactive space.

Fingerprint analysis

A critical method for incorporating ligand flexibility in a vHTS (Virtual High Throughput Screening) process is to generate multiple conformers of each molecule. One of the key things to take into account at this stage in the process is the balance between the number of conformers generated and the diversity of each with regard to conformational space sampling. Generally, less conformers with a wider degree of RMSD separation between each is beneficial.

To investigate the uniqueness of the conformers generated from each permuted SMILES string and the software used, all the conformers were read into MOE and fingerprint analysis carried out using Esshape3D. By exploring the percentage of conformers that were identical in each set for a molecule it was possible to determine whether this was due to the orientation of each conformer. For example, it was possible that the ensemble of conformers derived from different SMILES permutations of 1AQ1, contained duplicates.

Figure 7 shows the result from the fingerprint analysis for Omega, Catalyst and Rubicon. The numbers of conformers produced by each software package were normalized to the one with the highest value. Thus, Rubicon (3,845) was 100%, Catalyst (3,623) 94.22% and Omega (2,406) 62.57%. The percentage of unique conformers generated from the permuted SMILES strings by Omega was 99.38% of the total. If the conformers were generated from only the canonical SMILES string, the percentage of the total would be just 11.33%. It is apparent at this stage that generating conformers using Omega with default settings is highly efficacious when combined with permuted SMILES strings. In the case of Rubicon, 87.73% of the total numbers of conformer were unique in conformational space.

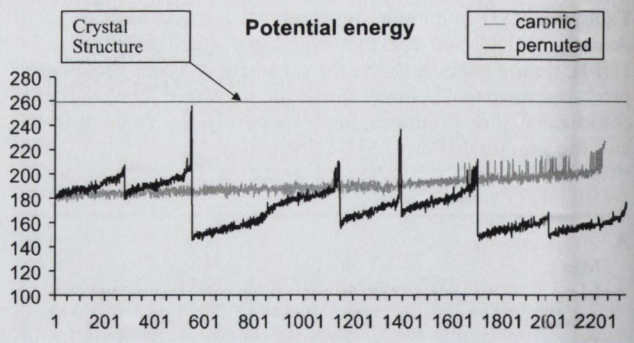


Fig. 5 Potential energy graph for conformer set B, generated using omega (ca. 2200 members). Black line illustrates result for permuted input string, grey illustrates result for canonical

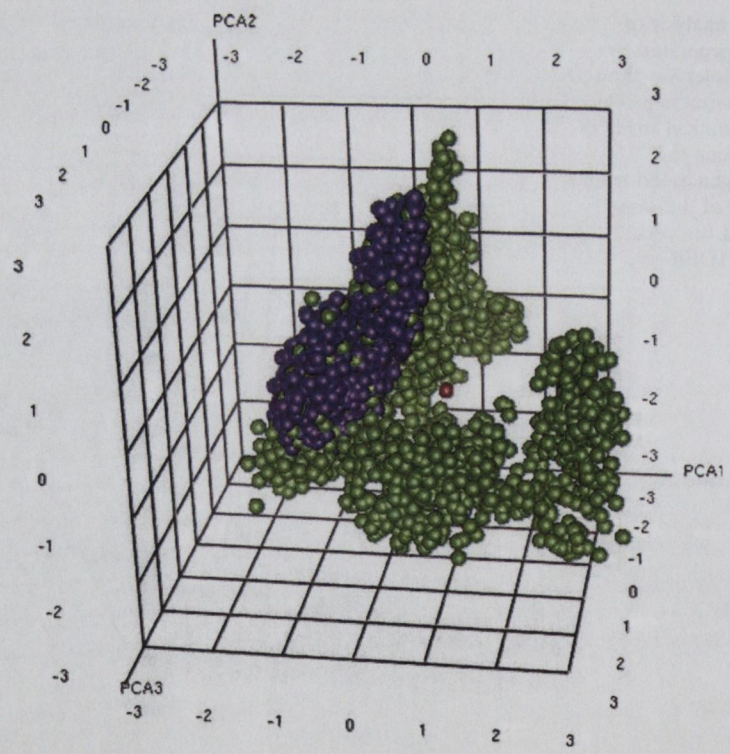


Fig. 6 Fingerprint analysis of conformers generated by Omega. In violet the conformers (about 1200) generated setting an energy window equal of 60 and a RMSD cut off of 0.5 from the single canonical string of the co-crystal ligand from 1DFR. In green the conformers generated from the 8 permutations of the 1DFR ligand keeping all the parameters in OMEGA as default (with the exception of –max_conf, that was set to 310 to generate a comparable number of final conformers with the canonic set. In red is showed the PCA coordinate calculated for the crystal structure

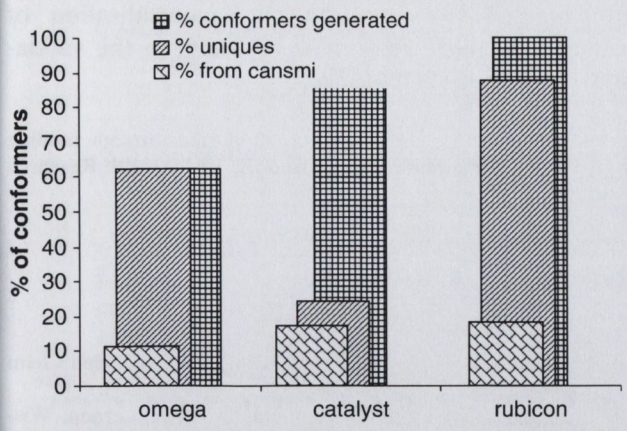


Fig. 7 Fingerprint analysis of conformers generated by Omega, Catalyst and Rubicon. The height of the 3 bar graphs is proportional to the total amount of conformers generated. Cansmi: Canonical SMILES string

Again starting from a single SMILES string, only 18.11% would be unique. To a lesser extent, but still evident, was Catalyst with the percentage of unique conformers being 23.98% of the total. Beginning from a single SMILES string would allow 17.06% unique conformers to be produced.

Since Corina produced just one conformer per string (94 total amount = 100%) it is represented on its own in Fig. 8 so as to keep it to scale. The percentage of unique conformations using Corina was 71.28%. With only the canonical SMILES string it would have been 18.18% of the total.

Connection table analysis

Table 10 depicts the RMSD difference between the co-crystal structure conformer and the conformers produced using permuted connection table input. As the values suggest, the same result obtained above for the permutation of input SMILES strings is reproduced for the permutation of more complex (SDF) connection table input. As explored for the SMILES strings, different conformers were generated for the input sets (different standard deviation). Moreover, beginning from 10 unique input permutations, the 280 output conformers generated are closer to the crystal structure (average: 1.78, min: 0.37) than the 280 conformers originating from a single connection table (average: 2.37, min: 1.14).

Conclusion

In this study we have shown that permutation of input strings leads to the generation of different

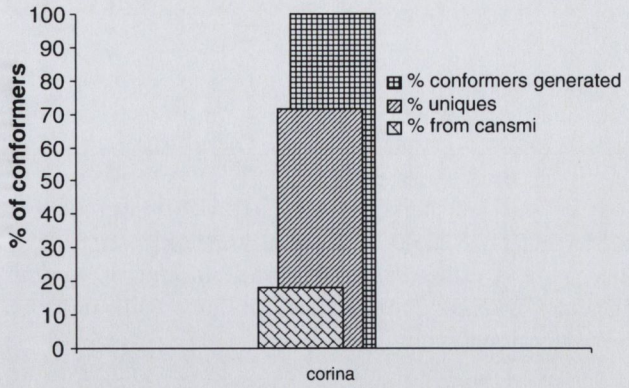


Fig. 8 Fingerprint analysis of conformers generated by Corina

Table 10 RMSD minimum, maximum, average and standard deviation calculation for 1DFR: the table shows the set where 280 conformers were produced using one connection table (one CT) and 10 permutation of a single connection table (Permuted)

1DFR	One CT	Permuted
Min	1.14	0.37736
Max	3.26	3.26
Average:	2.373614865	1.787744
Dev st.	0.377360288	1.280565

numbers of conformers and that some software programs are affected more than others by this depending on their innate algorithm. Depending on the software used the conformers generated from permuted SMILES strings are more diverse than if they were generated from a single representation. The differing amounts of conformations generated are not an inherent fault of the software packages used, rather it is an outcome of the input SMILES string and how the software deals with that SMILES string. When a SMILES string is permuted this possibly has an incidental effect on the order of the torsions which are subsequently used to generate the conformations.

Conformers produced from permuted SMILES strings exhibit a smaller RMSD value, when compared to the X-ray structure, than do conformers generated from the canonical SMILES string. Other work carried out by our group has shown the dramatic effect that this phenomenon has on Enrichment and False Positive rates. Thus, many validations of rigid docking/scoring combination platforms where the same dataset was used, but converted to SMILES strings using different software, may show better or worse performance as a direct result of this phenomenon and not due to the platforms utilised. We suggest that all validations of rigid docking/scoring platforms account for this effect prior to the docking step. Increasing the

number of conformers generated, may possibly reduce this effect.

Importantly, all of the software tested compares the generated conformers and discards those that are judged to be too closely related. The ultimate measure of this is an RMSD assessment. For example, one of the comparison steps in Omega involves applying a conformation RMSD exclusion threshold of 0.8 Å. If this value is reduced an increased amount of conformations will be accumulated but they will be more conformationally similar.

Conversely, when using permutation of the same SMILES string those conformers would not be discarded and thus some of the conformers obtained might have similar RMSD values.

Finally, we suggest a novel way of generating conformers when using Omega, Rubicon or Corina with their default settings. Permuted input strings, in conjunction with the canonical SMILES string, could be utilised to more rapidly and comprehensively sample conformational space of a molecule. This method extends to connection table representations of 2D molecular structure such as SDF. PCA analyses clearly illustrate superior conformational sampling arising from use of permuted input. To ensure that the final result does not include any duplicates a quick fingerprint analysis can be carried out as we have shown. We also propose that future versions of software employed in the generation of conformers from SMILES strings could include a switchable flag allowing the canonicalisation of the SMILES strings before conformer generation. The same flag, when off, would allow the users to generate multiple molecular conformations starting from permutations of the same SMILES string.

The effect is also observed using more complex molecular connection tables rather than SMILES strings. To this end, the same method described above may be used for the efficient generation of a diverse unique set of conformers. Clearly, SMILES notation presents a far more limited set of possible atom/bond orderings unless the use of dot disconnections or closure bonds are employed. Thus, exploiting SD or indeed any format incorporating a connection table may be used to generate (#atoms!)*(#bonds!) possible orderings, resulting in efficient sampling of conformational space. Finally, although not currently implemented in any software, canonicalisation of connection tables would also be useful in the validation process of vHTS tools.

Acknowledgement This work was supported through funding from Science Foundation Ireland and the Irish Health Research Board.

References

1. Hou T, Xu X (2004) Curr Pharm Des 10(9):1011
2. Liao C, Liu B, Shi L, Zhou J, Lu XP (2005) Eur J Med Chem 40(7):632
3. Bringmann BKA (2004) Frequent SMILES. Lernen, Wissensentdeckung und Adaptivität, Workshop GI Fachgruppe Maschinelles Lernen, part of LWA 2004
4. Weininger D (1988) J Chem Inf Comput 28:31
5. Knox AJS, Meegan MJ, Carta G, Lloyd DG (2005) J Chem Inf Model 45(6):1908–19
6. Vigers GP, Rizzi JP (2004) J Med Chem 47(1):80–89
7. Kauppi B, Jakob C, Farnegardh M, Yang J, Ahola H, Alarcon M, Calles K, Engstrom O, Harlan J, Muchmore S, Ramqvist AK, Thorell S, Ohman L, Greer J, Gustafsson JA, Carlstedt-Duke J, Carlquist M (2003) J Biol Chem 278(25):22748
8. Cronet P, Petersen JF, Folmer R, Blomberg N, Sjoblom K, Karlsson U, Lindstedt EL, Bamberg K (2001) Structure (Camb) 9(8):699
9. Daylight Chemical Informations Systems Inc. (URL: http://www.daylight.com)
10. Molecular Operating Environment (MOE), developed and distributed by Chemical Computing Group (http://www-chemcompcom)
11. Babel v2.0A3, distributed by Openeye Scientific Software
12. Chemsketch v8.17, www.acdlabs.com
13. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261(3):470
14. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267(3):727
15. CORINA 3.6, distributed by Molecular Networks GmbH
16. Cambridge Structural Database, http://www.ccdc.cam.ac.uk/
17. OMEGA 1.8.1, distributed by Openeye Scientific Software
18. Catalyst v4.9.1, www.accelrys.com
19. RUBICON, distributed by Daylight Chemical Informations Systems Inc
20. Shanno DF, Phua KH (1980) ACM Trans Math Software 6:618
21. Bostrom J, Greenwood JR, Gottfries J (2003) J Mol Graph Model 21(5):449
22. Bostrom J (2001) J Comput Aided Mol Des 15(12):1137
23. Ivanciuc O (2003) In: Gasteiger J (ed) Handbook of chemoinformatic , vol. 1. p 103
24. Kudo Y, Sasaki S (1974) J Chem Document 14(4):200
25. Babel 1.100.2, Distributed by Openeye Scientific Software
26. oechem, RMSD, Distributed by Openeye Scientific Software

Unbiasing Scoring Functions: A New Normalization and Rescoring Strategy

Giorgio Carta, Andrew J. S. Knox, and David G. Lloyd*

Molecular Design Group, School of Biochemistry & Immunology, Trinity College Dublin, Dublin 2, Ireland

Received October 31, 2006

Ligand bias can contribute significantly to the number of false positives observed in a virtual screening campaign. Using a receptor-based docking approach against a well-established target of therapeutic importance, estrogen receptor α (ERα), coupled with several common scoring functions (ChemGuass, ChemGauss2, ChemScore, ScreenScore, ShapeGauss, and PLP), taken both individually and as a consensus, we sought to examine the characteristics of molecules retrieved by each. It has been previously shown that scoring functions (mainly empirical) exhibit bias in prioritizing more complicated molecules arising from additive components within the function. Using Spearmen's correlation coefficient, we show that a large set of descriptors calculated for a docked set of molecules exhibit positive correlation with the ranked position in a hitlist. Moreover, most of these descriptors correlate well with MW. To this end, rather than penalizing the docked score of all heavy molecular weight (MW) molecules and rewarding those of lower MW, as is common practice, we examine the impact of penalizing the score only of those molecules which were of higher MW, leaving lower MW molecules unaffected. Here, we introduce a new power function to aid the process. Using scoring frequency analysis and SIFt fingerprints, we acheived a more meaningful analysis of virtual screening (VS) performance than with enrichment calculations, facilitating target-specific VS method development.

INTRODUCTION

Improvements in crystallography, genomics, and proteomics have permitted a deeper understanding of the interaction between proteins and their endogenous ligands. More specifically, in the realm of computational chemistry, development of algorithms and new mathematical approaches to describe the binding between receptor and ligand has allowed characterization of many important interactions such as hydrogen bonding, hydrophobic contacts, and van der Waals interactions. This knowledge is now commonly used in processes such as receptor-based virtual screening (RBVS), whereby the goal is to be able to simulate and accurately predict the docked position of a molecule inside a binding pocket allowing identification of "binders" from "nonbinders".

Subsequently, each docked compound receives a score equivalent to the free energy of binding, generated from the predicted binding pose. The general assumption is that molecules that bind tightly to the receptor should be given a higher score. Once scored, molecules can be ranked and sorted, and the top x% is generally retained for biological evaluation.

A large variety of methods have been described in literature to calculate and predict in silico the binding affinity between a receptor and ligand; however, it is also well-known that most fail to accurately predict it.[1] Currently, scoring functions can be classified in four main categories: force-field, semiempirical, empirical, and knowledge-based methods.[2] Empirical and knowledge-based methods are the most commonly used scoring functions in a VS campaign because of their inherent speed and ease of calculation.

* To whom correspondence should be addressed. Phone: +353-1-8962904. Fax: +353-1-6772400. E-mail: lloyddg@tcd.ie.

In this work, we focused our attention on the empirical method, such as that described originally by Bohm,[3] who dissected the overall Gibbs free energy of binding into singular contributions

$$\Delta G_{\text{bind}} = \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{h-bonds}} f(\Delta R,\Delta\alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic_int}} f(\Delta R,\Delta\alpha) + \Delta G|A_{\text{lipo}}| + \Delta G_{\text{rot}} N_{\text{rot}}$$

where ΔG_{bind} is a summation of hydrogen bonds interactions, ionic interactions, and hydrophobic interactions, minus an entropy penalty depending on the number of rotatable bonds. A wide variety of empirical scoring functions now exist as a result of Bohm's pioneering work, several of which we have used for the purpose of this study (ChemGauss,[4] ChemGauss2,[4] ChemScore,[5] ScreenScore,[6] ShapeGauss,[7] and PLP[8]) to rank a set of docked molecules generated from the FRED docking algorithm. Empirical scoring functions by nature are additive and generally exhibit a dependence on molecular weight resulting in a bias toward heavier molecules in the top of a ranked list of molecules postscoring.[9] It has also been widely shown in literature how different scoring functions produce markedly different results when applied to same target.[1,10−13] The combination of two or more of these individual scoring functions into a consensus score can be more effective in the retrieval of active molecules and ameliorate this effect.

By plotting the ranked position produced by each scoring function for each molecule versus their associated calculated descriptors, we employed Spearmen's correlation coefficient to evaluate their intrinsic ligand bias. The descriptors which exhibited a trend toward the higher ranking molecules were selected, and the correlation among them was explored. A normalization procedure was applied to each of the scores,

10.1021/ci600471m CCC: $37.00 © 2007 American Chemical Society
Published on Web 06/07/2007

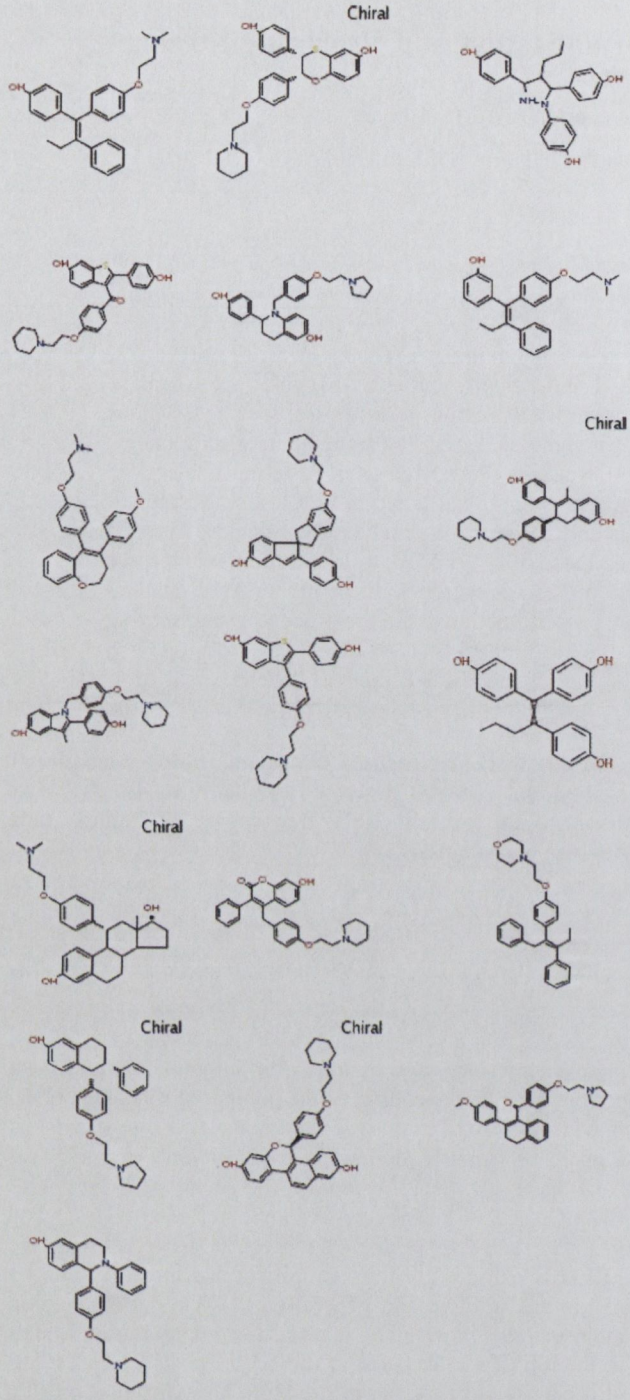


Figure 1. Structure of 19 Estrogen Receptor antagonists selected from literature.

depending on the strength of the trend exhibited by selected descriptors.

To minimize the occurrence of false positives at the top of a ranked hitlist caused by high MW, we introduce a power function whereby only high MW molecules are penalized according to their score. This serves to evenly distribute the probability of a molecule being retrieved because of disproportionate MW contribution and is shown to reduce false positives overall. The impact of this rescoring strategy within a consensus scoring scheme is also investigated.

Comparison of enrichment rates before and after the rescoring procedure was carried out to illustrate the benefits

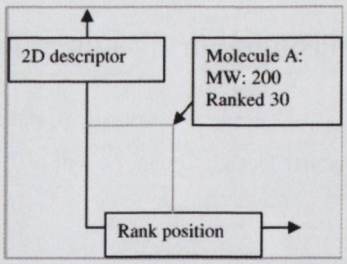


Figure 2. Descriptor (y-axis) plotted against ranked position (x-axis).

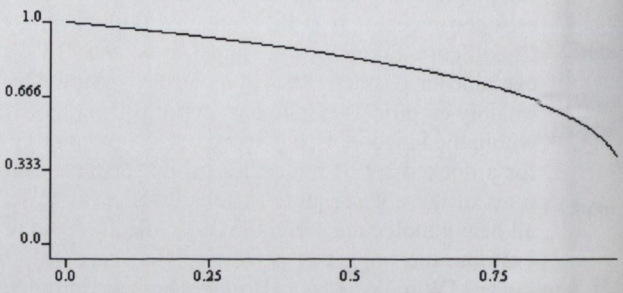


Figure 3. Plot of the function $y = (1 - x)^{0.25}$.

in reducing the number of false positives ranked highly because of increasing molecular weight. Enrichment is a commonly used measure of VS performance but it fails to depict the effect that each adjustment to a VS protocol has on the whole of the database, not just the actives In line with this, Warren et al.[2] have observed that enrichment alone is not sufficient for optimizing a VS protocol. Thus, we propose that scoring frequency, that is, the distribution of scores observed for a set of ranked ligands, is a better method for assessing the discriminatory power of these scoring functions.

Finally, to analyze the binding modes of the ligands used in the docking studies, structural interaction fingerprints (SIFt[14]) were employed as a means of determining whether ligands with better interactions are shifted to better ranks following application of our rescoring strategy.

COMPUTATIONAL METHODS

Multiconformers Database Preparation. The following protocol was fully automated through Pipeline Pilot (PP).[15] Using PP, we trained a Bayesian to extract compounds with similar characteristics to known active ligands according to the following: number of H-Bond donors/acceptors, number of rotatable bonds, A log P, molecular Weight, and FCFP_4 fingerprint.

The Bayesian algorithm was run against Maybridge, Bionet, and IBS2004 databases to select similar molecules to known ERα actives. All molecules were standardized (e.g., keep largest fragment, check valences) and ionized at pH 7.4.

Nineteen estrogen antagonist structures were retrieved from literature (Figure 1) and standardized as above. The active and decoy set were merged and duplicates were removed to ensure that the final database size was 10 000. Omega 1.81[16] was employed for the generation of conformers. Default settings were kept with the exception of maxconfs = 10, finalopt = true, and output = mol2 to retain charges.

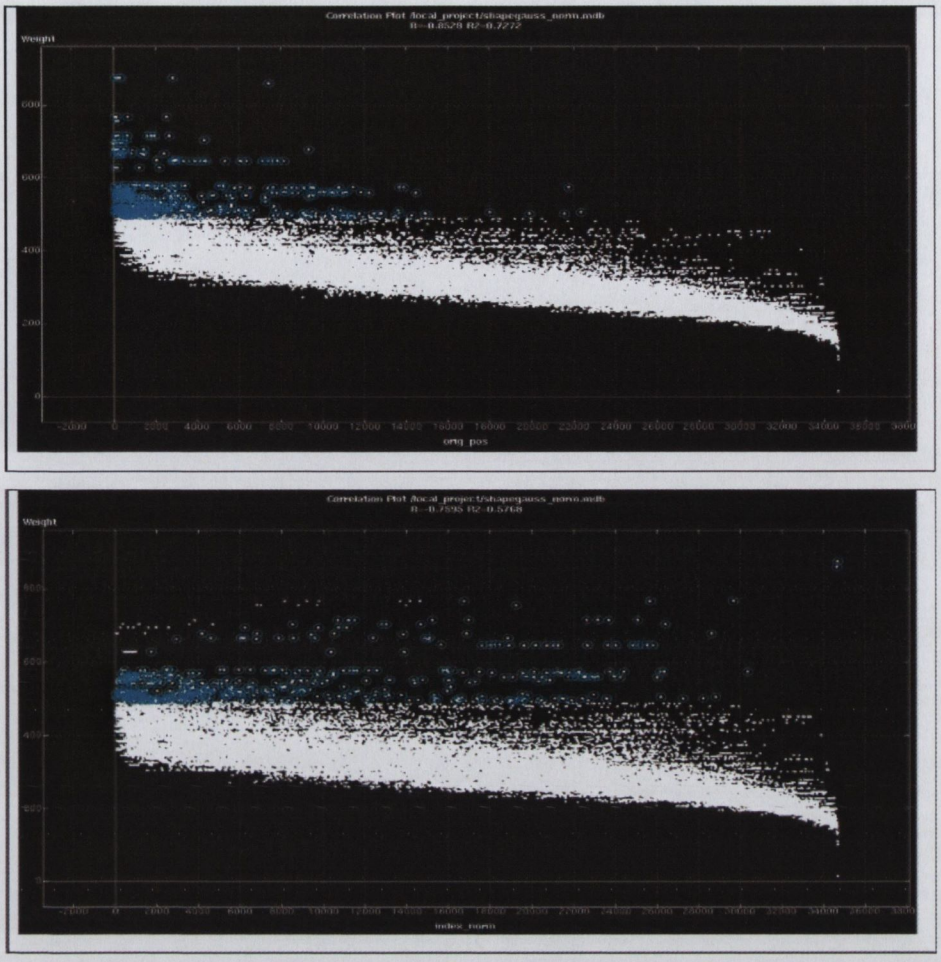


Figure 4. Effect of the rescoring procedure on the molecular-weight distribution plotted versus the rank position for the Shapegauss scored set. The *x*-axis depicts number of conformers of a 10 000 molecule set. Top: Before the normalization. Bottom: After normalization. Molecular weight above 500 is highlighted in both the plots.

Docking. The coordinates of estrogen α cocrystal structure 3ERT[17] with agonist 4-hydroxytamoxifen were downloaded from Protein Data Bank (PDB).[18] Water molecules were removed. MOE.2005.06[19] was used to add hydrogens to the protein, and a minimization protocol using MMFF94 forcefield was implemented to adjust the positions of hydrogen keeping the heavy atoms fixed at their coordinates. The ligand was extracted and used to define the 5 Å search box for docking. FRED2.11[4] was then used to generate poses inside the binding pocket using the standard default settings.

Scoring. Scoring of the docked poses for each ligand was achieved using the scoring functions implemented in FRED2.1.1 (ChemGauss2, ChemGauss, ChemScore, PLP, ScreenScore, and ShapeGauss). Consensus scoring (testconf = none) was also calculated. Knox et al.[20] have previously shown FRED2.01 in conjunction with ChemGauss and PLP to provide excellent enrichment with ERα validating our choice of docking algorithm and set of scoring functions.

Descriptor Correlation. A variety of 2D descriptors (available in MOE)[19] were calculated for all molecules and plotted against the ranked position produced by each scoring function for each molecule (Figure 2).

A trend line was calculated for each scoring function. The descriptors which exhibited a trend toward the higher ranking molecules were selected, and the correlation among them was explored. Correlation matrices were calculated using MOE.2005.06 by application of Spearmen's formula

$$R_s = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Spearmen's rank correlation measures the correlation between two sequences' *f* values. The two sequences are ranked separately and the differences in rank are calculated at each position (D). Spearmen's correlation is from -1 to 1. N represents the total number of elements in the list.

Rescoring Procedure. To negate the trend toward higher molecular weight compounds being prioritized at the top of the database, a normalization procedure was introduced as follows. First, the scores resulting from each scoring function and the corresponding MW of each were normalized according to

$$\text{normValue} = (x - \text{min})/(\text{max} - \text{min})$$

where x represents either the MW or the actual score for each compound and min and max are the minimum and maximum values, respectively, of the ranked set.

The ranked list was then rescored to reduce the dependence of the scoring function on molecular weight. A new power function formula was introduced to penalize only those

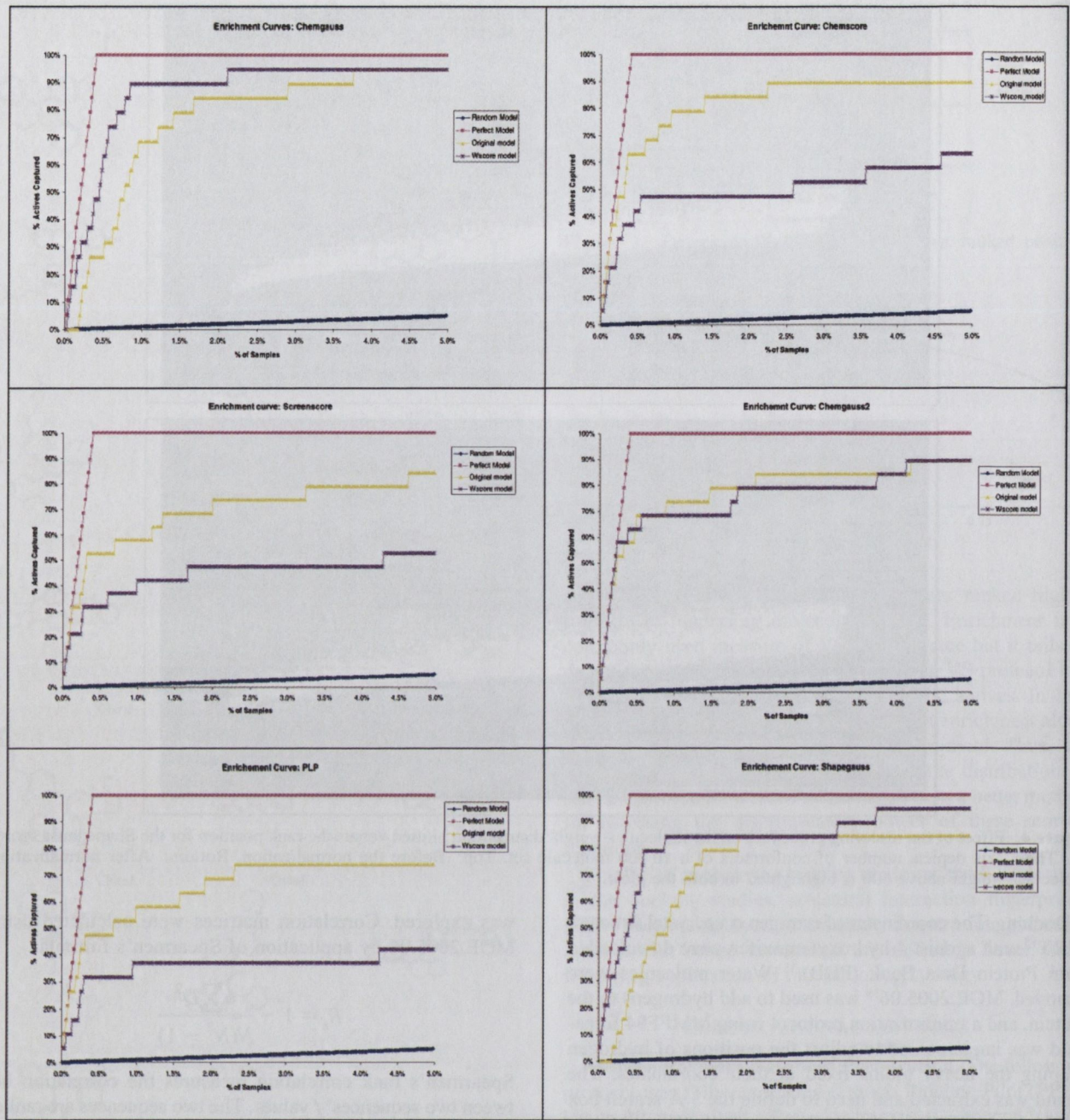


Figure 5. Enrichments curves before and after rescoring. Wscore represents the enrichment curve post rescoring. Original model represents the enrichment curve prior to rescoring.

molecules which exhibit high MW and leave the remainder relatively unchanged

$$\text{corrScore} = \text{normScore}\,(1 - \text{normMW})^{0.25}$$

where normScore is the normalized score and normMW is the normalized MW according to the previous formula.

A visual representation of the power function is depicted in the Figure 3.

ANALYSIS OF THE RESULTS

Enrichment. The potential of the scoring functions in the identification of a set of known active ligands from a set of druglike "decoys" is typically measured using the metric enrichment (E). Enrichment is a measure of the proportion of hits retrieved in a subset of compounds compared with the proportion of hits expected from a random sample of compounds

$$E = \frac{\text{Hits}_{\text{sampled}}/N_{\text{sampled}}}{\text{Hits}_{\text{total}}/N_{\text{total}}}$$

Enrichments were calculated over all the scoring functions, before and after rescoring procedure.

Scoring Frequencies. Scoring function results can be analyzed by plotting the score frequencies. In other words, how many molecules (%) receive a certain score. Scoring frequencies were normalized for ease of plotting, and results prior to and post rescoring were analyzed by this method.

SIFt Fingerprints. Structural interaction fingerprints (SIFt)[21] are used in this study to capture the interactions occurring between each docked pose of a compound and the residues of the active site of ERα. VS results for a set of protein-kinase inhibitors have previously been shown to produce superior enrichment using SIFt compared with traditional scoring functions.[14] It has also been used to design target-focused libraries because of its inherent sensitivity.

In this light, we have employed SIFt, available in MOE.2005.06, to calculate the 1D binary strings. The Tanimoto coefficient[22] was used to discern the similarity between the string generated for the PDB entry 3ERT and the ranked poses. The top scoring 200 molecules for each scoring function were assessed for similarity to the crystal structure with an overlap of 65, 70, 75, 80, and 85%.

RESULTS

A large number of descriptors correlated well with MW according to Spearmen's coefficient (>0.60 × 100), and for this reason, we selected MW as the descriptor to apply our power function. It is important to note however that additional descriptors such as Oprea's number of rigid bonds (opr_brigid), number of rings (opr_nrings), and number of rotatable bonds (nrot) did not correlate well with MW, giving r^2 values of 0.59, 0.52, 0.49 × 100. Thus, limiting of the score normalization to solely MW values with a view to the removal of ligand bias may not be wholly inclusive to effect maximal impact on performance. However, given the clear correlation with MW in this study for the ER, for the purposes of this article we apply our methodology to only MW for illustrative purposes. The concept/methodology can however be expanded and applied to any descriptor with an associated score or rank in a hitlist.

The graphics depicted in Figure 4 show the effect of the rescoring procedure on the molecular-weight distribution plotted versus the rank position for the Shapegauss scored set.

Molecules with MW above 500 are highlighted, and moving from the first plot to the second, we can observe these molecules are more evenly distributed throughout the ranked set. This will serve to concentrate the scoring by each function on actual interactions rather than because of MW.

The enrichments curves retrieved for the scoring functions as implemented in FRED2.1.1 before and after rescoring, are showed in Figure 5. Shapegauss and Chemgauss rescored results (Wscore) show higher enrichment values when compared with the original scoring functions. Chemgauss2 exhibits neither improvement nor a reduction of enrichment. PLP, Chemscore, and Screenscore original score enrichments outperform the rescored enrichment results.

To further investigate the increase in enrichment observed for Shapegauss and Chemgauss, but reduction for PLP, Chemscore, Screenscore, MW versus the ranked position was plotted for all scoring functions. The upper and lower MW for the active set lies between ~340−480, and from Figure 6, it is clear that prior to the rescoring procedure, both Shapegauss and Chemgauss retrieve molecules with a MW outside this range. After implementation of our rescoring method, the molecules retrieved by Shapegauss and Chemgauss are now in the same MW range as the active set, thus producing a higher enrichment. It is also shown that

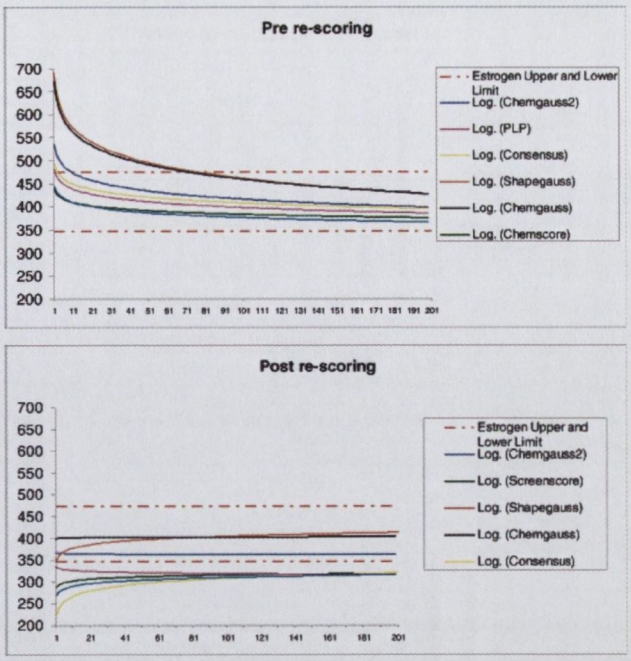


Figure 6. Molecular weight versus the ranked position of each molecule plotted for all scoring functions. Broken-line indicates the upper and lower MW for the set of ER actives.

Chemgauss2 remains unaffected, correlating well with lack of change observed in enrichment.

However, there is an appreciable shift toward the retrieval of molecules that have a lower MW than the active set. This would obviously result in an enrichment reduction for those particular scoring functions (Chemscore, PLP, and Screenscore).

To examine the full effect of our re-scoring procedure on both the active and decoy sets, it is necessary to look at the score distribution of each molecule for all the scoring functions (Figure 7).

The graphs in Figure 7 represent the score distributions for Chemgauss, Chemscore, and Chemgauss2. For conciseness, the results for just three scoring functions are depicted, being expressive of three the different scenarios obtained from the enrichment results: (1) rescoring outperforms original score (Chemgauss), (2) the original score outperforms the rescoring method (Chemscore), and (3) no appreciable change (Chemgauss2)

In the case of Chemgauss, a shift for the whole ranked set can be seen, with greater separation of the actives from decoys also. The discrimination between actives and decoys appears to decrease significantly for Chemscore, with the distribution of the estrogen remaining almost unchanged, and the decoy set shifting higher. In the case of Chemgauss2, no significant difference is observed in the scoring distribution of the dataset.

Finally, to critically analyze and classify the type of compounds being retrieved from each screen, we introduced the SIFt analysis to the procedure. The goal was to establish if compounds that exhibit the best interactions within the binding site of ERα are shifted higher in the resultant hitlist post rescoring. The *x*-axis indicates the allowed overlap and in the *y*-axis the numbers of molecules with that percentage overlap. Thus, the higher the number of molecules retrieved for a given overlap translates into closer protein−ligand

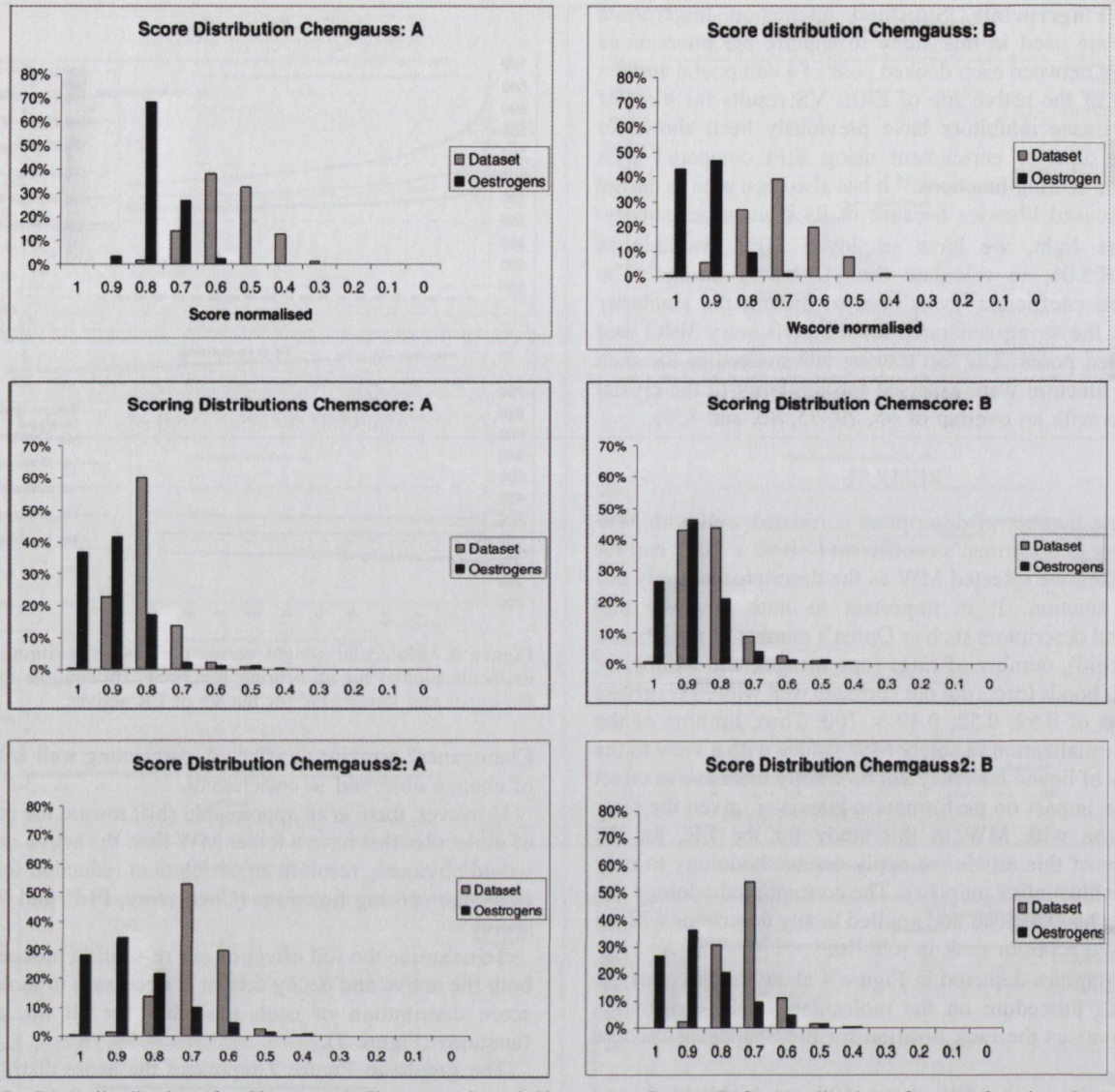


Figure 7. Score distributions for Chemgauss, Chemgauss2, and Chemscore. Set A: Before application of power function. Set B: After application of power function.

interactions to the ones observed in the crystal structure now being made for the top molecules in the ranked dataset.

The graphs (Figure 8) illustrate that after the rescoring procedure the molecules recovered with an overlap of 65−70% increase for all the scoring functions. For the scoring functions, Shapegauss, Chemgauss, and Chemgauss2, better interactions are observed for all the higher ranked molecules with better overlapping at 70−85% also.

As we have seen, after rescoring there are improvements in the abilities of Chemgauss and Shapegauss to discriminate between actives and decoys in the database used. To further illustrate the value of unbiasing the scoring functions with respect to MW, we formed a consensus score consisting of adding the rank by rank values from the original positions of PLP, Chemscore, Screenscore, and Chemgauss2 and the rescored ranks of Chemgauss and Shapegauss.

The new consensus enrichment (Table 1) obtained higher enrichment value in the first 0.20% of the database, while the values in the rest of the database remained the same. The benefit of this procedure is evident considering the formula can be tuned accordingly by altering the exponential value to penalize more heavily compounds with high or low MW. We have not attempted to tune the function to its maximum potential because the purpose of the study was not to achieve high enrichment values but to prevent high MW compounds from ranking highly in a database screen.

DISCUSSION

In this work, we propose a procedure to unbias scoring functions with regard to MW. It has been noted previously by Jacobsson et al.[23] that ligand bias is a source of false positives in structure-based virtual screening. Importantly, scoring functions are generally correlated with molecular size because of their inherently additive nature. Par et al.[9] proposed size normalization by dividing by the square or cubic root of the heavy atom count (HAC) to negate the effect of this dependence.

Plotting the ranked position of compounds versus their associated descriptors is a fast way to inspect the characteristics of the compounds which scoring functions rank better than others. We found the Spearmen coefficient can be readily employed to explore possible relations between

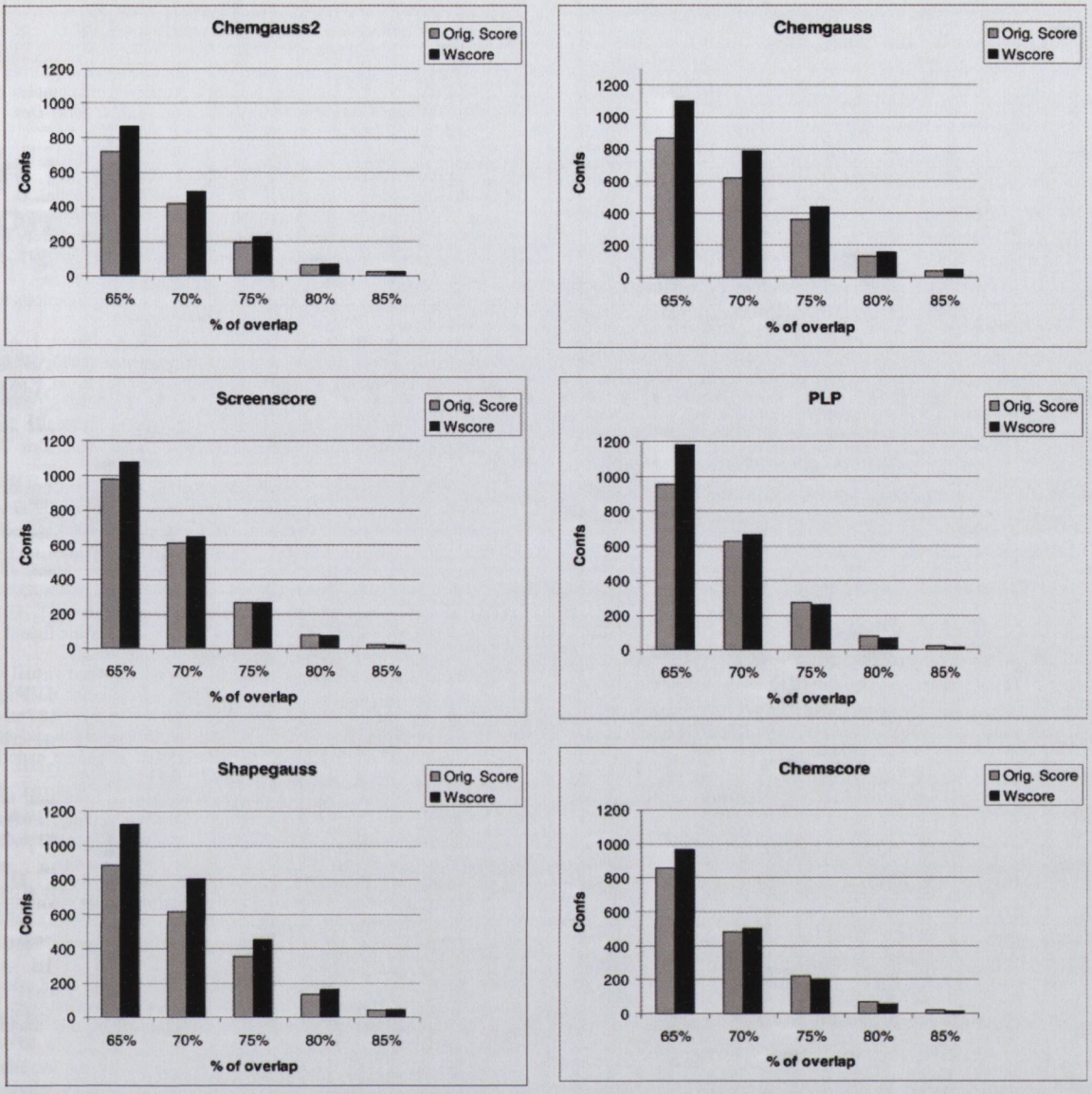


Figure 8. SIFt analysis results for all the scoring functions. The *x*-axis indicates the percent overlap of the first 200 molecules prior to and after application of our power function.

Table 1. Enrichment Calculated for 0.2, 0.5, 1, 1.5, 2, 3, and 5% of Conformer Database[a]

	0.20%	0.50%	1%	1.50%	2%	3%	5%
ligand no.	9	23	47	70	94	142	236
enrichment (pre-rescoring)	193	140	79	57	42	28	18
enrichment (post-rescoring)	220	140	79	57	42	28	18
theoretical maximum	248	204	100	67	50	33	20

[a] Consensus enrichment obtained with the original consensus scoring function versus that obtained after rescoring.

descriptors values and ranking position. Using this knowledge, we have introduced a function that can be tailored or tuned to lightly or heavily impact this MW bias. This formula penalizes exponentially higher molecular weight molecules and leaves other molecules untouched. Of great importance in this respect is that it is calculated post-scoring and thus can be reevaluated easily. The ranked set can also be easily heavily biased toward the retrieval of more leadlike MW compounds without loss in their ability to interact correctly with residues of the active site by increasing the value of the exponential in our function.

It is important to note that not all descriptors correlated well with MW, and application of our function to these descriptors would probably enhance the prioritization of actives. In fact, typical molecules containing heavy elements such as Cl or S may be treated unfavorably by this strategy, and so we suggest application of our strategy to a broader range of descriptors. We suggest its use in a consensus weighting scheme to remove ligand bias overall.

Although improvements in enrichments were observed for Chemgauss and Shapegauss after the application of our formula, a reduction was also seen for the remaining scoring functions with the exception of Chemgauss2. However, SIFt analysis showed that, although there is a fall in enrichment for some of the scoring functions, there is conversely an improvement in the quality of possible false positives in the ranked dataset.

A better understanding of the full impact of applying our formula can be derived from the score distributions because

it brings more information about how scoring functions rank all the compounds in the database and not just the positive control. Important feedback results from analysis of the scoring frequencies. For example, for Chemscore, it was clear that it outperformed most of the scoring functions in discriminating between actives and decoys. The rescoring procedure clearly decreases the enrichment value; however, analysis of the score distribution shows that on average the score of the positive controls remain more or less unchanged. Thus, the rescoring procedure brings up more molecules from the decoy set among the higher rankings. It is possible that some of these molecules that are now ranked highly with equivalent scores to those of the antiestrogens are actually actives because the database chosen was one compiled from commercially available compounds.

Finally, we demonstrate the benefit of adjusting the bias of individual scoring functions for incorporation in a consensus scheme. Using the ranks obtained for rescored Shapegauss and Chemgauss in conjunction with the ranks obtained for the original versions of Chemgauss2, PLP, Screenscore, and Chemscore, we generated higher enrichment values.

Our findings, though centered on examining the phenomenon in relation to a single receptor, are derived from an examination of the additive nature of the scoring functions selected. As such, we anticipate similar improvements can be attained in other target classes: indeed, we advocate similar target-specific examination of scoring bias during method development to maximize the performance of virtual screening.

These methods can be readily applied to any descriptors (QSAR, etc.) to tune the scoring functions toward retrieval of molecules with defined properties and unbias particular parameters within a scoring function if the user has prior knowledge of the type of ligands needed to be selected.

ACKNOWLEDGMENT

We acknowledge the funding support from Science Foundation Ireland (SFI). We thank the software vendors for their continuing support of our academic research efforts, in particular, the contributions of Openeye Scientific Software, Chemical Computing Group (CCG), and Accelrys/Scitegic (Pipeline Pilot).

REFERENCES AND NOTES

(1) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B.; Glen, R. C. Predicting protein−ligand binding affinities: A low scoring game? *Org. Biomol. Chem.* **2004**, *2*, 3267−73.

(2) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C.E.; Head, M. S. A critical assessment of docking programs and scoringfunctions. *J. Med. Chem.* **2006**, *49*, 5912−31.

(3) Bohm, 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*, 243−56.

(4) *FRED (Fast Rigid Exhaustive Docking)*, version 2.11; Openeye Scientific Software: Santa Fe, NM, 2006.

(5) 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*, 425−45.

(6) Stahl, M.; Rarey, M. Detailed analysis of scoring functionsfor virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−42.

(7) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*,76−90.

(8) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand−protein complexes. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 731−51.

(9) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−72.

(10) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−303.

(11) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., 3rd. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 755−63.

(12) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−67.

(13) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comput.-Aided. Mol. Des.* **2004**, *18*, 333−44.

(14) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: A new approach to organizing, mining, analyzing, and designing protein−small molecule complexes. *Chem. Biol. Drug. Des.* **2006**, *67*, 5−12.

(15) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *10*, 283−99.

(16) *OMEGA*, version 1.8.1; Openeye Scientific Software: Sana Fe, NM, 2006.

(17) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927−37.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−42.

(19) *MOE*, version 2006.08; Chemical Computing Group: Montreal, Quebec, Canada, 2006.

(20) Knox, A. J.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in compound database preparation−"Hidden" impact on virtual screening results. *J. Chem. Inf. Model.* **2005**, *45*, 1908−19.

(21) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337-44.

(22) Tanimoto, T. T. IBM Internal Report. Nov. 17, 1957.

(23) Jacobsson, M.; Karlen, A. Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1134−43.

CI600471M

Chem Biol Drug Des 2007; **70:** 123–133

Research Article

© 2007 Johnson & Johnson Pharmaceutical Research & Development LLC
Journal compilation © 2007 Blackwell Munksgaard

doi: 10.1111/j.1747-0285.2007.00538.x

On the Effects of Permuted Input on Conformational Sampling of Drug-like Molecules: an Evaluation of Stochastic Proximity Embedding

Dimitris K. Agrafiotis[1,*], Deepak Bandyopadhyay[1], Giorgio Carta[2], Andrew J. S. Knox[2] and David G. Lloyd[2]

[1]*Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, PA 19341, USA*
[2]*Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland*
**Corresponding author: Dimitris K. Agrafiotis, dagrafio@prdus.jnj.com*

Conformational sampling is a problem of central importance in computer-aided drug design. A good conformational search method must not exhibit any intrinsic bias, and must provide confidence that important regions of conformational space are not missed during the search. A recent study by Carta *et al.* showed that this is not always the case, and that several popular conformational search methods, such as Omega, are very sensitive to the relative ordering of atoms and bonds in the connection table. Here, we examine the performance of a newer method known as stochastic proximity embedding, or SPE, using five diverse bioactive ligands extracted from the PDB. Our results confirm that the conformational ensembles produced by SPE using different permuted inputs are statistically indistinguishable, and well within the range of variability that would be expected from the stochastic nature of the method itself. This, along with the results of a more comprehensive comparative study (Agrafiotis *et al., J. Chem. Info. Model,* 2007, in press), provides further evidence that SPE is one of the most robust and competitive conformational search methods described to date.

Key words: bioactive conformation, boosting, conformational analysis, conformational sampling, distance geometry, SPE, stochastic proximity embedding

Received 10 May 2007, revised 15 June 2007 and accepted for publication 17 June 2007

The properties and reactivity of organic molecules depend intimately on their three-dimensional shape. Most organic molecules of non-trivial size can assume a multitude of three-dimensional conformations. Identifying which of these conformations are relatively stable and likely to be populated at room temperature has been the subject of extensive study in the computational chemistry literature (1). This problem is particularly critical in computer-assisted drug design. Recent studies of crystal structures of protein–ligand complexes have shown that bioactive conformations tend to be more extended than random ones (2), and may lie several kcal/mol higher in energy than their respective global minima (3). Several applications depend critically on the diversity of conformations sampled during the search, including protein docking, pharmacophore modeling, 3D database searching, and 3D-QSAR, to name a few. Since the bioactive conformation of a ligand also depends on the geometry of its host, it is imperative that the search for conformational minima casts a wide net over the potential energy surface.

Conformation generation algorithms fall into two broad categories: deterministic, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic, which use a random element to explore the molecule's conformational space. Validation studies (4–7) and practical experience suggest that no individual method is decisively better than the others. Systematic search can be very effective for molecules with limited conformational flexibility, but the exponential growth of the search space with the number of rotatable bonds and problems associated with ring closures limit its utility as a general conformational sampling technique (8–11). Stochastic methods, particularly those based on iterative perturbation such as molecular dynamics and Monte Carlo sampling (12–14), have broader applicability, but tend to generate many transitional conformations between local minima and, as a result, they are generally slower.

An alternative approach, known as distance geometry (15,16), is to generate conformations that satisfy a set of interatomic distance constraints derived from the molecular connectivity table and defined in the form of lower and upper distance bounds $\{l_{ij}\}$ and $\{u_{ij}\}$.[†] Distance geometry methods involve four basic steps: (i) generation of interatomic distance bounds, (ii) assignment of a random value to each distance within the respective bounds, (iii) conversion of the resulting distance matrix into a starting set of Cartesian co-ordinates, and (iv) refinement of the co-ordinates through minimization of an error function measuring the violation of the input

[†]In addition to distance constraints, distance geometry methods also employ volume constraints that prevent the signed volume formed by four atoms from exceeding certain limits. These constraints are used to enforce planarity of conjugate systems and correct chirality of stereocenters.

constraints. To ensure that reasonable conformations are generated, the original upper and lower bounds are usually refined using an iterative triangular smoothing procedure. Although this process improves the initial guess, the randomly chosen distances may still be inconsistent with a valid 3-dimensional geometry, necessitating expensive metrization schemes (16,20,21) or higher dimensional embeddings (13) prior to error refinement, or lengthy refinement procedures if random starting co-ordinates are used.

Stochastic proximity embedding (SPE) is a relatively recent technique for finding molecular geometries that satisfy distance constraints in an efficient manner (17–19). The method starts from random initial atomic positions, and gradually refines them by repeatedly selecting a pair of atoms at random, and updating their respective co-ordinates toward satisfying the corresponding distance bounds. This procedure is performed repeatedly until a reasonable conformation is obtained.‡

While SPE was originally shown to provide a good sampling of conformational space, it was observed that 'extreme' conformations located near the periphery of conformational space were not as likely to be visited, and in principle some important conformations could be missed. To remedy this problem, a boosting heuristic was introduced that can be used in conjunction with SPE to bias the search toward more extended or more compact geometries (20). The method generates increasingly extended (or compact) conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed, generating a chemically sensible conformation c_1. The lower bounds of all atom pairs $\{l_{ij}\}$ are then replaced by the actual interatomic distances $\{d_{ij}\}$ in conformation c_1, and used along with the unchanged upper bounds $\{u_{ij}\}$ to perform a second embedding to generate another conformation, c_2. This process is repeated for a prescribed number of iterations, yielding additional conformations. The lower bounds are then restored to their original default values, and a new sequence of embeddings is performed using a different random number seed. This process will never yield a set of distance constraints that are impossible to satisfy, because there exist at least one conformation (i.e. the one generated in the preceding iteration) that satisfies them. An analogous procedure can be used to generate increasingly compact conformations. The method was validated against seven widely used conformational sampling techniques, and was found to be significantly more effective in sampling the full range of geometric sizes attainable by any given molecule compared to the other methods (21).

However, it is well known that many stochastic 3D modeling techniques are very sensitive to starting configurations and random number effects, and the results are often difficult to reproduce when the search is repeated under slightly different initial conditions. Carta *et al.* (22) demonstrated that this reproducibility problem plagues systematic methods as well. More specifically, they examined how different permutations of the connection table affected the conformations generated by Corina[a], Omega[b], Catalyst[c], and Rubicon.[d] The authors used Daylight and in-house utilities to generate different (non-canonical) variants of SMILES and SDF representations for 17 bioactive ligands, effectively changing the order of the atoms and bonds while keeping the topology intact. Each variant was subjected to conformational search, using the same set of parameters for conformer generation. The results were evaluated by comparing the distributions of several 3D descriptors in the conformational ensembles produced by each permuted input. While Catalyst showed negligible sensitivity, Omega and Rubicon produced substantially different ensembles, suggesting that these methods have an intrinsic bias that is highly dependent on the atom and bond ordering. The work clearly indicated the benefit of including multiple input permutations in improving the sampling of conformational space by established methodologies.

Here, we examine the sensitivity of SPE to permutation of input structure representation using a diverse subset of five bioactive ligands from our reference collection. Our results confirm that the conformational ensembles produced by different permuted inputs are statistically indistinguishable, and within the range of variability that would be expected from the stochastic nature of the method.

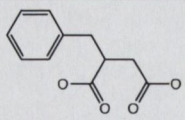
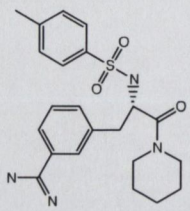
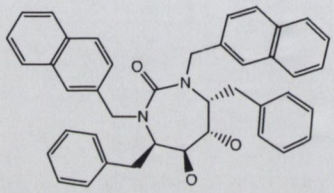
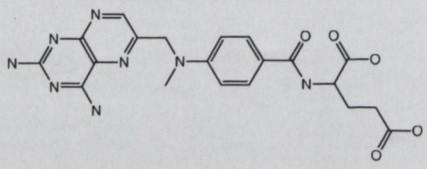
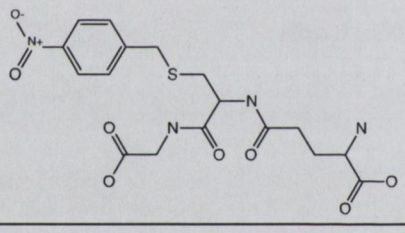
Methods

The comparison was based on five bioactive ligands extracted from the Protein Data Bank (23,24), containing between 5 and 13 rotatable bonds. Their structures, PDB codes, and number of rotatable bonds are listed in Table 1, along with the number of SPE trials and the number of conformational boosts employed in each trial (vide infra). Each ligand was neutralized and the proper chirality was assigned to each chiral center, as inferred from the PDB 3D co-ordinates. For 1CBX, 4DFR, and 1GLQ, these chiral flags were subsequently cleared to study the performance of SPE in the presence and absence of stereochemistry. For 1ETT and 1HVR, the chiral flags remained as shown in Table 1.

For each ligand, five different SMILES and five different SDF variants were generated, using the Daylight utility permsmi (25) for the former and an internally developed program for the latter. Each variant was used as input to SPE, which was instructed to produce between 1000 and 10 000 conformations for each one, depending on the size of the molecule (1000 conformations for each 1CBX variant, 5000 for each 1ETT and 1HVR variant, and 10 000 for each 4DFR and 1GLQ variant). The SPE conformer generator reads both SMILES and SDF files, and does not necessitate explicit file conversion. The input structures are read into an internal c++ data structure, which preserves the order in which the atoms and bonds were specified in the input stream (i.e. no canonicalization takes place). Thus, we are certain by virtue of the

‡The actual algorithm involves stochastic refinement of all geometric constraints, both distance and volume. The co-ordinate adjustments differ for each type of constraint, for example, a volume constraint is minimized by moving the positions of the four atoms associated with that constraint in the direction opposite to the gradient of an error function that measures the deviation from the target value. The algorithm also uses a number of parameters such as learning rates and sampling frequencies to control the self-organization process. We refer the reader to ref. 19 for a complete description, validation, and performance benchmark of the algorithm, as well as a comparison to other distance geometry methods.

Table 1: PDB codes, structures, number of rotatable bonds, number of SPE trials, and number of boosts for each of the molecules in the data set

PDB code	Structure	Rotatable bonds	Trials	Boosts[a]
1CBX		5	1000	0
1ETT		7	5000	1
1HVR		8	5000	1
4DFR		9	10 000	1
1GLQ		13	10 000	1

[a]Number of boosts per trial. For example, the 5000 conformations for 1ETT were obtained by generating 2500 conformations using standard distance bounds, and boosting each of these conformations once to obtain 2500 additional conformations.

architecture of our program that each variant resulted in a unique internal atom and bond numbering order as well.

Given the size and flexibility of the molecules in our data set, we used one level of boosting for 1ETT, 1HVR, 4DFR, and 1GLQ, and no boosting for 1CBX (Table 1). For example, the 10 000 conformations for 4DFR were obtained by generating 5000 random conformations using the standard distance bounds derived from the connection table, and boosting each of these conformations once to obtain 5000 additional conformations (10 000 = 2 × 5000). As SPE proper favors compact geometries, boosting was applied to the lower bounds only, leading to more extended conformations.

All SPE embeddings were carried out on hydrogen-depleted graphs using the following parameters: $\lambda_d = \lambda_v = 1$, $C = 50$, $\delta\lambda_d = \delta\lambda_v = 0.9/(C - 1)$, $S = 50 \times N$, and

$$v = \min\left(0.5, 8.0 \times \frac{N_V}{N(N+1)/2 + 8N_V}\right),$$

where λ_d and λ_v are the initial distance and volume learning rates, C is the number of cycles, S is the number of steps, $\delta\lambda_d$ and $\delta\lambda_v$ are the decrements of the distance and volume learning rates in each cycle, v is the ratio of volume to distance refinements, N is the number of atoms in the molecule, and N_V is the total number of volume constraints [for a detailed description of these parameters, see Ref. (19)]. Each conformation was derived by running the algorithm with a different random number seed.

All the raw conformations generated by SPE were minimized using the MMFF94s force field (26–30) and the BFGS variable metric

minimization algorithm, as implemented in the DIRECTEDDIVERSITY (31) software. This software has been tested thoroughly and has passed the entire MMFF94s validation suite.[e] For each permuted input, duplicate conformations were identified and eliminated in a post-processing step using an RMSD threshold of 1.0 Å. More specifically, each new minimized conformation c was compared to all previously unique conformations, and if the RMSD to all of them was >1.0 Å, c was added to the unique set. If, on the other hand, there was an existing conformation c^* in the unique set such that RMSD(c, c^*) ≤ 1.0 Å, the one with the lowest energy between the two was retained, and the other was discarded as a duplicate. The RMSD was computed by taking topological symmetry into account, i.e. all possible mappings of topologically equivalent atoms were considered, and the alignment with the lowest RMSD was used to measure the similarity of two conformations. From the resulting non-duplicate lists, the lowest energy conformation for each variant was identified, and all conformations with energies >20 kcal/mol from that conformation were discarded.

All calculations and postprocessing were performed within the DirectedDiversity (31) and Third Dimension Explorer (32) environments, except ANOVA, which was carried out in Microsoft Excel. All calculations were carried out on an IBM Intellistation running Windows XP Professional, and equipped with two 3.2 GHz Xeon processors and 2048 Mb of RAM.

Results and Discussion

Following the approach of Carta *et al.* (22) we analyzed the conformational ensembles produced by the different permuted inputs through both statistical and visual means. Being a stochastic method, SPE is not guaranteed to produce the same results when run under different initial conditions. Different runs, starting from different random number seeds, are expected to produce slightly different sets of conformations, but these should converge as sampling becomes more exhaustive and the search space becomes saturated. This is illustrated in Figure 1, which shows the number of unique conformations discovered at the end of each trial for each SMILES (black) and SDF (red) permuted input for each of the five molecules. When contrasted with multiple independent random runs of only one of the permuted inputs (Figure 1E for 4DFR), it becomes clear that the input representation does not have any discernible impact, and that the variation in the sampling profiles is within the range expected by the stochastic nature of the method.

The lowest energy and the number of unique conformations identified after the prescribed number of trials for each molecule and each permuted input are summarized in Table 2. These numbers represent the conformations within 20 kcal/mol from the lowest energy structure identified by each permuted input. For 1CBX, which is a small molecule with few conformational states, every run identified the same global minimum (E = −50.58 kcal/mol). However, there were some differences in the number of local minima identified by each variant, which ranged from 21 to 25, with no apparent dependence on the input format. As the size of the molecule and the number of accessible conformations increases, differences between variants become more likely, and this is reflected in both the lowest energy minimum and the number of local minima identified by each method (170–189 for 1ETT, 340–353 for 1HVR, 411–433 for 4DFR, and 3103–3543 for 1GLQ). Again, there is no apparent interdependence between the input representation (SMILES versus SDF) and the number of conformers generated.

The ensembles produced by each permuted input were further compared using three different conformational properties: (i) the potential energy computed by MMFF94s; (ii) the root mean square deviation (RMSD) from the X-ray crystal structure; and (iii) the radius of gyration, which measures a conformation's degree of compactness (or extendedness). The cumulative distributions of the RMSD from the crystal structure of the low-energy conformations identified for each molecule and each permuted input are illustrated in Figure 2. In all cases, the conformational ensembles seem to be drawn from the same probability distribution and the variation appears to be random and not dependent on the input representation.[§] To obtain a more robust statistical estimate, we compared these distributions using single-factor ANOVA. ANOVA is a well-established technique that attempts to test the null hypothesis that two or more mean values are equal, by comparing the variation observed within each class against the one observed across the entire sample. The key parameters of an ANOVA study include the F-statistic which is compared to the critical value of the F distribution, F_{crit}, and the p-value which represents the smallest level of significance for which the observed sample information becomes significant, provided the null hypothesis is true. The results, which are summarized in Table 3, show that the differences in the conformational property distributions of the permuted inputs are statistically insignificant, except for the energy distribution of the conformations of 1GLQ.[¶] This result is not surprising in light of the fact that the conformational space of that molecule is grossly undersampled. As seen in Figure 1, the rate at which new conformations are discovered decreases exponentially with the number of trials, as the probability of visiting a previously seen conformation increases. While for 1CBX, 1ETT, 1HVR, and 4DFR the search appears to have converged, for 1GLQ new minima continue to be discovered at a fast rate even after the 10 000 trials. These plots also show that the permuted inputs explore the same range of conformational space throughout the sampling process, i.e. the number of unique conformations identified by each variant is comparable after any number of trials.

These results contrast with the original findings of Carta *et al.*, who showed that the low-energy conformations generated by Omega and Rubicon for each permuted SMILES string exhibited a

[§]Similar distributions were observed for the potential energy and radius of gyration as well. The results are not shown here to conserve space.

[¶]Strictly speaking, ANOVA tests the equality of mean values and not the general similarity between two distributions, for which other statistics such as Kolmogorov–Smirnov's may be more suitable. Although one could imagine different distributions with the same mean but very different shapes, visual inspection of the plots in Figure 2 (and those of the other conformational properties which are not included here) show that their shapes are virtually identical.

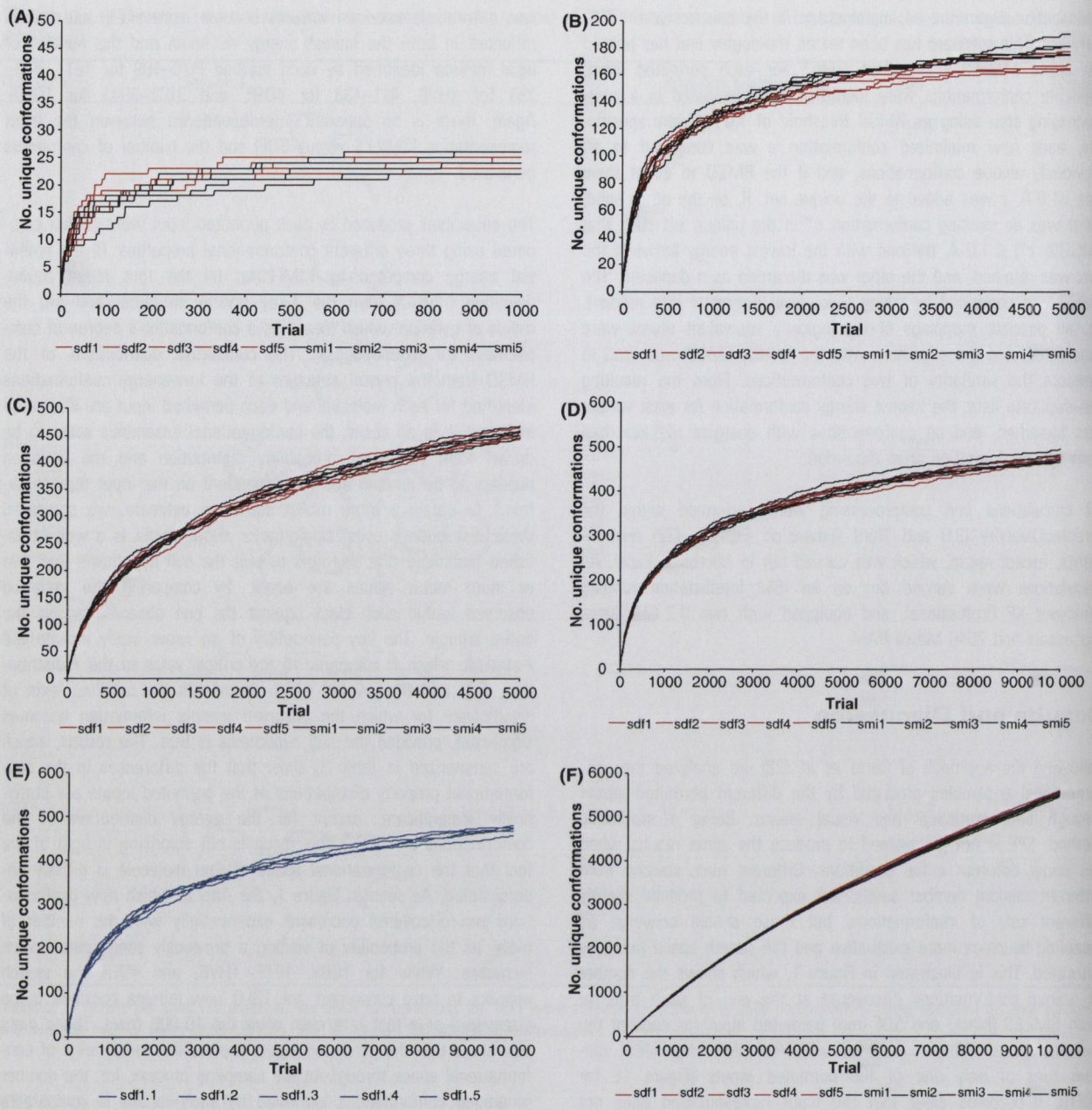


Figure 1: Number of unique conformations (within 1 Å RMSD) discovered at the end of each SPE trial for the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (F) 1GLQ. (E) Five independent runs of the first SDF permuted input of 4DFR, each starting from a different random seed.

different distribution of RMSD from the X-ray co-crystal conformation (22). Catalyst, on the other hand, was found to be insensitive to permuted input, yielding identical distributions for all SMILES variants. This is consistent with our recent comparison of SPE to seven other widely used conformational analysis programs, which showed that Catalyst's sampling capacity was comparable to SPE's, and significantly higher than all the other methods evaluated (21).

We should point out, parenthetically, that for the two chiral molecules (1ETT and 1HVR), not all trials resulted in valid geometries. SPE uses volume constraints to generate conformations that preserve the chirality of all explicitly defined stereocenters in the input file. As we noted in our original publication (19), the probability of failing to generate a correct 3D geometry increases with the number and proximity of chiral centers. As shown in Table 4, this has not been a serious issue for 1ETT and 1HVR. On average,

Table 2: Global minimum and number of unique conformations within 20 kcal/mol from the 'global' minimum identified for each permuted input

PDB code	Variant	E_{min}	Number of conformations
1CBX	SDF 1	−50.58	23
1CBX	SDF 2	−50.58	22
1CBX	SDF 3	−50.58	22
1CBX	SDF 4	−50.58	25
1CBX	SDF 5	−50.58	22
1CBX	SMI 1	−50.58	22
1CBX	SMI 2	−50.58	22
1CBX	SMI 3	−50.58	23
1CBX	SMI 4	−50.58	21
1CBX	SMI 5	−50.58	22
1ETT	SDF 1	−34.37	173
1ETT	SDF 2	−36.05	182
1ETT	SDF 3	−35.60	178
1ETT	SDF 4	−34.46	179
1ETT	SDF 5	−33.98	170
1ETT	SMI 1	−33.90	174
1ETT	SMI 2	−36.05	189
1ETT	SMI 3	−35.55	181
1ETT	SMI 4	−33.95	182
1ETT	SMI 5	−34.43	189
1HVR	SDF 1	102.37	341
1HVR	SDF 2	102.32	346
1HVR	SDF 3	102.32	344
1HVR	SDF 4	102.32	347
1HVR	SDF 5	102.32	351
1HVR	SMI 1	102.32	348
1HVR	SMI 2	102.32	352
1HVR	SMI 3	102.32	340
1HVR	SMI 4	102.32	342
1HVR	SMI 5	102.37	353
4DFR	SDF 1	−30.02	416
4DFR	SDF 2	−30.00	422
4DFR	SDF 3	−28.80	420
4DFR	SDF 4	−28.19	421
4DFR	SDF 5	−27.55	423
4DFR	SMI 1	−30.00	414
4DFR	SMI 2	−30.00	420
4DFR	SMI 3	−29.38	433
4DFR	SMI 4	−30.00	418
4DFR	SMI 5	−31.16	425
4DFR	SDF 1.1	−29.36	427
4DFR	SDF 1.2	−28.34	411
4DFR	SDF 1.3	−30.02	419
4DFR	SDF 1.4	−31.20	416
4DFR	SDF 1.5	−30.00	412
1GLQ	SDF 1	59.70	3543
1GLQ	SDF 2	58.62	3225
1GLQ	SDF 3	58.40	3103
1GLQ	SDF 4	58.40	3226
1GLQ	SDF 5	58.62	3197
1GLQ	SMI 1	58.62	3197
1GLQ	SMI 2	58.40	3151
1GLQ	SMI 3	58.40	3154
1GLQ	SMI 4	59.25	3421
1GLQ	SMI 5	59.25	3418

only 17 of 5000 trials failed to produce the correct stereochemistry for 1ETT, and only 38 of 5000 for 1HVR (which contains four adjacent chiral centers), and in both cases that failure rate was insensitive to permuted input. Note that energy minimization cannot correct such problems because of the enormous energy barriers involved in inverting a stereocenter. Because SPE is extremely fast (it can generate hundreds of conformations per second on a modern PC for a typical drug-like molecule), the raw conformations produced by SPE can be tested for the correct stereochemistry in a postprocessing step, and discarded if they fail to reproduce the input specification.

To provide a more visual assessment of the overlap in the conformational space covered by each permuted input, the unique low-energy conformations produced by the 10 variants of 4DFR were concatenated into a single list and were embedded into a two-dimensional nonlinear map using SPE. As mentioned in the Introduction, SPE was originally devised as a method to generate low-dimensional Euclidean embeddings that best preserve the similarities between a set of related observations (17,18). Unlike classical multidimensional scaling and nonlinear mapping, SPE scales linearly with respect to sample size, and can be applied to very large data sets that are intractable by conventional embedding procedures. In the case at hand, the map was constructed in such a way that the distances of the points on the map approximated as closely as possible the (dis)similarities of the respective conformations, as measured by their pairwise RMSD. As 4DFR is a flexible molecule with many conformational degrees of freedom, this projection leads to some loss of information, in the sense that the distances cannot all be preserved to perfection (the intrinsic dimensionality of the conformational space of 4DFR is roughly equivalent to the number of rotatable bonds, which is considerably >2). That loss of information is measured by the stress function, which in this particular case was 0.262.

Embedding all the conformations produced by all the 4DFR variants (4221) into a single map allows them to be placed on a common reference frame. The 10 panels in Figure 3 show the conformations generated by each permuted input on that common reference frame, color-coded by energy (from blue to red, in order of decreasing energy). Unlike other conformational search methods (22), the ensembles generated by SPE for each permuted input cover the same regions of conformational space and with the same relative density. Minor differences in the maps are attributed to two factors related to the stochastic nature of the method. First, if each unique conformation is seen as a representative of a cluster of related conformations within an RMSD radius of 1 Å, different ensembles may contain different representatives from each cluster, which are closely related but not identical. Secondly, because of the 20 kcal/mol energy cap, an ensemble that contains a 'global' minimum of lower energy than the other ensembles will be missing conformations on the high end of the energy spectrum. For example, the 'global' minimum identified by the fifth SMILES variant of 4DFR was −31.16 kcal/mol, whereas that identified by the first SDF variant was only −27.55 kcal/mol (Table 2). As seen in Figure 4, there

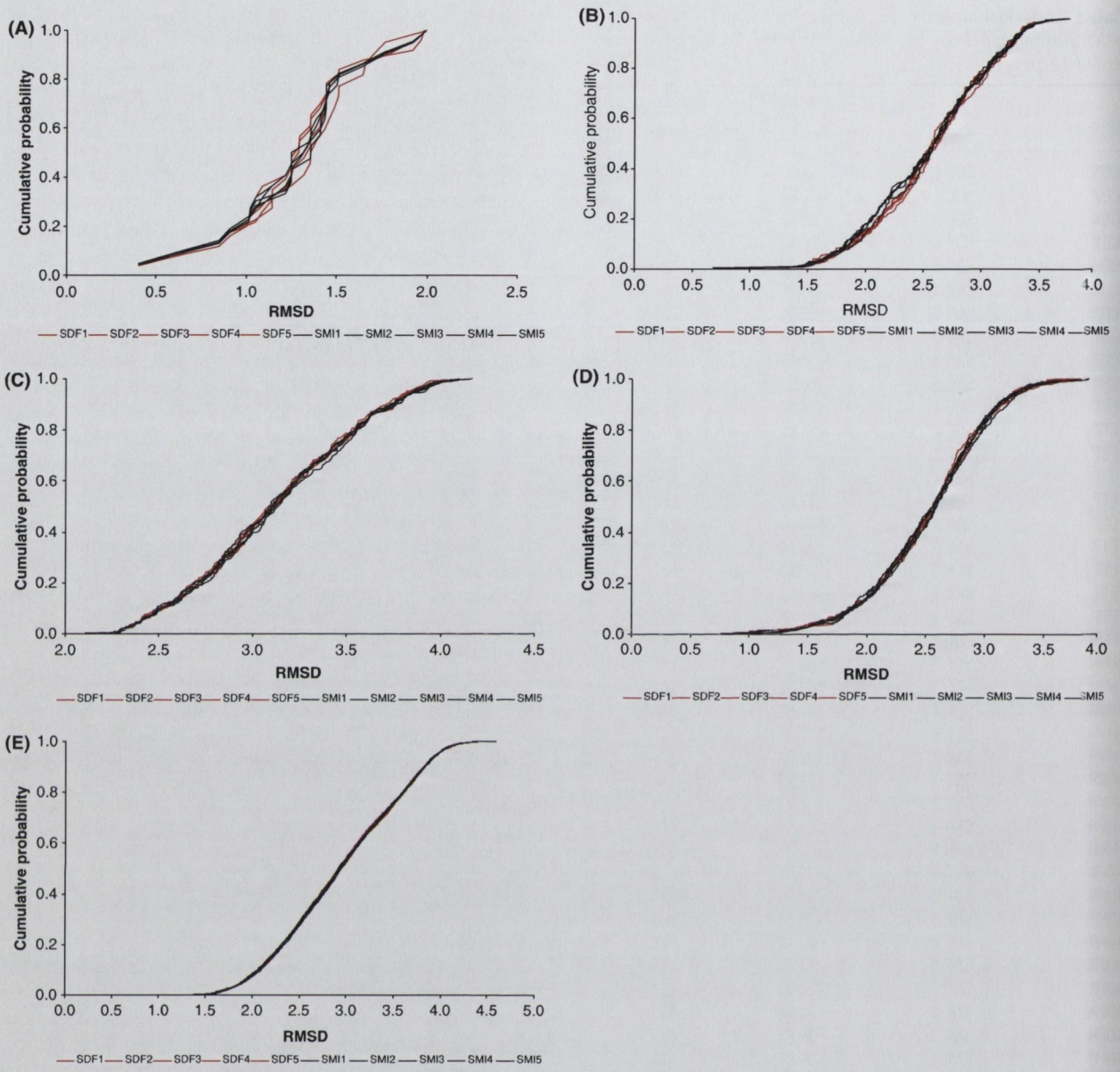


Figure 2: Cumulative distribution of RMSDs from the bioactive conformation of the conformational ensembles generated from the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (E) 1GLQ.

Table 3: ANOVA of the conformational properties of the ensembles produced by the permuted inputs for each molecule under investigation (when F is less than F_{crit}, the differences are considered to be statistically insignificant)

	Energy			RMSD			Radius of gyration		
PDB code	F	F_{crit}	p-Value	F	F_{crit}	p-Value	F	F_{crit}	pValue
1CBX	0.269	1.924	0.982	0.027	1.924	0.999	0.073	1.924	0999
1ETT	0.124	1.885	0.999	0.214	1.885	0.992	0.189	1.885	0995
1HVR	0.281	1.883	0.980	0.189	1.883	0.995	0.283	1.883	0980
4DFR	0.834	1.693	0.632	0.316	1.693	0.992	0.185	1.693	0999
1GLQ	13.337	1.880	1.45E-21	0.346	1.880	0.960	0.499	1.880	0376

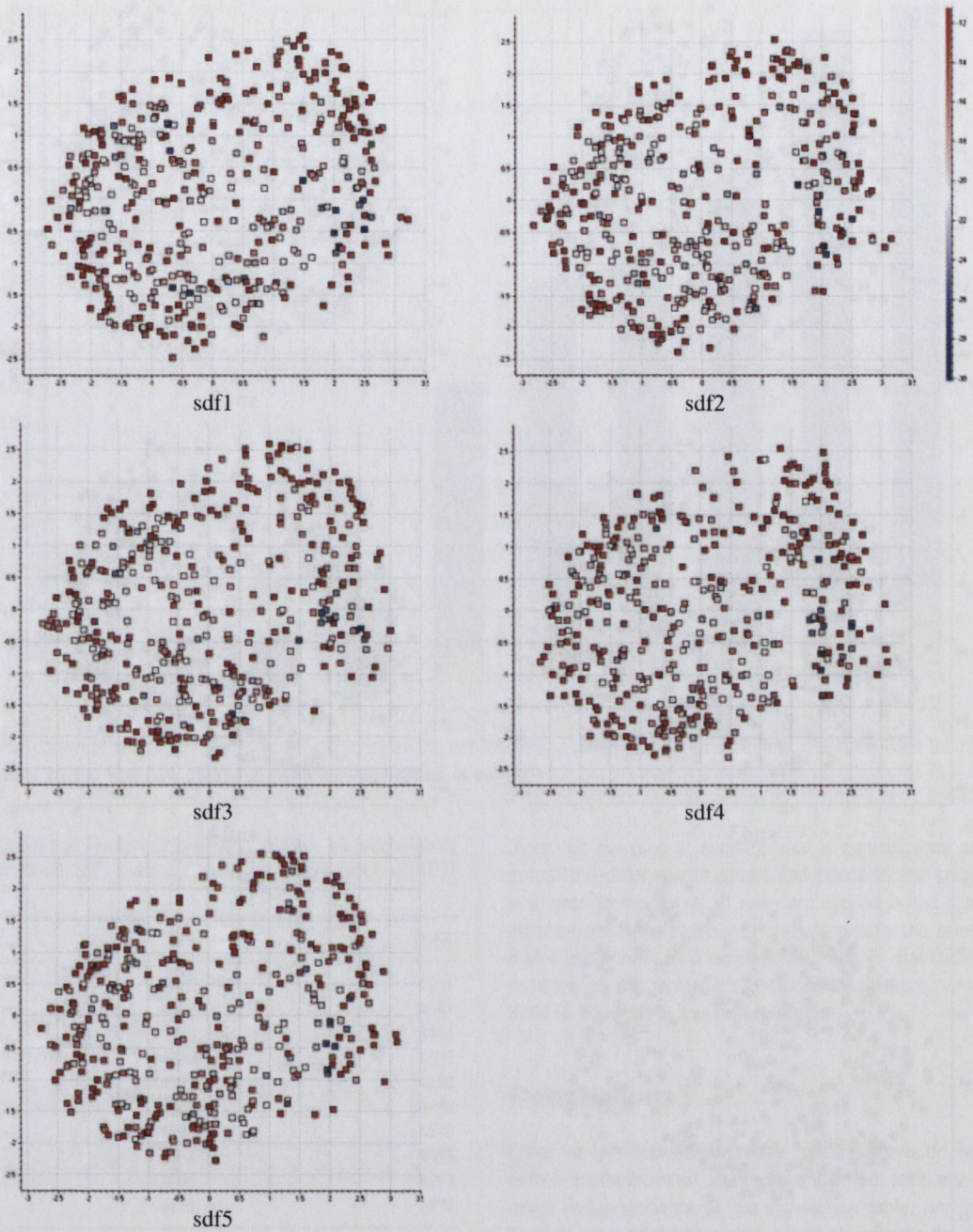


Figure 3: Two-dimensional SPE maps of the conformational ensembles generated from each SDF and SMILES permuted input of 4DFR, color-coded by their MMFF94s potential energy. These maps are constructed in such a way that the distances of the points in the map match as closely as possible the RMSDs of the respective conformations.

are considerably more conformations on the high end of the energy scale, which explains some of the observed differences in the maps`.**

These results are not at all surprising, if one considers the way in which the SPE conformer generator works. The process starts by establishing a set of distance and volume constraints from the molecular connectivity table. This is accomplished using a set of rules encoded in the form SMARTS patterns. These rules define groups of atoms that form a local substructure, along with the geometric constraints imposed by the local geometry of that substructure. The rules are supplied to the program in a separate file and can easily be extended and customized by the users. When a new molecule is read in, the patterns in the rule base are mapped onto the target molecule in sequence, and the corresponding constraints

**To conserve space, the SPE maps for the other four molecules are not included, but they do exhibit similar qualitative trends.

References

1. Leach A.R. (1991) A survey of methods for searching the conformational space of small and medium-sized molecules. In: Lipkowitz K.B., Boyd D.B., editors. Reviews in Computational Chemistry, Vol. 2. New York: VCH; p. 1–55.
2. Diller D.J., Merz K.M. Jr. (2002) Can we separate active from inactive conformations? J Med Chem;16:105–112.
3. Kirchmair J., Laggner C., Wolber G., Langer T. (2005) Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. J Chem Inf Model;45:422–430.
4. Bostrom J., Greenwood J.R., Gottfries J. (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model;21:449–462.
5. Bostrom J. (2001) Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comput.-Aided Mol Des;15:1137–1152.
6. Feuston B.P., Miller M.D., Culberson J.C., Nachbar R.B., Kearsley S.K. (2001) Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. J Chem Inf Comput Sci;41:754–763.
7. Perola E., Charifson P.S. (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem;47:2499–2510.
8. Lipton M., Still W.C. (1988) The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. J Comput Chem;9:343–355.
9. Bruccoleri R.E., Karplus M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers;26:137–168.
10. Bruccoleri R.E., Karplus M. (1985) Chain closure with bond angle variations. Macromolecules;18:2767–2773.
11. Go N., Scheraga H.A. (1970) Ring closure and local conformational deformations of chain molecules. Macromolecules;3:178–187.
12. Saunders M. (1987) Stochastic explorations of molecular mechanics energy surfaces. Hunting for the global minimum. J Am Chem Soc;109:3150–3152.
13. Ferguson D.M., Raber D.J. (1989) A new approach to probing conformational space with molecular mechanics: random incremental pulse search. J Am Chem Soc;111:4371–4378.
14. Chang G., Guida W.C., Still W.C. (1989) An internal coordinate Monte Carlo method for searching conformational space. J Am Chem Soc;111:4379–4386.
15. Crippen G.M. (1978) Rapid calculation of coordinates from distance matrices. J Comput Phys;26:449–452.
16. Spellmeyer D.C., Wong A.K., Bower M.J., Blaney J.M. (1997) Conformational analysis using distance geometry methods. J Mol Graphics Modell;15:18–36.
17. Agrafiotis D.K., Xu H. (2002) A self-organizing principle for learning non-linear manifolds. Proc Natl Acad Sci USA;99:15869–15872.
18. Agrafiotis D.K. (2003) Stochastic proximity embedding. J Comput Chem;24:1215–1221.
19. Xu H., Izrailev S., Agrafiotis D.K. (2003) Conformational sampling by self-organization. J Chem Info Comput Sci;43:1186–1191.
20. Izrailev S., Zhu F., Agrafiotis D.K. (2006) A distance geometry heuristic for expanding the range of geometries sampled during conformational search. J Comput Chem;27:1962–1969.
21. Agrafiotis D.K., Gibbs A., Zhu F., Izrailev S., Martin E. 2007) Conformational sampling of bioactive molecules: a comparative study. J Chem Info Model;47:1067–1086.
22. Carta G., Onnis V., Knox A.J. S., Fayne D., Lloyd D.G. (2006) Permuting input for more effective sampling of 3D conformer space. J Comput Aided Mol Des;20:179–190.
23. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. Nucleic Acids Res;28:235–242.
24. Berman H.M., Bhat T.N., Bourne P.E., Feng Z., Gilliland G., Weissig H., Westbrook J. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol;7(Suppl):957–59.
25. Daylight Chemical Informations Systems Inc. (http://www.daylight.com).
26. Halgren T.A. (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem;17:490–519.
27. Halgren T.A. (1996) Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem;17:520–552.
28. Halgren T.A. (1996) Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem;17:553–586.
29. Halgren T.A., Nachbar R.B. (1996) Merck molecular force field. IV. conformational energies and geometries for MMFF94. J Comput Chem;17:587–615.
30. Halgren T.A. (1996) Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem;17:616–641.
31. Agrafiotis D.K., Lobanov V.S., Salemme F.R. (2002) Combinatorial informatics in the post-genomics era. Nat Rev Drug Discov;1:337–346.
32. Agrafiotis D.K., Alex S., Dai H., Derkinderen A., Farnum M., Gates P., Izrailev S. *et al.* Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. Submitted.

Notes

a. CORINA 3.6, distributed by Molecular Networks GmbH (http://www.mol-net.de).

b. OMEGA 1.8.1, distributed by Openeye Scientific Software (http://www.eyesopen.com).

c. Catalyst v4.9.1, distributed by Accelrys, Inc. (http://www.accelrys.com).

d. RUBICON, distributed by Daylight Chemical Information Systems Inc. (http://www.daylight.com).

e. http://www.ccl.net/cca/data/MMFF94s/