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.

Rational discovery of second generation anti-cancer ligands

A thesis submitted to the

University of Dublin

for Doctorate of Philosophy.

Presented by

Georgia Golfis, MSc.

April 2008

Based on research carried out under the supervision of

Dr. David G. Lloyd

at the School of Biochemistry and Immunology,

Trinity College Dublin.

TRINITY COLLEGE
0 6 NOV 2008
LIBRARY DUBLIN

THESIS.
8645

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: G. Golfis

DECLARATION

This Thesis has not been submitted as an exercise for a degree at any other University. The work described is entirely my own work except where duly acknowledged.



Georgia Golfis

Στην οικογένειά μου, τους γονείς μου και

τον ξαδερφό μου*

* To my lovely family, my parents and my cousin.

Abstract

Recently, a novel series of Pyrrolo-1,5-BenzOXazepine (PBOX) compounds have shown apoptotic activity upon numerous cancer cell lines. Importantly, they have been suggested as potential antineoplastic agents in particular for the treatment of Leukaemia chemotherapeutic resistant. This thesis combines *in silico* and *in vitro* screening technologies for the discovery and mechanistic elucidation of a second generation of PBOX compounds. It is wished to scaffold hop away from the original chemotype due to an adverse *in vitro* profile. Cancer medicinal chemistry space is examined to focus the virtual screens towards relevant regions of chemical space and enhanced the success of chemotype switching. To overcome the inherent difficulties of working with active analogues set of actives, this work also aims to rationalise the PBOX tubulin interaction so as to guide biochemical research on the understanding and description of PBOX anti-cancer mechanism of action. Finally, a High Content Screening (HCS) platform was employed to determine a Structure Activity Relationship (SAR) for this group of compounds, validate the *in silico* work and assess the bio-activities of novel molecules based on the PBOX pharmacophore..

Acknowledgements

I would like to thank my supervisor Dr. David G. Lloyd and also our several collaborators: Dr. Danny Zisterer for her advice in the biochemistry field with a particular thank to Sandra Bright who did a lot of *in vitro* testing, Dr Lisa Greene and Dr. Seema Nathwani who familiarized us with cell culture. Thanks to Professor Dr. Clive Williams for his help on biodata interpretation and Susan Cloonan and Trevor Price for being willing to relinquish possession of a never-ending list of items. Also, a sincere thank to all our collaborators, Dr. Mark Lawler and Dr. Tony McElligott from the department of haematology who supported and facilitated my experimental work in St. James Hospital and Dr. Anthony Davies, Dr. Yuri Volkov , Dr Bashir Mohamed and Dr. Yvonne Williams who all helped me in understanding and using the brand new HCS technologies.

Especially thanks to Dr. Tim Mantle who gave me the opportunity to discover this fabulous country: Ireland. Thanks to his entire group Dr. Seamus Browne, Dr. Edward Franklin, Jerrard Haye, Dr. Tania O'Connor, Dr. Vincent P. Kelly for all the "craic" and a particular thank to Dr. Gavin McManus who helped me so much during my first steps in biochemistry and became a close friend. And thanks to our lovely secretary, so efficient, what a super-woman, thank you Mary Path for all your help.

Also thanks to the numerous members of the school of immunology and biochemistry in particular; Dr. H. Paul Voorheis for his precious advice, guidance and infectious passion for science. Thank to all his team: Robert O'Byrne, Dr. Orla Hanrahan and Conor Breen respectively for the fun, the sensual confocal teaching and the competitive spirit in body building.

Thanks to Dr. Mary J. Meegan from the pharmacy department team for giving me many collaborative opportunities and Dr. Miriam Carr for her joie de vivre!

Thanks to the structural group including Dr. Rosario Recacha-Castro for her Spanish teaching during the coffee breaks and Dr. Raymond Chambers for not robbing me of all my money at our poker games and of course their scientific support.

I would also like to thank all the MDG group for their help and support in numerous fields from computational to experimental and finally to the libational. Also to all the post-docs for their pertinent and strategic corrections: Dr. Natalia Artemenko, Dr. Darren Fayne, Dr. Timothy James, Dr. Gemma Kinsella, Dr. Martin Peters, Dr. Giorgio Carta, Dr. Paul McKay, Dr. Daniel Nevin, Dr. Valeria Onnis, Dr. Yidong Yang, Dr. Nick Jagoe. Thank you all very much. Especially thanks to Dr. Andrew Knox to whom I owe so much, I am so grateful for everything.

Thanks to my Greek team Dr. Constantinos Brikos, Dr. Alexis Laios and Dr. Dimitri for all the fun, food and care. Thanks to all my friends from France and Greece for supporting me during this endless marathon. Merci à vous tous pour tout le soutien moral, sentimental, et pour tous ces rires. Je vous remercie tous de tout mon coeur. Merci Sarah pour m'avoir fait confiance et suivi dans cette fantastique aventure et Tom bien entendu! Merci Aurélien pour …que dire…. absolument tout, toi même tu sais!

Ευχαριστώ την οικογένειά μου για την αγάπη τους, την υποστήριξή τους και όλη την προσπάθεια τους να με βοηθήσουν με οποιδήποτε τρόπω μπορούν.

Thanks to the Health Research Board for their financial contributions to this project.

Abbreviations

siRNAs short-interfering Ribonucleic acids
cDNA Deoxyribonucleic acid clones
CML Chronic Myeloid Leukaemia
CNS Central Nervous System
CV covariance
DHFR Dihydrofolate Reductase
DMSO Dimethyl sulfoxide
DNA Deoxyribonucleic acid
E Enrichment
ER Estrogen Receptor
FACS : Fluorescence Activated Cell Sorting
FDA Food and Drug Administration
FN False Negatives
FP False Positive
FRED Fast Rigid Exhaustive Docking
GPCRs G protein-coupled receptors
HCA: High Content Analysis
HCS High Contentt Screening
HIV-1 IN HIV 1 integrase
HTS High Throughput Screening
KSR Kinetic Scan Reader
MDG Molecular Design Group
MOE Molecular Operating Environment
MT: Microtubule
MTAs Microtubule Targeting Agents
NCI National Cancer Institute
PARP poly(ADP-ribose)
PBOX Pyrrolo-1,5-benzoxazepine analogues
PBR Benzodiazepine Receptor
PCA Principal Component Analysis
PDB Protein Data Bank
PDGFR Platelet-Derived Growth Factors Receptor
RMSD Root Mean Square Deviation
RNA Ribonucleic acid
RNAi Ribonucleic acid interference
RO3 Rule Of Three
RO5 Rule Of Five
ROC Receiver Operating Characteristic
ROCS Rapid Overlay of Chemical Structures
SAR Structure Activity Relationship
SARM Selective Androgen Receptor Modulators

SBVS Structure Based Virtual Screening
SD Standard Deviation
Se Sensitivity
SE Standard Error
SERM Selective Estrogen Receptor Modulators
SMILES Simplified Molecular Input Line Entry System
Sp Specificity
TN True Negatives
TP True Positives
vdW van der Waals
vHTS Virtual High Throughput Screening
VS Virtual Screening
VSA van der Waals Surface Area

Table of contents

Abstract I
Acknowledgements II
Abbreviations III
Table of contents V
Table of figures VIII
Table of tables X

Chapter I

I. General Introduction 1

1.1 Chronic Myeloid Leukaemia 2
1.2 Pyrrolobenzoxepines 3
1.3 Drug Discovery Overview 5
1.4 Virtual Screening approach 6
1.4.1 Virtual Screening methods overview 7
1.4.2 Similarity Methods 7
1.4.3 Three Dimensional Representation 8
1.4.4 Ligand Based Methods 9
1.4.5 Receptor Based Methods 11
1.4.6 Other Methods 11
1.4.7 Evaluation metrics 13
1.5 HTS/HCS approach 15
1.6 Objectives 17
References 19

Chapter II

2. Oncology exploration 23

Introduction 24
2.1 Cancer drugs 24
2.2 Managing chemical space for virtual screening 30
2.3 Cancer Medicinal Chemistry Space 32
2.4 Methods 33
2.4.1 Calculating descriptors 34
2.4.2 Filtering 34
2.4.3 Mapping Methods 34
2.5 Projection results 36
2.5.1 Drug like space vs. cancer medicinal space 36
2.5.2 Focus on target based spaces 38
2.5.3 Reasons for rejections 40
2.6 Filtering PBOX compounds 41
2.7 Conclusion 43
References 44

Chapter III

3. Virtual Screening protocol for PBOX scaffold hopping 23

Introduction 47
3.1 Virtual screening strategy breakdown 48
3.1.1 Training sets structures selection 49
3.1.2 Haystack generation 52
3.1.3 Pre-processing compounds for VS 54
3.1.4 VS1- the shape approach 54
3.1.5 VS2- Refined pharmacophore generation 55
3.2 Computational methods 56
3.2.1 Descriptors calculation 56
3.2.2 Evaluation metrics 58
3.2.3 Principal Component Analysis 58
3.2.4 Fingerprint calculations 60
3.2.5 Bayesian model 60
3.2.6 Molecular clustering 61
3.2.7 Ionisation and isomerisation 61
3.2.8 Conformer generation 62
3.2.9 ROCS and EON 63
3.2.10 HipHop and catShape 65
3.2.11 Catalyst HypoGen and HypoRefine 67
3.2.12 Pharmacophore clustering 69
3.2.13 Database searching with Catalyst pharmacophores 70
3.3 Results 70
3.3.1 Haystack elaboration 70
3.3.1.1 Filter elaboration 71
3.3.1.2 Principal Component Analysis 72
.3.3.1.3 Bayesian model 74
3.3.1.4 Haystack validation 75
3.3.2 3D-pharmacophore 82
3.3.2.1 Pre-processing 82
3.3.2.2 In silico validation of 3D shape models 84
3.3.2.3 Pharmacophore refinement 96
3.4 Conclusion 102
References 103

Chapter IV

4. Investigation of the PBOX binding site in tubulin 105

Introduction 106
4.1. Microtubule structure and dynamics 108
4.2 Tubulin sequences 111
4.2.1 Role of Tubulin Isotypes 111
4.2.2 Homology and origins 112
4.3 Tubulin 3D structure and interactions 112
4.3.1 Crystallographic data 113
4.3.2 Domain descriptions 115
4.3.3. Tubulin-GTP interactions 117
4.3.4 Tubulin-tubulin interactions 119

4.3.5 Interaction with other proteins 121
4.3.6 Tubulin Binding Sites and Ligands 125
4.4 On target destabiliser ligands 125
4.4.1 The colchicine binding site 128
4.4.2 The vinblastine site 131
4.4.3 Non localised binding sites destabilisers 135
4.5 On target stabilisers ligands 136
4.5.1 The taxol binding site 138
4.5.2 Non localised binding sites stabilisers 141
4.6 Mapping tubulin sites. 142
4.6.1 PBOX binding site discovery strategy 143
4.6.2 Computational methods 147
4.6.3 Results 152
4.7 Conclusion 174
References 175

Chapter V

5. Coupling High Content Screening to Virtual Screening 185

Introduction 186
5.1 Biological background 187
5.1.1. The cell cycle 187
5.1.2 Apoptosis 189
5.1.3 High-throughput fluorescence microscopy 190
5.2 Strategy for screening of PBOX and MDG compounds 191
5.3 Material and methods 195
5.3.1 Materials 195
5.3.2. HCS theory 196
5.3.3. Cell cycle HCA experimental procedure 204
5.3.4. Screening of compounds for effects on the microtubule network 208
5.3.5 IC 50 value determination 209
5.3.6 Tubulin polymerisation assay 210
5.3.7. Cell culture 211
5.3.8 Cell count 211
5.3.9 Preparation of drug solutions 212
5.5 Results 212
5.5.1 HCA cell cycle analysis validation 212
5.5.2 The effect of PBOX compounds on the cell cycle profile of MCF-7 cells 214
5.5.3 The effect of MDGs on the cell cycle profile of MCF-7 cells 220
5.5.4 HCA for tubulin effect detection and validation 224
5.5.5 Screening by HCA of MDG compounds for effect on the MT network 227
5.6. Conclusion 230
References 232

6. General conclusion 234

Appendix I: PBOX structures 235

Appendix II: MDG structures 238

Table of figures

Chapter I

Figure 1.1: Chemical structures of established CML chemotherapeutics. 3
Figure 1.2: Pyrrolo-1,5-BenzOXazepines (PBOX) general structure. 3
Figure 1.3:Structures of diverse PBR ligands 4
Figure 1.4: Example of ligand based virtual screening workflow 10
Figure 1.5: Theoretical distributions for active molecules and decoys according to their score ... 14
Figure 1.6: Example of ROC curve. 15

Chapter II

Figure 2.1: Chemical structures of non-selective chemotherapy drugs acting through different mechanisms. 26
Figure 2.2 Chemical structures of small molecule cancer target selective inhibitors. 27
Figure 2.3: Example of antimitotic, hormonal, SERM, SARM and aromatase targeting ligands. 30
Figure 2.4: Charted Cancer Medicinal Chemistry Space 37
Figure2.5 Anticancer kinase-targeted space 38
Figure2. 6: Charting cancer medicinal chemistry space. 39
Figure 2.7 RO5 descriptors distribution 40
Figure 2.8: PBOX RO5 descriptor distributions 42

Chapter III

Figure 3.1: Ligand-based workflow scheme 48
Figure 3.2: Pyrrolo-1,5-BenzOXazepines (PBOX) general structure. 51
Figure 3.3: Correlation between the estimated activity and the calculated geometric Fit 66
Figure3.4: Description of automated pharmacophore generation runs 67
Figure 3.5: PBOX van der Waals surface area and polar surface area distributions 71
Figure 3.6: PBOX projection onto the two first PCA axes 72
Figure 3.7: ROC curve for the Bayesian model 74
Figure 3.8: NCI compound classification according to the different models and total number of compounds selected for each method 76
Figure 3.9: Classification of WOMBAT molecules according to the different 77
Figure 3.10: Conformation generation trials on PBOX compounds 83
Figure 3.11: ROCS method determination 84
Figure 3.12: ROCS and EON effect upon the haystack 87
Figure 3.13: PBOX structures from the training set 88
Figure 3.14: Shape addition evaluation upon HipHop best pharmacophore. 92
Figure 3.15: Effect of HipHop pharmacophore on MDL fingerprints distribution. 93
Figure 3.16: HypoRefine exclusion volume distribution along all pharmacophores generated 97
Figure 3.17: (a) Pharmacophore correlation to the training set, (b) percent of actives retrieved .. 98
Figure 3.18: (a) Best Fit of PBOX 16 (b) HypoRefine vs.2D descriptor. 99
Figure 3.19: HypoRefine and ROCS/EON hits molecular distribution comparison 100
Figure 3.20: New Maybridge hit structures 100
Figure 3.21: Effects of hits on Leukaemia cell lines, ... 101
Figure 3.22: Alignment PBOX6, PBOX16 101

Chapter IV

Figure 4.1: Phases of Mitosis. 106
Figure 4.2: Computational publication on tubulin specific site binder since 1998-2008 107
Figure 4.3: Microtubule protofilament 108
Figure 4.4: Microtubule stable phase 109
Figure 4.5: Microtubules unstable phases 110
Figure 4.6: Microtubule rings formation 110
Figure 4.7: Tubulin heterodimer from the crystal structure 116
Figure 4.8: Curved and straight tubulin conformations 117
Figure 4.9: (a) GDP in the E site and (b) GTP in the N site. 119
Figure 4.10 The lateral interaction between protofilaments 120
Figure 4.11: MTs and main MAPs interactions 122
Figure 4.12: Structural diversity of tubulin destabiliser compounds. 127
Figure 4.13: Colchicine binding site 129
Figure 4.14: Vinblastine binding site 132
Figure 4.15: Structural diversity of tubulin stabilisers. 138
Figure 4.16 Taxol binding site 139
Figure 4.17: Binding site search - strategy break down 143
Figure 4.18: Known binding sites found by Site Finder 153
Figure 4.19: Site Finder mapping of the nucleotic sites. 154
Figure 4.20: Site Finder mapping of vinblastine site 155
Figure 4.21: Site Finder mapping of taxol site. 156
Figure 4.22: Site Finder mapping of colchicine site 157
Figure 4.23: Site volume distributions for the curved and straight conformation. 158
Figure 4. 24: Tubulin destabilisers and stabilisers spaces 159
Figure 4.25: Known drugs sites spaces 160
Figure 4.26: WOMBAT tubulin binders distributions 162
Figure 4.27: PBOX compound space. 164
Figure 4.28: Known drugs binding site selection after filtering 166
Figure 4.29: PBOX compounds binding site selection after filtering 167
Figure 4.30: PBOX structures docked. 168
Figure 4.31: Optimised binding pose of PBOX16 in the E site 169
Figure 4.32: Optimised binding pose of GDP in the E site and 170
Figure 4.33: PBOX ligand based pharmacophore superposed on the E site 171
Figure 4.34. Comparaison docked structures and ligand based phamacophore 173

Chapter V

Figure 5.1: The Phases of the Cell Cycle 188
Figure 5.2: Physical screen general workflow 194
Figure 5.3: Principle of labelled objects selection 197
Figure 5.4: (a) Bounding Box, (b) Convex hull 198
Figure 5. 5: Cell cycle healthy profiles 201
Figure 5.6: Nocodazole cell cycle profiles 214
Figure 5.7: PBOX6 effects on the cell cycle 215
Figure 5.8: Example of images from scanned cell nuclei. 218
Figure 5.9: Morphological nuclei changes. 219
Figure 5.10: Morphology of apoptotic nuclei. 220
Figure 5.11: MDG antiproliferative and G2 IC50 values 222
Figure 5.12: Active compounds tested on leukaemia cell lines 224
Figure 5.13: Controls images from KSR at x 20 magnificence. 225

Figure 5.14: Effect of MTA on cell parameters following tubulin staining 227
Figure 5.15: MDG Mean difference of tubulin stain intensity density. 227
Figure 5.16: MDG61 tubulin morphological profile 228
Figure 5.17: MDG41 tubulin morphological profile 228
Figure 5.18: MDG36 tubulin morphological profile 228
Figure 5.19: Tubulin polymerisation assay. 230

Table of tables

Chapter II

Table 2.1: Breakdown of cancer compound drug-like nature 37

Chapter III

Table 3.1: Summary of PBOX experimental results. 49
Table 3.2: Pharmacophore training sets. 50
Table 3.3: MOE 2D main descriptors calculated. 58
Table 3.4: Variation of the PCA model along the diverse datasets 59
Table 3.5: Bayesian scores, (a) PBOX compound scores (b) databases scores. 75
Table 3.6: Top target frequencies among compounds retrieved from WOMBAT by the various classification methods. 78
Table 3.7: Structures representative of the different target classes retrieved from WOMBAT by the Bayesian classification model. 80
Table 3.8: Haystack content 81
Table 3.9: Bayesian active compounds 82
Table 3.10: Preferred conformation chosen by the training set compounds 85
Table 3.11: Mapping of training set features; 89
Table 3.12: Pharmacophore clustering; 90
Table 3.13: ROC scores per pharmacophore 91
Table 3.14: Training set prediction 93
Table 3.15: Activity prediction by HipHop pharmacophore 5 of active compounds found from ROCS/EON screen 95
Table 3.16: PBOX structures of inactives present in the training set 96

Chapter IV

Table 4.1: .α/β Tubulin crystal structures 113
Table 4.2: Validation set diversity and size 148
Table 4.3: WOMBAT's tubulin binders projection. 163
Table 4.4: Co-crystallized ligand volume 165
Table4. 5: Chemgauss2 scores for all docked compounds. 168

Chapter V

Table 5.1: Materials 195
Table 5.2: Output features categories 199
Table 5. 3: Cellomics bioapplication cell cycle features 200
Table 5.4: Morphology explorer bioapplication output features 203
Table 5.5: PBOX active IC50 values. 216
Table 5.6: PBOX with showing intermediate activity. 216

Table 5.7: PBOX that did not show any activity. 217
Table 5.8: Stage 3, MDG compounds first filter. 221
Table 5.9: MDG IC50 values for G2/M and reduction in cell nuclei in μM. 223
Table 5.10: Calculated cytoskeleton morphological features 229

Chapter I

General introduction

1.1 Chronic Myeloid Leukaemia

Cancer of the blood or bone marrow more commonly known as Leukaemia is characterized by an abnormal proliferation of white blood cells (leukocytes) and is part of the broad group of diseases called hematological neoplasms. Chronic Myeloid Leukaemia (CML) is a form of Leukaemia where an increase and unregulated clonal production of predominantly myeloid cells occurs in the bone marrow. CML represents about 15-20% of all cases of adult Leukaemia in western populations. It affects all age groups but mostly middle-aged and elderly with men being slightly more predisposed. It occurs in 1 to 2 per 100,000 people per year[1]. CML is an overproduction of immature blood cells due to an abnormal increase of granulocytes in the blood. Hence, it increases the number of white blood cells and induces enlargement of the spleen. It is accompanied by thrombocytopenia and anaemia, with consequent symptoms of weakness, fatigue and abdominal discomfort due to splenic enlargement and can cause people to look pale. CML can be broken into three phases:

- Chronic: The cancer is stable and lasts a period of 3-4 years.
- Accelerated: The cancer grows more quickly and it lasts usually 5-6 months. There are between 15-30% of blast cells in the blood. (Blast cells can be defined as immature, undifferentiated cell, precursor of human blood cell).
- Blast: Immature cells have replaced the bone narrow and more than 30% of blast cells can be found.

In this myeloproliferative disease an abnormal chromosome known as the "Philadelphia" chromosome is observed in most patients affected (95%). The Philadelphia chromosome develops when part of chromosome 9 (the ABL gene) mistakenly attaches itself to chromosome 22 (the BCR gene) during cell division. These structural abnormalities create a new gene, known as BCR-ABL. Three different proteins from BCR-ABL fusion can exist depending on the breakpoint on the BCR fragment. Here, the chromosomal translocation leading to BCR-ABL, encodes for the fusion protein (p210)[2] and the produced specific enzyme (ABL proteins) is a non receptor tyrosine kinase. ABL proteins play an essential role in signal transduction and in the regulation of cell growth[3]. Deregulation and increase of tyrosine kinase activity blocks normal integrin function and consequently reduces adhesion of progenitor cells to stromal elements. Hence, stem cells escape physiological inhibitory regulation[4] while BCR-ABL inhibits apoptosis and leads to an accumulation of cells with associated symptomatic manifestation[5].

Historically, CML was treated with radiotherapy but for the last 50 years the principal mode of treatment has been palliative chemotherapy. Chemotherapy (*eg.* busulfan and hydroxyurea illustrated in Figure 1.1) and allogenic bone marrow transplantation[6-10] are the two alternatives available. Both remain unsatisfactory, with the condition-proving terminal for a majority of patients. Despite the arrival of number of putative chemotherapeutic agents, treatments are still

limited[11, 12]. Considerable research is being carried out in the genetics area with the goal of developing new gene therapies and also anti-sense nucleotide[13] treatments. Technological advancements in targeted therapies of the 21st century have radically boosted the management of CML. Lately, high-throughput biological screening at Novartis identified a class of 2-phenylaminopyrimidine kinase inhibitors (imatinib or gleevec or STI571 illustrated in Figure 1.1), which showed activity against CML cell lines[14]. However, recent reports on the emergence of imatinib resistance in patients[15] highlight the need for multiple alternate chemotherapeutic agents for CML[16] and in particular compounds active against multi-drug resistant instances of the disease.

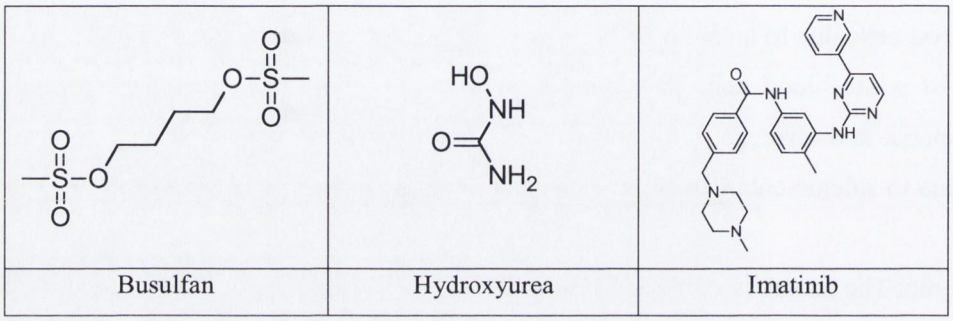


Figure 1.1: Chemical structures of established CML chemotherapeutics.

1.2 Pyrrolobenzoxepines

Recently, the potential use of a novel series of Pyrrolo-1,5-BenzOXazepines (PBOX) compounds has been suggested for the treatment of CML (*cf.* Figure 1.2). They have shown apoptotic activity upon numerous cell lines including CML that are resistant to many chemotherapeutic agents.

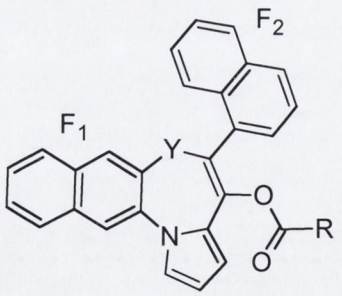
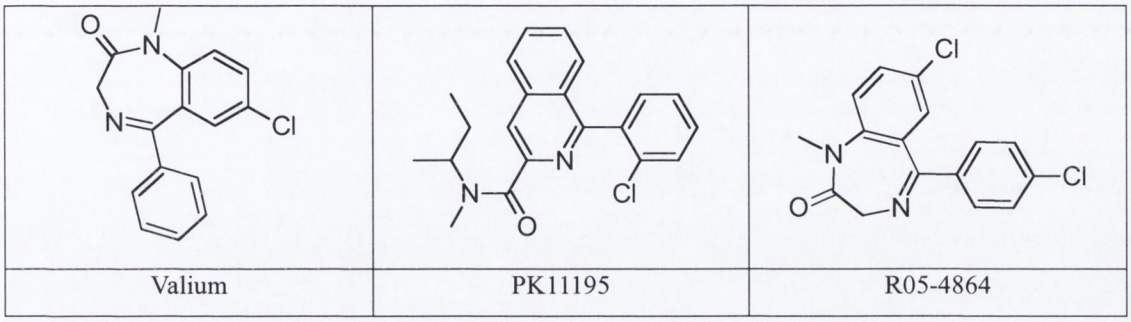


Figure 1.2: Pyrrolo-1,5-BenzOXazepines (PBOX) general structure.
Analogues presents modification of rings (F_1 and F_2), central atom Y and the side chain R.

At first PBOX compounds were synthesized[17] and developed for probing the Peripheral-type Benzodiazepine Receptor (PBR) present in peripheral tissues and in the Central Nervous System (CNS)[18]. These pyrrolobenzoxazepines were expected to induce similar pharmacological effects as seen with other drugs targeting the PBR[19] such as Valium (illustrated in Figure 1.3) which exhibits

anxiolytic, relaxant and sedative effects or the well known PBR high affinities ligands PK 11195 and R05-4864[20] (shown in Figure 1.3). Although, the PBR has been shown to be involved in many physiological functions and more lately in apoptosis[21], further studies showed that PBOX anti-proliferative effect did not correlate with binding affinity to the PBR binding site[22]. However, certain members of the PBOX series appeared to potently induce apoptosis in diverse cancer cell lines such as breast cancer cell lines and human chemotherapy-resistant cancer cell lines such as a number of CML cell lines such as K562, KYO.1, and LAMA 84.[23]



Valium	PK11195	R05-4864

Figure 1.3: Structures of diverse PBR ligands

PBOX induced apoptosis occurs by bypassing the apoptotic suppressor, BCR-ABL[23]. Further delineation of the mechanism of action of these compounds has indicated activation of the c-Jun N-terminal Kinase (JNK) signalling pathway as essential during ligand-induced apoptosis in the cell lines studied[24]. Extrapolations of the investigation to other human cancer cell lines demonstrated that caspase-3 was not essential for DNA fragmentation in MCF-7 or CML cells during apoptosis induced by the compounds, and initially implicated caspase-7 in the apoptotic mechanism of action[25]. However, subsequent mechanistic investigations have now eliminated the role of caspases. Cell cycle studies indicated that G1 arrest is preceded by both a decrease in CDK2 kinase activity, which is critical for the G1/S transition, and a down regulation in cyclin D(3) protein expression levels, suggesting that these two events may be crucially involved in the mediation of the cell cycle arrest[26].

Lately, further studies carried out to elucidate the mechanism of action of these compounds showed that a representative pro-apoptotic PBOX compound induced an accumulation of cells in G2/M phase preceding apoptosis. Thus, prometaphase arrest and the accumulation of cyclin B1 levels and activation of cyclin B1/CDK1 kinase were observed.

Furthermore, immunofluorescence showed microtubule depolymerization in MCF-7 cells following PBOX treatment. These last observations were very similar to the effects of nocodazole and paclitaxel. This suggested tubulin as the target of PBOX compounds according to the polymerization tubulin assay. Yet, competitive assays run against well known tubulin binders such

as colchicine and vinblastine did not show any replacement of the ligands suggesting that PBOX compounds bind to an as-yet uncharacterised novel binding site on tubulin[27].

Finally, PBOX anti-cancer pharmacological effect was also observed in primary *ex vivo* material derived from cancer patients. Hence, PBOX compounds could possibly be used for the treatment of both solid tumors and tumors derived from the haematopoietic system. All this suggests PBOX compounds as novel anticancer therapies in the treatment of both solid tumors and hematological neoplasias and as novel candidates for antineoplastic therapy.

While there have been many exciting developments in the treatment of CML as we start to understand more about the disease, unfortunately the only treatment that promises hope of a cure at present is a bone marrow transplant, which is not possible for everyone. Deconvolution of the apoptotic mechanistic pathway(s) of PBOX compounds may serve to aid in the development of compounds that might add to the current therapeutic arsenal.

The use of computational software in drug discovery is now growing in popularity; validation cases and reviews have recently been published highlighting its utility in the discovery process, and much effort has been invested in the formulation of *in silico* screening routines to complement traditional *in vitro* screening.

1.3 Drug Discovery Overview

Nowadays, both pharmaceutical companies and Research and Development (R&D) centres, adapt the drug discovery process in multiple ways according to their respective constraints, needs and aims. Obviously, prior to the elaboration of such a procedure, the determination of the biochemistry underlying a particular disease is essential. Usually years of scientific investigations are necessary for the identification of a particular target protein where activity modulation is desirable. Then, the aspiration common to all is to find new molecules capable of interacting with the particular receptor. The variable pipelines leading to their discovery can always be formulated as a succession of more general phases. Once a biological assay has been developed to confirm or reject candidates are binding, a myriad of molecules are *in vitro* screened. Nowadays High Throughput Screening (HTS) technology can assess the bioactivity for several thousand molecules per day. Successful molecules denoted as "lead compounds" will be then proceed through iterative molecular modelling and biological testing to optimise their usually weak native potencies. Finally, the selected drug candidate must pass rigorous pre-clinical testing as well as detailed clinical trials. Often pre-clinical studies involve *in vitro* and *in vivo* studies on animals to obtain preliminary efficacy, toxicity and pharmacokinetic information. Subsequently, Phase 0 consists of human microdosing studies on 10 to 15 patients and to gather preliminary data on the agent's pharmacokinetics and pharmacodynamics. Phase I determines the safety, tolerability, pharmacokinetics, pharmacodynamics and dose range of a drug, on inpatients (full time followed

patients). Phase II applies to larger group of patients up to 300 and consists of confirming and refining the required dosing (Phase IIA) and the drug efficacy (Phase IIB).

Phase III is the most expensive and time-consuming phases employing up to 3000 patients to compare the actual efficiency of the new candidate to the standard marketed drug. Finally, Phase IV also known as the Post Marketing Surveillance Trial, which depends on regulatory authorities, the new drug is on the market and pharmacovigilance tracks safety issues such as potential multi-drugs interactions.

Maximizing the success rate of hits thus minimizing the number of candidates subject to the expensive and long clinical trails has become a general preoccupation as HTS technology allows the screening of a constantly increasing number of chemical compounds libraries. In order to reduce R&D costs and the time taken between idea and identification, of the field of Virtual Screening (VS) has attracted much attention.

1.4 Virtual Screening approach

The main objective of VS or virtual High Throughput Screening (vHTS) is to improve the drug discovery process and the quest for finding new leads[28]. It takes place in the iterative molecular modelling based optimization phase prior to the diverse (pre) clinical trials assessments. Indeed, it involves the rapid assessment of large libraries of chemical structures in order to guide the selection of likely drug candidates. Typically, only one drug will come out after lead optimization if a library of the order of 1,000,000 compounds is screened.

Several successful applications have been reported since the development of VS in ligand based and structure based methods,[29] and with the advances made in structural biology have led to the elucidation of a plethora of protein structures giving invaluable information to drug designers. As promised, HTS coupled to VS have resulted in an increase of hit retrieval[30] and decreased both cost and time problems.

VS technologies as known today first appeared in 1997 and had their roots in the enhancing computational chemistry field due to the recent explosion in informatics' programs. Nevertheless, *in silico* technology is not yet mature enough to accurately describe complex molecular interactions. Some fundamental parameters used in VS such as entropy and the dielectric constant are not defined in a unique way and are still being debated. These are just two examples from many and one of the implications of those remaining problems without a clear answer is the appearance of diverse methods, algorithm and consequently software.

1.4.1 Virtual Screening methods overview

To be able to predict bioactivity, the parameters related to protein-ligand complex action have to be elucidated. This involves the prediction of the conformation of the ligand bound to the receptor. The conformational space being infinite, guiding lines or rules of reactivity and acceptance are needed therefore a bioactive binding mode has to be described by accessible conformational states for both ligand and receptor. The ligand should be selective and hence have a high degree of efficacy. This describes the complexity of the quest of finding the bio-active conformation. With regards to this, many analytical methods are being developed such as, ligand based VS and receptor based VS with the main objective of improving the drug discovery process and the quest for finding new leads[28]. It takes place in the iterative molecular modelling based optimization phase prior to the diverse clinical trials assessments. It involves the rapid assessment of large libraries of chemical structures in order to guide the selection of likely drug candidates.

Recently, protein Nuclear Magnetic Resonance (NMR) also actively contributes in the delivery of 3D structures whose X-Ray crystallographic resolution failed. Since the human genome publication a considerable number of targets have been submitted to homology modelling for the prediction of their three dimensional structures opening more possibilities to structure based virtual screening.[31]. VS technologies has its roots in computational chemistry. Computational chemistry, using theoretical chemistry concepts, aims to describe experimentally measurable quantities and to predict measurable and immeasurable properties[32]. This field is in constant development[33]. Each class of tool is generally dedicated to a particular mission to reduce the chemical space to a short selection of molecules ranked according to their likelihood in being "hits". An example of virtual screening protocol can be depicted as firstly the reduction of the chemical space to a drug like compounds space using molecular descriptor based methods. Consecutively three dimensional methods can either filter out unlikely "hits" on the bases of their three dimensional descriptions and/or only retain compounds that fit the receptor. Finally, for instance, a binding energy with solvation and accurate charges evaluation can select and rank the initial hit list.

1.4.2 Similarity Methods

Similarity methods were first developed for substructure searching to retrieve molecules from databases containing a predefined query structures. The definition on how two entities are "similar" can vary considerably. For instance two molecules can have similar solubilities or they can be similar from a structural point of view such as having common chemical features. Such similarity properties are often called descriptors. Descriptors are parameters that can distinguish molecules from one another. These descriptors can roughly be classified into 1D, 2D and 3D based upon how the molecules are stored. For example, calculable electronic properties such as the highest occupied

molecular orbital require the 3D structure of a molecule whereas a substructure property such as ring count can be calculated from 1D information.

Descriptor space can be described as the ensemble of molecular properties and fingerprints encoding for ligand scaffolds. This space is becoming more and more populated-proportional to the elaboration of new property reflecting descriptors and the elaboration of new fingerprints codes (*eg* Extended Connectivity Fingerprint developed by Pipeline Pilot[34] and explained in Chapter 3). An analysis carried out by Bender *et al.* [35] illustrated this constant increase, mirrored in the number of publications treating of molecular similarity as main subject. In the twelve years 1981 to 2003 such output was multiplied by a factor of 80. The most interested users appeared to be pharmaceutical companies[36] in applications such as virtual screening, estimation of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME/Tox) and prediction of physicochemical properties (solubility, etc.). In the context of lead like libraries where lead like compounds are exhaustively enumerated, similarity methodology can be used the reduce the total number of compound to be screened according to particular properties, aiming to retrieve new molecules with enhanced potency stripped of inconvenient features[37].

Inherent to similarity is the association of its measurement. Multiple coefficients are currently used to translate similarity "distances" between molecules among which the most common is the Tanimoto coefficient. While traditionally similarity was measured between two molecules, it has been shown by several studies that the consideration of the occurrence of a feature within a large ensemble of molecules was more efficient in similarity searching[38]. This more probabilistic approach is used by in the Bayesian method discussed in Chapter 3.

However, to be able to predict bioactivity, most often the protein-ligand complex structure has to be elucidated. This involves the prediction of the conformation of the ligand bound to the receptor. The conformational space being infinite, guiding lines or rules of reactivity and acceptance are needed therefore a bioactive binding mode has to be described by accessible conformational states for both ligand and receptor. The ligand should be selective and hence have a high degree of efficacy. With regards to this, many three dimensional analytical methods are being developed roughly classified in two classes: Ligand Based VS (LBVS) and Receptor Based (RBVS).

1.4.3 Three Dimensional Representation

The flexibility of ligand molecules can be accounted for through the generation of multitude of conformers prior to the VS screening method. Conformation search techniques aim to intelligently sample and reduce the huge conformational space into a more manageable subset keeping storage and subsequent use by screening methods in mind. Searching methods usually employ the Root Mean Squared Distance (RMSD) between two conformers as a cut-off value to decide to retain or not the new conformer generated. Methods can roughly be classified in two categories. Explicit conformational storage promotes conformational variation by the addition of a poling function to a

standard molecular mechanic equation. The function affects the energy surface being minimized to penalize conformational space around any previously retained conformer. Catalyst[39] software from Accelrys falls in this first category. The second category is much faster where torsion look up tables are used in the building phase and van der Waals clash checks remove unrealistic conformations. OMEGA2[40] from Openeye falls in the second category. The catBEST and catFAST algorithms from Catalyst and OMEGA2 will be described in Chapter 3. There are a multitude of conformer generators all presenting several differences[41]. The choice of the adequate tool lies in the subsequent use of the conformational database. Generally, either the total number or the conformer energies are the directing factors toward one method or another. For instance, it has been shown by Knox *et al.*[42] that the use of 10 conformers was sufficient for rigid docking with ERα while to generate a 3D predictive model based on a small set of active molecules it is suggested to use up to 250 conformers[43].

1.4.4 Ligand Based Methods

Ligand based approaches for virtual screening begins by developing a model describing a set of active compounds. When the model is built in a three dimensional context, the model is refer to as a pharmacophore. "A pharmacophore is a specific, three dimensional map of biological properties common to all active conformations of a set of ligands which exhibit a particular activity. Conceptually, a pharmacophore is a distillation of the functional attributes of ligands which accomplish a specific task"[44]. In other words, pharmacophores are a representation of the chemistry necessary for the ligand to bind its receptor and include features that carry biological data.

The elaboration of a pharmacophore, as for structure based approaches, requires a 3D molecular representation. A pharmacophore can be used to select compounds based on 3D similarity and to predict and estimate an actual bioactivity value. The latter falls in what is generally referred to as a 3D Quantitative Structure Activity Relationship (3D QSAR). Other QSAR methods will be explained in the next section. An example of the general workflow of a LBVS is illustrated in Figure 1.4.

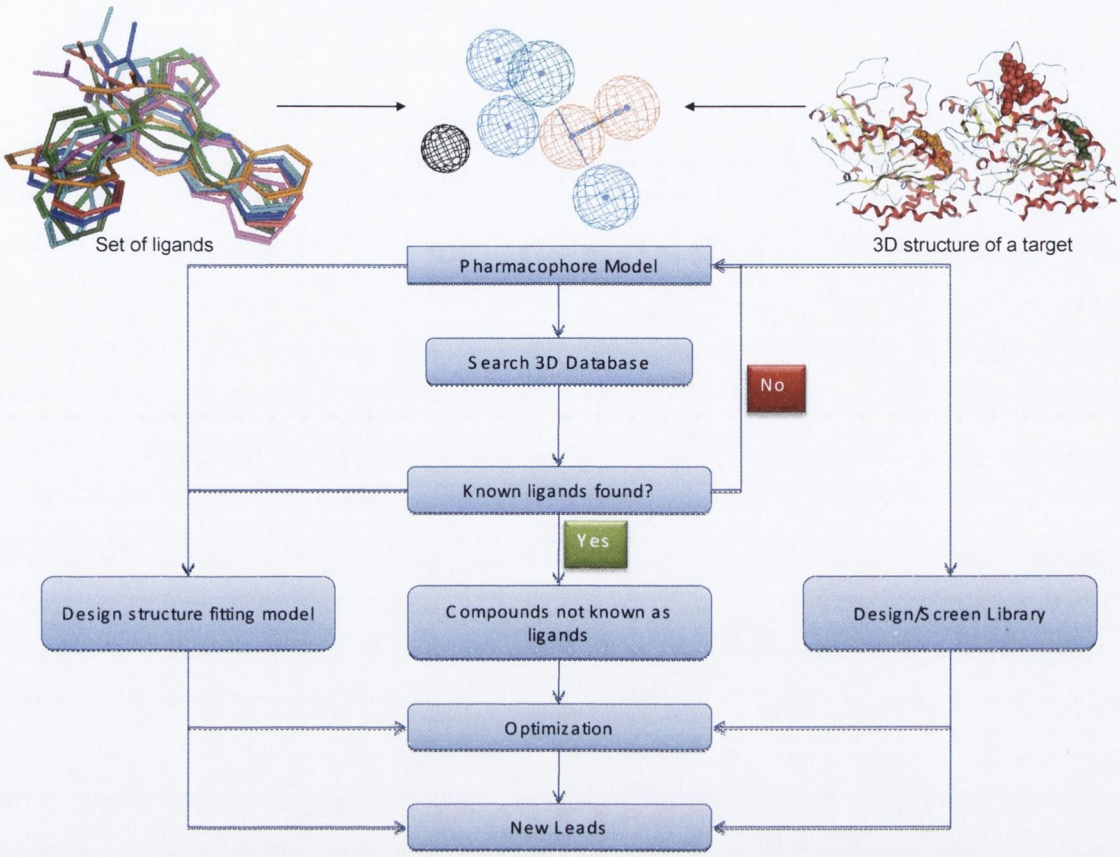


Figure 1.4: Example of ligand based virtual screening workflow *(adapted from T. Langer)*

In general, the model is built using a set of active ligands (some tools can use information described by verified inactive ligands). This set of ligand is called a training set. In addition, information from the receptor when available can complete the model. As a consequence, the model accuracy in predicting activity is dependent on both the initial molecular representation of the training set and the bioactivity data when the latter is used in the building process. Traditionally, the initial set of known activity structures is split into three: one third to compose the training set and for two thirds to compose the testing set. The assessment of the quality of the model can be tested in various ways. The testing set can obviously be employed to test the model in estimating known ligand activities. Another possibility would be to seed a small dataset of inactives or decoys, informally called a "haystack", and test the model in retrieving the hidden actives (needles) and the rates of inactives selected. Measurement of the model quality can be assessed in several ways that will be discussed in detail later. The model is submitted to optimisation iterations until a reasonable quality is reached. Once the model is validated (*in silico)* databases of compounds can be screened and hits selected according to their model fit.

The fitting can be expressed also in several ways among which the traditional Tanimoto coefficient reflecting the overlap of the model and the molecule fitted, a Root Mean Square Deviation (RMSD) reflecting the deviation from the model or the correlation of the predicted activity with the

experimental one. Each fitting method used in this work will be explained in the method section of the following Chapters.

1.4.5 Receptor Based Methods

When the 3D structure of the target protein is known a receptor based screening or docking protocol can be applied. The final aim is to fit each molecule into the protein binding site and output a ranked list of the input molecules which prioritises molecules according to binding affinity[33, 36]. Independent of the choice of the method the general docking process can be split into two successive steps. The first step consists of fitting each conformer of each molecule or each molecule into the binding site and scoring them. Theoretically, all possible poses should be generated to ensure fairness between all molecules. As one of the main considerations in VS is time and thus computational speed is of the essence. Consequently algorithms are used to restrict the number of poses. Different methods can be applied depending to the tool choice and they can differ in their docking algorithms. The second step consists of a scoring function assessing the fit of each pose generated. Most scoring functions estimate the free energy of binding for a receptor-ligand complex in aqueous solution. The *in silico* validation of a docking protocol is carried out in two phases. Firstly, the ability of a docking function to retrieve the actual binding mode of a known crystallized ligand. Secondly, the scoring function should be able to attribute appropriate scores to each docked ligand permitting the retrieve of known active ligands hidden in a haystack.

Docking tools have been used successfully[31] in drug discovery but many limitations remain inherent to the scoring function method used or to the quality of the available 3D protein structures. In some cases, ligand based tools can be used in conjunction with docking which then is named guided docking[45] or used as an alternative when no receptor structure is available

1.4.6 Other Methods

The previously described pharmacophore model is part of a large family of methods denoted QSAR methods. QSAR is directly related to the biological activity and based on the first assumption that similar structures of ligands will posses similar bioactivities for similar receptors[46]. QSAR is largely used to increase the relevance of drug design by eliminating as soon as possible unfavourable physicochemical properties and pharmacokinetic profiles leading to an increasing integration of *in silico* technologies that sift through enormous numbers of virtual chemicals based on 'soft modelling' QSAR techniques[47]. It tries to estimate properties deemed important for orally available drugs by incorporating for instance structure-permeation, structure-distribution, structure-metabolism, and structure-toxicity relations into drug-design strategies. QSAR can be applied at three different levels:

- 2D QSAR method first purpose was to investigate binding affinity and activity[48]. It uses calculable 1D descriptors such as the molecular weight of a molecule and measurable parameters such as lipophilicity and acidity.
- 3D-QSAR methods incorporate 3D descriptors from 3D structures from X-rays such as Molecular Shape Analysis [49], steric parameters, electrostatic parameters and so on. 3D-QSAR can be split into two sub-fields which are independent and dependent on the receptor.
- *n*D-QSAR is then 3D-QSAR plus the integration of other unique physical parameters. Examples of this method are 4D-QSAR using an ensemble of molecular conformation and EigenValue-QSAR Analysis (EVA-QSAR) using infrared spectra.

Each QSAR package available on the market can be characterized by having a particular method of calculating descriptors, a procedure for selecting descriptors (evolutionary[50] and genetic[51]) and a statistical algorithm for constructing the model. The QSAR model can be constructed/analysed using multivariable analysis[52, 53] including Principal Component Analysis (PCA)[54, 55] which will be discussed in Chapter 2 and 3, Principal Component Regression (PCR)[56], Partial Least Squares (PLS)[57, 58], Multiple Linear Regression (MLR) or even Artificial Neural Networks (ANN)[59]. Nowadays, there is still, of course, no perfect package and one must try various algorithmic functions and analytical methods available in order to build ones own screening protocol and platform.

VS protocols are usually directed toward diverse QSAR method as an alternative to SBVS when no 3D structure of the receptor is available. In opposition, when receptor information is at hand and ligands bioactivity determined, both approaches can be employed in conjunction taking advantage of all the available knowledge. Both methods can be therefore coupled sequentially or in parallel. Most of the times the first combination consists in filtering out compounds not likely to be "hits" on a pharmacophore basis. An example of the second, termed "Tiered" VS evaluates the final VS rank on the consensus of the two methods[60].

Usually, docking methods finds their applications in VS for the identification of novel drug candidates. However, docking capabilities can go beyond the traditional use and guide experimental research for the discovery of novel targets. Several publications have been treating the subject under the name of "blind docking" or "inverse docking"[61, 62]. For instance, in the context of ADMETox evaluation, the identification of novel potential target responsible for toxicity would give supplementary valuable information. Lead compounds could then be optimized to remove the binding to the additional target. In this situation the question asked is to which target could fit a particular ligand instead of which ligands would fit a particular target. In general, databases of target known to be involved in toxicity, were extracted from available ligand-protein complex crystal structures and the target final rank was based on the calculation of ligand protein interaction of hydrogen bond and non-bonded interactions only and compared to the actual ligand co crystallized to the corresponding structure[63].

1.4.7 Evaluation metrics

Independent of the computational method chosen, it is essential to test and validate the model developed. Measurement of the model quality can be assessed by several metrics discussed in the litterature[64, 65]. They all aim to describe the model's aptitude in retrieving the active compounds from a database containing both active molecules and decoys. The accuracy of the model is defined as the percentage of molecules that are correctly classified evaluating the fraction of true positives and true negatives yet ignoring the hit list size. Thus, the enrichment factor, certainly the most commonly used metric, measures the improvement of the hit rate compared to a random selection. 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:

$$\text{Enrichment} = \frac{Hits_{sampled} / N_{sampled}}{Hits_{Total} / N_{Total}}$$

where;

- Hits sampled = Actual number of hits
- Hits Total = Total number of hits
- N sampled = Actual number of compounds sampled
- N Total = Total number of compounds

Unfortunately, the enrichment factor does not take into account whether the retrieved actives were ranked at the end or the top of the VS output hit list. Facing this consideration, a plethora of metrics have been developed, one trying to overcome the defaults of the other. Currently, the trend is to apply statistical methods trying to measure significance of diverse trend. Among the derivate metrics the Receiver Operating Characteristic (ROC) curve seems to comply to all needs by measuring all true and false negatives and positives quantities and ranks. Indeed, an ideal model function allows the choice of some cutoff value that completely separates active molecules from inactives or decoys. The Figure 1.5 illustrates the actual results in practice.

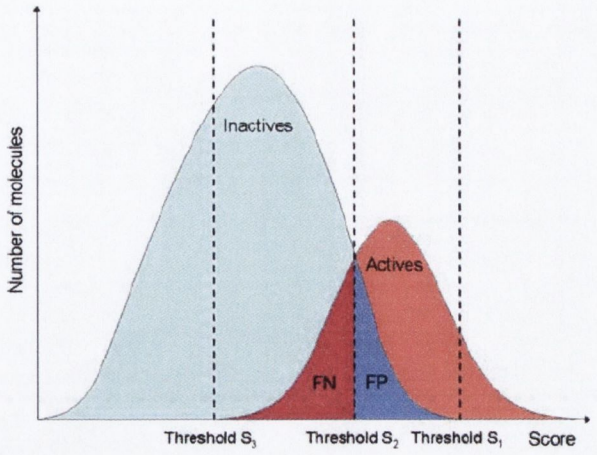


Figure 1.5: Theoretical distributions for active molecules and decoys according to their score[64].

Different ratios of false positives (FP) and false negatives (FN) are retrieved depending on the selection of the threshold score *S*

ROC plots illustrate the performance of a scoring function by comparing sensitivity with specificity for different cutoff values (*cf.* Figure S_1, S_2 and S_3). Sensitivity (*Se*) is a measure of the ability to avoid false-negatives and Specificity (*Sp*) the ability to predict true inactives. These two parameters are defined as followed:

$$Se = \frac{N \text{ selected actives}}{N \text{ total actives}} = \frac{TP}{TP + FN}$$

$$Sp = \frac{N \text{ discarded inactives}}{N \text{ total inactives}} = \frac{TN}{TN + FP}$$

where

- TP is the number of True Positives (actives) selected by the VS protocol,
- FP is the number of False Positive (inactives) or decoys selected by the VS protocol,
- FN is the number of False Negatives (actives) not selected by the VS protocol, and
- TN is the number of True Negatives (inactives) not selected by the VS protocol.

As sensitivity is defined as the percentage of actives selected (TP), it is also called "true positive rate", and as specificity is defined as the percentage of inactives not selected calculated over all inactives, (1 - specificity) is also called the "false positive rate". Thus, an ROC plot can be thought of as a plot of the true positive rate against the false positive rate, as illustrated in Figure :

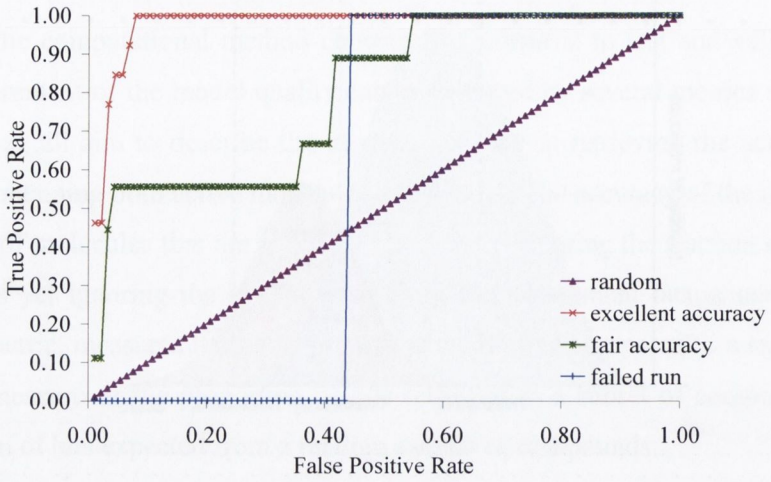


Figure 1.6: Example of ROC curve.

In purple, a random model, in red, a close to ideal model, in green a fair, model and in blue, a model that incorrectly selects decoys in preference to active molecules.

The closer the curve follows the left-hand border and then the top border of the ROC space as illustrated by the red line, the more accurate the model. A curve equal to the 45-degree diagonal illustrates a random model (purple line). This random distribution of scores causes the curve to tend towards the Se = 1 - Sp line asymptotically with increasing number of actives and decoys. The area under the curve (AUC), also denoted the ROC score, is a measure of the VS protocol accuracy. An area of 1 represents a perfect model whilst one of 0.5 represents a worthless model. In general, a ROC score of 0.9 to 1 is an excellent accuracy, 0.8 to 0.9 good, 0.7 to 0.8 fair, 0.6 to 0.7 poor and 0.5 to 0.6 a fail.

1.5 HTS/HCS approach

Nowadays, more than 100 diseases therapies are based on approximately 500 molecular targets[66] showing the necessity to better characterize the relevant drug-target to a particular disease. The constant genetics advances enforce biochemical and pharmaceutical research by strongly contributing to the elucidation of biochemical pathways and target identification. Indeed, genomic knowledge in sequences determination has extraordinary extended in several species from simple organisms such as yeast and worm to more complex organisms such as flies, mammals and last but not least in the human genome sequencing[67-75]. Nonetheless, more than one third of genes functions reminds unknown at the present time.[72]

To overcome this, using HTS methods allied to proteomics and genetics as well as DNA microarrays and protein global expression and transcription profiling, we have been able to enhance our understanding of protein-protein interactions and genes functions knowledge.

However, in practice, these techniques are based on intact cells or organisms, not representative of the actual studied disease system. Consequently, fluorescence microscopy has emerged as a suitable complementary approach to cover the lack of information inherent to all previous methods. Fluorescence microscopy used by High Content Analysis (HCA) could, on large scales, and under more realistic conditions, explain complex molecular interactions in whole cells. Moreover, HCA can, in real time, follow diverse biological processes simultaneously, quantify changes and locate molecular targets within suitable samples such as human tissue, parasites, diseased cells etc.

All together, biochemical markers, fluorescence microscopes, imagine technology and image data analysis methods stunning improvements have led to an interest in HCA for a plethora of bio-applications.

Protein functions[76] and target elucidation are two novel possible applications. Allied to the production of full length DNA clones denoted cDNA and RNA molecules, target discovery and pathway elucidation have been largely enhanced. World-wide several projects have proved the capacity of such imaging technology in target discovery[77-79]. For instance, the eminent technology to knockdown target denoted RNA interference (RNAi)[80 81, 82] has led to new drug target identification. Indeed, as short-interfering RNAs (siRNAs) target the complementary mRNA, the later is subsequently cleaved and degraded. In this manner, a specific loss of protein function can be detected therefore identifying protein functions[83].

Furthermore, the follow up of living cells fluorescence marker can provide temporal and spatial information at the sub-cellular level enabling cell phenotypes population classification. These phenotypes can be directly linked to changes due to drug treatment. When applied to screens of library of compounds, the automation and quantification provided by such methods can deliver large collections of data valuable for drug discovery data.

Apart from living cells, the use of fixed cells can ensure reproducibility of the assay undertaken limiting the variability in cell detection. Assaying becomes much faster and experimental issues are considerably reduced permitting high quantities of imaging analysis. This in turn can, for instance, track changes at the sub-cellular level and cellular phenotypes. These new information can then be use to enrich the still too poor variety of drug targets. At present in pharmaceutical targeting is restricted to mainly GPCRs (> 45%), various enzymes (28%), hormones and growth factors (11%) and ion channels (5%)[66]. For example, research in drug transportation, secretory pathways, and target functionality, applied with the large compound library screens has delivered not only new targets but also new active compounds.

In addition, lately, toxicity assessment using HCA technology has attracted much attention[84, 85]. As a matter of fact, assessment of genotoxicity is of main importance to exclude the mutagenic or carcinogenic effects of potential drugs as side effect reduction by optimal drug dosage. All require an arsenal of physiological, biochemical and pharmacological tests[86] and often underlie drug failure in the progression to clinical trials. At the moment, the rate of drug-candidate progression reaches not more than 30% at the early and less than 10% at the late stages in drug development[87]. The fast

large clustering of diverse phenotypic population had permitted a net improvement in the early detection of toxicity and had also considerably reduced the general costs of such studies by reducing the number of drug candidates[88]. Linking VS to HCA is both attractive and sensible to enhance the discovery process.

1.6 Objectives

As mentioned previously, PBOX compounds would be greatly enhanced if the intrinsic metabolic liabilities of the series were overcome through a new scaffold. 'Scaffold-hopping' or chemotype switching is a relatively new technique in computational biology and drug design where the pharmacologically relevant features of an active ligand are used for *in silico* virtual screening to identify alternate chemotypes with equivalent potency and mechanism of action. This thesis integrates complementary *in silico* and *in vitro* screening technologies for the discovery and mechanistic elucidation of novel inducers of apoptosis, through the application of chemotype switching from the PBOX parent.

Chapter 2 presents an examination of the first step generally used prior to a VS campaign. Indeed, it is of common use to pre-filter compounds libraries, for instance in order to create target focus libraries or drug-like libraries [47]. Therefore, the concept of drug likeliness and the appliance of such filters in oncology research will be discussed in Chapter 2. This section includes a short review of currently known cancer drugs as an introduction prior to their mapping in chemical medicinal space. This section concludes on whether or not to apply such standard filters prior to using a scaffold hoping protocol for PBOX.

Chapter 3 presents from one hand, the LBVS protocol used for the discovery of novel anti-cancer agents and from the other hand, the elaboration of a test set used for *in silico* validation. Compound databases of commercially available compounds are virtually enumerated as multimeric forms and indexed for 2D and 3D searching. The final expectation is to find new diverse structures mechanistically similar to PBOX to enhance apoptotic specificity on cancer cell lines.

Chapter 4 aims to rationalize the PBOX tubulin interaction therefore seeking its binding site on the protein. Elucidation of PBOX's tubulin binding site will deliver two outcomes. This could guide biochemistry research in the understanding and description of PBOX anti-cancer mechanism of action. In addition, identification of PBOX binding site would enable the application of SBVS that could overcome the inherent difficulties of working on small and highly similar analogues set of actives. This section first presents a complete review on tubulin essential to the understanding of its structure. Tubulin's role, mechanistic and specific binders known to date are discussed in the context of cancer therapy. This introduces a novel computational method involving 3D protein binding site mapping to attempt to elucidate PBOX binding site on tubulin.

Finally, Chapter 5 presents the experimental work undertaken including the set up and development of a High Content Screening (HCS) platform based on a novel, fast and cheap assay.

PBOX compounds potencies was assessed as their structure activity relationships was of first importance in 'Scaffold-hopping' methodologies. PBOX compounds structure activity relationships were evaluated *in vitro* by choosing noteworthy parameters and using novel High Content Analysis (HCA) techniques. *In silico* screening hits were validated *in vitro* towards the identification of novel viable agents for CML chemotherapy, validating the scaffold hopping methodology.

References

1. Yoshida, C.; Melo, J. V., Biology of chronic myeloid leukemia and possible therapeutic approaches to imatinib-resistant disease. *Int J Hematol* **2004,** 79, (5), 420-33.
2. Sinclair, P. B.; Nacheva, E. P.; Leversha, M.; Telford, N.; Chang, J.; Reid, A.; Bench, A.; Champion, K.; Huntly, B.; Green, A. R., Large deletions at the t(9;22) breakpoint are common and may identify a poor-prognosis subgroup of patients with chronic myeloid leukemia. *Blood* **2000,** 95, (3), 738-43.
3. Wang, J. Y., Abl tyrosine kinase in signal transduction and cell-cycle regulation. *Curr Opin Genet Dev* **1993,** 3, (1), 35-43.
4. Verfaillie, C. M., Biology of chronic myelogenous leukemia. *Hematol Oncol Clin North Am* **1998,** 12, (1), 1-29.
5. Bedi, A.; Griffin, C. A.; Barber, J. P.; Vala, M. S.; Hawkins, A. L.; Sharkis, S. J.; Zehnbauer, B. A.; Jones, R. J., Growth factor-mediated terminal differentiation of chronic myeloid leukemia. *Cancer Res* **1994,** 54, (21), 5535-8.
6. Antonelli, G.; Currenti, M.; Turriziani, O.; Dianzani, F., Neutralizing antibodies to interferon-alpha: relative frequency in patients treated with different interferon preparations. *J Infect Dis* **1991,** 163, (4), 882-5.
7. Hehlmann, R.; Heimpel, H.; Hasford, J.; Kolb, H. J.; Pralle, H.; Hossfeld, D. K.; Queisser, W.; Loffler, H.; Hochhaus, A.; Heinze, B.; et al., Randomized comparison of interferon-alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. The German CML Study Group. *Blood* **1994,** 84, (12), 4064-77.
8. Hrabanek, J.; Lukasova, M.; Vitek, A.; Sajdova, J.; Jelinek, J.; Vanasek, J.; Chudomel, V.; Korinkova, P.; Michalova, K.; Haskovec, C.; et al., [Treatment of chronic myeloid leukemia with bone marrow transplantation at the Institute of Hematology and Blood Transfusion in Prague]. *Vnitr Lek* **1995,** 41, (10), 682-7.
9. Rushing, D.; Goldman, A.; Gibbs, G.; Howe, R.; Kennedy, B. J., Hydroxyurea versus busulfan in the treatment of chronic myelogenous leukemia. *Am J Clin Oncol* **1982,** 5, (3), 307-13.
10. Clarkson, B.; Strife, A.; Wisniewski, D.; Lambek, C. L.; Liu, C., Chronic myelogenous leukemia as a paradigm of early cancer and possible curative strategies. *Leukemia* **2003,** 17, (7), 1211-62.
11. Hitt, E., Homoharringtonine effective in CML patients. *Lancet Oncol* **2002,** 3, (5), 259.
12. Johnson, S. A., Nucleoside analogues in the treatment of haematological malignancies. *Expert Opin Pharmacother* **2001,** 2, (6), 929-43.
13. Verfaillie, C. M.; McIvor, R. S.; Zhao, R. C., Gene therapy for chronic myelogenous leukemia. *Mol Med Today* **1999,** 5, (8), 359-66.
14. Zimmermann, J.; Caravatti, G.; Mett, H.; Meyer, T.; Muller, M.; Lydon, N. B.; Fabbro, D., Phenylamino-pyrimidine (PAP) derivatives: a new class of potent and selective inhibitors of protein kinase C (PKC). *Arch Pharm (Weinheim)* **1996,** 329, (7), 371-6.
15. von Bubnoff, N.; Peschel, C.; Duyster, J., Resistance of Philadelphia-chromosome positive leukemia towards the kinase inhibitor imatinib (STI571, Glivec): a targeted oncoprotein strikes back. *Leukemia* **2003,** 17, (5), 829-38.
16. Faderl, S.; Talpaz, M.; Estrov, Z.; O'Brien, S.; Kurzrock, R.; Kantarjian, H. M., The biology of chronic myeloid leukemia. *N Engl J Med* **1999,** 341, (3), 164-72.
17. Campiani, G.; Fiorini, I.; De Filippis, M. P.; Ciani, S. M.; Garofalo, A.; Nacci, V.; Giorgi, G.; Sega, A.; Botta, M.; Chiarini, A.; Budriesi, R.; Bruni, G.; Romeo, M. R.; Manzoni, C.; Mennini, T., Cardiovascular characterization of pyrrolo[2,1-d][1,5]benzothiazepine derivatives binding selectively to the peripheral-type benzodiazepine receptor (PBR): from dual PBR affinity and calcium antagonist activity to novel and selective calcium entry blockers. *J Med Chem* **1996,** 39, (15), 2922-38.
18. Braestrup, C.; Squires, R. F., Specific benzodiazepine receptors in rat brain characterized by high-affinity (3H)diazepam binding. *Proc Natl Acad Sci U S A* **1977,** 74, (9), 3805-9.
19. Zisterer, D. M.; Williams, D. C., Peripheral-type benzodiazepine receptors. *Gen Pharmacol* **1997,** 29, (3), 305-14.
20. Campiani, G.; Butini, S.; Gemma, S.; Nacci, V.; Fattorusso, C.; Catalanotti, B.; Giorgi, G.; Cagnotto, A.; Goegan, M.; Mennini, T.; Minetti, P.; Di Cesare, M. A.; Mastroianni, D.; Scafetta, N.; Galletti, B.; Stasi, M. A.; Castorina, M.; Pacifici, L.; Ghirardi, O.; Tinti, O.; Carminati, P., Pyrrolo[1,3]benzothiazepine-based atypical antipsychotic agents. Synthesis, structure-activity relationship, molecular modeling, and biological studies. *J Med Chem* **2002,** 45, (2), 344-59.
21. Hirsch, T.; Decaudin, D.; Susin, S. A.; Marchetti, P.; Larochette, N.; Resche-Rigon, M.; Kroemer, G., PK11195, a ligand of the mitochondrial benzodiazepine receptor, facilitates the induction of apoptosis and reverses Bcl-2-mediated cytoprotection. *Exp Cell Res* **1998,** 241, (2), 426-34.

22. Zisterer, D. M.; Hance, N.; Campiani, G.; Garofalo, A.; Nacci, V.; Williams, D. C., Antiproliferative action of pyrrolobenzoxazepine derivatives in cultured cells: absence of correlation with binding to the peripheral-type benzodiazepine binding site. *Biochem Pharmacol* **1998,** 55, (4), 397-403.
23. Mc Gee, M. M.; Campiani, G.; Ramunno, A.; Fattorusso, C.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Pyrrolo-1,5-benzoxazepines induce apoptosis in chronic myelogenous leukemia (CML) cells by bypassing the apoptotic suppressor bcr-abl. *J Pharmacol Exp Ther* **2001,** 296, (1), 31-40.
24. Mc Gee, M. M.; Campiani, G.; Ramunno, A.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Activation of the c-Jun N-terminal kinase (JNK) signaling pathway is essential during PBOX-6-induced apoptosis in chronic myelogenous leukemia (CML) cells. *J Biol Chem* **2002,** 277, (21), 18383-9.
25. Mc Gee, M. M.; Hyland, E.; Campiani, G.; Ramunno, A.; Nacci, V.; Zisterer, D. M., Caspase-3 is not essential for DNA fragmentation in MCF-7 cells during apoptosis induced by the pyrrolo-1,5-benzoxazepine, PBOX-6. *FEBS Lett* **2002,** 515, (1-3), 66-70.
26. Mulligan, J. M.; Campiani, G.; Ramunno, A.; Nacci, V.; Zisterer, D. M., Inhibition of G1 cyclin-dependent kinase activity during growth arrest of human astrocytoma cells by the pyrrolo-1,5-benzoxazepine, PBOX-21. *Biochim Biophys Acta* **2003,** 1639, (1), 43-52.
27. Mulligan, J. M.; Greene, L. M.; Cloonan, S.; Mc Gee, M. M.; Onnis, V.; Campiani, G.; Fattorusso, C.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Identification of tubulin as the molecular target of proapoptotic pyrrolo-1,5-benzoxazepines. *Mol Pharmacol* **2006,** 70, (1), 60-70.
28. Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004,** 432, (7019), 862-5.
29. Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **2002,** 1, (11), 882-94.
30. Mestres, J., Virtual screening: a real screening complement to high-throughput screening. *Biochem Soc Trans* **2002,** 30, (4), 797-9.
31. Stoermer, M. J., Current status of virtual screening as analysed by target class. *Med Chem* **2006,** 2, (1), 89-112.
32. Hann, M., A perspective on what we can and can't do with Computational Chemistry in support of Drug Discovery. *presentation at the MGMS conference,Trinity College Dublin, Ireland* **2005**.
33. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **2006,** 11, (13-14), 580-94.
34. Pipeline Pilot, *http://accelrys.com/products/scitegic/* **Software developed by Scietegic, a product from Accelrys.,** Pieline client version 6.0.2.0 , Pipeline Pilot Server version 6.5.1.0.
35. Bender, A.; Glen, R. C., Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem* **2004,** 2, (22), 3204-18.
36. Schneider, G.; Bohm, H. J., Virtual screening and fast automated docking methods. *Drug Discov Today* **2002,** 7, (1), 64-70.
37. Schuffenhauer, A.; Gillet, V. J.; Willett, P., Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J Chem Inf Comput Sci* **2000,** 40, (2), 295-307.
38. Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W., "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J Chem Inf Model* **2006,** 46, (6), 2445-56.
39. Catalyst Version 4.11 available from Accelrys Inc. San Diego.CA.USA
40. OMEGA2, http://www.eyesopen.com. **developed by Openeye scientific software,**.
41. 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.
42. 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, (6), 1908-19.
43. Catalyst Tutorials Version 4.11
44. http://cnx.rice.edu/content/m11538/latest/.
45. Yang, J. M.; Shen, T. W., A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **2005,** 59, (2), 205-20.
46. Johnson, M.; Lajiness, M.; Maggiora, G., Molecular similarity: a basis for designing drug screening programs. *Prog Clin Biol Res* **1989,** 291, 167-71.
47. Kubinyi, H., [The key to the castle. II. Hansch analysis, 3d-QSAR and de novo design]. *Pharm Unserer Zeit* **1994,** 23, (5), 281-90.
48. Kubinyi, H., Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. *J Cancer Res Clin Oncol* **1990,** 116, (6), 529-37.
49. Hopfinger, A. J.; Burke, B. J., Molecular shape analysis of structure-activity tables. *Prog Clin Biol Res* **1989,** 291, 151-9.
50. Clarck, D. E., Evolutionary Algorithm in Molecular Design, in Methods and Principles in Medicinal Chemistry. *R. Mannhold, H. Kubinyi H. Timmeran (Eds)*, (Wiley-VCH: Weinheim, Germany.).

51. Holland, J. H., Adaptation in natural and Artificial systems. *University of Michigan press Ann Ardbor, MI,* **1975**.
52. Manly, B. F. J., Multivariate Statistical Methods: A prime. *Chapman and Hall* **1994,** 2nd Ed, (New York).
53. Grimm, L. G., Yarnold, P. R.., Reading and understanding Multivariate Statistics; . *American Psychological Association; Washigton, Dc,* **1995**.
54. Pearson, K., On lines and planes of closest fit to a system of points in space. *Philosophical Magazine* **1901,** 2, 557–572.
55. Hotelling, H., Analysis of a complex of statistical variables principal components into in space. *Educational Psycology* **1901,** 24.
56. Glen, W. G. et al., Principal components analysis and partial least squares regression. *Tetrahedron Computer Methodology* **1989,** 2, 349-376.
57. Lindberg, W. at al. Wold, Partial least squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Anal. Chem.* **1983**, (55), 643-648.
58. Dunn III, W.J., Wold., S., Edlund U., Hellberg S. and Gasteiger, J. Multivariate Structure-Activity Relationships Between Data from a Battery of Biological Tests and an Ensamble of Structure Descriptors: The PLS Method. . *Quant. Struct.-Act. Relat. 3* **1984**, 131–137.
59. Gasteiger, J. Z., J. Angew, Neural Networks for Chemists: An Introduction. *VCH-Verlag* **1993,** Weinheim, 300.
60. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** 7, (20), 1047-55.
61. Hetenyi, C.; van der Spoel, D., Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett* **2006,** 580, (5), 1447-50.
62. Paul, N.; Kellenberger, E.; Bret, G.; Muller, P.; Rognan, D., Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **2004,** 54, (4), 671-80.
63. Chen, Y. Z.; Ung, C. Y., Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J Mol Graph Model* **2001,** 20, (3), 199-218.
64. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T., Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes? *J Comput Aided Mol Des* **2008,** 22, (3-4), 213-28.
65. Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A., How to do an evaluation: pitfalls and traps. *J Comput Aided Mol Des* **2008,** 22, (3-4), 179-90.
66. Drews, I. I., Drug discovery today - and tomorrow. *Drug Discov Today* **2000,** 5, (1), 2-4.
67. Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G., Life with 6000 genes. *Science* **1996,** 274, (5287), 546, 563-7.
68. www.yeastgenome.org.
69. http://www.sanger.ac.uk/Projects/C_elegans and www.sanger.ac.uk/Projects/C_elegans/ Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **1998,** 282, (5396), 2012-8.
70. www.wormbase.org.
71. Adams, M. D.; Celniker, S. E.; Holt, R. A.; et al, The genome sequence of Drosophila melanogaster. *Science* **2000,** 287, (5461), 2185-95.
72. Lander, E. S.; Linton, L. M.; Birren, B.; et al, Initial sequencing and analysis of the human genome. *Nature* **2001,** 409, (6822), 860-921.
73. Venter, J. C.; Adams, M. D.; Myers, E. W.; et al, The sequence of the human genome. *Science* **2001,** 291, (5507), 1304-51.
74. http://www.nature.com/nature/journal/v431/n7011/suppinfo/nature03001.html Finishing the euchromatic sequence of the human genome. *Nature* **2004,** 431, (7011), 931-45.
75. www.ensembl.org.
76. Pepperkok, R.; Ellenberg, J., High-throughput fluorescence microscopy for systems biology. *Nat Rev Mol Cell Biol* **2006,** 7, (9), 690-6.
77. Gerhard, D. S.; Wagner, L.; Feingold, E. A.; et al, The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* **2004,** 14, (10B), 2121-7.
78. Kikuchi S. et al, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* **2003,** 301, (5631), 376-9.
79. Wiemann S. et al. , Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res* **2001,** 11, (3), 422-35.
80. Foster City, www.ambion.com. *CA, USA.*
81. Dharmacon, ONTARGETplus siGENOME siRNAs *www.dharmacon.com* Lafayette, CO, USA.
82. HP Genome Wide siRNAs Qiagen. *Hilden, Germany.*

83. Meister, G.; Tuschl, T., Mechanisms of gene silencing by double-stranded RNA. *Nature* **2004,** 431, (7006), 343-9.
84. Cellumen Inc, www.cellumen.com. *Pittsburgh, PA, USA.*
85. ArrayScanVTI HCS Reader, www.cellomics.com. *Pittsburgh, PA, USA.*
86. Pritchard, J. F.; Jurima-Romet, M.; Reimer, M. L.; Mortimer, E.; Rolfe, B.; Cayen, M. N., Making better drugs: Decision gates in non-clinical drug development. *Nat Rev Drug Discov* **2003,** 2, (7), 542-53.
87. Kola, I.; Landis, J., Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* **2004,** 3, (8), 711-5.
88. Perlman, Z. E.; Slack, M. D.; Feng, Y.; Mitchison, T. J.; Wu, L. F.; Altschuler, S. J., Multidimensional drug profiling by automated microscopy. *Science* **2004,** 306, (5699), 1194-8.

Chapter II

Oncology exploration

The work in this chapter was published as :

Lloyd, D. G., **Golfis, G.**, Knox, A. J., Fayne, D., Meegan, M. J., Oprea, T. I., Oncology exploration: charting cancer medicinal chemistry space, *Drug Discov Today*, **2006**, 11, 3-4 149-59

Introduction

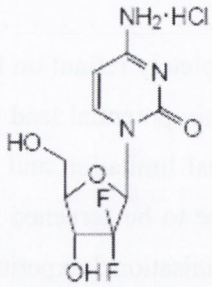
In a virtual screening campaign, the performance of the protocol is completely reliant on the quality of the library screened. This library, either virtual or physical, must contain potential lead candidates within it. Still in both cases its size remains constrained by computational limitation and laboratory facilities respectively. Therefore, trying to reduce the size of virtual space to be screened is of great importance, especially considering the capabilities of most research organisations' experimental and computational in-house facilities. In the last few years, significant research has examined various protocols to reduce the size of chemical space and to design focused virtual libraries or libraries of available compounds[1]. There have been important advances in this field, mainly in the realm of describing and differentiating 'drug' or 'lead'-like molecule subspaces of the complete chemical space.

PBOX compounds, have been suggested as potential leads for the treatment of leukaemia and could also find application in treating other types of cancer. Our final aim is to use Virtual Screening (VS) techniques for scaffold hopping using the PBOX core as a starting point. Focusing our future screens on relevant parts of chemical space would increase our chances of finding appropriate new hits. Hence, in this chapter our aim will be to describe oncology space, in relation to medicinal chemistry space, and investigate how to navigate space to converge on potential cancer therapeutic molecules.

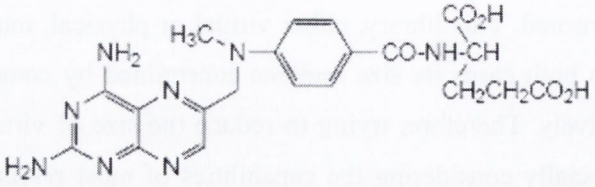
2.1 Cancer drugs

Great advances were made in the last decades in oncology in general and more precisely towards developing a deeper understanding of cancer mechanisms[2]. A plethora of molecules have been indexed as cancer active compounds and a large number of these have seen their target validated leading to some passing successfully through all clinical and toxicity tests[3].In order to describe cancer medicinal space, it is essential to classify cancer chemotherapeutics by scaffold, effect and potency. These molecules can be firstly divided into two groups, selective and non-selective targeting drugs.

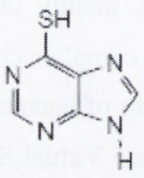
Selective targeting drugs induce cell death by mainly blocking DNA replication through different mechanisms during the cell mitosis[4, 5] (see figure 2.1). This first group encloses molecules such as antimetabolite drugs. Antimetabolites become building blocks of DNA during the "S" phase of the cell cycle and stop normal development and division[6, 7]. In addition, genotoxic drugs are considerably electrophilic so that they form strong covalent bonds with the DNA, causing mutations and preventing accurate replication[8-10].



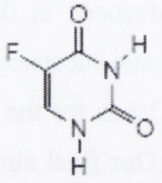
Gemcitabine



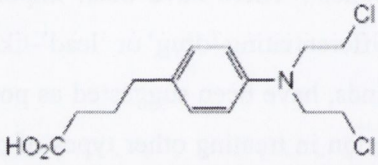
Methotrexate



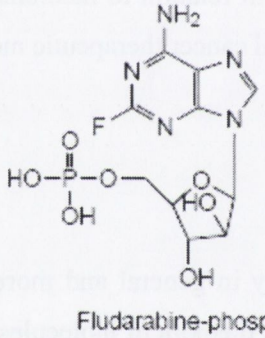
6-Mercaptopurine



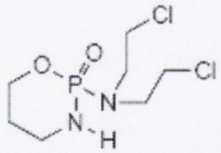
5-Fluorouracil



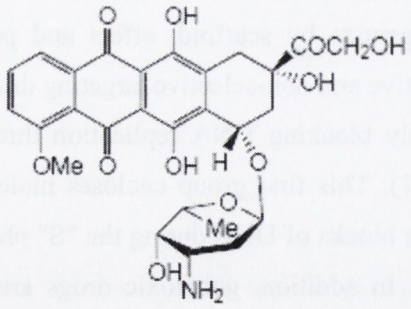
Chlorambucil



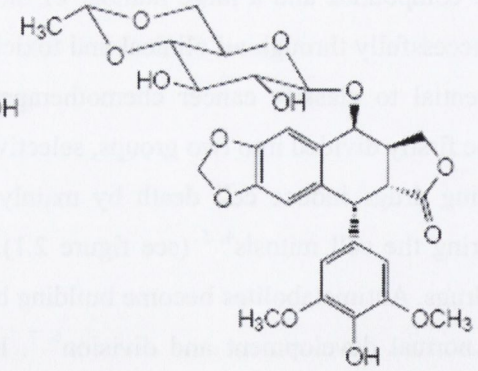
Fludarabine-phosphate



Cyclophosphamide



Doxorubicin



Etoposide

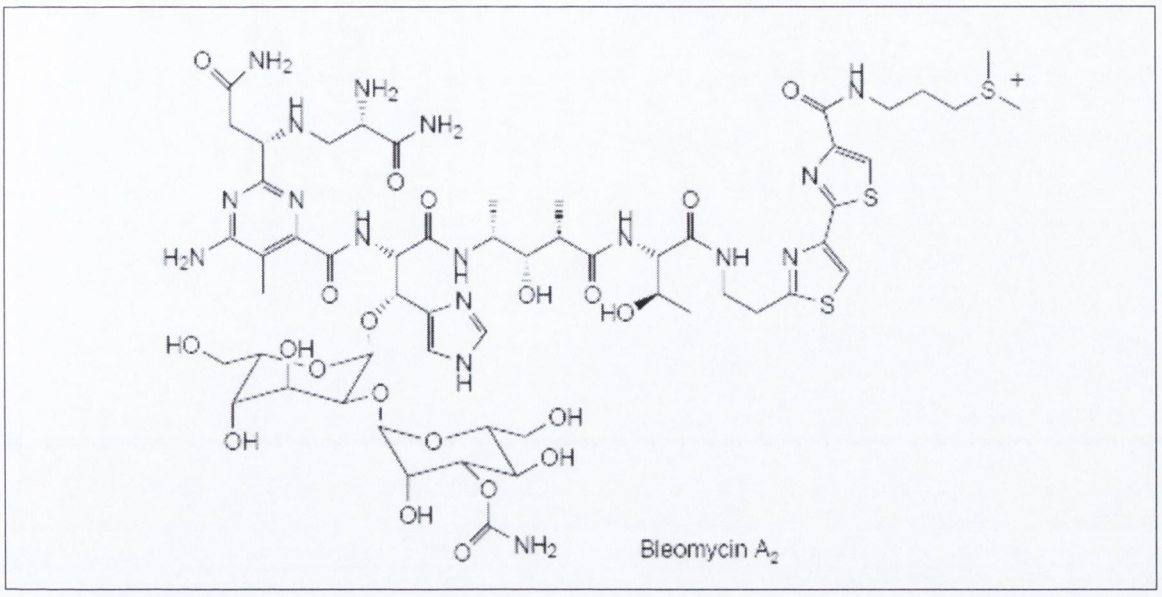


Figure 2.1: Chemical structures of non-selective chemotherapy drugs acting through different mechanisms.

A selection of genotoxic and antimetabolic drugs are depicted, illustrating the great variety in structural form of anticancer ligands.

The second group of cancer chemotherapeutics comprises drugs which can selectively target tumour cells to minimise associated toxicity[11, 12] (see Figure 2.2 and 2.3).

Kinases regulate proteins and cellular activity such as proliferation by addition of a phosphate group to other proteins which induces conformational changes[11]. Gleevec (imatinib) and iressa (genfinitib) are two different tyrosine kinase inhibitors respectively prescribed for chronic myeloid leukaemia treatment and refractive non-small cell lung cancer treatment[13].

Proteasome inhibitors find their medicinal application in treatments of, for instance, myeloma[14]. They may cause cell cycle arrest and cell death by interfering with the regulation of cell cycle proteins and proteins involved in the degradation of damaged proteins[15].

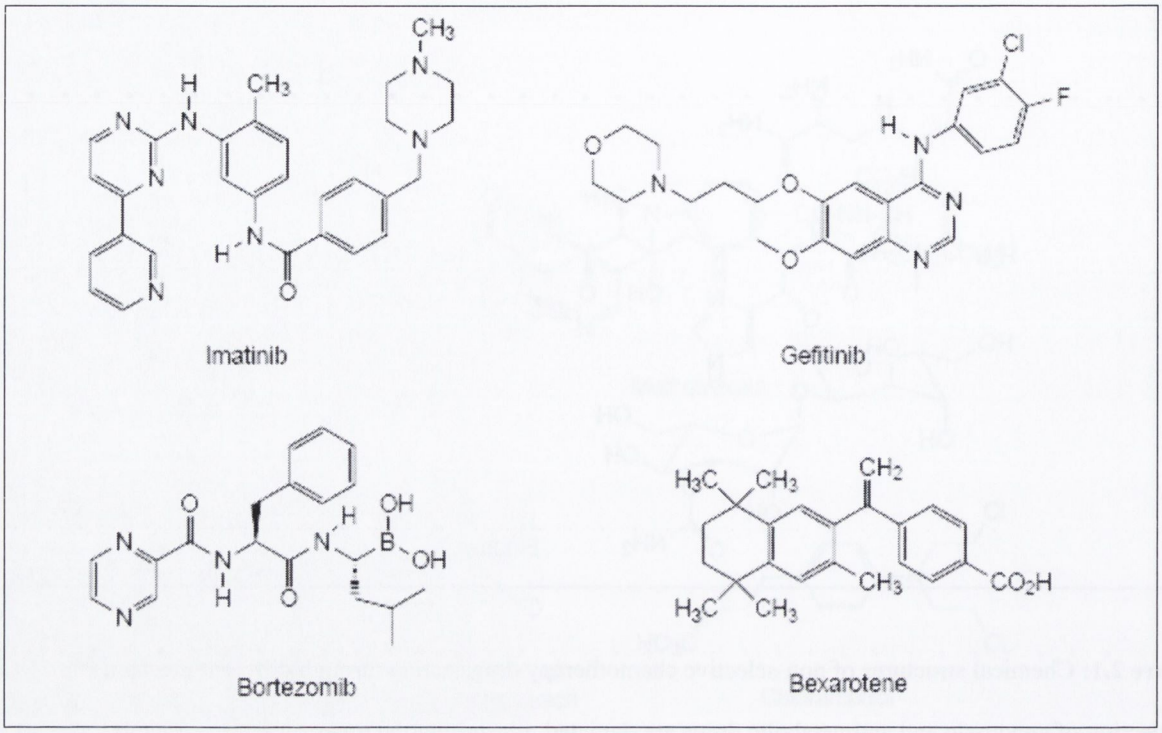


Figure 2.2 Chemical structures of small molecule cancer target selective inhibitors.

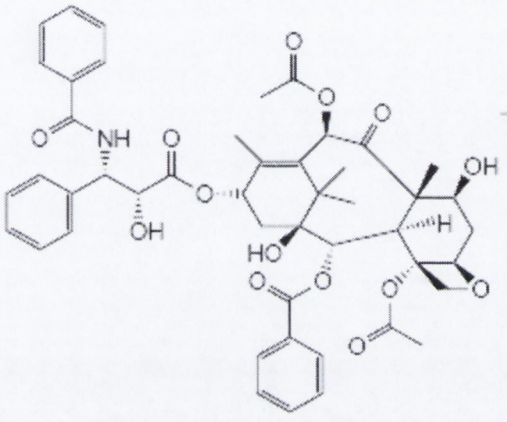
Kinase and proteasome inhibitors bind to specific protein targets rather than to proteins involved in basic biochemical processes.

Hormonal agents are in use in the treatment of cancers such as breast, ovarian and prostate cancer which all are hormonal signal dependent[16]. These drugs, by targeting hormonal receptors, block their action by changing the 3D shape of the receptor. A sub-group comprises of Selective Estrogen Receptor Modulators (SERMs) including estrogen antagonist such as tamoxifen, toremiphe and raloxifene) and Selective Androgen Receptor Modulators (SARMs) including antiandrogens such as flutamide, bicalutamide and nilutamide[17].

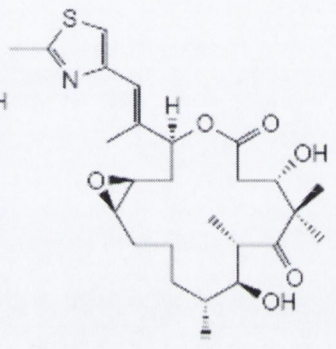
The biosynthesis of estrogen from androgen precursors may be prevented by blocking the action of the enzyme aromatase[18]. There are two types of aromatase inhibitors used clinically for postmenopausal ER positive breast cancer: the steroid examestane and the nonsteroidal aromatase inhibitors, anastrazole and letrozole.

Last but not least, antimitotic agents are a growing class of cancer therapeutics. Tubulin is the building block protein of microtubules; one of the essential fibres that the cytoskeleton is composed of. Microtubules are involved in the separation of the two daughter cells during mitosis by forming the mitotic spindle. Antimitotic drugs interfere with the normal process of cell division. On one hand, drugs such as vincristine bind to tubulin monomers and inhibit the formation of microtubules. On the

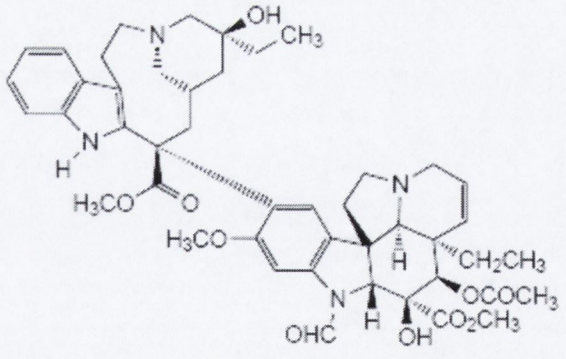
other hand, drugs such as paclitaxel (Taxol) and the epithiolones stabilise the microtubules thus preventing the normal cell division process[14]. Paclitaxel has a number of indications within various chemotherapies including combination regimens with cisplatin for lung and ovarian cancers and also an application with the monoclonal antibody Herceptin for breast cancer. Microtubule targeting drugs will be discussed in detail in chapter 4.



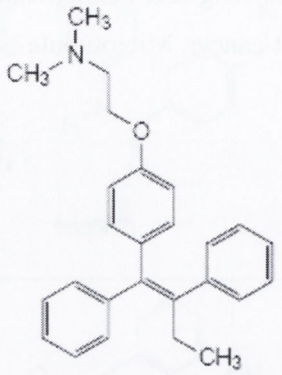
Paclitaxel



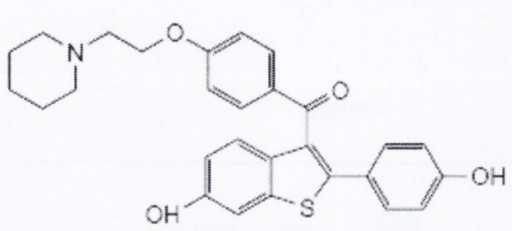
Epothilone A



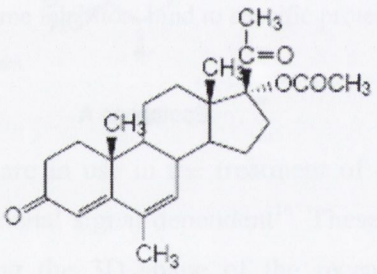
Vincristine



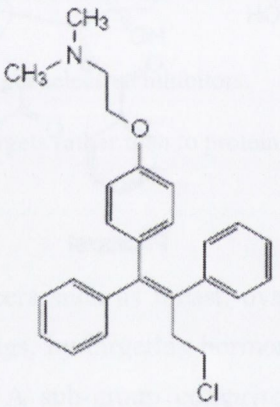
Tamoxifen



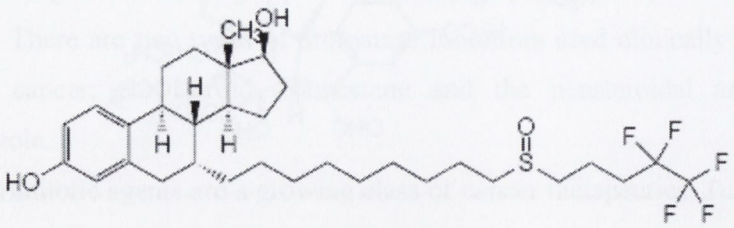
Raloxifene



Megestrol



Toremiphene



Fulvestrant, ICI 182780

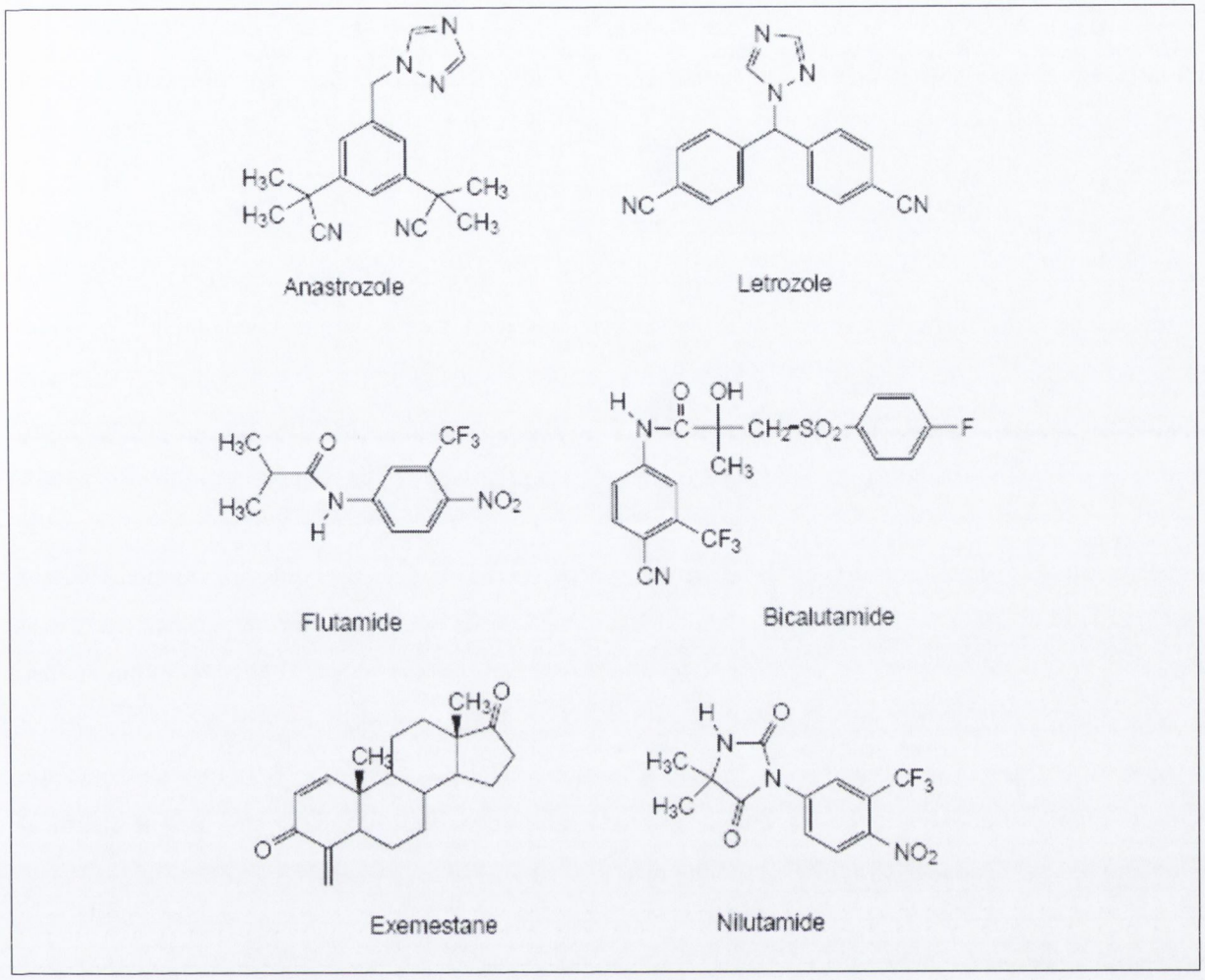


Figure 2.3: Example of antimitotic, hormonal, SERM, SARM and aromatase targeting ligands. These drugs specifically target different classes of cancer-related proteins.

2.2 Managing chemical space for virtual screening

Douglas Adams pointed out to would-be intergalactic hitch-hikers that space is mind-bogglingly big[19]. Similarly, chemical space is vast, so vast that it can be conceptually advantageous to notionally partition it into smaller, more-manageable sections. From a drug design perspective, these areas could include biologically relevant or medicinal chemical space, which would encompass smaller areas of theoretical chemical space and synthetically accessible space. These subsets further contain places such as drug space and lead space. The challenge for drug discovery is to explore areas of value in medicinal chemical space and, within these areas, retrieve bioactive, workable molecules. The navigational use of 'chemography' (GPS for drug discovery), mapping compounds onto chemical descriptors, allows us to

find our way among the vast number of chemical entities (10^{60}) that exist *in potentia*[20]. Paradoxically, this landscape is so large that we can never hope to explore it fully; however, the biologically relevant and valuable subset of hit space is much smaller than chemical space as a whole but elements of hit space are dispersed within the larger system.

Hence, our aim is to focus and limit the breadth of screening that must be undertaken to find a hit. An effective pre-selection of the screening candidates is critical and computational methods can be used for this purpose[21, 22]. These methods enable us to restrict virtual screens and consequently physical testing to a smaller sub-ensemble. Only a small part of the whole library, expected to include the most potentially active compounds, would be kept. These actives would be then selected for further more intense assaying[23]. Nowadays, cheminformatic filters are largely used to reduce compound datasets. Indeed, it is preferred to have one unique set of rules of filtering to apply to our libraries to direct the search towards a particular region of the chemical space. Therefore, selection depends on the choice of the correct parameters and thresholds for filtering. The last in part relies on how good we are at describing molecules within this particular area in order to differentiate them from random molecules. To do so, diverse biological fields can be investigated such as cell toxicity, cell membrane properties, receptor structures and so on[1] and statistical analysis on known drugs carried out. Two sets of rules are most commonly used in the pharmaceutical industry. They describe two different parts of chemical space: the 'drug-like space' and the 'lead-like space'.

The 'Rule of Five' (RO5)

Based on a study of the properties of orally available drugs Lipinski related in his key review an analysis of the reasons why compounds fail in progression. The necessity to consider pharmacokinetic properties in compound library design was highlighted[24]. This work furnished drug designers with a rule which essentially directed library creation and our consideration of screening collection diversity towards oral drug space. Application of the rule of five would enforce not breaking more than two of the following properties on a screening collection:

- Molecular weight (MW) $\leq$ 500 Da
- Octanol/water partition coefficient, logP $\leq$ 5
- number of H-bond donors $\leq$ 5
- number of H-bond acceptors $\leq$ 10

The 'Rule of Three' (RO3)

This rule set was proposed by Congreve et al.[25], and is designed for compliance in fragment-based drug discovery[26]. Application of the rule of three would enforce the following properties on a screening collection:

- Molecular weight ≤ 300 Da
- Octanol/water partition coefficient, logP ≤ 3
- number of H-bond donors and acceptors < 3
- flexible bonds < 3

Broadly speaking, compounds sought after in the drug discovery process can be split into two categories, drug-like and lead-like. The concept of lead-likeness implies a physicochemical profile in chemical libraries where the members have reduced complexity (e.g. MW <400) and other more restricted properties than those deemed drug-like. This leaves room for chemical modification in lead optimisation rounds, which subsequently modify the properties toward drug likeness[27]. Various authors have proposed different concept of what constitutes drug-like compounds. These models refer to having molecular similarities to known drugs[28] or acceptable absorption, distribution, metabolism, excretion and toxicology (ADME-Tox) properties[29]. Drug-likeness is often entirely dependant on the mode of administration. Other researchers have utilised the application of hard and fast cheminformatic rule systems to partition chemical space into drug-like and non drug-like bins. It was previously illustrated that RO5 compliance alone does not necessarily imbue drug-likeness on the post partitioned cohort[21]. It is generally accepted that the initial active molecules discovered in VS and HTS are often far removed from the drugs into which they might evolve through optimisation. In such practices the best we can hope for is the identification of screening hits.

2.3 Cancer Medicinal Chemistry Space

To accurately describe what we term cancer medicinal chemistry space, it was first necessary to amass a set of compounds known to be active against cancer – both clinical and pre-clinical compounds – as well as quality data on compounds purported to be anticancer active, but later shown to be inactive in assays.

A study set of clinical cancer compounds (34 ligands) and literature claimed active (4026 ligands) and verified inactive (4285) anticancer compounds were selected from the WOMBAT database (WOrld of Molecular BioAcTivity[30, 31]). This database contains chemical and biological data from 4773 papers published in medicinal chemistry journals between 1975 and 2005).

The NCI activity database (National Cancer Institute Database of 41086 compounds) was also used to identify a subset of anticancer actives.

In this database, compounds were tested against diverse cancer cell lines and tagged by GI50. The NCI renamed the IC50 value, the concentration that causes 50% growth inhibition, the GI50 value to emphasise the correction for the cell count at time zero; thus, GI50 is the concentration of test drug where $100 \times (T - T0)/(C - T0) = 50$. The optical density of the test well after a 48-h period of exposure to test drug is T, the optical density at time zero is T0, and the control optical density is C. The ``50" is called the GI50PRCNT, a T/C-like parameter that can have values from +100 to -100. The GI50 measures the growth inhibitory power of the test agent.

We classified the compounds according to their GI50 (8688 ligands with GI50 in > 6 assays in NCI screens) and verified inactive ligands (32398 ligands GI50 in < 6 assays) from compounds which have undergone general anticancer assays at NCI.

Finally, to contextualise cancer space in comparison to general medicinal chemical space, it was necessary to describe and populate generic drug-like space.

To achieve this, a filtered subset of the ZINC[32] database was used containing 109,432 commercially available 'drug-like' molecules, with no specific activity claimed. ZINC was our choice of database because it is pre-partitioned into various regions (e.g. drug-like and lead-like). ZINC has been widely adopted by the VS community because of the quality of its content and the availability of compounds therein for validation studies, and for this reason it has been used as a benchmark set.

Once unity of the data had been verified through cross correlating all dataset members, the study set of medicinal chemical space equated to a total of 158,863 compounds among which 12,714 are known cancer actives, 36,683 are known cancer inactives and the remainder can be described as classical drug-like ligands in nature. Unity of each compound was achieved using MOE toolkit based on SMILES strings (Simplified Molecular Input Line Entry Specification). Duplicate molecules were thereby detected and removed.

Last but not least, we used a set of 85 PBOX analogues with known and unknown apoptotic activities introduced to the screening set above.

2.4 Methods

The design of a focused library can be broken down into 2 steps. Firstly, parameters describing each molecule needs to be calculated and stored for each member of the database. These parameters are called descriptors. Secondly, a set of descriptors is chosen and thresholds for filtering determined[33] and the visualisation of the segregation of each data-subset can be illustrated using mapping methods.

2.4.1 Calculating descriptors

Molecular descriptors can be partitioned into dimensional classes:

- 1D descriptors: they can be physical properties or chemical properties attributed per compound. These descriptors only use the atom types of each molecule or can translate biological information such as GI50.
- 2D descriptors: they only use the atom information and their connection information or bond types.
- i3D. Internal 3D descriptors: They use 3D coordinate information about each molecule however they are invariant to rotations and translations of the conformation. In extension, x3D. (External 3D) descriptors also use 3D coordinate information but also require an absolute frame.

To partition the dataset on this basis, it was necessary to first calculate 2D descriptors for all members in MOE (CCG) and to identify those descriptors which related to adherence to the RO5. A total of 148 2D molecular descriptors were calculated describing atomic nature and molecular size, polarity, lipophilicity, and flexibility[34].

2.4.2 Filtering

It was also useful for us to quantify the subsets within our data that would conform to our working definition of hit-like (i.e. those compounds which would pass an application of the FILTER software protocol including RO5 compliance).

For the purpose of this study, we are conferring the loose description of 'hit-likeness' to those small molecule compounds that pass the unmodified OpenEye FILTER (which uses XLOGP[35] as a measure of the hydrophobicity partition coefficient with a maximum cut-off value of six) cheminformatic criteria including a strict application of the RO5. These criteria are now widely adopted in database filtering applied in the early stages of VS.

We partitioned the dataset according to the rule of five describing drug-like space using Filter toolkit enabling us to apply rapid partitioning of compound databases using these tuneable cheminformatic parameters based on the classical rules [36].

2.4.3 Mapping Methods

To present the relationship of cancer medicinal chemistry space in the context of wider chemical space in a meaningful and accessible way, it is necessary to construct graphical distributions of the

dataset in the 3D space. PCA is a relatively easy way to transform an n-descriptor space into a more-manageable 3D space. PCA reduces a multidimensional space by linearly transforming the data. This method is capable of calculating a new table of descriptors, often smaller, that are uncorrelated and normalised (mean 0 and variance 1).

PCA theory description[37]

Given m molecules each of which described by an n-vector of real numbers xi=(xi1,...,xin), consisting of the descriptors for molecule i.

Each molecule has an associated importance weight, wi, a non-negative real number. These weights can be thought of as the relative probability that the associated molecule will be encountered and are usually all equal to 1; however, in some applications, unequal weights are used.

Let W denote the sum of all the weights.

Dimension reduction through principal component analysis can be interpreted in the following manner.

Let X denote a random n-vector and let Z denote a random p-vector with mean 0 and covariance matrix equal to the identity matrix.

If we assume that X=RZ+x0 for some n by p linear transform R and some n-vector x0 then PCA is the estimation of the Z vectors from a sample of X vectors.

If we integrate both sides of the supposed affine transform we obtain:

$$E(X) = \int x \Pr(X = x) = \int (Rz + x_0) \Pr(Z = z) = R0 + x_0 = x_0$$

which shows that x0 is the mean of the distribution of the X vectors. Turning to the covariance of the X vectors, we observe that:

$$\mathrm{Cov}(X) = \int (x - x_0)(x - x_0)^T \Pr(X = x) = \int Rzz^T R^T \Pr(Z = z) = RR^T$$

These equations suggest the following method for estimating the Z vectors. We use the sampled data to approximate both E(X) and Cov(X):

$$E(X) \approx \bar{x} = \frac{1}{W} \sum_{i=1}^{m} w_i x_i, \quad \mathrm{Cov}(X) \approx S = \frac{1}{W} \sum_{i=1}^{m} w_i x_i x_i^T - \bar{x}\bar{x}^T$$

The sample covariance matrix, S, is now symmetric semi-definite; hence, all of its eigenvalues are real and non-negative. We can therefore diagonalise S so that:

$$S = QT\,D\,D\,Q$$

where

- Q is orthogonal
- D is diagonal-sorted in descending order from top left to bottom right.

Let p be the number of non-zero diagonal values in D (the square roots of the eigenvalues of S).

We can estimate R as the first p columns of QTD and say that the X vectors have p principal components. In practice, we restrict the selection of p further by limiting the condition of the estimated RTR matrix; that is, we choose p so that the largest eigenvalue divided by the smallest eigenvalue is less than some specified threshold. The foregoing considerations are integrated into the PCA computational procedure by an algorithmic procedure.

Algorithm description

- Calculate the sample average vector x0 and covariance matrix S.
- Diagonalise S.
- Remove all components associated with zero eigenvalues.
- Remove all components that exceed the matrix condition.
- Remove all components above the minimum variance threshold.
- Remove all components past the component limit.
- Form the PCA transform Q from the p remaining components so that Z=Q(x- x0) has identity covariance and zero mean.
- For each i write the p-vector zi = Q(xi- x0) to the output.

Source ccg.

In our analysis the multidimensional space was represented by the descriptors calculated and was reduced to three[34]. We transformed the 48 vectors space into a 3D space described by 3 principal component vectors, where each of the 3 vectors is a combination of the 48 weighted descriptors. These operations facilitated the creation of graphical representations of the 3D space spanned by the compound set. Finally results were plotted using MOE software for 3D visualisation.

2.5 Projection results

2.5.1 Drug like space vs. cancer medicinal space

Table 2.1 illustrates the nature of cancer compound space. In a total of 12,714 verified active anticancer compounds, only 33.4% pass a cheminformatic hit filter, positioning almost two-thirds of cancer actives outside what is traditionally accepted as hit-like chemical space.

Dataset / Source	Active	Inactive	Total Members	'Drug-like'	Rejected*
Clinical drugs (WOMBAT)	34	-	34	2	32 (94%)
'designed' compounds (WOMBAT)	4026	4285	8311	4150 (50% actives 50% inactives)	4161 (50%)
NCI assayed compounds (NCI)	8688	32398	41086	13042 (17% actives 83% inactives)	28044 (68%)
Anticancer active (WOMBAT & NCI)	12714	-	12714	4248	8466 (66.6%)

Table 2.1: Breakdown of cancer compound drug-like nature

rejected = fails on application of cheminformatic drug-like filter in FILTER.

The implication when looking for actives in filtered compound collections is immediately clear. When we examine the graphical distribution of cancer medicinal chemistry space, we can see a similar lesson emerging. The overall distribution of active anticancer compounds spans an area of medicinal chemistry space far beyond that described by traditional 'drug-like' definitions (Figure 2.4).

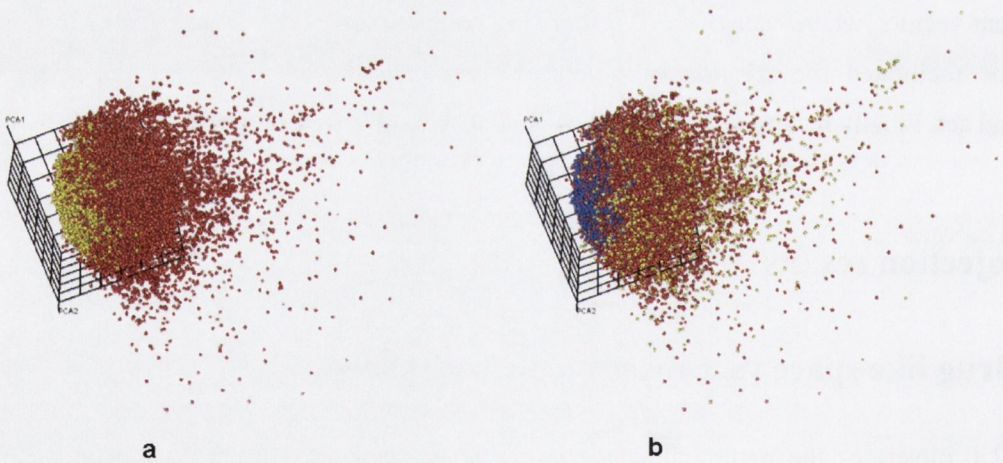


Figure 2.4: Charted Cancer Medicinal Chemistry Space.

(a) yellow sphere: generic "drug-like space", red spheres: medicinal chemistry space; (b) blue spheres: generic "drug-like" space, red spheres: cancer inactive medicinal chemistry space, yellow spheres: cancer active medicinal chemistry space.

Figure 2.4(a) contextualises the nature of medicinal chemistry space considered in this study. The yellow spheres are nonspecific hit-like compounds taken from the filtered ZINC database, illustrating the relatively compact nature of the RO5 space when compared with the wider medicinal chemical space. The red spheres are compounds which were claimed, assayed or demonstrated as anticancer agents from WOMBAT and the NCI databases. These compounds represent charted cancer medicinal chemical space. Figure 2.4(b) illustrates the distribution of active anticancer compounds (yellow) in comparison with inactive cancer medicinal chemical space (red) and in relation to generic hit-like compound space (blue).

2.5.2 Focus on target based spaces

When examining familial distributions, as for the cancer kinome, the breadth of cancer space spanned by actives is considerable (Figure 2.5). Figure 2.5 shows a view of the distribution of anticancer kinase-targeting compounds in medicinal chemistry space: from a total of 915 active compounds examined (blue spheres), only 156 (15%) lie within our defined hit-like space (yellow cloud), whereas 759 (83%) lie outside (red cloud) and do not pass application of FILTER incorporating RO5 compliance.

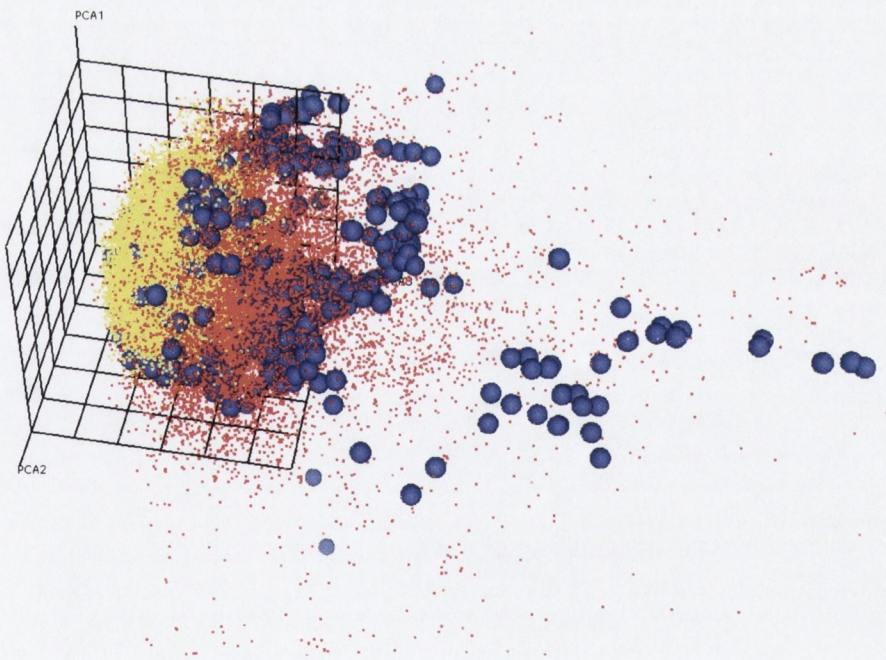


Figure2.5 Anticancer kinase-targeted space

Active ligands: yellow cloud, generic "drug-like" space - red cloud, non-kinase targeted medicinal chemistry space - blue spheres, anticancer kinase-targeted space.

Such a general spatial distribution of actives outside the boundaries of the traditional hit-space precludes the creation of all encompassing cancer generic filtering rules for database pre-processing. By adopting a class-by-class focus on targeted compound sets (e.g. antitubulin or anti-EGF receptor), compounds could be used to craft tailored cheminformatic filters biased to the target of study for the creation of more rationally focused screening collections. These filters can be utilised in the exploration of target-relevant chemical space. The caveat here is the need for unambiguous target information for selecting the regions of space that are to be explored. In Figure 2.6 the comparative distribution of targeted actives is presented, with reference to the wide spatial distribution of clinically used oncology compounds. It is clear that clusters exist in targeted medicinal chemistry space, in some instances these clusters are not far removed from hit-space and they could potentially be optimised into orally available drug-like space through design.

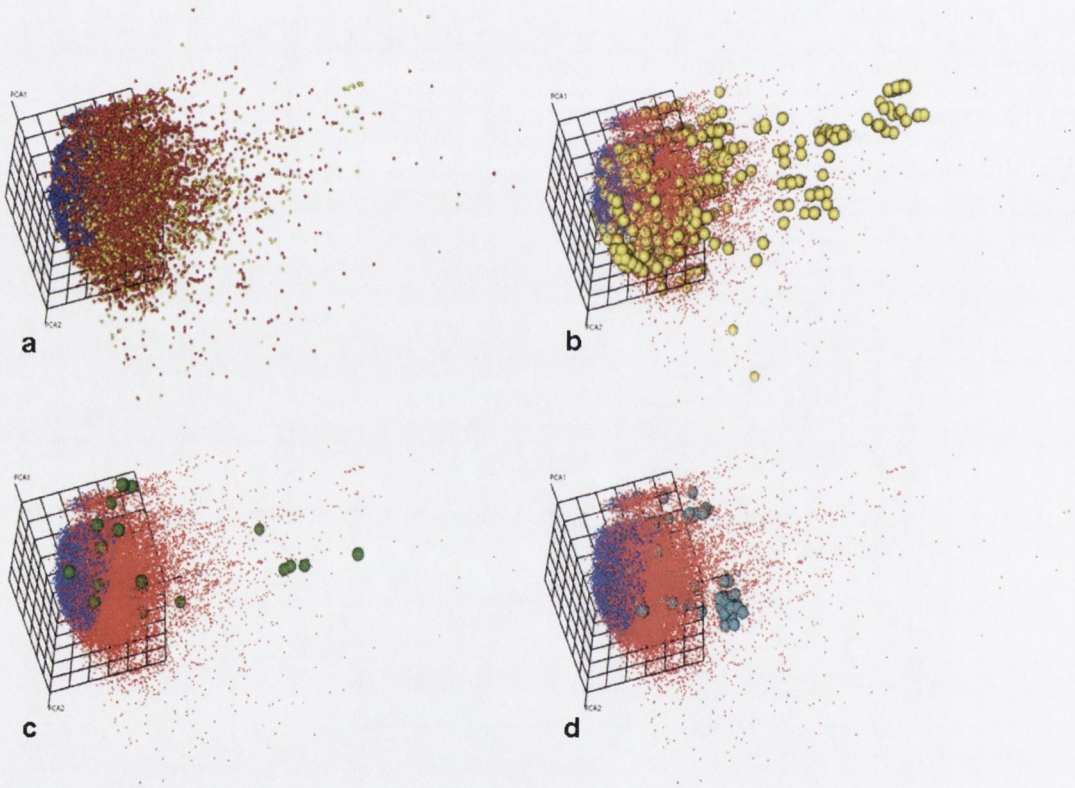
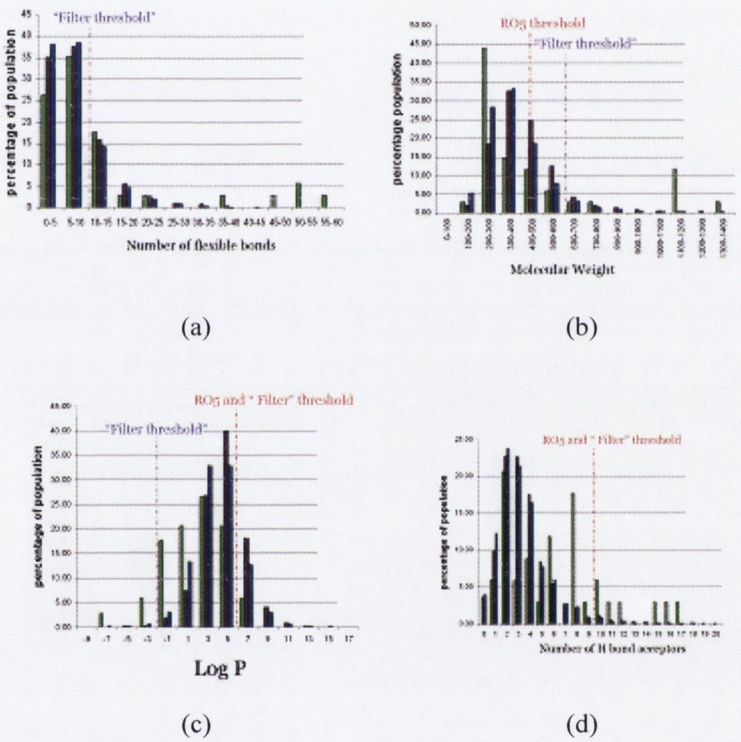


Figure2. 6: Charting cancer medicinal chemistry space.

(a) yellow spheres: all cancer actives, red spheres: cancer inactives, blue spheres: traditional 'drug-like' space (indeterminate activity). (b) yellow spheres: cancer actives known to target kinases, red cloud: cancer medicinal chemistry space, blue cloud: traditional 'drug-like' space. (c) green spheres: clinically used anticancer compounds, red cloud: cancer medicinal chemistry space, blue cloud: traditional 'drug-like' space. (d) cyan spheres: cancer actives known to target tubulin, red cloud: cancer medicinal chemistry space, blue cloud: traditional 'drug-like' space. ('Drug-like' in this instance is defined as compounds adhering to the Rule of Five)

2.5.3 Reasons for rejections

One would expect a modicum of attrition in such a process when utilities such as FILTER are designed to remove not only RO5 fails but also specific molecules containing toxic and reactive functionalities by removing staples such as alkylating agents, nitrogen mustards that make up a large proportion of our anticancer arsenal. In our analysis, however, the vast majority of compound failures stemmed from a lack of MW, log P and H-bond acceptor compliance with regard to the RO5. Figure 2.7 illustrates a sample of the descriptors responsible for failure of the sub ensembles passed through the exact rule of five filter and the default values of the toolkit 'Filter' mentioned previously in this chapter which is broadly used with its default settings.



(a) (b)

(c) (d)

Figure 2.7 RO5 descriptors distribution

(a) Flexible bond, (b) Molecular weight, (c) log P, (d) H bond acceptor

Blue bars: cancer inactives, purple bars cancer actives and green bars antineoplastic drugs. Dash red lines indicated Lipinski's rule thresholds and purple dash line "Filter" thresholds.

For each of the sub ensembles, antineoplastic drugs, cancer actives and cancer inactives, the 2D descriptor rejection trends are the same for the RO5 and FILTER. According to their number of H bond acceptors, the Log P value and their molecular weights, the antineoplastic drugs failed about 15-21% per descriptor and failed up to ~38% on the number of rotatable (flexible) bonds parameter. These

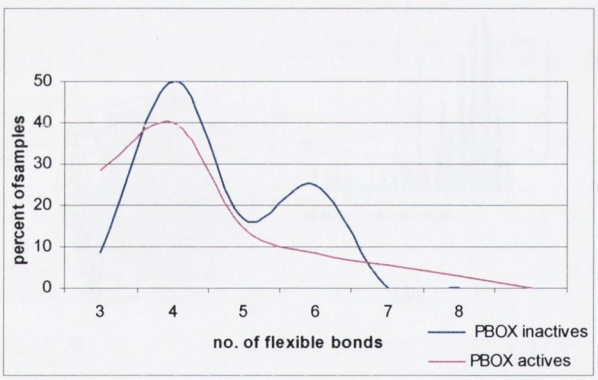
drugs are marketable drugs and have been subjected to diverse optimisation phases which almost always consist of an enhancement in their potency. In comparison to RO5, FILTER accepts more molecules through its nets, increasing the likelihood of increasing hit diversity and leaving space for optimisation growth. Cancer actives and inactives are rejected in the same proportions according to the four descriptors presented in figure 2.7.

Despite the generally low rates (2-27%) of discarded cancer active molecules per descriptor, addition of all the rejected molecules per descriptors together ends up representing almost their totality.

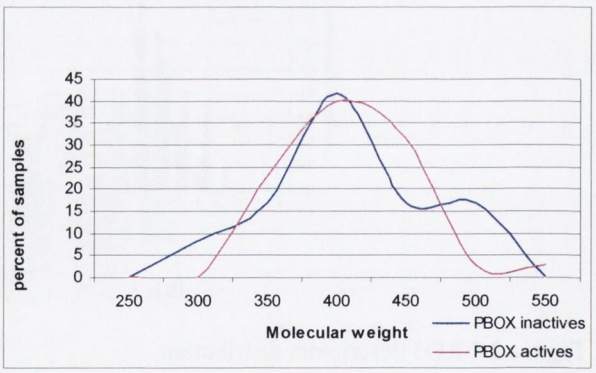
In conclusion, none of these filter parameters appear to be applicable to designing an oncology focused library. The actual level of attrition in these circumstances is significant when the diversity of the active molecules is considered in 3D space; this is not simply a matter of 'nasty' groups removing cancer actives from the search space. The implication when looking for cancer actives in prefiltered compound collections is immediately clear.

2.6 Filtering PBOX compounds

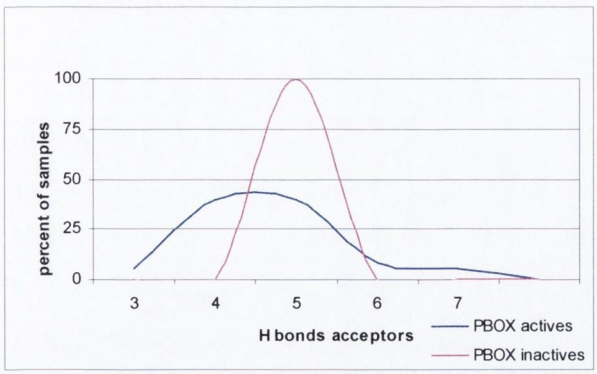
RO5 descriptors was similarly calculated for 85 PBOX compounds which display a mixed actives /inactives profile. Figure 2.8 clearly shows that the quasi totality of PBOX compounds adhere to the RO5.



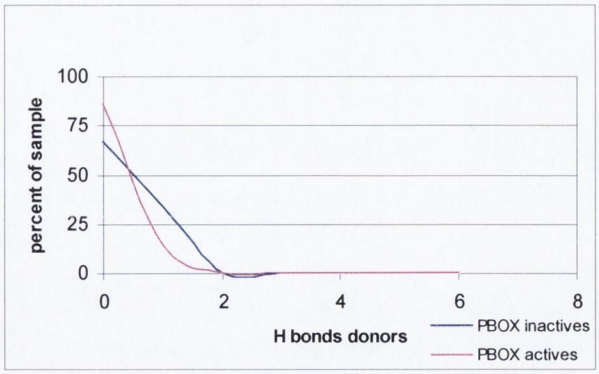
(a)



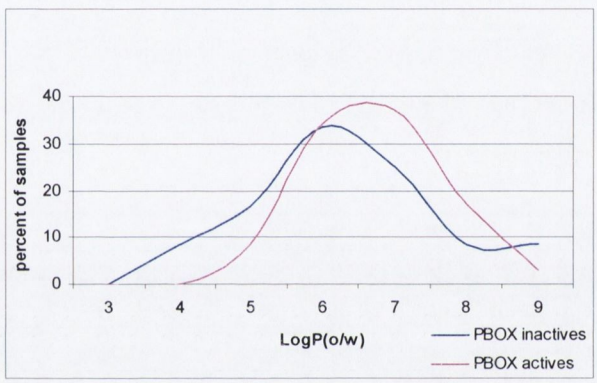
(b)



(c)



(d)



(e)

Figure 2,8: PBOX RO5 descriptor distributions

Active PBOX in pink and inactive PBOX in blue.

(a) Flexible bond distribution, (b) Molecular weight distribution, (c) Hydrogen bond acceptor distribution, (d)Hydrogen bond donor distribution. (e)Lop(o/w) distribution

PBOX compounds are consequently drug-like and would not be rejected by a drug-like filter. They are in a region of chemical space which is common to oncology space, tubulin space and also 'drug-like' space. This overlap in space translates to the possibility of designing improved PBOX compounds possessing drug-like properties which can be used in the arsenal of cancer therapeutics.

2.7 Conclusion

Cancer medicinal chemical space is far broader than just hit space or orally available drug space and, although it shares common areas to these spaces, it has unique untapped pockets still ripe for exploration. To explore cancer space, drug designers must bear in mind that cancer medicinal chemistry space is not simply a subset of hit- or drug-like space and application of ubiquitous rules and generic filters in these instances will seriously limit the realm of exploration, particularly when dealing with novel targets in the earliest phases of discovery, perhaps to the detriment of the discovery program underway.

In addition, this preliminary work illustrates that cancer active compounds are not within the commonly used drug-like space. Using filters based on rules for "drug-like" or "lead-like" compounds may not be appropriate in the context of our initial leukaemia research.

In the next chapter, a predictive model will be built utilising PBOX compounds to rationalise their activities and for the purposes of scaffold hopping.

References

1. Hou, T.; Xu, X., Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* **2004,** 10, (9), 1011-33.
2. Desany, B.; Zhang, Z., Bioinformatics and cancer target discovery. *Drug Discov Today* **2004,** 9, (18), 795-802.
3. Wong, C. F.; Guminski, A.; Saunders, N. A.; Burgess, A. J., Exploiting novel cell cycle targets in the development of anticancer agents. *Curr Cancer Drug Targets* **2005,** 5, (2), 85-102.
4. Martinez, R.; Chacon-Garcia, L., The search of DNA-intercalators as antitumoral drugs: what it worked and what did not work. *Curr Med Chem* **2005,** 12, (2), 127-51.
5. Piazza, F.; Semenzato, G., Molecular therapeutic approaches to acute myeloid leukemia: targeting aberrant chromatin dynamics and signal transduction. *Expert Rev Anticancer Ther* **2004,** 4, (3), 387-400.
6. McGuire, J. J., Anticancer antifolates: current status and future directions. *Curr Pharm Des* **2003,** 9, (31), 2593-613.
7. Brathe, A.; Gundersen, L. L.; Nissen-Meyer, J.; Rise, F.; Spilsberg, B., Cytotoxic activity of 6-alkynyl- and 6-alkenylpurines. *Bioorg Med Chem Lett* **2003,** 13, (5), 877-80.
8. Izbicka, E.; Tolcher, A. W., Development of novel alkylating drugs as anticancer agents. *Curr Opin Investig Drugs* **2004,** 5, (6), 587-91.
9. Brana, M. F.; Cacho, M.; Gradillas, A.; de Pascual-Teresa, B.; Ramos, A., Intercalators as anticancer drugs. *Curr Pharm Des* **2001,** 7, (17), 1745-80.
10. Larsen, A. K.; Escargueil, A. E.; Skladanowski, A., Catalytic topoisomerase II inhibitors in cancer therapy. *Pharmacol Ther* **2003,** 99, (2), 167-81.
11. Benvenuti, S.; Arena, S.; Bardelli, A., Identification of cancer genes by mutational profiling of tumor genomes. *FEBS Lett* **2005,** 579, (8), 1884-90.
12. Saglio, G.; Morotti, A.; Mattioli, G.; Messa, E.; Giugliano, E.; Volpe, G.; Rege-Cambrin, G.; Cilloni, D., Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia. *Ann N Y Acad Sci* **2004,** 1028, 423-31.
13. Druker, B. J., Imatinib as a paradigm of targeted therapies. *Adv Cancer Res* **2004,** 91, 1-30.
14. Jordan, M. A., Mechanism of action of antitumor drugs that interact with microtubules and tubulin. *Curr Med Chem Anticancer Agents* **2002,** 2, (1), 1-17.
15. Mitsiades, C. S.; Mitsiades, N.; Hideshima, T.; Richardson, P. G.; Anderson, K. C., Proteasome inhibition as a therapeutic strategy for hematologic malignancies. *Expert Rev Anticancer Ther* **2005,** 5, (3), 465-76.
16. Meegan, M. J.; Lloyd, D. G., Advances in the science of estrogen receptor modulation. *Curr Med Chem* **2003,** 10, (3), 181-210.
17. Balk, S. P., Androgen receptor as a target in androgen-independent prostate cancer. *Urology* **2002,** 60, (3 Suppl 1), 132-8; discussion 138-9.
18. Narashimamurthy, J.; Rao, A. R.; Sastry, G. N., Aromatase inhibitors: a new paradigm in breast cancer treatment. *Curr Med Chem Anticancer Agents* **2004,** 4, (6), 523-34.
19. Adams, D., The hitch-hiker's guide to the galaxy. *Pan Books* **1979**.
20. Oprea, T. I.; Gottfries, J., Chemography: the art of navigating in chemical space. *J Comb Chem* **2001,** 3, (2), 157-66.
21. Oprea, T. I., Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* **2000,** 14, (3), 251-64.
22. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** 7, (20), 1047-55.
23. Mestres, J., Virtual screening: a real screening complement to high-throughput screening. *Biochem Soc Trans* **2002,** 30, (4), 797-9.
24. 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 Deliv Rev* **2001,** 46, (1-3), 3-26.
25. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **2003,** 8, (19), 876-7.
26. Fattori, D., Molecular recognition: the fragment approach in lead generation. *Drug Discov Today* **2004,** 9, (5), 229-38.

27. Hann, M. M.; Oprea, T. I., Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **2004,** 8, (3), 255-63.
28. Walters, W. P.; Murcko, M. A., Prediction of 'drug-likeness'. *Adv Drug Deliv Rev* **2002,** 54, (3), 255-71.
29. Lipinski, C. A., Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* **2000,** 44, (1), 235-49.
30. WOMBAT database, Sunset Molecular Discovery LLC,. *Santa Fe, New Mexico, http://www.sunsetmolecular.com/,* **2004.**
31. Olah M, M. M., Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI In: Oprea TI (ed) WOMBAT: world of molecular bioactivity, in chemoinformatics in drug discovery. *Wiley-VCH, New York,* **2004**, 223–239
32. Irwin, J. J.; Shoichet, B. K., ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005,** 45, (1), 177-82.
33. Harper, G.; Pickett, S. D.; Green, D. V., Design of a compound screening collection for use in high throughput screening. *Comb Chem High Throughput Screen* **2004,** 7, (1), 63-70.
34. Haggarty, S. J.; Clemons, P. A.; Wong, J. C.; Schreiber, S. L., Mapping chemical space using molecular descriptors and chemical genetics: deacetylase inhibitors. *Comb Chem High Throughput Screen* **2004,** 7, (7), 669-76.
35. Wang, A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.,* **1997,** 37 615 -621, .
36. Chen, G.; Zheng, S.; Luo, X.; Shen, J.; Zhu, W.; Liu, H.; Gui, C.; Zhang, J.; Zheng, M.; Puah, C. M.; Chen, K.; Jiang, H., Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J Comb Chem* **2005,** 7, (3), 398-406.
37. Molecular Operating Environment, Software developed by Chemical Computing Group Inc. *www.chemcomp.com*.

Chapter III

Virtual Screening protocol for PBOX scaffold hopping

Introduction

PBOX compounds have been suggested as potential antineoplastic agents because they exhibit potent apoptotic activity in particular on Leukaemia cell lines[1]. They have been confirmed to bind tubulin protein yet their exact binding site remains undetermined[2]. Indeed, as PBOX compounds were firstly developed to bind the PBR protein they certainly exhibit a binding affinity for the PBR, nonetheless their apoptotic activity did not show any correlation with this binding[3]. Undertaking a scaffold hopfping approach via virtual screening (VS) could lead to the efficient removal of PBR binding responsible for side effects in the context of chemotherapy. Thus, this negation could enhance the apoptotic activity and facilitate the delineation of their actual apoptotic mechanism. When no receptor information is at hand, a structure based virtual screening approach (SBVS) is not feasible. Then, the key to VS is to engage a process whereby one examines all available compound data. The Ligand Based Virtual Screening (LBVS) alternative was therefore undertaken in order to screen for scaffold hopping.

The success of a VS campaign depends on the accuracy of the method employed to assign scores to each molecules processed. The end result of a VS is a ranked database where the molecules predicted to have the highest activity are displayed at the top of the list. There are multiple variations of the general course used to reach this goal. Unfortunately, the search for new active molecules having similar activities to the PBOX compounds presented a couple of difficulties complicating the LBVS. At the time this work was undertaken, a number of PBOX analogues had been developed and synthesised but only a small fraction of them had biological data assigned, restricting the training set selection and analogously the formation of a validation set. Hence, the choice of methods and tools in the modelling protocol was strongly constrained by the restricted starting data.

Figure 3.1 illustrates the protocol followed during the present work. The workflow was split into two parts. The first part, denoted VS1, consisted of the primary LBVS screen whereas the second part, denoted VS2, consisted of a secondary, more focused screen using refined models. While more PBOX compounds were being biologically assessed, the VS1 protocol, consisting of a primary shape search with electrostatic refinement, was undertaken. In order to validate this protocol *in silico* a "haystack" was elaborated as an alternative to the missing classical testing set. A 3D quantitative structure-activity relationship (3D QSAR) model *i.e* a pharmacophore was also deployed and used in parallel to the shape protocol with the freshly evaluated PBOX bioactivities. The activities of the newly compounds identified, using the shape based approach, were biologically assessed and estimated using the 3D pharmacophore model. The second screen, VS2, consisted of a more elaborate pharmacophore that included the new identified and biologically confirmed hits in its validation set. VS2 was subsequently

used to screen analogues of the new identified hits, novel hits, and kept for further studies presented in Chapter 4 that aimed to identify a binding site on tubulin. With the lack of binding site structural information, Structure-Activity Relationships (SARs) extracted from the available biological data, allied with the present LBVS strategy, appeared to be the best alternative to a VS campaign.

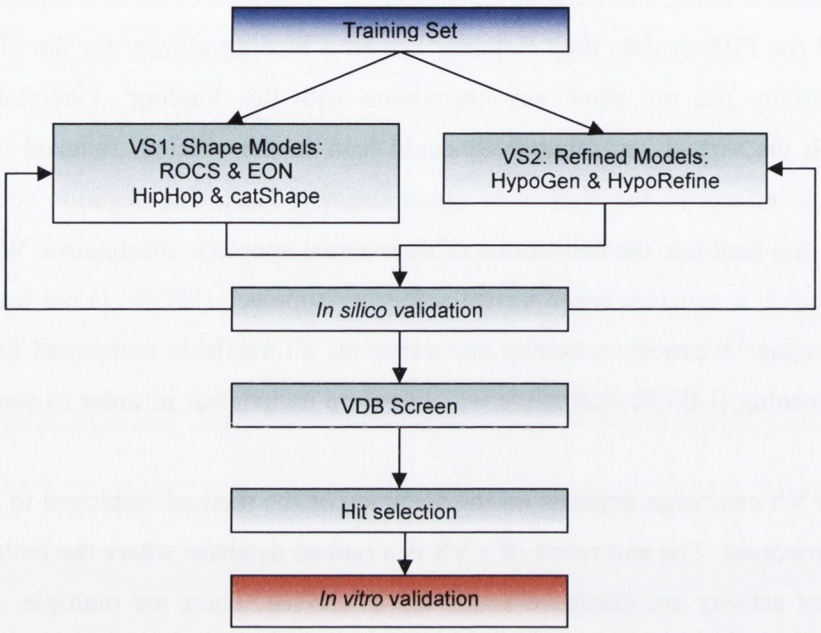


Figure 3.1: Ligand-based workflow scheme

VDB: Vendor database of molecules, *: VDB and analogue screens.

3.1 Virtual screening strategy breakdown

Both, the validation set and training set consist of the fusion of active and inactive compounds. The choice of the members populating each set is of crucial importance for the success of VS campaign. Methods to select these compounds are extensively discussed in the literature[4, 5]. In the present work, the quantity and structural diversity selection were restricted by the small fraction of PBOX compounds with known activity and high similarity. Therefore, it was necessary to:

1) Select PBOX compounds with activities rigorously checked to train and test the LBVS model and 2) Complete the lesser validation set using an appropriate method to select decoys (playing the role of inactive compounds).

3.1.1 Training sets structures selection

Prior to the initialisation of the modelling study, the available set of PBOX compounds was submitted for *in vitro* testing to assess their apoptotic potency in breast cancer cell line. Table 3.1 shows a summary of the data gathered:

Training set	Name	No. compounds	Actives	Inactives
1	Test 1	52	19	33
2	Test 2	49	19	30
3	Confident set	18	12	6

Table 3.1: Summary of PBOX experimental results.

The first test was a qualitative evaluation of apoptosis determined by co-workers, whereas the second test was a more quantitative classification of the compounds' influence on cell division. The complete bio-protocol used to determine these latter activities is presented in chapter 5 and employed a novel high content screening technology. Importantly, either a subjective or a statistical source of error was inherent in each test and consequently in the activity attributed to the molecules.

For the elaboration of haystacks, all compounds tested in the present work by test 2 were used on the basis of a preliminary binary classification of active or inactive. All methods of haystack determination used a total of 49 compounds of which 19 active and 30 inactive PBOX completing a reasonably large set.

The selection of the training set structures is of high importance in VS as they guide the computational models from a very early stage. In particular, compounds emphasizing activity will strongly direct the building of the model by highly weighing particular features. In test 2, the correlation factor associated with each IC50 value evaluated was used to determine each values accuracy. This parameter is explained in more details in Chapter 5. As 3D-QSAR models employed in the present chapter used IC_{50} values for input, to reduce the negative impact of inaccurate or variant data, it was decided to retain those compounds that showed IC_{50} correlation factors of at least 0.5 and where both tests agreed on their bio-activities. This filtering resulted in a subset of 18 PBOX compounds, of which 12 were active and 6 were inactive. This appeared to be the best solution and the final subset was denoted as a confident set. Table 3.2 illustrates the structures contained in the confident training set.

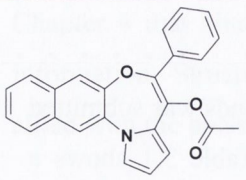
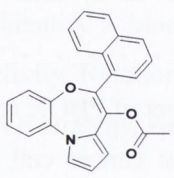
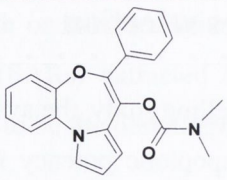
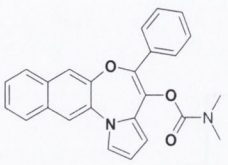
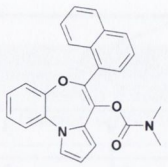
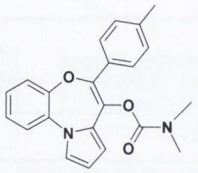
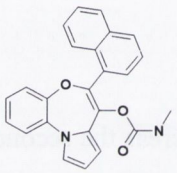
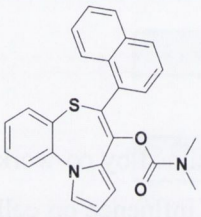
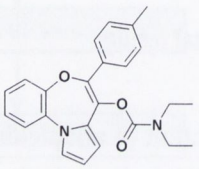
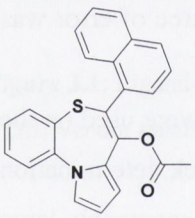
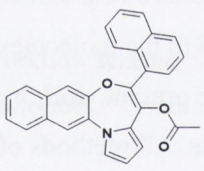
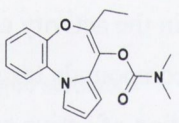
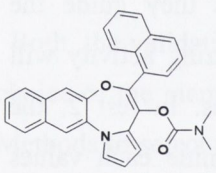
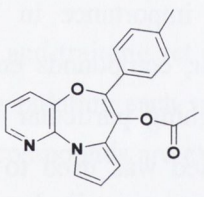
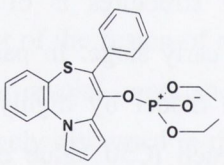
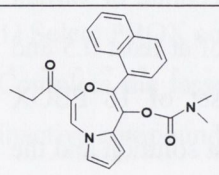
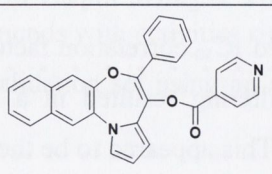
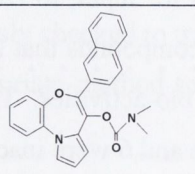
Training set actives (IC50 in μM)		Training set inactives
		
PBOX 3 (4.8)	PBOX 4 (4.0)	PBOX 1
		
PBOX 5 (3.4)	PBOX 6 (3.0)	PBOX 2
		
PBOX 7 (4.0)	PBOX 8 (3.1)	PBOX 21
		
PBOX 9 (15.7)	PBOX 15 (1.4)	PBOX 31
		
PBOX 16 (2.6)	PBOX 68 (80.7)	PBOX 38
		
PBOX 70 (3.0)	PBOX 71 (3.0)	PBOX 42

Table 3.2: Pharmacophore training sets.

First two columns illustrate PBOX actives and the third column PBOX inactives used for the training sets..

On the basis of this training set a first SAR study could be undertaken. For more clarity, the PBOX core is represented in Figure 3.2.

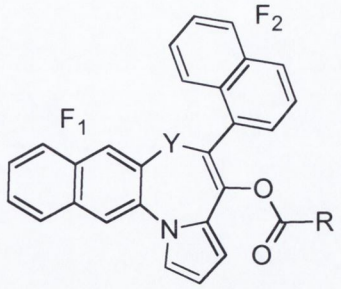


Figure 3.2: Pyrrolo-1,5-BenzOXazepines (PBOX) general structure.
Analogues present modification of rings (F_1 and F_2), central atom Y and the side chain R.

SARs

Compounds such as PBOX4 and PBOX9 presented the lowest correlation (R^2=0.5) and PBOX68 exhibited a considerable error (see Chapter 5). Hence, these compounds were not taken into account for further consideration. Focussing on the Y atom position of PBOX6 (3μM) and PBOX8 (3.1μM), it is evident that although they differ by the incorporation of either an oxygen or a sulphur atom, both have comparable activities. Therefore, the choice of Y atoms in this case appears to be negligible with respect to activity. Importantly, it is also shown that incorporation of a diphenyl ring rather than a phenyl ring (F1) has an additive effect on activity. This effect is nicely observed in the case of PBOX16 and PBOX6 whose activities are 2.6μM and 3μM respectively. Comparing PBOX15 (1.4μM) with PBOX3 (4.8μM), and also PBOX16 (2.6μM) with PBOX5 (3.4μM), the only difference is the inclusion of a diphenyl ring instead of a phenyl ring at position F2 illustrated. Several notable differences are apparent when considering the side chain R.

1) Both PBOX5 (3.4μM) compared with PBOX3 (4.8μM) and PBOX15 (1.4μM) compared with PBOX16 (2.6μM) differ by $N(CH_2)_2$ and a methyl group. Thus, it was concluded that the $N(CH_2)_2$ and methyl were equivalent.

2) PBOX71 (3μM) and PBOX3 (4.8μM) differ respectively from a pyridine and a methyl and it was concluded that the pyridine bulk had the effect of reducing activity.

The overall conclusion can be summarised as follows: Hydrogen bond acceptor on Y (not essential), hydrophobic aromatic on F1 and F2 and small bulk on R pertain to optimal activity.

3.1.2 Haystack generation

An ideal haystack should contain actives and verified inactives, yet the gathering of a large enough ensemble for *in silico* validation purposes was, as usual in academia, restricted by the availability of data. The selection of decoy compounds has now become a common use as an alternative. It has been shown by several studies that a random selection of these compounds could bias the validation results[6, 7]. Indeed, compounds completing the validation set should be similar enough to the training set compounds. And this, especially in the case of PBOX compounds where there is no structural diversity in the training set. In other words, if the model has been trained to distinguish apples and oranges, measuring its performance on separating pears and grapes will not be appropriate. The following section demonstrates how to select the most appropriate compounds which would mimic the inactives. Chemoinformatics can provide several concepts and tools to find relationships between molecules and define similarity. In the present work, three different techniques were used, all tending to describe a particular similarity. Firstly, a statistical analysis of generic 2D descriptors traditionally used in similarity was carried out, in addition a principal component analysis based on a larger set of 2D descriptors was employed to maximize the similarity between molecules. Finally a fingerprint similarity was express using a Bayesian statistical approach.

Filter elaboration

In addition to the set of descriptors on which the "rule of three"[8] and "rule of five"[9] are based (discussed in Chapter 2), the use of supplementary descriptors have been suggested in the literature.[10] The filter elaborated here was based on descriptors' thresholds calculated to suit PBOX compounds. Therefore databases of decoy molecules were filtered, retaining molecules fitting these thresholds for the following properties: lipophilicity, molecular weight, number of hydrogen bond donor and acceptors, and number of rotatable bonds and total polar surface area and the van der Waals area. PBOX similar compounds were selected while adhering to these PBOX rules and were expected to have some level of similarity, at least, among these properties.

Principal component analysis

In order to extend the previously generic set of descriptors, a principal component analysis was carried out. This projection method finds information in databases of multiple molecules with large number of descriptors associated (data table) by projecting (reducing) these molecules to a few components. These components provide a very good summary of the data table and plotting one

component against another illustrates this in an understandable manner. The coefficients of the projection, *i.e.* how the descriptors are combined to form the components, are called loadings or weights. Plotting the weights can show the relative importance of the descriptors, their similarity and their dependence. To do so, the technique uses multivariate analysis and linear regression. In this work, it was applied to PBOX compounds associated to more than a hundred descriptors depicting size, shape, path, physical and chemical properties. The advantage of this method over the previous one, is that each of the selected descriptors contribute to the explanation of PBOX activity. This technique automatically attributes an appropriate weight to each descriptor in order to best explain the maximal variance between the active and inactive PBOX. If successfully used, the method can separate actives from inactives in the reduced space and give information on where differences lie. This method offers the possibility to detect outliers and check the bio-data used such as PBOX compounds wrongly tagged actives or inactives. Analogously the analysis of the descriptors can provide information about their validity for the present study. Furthermore, using this method to project a new set of molecules using the same settings would enable a graphical, easy way to select molecules close to PBOX and hence similar.

Bayesian model

To classify PBOX molecules among decoys, a naïve Bayesian classifier was used as an alternative approach to modelling the difference between active and inactive PBOX compounds. The naïve Bayesian classifier can be used as a simple and accurate machine-learning method [11]. A Bayesian classifier generates a model with an output that reflects the relative likelihood for a molecule to be active. A detailed statistical explanation of this approach, as applied to molecular structural recognition, was presented by Bender et al.[12]. Briefly, Bayesian statistics compares the frequency of occurrence of features contained in the active molecules against the overall frequency across the whole training set. Inactive molecules are processed as well and their statistics provide a baseline. The model weights the input features, giving more weight to features that differentiate between the active and inactive classes. Consequently, the Bayesian classifier scores each feature in decreasing order (from more likely active to inactive) and predicts the class (active or inactive) that a new feature is likely to belong to. Extrapolating features to molecules, the overall molecular Bayesian score permits the prediction of activity classes for new molecules – the more negative the score is, the most likely the molecule is to be inactive. If applied to a larger decoy database, one would expect to find, at the top of the hit list, compounds similar to PBOX. To complete the previous PCA approach, here similarity was explored from a structural point of view rather than a property one, with the use of path-based fingerprint.

Validation

To elaborate the molecular haystack it was necessary to select inactive compounds. To ensure inactivity of the selected compounds the two databases of known ligands used in Chapter 2 were analysed. Compounds were selected from the cancer inactive sub-ensemble from NCI (National Cancer Institute) database (as described in Chapter 2) and from WOMBAT (WOrld of Molecular BioAcTivity Oprea, T.I. ed.).non tubulin and non PBR binders. In addition, compounds similar to PNOX compounds were taken from SPECS[13] to complete the validation set. Finally, to evaluate the quality of the Bayesian model and of all three dimensional models Receiver Operating Characteristic plots (ROC plots) were calculated which takes into account both inactive and active rates.

3.1.3 Pre-processing compounds for VS

Prior to any modelling and screening, it was essential to obtain adequate three dimensional representations of the molecules. In addition, for any 3D approach, performance relies on the quality of ligand flexibility representation and the accurate representation of the bio-active conformation. Therefore, each molecule was submitted to conformational sampling. Three conformer generators were considered in the work, catFAST and catBEST from Catalyst[14] and OMEGA2 from openeye[15]. These three methods differ from their respective sampling method as explained in Chapter1. OMEGA2 has been demonstrated to be extremely fast. However, as again the representation of the training set is of main importance, the three generators efficiencies were compared for PBOX conformers sampling.

3.1.4 VS1- the shape approach

Shape and electrostatic model

VS1 utilized as a template PBOX16, which is the biggest of the most active compounds in order to search a vendor database. To accomplish this, Rapid Overlay of Chemical Structures[16] (ROCS) was used. The first step, consisted in aligning all molecules conformers on the PBOX16 conformers and evaluating all overlaps according to which molecules were scored and ranked. This method of screening has been proved to be efficient in the literature[17] in particular for chemotype switching[18]. This tool offers the possibility to refine the shape match using color force fields which can be of the following types: hydrogen-bond donors, hydrogen-bond acceptors, hydrophobes, anions, cations, and rings. This leads to scores that mirror the features matching used in pharmacophore matching programs. Following this, EON[19] was applied to refine the final alignment using electrostatic

mappings. This consists of systematically rotating terminal rotatable bonds until an optimum electrostatic overlap is found. Subsequently, all overlaps are re-ranked.

Activity estimation

Another approach for LBVS was also carried out using pharmacophore modelling. Using the HipHop algorithm a module of Catalyst[14] pharmacophores were generated with only common chemical features from the training set of PBOX active molecules. This algorithm does not require activity value associated to PBOX actives to determine three-dimensional chemical features arrangement in space. HipHop has largely been utilized in the literature successfully[20]. The pharmacophores generated are scored according to the degree to which one is common to all actives and the estimated rarity of the pharmacophore itself[21]. Furthermore, the PBOX16 shape can be added as a constraint for shape matching using the catShape algorithm of Catalyst. Finally, the pharmacophore evaluated as the best can be regressed by a linear structure-activity regression equation which uses the activity fields of the training set. In this way the pharmacophore elaborated is able to predict the activity of novel compounds such as hits subsequent to the ROCS-EON screen.

3.1.5 VS2- Refined pharmacophore generation

The refined pharmacophores elaborated for VS2 used HypoGen and HypoRefine technology from Catalyst[14]. Both methods are able to generate pharmacophores in which each feature is associated to a weight. This translates the importance of each feature to describe the best elements of activity of compounds in the training set. Indeed, such pharmacophores are able to estimate activity of screened compounds. Contrary to HipHop, these algorithms require input of at least a number of 16 PBOX compounds with associated activities. In addition, HypoRefine can also take into account the molecular shapes of inactives and distribute exclusion volumes in space. These are regions in space which are biologically unfavourable if populated. While the database is searched for hits, molecules are not allowed to overlap exclusion volumes.

3.2 Computational methods

3.2.1 Descriptors calculation

2D molecular descriptors and numerical properties that can be calculated from the connection table of a molecule, and therefore are not dependent on its 3D conformation. All descriptors used in the 2D study of the PBOX compounds were calculated using the MOE[22] QSAR descriptor module. Table 3.3 provides a list of those descriptors that are referred to in the results section.

Name	Description
a_count	Number of atoms (including implicit hydrogens). This is calculated as the sum of $(1 + h_i)$ over all non-trivial atoms i.
a_nC	Number of carbon atoms: $(Z_i \mid Z_i = 6)$.
b_count	Number of bonds (including implicit hydrogens). This is calculated as the sum of $(d_i/2 + h_i)$ over all non-trivial atoms i.
a_aro	Number of aromatic atoms.
b_ar	Number of aromatic bonds.
b_heavy	Number of bonds between heavy atoms.
b_rotN	Number of rotatable bonds. A bond is rotatable if it has order 1, is not in a ring, and has at least two heavy neighbors.
lip_acc	The number of O and N atoms.
lip_don	The number of OH and NH atoms.
opr_brigid	The number of rigid bonds bonds[23].
TPSA	Polar surface area (Å^2) calculated using group contributions to approximate the polar surface area from connection table information only. The parameterization is that of Ertl *et al.*[24]
vdw_area	Area of van der Waals surface (Å^2) calculated using a connection table approximation.
vsa_hyd	Approximation to the sum of VDW surface areas of hydrophobic atoms (Å^2).
Weight	Molecular weight (including implicit hydrogens) in atomic mass units[25].
logP(o/w)	Log of the octanol/water partition coefficient

	r^2 = 0.931, RMSE=0.393 on 1,827 molecules.
logS	Log of the aqueous solubility (mol/L). This property is calculated from an atom contribution linear atom type model[27] with r^2 = 0.90, ~1,200 molecules.
apol	Sum of the atomic polarizabilities (including implicit hydrogens)[25].
mr	Molecular refractivity (including implicit hydrogens). This property is calculated from an 11 descriptor linear model[28] with r^2 = 0.997, RMSE = 0.168 on 1,947 small molecules.
chi0	Atomic connectivity index (order 0) from[29, 30]. This is calculated as the sum of 1/sqrt(d_i) over all heavy atoms i with $d_i > 0$.
chi0v	Atomic valence connectivity index (order 0)[29, 30]. This is calculated as the sum of 1/sqrt(v_i) over all heavy atoms i with $v_i > 0$.
chi0v_C	Carbon valence connectivity index (order 0). This is calculated as the sum of 1/sqrt(v_i) over all carbon atoms i with $v_i > 0$.
chi1	Atomic connectivity index (order 1)[29, 30]. This is calculated as the sum of 1/sqrt($d_i d_j$) over all bonds between heavy atoms i and j where $i < j$.
Kier1	First kappa shape index[30]: $(n\text{-}1)^2 / m^2$.
Kier2	Second kappa shape index[30]: $(n\text{-}1)^2 / m^2$.
Kier3	Third kappa shape index[30]: $(n\text{-}1)\ (n\text{-}3)^2 / p_3^2$ for odd n, and $(n\text{-}3)\ (n\text{-}2)^2 / p_3^2$ for even n
KierA1	First alpha modified shape index[30]: $s\ (s\text{-}1)^2 / m^2$ where $s = n + a$.
KierA2	Second alpha modified shape index[30]: $s\ (s\text{-}1)^2 / m^2$ where $s = n + a$.
KierA3	Third alpha modified shape index[30]: $(n\text{-}1)\ (n\text{-}3)^2 / p_3^2$ for odd n, and $(n\text{-}3)\ (n\text{-}2)^2 / p_3^2$ for even n where $s = n + a$.
balabanJ	Balaban's connectivity topological index[31] .
VDistMa	If m is the sum of the distance matrix entries then VDistMa is defined to be the sum of $\log_2 m$ - $D_{ij} \log_2 D_{ij} / m$ over all i and j
VAdjMa	Vertex adjacency information (magnitude):

	$1 + \log_2 m$ where m is the number of heavy-heavy bonds. If m is zero, then zero is returned.
VAdjEq	Vertex adjacency information (equality): $-(1-f)\log_2(1-f) - f \log_2 f$ where $f = (n^2 - m) / n^2$, n is the number of heavy atoms and m is the number of heavy-heavy bonds. If f is not in the open interval (0,1), then 0 is returned.
VDistEq	If m is the sum of the distance matrix entries then VdistEq is defined to be the sum of $\log_2 m - p_i \log_2 p_i / m$ where p_i is the number of distance matrix entries equal to i.
weinerPol	Wiener polarity number: half the sum of all the distance matrix entries with a value of 3 as defined in [32]
weinerPath	Wiener path number: half the sum of all the distance matrix entries[32, 33].
zagreb	Zagreb index: the sum of d_i^2 over all heavy atoms i.

Table 3.3: MOE 2D main descriptors calculated.

3.2.2 Evaluation metrics

In the present study, measurement of model quality was calculated using the Receiver Operating Characteristic (ROC) technique and calculating ROC scores with component from Pipeline Pilot[34] (PP).

3.2.3 Principal Component Analysis

The Principal Component (PC) model was built starting from an initial set of 49 PBOX compounds, classified as actives, inactive and intermidiates as followed:

- 12 active PBOX numbered [3, 4, 5, 6, 7, 8, 9, 15, 16, 68, 70, 71]
- 7 intermediate PBOX numbered PBOX [32, 43, 46, 66, 67, 72, 78]
- 30 inactive PBOX numbered [1, 2, 12, 23, 24, 25, 26, 27, 31, 33, 34, 35, 38, 42, 48, 65a, 65b, 69, 73, 74, 75, 79, 80, 81, 82, 83, 84, 85, 86]

All structures can be found in appendix' and detail on their activity categorisation is presented in Chapter 5. The PC model and the Principal Component Analysis (PCA) were performed using SIMCA-P package[35], iterating through the following steps:

1. Application of PCA to reduce the dimensional space of the descriptors, starting from 150 2D descriptors. No limit on the number of components was set.
2. Projection of all PBOX compounds onto the first two PC's to evaluate the quality of the separation between actives and inactives.
3. Evaluation of PBOX outliers by calculation of the residual standard deviation (RSD) of all PBOX in the new PC space. Compounds exceeding a critical limit were considered to be outliers. The Hoteling range (confidence interval) was kept at the default value of 95% equivalent to a critical distance of 0.05.
4. Analysis of the loadings plot illustrating the descriptor weights for the first two PC's and their correlations. Descriptors giving a similar explanation of the data (PBOX compounds) were made unique, therefore reducing the total number of descriptors before the next iteration of PCA.

After several iterations, a total of 34 descriptors remained and a total of 20 PBOX compounds of which 11 were actives and 9 inactives. PBOX actives [3, 4, 5, 6, 7, 8, 9, 15, 16, 68, 71] and PBOX inactives [1, 2, 21, 27, 73, 79, 83, 84, 86]. The iterative process reduced the number of components before PCA focused attention on those descriptors that seem to be of most relevance. Dimensionality was finally reduced to two PC components, which reasonably represented the variance of the dataset, as illustrated in Table 3.4.

Dataset	R^2	R^2 cumulative
PBOX	0.87	0.94
WOMBAT	0.76	0.88
NCI	0.66	0.76
SPECS	0.74	0.82

Table 3.4: Variation of the PCA model along the diverse datasets.

R^2 is the fraction of the variation of the compounds in the dataset explained by the first PC, and R^2 cumulative is that explained by the two first PC's.

R^2 is a measure of how well the model fits the data - the higher the value the better the model. In addition, molecules from the WOMBAT (120488 compounds) and NCI (41086 compounds) databases were evaluated with the same set of 34 descriptors to probe the model surface and from SPECS[13] to complete the haystack molecular diversity. PBOX like space was described by defining the smallest circle enclosing all of PBOX actives in the space defined by the two first PC's. Compounds from the databases were considered to be similar to the PBOX's if they lay within this circle.

3.2.4 Fingerprint calculations

MDL fingerprints were used to gauge the similarity between the PBOX compounds and the specific database subset identified by HipHop[14] pharmacophore selection of compounds. The similarity between two fingerprints was measured using a traditional Tanimoto coefficient. MDL public keys are path-based fingerprints that were first developed for rapid substructural searching of ISIS databases[36]. They maximally consist of a set of 960 features, of which most are substructural. In the present work, a 166 feature version as implemented in PP[34] was used. The MDL public keys fingerprint calculated for a molecule contains a list of key numbers for features that exist in that molecule.

Extended Connectivity Fingerprints (ECFP) from PP[34] were also used to cluster molecules. In these fingerprints each bit represents the presence of a structural feature rather than a substructural feature allowing them to represent additional information. Thus, the ECFP features are generated directly from the input molecules and are not pre-selected. The ECFP generation method is based on the Morgan algorithm[37], which is also generally used for the generation of canonical molecular representations. ECFP generation method lies firstly in the use of a hashing scheme to generate codes comparable across molecules and secondly in the use of an iteration threshold that limits the number of bonds included in a feature. Each iteration describes a node (atom) and its neighbours (connected atoms) Each feature is defined as a specific ensemble of atoms and bonds and the size of the feature (molecular substructure) can be described as the maximal bond path length. In the present work the substructure size was set to four-that is to say each feature was describing a substructure of a maximum of four consecutive bonds.

3.2.5 Bayesian model

In this work, the initial set of 49 PBOX compounds from test 2 was split into two subsets a training set of 33 compounds and a testing set of 16 compounds. The Bayesian learner was trained using 13 active and 20 inactive PBOX. Subsequently, the model was tested using 6 active and 10 inactive PBOX. In an attempt to improve the baseline when building the model a total random set of a 10 000 decoy molecules, randomly and evenly selected from SPECS, WOMBAT and NCI was also added to both training and validation sets. The feature on which a Bayesian model is built can be a string, a fingerprint, a boolean, an integer, floating point property such as the molecular weight, the number of hydrogen bond donors/acceptors, the number of rotatable bonds etc. In the present work MDL public key fingerprints were used. This fingerprint is a long integer value. The Bayesian method bins each feature present in the input molecules, so that finally each bin contains the number of occurrences of a given feature across the input set. In the present PBOX model post-processing was performed to

remove low-information bins. Low-information bins were defined as those that had: normalized estimates in the range [-0.05, 0.05]. The original number of bins was 163; none was removed due to too few samples but the number removed due to a poor normalized estimate was 10. The final number of bins saved in the model was therefore 153. Cross validation was applied using the Leave One Out (LOO) method, and the final model obtained a ROC score of 0.994. The final validation of this model is presented in the results section below.

3.2.6 Molecular clustering

To cluster the Bayesian hit list extracted from the WOMBAT database, the clustering component from PP[34] was employed. In this component, a number of representative objects (cluster centres) are chosen from the data set. The first of these is selected randomly. Subsequent cluster centres are chosen using a maximum dissimilarity method, in which each successive centre is the record maximally distant from the centres that have already been selected. The distance function between the objects can be a Euclidian distance, a Tanimoto distance or a combination of the two. The process is repeated until there are a predefined number of cluster centres. The non-selected objects are then assigned to the nearest cluster centre to determine the cluster membership.

In the current study, members of each data set (WOMBAT selected molecules per target class) was clustered with standard settings to give 3 clusters using extended connectivity fingerprints (ECFP_4) and cluster distances were measured using Tanimoto distances. Cluster centroids were then presented as representative of their clusters.

3.2.7 Ionisation and isomerisation

All compounds entering into computational models were first passed through a PP[34] component ensuring ionisation at physiological pH=7.4. This component identifies potential ionisation sites using a predefined set of molecular query and pKa values are calculated for those sites using a series of Partial Least Squares (PLS) Models.

All compounds were also passed through a PP component to enumerate all possible stereoisomers, as in many cases the stereochemistry is not correctly represented in the databases from which these molecules are obtained.

3.2.8 Conformer generation

Training and validation set conformers were generating using the Catalyst[14] catBEST algorithm whereas for molecules from the vendor database SPECS they were generated using OMEGA2[15]. For

the purpose of comparing algorithms, all PBOX compounds were also submitted to the catalyst catFAST algorithm. This conformation generation method uses one of three different algorithms; depending on the size of the molecule:

- Small molecule conformers are generated using a quasi-exhaustive systematic search in which discretized rotations were performed. The theoretical number of conformers depends on the number of rotatable bonds, the number of possible torsion values for each rotatable bond, and the number of ring conformations. A systematic search is used if the theoretical number of conformations is less than a default threshold value. Conformers with excessive van der Waals clashes are removed and the diversity of the conformational space is controlled by computing pairwise Root Mean Squared Deviations (RMSD) which removes any conformers that are very similar to those that have already been considered.
- Medium-sized molecule conformers are generated with a random search method that uses poling. Conformer generation stops when no conformers within an energy threshold are found after consecutive attempts.
- For large molecules, only one conformation is generated for each possible combination of stereocenters.

FAST stops generating conformers once the desired maximum number of conformers is generated, or the energy of any new conformer is too high according the predefined energy threshold or when there are no possible new conformers after a certain number of trials. All default values were kept unchanged in this work.

The BEST algorithm generates conformers by performing a more rigorous energy minimization in both torsional and Cartesian space and by using a poling algorithm. The later promotes conformational variation as the function being minimized is modified to force similar conformers away from each other. The addition of poling eliminates much of the redundancy in conformer generation and improves the coverage of conformational space[38]. All default values were kept unchanged in this work, so that all PBOX compounds including the training and validation set were submitted to internal strain energy minimization and conformational analysis (max. number of conformers = 250, generation type: best quality, energy range = 20 kcal/mol above the calculated global minimum).

The OMEGA[15] software has been previously shown by many groups, including our own, to generate diverse and sensible 3D conformations of molecules [39]. OMEGA conformer generation proceed in steps:

1. Fragment molecules on their sigma bonds.
2. Reassembles fragments in three dimensional space, using a constraint library or by generating fragment conformations on the fly using the same distance constraints.

3. Generate additional models by enumerating ring conformations attempting to generate every possible combination of ring conformations possible for a given structure.
4. Torsion driving search is initiated by examining the molecular graph and determining the bonds that may freely rotate.
5. The final ensemble selection is based on heavy atom RMS distance and torsions are grouped into sets of up to five contiguous rotatable bonds
6. An exhaustive, depth first torsion search is performed on each of the fragments, and the resulting conformers are placed into a list, sorted by energy.
7. Entire structures are assembled by first combining the lowest energy set of fragments and subsequently the next lowest set, until the search is terminated.

OMEGA stops when the limit on the total number of conformers is reached, when the fragment list is exhausted, or when the sum of the fragment energies exceeds the energy window of the global minimum structure. A final ensemble is selected by sequentially testing the conformers using the RMS distance cut-off.. In the present work, all default values were kept except the total number of conformers which was set to 250.

3.2.9 ROCS and EON

Additional shape matching of molecules was performed using Rapid Overlay of Chemical Structures (ROCS)[16], and subsequently EON[19] electrostatic refinement.

ROCS is a shape-based superposition method. It effects a rigid alignment of two structures by maximizing the overlap volume. To gain in computational time the volume overlap is not the hard-sphere overlap volume, but rather a Gaussian-based overlap parameterized to reproduce hard-sphere volumes. Only the heavy atoms of a structure are taken into account, hydrogens are ignored. Chemical features can be included in the superposition and similarity analysis process which facilitates the identification of those compounds that are similar both in shape and chemistry. The Chemical Force Field (CFF) can be used to measure chemical complimentarity, and to refine shape superpositions based on chemical similarity. The CFF is composed of fragmental rules that determine chemical centers, plus rules to determine how such centers interact. ROCS relies on shape matching, where two molecules will have the same shape if their volumes exactly correspond. The more the volumes differ, the more the shapes will differ. Shape is defined as a relative quantity, depending on references to other shapes. Several measures of similarity can be claculated such as :

- The Shape Tanimoto coefficient:

$$Tanimoto_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}$$

Where the ***I*** terms are the self-volume overlaps of each molecule, ***f*** and ***g***, and the *O* term is the overlap between the two Gaussian functions representing the volumes of molecules ***f*** and ***g***. The optimal value of this function is 1.

- The two shape Tversky coefficient:

$$Tversky_{f,g} = \frac{O_{f,g}}{\alpha I_f + \beta I_g}$$

Usually with this coefficient **α+ β = 1,** but ROCS sets **α = 0,95** therefore introducing an asymmetry. Therefore two Tversky coefficients are calculated; one for the query molecule (Tversky_q) and one for the database molecule (Tversky_d). This coefficient can be greater than 1.0 because the overlap between the two molecules, can be larger than a molecule's self-overlap.

- The overlap equal to the absolute value of the volume overlap between the query and the database molecule. The overlap is in arbitrary units.
- The ColorScore, which measures the chemical complimentarity by summing the single best color interaction of all possible matches between the query and database molecules. It has no limit values.
- The ScaledColor, which is equal to the ColorScore of a database molecule normalized over the ColorScore of the query molecule with its-self. Its optimal value is 1.
- The ComboScore, which is the sum of the shape Tanimoto and the scaled color score. Its optimal value is 2.

In this work, ROCS consisted either of PBOX conformers generated by BEST or database molecules output from OMEGA2. All default parameters were kept, and the default CFF was turned on. This CFF includes a simple pKa model that assumes pH=7. All cut-offs were set to 0, so that all overlays were output for the validation step (choice of ranking coefficient), and then the shape Tanimoto threshold was set to 0,75 prior to EON refinement to minimize the hit list size and gain in computational time.

EON[19] calculates the Electrostatic Tanimoto (ET) coefficient between a query molecule and a database molecule already pre aligned with ROCS, calculating partial charges on both molecules using MMFF94[40]. It reads the query molecule conformer and uses technology from OMEGA to expand terminal torsions to search for subtle changes in conformation that might increase the electrostatic overlap without changing the overall shape overlap. EON uses a Tanimoto coefficient (ET_pb) to compare electrostatic potentials calculated from an OpenEye's electrostatics function[41]. In the present

work, molecules were ranked by ET_pb coefficient using an outer dielectric of 80 (equivalent to the water dielectric constant) as usually proteins essentially act to compensate the aqueous desolvation of well-bound ligands. All other default parameters values were kept and the ROCS output was used for EON input.

3.2.10 HipHop and catShape

The HipHop algorithm within Catalyst[14], was employed to generate ten common-feature pharmacophore models for highly active PBOX compounds. Each pharmacophore contained a combination of features such as hydrogen bond donors, hydrogen bond acceptors, aromatic rings, hydrophobic and hydrophobic aromatic for all training set compounds and the "principal" column was set to a value of 2 for all actives, ensuring that all of their chemical features were considered in building the pharmacophore space. The resulting pharmacophores were described by a three-dimensional arrangement of features located at defined positions. Features are assigned spatial tolerances (spheres) for subsequent matching, and these were all kept as default values. In addition, hydrogen bond acceptors, donors and ring aromatic features include a vector indicating the direction of the interaction.

The HipHop algorithm is a module of Catalyst that generates pharmacophores with only common chemical features from a set of active molecules. Importantly, the algorithm determines three-dimensional configurations of chemical features that are common to a set of active molecules without the need for an activity value field per active. The scoring methods of the resulting pharmacophores uses standard techniques for assessing statistical hypotheses against experimental data, and is based on both the degree to which the configuration is common to all actives and the estimated rarity of the configuration itself[21]. By default, ten configurations are returned and ranked from the most statistically significant to the least.

Spatial information from compounds can also be added as steric constraints (shape features) for shape matching using the catShape algorithm of Catalyst. The hit list is reduced by eliminating compounds that do not comply with these constraints once aligned to the pharmacophore. In this work, PBOX16 was aligned to the chosen pharmacophore using the BEST algorithm and used to define a volume constraint.

Finally, the most suitable pharmacophore for PBOX study, was regressed using Catalyst. This assigned a weight to each feature so that when matching a molecule to the pharmacophore, the Fit value is defined as a geometric fit and as the sum over all features *f* as follows:

$$Fit = \Sigma\, \mathbf{w}(f) * [\, 1 - SSE(f) \,]$$

where

$\mathbf{w}(f)$ is the weight associated to the feature *f*,

and $SSE(f)$ is defined as the sum over the location constraints *c* on a feature *f* and is calculated as follows:

$$SSE(f) = \Sigma\, (\boldsymbol{D}(c) / \boldsymbol{T}(c))^2$$

where

$\boldsymbol{D}(c)$ is the displacement of the feature *f* from the center of the location constraint *c*,

and $\boldsymbol{T}(c)$ is the radius of the location constraint sphere *c* for the feature *f* denoted Tolerence $\boldsymbol{T}$.

The Fit value of a molecule is then used to predict its activity as follows:

$$-\log(\mathbf{Activity})_{Est} = Fit * Slope + Y\ interception$$

The slope and intersection correspond to the line illustrated in figure 3.3

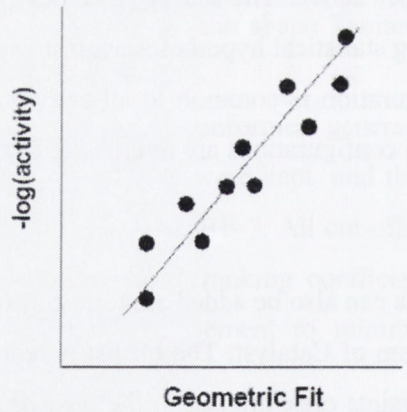


Figure 3.3: Correlation between the estimated activity and the calculated geometric Fit [42]

X axis: Geometric fit of each molecular structure to a given pharmacophore

Y axis: Estimated activities for molecular structures by a given pharmacophore

A good regression consists of attributing an appropriate weight to each feature so that the subsequent estimated activity per molecule is the closest possible to the experimental one.

3.2.11 Catalyst HypoGen and HypoRefine

In the present work, the same training set used for the HipHop algorithm was used (12 actives) with the addition of 6 inactive compounds. As PBOX compound activity range covered only 1.7 orders of magnitude, inactive compounds were artificially assigned an activity value of 90000, enhancing the range to 4.8 orders of magnitude. This larger activity range enabled to widen the predicted activity range and its accuracy. All default parameters were kept, and five runs were performed as presented in figure 3.4.

Run	Algorithm	settings
Run 1	HypoGen	Min sub set features= 4
Run 2	HypoGen	Min sub set features = 3
Run 2	HypoRefine	No. exclusion volumes = 1
Run 4	HypoRefine	No. exclusion volumes = 10
Run 5	HypoRefine	No. exclusion volumes = 100

Figure3. 4: Description of automated pharmacophore generation runs.

HypoGen and its extension HypoRefine are two other pharmacophore generation algorithms implemented in Catalyst. In contrast to HipHop, their output pharmacophores can estimate the activity of a compound by calculating the geometric fit of the molecular structure to the given pharmacophore model as previously described in the regression section. HypoGen is trained to build pharmacophores common to the active compounds without reflecting the inactive ones. By default, the ten pharmacophores best able to correlate estimated and experimental activities are output, ranked by statistical significance. Preferably, at least 16 molecules should be included in the training set, with activity values issued from the same experimental procedure. An activity range of four orders of magnitude is also by default preferable[43].

A variable weight is assigned to each feature of a pharmacophore, which is proportional to its relative contribution to biological activity. The algorithm constructs pharmacophores with optimal correlation with the experimental activity values trying to use as few features as possible. HypoGen generates the models in three steps[20]:

- The constructive step: identification of pharmacophore features common to the most active* compounds (usually the 2 most active). This proceeds by an exhaustive enumeration of all possible pharmacophores (with all the input features) that match all of the most active conformers. Coverage of a large pharmacophore space is thus ensured. Constraints on the number of pharmacophores to be generated can be controlled by a minimum subset of features parameter that specifies a minimum number of matches for the remaining set of most actives molecules.
- The subtractive step: all pharmacophore common to the least active set of molecules are removed.
- The optimization step: enhancement of the pharmacophore score. Linear regression is performed, allowing pharmacophore scoring based on errors in the estimated activities and the pharmacophore complexity. To optimize the scores, feature selection and location are varied and the "total cost" for each pharmacophore is also calculated and the optimization stops when the score stops improving, resulting in the output of the top ten unique pharmacophores.

*: Catalyst defines the most active set as those compounds meeting the following condition:

$$Activity/Uncertainty \leq Activity\ of\ most\ active\ compound \times Uncertainty$$

where the uncertainty is defined as the ratio of the maximum and minimum experimental activity value errors. The default and recommended threshold is 3.

Pharmacophores are scored according to their "cost analysis"[44]. The overall cost of a pharmacophore consists of three components:

- The weight cost: increasing in a Gaussian form as this function weights in a model deviate from its ideal value.
- The error cost: representing the root-mean-squared difference between the estimated and measured activities of the training set.
- The configuration cost: quantifying the entropy of the pharmacophore space. This should not exceed a maximum value of 17 as it corresponds to a number of 2^{17} pharmacophore models. It has been empirically determined that higher values often lead to a good correlation by chance.

In addition, fixed and null cost values were evaluated for each HypoGen run. The first of these is the minimum possible cost representing a hypothetical simplest model that fits all data perfectly. Fixed costs are calculated by adding the minimum achievable error and weight cost and the constant configuration cost. The null cost represents the maximum cost of a pharmacophore that no features and estimates all activities to be the average of the training set activity data. The null cost value is equal to the maximum possible error cost. The greater the difference between these two cost values and the

closer the total cost of a generated hypothesis is to the fixed cost, the more statistically significant that hypothesis is supposed to be. According to randomized studies, a cost difference of 40-60 between the total cost and the null cost indicates a 75-90% chance of representing a true correlation in the data.[45]

HypoRefine, the second automated pharmacophore generator used, has the additional capability to utilise exclusion volume features as spheres. This algorithm also proceeds in three steps:

1. Identical to the HypoGen contructive step.
2. A step which again enumerates all pharmacophores found from the most active compounds in the training set including excluded volumes features. Identification of excluded volume spheres is automated and performed first by aligning the active molecules to identify the Active Space (AS). Alignment of the inactive molecules with the highest fit scores is used to identify the Inactive Space(IS). Identification of those atoms in the aligned inactive compounds that are far away from those in the aligned actives is then used to determine the Candidate Space (CS), where CS = IS –AS. Finally, a random selection from the CS points determines the locations of the excluded volume spheres.
3. A final pharmacophore penalization step is accomplished by discarding models that fit both active and inactive molecules.

Pharmacophores generated with both HypoGen and HypoRefine were validated by verifying that all differences between the generated pharmacophore costs and the null costs were greater than 60 bits, which accounts for a 90% statistical probability that the generated models represents a real correlation with biological activity. Finally, Catalyst offers the possibility to output a Fisher randomization test, which generates negative control models where the activities are randomly assigned to the training set compounds before the calculation of the model significance. Catalyst can generate up to 99 random spreadsheets equivalent to a 99 % confidence level. In the current work 19 random spreadsheets were generated, ensuring a 95% confidence in the results for each run.

3.2.12 Pharmacophore clustering

In order to assess the diversity of pharmacophores generated by HipHop, pharmacophores were clustered by calculating pair-wise root-mean-square (rms) fits. This is performed on all pharmacophores pairs using the mapping function:

$$F_{ij} = \text{pairs} + 1/(\text{rms} + 1.01)$$

where pairs is the number of features common to both pharmacophores *i* and *j*,
and rms is the minimum root-mean-square displacement between the pharmacophores.

The similarity function is defined as:

$$S_{ij} = F_{ij}/\text{sqrt}(F_{ii} * F_{jj})$$

and the distance function, or distance between a model pair, as:

$$D_{ij} = 1 - S_{ij}$$

In the present work, clustering of the pharmacophores was performed using a hierarchical average linkage method. This analysis involved a nested family of clustering operations, organized according to the arithmetic mean of all distances between model pairs taking representative model from each cluster (smallest distance between clusters).The algorithm proceeds by first assigning each model to its own cluster. All model pairs are then examined in turn and the pair that yields the smallest value of the arithmetic mean of distance is fused into a new cluster. This process is repeated until all models are fused into one cluster.

3.2.13 Database searching with Catalyst pharmacophores

All preliminary database searches were performed using the Fast Flexible search algorithm from Catalyst, and all fit and final estimated activity values were calculated using the Best algorithm.
The Catalyst Fast Flexible search method uses pre-generated conformations to match a molecule to a pharmacophore, whereas the Best Flexible search can modify the conformations during the matching process. The Best Flexible search finds the best fit among pre-generated conformations and alters the fitted conformer not permitting its energy to rise by more than the default of ~9.5 kcal.

3.3 Results

In order to be able to validate the ligand-based model, a molecular haystack was first established. The haystack elaboration was undertaken using the PBOX activity data that was available at the time.

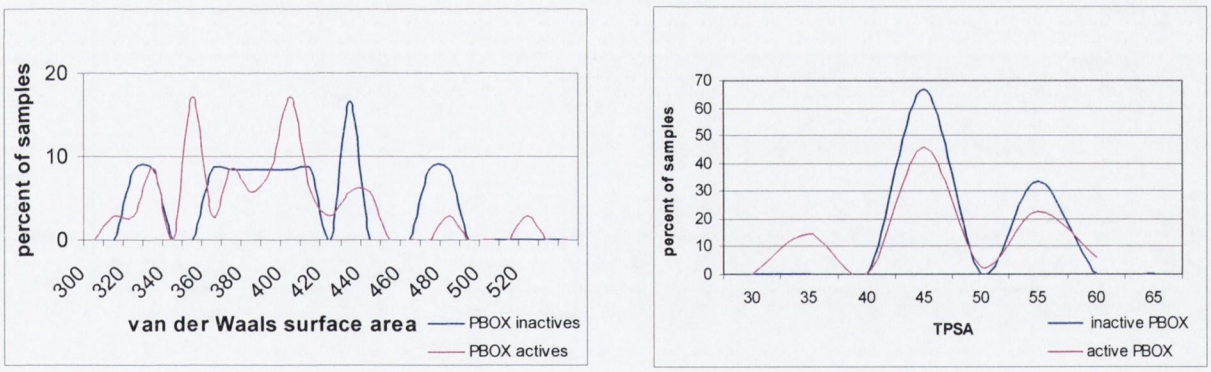
3.3.1 Haystack elaboration

In order to develop the haystack, three different approaches were studied, all of which were considering only 2D descriptors for each molecule. As discussed in Chapter 2, it is common in the chemoinformatics field to employ rule based filters hoping to pre-clean a database of vendor

molecules. Therefore, the first step undertaken in the current work consisted of selecting an adequate set of 2D descriptor thresholds, hoping to filter out non PBOX-like molecules.

3.3.1.1 Filter elaboration

Generally, filters on specific properties exhibited by the active compounds are used to prune molecules from a screening database. Chapter 2 already presented a set of descriptors that is generally used by "lead like" and "drug like" filters. As seen in the last figures in Chapter 2 (Figure2.8), showing the distributions of the molecular weights, number of hydrogen bond donors and acceptors, rotatable bonds and lipophilicity, no significant difference between the active and inactive sets of PBOX compounds in terms of these properties was observed. Importantly, the active and inactive distributions s are similar to or overlapping with one another, with the exception of a slight tendency to fewer rotatable bonds and more hydrogen bond donors for the active set. Additionally, the van der Waals surface area and the Total Polar Surface Area (TPSA) distribution were calculated, but also showed no separation between actives and inactives as illustrated in Figure 3.5. Therefore, descriptor thresholds to pre-filter for PBOX like compounds were set to include all PBOX compounds regardless of their activities.



(a) (b)

Figure 3.5: (a) PBOX van der Waals surface area distributions, (b) PBOX polar surface area distributions
In blue are plotted the inactive PBOX and in pink the active PBOX.

The PBOX lipophilicity threshold was set as $4.65 < \log P (o/w) < 8.63$, the number of hydrogen bond donors (HD) and acceptors (HA) were set as $HD < 2$ and $HA < 6$, the molecular weight (MW) as $55<MW<317$ and the number of rotatable bond (brot) as $3< brot <8$. In addition the total polar surface

area (TPSA) was set as 31.23< TPSA < 70 and the van der Waals area (vdW) as 308 < vdW < 518. In this manner, a PBOX filter was elaborated using seven descriptors filtering out non similar compounds.

3.3.1.2 Principal Component Analysis

The concept of similarity to a molecule is strongly dependant on the set of descriptors chosen. The filter described above selects compounds similar to the PBOX according to seven descriptors that were chosen from amongst those commonly used in the chemoinformatic field. In order to develop a filter more specific to the PBOX compounds, a large number of descriptors was calculated. The correlation vector between these descriptors and PBOX activity did not show any correlation superior to 50 % therefore the large set of descriptors calculated was processed by PCA reducing the total number of descriptors from 150 to 34. Figure 3.6 illustrates the projection of the PBOX compounds onto the two first PCA axes.

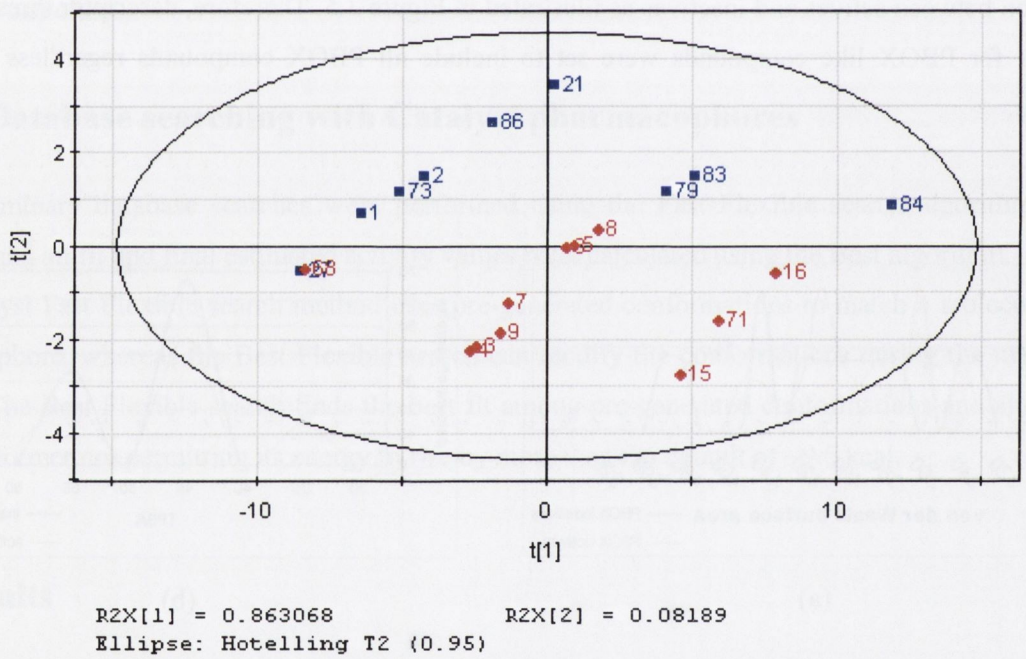


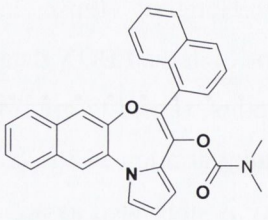
Figure 3.6: PBOX projection onto the two first PCA axes. In red active and in blue inactive PBOX.
*In the bottom left quarter PB68 and PBOX27 overlap.

The scatter plot of the first PC (t1) *vs.* the second (t2), is a projection of the descriptor space, displaying how the PBOX compounds are situated with respect to each other. PBOX near each other are similar whilst PBOX far away from each other are dissimilar. PBOX close to the origin are not well

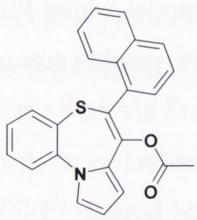
explained by the model whilst PBOX found in the outer part of the Hotelling ellipse can be outliers. This plot shows the possible presence of outliers, groups and similarities in the PBOX data set.

The PC's t1 and t2 are new variables summarizing the 2D descriptors. t1 and t2 are orthogonal, *i.e.* completely independent of each other. t1 explains the largest single variation in the dataset, and t2 explains the second largest. PBOX9 which presented an uncertainty is suggested as active because it overlaps the active space PBOX68 overlaps the PBOX27 which is inactive but are still in the active space. These two compounds could be intermediates as they are localized on the left side of the vertical axes away from the most actives that are grouped on the right. PBOX 6, PBOX5 and PBOX8 were not well explained by the model (close to the origin). More details on the activity evaluations can be found in Chapter 5. To manage this separation, all intermediate compounds were detected as outliers during the several iterations required. The compounds left are highly likely to be the best compounds meaning the most confident to chose for a 3D modelling as the two classes they define (active and inactive) are the best separated.

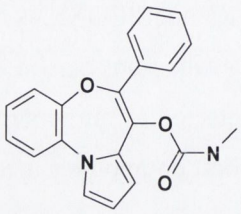
The contribution of each descriptor to a PC is controlled by a weighting factor. These weights or loadings are computed from the correlation matrix of the descriptors. The analysis of this matrix provides information such as which descriptors are negatively correlated, or not correlated to each other, which ones do not contribute to the model and which ones provide similar information. The normalized descriptor weights (loading scores) along the two PCA axes were therefore analysed and particular attention was given to those with normalized weights greater than 0.9 (after elimination of descriptors carrying similar information). These descriptors were considered to be the descriptors of main importance in the separation of the active and inactive PBOX subsets, and were identified as chi1v_C (the carbon valence connectivity index of order 1), opr_nrings (the number of rings), balabanJ (Balaban's connectivity topological index) and KierA3 (the third alpha modified shape index). They appeared to mostly describe the number of rings, the atom connectivity and the molecular shape. This concurred with the available structure activity relationship for PBOX, which showed the importance of the presence of two diphenyl groups, as can be seen when comparing the structure of the active PBOX16 with the inactive PBOX2 and the diphenyl group positioning, seen by comparing the active PBOX16 with the inactive PBOX27.



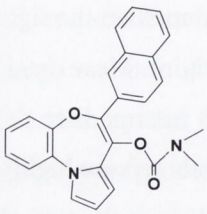
PBOX16 active



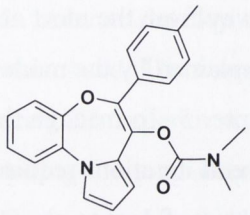
PBOX9 active



PBOX27 inactive



PBOX42 inactive



PBOX2 inactive

The aptitude of the PCA method in describing PBOX-like space was assessed by comparison to a third method of selection, the Bayesian model

3.3.1.3 Bayesian model

Bayesian scores for the PBOX actives and inactives were calculated using the model described in section 3.2.5, and a ROC curve plotted in Figure 3.7.

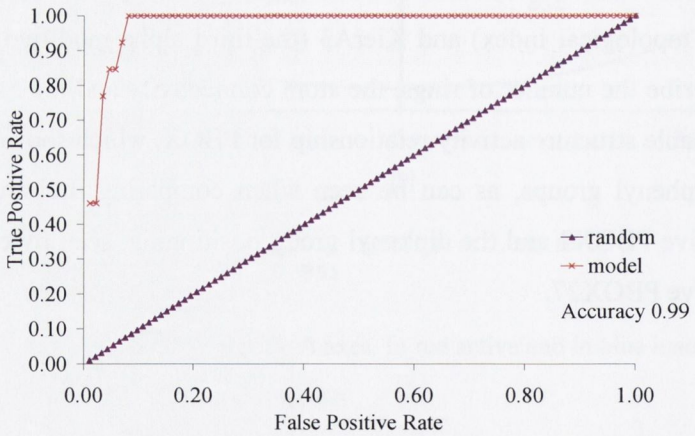


Figure 3.7: ROC curve for the Bayesian model

The score found for the model built showed an accuracy of 0.99 classified as "excellent" and therefore was kept for further studies. Subsequently, several databases were processed to confirm the confidence

in the model in selecting PBOX like compounds. A summary of the scores obtained is illustrated in Table 3.5. It can be seen that active compounds were assigned better scores, on average, than either inactives or molecules from the screened databases, giving some confidence in the model.

Database	Top compound score
WOMBAT	14.85
NCI	19.30
SPECS	13.80

(a)

Set	Top compound score	Bottom compound score
PBOX active	23.96	17.38
PBOX inactive	21.20	13.08

(b)

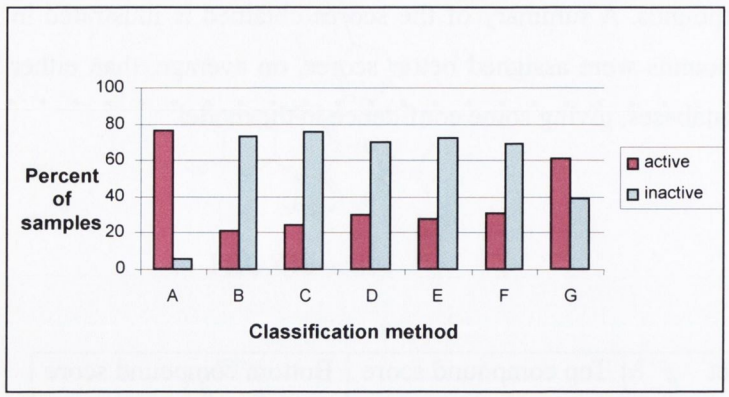
Table 3.5: Bayesian scores, (a) PBOX compound scores (b) databases scores.

3.3.1.4 Haystack validation

The selection of compounds to include in the haystack can be done using either one of the individual models and the filter or a combination of all of them. To test which solution was the most suitable, the WOMBAT and NCI databases were analysed. The NCI database was used to assess the model's aptitude in retrieving cancer active compounds, and WOMBAT to assess the model's ability to retrieve compounds that hit the same targets as PBOX *i.e.* tubulin and PBR.

Cancer activity considerations

Firstly, the NCI database was screened with all possible combinations. The results concerning the active and inactive subsets are presented in Figure.



(a)

Tag	Classification Method	Total no. of compounds
A	Bayesian	4863
B	PCA	3508
C	Filter	2674
D	Bayesian ∩ PCA	429
E	Bayesian ∩ Filter	550
F	PCA∩ Filter	267
G	Bayesian ∩ PCA∩ Filter	54

(b)

Figure 3.8: (a) NCI compound classification according to the different models and (b) Total number of compounds selected for each method

It can be seen that the Bayesian model (A) retrieves more cancer active compounds than the PCA method (B). The use of the filter did not seem to improve the model, as seen by the drop in the proportion of actives (E). Finally, the intersection of all methods (G) was given a worse ratio again, and decreased considerably the total amount of molecules. Therefore, it was concluded that the Bayesian model was the best at retrieving cancer actives,

Target considerations

Table 3.9 shows a summary of the contents of the WOMBAT database, before and after applying the various filtering methods. Particular attention is paid to these compounds labelled as hitting tubulin or PBR, as these are the two proteins known to interact with the PBOX.

	Filtering methods					
	Native (unfiltered content)	Bayesian	PCA	Bayesian ∩ PCA	Bayesian ∩ Filter	PCA∩ Filter
Total number of target entries	135 673 entries* 100%	6 708 entries 5%	31 077 entries 23%	1 174 entries 1%	953 entries 1%	4 115 entries 3%
Total number of different targets	839 targets 100%	298 targets 35%	528 targets 63%	147 targets 17%	112 targets 13%	245 targets 29%
Occurrence of "tubulin" entries	1 167 entries 100% rank 27	485 entries 42 %** rank 1***	183 entries 15.7 % rank 47	67 entries 6% rank 3	46 entries 4% rank 4	17 entries 1% rank 65
Occurrence of "PBR" entries	405 entries 100% rank 80	164 entries 0.5 % rank 8	266 entries 65.7 % rank 27	103 entries 25% rank 3	134 entries 33% rank 1	166 entries 41% rank 2

Figure 3.9: Classification of WOMBAT molecules according to the different

*: this is the total number of target entries retrieved, the sum of all target entries can be different of the total number of compounds as a single compound can be associated with more than one target.

**: this is the percent of target entries retrieved by a particular method with regards to the initial number of target entries prior to any filtering.

***: this is the rank of a particular target retrieved per a particular filtering method.

As shown in Table 3.10, no compound could pass the three models simultaneously. The Bayesian models and PCA both retrieved tubulin and PBR compounds. The PCA method preferably selected PBR ligands (266) to tubulin ligands (183), retrieving the majority of PBR ligands that were present in WOMBAT (65.7 %). As the final aim of the current VS work is to find new scaffolds having similar effects to the PBOX, the underestimation of one target's importance over another was not acceptable. Furthermore, the PCA method was not selective enough as it retrieved other targets in preference to the tubulin and PBR, as seen from their respective ranks 47 and 27. When considering only the top frequency target retrieved tubulin and PBR were not present anymore as shown on Table which lists the target frequencies retrieved for each model.

	Filtering methods						
Rank	Native (unfiltered content)	Bayesian	PCA	Filter	Bayesian ∩ PCA	Bayesian ∩ Filter	PCA ∩ Filter
1*	HIV-1 P 3 318 entries 100%	Tubulin 485 entries.** 42%***	A1 839 entries. 26%	CRF-R1 346 entries. 49%	PBR 103 entries. 25%	PBR 134 entries. 33%	CRF-R1 201 entries. 28%
2	A1 3 283 entries 100%	**PDGFR** 288 entries. 36%	ET-A 806 entries. 45%	D2 306 entries. 19%	ET-A 85 entries. 5%	RAR-alpha 60 entries. 21%	PBR 166 entries. 41%
3	5-HT1A 2 514 entries 100%	ET-A 283 entries. 16 %	5-HT1A 693 entries. 28%	5-HT1A 302 entries. 12%	Tubulin 67 entries. 6%	5-LOX 47 entries. 4%	D2 155 entries. 10%
4	DAT 2 354 entries 100%	**HIV-1 IN** 215 entries. 31.0%	D2 663 entries. 41 %	AChE 293 entries. 19%	IMPDH II 63 entries. 15%	Tubulin 46 entries. 4%	COX-2 153 entries. 16%
5	HIV-1 RT 2 183 entries 100%	**ERβ** 191 entries. 57 %	DHFR 653 entries. 36 %	COX-2 271 entries. 29%	RAR-alpha 49 entries. 7%	ET-A 34 entries. 2%	COX-1 150 entries. 15%
6	GABA-A/BzR 1 854 entries 100%	5-LOX 172 entries. 13.4%	fXa 650 entries. 41%	PBR 265 entries. 65%	GR 38 entries. 21%	PDGFR 34 entries. 4%	AChE 149 entries. 10%
7	AT1 1 848 entries 100%	GABA-A/BzR 170 entries. 9%	AT1 621 entries. 34%	COX-1 263 entries. 26%	HIV-1 IN 35 entries. 5%	RXR-alpha 31 entries. 22%	GAT1 116 entries. 45%
8	DHFR 1 816 entries 100%	PBR 164 entries. 40%	DAT 573 entries. 24%	5-LOX 218 entries. 17%	RXR-alpha 29 entries. 21%	TP 29 entries. 7%	DAT 107 entries. 5%

Table 3.6: Top target frequencies among compounds retrieved from WOMBAT by the various classification methods.

Targets known to hit by the PBOX's are highlighted in red *: this is the rank of a particular target retrieved per filtering methods *e.g* most compounds retrieved by the Bayesian method are associated with "tubulin" therefore "tubulin" as a target is ranked in the first position whereas initially most compounds present in the not filtrated database where associated to "A1" therefore was ranked in the first position for the natives. **: this is the number of target entries (here "tubulin") retrieved by a method (here the Bayesian method). ***: this is the percent of entries for a particular target *e.g* there was 1 167 entries associated to "tubulin" prior to any filtering and the Bayesian method retrieved 485 entries associated with "tubulin" which represent 42% of the initial "tubulin" entries.

Tubulin was the target retrieved most frequently by the Bayesian method and PBR was also retrieved in the top frequency targets. However, the PCA methods retrieved mostly targets not known to interact with the PBOX compounds. The filter methods retrieved the 67% of PBR ligands but did not retrieve the tubulin ones. Combination of the Bayesian method with either the filter or the PCA methods brought tubulin frequencies higher to the top 3% of the target hit list but considerably decreased their respective retrieval rates. Both the filter and the PCA method were discarded as methods to be used for the haystack elaboration and only the Bayesian model was kept.

Interestingly, the Bayesian method also retrieved 36% of the PDGFR ligands, 31% of the HIV-1 IN ligands and 57% of ERβ ligands from WOMBAT. As these ligands were selected as being similar to the PBOX compounds it is not excluded that PBOX compounds could hit these targets. It was interesting to analyse more closely the kind of structures selected amongst these novel, potential targets. Table 3.7 (next page) illustrates three representative members of each target of interest.

The WOMBAT database includes the PBOX structures because they were published and made available to the public. This kind of blind test was passed successfully by the Bayesian method, as illustrated in Table 3.7 where the PBOX structure appears in a PBR ligand (on the right) as well as a number of very similar compounds. In general, all the compounds contained ring systems (two to four rings) similar to those seen in tubulin binders. As an example, in Table 3.7, the centre compound from the tubulin row includes the combretastatin core. These types of compounds are discussed in more detail in Chapter 4. The Platelet-Derived Growth Factors Receptor (PDGFR) ligands retrieved contained features such as the fused five and six membered heteroaromatic rings common to PBOX compounds. As the PDGFR has been recognized as an important factor for cell regulation, therefore being important in diseases such as cancer, it could be suggested as a potential supplementary target for the PBOX compounds. In the case of ERβ, the oestrogen receptor β isoform, in general those compounds retrieved by the Bayesian classifier showed flavanoid cores. The central structure in table 3.7 contains the core common to Tamoxifen, which is a currently marketed anti-oestrogen. The perceived similarity to the PBOX compounds appears to have arisen because of the presence of multiple rings, including fused ring system. Finally, HIV 1 integrase (HIV1-IN) ligands were also highly matched. There is an obvious similarity between these ligands and the top right compound in the table 3.7. HIV1-IN protein is essential for integration of viral DNA into the host cell chromosomes for the human immunodeficiency virus type 1 (HIV-1). A review of the literature, confirmed that Campiani et al[46, 47]. (the team that first synthesised the PBOX core) moved from their original optimization of the PBOX scaffold to the PBR to an optimisation as anti-HIV agents, and recently have demonstrated the ability of PBOX derivatives to bind HIV-1 reverse transcriptase (HIV-1 RT).

On the basis of those compounds retrieved from WOMBAT, the types of ring substructures present in the PBOX's seem to be a very useful representation of molecule class, enabling both high retrieval rates of PBOX targets (tubulin and PBR) and the identification of new chemotypes or novel potential targets such as ERβ and HIV-1 IN.

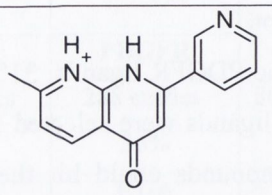
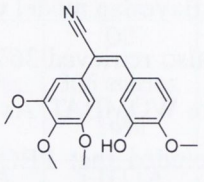
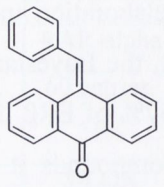
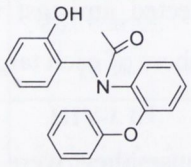
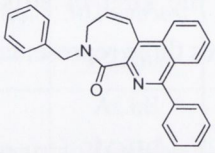
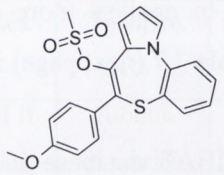
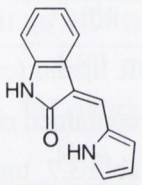
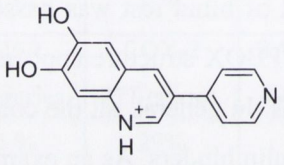
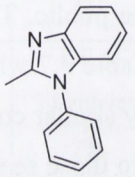
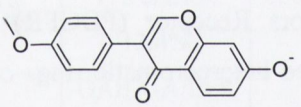
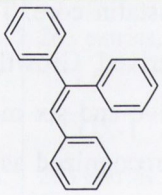
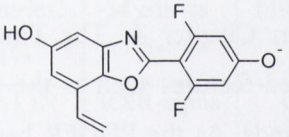
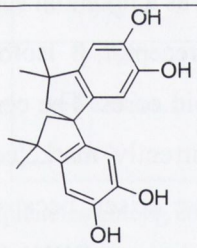
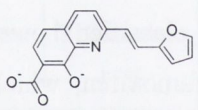
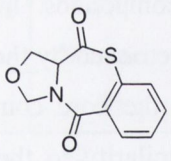
Target	Bayesian subset, cluster centres		
Tubulin			
PBR			
PDGFR			
ERβ			
HIV-1 IN			

Table 3.7: Structures representative of the different target classes retrieved from WOMBAT by the Bayesian classification model.

Haystack determination

Following this optimisation study, the NCI, WOMBAT and SPECS databases were screened using the Bayesian method and the most inactive compounds similar to the PBOX were selected to complete the haystack. Inactivity was defined as non tubulin binders for the WOMBAT database, as cancer inactives for the NCI database and as decoys with highest Bayesian scores from SPECS. Table 3.8 illustrates the selection of compounds from each database.

Database	WOMBAT Not tubulin binders	NCI inactives (GI50 in < 6 assays in NCI screens)	SPECS PBOX-like	Total molecules
Haystack	1053	295	3515	4863

Table 3.8: Haystack content

In addition, the SPECS 15 compounds with the highest scores from the Bayesian model were tested *in vitro* for their potency against a breast cancer cell line. These compounds were numbered from MDG20 to MDG35 and their structures can be found in this chapter's appendix. Details on the assay are presented in Chapter 5. The IC50 values calculated are reported in Table 3.9, they represent the mean and standard errors of three replicants.

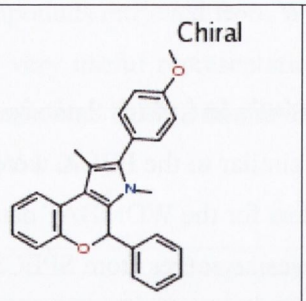
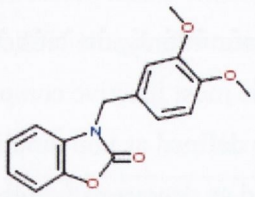
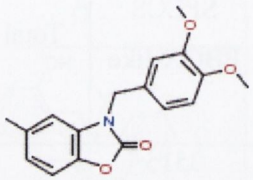
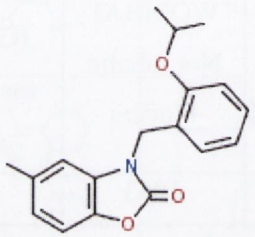
	
MDG31, IC50 < 1 μM	MDG35 IC50 < 1 μM
	
MDG32, IC50 = 2.9 +/- 1.5 μM	MDG29, IC50 = 39.1 +/- 4.3 μM

Table 3.9: Bayesian active compounds

These compounds did not show any effect on microtubules in MCF-7 breast cancer cell line and were also tested inactive in leukaemia cell lines. Indeed, the selection of compounds by the Bayesian method was validated by seeding a random database of molecules with known actives. This random decoy set includes only a very small fraction of all possible molecules. It is therefore simple for the algorithm to discriminate between actives and inactives and retrieving the seeded actives remains relatively easy as their features were memorised. In addition, these methods were limited to both a local and 2D description of the activity, so the Bayesian model could only score and predict localised a region of chemical space. In an attempt to cover a larger chemical space and find new active in Leukaemia cells, a 3D approach was therefore undertaken.

3.3.2 3D-pharmacophore

3.3.2.1 Pre-processing

Inherent to the development of a 3D model is the actual 3D representation of each molecule. In order to be able to choose a conformation generator we tested the aptitude of a number of algorithms in generating PBOX conformers (*cf.* Figure 3.10).

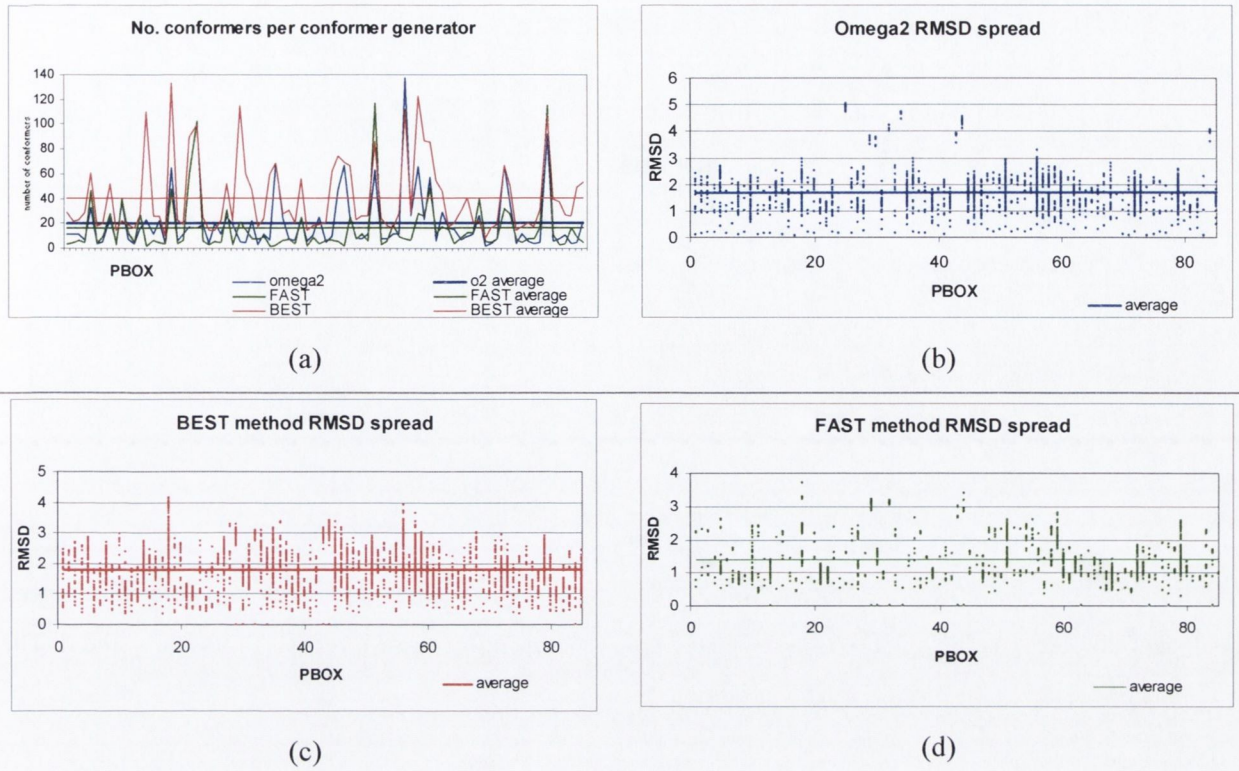


Figure 3.10: Conformation generation trials on PBOX compounds.

(a): number of conformers per generator, in red BEST, in green FAST and in blue OMEGA2. Straight lines are the respectives average number of conformers for the whole set of PBOX.

(b) RMSD spread for OMEGA2 with regard to the first structure generated.

(c) RMSD spread for BEST with regard to the first structure generated.

(d) RMSD spread for FAST with regard to the first structure generated.

As the development of the 3D models is strongly dependant on the 3D representation of the molecules in the training set and validation sets, it was essential to choose a 3D conformation generator that was able to generate a large number of conformers and diverse conformers to cover the maximum conformational space possible. The graph in Figure 3.10(a) shows that BEST outperformed the other tested generators in terms of the number of conformers per structure. In addition, as illustrated by the RMSD spread of PBOX conformers obtained by the three generators (*cf.* Figure (b,c,d)), BEST slightly outperformed the other methods in terms of the diversity of conformers very close to OMEGA. In consequence, all PBOX compounds were processed through BEST. Nonetheless, as BEST is a relatively heavy algorithm in terms of computational time, the large screening databases screened were pre-processed with OMEGA2, which appeared to be the second best generator, pretty close to BEST in quality.

3.3.2.2 *In silico* validation of 3D shape models

The 3D approach chosen was shape based method using ROCS. PBOX16 was used as a query to screen the haystack screened for two main purposes:

1) the ranking method

2) the best query conformer or the combination of all conformers.

3.3.2.2.1 ROCS and EON

Method determination

To determine the best ranking method, the haystack was screened using the combination of all conformers as query and the output was ranked successively using the seven possible parameters: shape Tanimoto, overlap, Combo score, scaled color scores, color score Tversky Tanimoto q and d. The results of these 7 trails are reported in Figure 3.11.

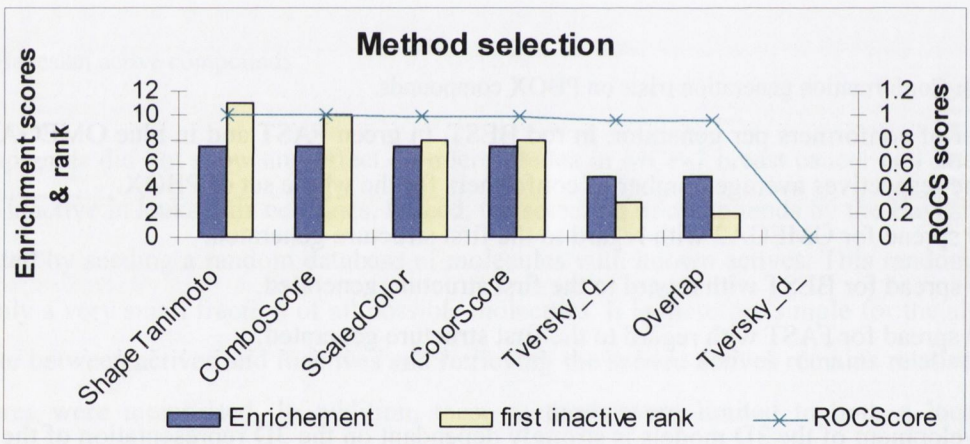


Figure 3.11: ROCS method determination

Blue bars are scaled on the left axe and represent the enrichment score, yellow bars are scaled on the right axe and represent the rank of the first inactive, the curve in light blue represent the ROC score and is scaled on the right axe.

The shape Tanimoto ranking methods ranked the first inactive of all PBOX inactives, the furthest from the hit list top if compared to the ComboScore and consequently to the scaled score which all obtained good ROC scores and enrichments. Therefore the method chosen was the shape Tanimoto method which is purely shape based.

Query determination

The next step consisted in analysing in the hit list delivered with the Shape Tanimoto method, if one conformation was most frequently chosen by the actives PBOX or not as shown in Table 3.10.

PBOX ID	Activity	Shape Tanimoto	Query
15	1.49	0.98	5
6	2.98	0.91	17
71	3.01	0.85	12
8	3.10	0.90	5
5	3.38	0.91	12
4	4.02	0.89	5
7	4.02	0.89	11
3	4.78	0.89	5
9	15.75	0.88	10
68	80.75	0.78	5
1	-	0.82	6
2	-	0.81	14
21	-	0.79	13
31	-	0.70	20
38	-	0.76	7
42	-	0.79	16

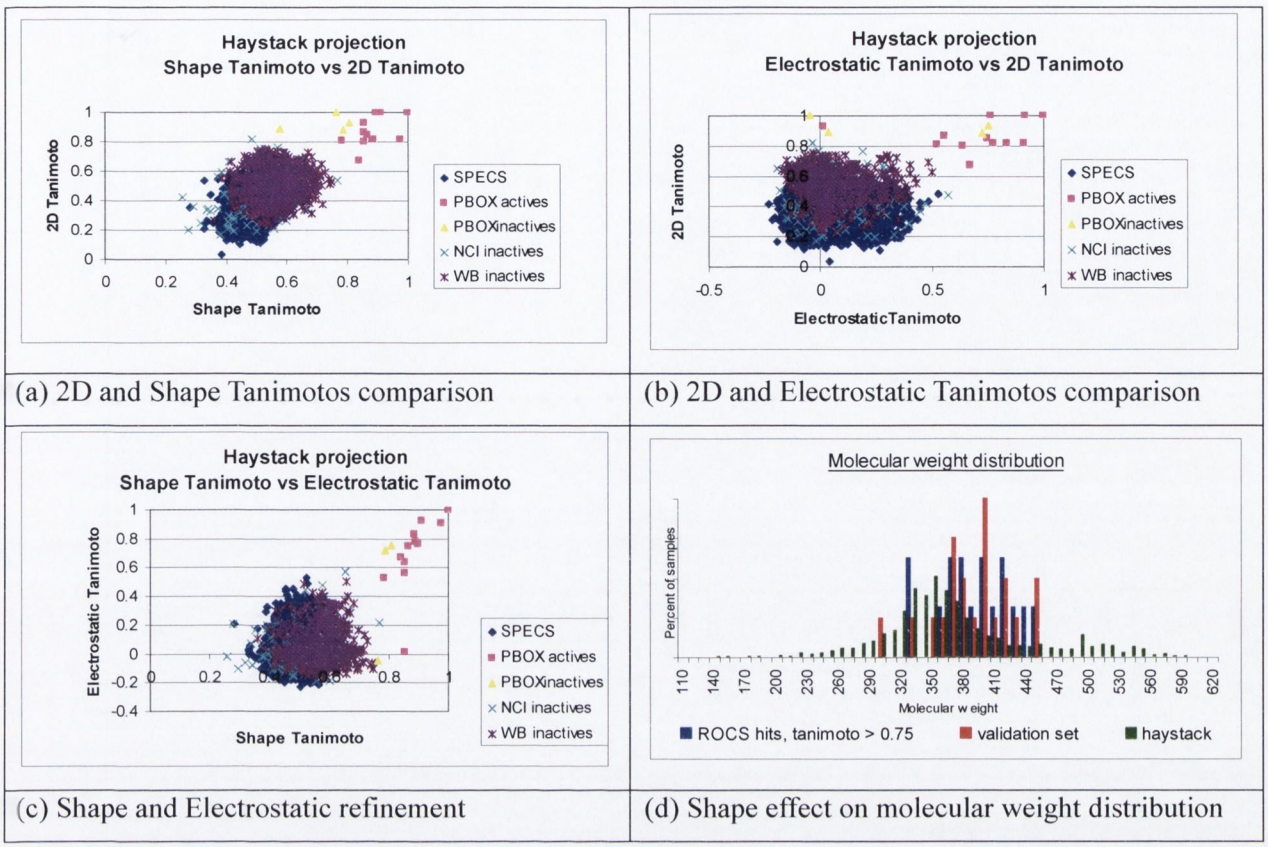
Table 3.10: Preferred conformation chosen by the training set compounds

PBOX16 conformation (denoted query 5) appeared to be chosen the most frequently by active compounds and the closer examination of the query conformation chosen by the remaining actives appeared to be considerably close to this same query. As the best match between all possible conformation of the actives and PBOX 16 all conformations was found more often for query 5 or close to query 5 conformers, query 5 was chosen as a query for the subsequent electrostatic refinement step with EON.

In silico validation

In order to validate the shape based method associated with query 5 and with EON refinement as a screening method, an *in silico* validation step was undertaken.

Figure 3.12 (a) illustrates the effect of the shape Tanimoto method when compared to a 2D descriptor method. ROCS shape method is able to accurately separate PBOX actives (in pink) from the haystack (large blue cloud) and can separate in two clusters, the inactive (in yellow) from the active PBOX (in pink) whereas the 2D method only separates all PBOX compound from the haystack regardless of their activity. The same phenomenon can be observed on Figure 3.12 (b) after EON refinement. Despite, one active is ranked with the inactives the spread of ranking values is considerably enhanced as PBOX compounds are separated one from the other in terms of electrostatic Tanimoto (horizontal axe) better than the 2D method (vertical axis) does. In other words, PBOX compounds are more spread along the X axis rather than the Y axis. Analogously, EON performance is retrieved in Figure 3.12 (c) when compared to the pure shape match. Finally, the last graph in Figure 3.12 (d) illustrates the effect of the screening process on the whole database. For each subset, the training set, the haystack and the hit list of the haystack the distribution of the similarity to a PBOX active is plotted showing that the hit list compounds (in blue) are closer to the training set (in red) when compared to the original haystack. This distribution plot was using molecular weight as a descriptor showing that compounds selected are close to PBOX in shape and size. Consequently, the method was validated for further screens.



(a) 2D and Shape Tanimotos comparison

(b) 2D and Electrostatic Tanimotos comparison

(c) Shape and Electrostatic refinement

(d) Shape effect on molecular weight distribution

Figure 3.12: ROCS and EON effect upon the haystack.

To examine the ability of Catalyst to predict the affinities of actives retrieved by ROCS/EON an initial study was undertaken.

3.3.2.2.2 Catalyst approach

To fully validate the Catalyst approach it is necessary to examine whether catalyst can correctly map features indicative of the training set which is presented in Figure 3.13.

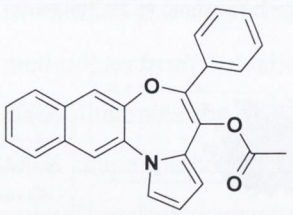
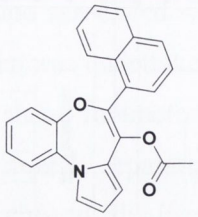
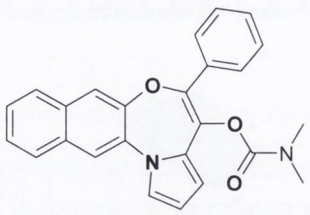
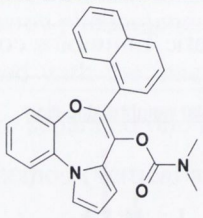
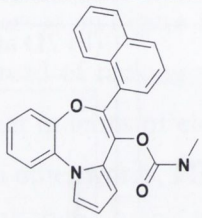
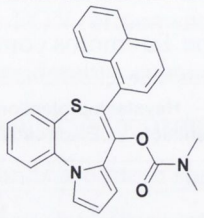
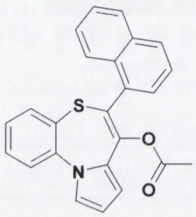
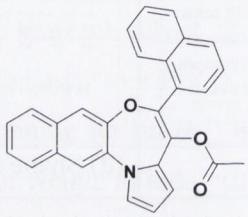
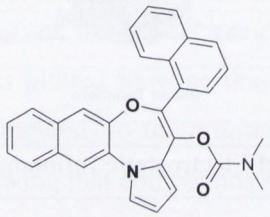
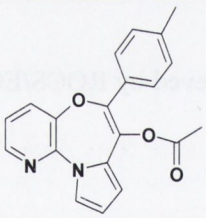
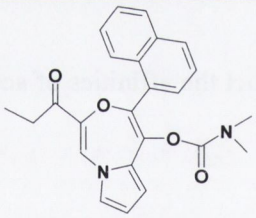
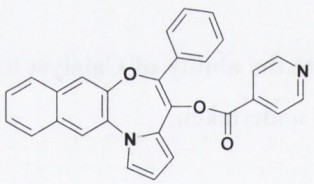
Training set actives (IC50 in μM)		
		
PBOX 3 (4.8)	PBOX 4 (4.0)	PBOX 5 (3.4)
		
PBOX 6 (3.0)	PBOX 7 (4.0)	PBOX 8 (3.1)
		
PBOX 9 (15.7)	PBOX 15 (1.4)	PBOX 16 (2.6)
		
PBOX 68 (80.7)	PBOX 70 (3.0)	PBOX 71 (3.0)

Figure 3.13: PBOX structures from the training set

Feature mapping

Importantly, as portrayed in Table 3.11, all molecules from training set match all features of the pharmacophores for all pharmacophores generated.

PBOX	HBA	HBD	HYDROPHOBIC	HYDROPHOBIC aromatic	RING AROMATIC
1	3.00	0	4	3	6
2	2.90	0	5	3	6
3	4.07	0	4	4	8
4	3.81	0	4	4	8
5	2.96	0	5	4	8
6	2.79	0	5	4	8
7	2.86	0.73	5	4	8
8	2.96	0	5	4	8
9	4.16	0	4	4	8
15	3.86	0	5	5	10
16	2.81	0	6	5	10
21	2.79	0	5.98	3	6
31	3.11	0	4	2	4
38	6.82	0	4.95	3	6
42	3.11	0	5	4	8
68	5.29	0	4	3	6
70	5.54	0	6	4	8
71	4.14	0	4.68	4.68	10

Table 3.11: Mapping of training set features;

HBA: Hydrogen Bond Acceptor, HBD: Hydrogen Bond Donor,

The number in each column correspond to the average number of hits per conformer.

It is clear from Table 3.11 that PBOX7 possesses the ability to donate H-bond and thus is mapped appropriately by the HBD feature. Although PBOX71 appears to have common features with PBOX 15/16, looking at the structures, it is noticeable that the naphthalene moiety is crucial to the apoptotic process. PBOX71 has certainly an additional aromatic feature, however, it is positioned on the side-chain making it more devoid of activity. Subsequently, ten pharmacophores were generated extracting only common features of the training set with HipHop.

HipHop

As PBOX compounds are structurally very close analogues and the most confident activity spread is quite restricted (from 1,4μM to 80μM) all 12 actives were selected to represent the training set. In Table 3.12 summarises all pharmacophores generated using HipHop algorithm and indicates in the numbered columns the total number of clusters generated and in the rows the pharmacophore ID. The entries in the table indicate which cluster a pharmacophore belongs to and the pharmacophores with the same number belong to the same cluster

Pharmacophore ID	Pharmacophore Features	2	3	4	5	6	7	8	9
01	ZZHAA	1	1	1	1	1	1	1	1
02	ZZHAA	1	1	1	1	2	2	2	2
04	ZZHAA	1	1	1	1	2	3	3	3
07	ZZHAA	1	1	2	2	3	4	4	4
08	ZZHAA	1	1	2	3	4	5	5	5
09	ZZHAA	1	2	3	4	5	6	6	6
10	ZZHAA	1	2	3	4	5	6	7	7
05	RZHAA	2	3	4	5	6	7	8	8
06	RZHAA	2	3	4	5	6	7	8	9
03	RZHAA	2	3	4	5	6	7	8	8

Table 3.12: Pharmacophore clustering;
A: Hydrogen Bond Acceptor, D: Hydrogen Bond Donor,
H: Hydrophobic, Z: Hydrophobic aromatic, R: Ring aromatic

It is clear that as the hydrogen bond donor feature (HBD) was only attributed to PBOX7, HipHop generation did not consider it as a common feature. From the first column of the clustering, the ten pharmacophores are split in two clusters according to their feature combination. To move from one cluster (RZHAA) to another (ZZHAA), a ring aromatic group (R) is exchanged for a hydrophobic aromatic group (Z). In addition, as in total there are only two possible combinations of features, increasing the number of clusters from columns left to right, will instruct the pharmacophores to gather according to the feature positions and directions of the directional features (*e.g.* hydrogen bond acceptor directions). Taken as a whole, the two clusters are very similar. Having generated ten

pharmacophores from the previous step, it is necessary to assess the ability of each to discriminate between actives and inactives in the haystack.

In silico validation

Table 3.13, presents the ROC score calculated for haystack successive screen using each pharmacophore.

Pharmacophore	ROC score on fit values	Accuracy
1	0.81	Good
2	0.88	Good
3	0.89	Good
4	0.86	Good
5	0.92	Excellent
6	0.90	Excellent
7	0.80	Good
8	0.87	Good
9	0.74	Fair
10	0.78	Fair

Table 3.13: ROC scores per pharmacophore

As depicted in Table 3.13, Pharmacophore 5 exhibits the optimal ROC score, and thus was retained for further shaped-based analysis.

Subsequently, PBOX16, the larger of all most actives PBOX was used to create an overaching shape constraint which was merged to pharmacophore 5. Adding shape-based features (blue spheres) had a negative impact on the ROC score as the shape was based on the best-fit of PBOX16 from Pharmacophore 5 which restricted the number of conformers for the other PBOX that passed. Thus without shape-based features, it is possible for each pose to adapt in a better manner to each of the other features and consequently obtain a superior fit value.

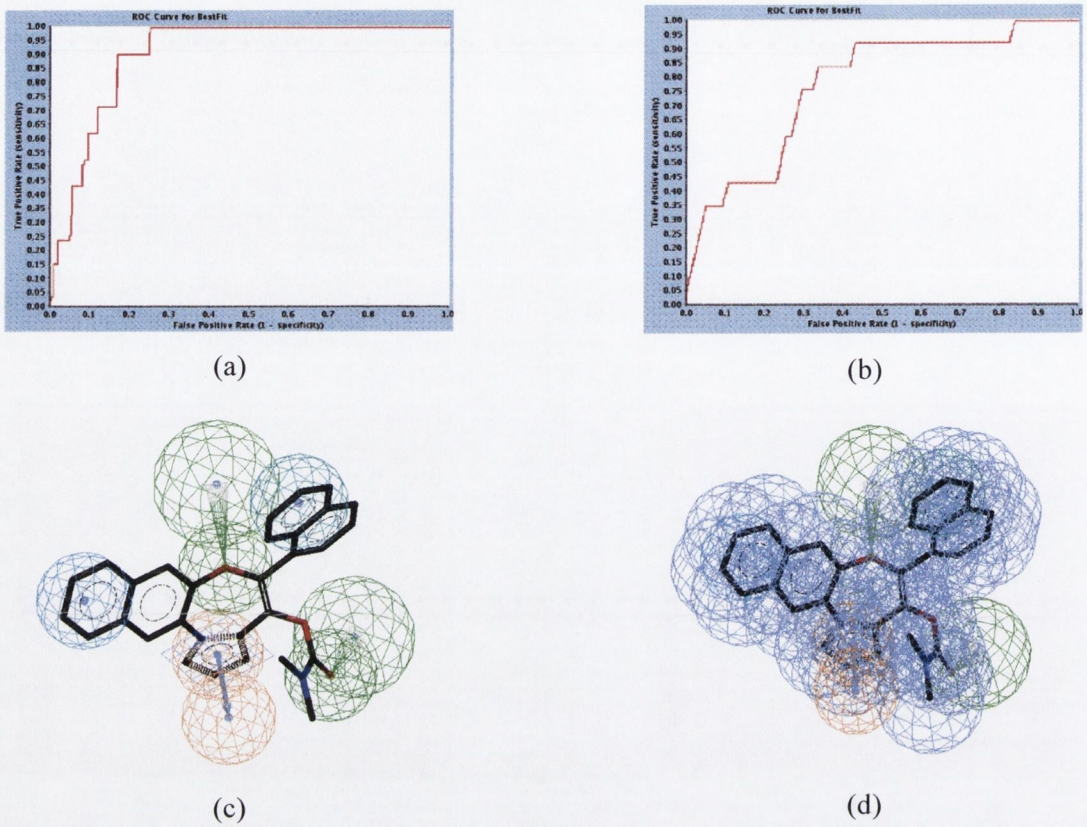


Figure 3.14: Shape addition evaluation upon HipHop best pharmacophore.
(a): ROC curve for HipHop pharmacophore 5, (b): ROC curve for HipHop pharmacophore 5 with shape added, (c): HipHop pharmacophore 5 where spheres represent features as in cyan hydrophobic, in blue a hydrophobic aromatic, in orange a aromatic and in green hydrogen bond acceptor and (d): HipHop pharmacophore 5 with shape added in blue spheres.

Although from Figure 3.14 (b) it is shown that some PBOX compounds see their fit scores dropping considerably on addition of shape-based features, addition of these features tends to shift the distribution of compounds that do pass towards compounds that are more similar to PBOX compounds as seen on Figure 3.15. Keeping the shape-features also reduced the peak size shown on the far right of the figure more dramatically than without, which reached the maximum similarity according to 2D fingerprinting.

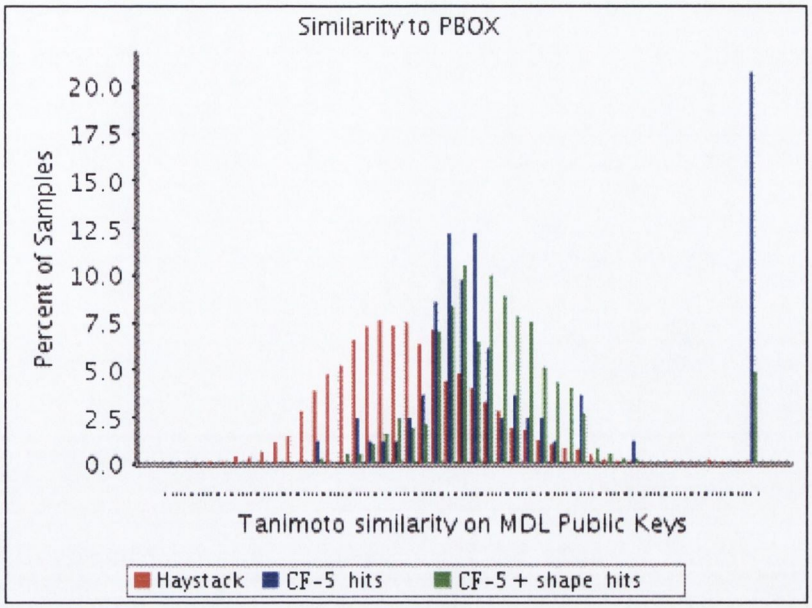


Figure 3.15: Effect of HipHop pharmacophore on MDL fingerprints distribution.

To further test the quality of HipHop pharmacophore, the later was regressed in order to predict training set compounds activities.(*cf.* Table 3.15)

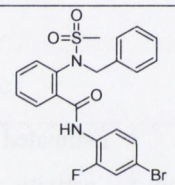
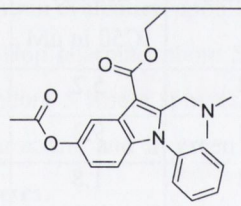
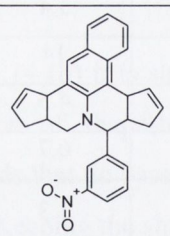
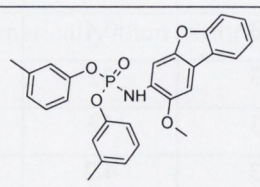
Training set compounds	Experimental activity IC50 in µM	Uncertainty	Estimated activity IC50 in µM	Error
PBOX3	4.78	3	5.2	1.1
PBOX4	4.02	3	9.3	2.3
PBOX5	3.38	3	1.8	-1.9
PBOX6	2.98	3	3.4	1.2
PBOX7	4.02	3	14	3.5
PBOX8	3.1	3	4.5	1.5
PBOX9	15.7	3	6.7	-2.4
PBOX15	1.49	3	1.3	-1.2
PBOX16	2.58	3	0.4	-6.5
PBOX68	80.7	3	32	-2.5
PBOX70	2.96	3	12	4.1
PBOX71	3.01	3	4.2	1.4

Table 3.14: Training set prediction.

In Table 3.14, the error column corresponds to the ratio of tested/estimated activity. A negative error corresponds to a tested activity higher than the estimated and a positive error to an estimated activity higher than the tested activity. Pharmacophore 5 is shown to adequately predict the training set activities and it is clear that the pharmacophore certainly assigns activity to each PBOX compound.

3.3.2.2.3 Shape models *in vitro* validation

The SPECS database of compounds was screened using ROCS and EON refinement and consequently compounds from the top of the hit list were tested *in vitro* for their effect upon cancer cell lines. Details of the in vitro testing procedure can be found in Chapter5 and these compounds were numbered MDG36 to MDG50 with the addition of MDG61. Consequently, hits retrieved were further evaluated by the same HipHop pharmacophore 5 regressed new pharmacophore.(*cf*.Table 3.15)

Hits	Structure	Experimental activity IC50 in μM	Uncertainty	Estimated activity IC50 in μM	Error
MDG36		5.9	3	5.6	-1.1
MDG39		50.7	3	46	-1.1
MDG40		230	3	380	1.6
MDG41		12.7	3	1.7	-7.4

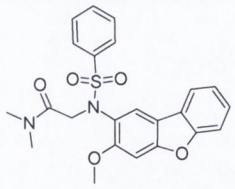
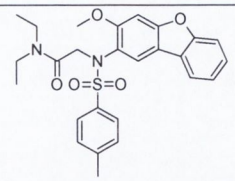
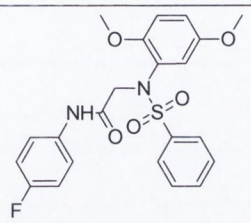
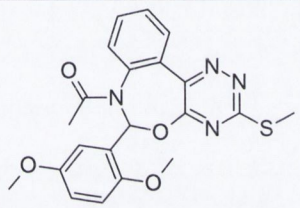
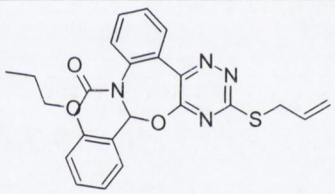
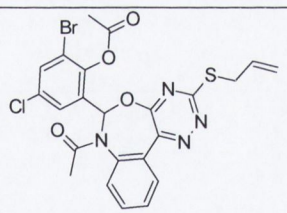
MDG42		1.4	3	6.1	4.3
MDG43		1.0	3	4.7	4.6
MDG44		9.5	3	3	-3.2
MDG45		26.7	3	15	-1.8
MDG49		7.5	3	13	1.7
MDG61		44.2	3	39	-1.1

Table 3.15: Activity prediction by HipHop pharmacophore 5 of active compounds found from ROCS/EON screen

It is evident that the hit selection found after screening with the ROCS method is quite diverse. Although, the three last hits are more structurally PBOX like (MDG45, MDG49 and MDG61) the remainder have amine links (MDG36, MDG42, MDG 43 and MDG44), reactive groups or fused rings (MDG40). Again pharmacophore 5 activities predicted are observed to be accurate with relatively low errors, especially in the case of those compounds with lower activities (high IC50 values) where

predictive ranges are more difficult. Indeed, the training set was composed of compounds whose activities spanned up to a maximum of 80μM and only for one compound (PBOX68), thus limiting the training of the model to a low μM range.

While the pharmacophore exhibited excellent initial results, at this stage further refinement was still possible by changing algorithm to HypoGen and exchanging shape-based features for exclusion volumes using HypoRefine algorithm.

3.3.2.3 Pharmacophore refinement

3.3.2.3.1 Refined pharmacophore generation

In order to obtain the optimal refined PBOX pharmacophore 5 distinct runs were undertaken using HypoGen method for run 1 and 2 and HypoRefine for run 3, 4 and 5, each run output the ten pharmacophore numbered from the most statistically significant to the less.

The active compounds included in training set were unchanged and six supplementary inactive PBOX added PBOX1, PBOX2, PBOX21, PBOX31, PBOX38 and PBOX42 illustrated in Table 3.16.

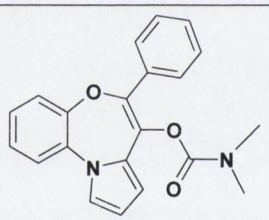
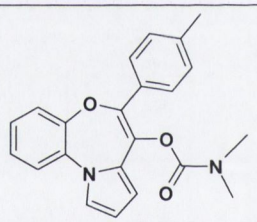
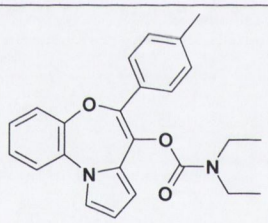
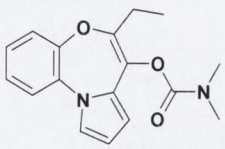
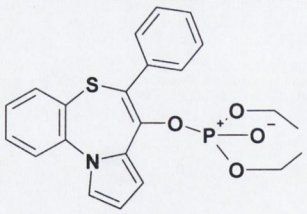
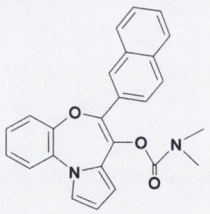
Training set inactives		
		
PBOX 1	PBOX 2	PBOX 21
		
PBOX 31	PBOX 38	PBOX 42

Table 3.16: PBOX structures of inactives present in the training set.

As mentioned previously, the training set activity range was quite small and PBOX molecules in general are quite rigid molecules. The assumption is that a good pharmacophore will map most of the

chemical features present in the most active compounds in the training set. Therefore, the parameter which defines the regions of pharmacophore space that are most likely to be relevant to the training set was restricted to 3 in run 2 instead of 4 in run 1. The last three runs using HypoRefine differed by the maximum number of exclusion volumes tolerated, run 3 was set to only one exclusion volume, run 4 ten exclusion volumes and run 5 one hundred exclusion volumes. The actual total number of exclusion volumes obtained per pharmacophore, for the three HypoRefine runs, is presented in Figure 3.16.

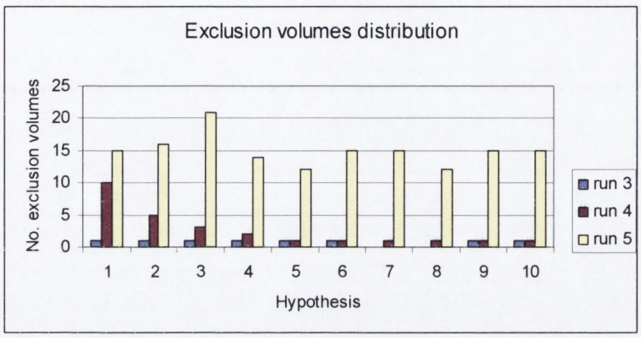
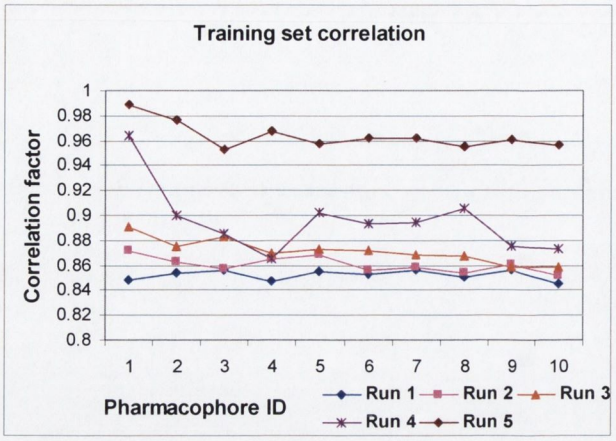


Figure 3.16: HypoRefine exclusion volume distribution along all pharmacophores generated

It is clear that for run 3 and 4 that Hyporefine included the requested number of exclusion volumes for the most significant pharmacophore *i.e* the top ranked ones whereas the less the pharmacophore is significant, the fewer exclusion volumes are attributed. The exclusion volumes distribution for run 5 (which requested 100 exclusion volumes) shows the limitation of HipoRefine generated models. The maximum exclusion volumes obtained was of 21 exclusion volumes for run 5 pharmacophore 3. Of course, the number of exclusion volumes depicted per model is strongly dependant on the training set inactive compounds diversity and their conformational representation. As PBOX are all very similar and pretty rigid the maximum number of volumes could not be reached.

Subsequently to the pharmacophore generation, the training set correlation was plot for the 50 pharmacophores and is presented in Figure 3.17 (a).



(a)

Pharmacophore ID	% of actives retrieved
R4.1*	32
R5.1	39
R5.2	61
R5.3	64
R5.4	61
R5.5	57
R5.6	**71**
R5.7	**71**
R5.8	43
R5.9	**71**
R5.10	**71**

(b)

Figure 3.17: (a) Pharmacophore correlation to the training set, (b) percent of actives retrieved

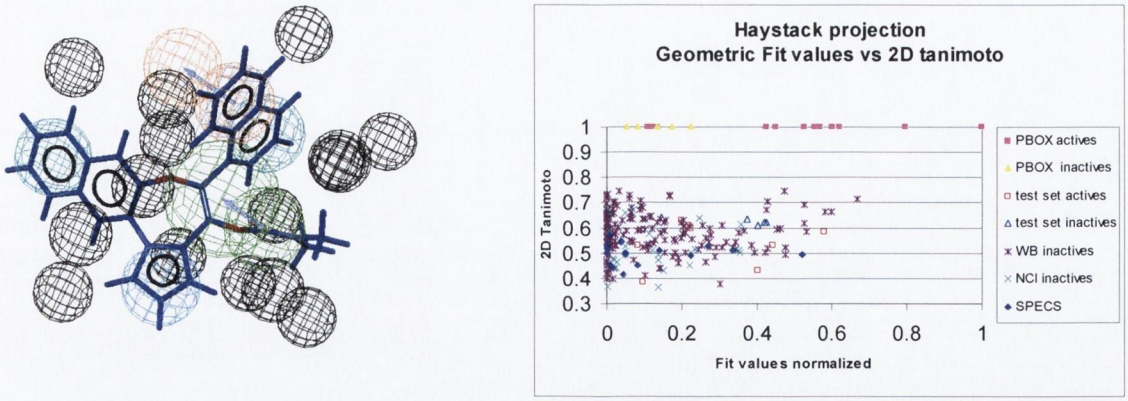
*RX.Y: R stands for Run, X corresponds to the run number (1 to 5 runs) and Y to the pharmacophore number (1 to 10 pharmacophores generated per run)

It appeared that run 5 presented in general the best correlations except for its pharmacophore 3 which had 21 exclusion volumes over-restricting the model. Run 4 presented an acceptable correlation for its first pharmacophore so that it was retained as well as the nine remaining run 5 pharmacophores.

To further test the quality of the 11 remaining pharmacophores, the previous hits found were added to the haystack and all models were tested for their aptitude in retrieving the 12 PBOX actives and 10 MDG actives (see Figure 3.17 (b)). Run 5 pharmacophore 6, 7, 9 and 10 presented the best over all yield therefore their ROC score based on the 'fit' values of compounds in the pharmacophore was evaluated for the four of them and was found the best for pharmacophore 7.

Run 5, pharmacophore 7, denoted R5.7, presented:

- A yield for PBOX actives of 100 %
- A yield for MDG actives of 80 %.
- ROC score = 0.85
- A training correlation factor of 0.96 with an RMS of 1.73.
- A cost of 115.058 where the Null hypothesis cost was of 412.7 and the fixed of 78.58.
- 15 exclusion volumes (in black), 3 hydrophobic (in cyan), 1 hydrogen bond acceptor (in green) and one aromatic (in orange) as depicted in Figure 3.18 (a).

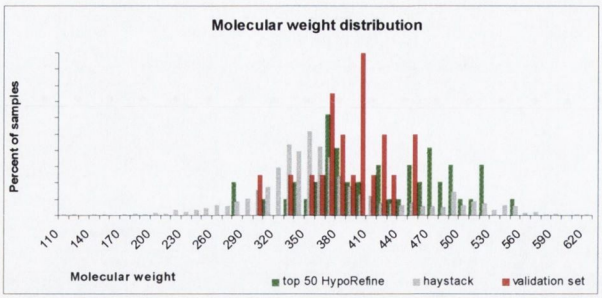


(a) (b)

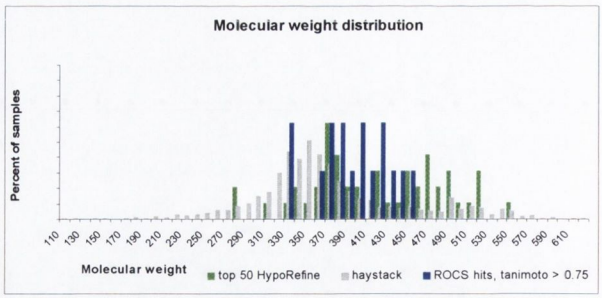
Figure 3.18: (a) Best Fit of PBOX 16 in blue in R5.7, (b) HypoRefine *vs*.2D descriptor.

Interestingly, as illustrated in Figure 3.18 (a) some exclusion volumes were placed around the side chain R which was concluded not to tolerate large groups from the preliminary SARs. This confirmed HypoRefines capacity in building an intelligent model. For instance, the exclusion volumes localised around the side chain prevent the alignment of PBOX21 (inactive) in the same fashion as PBOX16. PBOX21 is actually forced to flip, therefore cannot obtain a comparable Fit valsue.

The haystack hit list issued from R5.7 was then submitted to a 2D description. As illustrated in Figure 3.18 (b), R5.7 is able to separate the training set actives (pink) and inactives (yellow) better than the 2D description does (vertical axe). Nonetheless, only three actives from the previously found hits get Fit values (normalized) superior to 0,4 yet the total number of compound in this region of the graph remains seriously restrained. The first 20% of the top hit values is exclusively populated with actives.
In order to analyse what kind of compounds were selected by R5.7, the distribution of molecular weight was plotted in Figure 3.19 (a) for the training set in red, the haystack (in grey) and the hit list (in green).



(a)



(b)

Figure 3.19: HypoRefine and ROCS/EON hits molecular distribution comparison

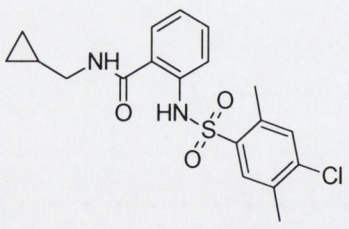
(a) Molecular weight distribution for in red the training set, in green the pharmacophore hits and in grey the non processed haystack.

(b) Molecular weight distribution for in green the pharmacophore hit list, in blue ROCS/EON hits with a Tanimoto > 0.75 and in grey the non processed haystack

R5.7 seems to select a wider variety of compounds in its hit list as shown by the larger molecular weight distribution of the hit list compared to the training set. Furthermore, when compared to the selection of ROCS/EON method, the molecular weight range remains larger for R5.7 as illustrated in Figure 3.19 (b). Consequently, screening with R5.7 which means including exclusion volumes, would lead a most diverse set of hits than screening on shape based methods alone.

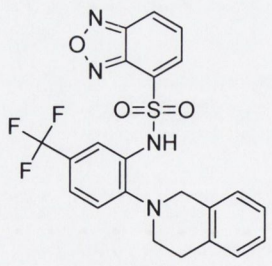
3.3.2.4 In vitro validation

Subsequently, Maybridge database of compounds was screened according to this new refined pharmacophore and the four compounds are at the moment undergoing *in vitro* testing (up-dated experimental results will be presented at the viva) The preliminary tests have been positive for MDG279 and MDG281 (structures are illustrated in Figure 3.20) yet experimental results presented in Figure 3.21 are not definitive as still repeats and dose-response assays are ongoing.



MDG279

Estimated activity 0.014μM and Best Fit 14.98



MDG281

Estimated activity 0,13 μM and Best Fit 14.04

Figure 3.20: New Maybridge hit structures

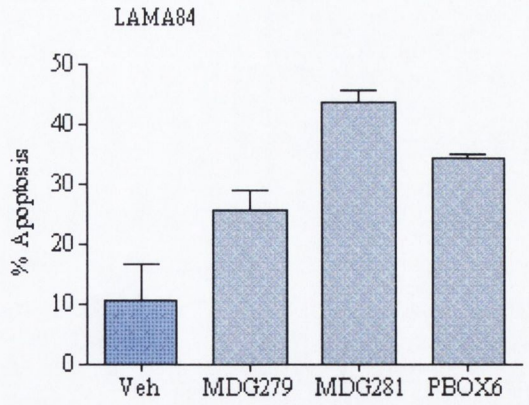


Figure 3.21: Effects of hits on Leukaemia cell lines,
Cells were treated for 48 hours at 50μM. Bars present the percent of apoptosis and are the mean of three independent experiments

As an example of fit PBOX 16 was mapped on the pharmacophore, illustrated in Figure, and obtained an optimal fit value of Best Fit 16.31. Despite the new hits IC_{50}s have not yet been evaluated to be compared to the estimated values, both structures were aligned with PBOX 16 as shown in Figure 3.22., MDG279 and MDG281 obtained respective lower Best Fit values of 14,98 and 14,04. Also PBOX6 which was used as positive control in the first *in vitro* test reached a Best Fit value of 13,01 (estimated activity 1,3μM). The preliminary experimental data agreed with the ranking of the fit values as both MDG compounds presented better fits than PBOX6 but lower than PBOX16.

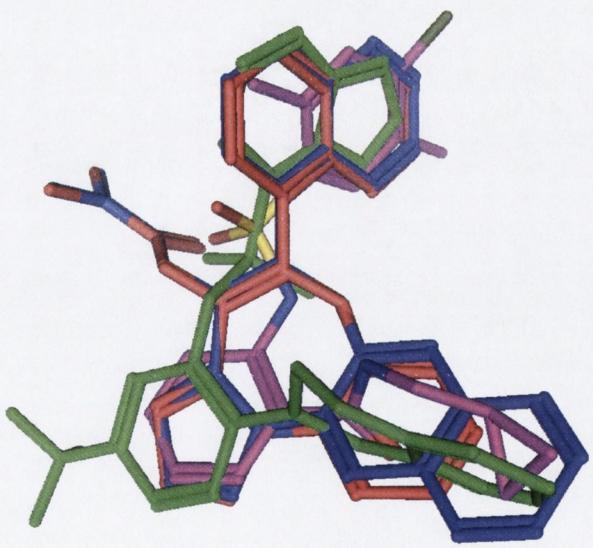


Figure 3.22: Alignment PBOX6 in red, PBOX16 in blue, MDG279 in pink and MDG281 in green

3.4 Conclusion

This chapter presented a method for selecting compounds similar to PBOX active which can be used as a haystack when the quantity or quality of compounds does not allow the gathering of the traditionally used "testing set". The most successful method in similarity selection appeared to be a Bayesian method based on simple 2D descriptors. The validation of this method using database of compounds which target were known, suggested the estrogen receptor β isoform and PDGFR as two potentially novel targets of PBOX compounds possibly finding applications in breast cancer and also many other cancers associated with PDGFR activities. Also a particular examination of HIV-IN as a target was suggested, and currently compounds very close to PBOX are already undergoing experimental investigations for the enhancement of current tri-therapy used on infected patients.

Thus, the shape based method refined by electrostatic matching was *in silico* and *in vivo* validated as a screening protocol as it led to the identification of ten novel hits. In addition, catalyst HipHop algorithm based on the attribution of only common features was probed for predicting these hits. As a consequence, with more confidence a refined pharmacophore using exclusion volumes was elaborated, *in silico* validated and almost fully *in vitro* validated. This last pharmacophore performed optimally in selecting active compounds and presented more diversity of hits. It could be used to screen new databases of compounds as the final actual hit rate is fully dependant on the screened database. In other words, if there is no active scaffold to find in the vendor database either another database should be screened or a combinatorial library of compounds in order to synthesise from 'scratch' novel potential hit molecules. This pharmacophore was used to screen a database of analogs of the ten hits previously found which are also undergoing further *in vitro* testing.

Tubulin presents more than one potential site for small molecules and therefore leaves PBOX compounds as binding orphans. In the next chapter a novel approach to determine the actual binding site will be presented using the new refined pharmacophore elaborated here with exclusion volumes.

References

1. Mc Gee, M. M.; Campiani, G.; Ramunno, A.; Fattorusso, C.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Pyrrolo-1,5-benzoxazepines induce apoptosis in chronic myelogenous leukemia (CML) cells by bypassing the apoptotic suppressor bcr-abl. *J Pharmacol Exp Ther* **2001,** 296, (1), 31-40.
2. Mulligan, J. M.; Greene, L. M.; Cloonan, S.; Mc Gee, M. M.; Onnis, V.; Campiani, G.; Fattorusso, C.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Identification of tubulin as the molecular target of proapoptotic pyrrolo-1,5-benzoxazepines. *Mol Pharmacol* **2006,** 70, (1), 60-70.
3. Zisterer, D. M.; Hance, N.; Campiani, G.; Garofalo, A.; Nacci, V.; Williams, D. C., Antiproliferative action of pyrrolobenzoxazepine derivatives in cultured cells: absence of correlation with binding to the peripheral-type benzodiazepine binding site. *Biochem Pharmacol* **1998,** 55, (4), 397-403.
4. Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A., How to do an evaluation: pitfalls and traps. *J Comput Aided Mol Des* **2008,** 22, (3-4), 179-90.
5. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T., Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes? *J Comput Aided Mol Des* **2008,** 22, (3-4), 213-28.
6. Schuster, D.; Steindl, T. M.; Langer, T., Predicting drug metabolism induction in silico. *Curr Top Med Chem* **2006,** 6, (15), 1627-40.
7. 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.
8. 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 Deliv Rev* **2001,** 46, (1-3), 3-26.
9. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **2003,** 8, (19), 876-7.
10. Hou, T. J.; Xu, X. J., ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J Chem Inf Comput Sci* **2003,** 43, (6), 2137-52.
11. Mitchelll, T. M., Machine Learning. *Ed.; McGraw-Hill: New York,* **1997**.
12. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S., Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J Chem Inf Comput Sci* **2004,** 44, (1), 170-8.
13. SPECS vendor database, *www.specs.net*.
14. catalyst Version 4.11 available from Accelrys Inc. San Diego.CA.USA
15. OMEGA2, developed by Openeye scientific software. *www.eyesopen.com*.
16. Rapid Overlay of Chemical Structures, developed by Openeye scientific software. *www.eyesopen.com*.
17. Jonas Boström, K. B., Thomas Elebringb, Peter J. Greasleya and Michael Wilstermanna, Scaffold hopping, synthesis and structure–activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: A novel series of CB1 receptor antagonists. *Bioorganic & Medicinal Chemistry* **2006,** Volume 15, (12), Pages 4077-4084.
18. Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **2005,** 48, (5), 1489-95.
19. EON electrostatic refinement, developed by Openeye scientific software. *www.eyesopen.com*.
20. Kurogi, Y.; Guner, O. F., Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem* **2001,** 8, (9), 1035-55.
21. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P., Identification of common functional configurations among molecules. *J Chem Inf Comput Sci* **1996,** 36, (3), 563-71.
22. Molecular Operating Environment, developed by Chemical Computing Group Inc. *www.chemcomp.com*.
23. Oprea, T. I., Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* **2000,** 14, (3), 251-64.
24. Ertl, P.; Rohde, B.; Selzer, P., Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem* **2000,** 43, (20), 3714-7.
25. CRC Handbook of Chemistry and Physics, *CRC Press* **1994**.
26. Labute, P., MOE LogP(Octanol/Water) Model. unpublished. . *Source code in $MOE/lib/svl/quasar.svl/q_logp.svl* **1998**.

27. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J., ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J Chem Inf Comput Sci* **2004,** 44, (1), 266-75.
28. Labute, P., MOE Molar Refractivity Model. *unpublished,Source code in $MOE/lib/svl/quasar.svl/q_mref.svl* **1998**.
29. Hall, L. H., Kier, L.B., The Nature of Structure-Activity Relationships and Their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **1977,** 12, 307.
30. Hall, L. H., Kier, L.B., The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. *Reviews of Computational Chemistry.* **1991,** 2.
31. Balaban, A. T., Balaban's connectivity topological index. **1982**.
32. Balaban, A. T., Five New Topological Indices for the Branching of Tree-Like Graphs. *Theoretica Chimica Acta.* **1979,** 53, 355-375
33. Wiener, H., Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947,** 69, 17-20.
34. Pipeline Pilot, developed by Scitegic, a product from Accelrys. *www.accelrys.com/products/scitegic/* Pipeline client version 6.0.2.0 , Pipeline Pilot Server version 6.5.1.0.
35. SIMCA-P developed by Umetrics, *www.umetrics.com* version 11.0..0.0,Jun 12 2005.
36. MDL Drug Data Report, MDL ISIS/HOST software,. *MDL Information Systems, Inc., San Leandro, CA.*
37. Morgan, H. L., The Generation of Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Services. *J. Chem. Doc.,* **1965,,** 5, 1, 07-113.
38. Smellie A. Teig, S. L. T., P. , Poling: Promoting conformational variation. *J. Comp. Chem* **1995,** 16, 171-187
39. 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, (6), 1908-19.
40. Halgren, T. A., The Merck Force Field,. *J. Comp. Chem.* **1996,** 17 490-512, 520-552, 553-586, 587-615, 616-641.
41. Zap electrostatics toolkit, developed by Openeye scientific software. *www.eyesopen.com*.
42. Jon M. Sutter, Users Group Meeting Catalyst, Frankfurt **2006**.
43. Li, H. S., J.; Hoffmann, R,, An Automated System for Generating 3D Predictive Pharmacophore Models. Pharmacophore Perception, development, and Use in Drug Design. *International University Line: La Jolla, CA, 2000*, 172-189.
44. Sutter, J. G., O.; Hoffmann, R.; Li, H.; Waldman, Effect of Variable Weights and Tolerances on Predictive Model Generation. Pharmacophore Perception, Development, and Use in Drug Design. *International University Line: La Jolla, CA, 2000*, 499-511.
45. Catalyst Tutorials Version 4.11 *atalyst Version 4.11 available from Accelrys Inc. San Diego.CA.USA ,*
.
46. Fattorusso, C.; Gemma, S.; Butini, S.; Huleatt, P.; Catalanotti, B.; Persico, M.; De Angelis, M.; Fiorini, I.; Nacci, V.; Ramunno, A.; Rodriquez, M.; Greco, G.; Novellino, E.; Bergamini, A.; Marini, S.; Coletta, M.; Maga, G.; Spadari, S.; Campiani, G., Specific targeting highly conserved residues in the HIV-1 reverse transcriptase primer grip region. Design, synthesis, and biological evaluation of novel, potent, and broad spectrum NNRTIs with antiviral activity. *J Med Chem* **2005,** 48, (23), 7153-65.
47. Campiani, G.; Nacci, V.; Fiorini, I.; De Filippis, M. P.; Garofalo, A.; Greco, G.; Novellino, E.; Altamura, S.; Di Renzo, L., Pyrrolobenzothiazepinones and pyrrolobenzoxazepinones: novel and specific non-nucleoside HIV-1 reverse transcriptase inhibitors with antiviral activity. *J Med Chem* **1996,** 39, (14), 2672-80.

Chapter IV

Investigation of the PBOX binding site in tubulin

Introduction

The previous ligand based approach, when applied to the PBOX series of compounds, was successful in finding new active compounds. One possibility for optimising the new scaffolds would be to increase the screening information available using the receptor structure. Recently, PBOX compounds have been shown to induce apoptosis and to arrest the cell cycle in the G2/M phase (more extensively explained in Chapter five) by interacting with the microtubular (MTs) network[1]. Although some PBOX compounds have been shown to inhibit tubulin polymerisation *in vitro* their exact binding location remains unknown. This chapter uses a computational approach to scan the tubulin structure for potential binding sites. For this purpose, an examination of the receptor structure must be undertaken and an exhaustive enumeration of its multiple binding sites explored.

Microtubules are the key components of the cytoskeleton and are found in all eukaryote cells. They are hollow tubes composed of polymerized dimers of α and β tubulin. They perform a variety of cellular functions, such as the maintenance of cell shape, structural integrity, cell signalling and cellular transport *i.e.* through vesicles or protein complexes within the cytoplasm[2, 3]. They are also extremely important in cell regulation and fundamental in cell division.

Cell division and proliferation occurs through the process of mitosis. During this ordered list of events, the complete genetic information is duplicated and split into two twin cells. In the numerous shapes that microtubules can form[4], the mitotic spindle arrangement is responsible for the separation, segregation and distribution of the duplicated chromosomes to the two daughter cells. The following Figure 4.1 illustrates the different phases of the mitotic process and emphasizes the evolution of the microtubular network.

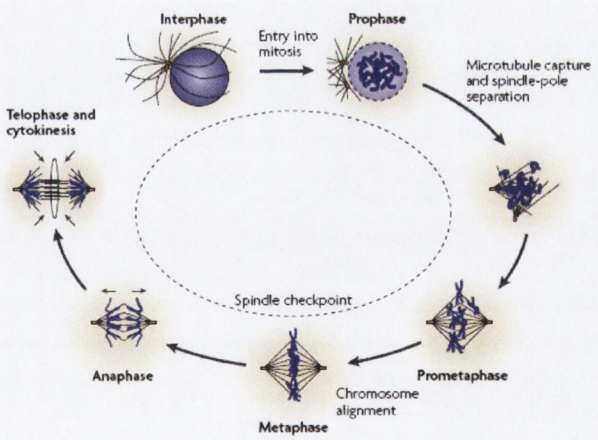


Figure 4.1: Phases of Mitosis.[5]

The reliable performance of microtubular functions, in particular for cell division, requires microtubular filament dynamics, which translates to a growth or shortening of microtubules[6]. Their dynamic apogee arises when the microtubule network arranges to form the mitotic spindle in contrast for instance to the interphase, which is composed of a more distributed state within the cytoplasm. Any modification of their dynamics during mitosis leads to the formation of damaged cells, doomed to death through apoptosis *ie*. programmed cell death. Targeting tubulin with small molecules can perturb the normal course of mitotic events and stop cell division. In this light, microtubular disruption appears to be a novel strategy to combat cancer.

Cancer is a collection of hundreds of diseases in which cells multiply abnormally. They avoid apoptosis, keep dividing abnormally and can be invasive or even metastatic. The involvement of MTs during cell division makes it an ideal target as this component of the mitotic spindle is found in all type of cells[7]. Several tubulin disrupting agents are already included in chemotherapies or are under clinical trials, the best known being paclitaxel (also called taxol).

The number of publications related to tubulin are constantly increasing especially in the modelling area since the first release of the 3D structure of a tubulin-taxol complex in 1998 by Nogales *et al.*[8] and after the release of the 3D structure describing a second site, termed the colchicine site[9]. Progressively, crystallograpic data expanded from the tubulin complex with taxol, to the colchicine complex and lately to a third site known as the vinblastine site was described[10]. Despite these new findings, the most probed site remains the taxol site for computational approaches (see Figure 4.2).

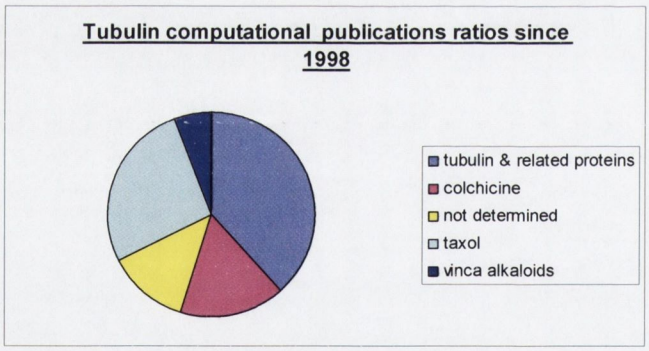


Figure 4.2: Percent of computational publication on tubulin specific site binder since 1998-2008
Pubmed key word search :"computational", "virtual","screening","pharmacophore",dock", followed by an abstract reading selection.

In this period, where experimental High Throughput technology is drastically improving, there was a large increase in the discovery of novel molecules with an anti-proliferative potency (*i.e.* National Cancer Institute Database) as well as discovery of tubulin binders to yet unknown binding sites, which

will be discussed in a following section. Novel elements recently found on the tubulin structure, such as the description of the interstice between heterodimers, could be linked to the target of some of these apoptotic compounds. A mapping of the binding sites of these compounds on tubulin would give supplementary information on the type of interactions involved and consequently aid in rationally modifying and optimizing drug structures to enhance their effects.

This chapter maps PBOX compounds onto tubulin binding sites using molecular modelling and statistical tools. Thus, it aims to propose possible binding sites in tubulin that could give rise to a new virtual screening approach *i.e.* a docking approach using newer inhibitors. A general overview of microtubules in a more biological context will first be given. Indeed, a complete understanding of how microtubules function is essential to describe in detail all of the locations and types of interactions found in the tubulin binding sites known to date.

4.1. Microtubule structure and dynamics

Microtubules are long hollow tubes of approximately 25 nm diameter. *In vitro,* each consists of less than nine to more than 16 protofilaments. The number of protofilaments is variable but *in-vivo* is almost always 13. Each protofilament is actually a linear polymer, itself constituted of an alternation of α-β dimers bound head-to-tail as shown in Figure 4.3.

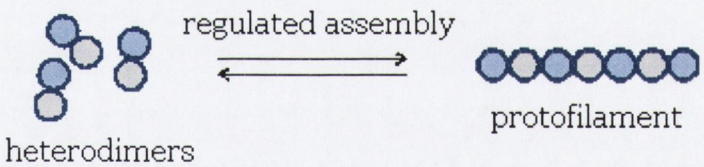


Figure 4.3: Microtubule protofilament where blue indicates an α monomer and grey a β monomer

In vivo, protofilaments run in a straight direction along the microtubule, facilitating the linear movement of microtubular associated proteins (MAPs) along microtubules. In this way, path deviations are avoided for optimal and efficient vesicle transport[11, 12]. Monomers in adjacent protofilaments are slightly staggered so that they form a set of shallow helices as shown in Figure 4.4.

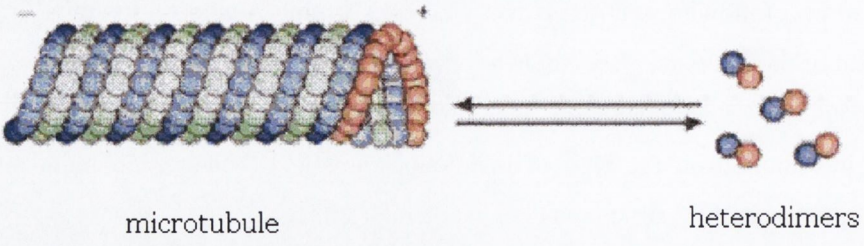


Figure 4.4: Microtubule stable phase[7]

Among other things, microtubular dynamics are driven by the hydrolysis of guanine nucleotides. Both α and β subunits bind a guanine nucleotide however the β subunit differs from the α subunit because it contains an exchangeable guanine nucleotide site (E site), while the alpha subunit contains a non-exchangable site (N site). In addition the β subunit site catalyzes the hydrolysis of guanosine triphosphate (GTP) to guanosine diphosphate (GDP) and inorganic phosphate. Inside the cell, tubulin dimers exist as microtubules and as free dimers, and in both cases the α subunit is always found with a non-exchangeable GTP. I will refer to tubulin dimers as GTP-tubulin or GDP-tubulin, depending on the nucleotide bound to the E site of the β subunit.

Finally, all protofilaments are parallel, in lateral contact with each other and orient their αβ-dimers in an identical fashion. Consequently, they have a polarity. The plus end with respect to growth is located at the β end whereas the α end is the minus end with respect to growth. This polarity plays an essential role in the function of microtubules and influences their dynamics.

Microtubular mechanisms of shortening and lengthening are respectively called polymerisation and depolymerization and are in general slow for the growing phase and rapid for the shortening one as illustrated in Figure 4.5. Microtubular dynamics are stable when there is no detectable polymerisation or depolymerization. However, MTs are unstable when they switch between polymerisation and depolymerization[13]. The passage from the stable state or from the polymerizing state to the depolymerizing state is called "catastrophe" whereas the transition from depolymerization to polymerisation is termed "rescue". Additionally, "Treadmilling"[6, 14] is the type of behaviour consisting of a net growth at the positive end and an equal net shrinkage at the negative end.

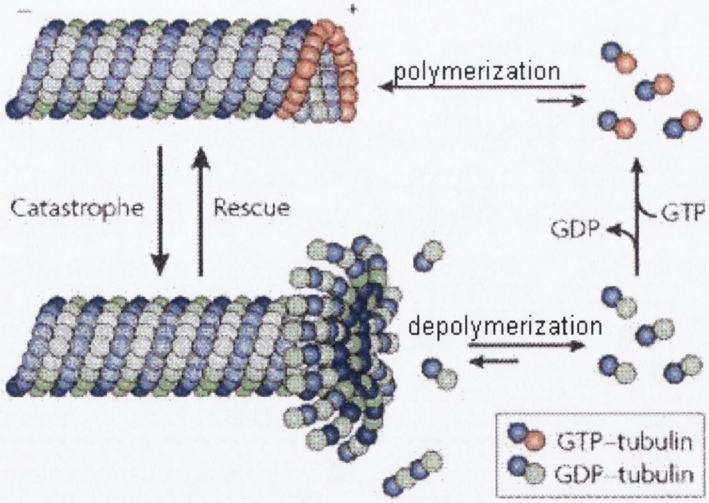


Figure 4.5: Microtubules unstable phases[7]

When the last dimer of a microtubule is found with a GTP bound, the protofilament is protected from unravelling so that this last dimer is called the GTP cap (Figure 4.4 and 4.5 with the GTP cap in red). In contrast to the GTP-bound state, if GDP is bound to the plus ends of the tubulin protofilaments, microtubules are unstable and will depolymerise unless "rescued". The principal event responsible for unbalancing the steady state of disassembly/assembly (dimer to polymer and vice versa) is nucleotide hydrolysis. GTP hydrolysis, which occurs at either the last dimer bound or close to the penultimate dimer, depending on the free tubulin dimer concentration. Consequently, GDP will always be part of the internal β-subunits of protofilaments or spirals and rings[15]. Spirals and rings (Figure 4.6(c)) are relatively stable depolymerization products, containing some γ-tubulin, and accounting for only a small proportion of the tubulin derived from the depolymerized microtubules (Figure (a)/(b)), along with tubulin dimers, which account for the majority of the tubulin derived from the depolymerized microtubules.

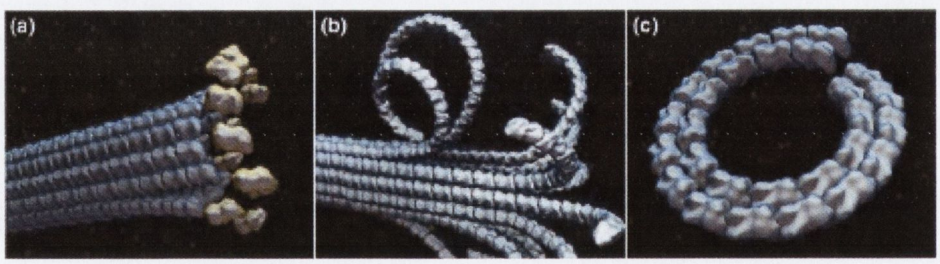


Figure 4.6: Microtubule rings formation[16]

(a) Loss of GTP cap, (b) Peeling curled protofilament MT and (c) ring disassembly intermediate.

4.2 Tubulin sequences

In the majority of eukaryote organisms more than one isoform and isotype of tubulin can be found. The existence of multiple forms of tubulin within a species and between species has suggested a relationship between their structures and their physiological roles[17]. Isotype expressions vary amongst cell types in tissues. Consequently, targeting a specific type of tubulin with a specific inhibitor, suggests that a particular type of cell might be eradicated. Consequently, one of the main pharmaceutical applications of interest is cancer research.

In addition, most organisms have multiple genes, which encode distinct isoforms or isotypes of tubulin. Usually the panel of tubulin types is reduced for simple organisms such as fungi or parasites so that targeting a particular isotype could, theoretically, lead to the elimination of a parasite within an infected mammalian host.

4.2.1 Role of Tubulin Isotypes

In eukaryotic cells, several isotypes and homologues have been identified and biochemical tools for recognition such as specific antibodies[18 19, 20] are in constant development.

All of the isotypes need to be present in order to coordinate microtubular dynamics in time and space. In fact, the role of tubulin seems to depend on the particular isotype present[21]. For instance, during the nucleation phase of MTs, γ tubulin, which is found in centrosomal MTs helps microtubular positioning [22]. Self-assembly suggests the presence of two extra types of tubulin δ and ε [23]. Together with the two last isotypes, η and ζ, whose roles are not yet fully understood, δ and ε tubulin could be involved in forming the structure of centrioles and cilia[24]. Moreover, β-tubulin isotypes modify microtubular dynamics[25] so that cells could adjust the rate of switching between polymerization and depolymerization of an individual MT by producing a particular tubulin type in excess.

Finally, the *in vivo* arrangement of MTs into 13 protofilaments appears to be regulated by the type of tubulin. In other words, MTs of more than 13-15 protofilaments appear when the α/β sequence is altered as for example in neurons [26, 27]. To date, in *Homo sapiens* (Hs), at least ten types of β-tubulin and eight types of α-tubulin have been identified and at least 5 isoforms (α, β, γ, δ, ε) are known. To date the three last isotypes γ, δ, ε have not been studied as much detail as the α an β isoforms. Alignment of the available Hs α-tubulin sequences, including some mRNA sequences, shows 91.6 % identity and 88.7 % similarity. The Hs β-tubulins have an 85.6 % identity and 94.9 % similarity. The difference between each type is mainly found at or close to the C-terminus, which could be involved functionally by recruiting other proteins from the environment and guiding them to specific points in microtubular arrays at specific times.

4.2.2 Homology and origins

Any of the tubulin isotypes, α, β, γ, δ, ε, ζ and η, show 22 to 30 % sequence identity to any other tubulin isotype. Presumably, the sequence divergence of the original precursor tubulin occurred by duplication and then independent evolution of the resulting copies over time. In addition the development of their individual specific functions are probably not due to the conserved structural similarity between the tubulin isotypes but, rather, is related to the non-homologous parts of each isotype.

One of the most studied orthologues to tubulin is the protein Fitsz, which was the first protein of the prokaryotic cytoskeleton to be identified. Its 3D structure, obtained by electron crystallography[28], confirmed the remarkable structural conservation of this protein to tubulin[29, 30]. The two first helices and strand parts of the protein can be superimposed almost perfectly on the solved bovine tubulin structure even if some loops are shorter and some residues are missing at both N- and C-terminal parts of the protein compared to tubulin. As the bovine sequences of α- and β-tubulin also exhibit a remarkable sequence identity to the respective Hs sequences, the use of the 3D bovine structures available seemed reasonable. Hence no homology model was built for use in this thesis; rather, the bovine 3D crystal structure was employed.

The crystallized[31] bacterial protein from *Prosthecobacte* (BtubA and BtubB) shows a similarity of 35 % to Hs tubulin and is believed to have had the same function before aquiring the reduced capacity of forming pairs and bundles of protofilaments rather than cylindrical MTs.

Homology model comparison amongst all available tubulin types showed that sequence and structure conservation are linked to conservation of physical properties[32]. For instance, the dipole moment of tubulin was shown to influence microtubular stability and dynamics through longitudinal interactions[33]. It was then suggested that electrostatic effects could also govern MAPs or tubulin dimer docking by generating the adequate steric configuration[32].

An important consequence of the existence of isotypes is drug resistance with respect to the level of expression of a particular tubulin type[34, 35]. Changes of these levels, the existence of diverse isotypes, mutations or MAPs alterations [36], all can perturb microtubular dynamics and have largely been studied in the context of drug resistance[37-39]. Multi-drug resistance (MDR) seriously limits the use of compounds such as the Vinca alkaloids[35] or the taxanes[40] to be discussed later.

4.3 Tubulin 3D structure and interactions

To examine further the microtubules' mechanism of action and for more aspects on drug-tubulin interactions, it is important next, to focus on the actual 3D structure of tubulin at an atomistic level.

4.3.1 Crystallographic data

All seven of the available crystal structures for α/β tubulin heterodimers employed bovine tubulin for collecting x-ray data and four of these used the bovine sequence for constructing the model while the remaining two used the pig sequence for constructing the model (Table 4.1).

PDB code	Chains	Ligands	Resolution Å	Protein Source	Sequence used in model	Year
1Z2B[10]	ABAB SLD	GTP,GDP, vinblastine DAMA –colchicine	4.10	Bos Torus (bovine)	Bos Torus (bovine)	2005
1SA0[9]	ABAB SLD	GTP,GDP DAMA –colchicine	3.58	Bos Torus (bovine)	Bos Torus (bovine)	2004
1SA1[9]	ABAB SLD	GTP,GDP Podophyllotoxin	4.20	Bos Torus (bovine)	Bos Torus (bovine)	2004
1JFF[41]	AB	GDP,GTP Taxol	3.50	Bos Torus (bovine)	Sus scrofa (pig)	2001
1FFX[42]	ABAB SLD	GDP,GTP	3.95	Bos Torus (bovine)	Bos Torus (bovine)	2000
1TVK[43]	AB	GDP,GTP Epothilone A	2.89	Bos Torus (bovine)	Bos Torus (bovine)	2004
1TUB[8]	AB	GDP,GTP Taxotere	3.70	Bos Torus (bovine)	Sus scrofa (pig)	1998

Table 4.1: .α/β Tubulin crystal structures

AB: αβ-tubulin dimer, SLD: Stathmin-like domain, GDP,GTP: Guanosine di & tri phosphate.

The mobility of tubulin within the protein crystal due to its inherent dynamics makes it difficult to determine crystal structures at high resolution. The first image of microtubule disposition was obtained by Electron Microscopy (EM) [44-46] and then by X-ray diffraction methods from oriented microtubules at ~18 Å resolution [47]. Electron crystallography (EC) of zinc-induced two-dimensional sheets

(monolayer crystals) of the protein also indicated a three-dimensional model of tubulin at a 6.5 Å resolution [48].

Owing to the high structural similarity of the α and β monomers (sequence identity of 40%), their clear discernment could only be done at an atomic level. The release of the three-dimensional crystallographic structure at a 3.7 Å resolution constituted a major advancement in the understanding of microtubular structure and function[8]. This was obtained using zinc-induced flat 2D sheets of tubulin (monolayer crystals) using Docetaxel in the crystallographic mixture for receptor stabilization. This procedure utilized mixed isotypes of tubulin and resulted in strict sequencial alternating protofilament polarity within the sheet (1TUB).

The next structure (1JFF) was obtained using Paclitaxel as the stabilizing drug. It refined the previous structure, filling in some residues missing as a consequence of the sometimes unclear electron density map. To date, the best resolution structure was obtained at 2.89 Å using another stabilizing agent: Epothilone[41]. This compound binds to the same binding site to which taxol binds (discussed later) but in a different mode making it an important and fast growing class of anti-mitotic agents (1TVK).

Due to the constant improvement of electron crystallography techniques, more structures of the heterodimer were elucidated including double dimer, apo and co-crystallized structures with diverse drugs.

When destabilizers were co-crystallized in the apo form, a stathmin protein was substituted for the stabilizing agent taxol previously described. Stathmin protein has the ability of sequestering heterodimers and keeping them together even if a depolymerization drug is bound. For instance, the stathmin-like domain (SLD) of RB3(a GTPase activating protein), was used when crystallizing the apo protein (1FFX) [42] or again when co-crystallising complexes with destabilizer compounds such as colchicine or podophillotoxin (1SA0 & 1SA1) [9].

The latest structure obtained at 4.1 Å resolution, used the same stabilizing method to reveal the previously ambiguous vinblastine site and comprises also two molecules of colchicine bound (1Z2B)[10].

In addition, other related structures have been found such as the human structure of the γ tubulin isotype (1Z5V&1Z5W)[49], prokaryotic structures such as *Prosthectobacter dejongeii* α- and β-tubulin complex (respectively 2BTQ & 2BTO)[31] and several structures of the prokaryotic tubulin homologue, FtsZ.

Finally, protein/protein docking techniques in association with cryo-electron microscopy were combined with the best solved 3D protein structures to explain better protein-protein interactions. An example is the latest description of the nucleotide binding pocket of the monomeric kenesin motor KIF1A (2HXF & 2HXH)[50].

Combining data from almost all of the crystal data summarised in Table 4.1. the α-β tubulin dimer will now be described in details.

4.3.2 Domain descriptions

Tubulin α and β monomers contain ~450 amino acids each. They are structurally very similar having an identical principal structure constituting of a central core helix 7 (H7) and on either side two globular parts consisting of β sheets surrounded by helices. Their structures are compact and the core of the structure contains two β sheets of six (B1-6) and four (B7-10) strands flanked by twelve α helices (H1-12) as shown in Figure 4.7. The structure can be split into three different domains according to their respective functional roles: the amino-terminal domain, an intermediate domain and the carboxy-terminal domain.

According to the residue numbering of Nogales *et al.* 1998[8], nucleotide-binding is localised at the N-terminal end and approximately extends from residue 1 to 205. The N-terminal is the larger globular domain of the two and consists of alternating β strands and helices forming a Rossman fold typical of nucleotide binding proteins such as GTPases[51]. There are six parallel β-strands (B1–B6) alternating with helices (H1–H6) and there is a direct involvement of the loops or turns (T1–T6) in nucleotide binding. These loops connect each strand with the start of the next helix. The N-terminal domain of the α subunit is found at the negative end of protofilament.

The intermediate domain represents the second smaller globular part and extends from residues 207 to 283. It is in longitudinal contact with the central helix, the nucleotide base and in lateral contact with other monomers. In this middle domain of the β subunit lies the Taxol binding site coloured in red in Figure 4.7.

Finally, the carboxy-terminal domain consists of the binding site for motor proteins. The C-terminal domain extends from residue 284 to the end. It is made up of two long helices (H11 and H12) and alternating loops. This domain ends with a section of around 10-18 often badly disordered residues. These last residues are acidic and negatively charged under physiological conditions. They are also important in isotype determination.

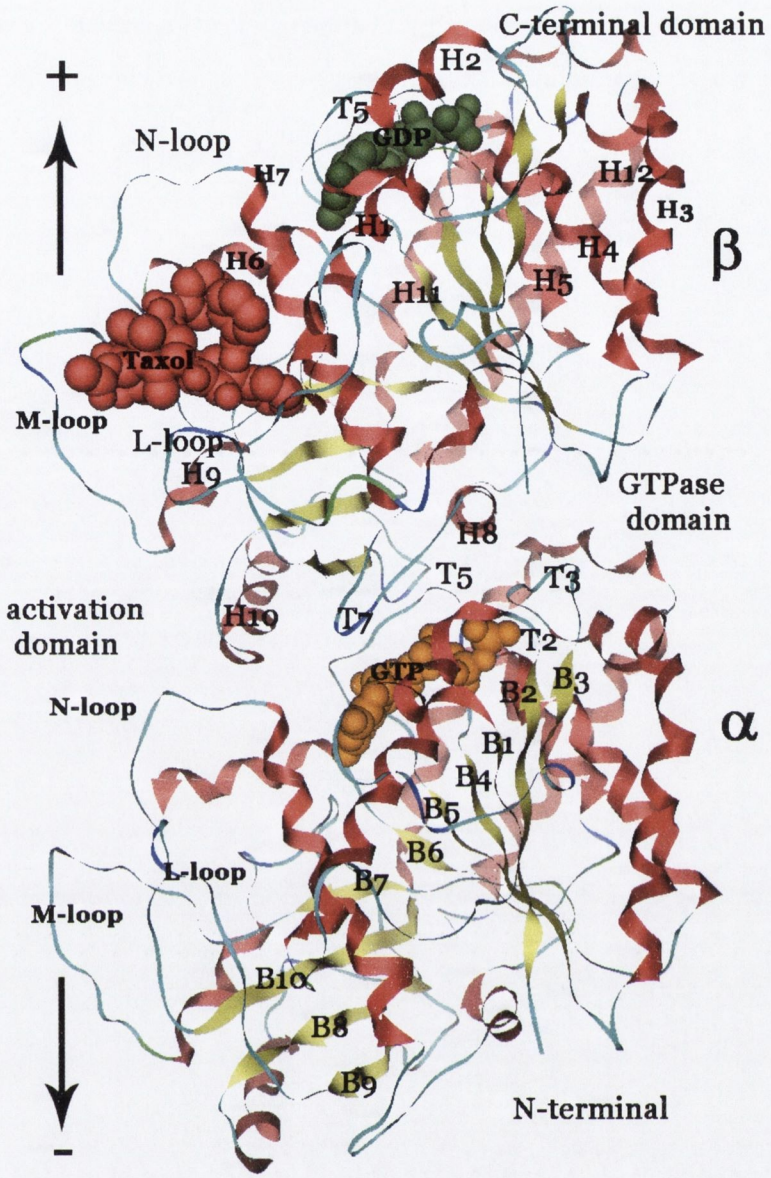
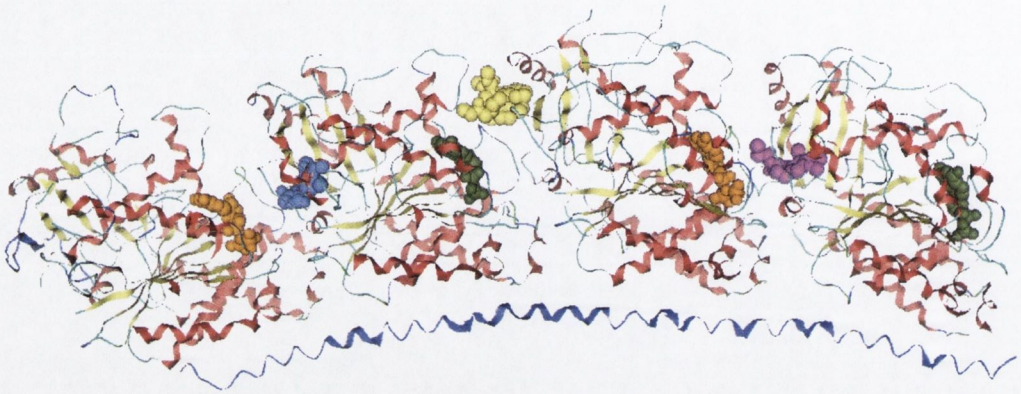


Figure 4.7: Tubulin heterodimer from the crystal structure 1JFF.[52]

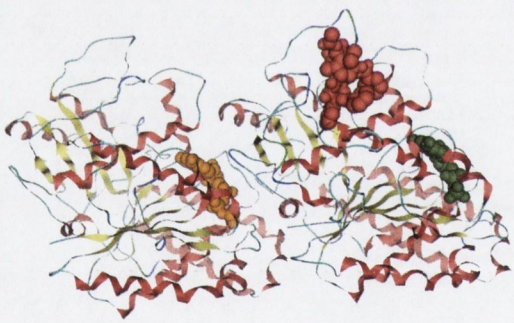
If the whole microtubule structure is compared to the crystallised dimers the general conformation remains similar. Longitudinal interactions are well known, nevertheless crystallized dimers are often found antiparallel whereas in microtubules they are parallel leaving us with some uncertainty concerning lateral interactions.

Nevertheless, a curvature at the protofilament ends and along dimers was observed. Observing either straight or curved conformations of the heterodimer structures can describe with more subtlety the movements occurring during growth or shortening as shown in Figure 4.8a&b. For instance, the whole heterodimer curvature can be observed when comparing side to side the colchicine tubulin complex

(Figure 4.8.a curved conformation) and the taxol tubulin complex (Figure 4.8.b straight conformation). Conformational changes to the protein are intrinsic to the dynamic mechanism.



(a)



(b)

Figure 4.8: Curved and straight tubulin conformations[52]

(a) Crystal structure 1Z2B of tubulin, two heterodimers in curved conformation complexed with the stathmin protein in a blue ribbon. In a space filling model are the following ligands: orange-GTP, green-GDP, blue-podophollotoxin, yellow-vinblastine and pink-colchicine. (b) Crystal structure 1JFF of tubulin in a straight conformation with taxol in a space filling model coloured in red.

4.3.3. Tubulin-GTP interactions

The GTP binding site is in direct contact with loops T1 to T6 of the GTPase domain. The non exchangeable GTP (at the N-site) lies between the two monomers α and β. It is confined in its pocket surrounded by protein, which minimises its accessibility for hydrolysis. The exchangeable β cap is at the positive end of the microtubule. There, GTP lies in its pocket in a way that it is exposed and has a large enough water contact region to be easily hydrolysed. Hydrolysis implies it conversion to GDP by releasing a phosphate. If the β cap phosphate is lost, depolymerization occurs rapidly and the curl observed in the whole protofilament could be the result of conformational changes observed in the interface where the hydrolysis occurred first[53].

During polymerisation, the loop T7 and the helix H8 from the activation domain (intermediate domain) of the approaching subunit are brought close to the phosphates of the nucleotide. Therefore, hydrolysis to GDP is enabled. This catalytic effect has been well studied in GTPase-activating protein (GAP). Once the GTP is trapped in a sandwich in the heterodimer, Glu254 of the α monomer finds itself close to the phosphate in β and could be involved in the catalysis[9] (see Figure 4.9). The fact that the corresponding residue to α Glu254 in the β monomer is of opposite charge could explain the difference in catalytic ability between the E-site and N-site of GTP[54].

As shown in Figures 4.8.a and 4.9, GTP's extra phosphate is in contact with the T4 loop of β tubulin, its hydrolysis to GDP breaks this interaction and induces conformational changes to this T4 loop. These changes are propagated to the helix in its direct vicinity, which is responsible for the lateral maintaince (dimer to dimer) of the protofilaments. As a result, lateral contacts become weaker so that there is a predisposition for detachment of the protofilaments from each other.

An explanation of the conformational changes transferred along polymers lies in the conversion to energy of the tension (or strain) of maintaining the system. Straight filaments held tightly together, as for instance in 2D sheets, can be described as being in a strained, straight state. During depolymerization, this strain is released when GTP converts to GDP and subsequently the structure will curve.

Also, the propagation of the conformational changes could then roll up sections of filaments into rings. This would add up to the massive speed of the depolymerization phase and is supported by the study of GTPases structurally similar to tubulin.

Furthermore, after hydrolysis to GDP, microtubules were observed to be shorter when they are straight due to the straightening tension that maintains them close to each other both laterally and longitudinally. The conformational change induced by hydrolysis could result in a shortening of the dimer (up to 4% of its original length)[55, 56].

Finally, the movement of some loops involved in longitudinal binding and of the core helix H7 can reflect the shortening effect and explain how hydrolysis can transmit conformational changes from β-tubulin to α-tubulin

A slow polymerization returns the system into a constrained shape responsible for the dynamic instability of microtubules.

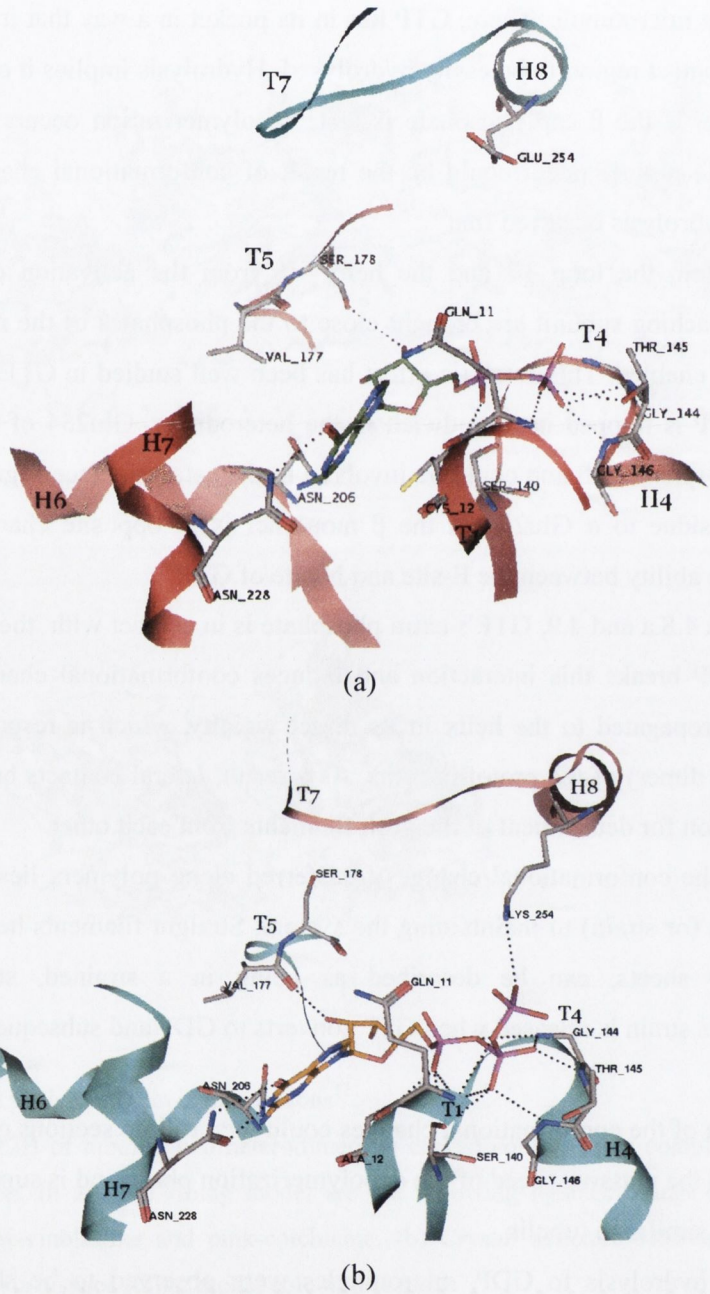


Figure 4.9: (a) GDP in the E site and (b) GTP on the right in the N site.
β tubulin is in red and α tubulin is in cyan. The pictures were generated from the 1Z2B structure[52].

4.3.4 Tubulin-tubulin interactions

Lateral interactions between protofilaments have been widely studied. The primary 3D microtubules map obtained from cryo-electron microscopy was refined by docking dimers in parallel using data from high resolution crystallography[57]. Several helical reconstitutions obtained by independent groups showed that the lateral interactions involved were mainly between the more rigid H3 helix of one dimer and the more flexible M-loops (B7-H9 loop or Microtubule loop) of the GTPase domain (activation domain) of the other dimer [57]. Basically, the region between helix H3 and the β-sheet of one monomer interacts with the GTPase domains of the other one. These docking studies are also in agreement with descriptions of the microtubule's surface consisting of an outside of longitudinal edges and lateral edges in the inner surface. The taxol binding site resides in the inside surface trapped between two parallel MTs.

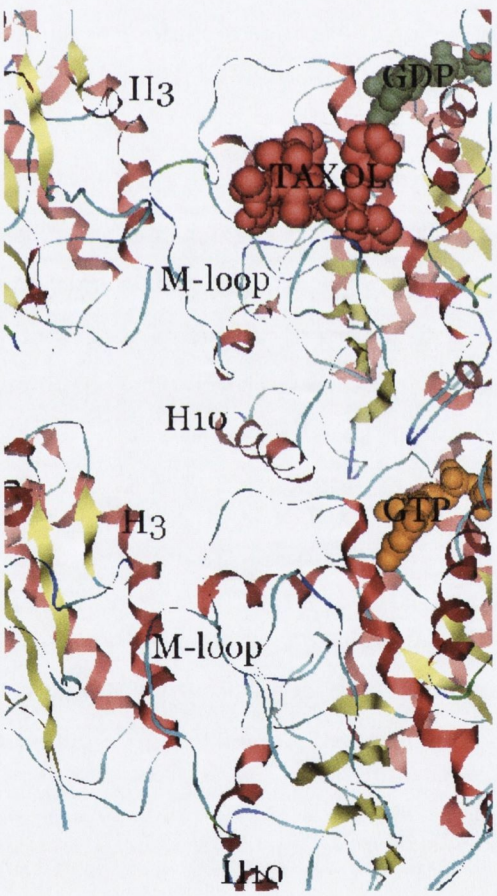


Figure 4.10 The lateral interaction between protofilaments[52].

The picture was generated using the 1JFF crystal structure. In a space filling model are coloured in green GDP, orange GTP and red Taxol.

The broad range of protofilaments per microtubule observed *in vitro*, is explained by both the M-loop flexibility allowing the angle of lateral attachment to vary and the curvature flexibility that therefore allows the diameter of the microtubule to vary.

The GDP-tubulin curved conformation does not allow polymerisation. In contrast, if GTP is bound to the end of the protofilament, the tubulin straight conformation is favourable for the approach of a free dimer to connect to its protofilament end. GDP-tubulin corresponds to a high curvature conformation favourable to depolymerization. As a result GDP is always found in the core of the protofilaments and in rings or spirals. However, lateral interactions are not to be neglected even if depolymerisation (hydrolysis from GTP to GDP) induces conformational changes that influence the longitudinal interactions.

In fact, when protofilaments are arranged into rings they bend not only at the borders between heterodimers but also between parallel protofilaments at the lateral edges [9, 15, 42, 58] as shown on Figure 4.6.

It has been suggested that GTP hydrolysis induces conformational changes to the protein, which in turn weakens both longitudinal and lateral bonding of the dimers[57]. For example, when microtubules are polymerising or depolymerising a curvature at the end of the protofilament is observed and this appears only with GDP-tubulin in rings and spirals. This shape is therefore associated with GDP-tubulin and was also verified with the swapping for non-hydrolysable GTP [59]. Furthermore, the extreme curvature induced by GDP, seems to also influence lateral contacts.

Yet, the existence of 2D sheets of protofilaments suggests that lateral interactions are more sensitive than longitudinal ones to hydrolysis. The coupling of these two directional consequences could explain the high depolymerisation speed[56].

4.3.5 Interaction with other proteins

Tubulin interacts with many other proteins including with active motor proteins for transportation of vesicle for instance. Cells that do not divide such as neurons use microtubules and microtubule-associated proteins (MAPs) to be transported along axons [60]. Figure 4.11 illustrates some of the major MAPs that interact with MTs.

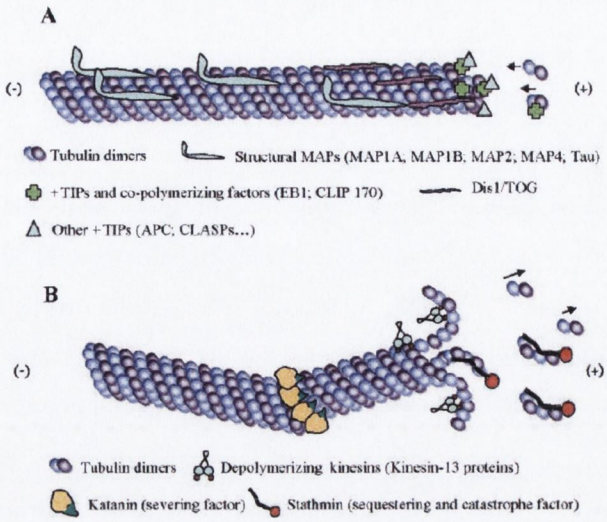


Figure 4.11: MTs and main MAPs interactions[61]

(A) A growing MT and various interactions with stabilizing factors. (B) A shortening MT and various interactions with destabilizing factors. (–) and (+) indicate the negative and the positive ends of the MT, respectively.

Microtubule dynamics can also be regulated by several proteins including stabilizer MAPs[61, 62] such as Tau protein, MAP1, MAP2, MAP4, or destabiliser MAPs such as Stathmin. They fine tune the microtubule dynamics in time and space.

In general, MAPs have repeated domains so that they can bind several tubulin dimers at the same time. For most MAPs, direct visualization of the protein/protein complex and genetics mapping, illustrated that interactions are found on the surface of the protein. Indeed, this area of tubulin is accessible from its environment and available to make contacts. Enzymatic interactions can separate the different domains of the α-β dimer. An example is the binding of Chymotrypsin or trypsin enzymes at the α-β interstice and the binding of subtilisin at the N-terminal domain. Finally, MAPs binding to tubulin are directed by kinases and phosphatases, which dephosphorylate them to induce detachment of the microtubule[63].

Some of the motor proteins are required in the folding process for dimer formation. They bind to the microtubule outer surface and perhaps at the intra dimer region where the β monomer is positioned in the dimer[64]. For instance, Kinesins such as Kinesin-13 bound to ATP, have loops interacting with β and α monomers at the interfilament region. In synergy with motor protein binding on tubulin they bend microtubules and can depolymerise them at their end[65].

Structural MAPs can bind two or more tubulin dimers at the same time to facilitate motor transport along microtubules. Tau and MAP2 or MAP4 are from the same category of MAPs, and all are

structural MAPs known to have a stiffening effect on microtubules[66]. The first two can be found in axons or dentrites and the last is their equivalent in non neuronal mammalian cells.

Tau's N-terminal negative regions are believed to be projected at the outside of the microtubule as its outside surface is negative as well. Their positive region is Proline rich and is thought to bind on the outside of the microtubule surface and to run along the protofilaments thereby playing a role in the separation of microtubules[67]. Their repeating region is less positive and is believed to bind at the inside surface of microtubules[67]. There, they might fill in inter-filament regions by binding shallow helices on tubulin. They have been shown to access a site on the β monomer overlapping with that of Taxol[68] suggesting Tau proteins as natural Taxol site substrates. This would concur with the tightening effect induced by their multi dimer binding, with their stabilising and stiffening effects. In this manner, Tau proteins help to generally maintain the microtubule structure and constrains dimers to their straight conformations[69, 70]. Another protein known as STOPs (Stable Tubulin Only Polypeptides) is thought to have its repeating region bind in a region of β monomer overlapping Taxol site. It is known to stabilize microtubules and protect them from depolymerisation induced by temperature drops[71].

Other proteins such as Doublecortin, can link several tubulin dimers by binding between protofilaments of either the same microtubule or different ones. In this way, microtubules are stabilized in a favourable position for polymerization and can maintain their specific arrangement of 13 protofilaments. However, the binding of MAPs to tubulin heterodimers is not always located at the exact same region therefore the induced effect is not always stabilizing but can also be destabilizing.

Stathmin[72] (also called OP18[73]) is known for sequestering tubulin and therefore has been used with tubulin in solution to aid in its crystallisation. Stathmin inhibits assembly but can promote hydrolysis of GTP and activates depolymerization[74]

Crystallography data explains the interactions involved when Stathmin protein comes to sequester tubulin dimers as illustrated in Figure 4.8.a. Its structure can be described as one repeating sequence motif and the usual N and C terminal region. Stathmin binds along a pair of tubulin dimers on their outside surface and the Stathmin N terminal domain caps the end of the tubulin polymer at the α tubulin negative end. Thus, stathmin could trigger catastrophe[75] . The Stathmin α helix consists of repeating regions which bind similarly to the two α monomers but slightly differently to the two β monomers. However, some of the residues involved in the interaction with Stathmin are conserved when comparing the α and β binding region sequences. These residues are equally spaced permitting the establishment of tubulin-stathmin complexes.

The comparison of the tubulin dimer stabilised by Taxol (a straight conformation) and tubulin stabilized by stathmin (a curved conformation) showed that the bending in the curve conformation corresponds to roughly 12°. The interactions between the α and β subunits are similar in both the curved and straight conformation. Helices H6-8 and the loop T5 adapt their position to conserve the

interactions between the two monomers α and β. For instance, interactions between βH8 and loops αT3, αT5 and αH11-H12 are conserved in both conformations. *In vivo,* many different curved conformations of the dimer exist although the curved structure obtained by crystallography is thought to be similar to the curved conformation of the free GDP-dimers (depolymerization products).

The Kinesin family represent a third family of MAPs binding the heterodimer and has been extensively studied[64, 65, 76]. These proteins attracted much interest as they are a large group of proteins, whose influence on microtubules could be more than regulating their dynamics. Some motors, such as Ncd that bind microtubules at the negative end, [77] [78] can depolymerise microtubules whereas other kinesins only transport other molecules to the ends of microtubules. Their possible application to oncology make them very attractive targets for drug design.[79]

Electron microscopy (EM) was applied to reconstitute in 3D the different dimer proteins of the kinesin family of motor molecules[76, 78, 80, 81]. It was observed that the kinesin orientations were constant and independent of the binding location toward the positive or negative ends[82]. The structure of the processive monomer motor protein KIF1 clarified that there are two main regions of interaction, one with the α monomer (retrieved by hybrid techniques [83, 84]) and a larger one with the β monomer comprised of the H11-H12 loop plus several exposed loops. An extra original region can be described beside the C-terminal region. K-loop Lysine residues were found close to the negatively charged C-terminal. They could allow the flexibility necessary for the generation of an intermediate state during the movement of the motor along microtubules and maintain proximity of the motor to the microtubule[50, 85]. This fashion of binding with the exception of the β binding region is believed to be general to kinesins[86, 87].

MAPs binding at microtubules ends

The C-terminal domain is relatively well known and involved in many proteins' binding so that it is an important part of tubulin. Different studies of MAPs binding to the C-terminal region concluded that there was not an universal way of MAP-tubulin binding[62]. Two modifications known to stabilise tubulin are 1) topological changes to the C-terminal to prevent surface interactions and 2) alteration of its charge. The C-terminal sequence varies considerably in the diverse tubulin isotypes. Its modification affects MAPs interactions and microtubule dynamics[88].

One of the performed modifications is its partial or total ablation using subtylisin that could promote polymerisation by preventing or weakening the MAP-tubulin interaction[89]. This modification could alter for instance some kinesin binders such as MKLP1 that could play a role in joining oppositely charged ends of microtubules. They can link tubulin heterodimers from different microtubules and target the positive C-terminal end of tubulin.

4.3.6 Tubulin Binding Sites and Ligands

The main interactions that tubulin has under natural and physiological conditions were previously described. In the last few decades targeting microtubule with new molecules has attracted much attention and overcoming microtubule drug resistance has lead to the discovery of a plethora of antitubulin agents[90-95]. Tubulin binders have widely been listed in reviews [5, 7, 96-101] and can be categorised according to either their effect on microtubule or their site of attack on tubulin. In only a few cases, their binding topology matches their effect. They can broadly be split into two groups: stabilising and destabilising compounds.

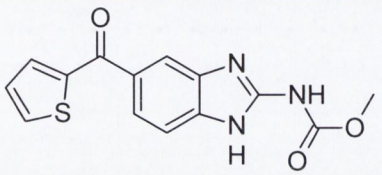
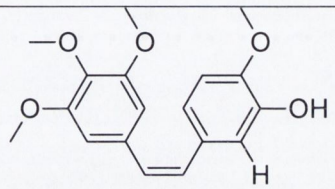
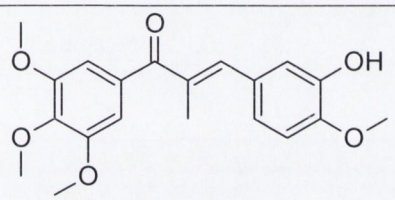
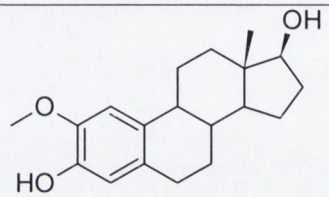
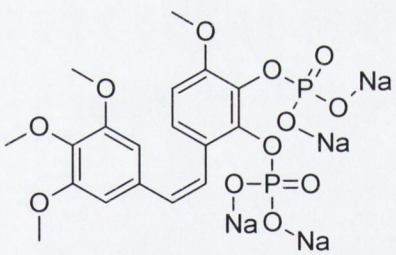
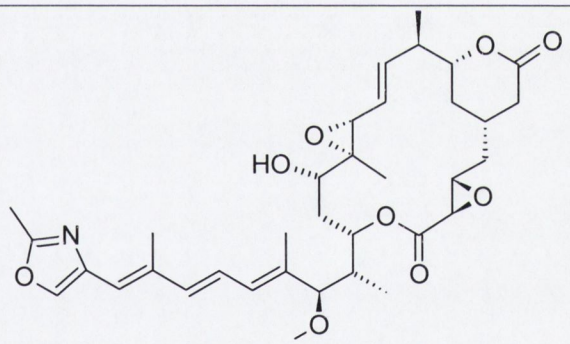
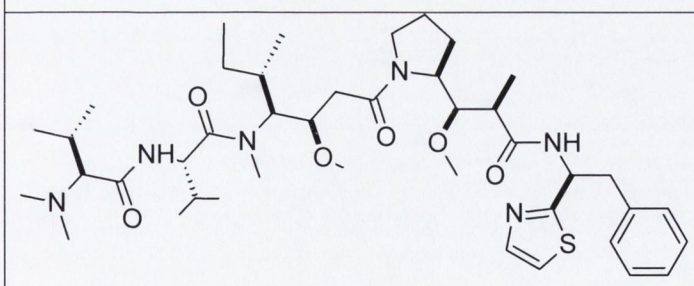
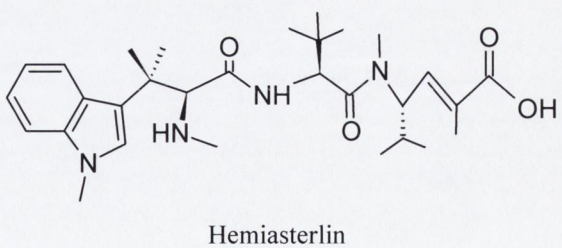
The sites well characterized by electron crystallography structures correspond to the respective regions where Taxol, colchicine and vinblastine bind. On one hand, stabilizing agents would prefer to bind to the Taxol site on β tubulin. On the other hand, destabilizing agents would bind to the interface between the α and β subunits of a single heterodimer at the colchicine site or at the interstice of two heterodimers again between the α and β subunits at the vinblastine binding site. It was also shown that other destabilisers and depolymerisers could bind in sites overlapping the vinblastine site, in the region around the drug localized at the positive end of microtubules.

In general, at higher concentrations stabilizing compounds can promote polymerization and destabilizing ones can induce depolymerization. Among the several synthetic and natural compounds that inhibit polymerization are: vinca alkaloids, colchicine, estramustine, combretastatins and haliochondrins, In addition, among the several stabilizing agents, are: taxanes, epothilones, laulimalides, discodermolides, eleutherobins, sarcodictyins, polyisoprenyl benzophenones and some steroids.

Despite the constant increase of new molecule discovery, cytotoxicity and potency remains intimately linked. Therefore, the already known scaffolds need to be improved to minimise undesirable side-effects while improving the anti-cancer potency. For instance, natural products and their derivatives represent more than two thirds of anticancer drugs and have, in general, complex macrostructures[102]. A better knowledge of their structure activity relationship (SAR) is required and therefore a mapping of the interactions known to date in each site will assist in progression.

4.4 On target destabiliser ligands

This section will discuss the two destabiliser sites known as the colchicine site and the vinblastine site. The depolymeriser compounds and destabilisers whose binding sites have not yet been determined, will also be listed. This whole collection of compounds will then be used in the computational section. A sample of the discussed ligands is shown on Figure 4.12.

DESTABILISER COMPOUNDS	
 Nocodazole	 Combrestatin A-4
 Chalcone	 2-Methoxyestradiol (2MEM2)
 Oxi-4503	 Rhizoxin
 Dolastatin 10	 Hemiasterlin

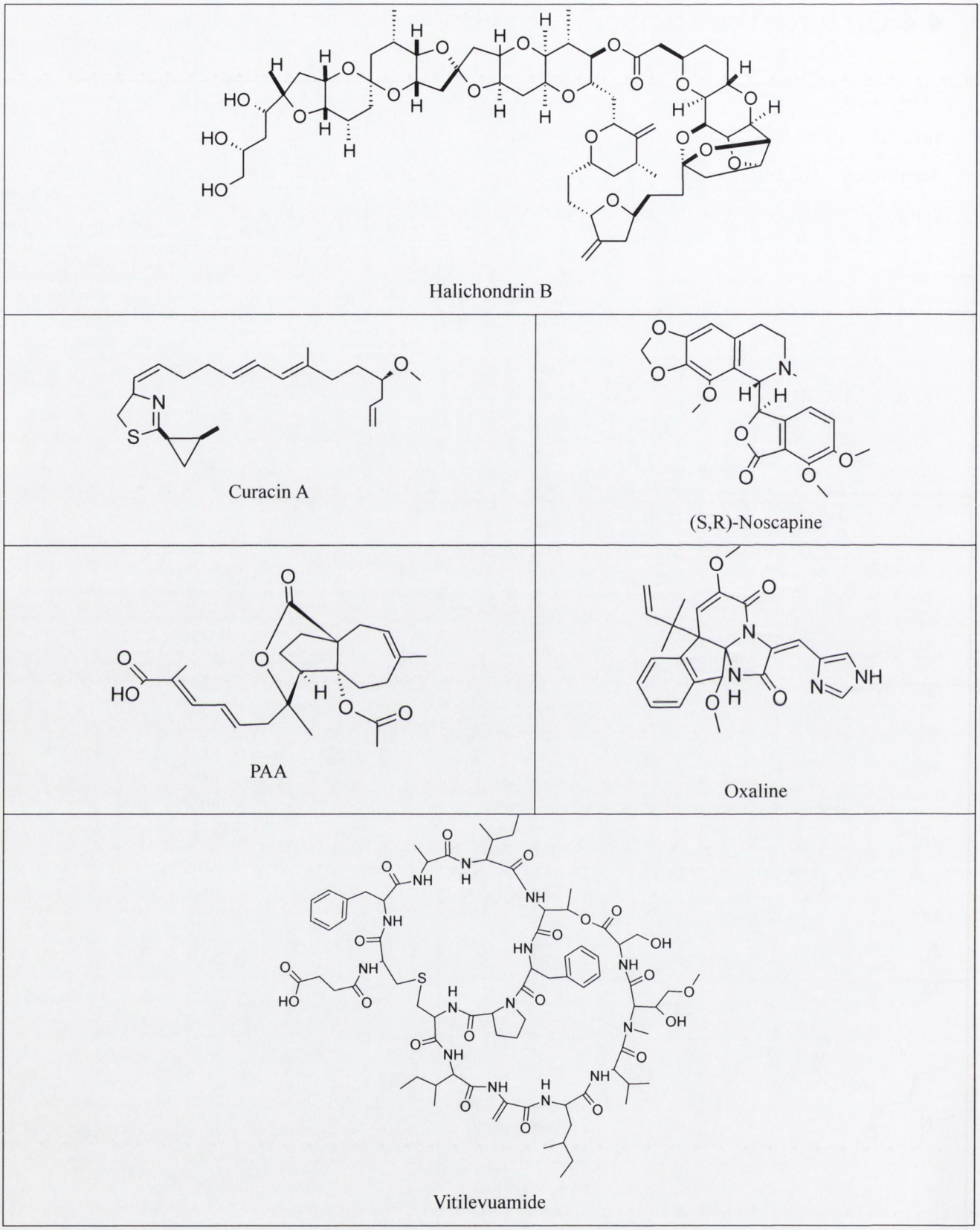


Figure 4.12: Structural diversity of tubulin destabiliser compounds.

4.4.1 The colchicine binding site

Colchicine was originally extracted from plants of the genus *Colchicum autmnale.* It was first used for the treatment of rheumatism and in particular for gout. Due to its anti-inflammatory, anti-mitotic and anti-fibrotic effects, colchicine and its derivatives were used for many disease conditions such as auto-inflammatory diseases, pseudogout, familial Mediterranian fever, cirrhosis of the liver or bile, amyloidosis and more recently is being studied in Alzheimer's dementia[103-105]. A considerable number of analogues were developed to overcome the toxicity effects such as vasculature damage causing haemorrhage and extensive necrosis[106]. Colchicine has the ability to inhibit proliferation of cancer cells by targeting microtubules. Colchicine depolymerises microtubules into dimers which show different sensitivity to the drug depending the phase of the cell cycle. Colchicines potency is in the nanomolar range and it blocks cells division in a stage called "C-mitosis" having as its main characteristic the total absence of spindle apparatus[107].

Binding description

The mechanism of action of colchicine depolymerisation involves a pre-formation of what is a tubulin-colchicine complex (TC-complex). Colchicine has a high affinity for soluble tubulin and the TC-complex formed is being added to the end of a protofilament prior to depolymerisation. This agrees with the fact that colchicine acts at a substoichiometric level.

First there is a reversible pre-equilibrium state between colchicine and tubulin which induces conformational changes. Then, colchicine binding becomes almost irreversible and is known to induce a conformational change in both the protein and the drug itself [108-110]. The protein develops a more curved conformation due to the local conformational changes propagated toward the positive end. This curved conformation sterically hampers growth. It is believed that the addition at the end of the protofilament of the TC-complex induces reduction of polymerization rather than depolymerisation. However, at higher concentrations depolymerisation is induced.

Recently, electron crystallography elucidated two tubulin structures with DAMA-colchicine (a modify colchicine at its B ring indicated on Figure 4.13) and the stathmin-like domain (SLD) of RB3: 1SA0 and 1Z2B. Formation of the TC-complex was performed before stabilisation with the stathmin protein so that the described binding location is realistic.

The colchicine site is buried into the interstice between the α and β monomers not far from the GTP Non-Exchangeable site. As shown in Figure 4.8.a, the binding pocket lies in the intermediate domain of the β monomer and is surrounded by strands B8, B9, loop T7 and helices H7, H8 of the β monomer. It also interacts with T5 of the α subunit. For instance, it was demonstrated that mutation of the β Val

318 residue induces a variation of tubulin sensitivity to colchicine addition. Colchicine binds to residues on the α monomer: Thr 179, and Val 181 as shown in Figure 4.13. The binding pocket is completed with residues from the β monomer: Ala 250, Asp 251, Leu 252 and Met 259.

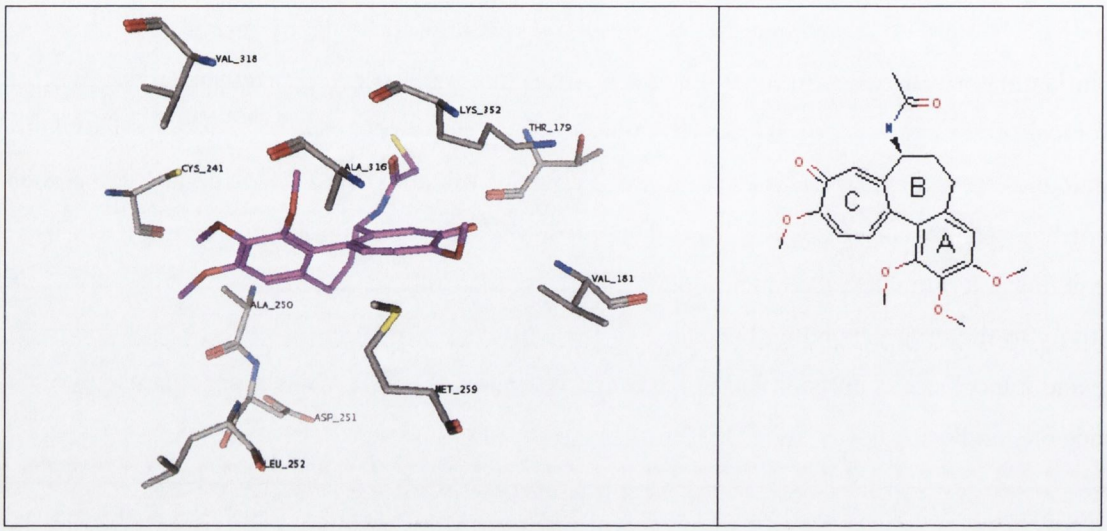


Figure 4.13: Colchicine binding site

The colchicine binding site[52] from 1Z2B on the left and the colchicine structure on the right

Comparing the TC-complex to the straight conformation reveals a displacement of the T7 loop and H8 helix, which move to let the drug penetrate the protein into the binding site. In this manner, sterical clashes between the two monomers are avoided but conformational changes are transmitted reaching to the M–loop, which is displaced by 9 Å in the curved conformation in comparison to the straight one. This explains the divergence found between the actual site described by the 3D structure data and the previous site predicted by linking diverse biochemical data including hybridisation experiments[111].

The different effects observed at low and high concentration can be explained by the M-loop movement. As mentioned previously, in microtubules, protofilaments are held together due to both lateral and longitudinal interaction between monomers. Lateral contacts are largely maintained by the M-loop of one protofilament with helix H3 plus other loops of the neighbouring protofilament. The movement of the M-loop therefore weakens lateral interactions. At low concentrations, the TC-complex added at the end of a protofilament will have its own M-loop displaced because of the bound colchicine ligand. The protofilament cannot stay in its native straight conformation. It has to curve to avoid steric clashes with residues α Asn 101, α Val181 and GTP. MTs perturbation occurs primarily by altering the tubulin–GTP cap. If the concentration of the TC-complex is low, the lateral perturbations are too small to induce a proper detachment whereas once the concentration is increased lateral contacts are lost so that the polymer disassembles. This is validated experimentally where it is also

noticed that at low concentration frequencies of "catastrophe" decreases and of "rescue" increases[108]. The "pause" state is also maintained where neither growth nor shortening of the protofilaments is observed and the global subunit exchange rate from the ends of microtubules is reduced. The stability of the ends is strongly dependent on the concentration of the TC-complex. Microtubules can grow back only once the TC-complex is dissociated from the end.

Colchicine site binders

SAR studies on allo-colchicines analogues and other analogues, revealed the importance for the binding and potency of the A ring with trimethoxy groups and the C ring, with more flexibility being allowed at the B ring (*cf.* Figure 4.13). The A and C rings contain the minimum structural features necessary for high affinity binding. In addition, the B ring can see its lateral chain grow within a certain length as it controls the binding kinetics and affects the thermodynamics[112-115]. Isocolchicines with B ring modifications such as additions of an hydrophobic group on the side chain, bind better to tubulin with improved protein ligand interactions and do not induce the conformational changes of the protein observed with Colchicine[96, 109, 116]. Furthermore, some colcemeids and deacetamidocolchicines have been found to bind in a reversible way[25]. Finally, it has been shown that colchicine binding varies for other isotypes.[111]

Numerous colchicine binders are known to date and have been described in the literature. Here the main families of compounds that were used in the next computational section are given.

Isolated from *Podophillum peltatum*, podophyllotoxin binds to the colchicine site and inhibits tubulin polymerisation[25]. Its cytotoxicity prevented it from being used as a cancer therapy although some water soluble analogues are in use but are suspected to act through a different pathway[117]. Superpositioning of the two complexed structures of colchicine and podophillotoxin as well as experimental studies indicates that both molecules share the trimethoxy group feature and that common interactions are shared when comparing the B and C ring of colchicine to the lactone and oxalone groups from podophillotoxin. However, podophillotoxin binding differs in the sense that it is reversible, faster and less dependant on the temperature[108].

Nocodazole (Figure 4.12) is the lead compound of the benzimidazole family with antimitotic and anti tumoural potencies. It has been shown to bind tubulin through a colchicine competitive assay. It also depolymerises tubulin and binds rapidly, reversibly and in a two step mode. It also causes conformational changes to the protein[118-121]. Its affinity varies across the isotypes of tubulin and it finds application on vinca and adriamicin resistant cell lines. Due to its potency for arresting the cell cycle, this compound was selected as one of the positive experimental controls detailed in chapter 5.

Combrestatin [122-124] (Figure 4.13) was first isolated from *Combretum.* It inhibits polymerisation and binds to the colchicine site. Their angiogenesis inhibition potency[125], makes these compounds considerably attractive, in particular the best analogue A-4. Its combination with a pro-drug permitted water solubility (A-4P) and enabled an impressive potency even in advanced clinical trials for solid tumour treatment. A-4P has the potential to soon become an actual drug[126].

In addition, the following groups of compounds were demonstrated to be colchicine site binders : Sulphonamides[127], indoles[128], Indanone[129] 2-aroyl-indoles derivatives[130], Chalcones[131] , anthracenones[132], 2-Methoxyestradiol analogues[133], Diaminothiazole[134], Benzophenones[135], thiophenes[136], Chromenes[137], Pyrimidinyl pyrazoles[138, 139] and Stilbenoids[140, 141], but shall not be discussed in this work.

4.4.2 The vinblastine site

Isolated from the plant *Catharanthus roseus*, vinblastine (Figure 4.14) is one of the most potent anti-mitotic compounds of the past few years[100]. It has application in the treatment of several cancer types such as leukaemia, non-Hodgkin's lymphoma, testicular germ cell tumour and Hodgkin's disease. It is also cytotoxic in particular involving bone marrow suppression, gastrointestinal toxicity, potent vesicant activity and extravasation injury. As observed with colchicine, vinblastine binds tubulin inhibiting microtubules formation and at higher concentrations disrupts microtubules even at substoichiometric concentrations in comparison with the quantity of protofilament ends[10].

Binding description

The mechanism of action of vinblastines anti-mitotic effect is not yet fully understood despite the latest crystallisation data of the tubulin-vinblastine complex, which shed light on some of its binding features[10]. A novel feature of vinblastine binding is that when disruption of the protofilaments occurs, spirals[142] and rings are observed rather than free dimers. They also seemed to generate paracrystalline aggregates[143]. Before the 3D structure of the tubulin complex was available, cross linking studies had suggested a binding site exposed at the positive end of a microtubule[144] and vinblastine was shown to interact with β monomers close to residues 177-215.

In 1Z2B, the vinblastine interactions were illustrated to take place between the two heterodimers sequestered by a stathmin protein in Figure 4.8.a. Vinblastine lies between the first β subunit and the second α subunit and 80% of its surface area is buried in the complex. Vinblastine interacts equivalently with both subunits[10]. The drug lies in the β subunit at the carboxy-terminal turn of helix H6 helix, T5 loop and the N loop (H6-H7 loop) and the rest of the molecule lies in the α subunit

between the T7 loop, H10 helix and B9 strand. Binding of vinblastine induces conformational changes on the protein. Vinblastine is in contact with residues Leu 248 and Asn 249 from the α subunit and Asp 179 from the β subunit, which are attracted closer to vinblastine therefore moving further away from the nucleotidic site. Conformational changes then occur along the protein as shown by the movement of residue α Asn 349[10].

In summary, the residues in contact with the drug from the β subunit appear to be residues βVal 177, βAsp 179 (loop T5), βTyr210 (helix H6) and Phe 214 (N-loop) and in the α subunit Lys 352 (B9 strand) which is not resolved in the crystal structure but has a Cα within a reasonable distance from the ligand (see Figure 4.14).

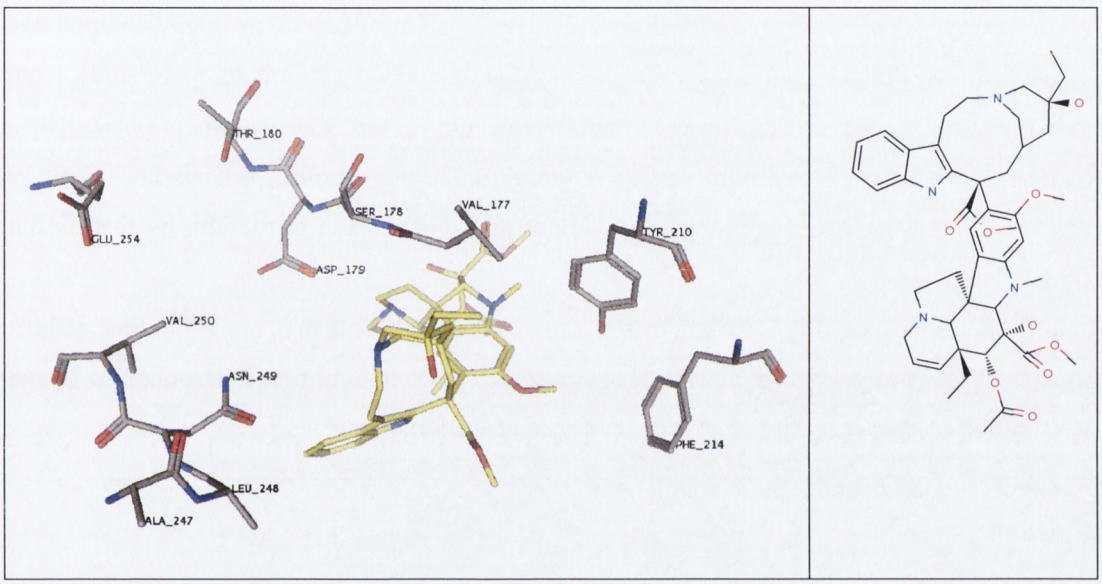


Figure 4.14: Vinblastine binding site
On the left is the Vinblastine binding site[52] from 1Z2B and on the right vinblastine structure.

In addition, close to both of the ligands, GDP and vinblastine, the following residues are found : αSer 178, αThr 180, βAla 247, βLeu 248, βAsn 249, βVal 250 and βGlu254 whereas Thr 349 is in the vinblastine side. All these residues are influenced by the drug binding and come closer to vinblastine to enclose it. Their conformational changes are propagated to the nucleotidic environment first and then to the rest of the protein[145].

It was shown experimentally that vinblastine binds to free dimers in two steps. First, the equilibrium vinblastine-tubulin complex forms, followed by a second slower step suggested to correspond to the conformational adjustments. Although, the binding has a ratio of 1:1 with the free dimer, it is well known that tubulin polymerisation inhibition is efficient at sub-stoechiometric levels. Indeed, a single

molecule of vinblastine binding at the end of one of the microtubule protofilaments is sufficient. The conformational changes caused by the ligand curve the last dimer and propagate longitudinally on the protofilament itself. Moreover, by curving the protofilament, lateral interactions maintained by the M-loop are weakened and consequently disturb the adjacent protofilaments. They then all curve and can split forming the observed rings. Finally, when used at higher concentrations, vinblastine binds at multiple protofilament ends and depolymerises microtubules quickly again forming rings. All of which makes this particular site on microtubules very attractive. The enhanced anti-mitotic effect resulting from binding at that particular site requires a far lower concentration of drug in comparison to the colchicine depolymerisation site.

The use of different tubulin isotypes does not seem to alter the binding. However, vinblastine still displays a cytotoxicity effect that contributed to the development of several analogue compounds.

The success of natural compounds as anti-mitotics generated a plethora of SAR studies on semi-synthetic compounds such as vindesine, vinorelbine and vinflunine. Vindesine is in phase IIIB and IV clinical trials for non small cell lung cancer. Vinorelbine is also in clinical trials and vinfluine is in phase III. They all work at sub-stoichiometric concentrations and inhibit tubulin by fast binding in a reversible manner.

Although the vinca-alkaloids all bind in the same region of β-tubulin, their binding affinities are dependent on the variable points of attachment responsible for tubulin self association. Nonetheless they all stabilize colchicines binding to tubulin as seen for vinblastine[10].

Vinca alkaloid sites

Other drugs that bind in the same domain with overlapping binding regions with Vinblastine are classified as vinca-alkaloid binders or vinca peptide binders. The area at the interstice of the two heterodimers is quite large and can accommodate a broad range of compounds from a scaffold point of view as well as from a size point of view. All binders are directed to a similar, quite large region of the β-monomer. Vinca-alkaloids are a large group of compounds and a substantial number of them are macromolecules such as macrocyclic and peptidic structures. Several are natural compounds extracted from marine and plant organisms. Their interactions can vary from molecule to molecule as they have diverse attachment points to the protein. The growing use of modelling studies and experiments permitted considerable attempts to determine their respective binding locations. In most of the cases, the exact binding position remains unclear.

Rhizoxin isolated from the fungus *Rhizopus chinensis* and Maytansine isolated from the *African plant Maytenus serrata* bind tubulin reversibly and inhibit polymerisation and vinblastine binding. They share some parts of their binding site with the vinblastine site. Phomopsin isolated from the fungus

Phomopsis leptostromiformis bind the same site of Rhizoxin and Maytansine. However it may also have affinities for other sites[146].

Cryptophycin, isolated from *cyanobacterium Nostoc sp*, is the most poisonous and binds almost irreversibly. It inhibits microtubule polymerization by binding at microtubules ends. Cryptophycin inhibits vinblastine noncompetitively and induces conformational changes on tubulin[15, 147].

Dolastatins are linear peptides isolated from the mollusk *Dolabella auricularia*. This group of compounds includes symplostatin and cematodin (analogues TZT-1027 and ILX-651). They stabilise tubulin, bind in a two step mode, generate aggregates of tubulin and inhibit vincristine non-competitively. They are postulated to bind in two classes of binding sites[147]. The best of these, Dolastatin 10 & 15 have their analogues in clinical trials[148] .

Rhizoxin, maytansine, phomopsin A, cryptophycin and Dolostatin 10, all can disrupt microtubules and some induce spiral or ring formation. As with vinblastine, some strongly inhibit GTP exchange, and they show varying degrees of inhibition of binding by each other[149]. Due to the differences in binding inhibition, the binding sites do not appear to overlap completely. Thus the region remains termed the vinca domain. Modelling studies using flexible docking for both ligands and protein were performed to investigate the active sites of cryptophycins, dolastatin 10, hemiasterlin, and phomopsin A[147]. The active site for all tested ligands was suggested to be near the nucleotidic site and overlaping with the vinblastine binding site.

Hemiasterlins are sponges *Cymbalstela sp.*-derivated to tripeptides and enclose analogues such as HTI-286, criamides and milnamide. They bind and depolymerise tubulin. Hemiasterlin, non-competitively inhibits the binding of vinblastine to tubulin whilst competitively inhibiting the binding of another peptide dolastatin 10 to tubulin; therefore it has been proposed that hemiasterlin binds to the Vinca peptide binding site of tubulin. It was shown that they bind at the β/α interface and shared some part with the dolastatine 10 [150] and vinblastine sites on β tubulin [151]. Molecular modelling was carried out to predict the binding mode of Hemiasterlins[147]. These docking studies used fast exhaustive rigid docking and predicted a binding site in agreement with experimental data showing hydrophobic and H bonding interactions. The binding site consists of an overlapped vinblastine binding site and is again not far from the nucleotidic site[152].

Tubulysins isolated from *myxobacteria Archangium gephyra* and *Angiococcus dicsiformis* are linear tetrapeptides of N-methyl pipecolic acid (Mep), Isoleucine (Ile), tubuvaline (Tuv) and tubutyrosine (Tut). They are potent inhibitors of tubulin polymerisation[153]. They strongly interfered with the binding of vinblastine to tubulin in a noncompetitive way and stabilise colchicine binding.[154] However, Tubulysin A induces the formation of rings, double rings, and curled structures as with vinca alkaloids.

Spongistatin, isolated from *Spirastrella spinispirulifera* binds to tubulin at a different site from the vinca alkaloid one[155, 156]. It does not generate either aggregates or morphologicaly distinctive polymers.

Scleritodermin, isolated from the lithistid sponge *Scleritoderma nodosum,* is a cyclic peptide and also inhibited tubulin polymerization[157].

Disorazols are macrocyclic dilactone acrocyclic polyketide compounds from the *myxobacterium Sorangium* cellulosum. Disorazol A1 inhibits tubulin polymerisation and leads to depletion of microtubules in intact cells. The mitotic cells showed abnormal spindles with multipolar configuration at low concentrations of disorazol[158].

Pyrimidines such as triazolopyrimidines[159] showed potency in the nanomolar range on cancer cell lines resistant to paclitaxel and/or vincristine. They enhanced the initial rate of MAP-rich tubulin polymerisation, induced polymerisation of pure tubulin in the absence of GTP and bound to the Vinca-peptide site of tubulin.

Halichondrin, isolated from the sponge *Halichondri okadai*, is a macrocyclic polyether. Its analogue E7389, has been shown to bind microtubules at their end and to form aggregates, which in turn would compete with free dimers. This could explain the lack of inhibition of tubulin shortening but only of the lengthening. Modelling studies suggested a binding site for halichondrin B and NSC 707389 at the inter heterodimer region near the nucleotidic and vinblastine site. This finding suggests also that by adapting to these compounds the residues involved could have moved enough to weaken vinblastine binding. This model suggests that these compounds most likely form highly unstable, small aberrant tubulin polymers rather than the large stable structures observed with vinca alkaloids and antimitotic peptides[160, 161].

Finally, the minimum pharmacophore of the halichondrin class of compounds as anti-mitotic agents is given by the two simplified structures tetrahydropyran and tetrahydrofuran. A novel mono-tetrahydrofuran COBRA (from a Curacin analogue) was design after 3D modelling studies. It is predicted to bind in a unique, hydrophobic cavity between the GDP/GTP binding site and the M-loop. The free space left could accommodate diverse analogues. One of these analogues (Cobra 1) is structurally close to colchicine and caused partial depolymerisation of tubulin and inhibited polymerisation in the presence of GTP. Additionally, complete depolymerisation was obtained by Cobra 5[162, 163].

4.4.3 Non localised binding sites destabilisers

For the following compounds the exact binding site remains unknown.

Noscapine analogues obtained from opium suppress microtubule dynamics by prolonging an attenuated 'pause' state in which growth or shortening is not detectable and inhibits cell cycle

progression at the G2/M phase and thus induces apotosis. Modelling studies have suggested a putative pocket on ß-tubulin near its interaction surface with α-tubulin [94]. The exact site is not yet elucidated.

Dihydronaphthalene compounds are inhibitors of tubulin polymerisation having some analogues with a potency comparable to CA4.[164]

Pseudolaric acid B, isolated from the root bark of *Pseudolarix kaempferi Gordon (Pinaceae)* inhibits the formation of mitotic spindles. Pseudolaric acids comprise of a large group of small molecule natural products[95].

A series of indole ketoamides have been demonstrated to be potent tubulin polymerisation inhibitors and cytotoxic against a panel of human cancer cell lines, including MDR-positive lines.[97]

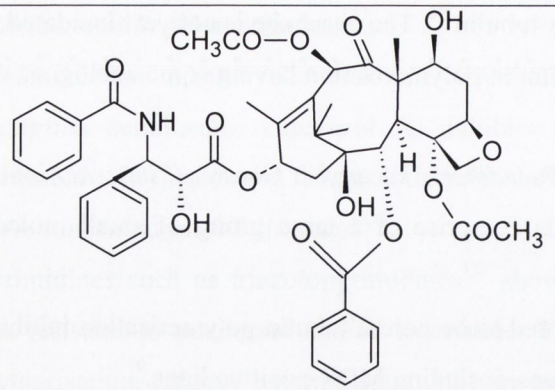
Oxaline and neoxaline, indole alkaloids isolated from the culture broth of *Penicillium oxalicum* and *Aspergillus japonicus*, respectively, are small molecule natural products. They display anti-proliferative activity and arrest the cell cycle at the G2/M phase. Oxaline inhibits tubulin polymerisation resulting in cell cycle arrest at the M phase[165]. Additionally, salvinal isolated from *Salvia miltiorrhiza Bunge* is a structurally-simple natural product, which has been shown to inhibit tubulin polymerisation in a concentration dependent manner[166]. Cellular microtubule network changes are similar to the features observed with colchicine.

Vitilevuamide, isolated from *Didemnum cuculiferum* and *Polysyncranton lithostrotum* is a bicyclic 13 amino acid peptide. It inhibits tubulin polymerisation and does not compete for the vinblastine or colchicine site although it does effect the GTP concentration. It is therefore suggested to bind to an unique site on tubulin which is as yet unknown[167]

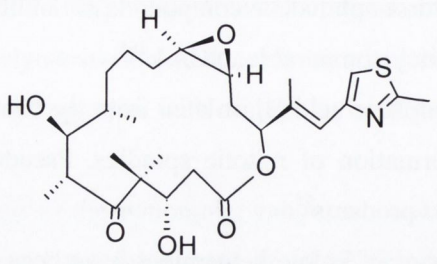
4.5 On target stabilisers ligands

A series of stabiliser ligands are now described, starting with those that have been determined to bird to an area termed the Taxol binding site. A sample of the following discussed structures is shown on Figure 4.15.

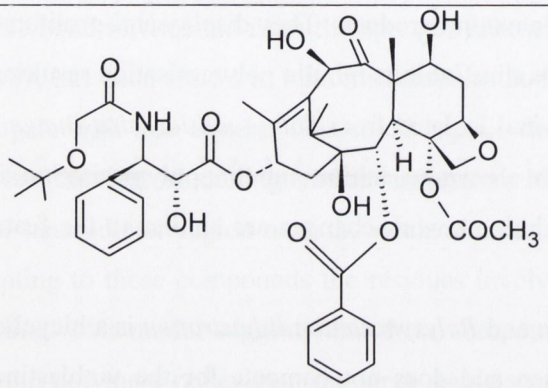
STABILIZER COMPOUNDS



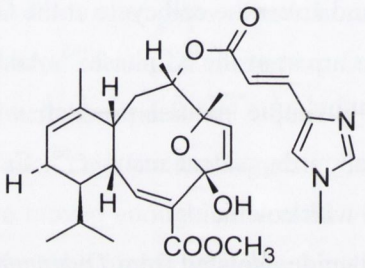
Taxol



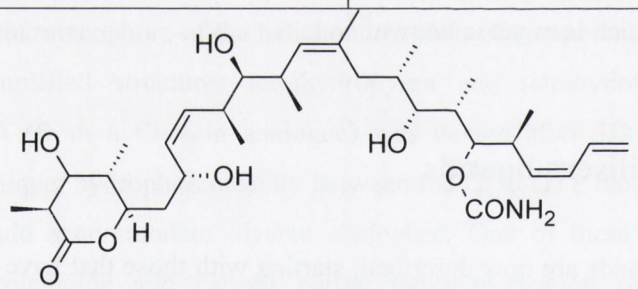
Epothilone A



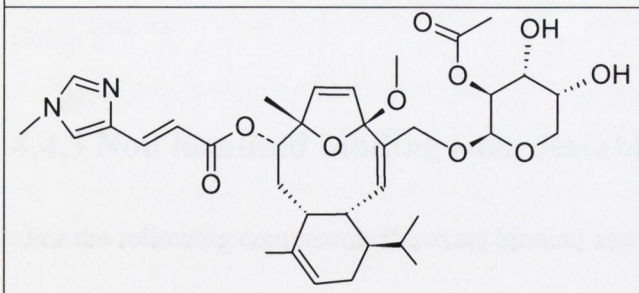
Taxotere



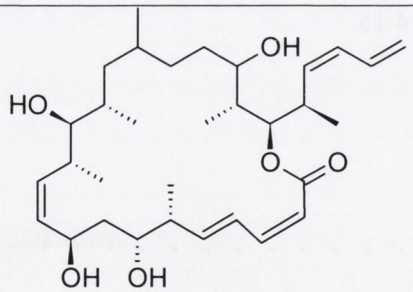
Sarcodictyn A



Discodermolide



Eleutherobin



Dictyostatin-1

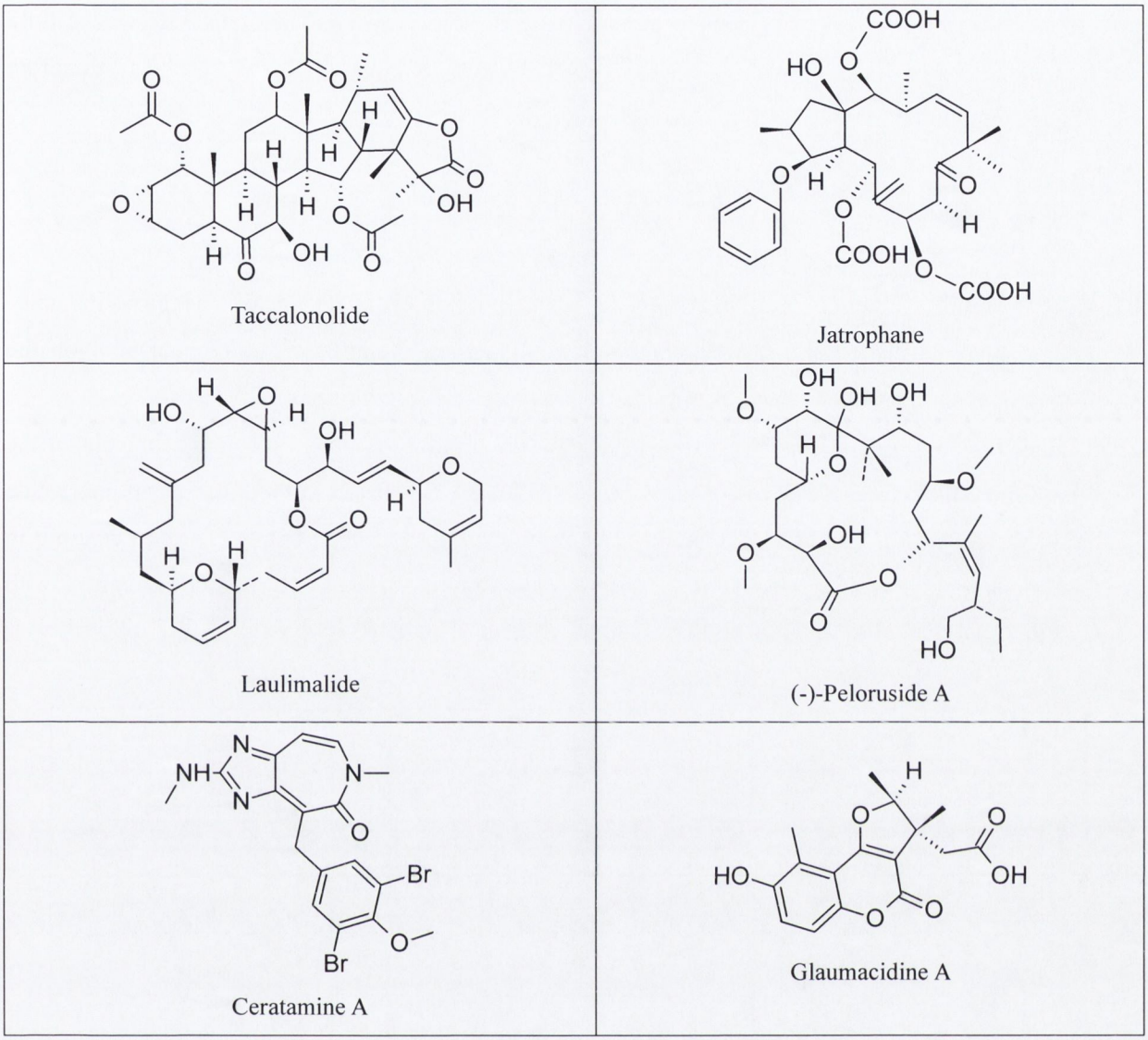


Figure 4.15: Structural diversity of tubulin stabilisers.

4.5.1 The taxol binding site

Paclitaxel (taxol; Bristol-Myers Squibb) and taxane compounds were the first class of molecules discovered to bind in this site and taxol was first extracted from the bark of the Pacific yew. These compounds stabilise microtubules and have found their application in various types of cancer chemotherapies[168, 169]. Since 1995, Paclitaxel was approved for clinical use and is now widely used to treat breast and non-small cell lung cancer. Cancer research was immediately directed towards finding new molecules having similar mechanisms of action and improved pharmacological, toxicological and physicochemical profiles[170]. As soon as the taxotere-tubulin complex and taxol interactions were identified[8], its pharmacophore was developed giving birth to a series of compounds with novel

chemotypes, having similar effects and similar interactions[98] Ergo, taxol and taxotere were successfully used in diverse cancer treatments[171] with Docetaxol derivatives having a better solubility therefore being the compound of choice for some cancers therapies.[172].

Binding description

Like other tubulin binders, taxol has been shown to bind free dimers as well as protofilaments. However, it has been recently shown, that if microtubules are inhibited by colchicine addition the conformational changes induced were sufficient to prevent taxol from binding.[48, 173]

The taxol binding site is located on the inside face of the microtubule wall approximately at the middle of the β monomer. It makes contact with the M loop and is between the middle and nucleotide domain (Figure. 4.8.b). The pocket is mainly hydrophobic with hydrogen bonds also participating in the binding and parts of the molecule are exposed to hydrophilic surfaces. Its interactions concur with most of the SAR available for taxol and its analogues (see Figure 4.16).

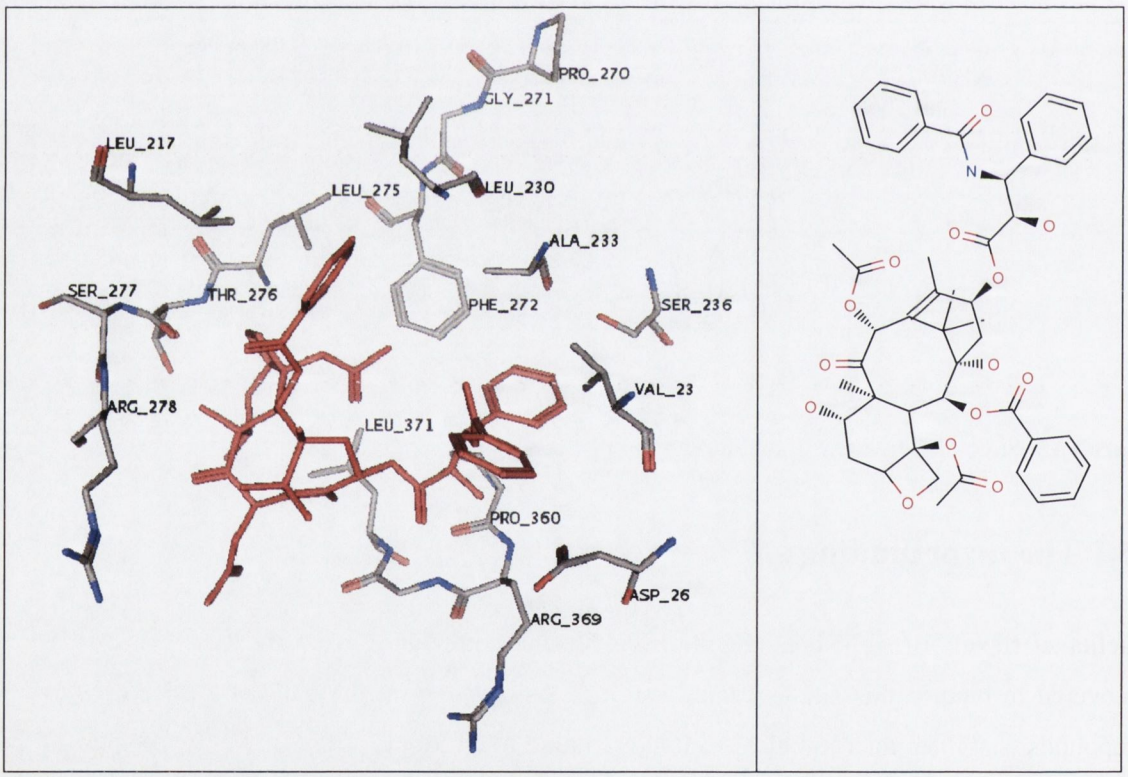


Figure 4.16 Taxol binding site

On the left taxol binding site[52] from1JFF and on the right the Taxol structure

Taxol has a stabilising effect that prevents the separation of adjacent protofilaments. As the M-loop is in part responsible for the maintenance of the cylindrical form of the microtubule, it is reasonable to suggest that taxol binding at the β tubulin stabilises the M-loop conformation in a way that lateral contacts between protofilaments are optimised. The protofilaments find themselves bound tighter with stronger lateral interactions[145].

Another explanation of the stabilising effect is through a counteracting process. Taxol could prevent the protofilament from curling. As it is in direct contact with the H1 and H7 helices, it could induce conformational changes propagated to the region near the nucleotide domain and counteract the conformational changes induced by the hydrolysis. An alternative explanation is that once taxol is bound, the second domain which comprises its site could be rigidified. This second domain would restrict the flexibility of the nucleotide site and therefore stabilise the whole system[174].

The taxol effect resides in the conformational changes it induces in its domain. It is believed that conformational modifications of H3 are the main conformational changes responsible for microtubule dynamics switching. Indeed H3 in tubulin corresponds to the switch II region of classical GTPases. The equivalent taxol site in the α subunit is filled with an eight residue section describing an extra loop for α when compared to β. As the α subunit comprises of the non-exchangeable GTP it is suggested that the lateral interactions induced from the α monomers are alike and stronger than the lateral interactions induced by the β monomer which enclosed the GDP exchangeable site.[57]

This can explain the fact that the negative end of the protofilament capped by an α monomer is less dynamic as it is laterally strongly stabilised in comparison to the positive end capped by a β monomer which has weaker lateral interactions.

Taxol site binders

Epothilone macrolide was isolated from *the mixobecterium Sorangium cellulosum*, and binds to the taxol site stabilizing the microtubules[175]. The complex structure with tubulin was resolved[43] clarifying the tubulin ligand interactions. This structure modified and enhanced the previously suggested pharmacophore common to Taxol and epothilone [90, 176]. Epothilones are easier to obtain and more soluble than taxol, however, they share comparable cytotoxic effects.[175] SAR on Epothilones led to more potent molecules such as cis-tmt-EpoB, which exhibited the highest binding affinity. A number of analogues including members from Bristol-Myers Squibb and Novartis are in phase I/II clinical trials.

Discodermolides, isolated from the marine sponge *Discodermia dissoluta* is in advanced clinical trials in the application of cancer therapy[177]. It greatly stabilises microtubules, binds to the taxol site and is slightly less cytotoxic in comparison to taxoids and epothilones [178]. However, the analogues generated did not exhibit better potency, antiproliferative effects or affinities, which remain similar or worse[179].

Further experiments showed that taxol and discodermolide shared a common binding region but discodermolide seemed to bind more tightly as its displacement by Taxol was not observed [92, 180, 181].

The polyketide dictyostatin-1, isolated from the marine sponge of the *genus Spongia sp.,* is a macrocyclic compound and revealed multiple aster formations and microtubule matrix bundling patterns similar to Paclitaxel on treated cells and is highly potent against paclitaxel-resistant cell lines[182, 183].

Eleutherobins isolated from *Eleutherobia sp.* and sarcodictyins, isolated from the *Sarcodictyon roseum,* share the same core structure and differ by the presence of an acetylarabinose group on the eleutherobins[182].

Additionally, Eleutherobins, sarcodictyins, Dictyostatin and cyclostreptin all bind to the Taxol site[182].

4.5.2 Non localised binding sites stabilisers

A wide range of stabiliser compounds are macrocyclics among which some of the following structures are related.

Jatrophane polyesters is isolated from the Sardinian *endemism Euphorbia semiperfoliata* and Pelorusides (A) is isolated from the marine sponge *Mycale sp.* Both families of compounds arrests diverse cancer cell lines in the G2/M phase of the cell cycle. *In vitro* Jatrophane polyesters induce tubulin assembly similar to paclitaxel-like microtubules, however, the rearrangement of the microtubule architecture remains different. Peloruside A induces tubulin polymerisation in the absence of MAPs and has been shown to bind to a different site on tubulin than that of paclitaxel[91].

Laulimalide is a potent, structurally unique microtubule stabilizing agent originally isolated from the marine sponge *Cacospongia mycofijiensis*. It does not compete for the taxol binding site [93] leaving its mode of action upon tubulin unclear [184]. 3D and 5D QSAR studies suggested a binding site on α tubulin, under the B9–B10 loop extension, which is a taxol site equivalent region on the α subunit [185]. The same modelling study suggested that Pelorusides would also bind to this alternative site on α tubulin.

Quinolones have polymerisation activities and quinolone treated cells behave in a similar, but not in an identical way to that observed after taxane treatment[186]. The indolinone family comprises an interesting novel indolinone which destabilises microtubules at high concentrations and stabilises microtubule dynamics at low concentration, which leads to abnormal spindle pole formation.[187]

Benomyl (used as an antifungal) and dinitroanilines (oryzalin, anti parasite trifluralin), used as a pre-emergent (herbicide) are both Benzimidasole compounds. Cross linking experiments mainly with Alanine mutations and scans have shown a certain constant part of the tubulin sequence to be responsible for sensitivity to these drugs[54, 188] but it was suggested the mutations themselves were

stabilising microtubules rather that interacting directly with the binding site [145]. These results and the sequence comparison were used to suggest a site on the surface of the β tubulin monomer localised between the Taxol and Colchicine sites.

A new class of coumarin antimitotic agents isolated from *Glaucidium palmatum* stabilises microtubules. Ceratamines A and B isolated from *Pseudoceratina sp.* are heterocyclic alkaloids. They are potent microtubule stabilising agents with antimitotic activity[189]. In breast cancer cell lines such as MCF-7, ceratamines stop mitosis in a concentration dependent manner and directly stimulates microtubule polymerisation in the absence of MAPs. Ceratamines do not compete with paclitaxel for binding to microtubules, but it has not been shown that they bind to another site on tubulin.[189]

This last section on tubulin stabiliser agents concluded the examination of small molecule binders. As illustrated by this review work, most of the compounds exhibit relatively serious cytotoxicities. New tubulin binders with a reduced side effect profile are urgently needed for potential applications in oncology. Indeed, the majority of the presented compounds failed in clinical trials because of this issue. The biological review section closes here, leading into the computational section of this chapter. In the modelling part, the focus will be returned on PBOX apoptotic compound and the attempt in finding new ways of approaching their development to avoid their side effects on the central nervous system.

4.6 Mapping tubulin sites.

The aim of this computational study is to identify the PBOX compound binding site in tubulin. Determination of their exact binding site would enable from one hand to improve the previous ligand based studies and from another hand could give the opportunity to use new Virtual Screening (VS) approaches using the receptor information. The previous section described tubulin dynamics and its diverse interactions. Under physiological conditions, tubulin interacts with many other proteins, *e.g.* MAPs, on various diverse parts of its surface. To date, only three binding sites have been elucidated for the binding of small molecules.

PBOX compounds are small molecules which have been shown[1]:

- to destabilize tubulin.
- not to compete with colchicines.
- not to compete with vinblastine.

By localising the region in tubulin where PBOX ligands interact, the previously built pharmacophore with Catalyst could be refined, thus opening a new door to research areas such as docking methods for VS. New scaffolds could then be found with expected similar activities occurring through the same pathway through which PBOX act. Thus, new small molecules would disrupt MTs network inducing apoptosis and their different structural form could prevent undesirable side effects. PBOX ligand as explained in chapter 1, were originally developed for the peripheral-type benzodiazepine receptor (PBR)[190] present in peripheral tissues and in the central nervous system (CNS)[191]. Using the tubulin receptor information together with the ligand based study, virtual screening will permit scaffold hopping, thus generating more efficiently novel more selective ligands for tubulin than for the PBR. Coupling these two computational approaches have already shown great success in cancer research such as the discovery of new molecules for selective inhibition of Glycogen Synthase Kinase (GSK-3[192]) or cyclin-dependent kinase (CDK[193]) and have also participated in Alzheimer research[194] and in the discovery of new potential narcotic analgesic inhibitors of G-protein coupled receptors(hKO[195]).

4.6.1 PBOX binding site discovery strategy

Determination of the binding region of these small molecules in such a large and flexible protein is not straightforward. A geographical mapping of the process to be undertaken is essential and advantageous at this stage. The computational strategy for identifying where PBOX bind is depicted in Figure 4.17.

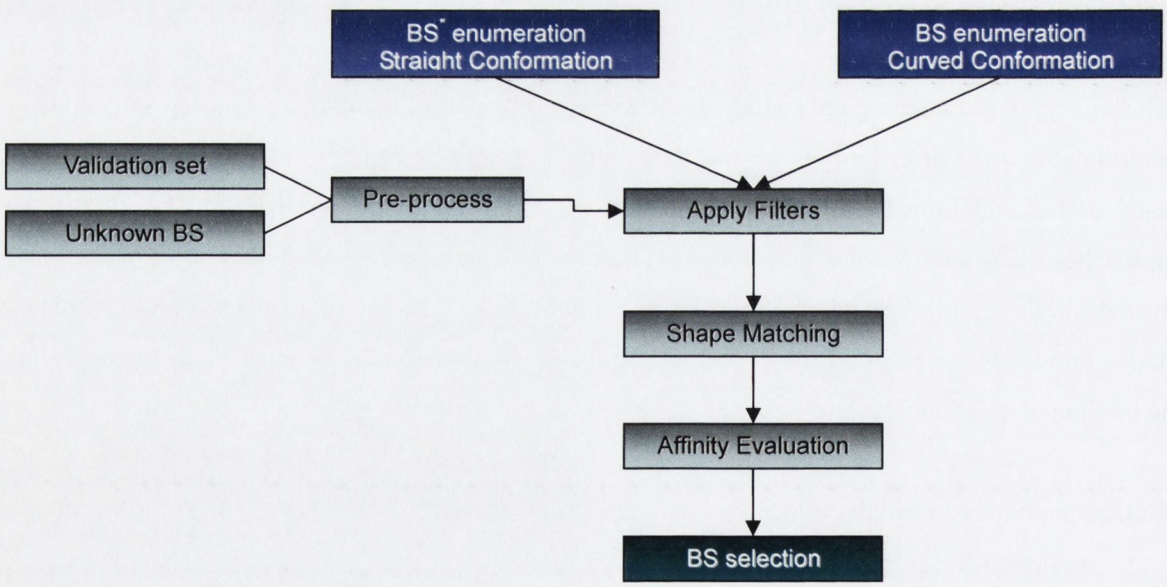


Figure 4.17: Binding site search - strategy break down*BS: Binding Site

The essence of this schematic lies mainly in a complete and comprehensive mapping of the receptor structure followed by a rational selection of the appropriate binding site with subsequent validation.

Validation set

Initially, molecules whose binding sites are known are used to test the efficiency of the method in retrieving their actual binding sites. What will be subsequently refered to as the validation set, is a database of molecules that includes tubulin 'binders'. This validation set consists for the most part of a vast, diverse number of ligands outlined in the first part of this chapter. Diverse structures were selected to avoid bias in the validation process, a pitfall commonly observed in such studies[196]. For instance, if the validation set consisted of mainly numerous analogues derived from one unique scaffold, then assigning their binding site would obviously only validate the method for this particular scaffold. To ensure pocket retrieval impartiality and fairness in molecular treatment, three known classes of molecules with elucidated binding sites were tested (*i.e.* taxol, colchicine and vinblastine sites). In addition, scaffold diversity for each group of molecules was maximized and a sample of the chosen scaffolds can be found on Figures 4.12 and 4.15.

Active site enumeration

The validation set comprises of stabilising and destabilising compounds. As explained in the previous section, the binding of a ligand from one or the other of the two effect categories would engender the protein to lock into a straight or curved conformation respectively. In this light, it is of the utmost importance to analyse the protein in its two forms. The availability of crystallographic data for both conformations permitted the enumeration of binding pockets for each case. Several modules have been developed to detect and characterize possible active sites from the 3D atomic coordinates of the receptor. They can roughly be classified into two classes according to the type of detection method they use.

The first class, consists of a number of methods based on the determination of the interaction energy (*i.e.* electrostatic and hydrogen bonding) between a probe at a given point and a protein, thereby locating energetically favourable sites [197]. Among these, the most established is the Goodford[198] method, which identifies sites of favourable interaction with specific probe types and which was tested using the hydrogen bonding potential of drug-like molecules using GRID[199, 200]. A second example detailed by Miranker[201] uses Multiple Copy Simultaneous Search (MCSS) method to detect favourable binding sites for different functional groups. The method of Silberstein et al.[202] distributes organic solvent molecules around the surface of an enzyme and optimizes interaction energies between them

using a conventional molecular mechanics function (including van der Waals, electrostatic and solvation terms) in a way similar to the MCSS method. For each type of organic molecule, the distances between the active site and the energy minima are calculated. Q-SiteFinder[203], uses the van der Waals interaction (of a methyl probe) and an interaction energy threshold to determine favourable binding clefts. Bate and Warwicker[204] predicted active site location based on the peak of the electrostatic potential which was shown to be close to the active site centre.

A second class of methods analyses the protein surface and defines binding site extents and locations using geometric criteria. Usually, the ligand binding site is in the largest pocket. This class of methods is represented by a broad range of different modules. A number of these methods were widely used in the detection of protein pockets such as APROPOS[205], Cavity Search[206], CAST[207], LigandFit[208], PASS[209], POCKET[210], SURFNET[211], VOIDOO[212] and the method of Del Carpio [213] which uses an analytical geometric algorithm to compute pocket sites. As an example, LIGSITE[214] uses a grid representation of the molecular volume and computes exterior site scores by projecting rays from the receptor exterior to the surface. The deeper and more surrounded a site is, the higher it scores. Geometric methods locate pockets without the use of energy models so that assignment to the receptor atoms of proton locations and partial charges are not required.

In the present work, Site Finder[52], a geometric method was employed. This module detects protein-ligand and protein-protein binding sites using a fast geometric algorithm, based on alpha shapes[215]. The description of the favourable interactions between the receptor atoms and "dummy" atoms within an active site is less restrictive than in the previously mentioned methods. Indeed, relative positions and accessibility of the receptor atoms are considered along with a rough classification of chemical type. Site Finder is widely used and has been validated in a variety of case studies[216]. Therefore, the two tubulin structures were scanned for possible binding sites (referred to as BS herein). Site Finder placed "dummy" atoms inside the pockets and labelled them as hydrophobic or hydrophilic depending on the environment they were surrounded by. Thus, all the sites found for both conformations were enumerated and each site was transformed into "dummy" molecules which were used later in the procedure for shape matching with either the validation set molecules or PBOX compounds.

Knowledge based filter for binding site elaboration

In advance of execution of the shape matching procedure, biological information was gathered and used to filter out the improbable sites. Generally, biological data can guide binding site identification. For example, in the case of Halichondrins and analogues, competitive assays against radiolabelled dolastatin-10, vinblastine and GDP, indicated a binding region that should not include the exchangeable site and should overlap the vinblastine site[161]. Another example is a charged-to-alanine

mutation of the protein followed by drug binding assay to measure, new sensitivity. Phenotypes and therefore sequence alteration were determined localizing for instance benomyl destabilising agent on beta tubulin between the non-exchangeable site and Taxol site[54].

In this work, limiting information was available. No genetic studies such as mutations or cross linking studies have yet been undertaken and the competitive assays completed did not define a site but at least pointed the search toward new directions. Therefore, each potential site was passed through the nets of a knowledge-based filter, rejecting sites according to:

Receptor conformation; sites were tagged as destabilisers when they were generated from the curved conformation and stabiliser when they were generated from the straight conformation. If a molecule was tagged as destabiliser or depolymeriser, it would only be compared to destabiliser sites.

The "goodness-of-fit"; the binding pocket volume had to be big enough to fit the molecule tested.

Binding sites and molecules shape matching

After filtering, to link each compound of the validation set to its own BS, a matching method was needed. The remaining BSs where sorted by optimal match using the Rapid Overlay of Chemical Structures (ROCS) [217] toolkit. This method relies on shape matching having optional chemical matching parameters available which were also turned on. Hawkins et al have previously compared the performance of ROCS with FRED/GLIDE (docking utilities) [218] and demonstrated that it generally outperforms both.

Rigid docking

Finally, binding affinities were examined by firstly rigidly docking molecules into the remaining BS and secondly by refining binding modes which allowed full flexibility of both the receptor and the ligands atoms. Rigid docking was employed using Fast Exhaustive Rigid Docking(FRED [221]) which exhaustively scores all possible poses of each ligand in the active site. The exhaustive search is based on rigid rotations and translations of each conformer. In this manner, the sampling issues associated with stochastic methods based dockings are avoided. FRED has previously been shown to generate reasonable binding modes of pyrimido[1,2-b]indazoles analogues in the binding site of colchicine guiding optimisation and corroborating experimental SARs[222]. To elucidate correct binding orientations, selection of a scoring function, post-docking, which correctly assigned known actives to their respective BS was initiated.

Binding mode selection

To validate the process, each bound ligand (*i.e.* Taxol, colchicine, vinblastine) was extracted from their respective binding sites and re-docked to assess the ability of the docking and scoring functions to reproduce known binding modes. Subsequently, Protein Ligand Interaction Fingerprints[223] (PLIF) were calculated for the outputted poses. The Tanimoto similarity between each pose and a known co-crystallised ligand was calculated enabling sorting of the binding poses by similarity to the co-crystallized ligands. Similar procedures involving interaction fingerprints have previously been utilized by several groups to overcome problems associated with scoring functions *i.e* their inherent ability to incorrectly rank optimal binding modes[223, 224].

Flexible refinement and binding affinity estimation

Subsequent flexible refinement of each retained pose was employed using the LigX module[52]. LigX is an interface module of MOE which facilitates active site analysis and the calculation of molecular properties such as binding affinity. It is usually used in the context of ligand modification and optimization in the active site. LigX estimations had already been compared to experimental data of 560 small molecules and excellent correlation was illustrated[225].

The previously described protocol was subsequently executed with a set of PBOX compounds post-validation. The next section delivers details on the settings and algorithms used by each of the different methods of PBOX BS quest.

4.6.2 Computational methods

Validation set preparation

The validation set was made up molecules whose binding sites were well established and were denoted as taxol, colchicine or vinblastine. Each category was represented by a minimum of 20 different scaffolds as shown in Table 4.2.

Ligand binding sites	Effect	set size	no. of scaffolds
Colchicine	Destabilise	180	~30
Vinca alkaloids	Destabilise	45	~20
Taxol	Stabilise	94	~20
Halichondrin	Destabilise	11	1

Table 4.2: Validation set diversity and size

The validation set was completed by addition of halichondrins, vitilavuamide, laulimalides and some other ligands, all of which have undergone independent modelling studies where binding sites have been suggested. A selection of diverse structural forms for the validation set is illustrated in Figure 4.16.

To translate the molecular information into 2D format, all molecules were drawn using ChemSketch[226] and passed to Pipeline Pilot (PP) to ensure uniqueness Protonation states were enumerated at pH7.4 using appropriate PP components[119] (detailed in Chapter 3). 3D representation of each molecule was created from PP output using OMEGA2. Further detail on the mechanism of conformer generation can be found in Chapter 3. All default parameters were retained with a maximum number of 150 low-energy conformers generated per structure.

Site Finder

Site Finder was used for active site identification and enumeration. This method, is invariant to rotation, and filters out sites that are "too exposed" to solvent. Also water sites are separated from likely hydrophobic sites where definitions of hydrophilicity and hydrophobicity are invariant to protonation and tautomeric states. Site Finder methodology is based upon alpha shapes[215] which are a generalisation of convex hulls described by Edelsbrunner[206].

Alpha sites are positioned in the actives site using a modified Delaunay triangulation of a collection of 3D points. Each alpha sphere is then fit in the four point simplex obtained and has a specific radius which can be infinite (corresponding to the planes of the convex hull of the point set). Inaccessible or over exposed alpha spheres are then discarded whereas small alpha spheres, corresponding to locations of tight atomic packing in the receptor, are retained. Alpha spheres are tagged as either "hydrophobic" or "hydrophilic" depending on whether the sphere is in a good hydrogen bonding spot in the receptor. Furthermore, hydrophilic spheres not near a hydrophobic sphere are eliminated. Finally, the alpha spheres are clustered using a single-linkage clustering algorithm to produce a collection of sites. Each site consists of one or more alpha spheres at least one of which is hydrophobic.

Site Finder allows several tuneable parameters to be altered regarding the size and nature of the sites to be enumerated. The probe radii and the minimum distance between the two categories of alpha spheres can be altered and also the level of separation of clusters of spheres can be controlled through a minimum distance parameter. Finally, the size of sites to be retained is controlled by the site radius. In the current work, all default values were kept. All co-crystallised ligands were extracted from their sites prior to Site Finder analysis.

Filtering: volume evaluation

The volumes calculated were the van der Waals volume of molecules, *i.e.* the sum of the volume of their atoms. The volume of an atom is defined as the volume of a sphere with the van der Waals radius of the atom. The radius is determined from measurements of atomic spacing between pairs of unbounded atoms in crystals. However, if the atoms are bonded by a covalent bond, the distance between their centres is smaller. Therefore, the van der Waals volume of a molecule with covalent bonds is smaller than the sum of the van der Waals volumes of the atoms.

All volumes were calculated (sites made up of alpha spheres or from molecules) using an SVL script using a grid approximation of 0.75 Å spacing. This descriptor depends on the structure connectivity and conformation (dimensions are measured in Å). In the case of binding site volumes, volumes were calculated before conversion of spheres to atoms.

Clustering molecules

To cluster the diverse data set the Clustering component from PP [119] was employed. This component partitioned the original data set into smaller regions that define the clusters. A number of representative objects (cluster centers) are chosen from the data set. The corresponding clusters are found by assigning each remaining object (cluster members) to each representative object, selecting the object that is the closest. The distance function between the objects can be a Euclidian distance, a Tanimoto distance or a combination of the two. To select the cluster centers a maximum dissimilarity method is employed. First, it randomly chooses a data record and assigns it as the first cluster center. The record maximally distant from the first point is selected as the next cluster center. The record maximally distant from both current points is selected after that. The process repeats itself until there are a sufficient number of cluster centers. The non-selected objects are then assigned to the nearest cluster center to determine the cluster membership.

In the current study, members of each data set were first manually selected in the region of interest and then clustered with standard settings to give 2 clusters. Extended Connectivity Fingerprints 4

(ECFP_4 discussed in chapter 3) were used and cluster distances were measured using Tanimoto distances. Cluster centroids were then presented as representative of their clusters.

ROCS

As detailed in Chapter 3, the ROCS[217] toolkit is a shape-based superposition method that rigidly aligns two structures maximizing the overlap volume between the two and where definitions of chemistry can be used. In this work, all default values were kept among which the ImplicitMillsDean as Chemical Force Field (CFF) which includes a simple *pKa* model that assumes pH=7. All cut-offs were disabled so that all the calculated overlays were output. The overlays were then ranked by ComboScore which is the sum of shape and colour Tanimotos. A value of 2 indicates that the two overlayed structures are identical. ROCS was employed in this study to match novel ligands with their respective binding sites on tubulin. Each BS was used as a "dummy" molecule query on which each compound of the validation set was superposed. Each superposition was evaluated using the Tanimoto coefficient and ranked accordingly. To realize these overlays, two molecular pre-processing steps were necessary:

Molecules to be matched were represented in 3D space using the conformer generator OMEGA2[219] to account for ligand flexibility. OMEGA software has been previously shown by many groups including our own group to generate diverse and sensible 3D conformations of molecules [220]. The actual nature of the bioactive conformation of a bound ligand (i.e. Taxol bound to tubulin) was sampled by generating a high and diverse enough number of conformers. In addition, all molecules were protonated at physiological pH (7.4) to more accurately mimick each molecule in its biological environment.

Each binding site was transformed to "dummy" molecules by converting each hydrophobic and hydrophilic "dummy" atom into either an oxygen atom or carbon atom respectively.

FRED docking

FRED2.1 uses a systematic, non-stochastic algorithm to ensure reproducible results are attained. FRED rigidly and exhaustively examines all poses in an active site, filters by shape complementarities and then ranks by 'fitness' prior to scoring using Gaussian functions which have chemical awareness incorporated (*eg*. Chemgauss2, Shapegauss). The final poses can be scored simultaneously utilising a number of scoring functions such as, Shapegauss, PLP, Chemgauss2, Chemscore, Screenscore, and Zapbind.

FRED2.1 [227] was utilized in this study to dock all pre-processed compound sets. For this study, default operational values were applied with the exception of –addbox=3.0. The addbox is the

extension of the 3D search space from the outermost points of the reference ligand. Chemgauss2 was found to be the most efficacious scoring method for ranking the docked poses. The Chemgauss2 scoring function combines the Shapegauss scoring function with additional potentials between chemically matched positions around the ligand pose. Schulz-Gasch et al [228] deemed FRED to be especially attractive as a docking tool because it docks at a high speed compared with other methods.

Clustering binding modes

Protein Ligand Interaction Fingerprints (PLIF) describes the interactions between ligands and proteins by classifying them according to the residue of origin. Six types of interaction are described and can take values of 00, 10 or 11 describing sidechain hydrogen bonds (donor or acceptor), backbone hydrogen bonds (donor or acceptor), ionic interactions, and surface interactions. Interactions are set as high or low dependent on their strength of interaction. If the strongest interaction passes the lower interaction threshold, the low order fingerprint bit is set and if the strongest interaction passes the higher interaction threshold, then the low order and high order bits are both set.

Briefly, hydrogen bonds between polar atoms are calculated using a method based on protein contact statistics[229]. Ionic interactions are scored by calculating the inverse square of the distance between atoms with opposite formal charge. Surface contact interactions are determined by calculating the difference of the solvent exposed surface area of the residue (in 1.4 Å distance) in presence and absence of the ligand. This is potentially indicative of a hydrophobic interaction.

In the presented work, all poses generated by FRED were clustered using PLIF and the highest scoring member of each cluster from the Chemgauss2 output was retained. This was deemed to be the most representative binding mode for the ligand.

LigX minimization

LigX module was used for pki estimations. LigX prepares active site and ligand atoms by applying diverse checks and finishes by minimising the system and evaluating properties.

The following steps were followed: 1) the 2D ligand structure was checked. 2) BS and ligand atoms were protonated, hydrogens added and the system minimised having all heavy atoms fixed and using the default molecular mechanics force field. 3) Additional artificial forces that pull the tethered atoms to their initial coordinates (tether restraints) were added to the receptor heavy atoms 4) The prepared system was submitted to a molecular mechanic minimisation and finally, an estimation of the binding affinity is reported as *pKi*.

Default values were not modified as they seemed reasonable in the current work. The molecular mechanics force field used for all minimisations was MMFF94s [230], an all-atom force field parameterised for small organic molecules. Partial charges are based on bond-charge increments. Receptor heavy atoms beyond 8Å of the ligand were tethered.

Binding affinity estimation

The binding affinity, is usually expressed as pK_i which is equal to -log(Ki) and characterised the interactions between a ligand with its binding site . In order to describe the potency of a ligand as a protein inhibitor, radiollabeled competitive assay are run. The potency of the tested compound is then expressed as its capacity in displacing the known radiollabeld ligand:

$$K_i = (IC_{50})/ (1+[L^*]/K_d)$$

Where IC_{50} is the concentration of the tested ligand needed to obtain 50% of the maximum effect, [L*]: concentration of radiolabelled ligand and K_d the dissociation constant. The dissociation constant is defines as:

$$K_d = [P][L]/[C]$$

where

[**P**], [**L**] and [**C**] represent the concentrations of the protein, ligand and complex, respectively in the context of the formation of a ligand-protein complex (**C**):

$$C \rightleftharpoons P + L$$

In general:

- A $pK_i > 7$: the ligand is active.
- A $5 < pKi < 7$: the ligand is a weak active.
- A $pK_i < 5$ the ligand is inactive.

All binding affinities were calculated using MOE code which uses the estimated binding affinity of the London dG scoring function reported in units of pKi (μM).

4.6.3 Results

Known binding sites enumeration

Site Finder was run on all available crystal structures listed in Table 4.1. In order to validate it as a binding site detection method probing both the curved and straight conformations of tubulin, it was first tested whether Site Finder could identify co-crystallised ligand binding sites. Results for the seven available pdb structures of all known drug sites *i.e.* taxol, colchicine, vinblastine and of the exchangeable (E site) and non exchangeable (N-site) sites are presented in Figure 4.18.

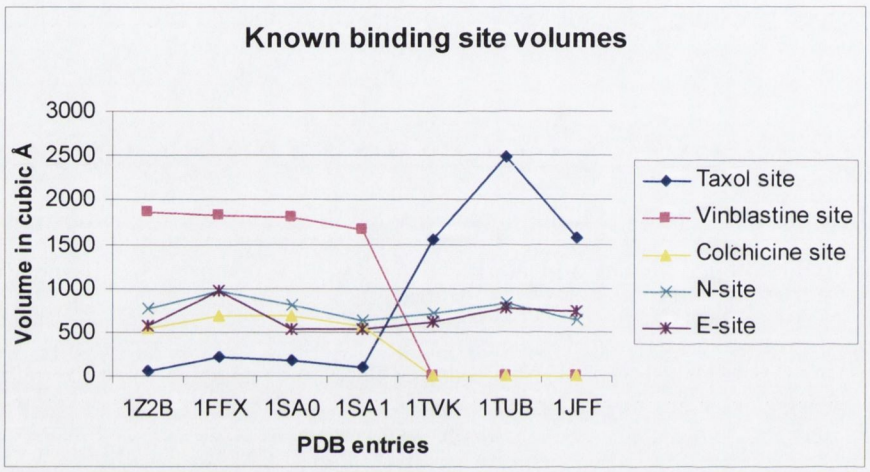


Figure 4.18: Known binding sites found by Site Finder

In some cases, the interstice between monomers, for e.g. where the N-site and the colchicine site meet or the E-site and vinblastine site meet, Site Finder generates alpha spheres that overlap both sites and represent them as a single unique large site rather than two separate ones. In order to separate these overlapping sites an X-ray structure where both sites are well defined was superposed with one where they overlapped and used to delimitate the alpha spheres to be retained for volume estimation. For instance, in the interstice between α and β monomers, GTP with its crystal structure coordinates was superposed and alpha spheres mapping GTP were used for the N-site volume estimation. Importantly, large alpha sites were observed only in curved conformation. Stathmin was used to aid crystallisation of the heterodimers, and as expected a curved complex was observed. Although used to keep heterodimers linked, the curvature of the protein tends to pull apart the two monomers creating more void space between them.

Site Finder still depicts correctly the active regions placing appropriate alpha spheres but does not manage to accurately split regions where close binding sites such as colchicine and GTP are observed. Nonetheless, as illustrated in Figure 4.19, for all the pdb entries studied the exchangeable E and N sites were correctly defined by alpha spheres and the small volume variability from one pdb to another was attributed to 1) the differences in resolution of the diverse crystal structures 2) to the difference of the two conformations. The N and E site volume variabilities were therefore not significant. In Figure 4.19, the re-entrant surface (Gaussian surface[231, 232]) was mapped around the crystallized ligand on the receptor to illustrate the general consensus between the selection of alpha spheres mapping the ligand.

Superposition of the N-sites of all seven pdb entries had a Root Mean Square Deviation (RMSD) of 1,06. Visual analysis of each residue pointed out residues βVAL 177 and βSER 178 (see Figure 4.9) to be the major cause of deviation. Their respective position gathers in two groups corresponding to the conformation types. The curved conformation corresponds to a bending of 12 Å which shifts these residues closer to the nucleotide.

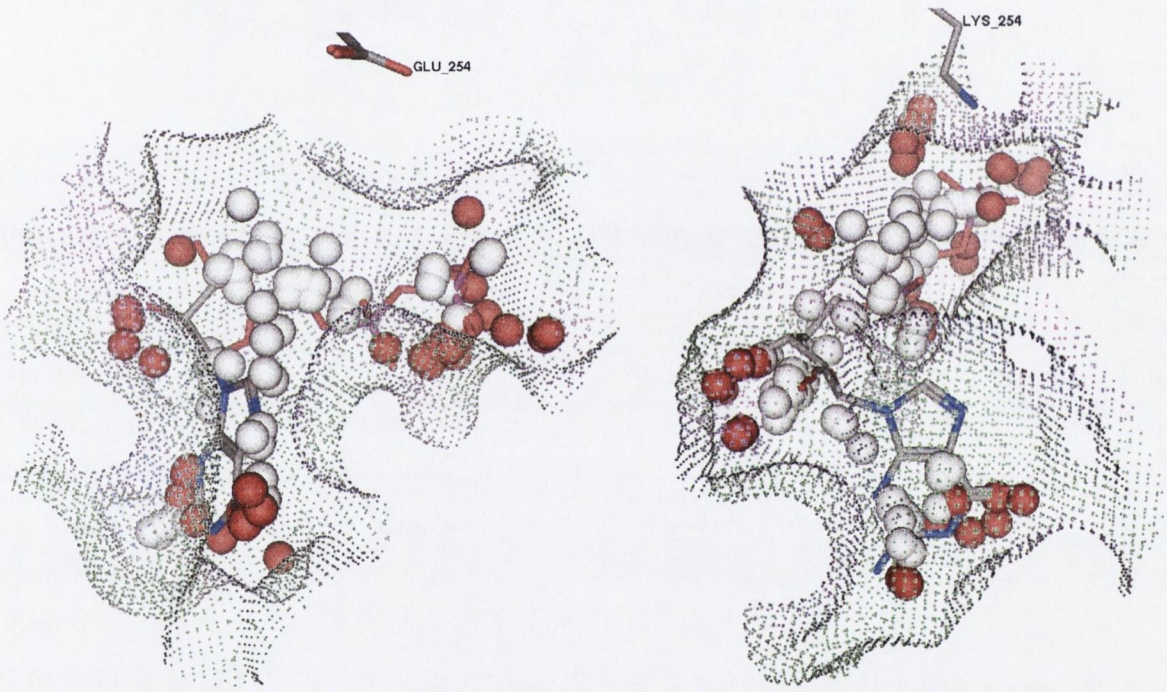


Figure 4.19: Site Finder mapping of the nucleotic sites.
Left = N-site mapping. Right = E-site mapping. Coloured in red, hydrophilic alpha spheres and in white hydrophobic alpha spheres. Gaussian surface[231, 232]: colored by H-Bonding (purple), Hydrophobicity (green) and Mild Polar (blue) regions.

The taxol binding site was found correctly only in the straight conformation PDB entries, whereas vinblastine and colchicine sites were only found in the curved conformations. The vinblastine site is not found in straight structures, as they are composed of only one heterodimer and vinblastine is

known to bind between a pair of heterodimers equivalently in both monomers and only, curved conformations satisfy these conditions. Comparison of the vinblastine site along the four curved structures showed an RMSD of 2.45 mainly attributed to the movement of the residues that lie between the drug and GDP (Asn249, Leu248, Pro175, Lys176, Val177, Ser 178, Asp 179). Site Finder is illustrated in Figure 4.20 under the same conditions as previously described for the nucleotidic sites.

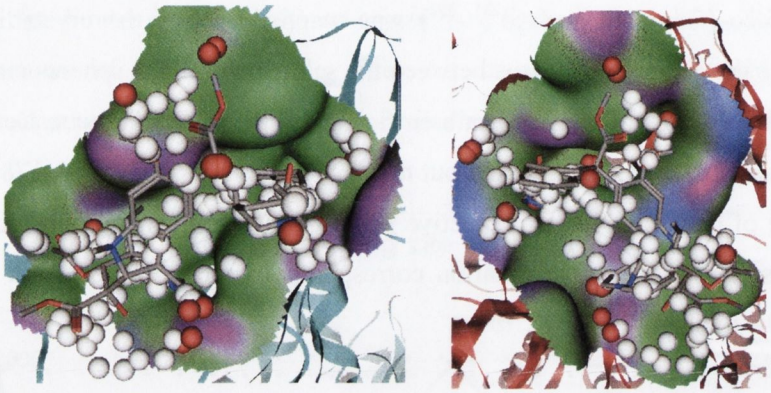


Figure 4.20: Site Finder mapping of vinblastine site(1Z2B).
On the right, α monomer view and left β monomer view of the mapping. Coloured in red: Hydrophilic alpha spheres and in white: hydrophobic alpha spheres. Gaussian surface[231, 232] colored by H-Bonding (purple), Hydrophobicity (green) and Mild Polar (blue) regions.

Interestingly, the taxol site was not detected in any of the curved structures. Indeed, in the curved conformations, the M and L loops divide the pocket into two smaller pockets. In fact, as free heterodimers are not in lateral contact with other dimers, the M-loop is not restrained by lateral forces staying close to the β monomer it belongs to. When the protein curves, the L-loop shifts due to propagation of conformational changes and closes also the pocket. This concurs with experimental data which demonstrated that if microtubules were inhibited by colchicine addition, the conformational changes induced were sufficient to prevent taxol from binding. Moreover, closer examination of taxol site region in 1FFX revealed the detection of a total of five sites, the biggest of which extends up to the multi loop region near βB3 on the top of the activation domain, and the second largest under the N-loop sandwiched between the M and L loops. The two sites are separated mainly by residues βLeu371 (from the L-loop), βPhe272 (from the beginning of the M-loop), βHis229 and βAla233 (from the core of H7). A second general observation is the higher Taxol site volume of the pdb 1TUB compared to the two other straight structures. Site finder here splits the region in four sites, two bigger and two smaller. The volume variability is attributed to the quality of the resolution. As illustrated in Figure 4.21 Site finder depiction was appropriate for the Taxol site when the pdb was of good quality i.e. 1JFF.

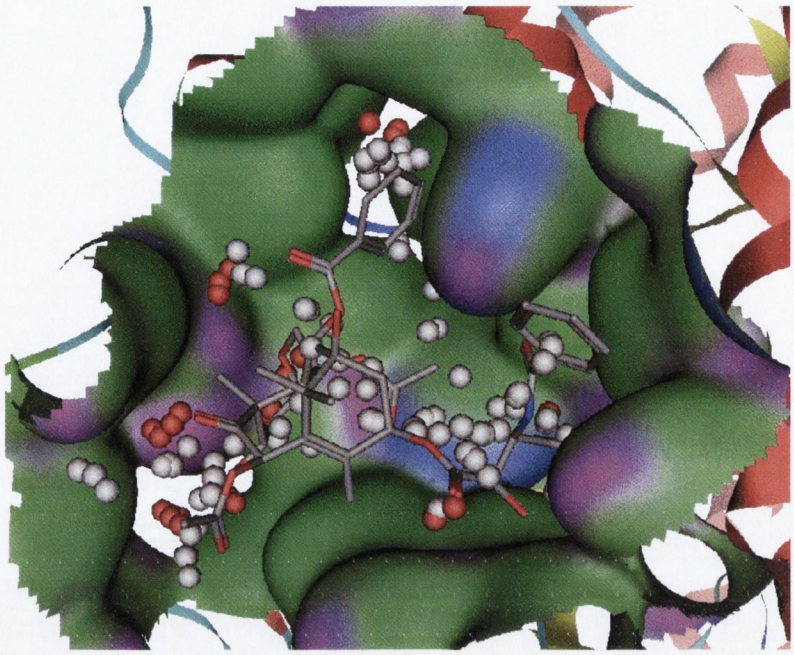


Figure 4.21: Site Finder mapping of taxol site.
Coloured in red, hydrophilic alpha spheres and in white hydrophobic alpha spheres. Gaussian surface[231, 232] colored by H-Bonding (purple), Hydrophobicity (green) and Mild Polar (blue) regions.

Similarly, the colchicine site was detected only in curved conformations. The mechanism of action of colchicine was described in the previous section describing that firstly the TC-complex is formed and then being added to the end of a protofilament prior to depolymerisation. The conformational changes induced by the addition of the TC-complex on the top of a protofilament (straight conformation) opens the colchicine pocket to enable its fitting. In the straight conformations, the pocket is therefore closed which explains that Site Finder did not enumerate it. Site finder depiction of the site is shown in Figure 4.22.

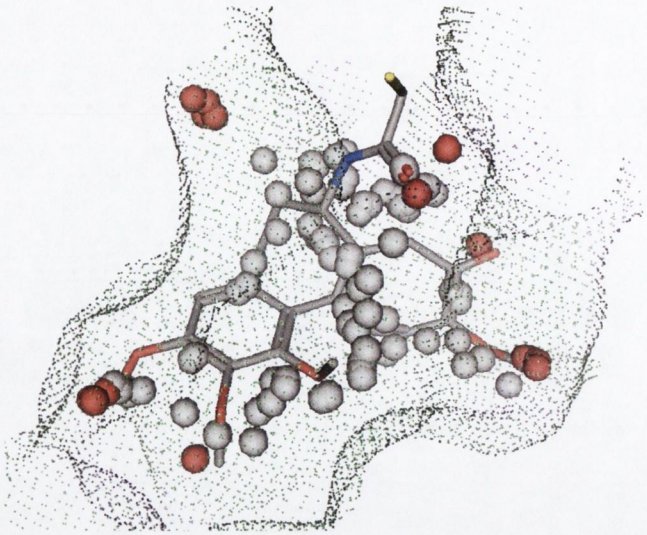


Figure 4.22: Site Finder mapping of colchicine site (1Z2B).
Coloured in red, hydrophilic alpha spheres and in white hydrophobic alpha spheres. Gaussian surface[231, 232] in dots: colored by H-Bonding (purple), Hydrophobicity (green) and Mild Polar (blue) regions.

In conclusion, Site Finder identified all known sites and alpha sphere mappings were in general in agreement with the complementary ligand properties. One should keep in mind that the final aim is to compare PBOX ligand volume to each pocket volume attributed by Site Finder. However, a perfect matching between alpha spheres and the site's co-crystallised ligand would represent the perfect theoretical case to retrieve the co-crystallised ligand. As each PBOX ligand once bound would induce specific conformational changes in tubulin a certain amount of tolerance in the comparative process may be necessary. Figure 4.23 illustrates all sites found for both conformations.

Active site enumeration

The study of these known sites demonstrates the relevance of using both conformations in order to map all possible sites as some sites are apparent in some cases and not in others. Concerning the straight conformation the PDB entry 1JFF was chosen as it the most highly resolved structure incorporating the Taxol site. In the case of the curved conformation, 1Z2B was chosen as it described the vinblastine site clearly. Only sites from the first heterodimer and the region between heterodimers were examined to avoid repetition.

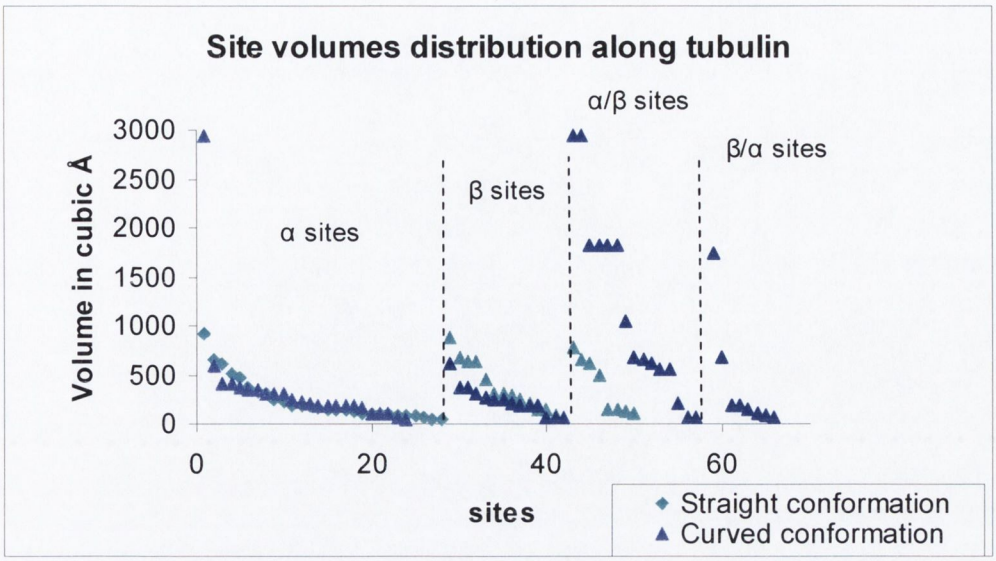


Figure 4.23: Site volume distributions for the curved and straight conformation.
α sites: sites on α monomer, β sites: sites on β monomer, α/β: sites between monomers of the same heterodimer, β/α: sites between heterodimers.

A total of 53 sites were identified for the straight conformation and 54 sites for the curved. In general, sites from the curved conformation were larger.
All sites were retained as defined by Site Finder and only known sites were manually split when their volume was overestimated due to overlapping (*i.e* split between colchicine and nucleotidic site). All of these volumes were then used to elaborate the knowledge based filter.

Validation Procedure - Shape matching

Molecules of the validation set were all submitted to the ROCS shape matching procedure on all sites for both curved and straight conformations and a Principal Component Analysis (PCA) was carried out based on Tanimoto scores outputted from ROCS to reduce the space described by all binding sites to a 3D space as depicted in Figures 4.24 and 4.25.

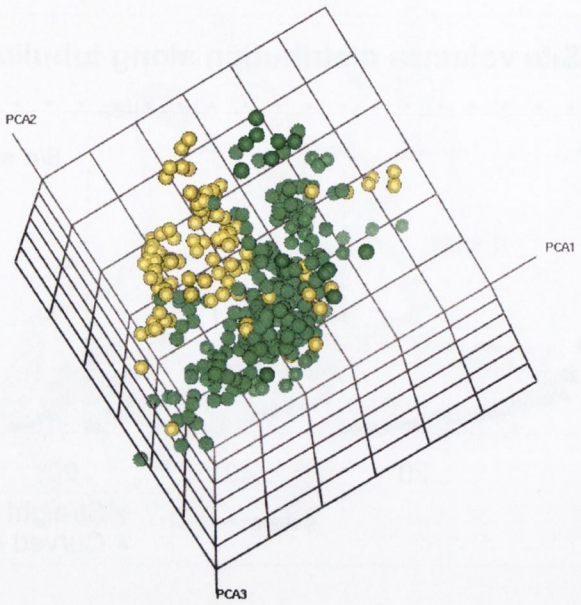


Figure 4. 24: Tubulin destabilisers and stabilisers spaces
Yellow spheres: validation set stabilisers and green spheres: validation set destabilisers

Importantly, following the ROCS shape-matching procedure, all tubulin binders gathered according to the effect they induce on tubulin (Polymerise/Depolymerise). It is important to note that no filter was used at this stage for the projection shown. The next projection in Figure 4.25, shows that the method also separates compounds according to their binding site preferences. Validation set compounds gathered into three sub-sets:

1) "taxol-site" space where compounds which prefer to bind the taxol site gather coloured in orange.
2) "Vinca Alkaloid-site" space where compounds which prefer to bind the vinca alkaloid site gather or the vinblastine site are coloured in blue.
3) "Colchicine-site" space where compounds which prefer to bind the colchicine site gather and are coloured in green

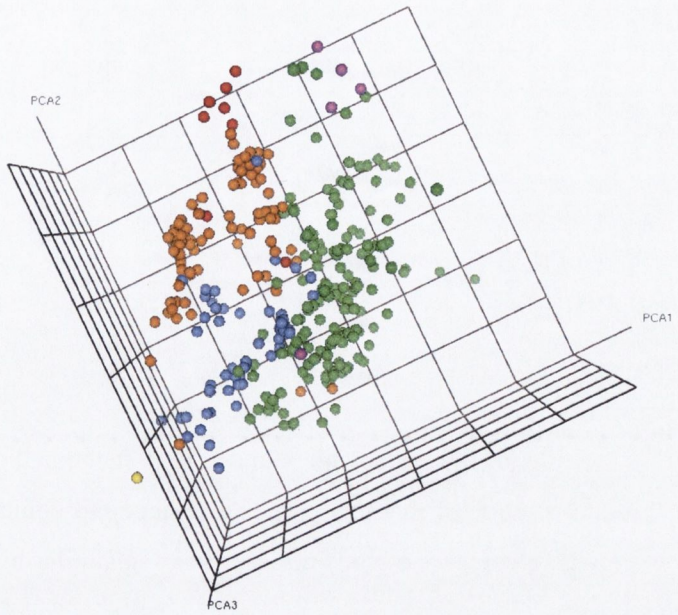


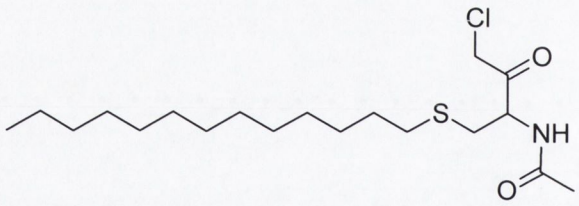
Figure 4.25: Known drugs sites spaces

Orange spheres: validation set Taxol site binders, blue spheres: validation set vinca alkaloid site binders, green spheres colchicine site binders, pink spheres: binders around nucleotidic area, red spheres: laulimalide binders and yellow sphere: vitilevuamide

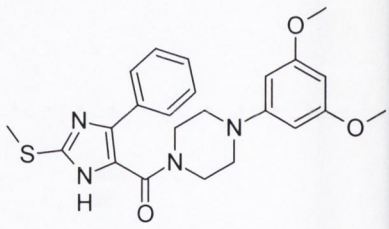
Additional to these three groups, compounds whose binding sites or regions have been suggested were projected using PCA.

4) "Laulimalide-site" space where all compounds projected are supposed to bind a common regions (coloured in red).
5) "Nucleotidic" space where compound have been shown to bind between the two monomers around the nucleotidic region, these compounds are coloured in pink.

Interestingly, the "nucleotidic" space assembled compounds such as COBRA and Curacin (see Curacin A Figure 4.12) analogues and is found in the depolymerising space. However, COBRA- 5[162, 163] is unusual as it does not follow the previous trend and is projected in "colchicine-site" space. This structure is the only COBRA compound which does not possess a ring and it overlaps in the PCA space, a 4-phenylpiperazine derivative shown to bind the colchicine site.



COBRA 5



4-phenylpiperazine

Common features such as the sulphur linker, the amide and ketone functional groups can be observed in both cases. Therefore, it can be suggest that this particular compound could bind between the two monomers around the N site where colchicine binds possibly sharing similar binding features with the 4-phenylpiperazine derivative.

Similarly, other compounds were projected in space differently to where the remainder of their family of analogs resides. For instance, one structure of the Peloruside [91] analogues (see Figure 4.15) was projected in the Taxol-like space. This compound if compared to other analogues of the same family possesses an additional six-membered ring. Corroborating this, it has been suggested in independent modelling studies that Pelorusides could bind Taxols' site in addition to the preferred site on α tubulin[185] (under B9–B10). Finally, it can be observed that an Epothilone and Eleutherobin derivative are projected incorrectly in 3D space as they both obtained higher tanimoto scores for the vinca domain region rather than in the Taxol site. This artefact is most likely due to receptor rigidity issues in the binding site identification process as no flexibility could be accounted for. Applying the knowledge-based filter would remove the possibility of either of those compounds binding to the curved conformation which is the only one containing the vinca domain site. This problem is also observed with one of the Halichodrin analogues.

Finally, Vitilavuamide, a known destabilizing agent has been previously demonstrated to bind to a site not observed as yet in tubulin. Interestingly, our PCA concurs with this as it is depicted to be completely separated from all others in space.

Blind Test

To attempt to profile a set of known tubulin binders (from WOMBAT) application of the same method was deemed appropriate and Figure 4.26 illustrates the projection results.

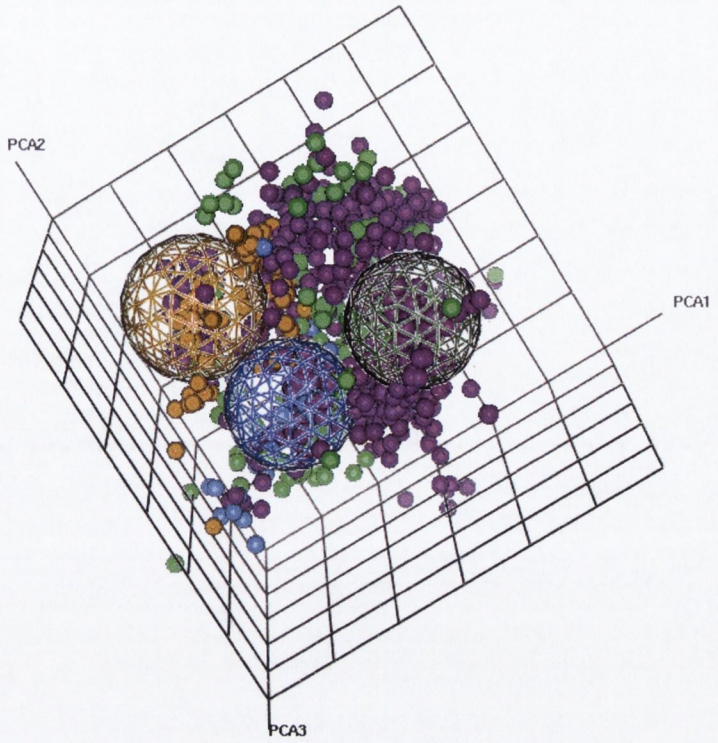


Figure 4.26: WOMBAT tubulin binders distributions.

Orange spheres: validation set Taxol site binders, blue spheres: validation set vinca alkaloid site binders, green spheres: colchicine site binders, purple spheres: WOMBAT compounds, orange large mesh spheres: WOMBAT Taxol-like, blue large mesh sphere: WOMBAT vinca-like and green large mesh spheres: WOMBAT colchicine-like.

A set of compounds that lay within each previously described binding site sub-set region was selected and the respective cluster centres are depicted in Table 4.3.

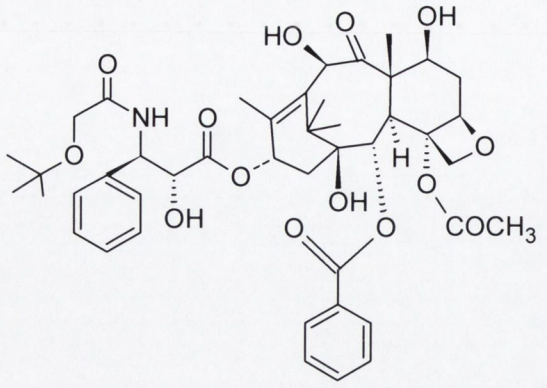
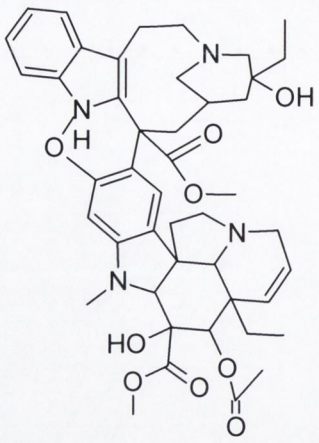
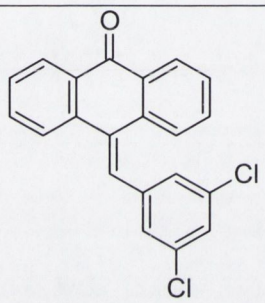
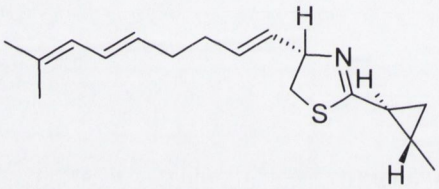
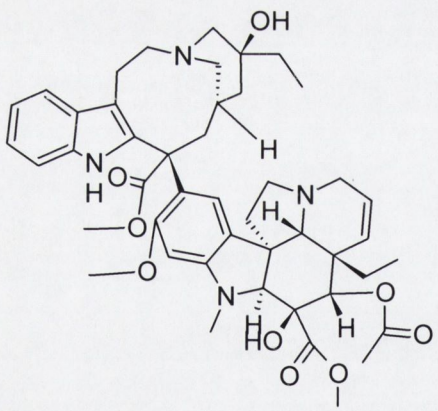
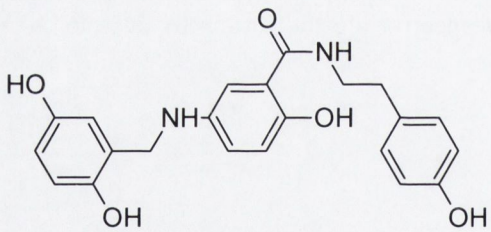
Molecules cluster centroids for the "taxol-site" space	
 Taxotere	 vinblastine
Molecules cluster centroids for the "colchicine-site" space	
 Antracenones moiety	 Curacin analogue
Molecules cluster centroids for the "vinca alkaloid -site" space	
 vinblastine analogue	 [(hydroxyphenyl)ethyl]benzamide

Table 4.3: WOMBAT's tubulin binders projection.

WOMBAT "Taxol-like", WOMBAT "vinca-like" and WOMBAT "colchicine-like" cluster centers

It is clear that this method only works in the case of assigning destabilisers and stabilisers but it is not sufficient enough to 'drill' down to assign sites of known binders whose sites were previously unknown. The method retrieved one appropriate cluster centroids per binding site space but as observed for the validation set projection, the projection presented the same artifacts when retrieving vinblastine structure in the Taxol site space. To overcome the limitations of this approach o knowledge based filter was deployed.

Projection of PBOX compounds was undertaken in the same manner. As illustrated in Figure 4.27, they all reside in destabilising space which corresponds to their actual effect. Nevertheless, as before, the method was not accurate enough to characterize PBOX binding site space as it overlapped more than one site space, among which the "vinca-alkaloid" space and "colchicine-like". At this stage it can however be surmised that PBOX might bind in the interstice between monomers but the exact whereabouts needs to be converged on.

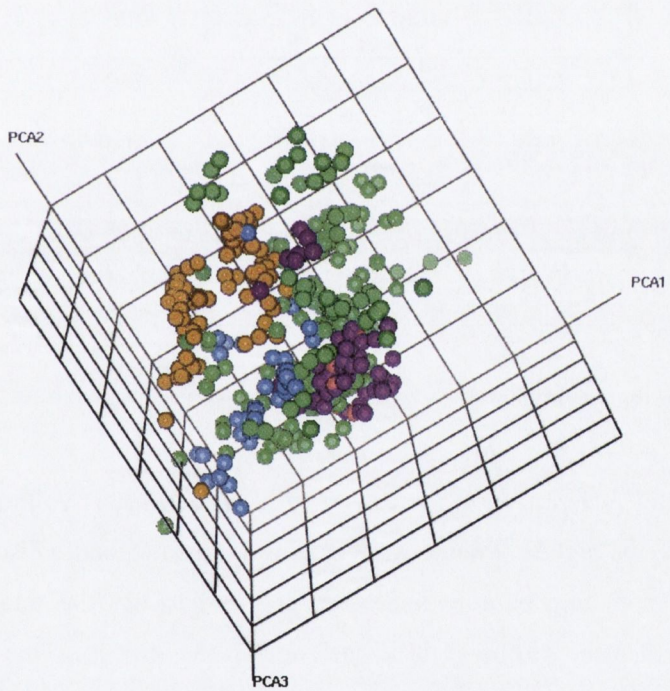


Figure 4.27: PBOX compound space.

Orange spheres: validation set Taxol site binders, blue spheres: validation set vinca alkaloid site binders, green spheres colchicine site binders and purple spheres PBOX compounds

Binding site filter elaboration

To overcome some defaults of the method a knowledge based filter was apply. For each molecule to be matched to a particular site, two filters were employed. The first, relied on knowledge of mechanism. For instance molecules known to destabilise tubulin were only matched to sites of the curved conformation. The second filter consisted in calculating the volume of the molecule and discarding every site with a smaller volume. Volumes calculated for the crystallised structures are shown in Table 4.4.

PDB ID	ligands	Volume in $Å^3$
1Z2B	GTP	338
1JFF	GDP	301
1Z2B	DAMA-colchicine	395
1Z2B	Vinblastine	760
1JFF	Taxol	800
1TUB	Taxotere	724
1TVK	Epothilone A	489

Table 4.4: Co-crystallized ligand volume

For new molecules, after generation of conformers, all van der Waals volumes were calculated and only the conformer with the highest volume was compared. This procedure discarded a number of sites per structure, *e.g.* the matching of colchicine sites reduced to 6% the total number of sites thus reducing the computational time to be spent in the comparative process. This filter was subsequently applied to the previous results and its effect on the validation set compounds is presented in Figure 4.28.

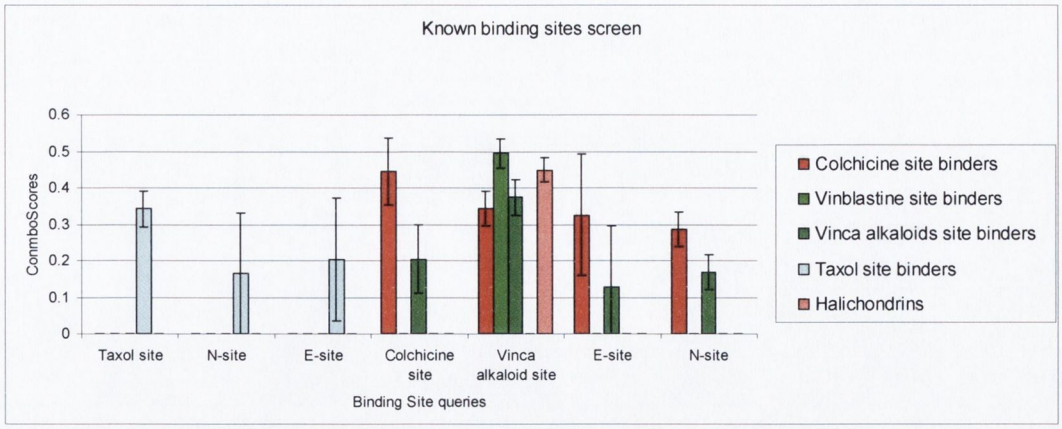


Figure 4.28: Known drugs binding site selection after filtering

The three first binding site bars on the left belong to the straight conformation and the remainder to the curved conformation, each bar is the average value obtained per family of compounds and error bars represent standard deviations.

The filtering method managed to reduce the ensemble of potential binding sites well, nevertheless, the final assignment of a binding site per family of ligands remained difficult as seen from the error bar overlapping. "Taxol-like" compounds on average prefer taxol site but the error remains quite large and does not allow for a definitive conclusion. Similarly, the other family of compounds tested such as colchicine, vinblastine and vinca alkaloids binders obtained an average highest score for their own binding site but error bars from one binding site to another introduce a level of uncertainty. However, halichondrins and vinblastine classes of compounds were the most sizeable compounds of the validation set and therefore their actual binding sites could be easily determined. In all other cases, numerous analogues with diversity in both volume and chemotype were gathered for the validation set making the assignment more difficult. It is important to keep in mind that compounds were matched to "dummy" molecules with only two categories of "dummy" atoms whereas, in reality, compound's features represent a much vaster range of features.

PBOX compounds were submitted to the same treatment, as shown in Figure 4.29.

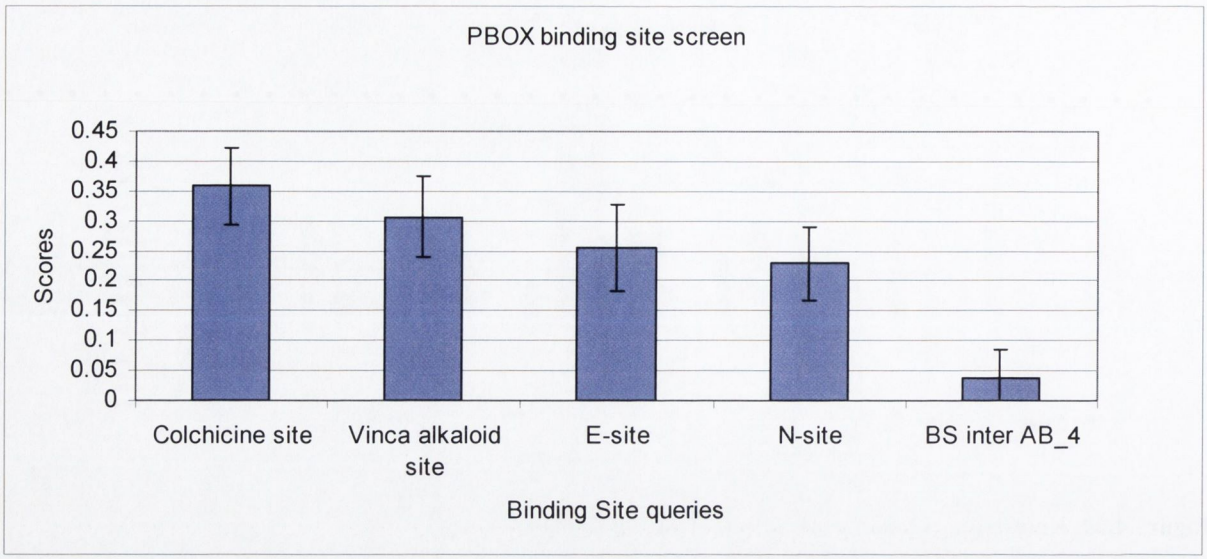


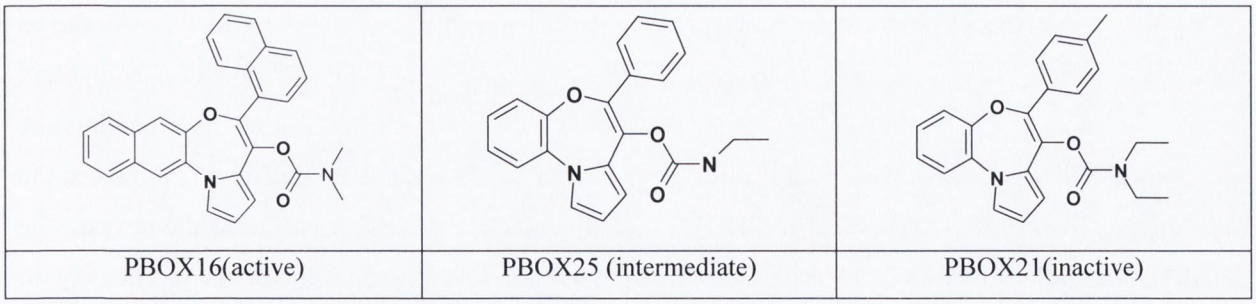
Figure 4.29: PBOX compounds binding site selection after filtering

Each bar is the average value calculated for the PBOX population, error bars represent standard deviations.

In Figure 4.29, BS inter AB_4 represents the query numbered 4 of a binding site situated at the interstice between the two α and β monomers. Based on score, this site can be discarded and the four binding sites left are colchicine, vinca alkaloid, the N-site and the E sites. Considering the scores values obtained and the large error bars, a docking study was undertaken to select a final BS.

Rigid docking, Flexible refinement and binding affinity estimation

In order to conclude on a particular binding site for the PBOX compounds, a rigid docking procedure was introduced to the overall process. Firstly all PBOX compounds were docked in the remaining binding sites, and secondly appropriate binding modes were selected based on the clustering of the docked poses. The results presented in the next Table 4.5 are these of some selected PBOX compounds representative of the complete set (structures are displayed in Figure 4.30) and of the co-crystallised structures used as a validation step for each BS studied.



PBOX16(active)	PBOX25 (intermediate)	PBOX21(inactive)

Figure 4.30: PBOX structures docked.

The rigid docking phase was validated for each binding site on its ability to retrieve the exact binding mode of the concerned co-crystallized structure. For all the binding sites studied, the X-ray coordinates of the template ligand were perfectly retrieved and ranked the highest for each set as illustrated in Table 4.5. The first stage of the rigid docking phase, consisted of choosing an exaggerated large protein volume as the binding area. For instance, concerning the colchicine site, the docking area was set at a 10Å area so that it would also enclose the N-site. Under these conditions, only the nucleotide compound was correctly placed in the N-site, colchicine was perfectly docked into its pocket while vinblastine was completely rejected from the whole area. In this light the N-site was discarded from further considerations. Further more, across the binding sites the Chemgauss2 scoring function was able to redistribute correctly each ligand to the appropriate BS as demonstrated for instance by the score of colchicine in its own site of -51.1014 vs. -47.0162 and -33.6531 in respectively vinblastine site and E-site. In addition, it has previously been shown by our group that PBOX compounds do not bind to the colchicine or vinblastine site based on competitive binding experiments and so the remaining sites selected for the docking studies were the E-site and N-site. In this light, only the E-site was considered for further studies and the docking box re-adjusted to 3 Å.

LIGANDS	E-SITE	COLCHICINE SITE	VINBLASTINE SITE
GDP	-51.6231	-50.8478	-45.4699
Colchicine	-33.6531	-51.1014	-47.0162
Vinblastine	NBM	NBM	-74.672
PBOX16	-43.7794		
PBOX25	-41.6047		
PBOX21	-39.2788		

Table 4. 5: Chemgauss2 scores for all docked compounds.

Blue: inactives PBOX, red: actives PBOX, pink: intermediate, NBM: No Binding Mode found.

Colchicine, Vinblastine and GDP affinity predictions distributed each compound to its own site as illustrated on Table 4.5. If no value is inscribed in the Table (NBM) then the corresponding compound was not docked by the rigid docking phase. The N-site was again discarded from PBOX binding site possibilities as better scores were obtained for the E-site for PBOX16. The E-site docking produced the best affinity for PBOX 16 (most active PBOX) while the inactive PBOX 21 produced the lowest. The final ranking was consistent with the activity of each compound. To estimate the propensity of PBOX16 to bind to the E-site, LigX minimization was carried out and compared with GDP in its own site. This permitted the calculation of pK_i's and ligand efficiency. The docked conformation of PBOX16 is illustrated below and compared with the GDP binding mode in Figure 4.31 and 4.32.

```
Ligand Properties:
      weight      :   446.506
      logP        :     6.700
      TPSA        :    43.700
      don         :     0
      acc         :     1
      reactive    :    no
      MM/GBVI     :   767.150 kcal/mol
      affinity    :    11.974 pKi
      efficiency  :     0.352
```

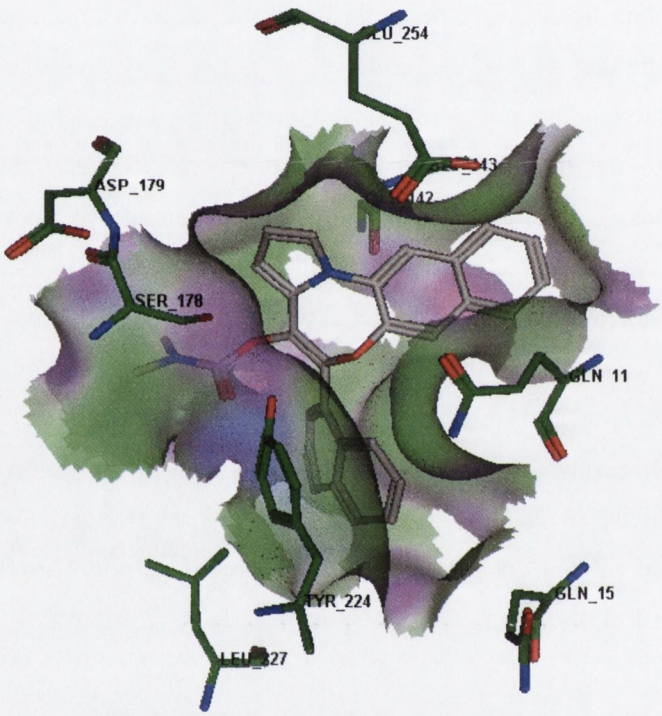


Figure 4.31: Optimised binding pose of PBOX16 in the E site$^+$.

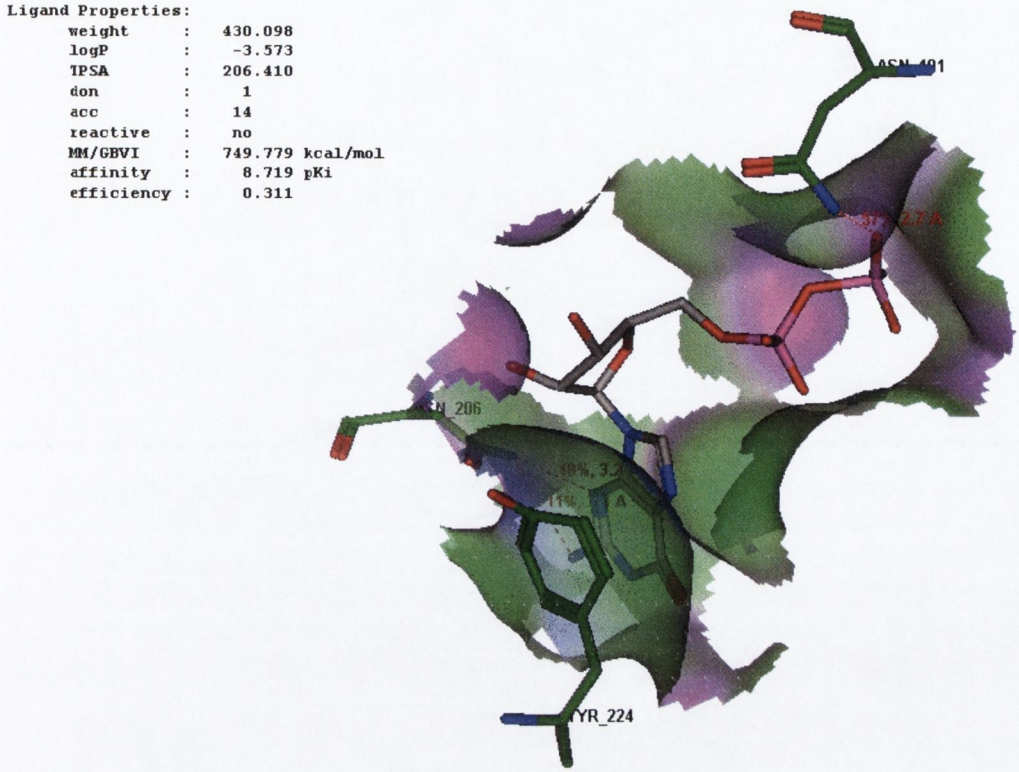


Figure 4.32: Optimised binding pose of GDP in the E site[+].

Although no hydrogen bonding patterns are observed in the LigX optimised complex of PBOX16, the pK_i is nonetheless (11.974) excellent and mainly due to fit and π-stacking interactions with Tyr224 which is also observed with GDP (8.719). The pocket appears to be well complemented in terms of both size and shape of PBOX16 leading to a pK_i values wich is usually indicative of active ligand. This could be a reason for the inactivity of the smaller PBOX25 and PBOX 21. Finally, GDPs lower pK_i suggests that PBOX16 could compete well with the nucleotide for binding to the E-site.

Pharmacophore matching

Finally, the refined pharmacophore elaborated in chapter 3 was superposed to the potential binding site by superimposing PBOX16 best fitted conformer on the conformer docked in the E-site. as illustrated in Figure 4.33,

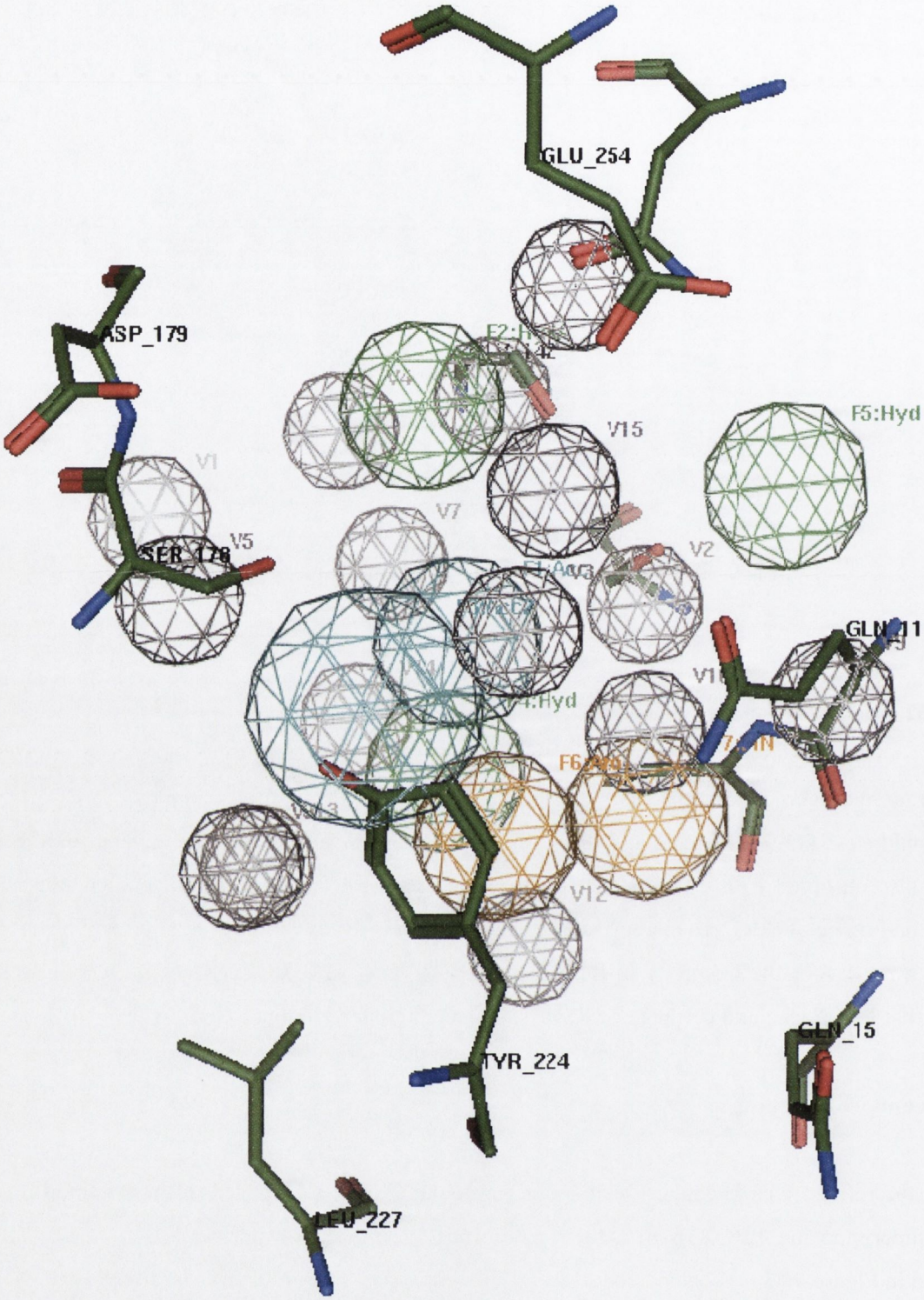


Figure 4.33: PBOX ligand based pharmacophore superposed on the E site^{+}.

The volume constraints of the ligand based pharmacophore are either overlayed on the following highlighted key residues Asn401, Ser178, Gln11 or are found tight between residues in inaccessible areas. The pharmacophore features are also logically placed as shown in Figure 4.34.a where the nucleotide and PBOX16 are overlayed on the pharmacophore. The π-stacking interaction of the nucleotide with residue Tyr224 is conserved for the PBOX16 docked structures whereas it is not allowed by PBOX21 (*cf.* Figure 4.34.(b).). This correlates with the SAR which demonstrated the importance of the diphenyl feature for activity. Figure 4.34.(c) shows the agreeing overlay of the two conformers from the docking study and the ligand based pharmaphore fit and Figure 4.34.(d) illustrates the overlay of all docked structures.

If the hypothesis made is correct and PBOX active compounds compete with the nucleotide for the E-site then addition of PBOX compounds are highly likely to depolymerise pre-polymerised MTs. Indeed, PBOX compound have been shown to depolymerise tubulin encouraging the proposition of a mechanism of displacement of the GTP-cap concordant by the affinity values. Of course, PBOX still could bind to an undetected area inducing conformational changes which would remove the GTP-cap. Indeed, this study is inherently restricted by the actual crystallisation model. Despite the flexibility introduced by the generation of multiple ligand conformations and the LigX refinement step which allows both receptor and ligand movements, the very first binding sites description of the whole protein is fully depending on the atoms protein coordinates. An alternate binding site might have been found if another version of the curved conformation had been chosen. However, according to the present study the most probable binding site is deemed to be the E-site.

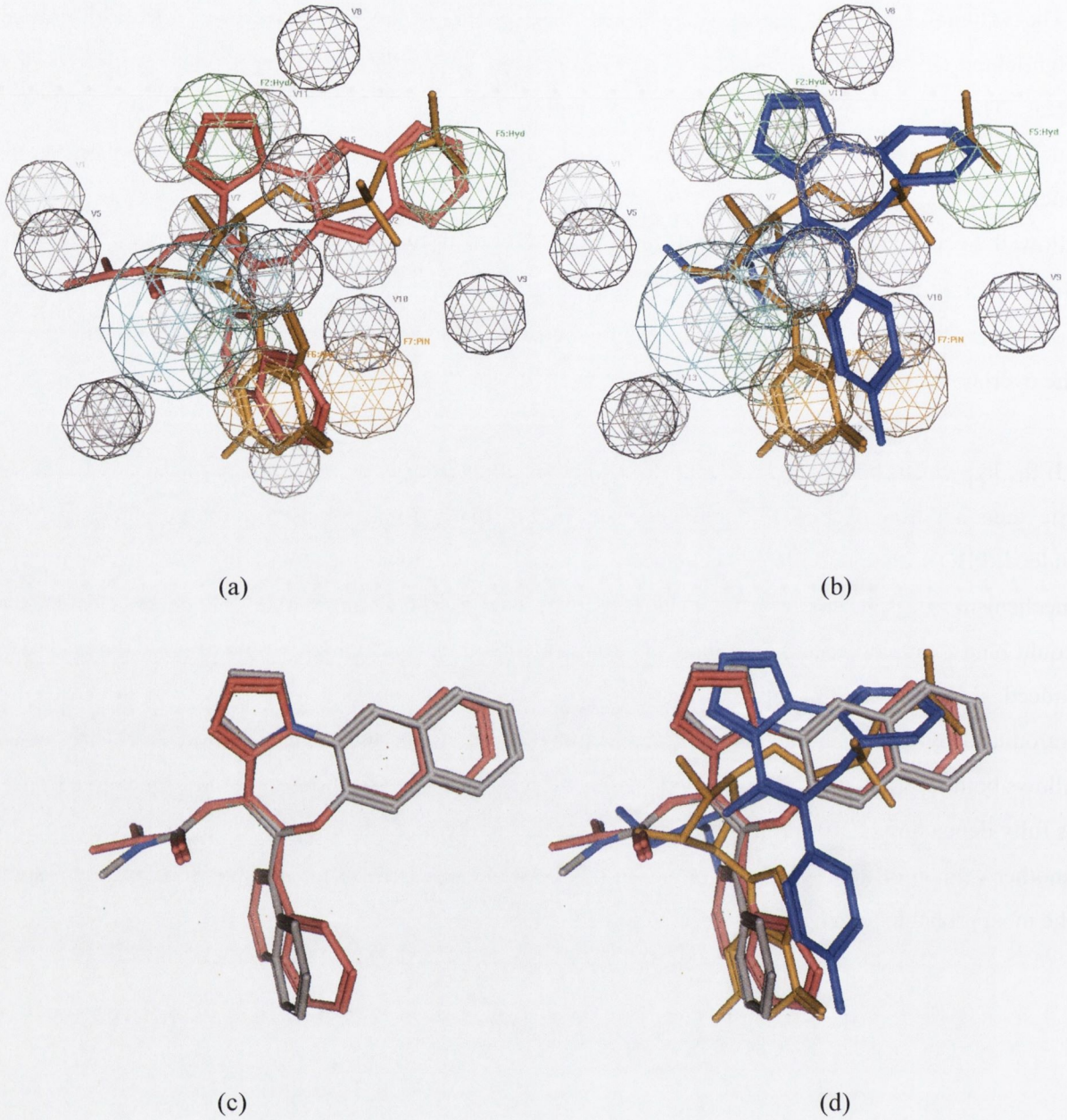


Figure 4.34. Comparaison docked structures and ligand based phamacophore

Red and grey : PBOX16 active, blue: PBOX21 inactive; orange: PBOXGDP

(a): coloured in orange GDP and in red PBOX16 docked structures and ligand based pharmacophore overlay,

(b): coloured in orange GDP and in blue PBOX21 docked structures and ligand based pharmacophore overlay,

(c): coloured red PBOX16 docked conformer and in grey PBOX16 best fit in the ligand based pharmacophore overlay, (d) coloured in orange GDP, in red PBOX16 and in blue PBOX21 docked conformers and in grey PBOX 16 best fit in the ligand based pharmacophore.

4.7 Conclusion

In this section, a novel method to redistribute tubulin binders to their appropriate binding site was developed, also incorporating biological knowledge of ligand effects. PBOX compounds were suggested to bind the nucleotidic site, however experimental verification is needed such as a competitive assay with, for instance, radiolabelled nucleotide. Furthermore, this hypothesis is supported by the ligand-based pharmacophore elaborated in Chapter 3. The absolute confirmation could only be attained through biological assay such as systematic rational mutation of the protein or ideally binding mode clarification through crystallisation of the PBOX16-tubulin complex.

In the presented work, several steps could be improved with the expectation of enhancing the final convergence on a particular binding pocket. Currently an automated pharmacophore builder is being developed as a refinement of the shape matching process. Thus, the enumeration of all possible binding pockets that tubulin contains was restricted by the conformation of the crystallized structure. To overcome this, molecular dynamics on the curved structure could be undertaken in order to loop the entire procedure and better represent the true conformational flexibility of the receptor. Finally, to improve the ligand flexibility, rigid structures from the validation set such as constrained macrocycles, should be submitted to a rigorous conformational analysis to maximize the conformational space explored thus avoiding unfairness in the treatment of all diverse validation set compounds. The presented work depicts a novel and interesting way of "inverse-docking" as it was both fast and efficient for the retrieval of known binding sites. This method can be extrapolated to other binders with as yet undetermined binding sites. This novel binding site suggested for PBOX compounds opens doors for lead optimisation through docking of virtual combinatorial libraries based on the PBOX scaffold within the tubulin E-site – directed design within the site constraints should yield ligands with enhanced activity.

References

1. Mulligan, J. M.; Greene, L. M.; Cloonan, S.; Mc Gee, M. M.; Onnis, V.; Campiani, G.; Fattorusso, C.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Identification of tubulin as the molecular target of proapoptotic pyrrolo-1,5-benzoxazepines. *Mol Pharmacol* 2006, 70, (1), 60-70.
2. Pichichero, M. E.; Avers, C. J., The evolution of cellular movement in eukaryotes: the role of microfilaments and microtubules. *Subcell Biochem* 1973, 2, (1), 97-105.
3. Hyams, J. S.; Stebbings, H., The mechanism of microtubule associated cytoplasmic transport. Isolation and preliminary characterisation of a microtubule transport system. *Cell Tissue Res* 1979, 196, (1), 103-16.
4. Cheeseman, I. M.; Desai, A., Molecular architecture of the kinetochore-microtubule interface. *Nat Rev Mol Cell Biol* 2008, 9, (1), 33-46.
5. Jackson, J. R.; Patrick, D. R.; Dar, M. M.; Huang, P. S., Targeted anti-mitotic therapies: can we improve on tubulin agents? *Nat Rev Cancer* 2007, 7, (2), 107-17.
6. Mandelkow, E. M.; Mandelkow, E., Microtubule oscillations. *Cell Motil Cytoskeleton* 1992, 22, (4), 235-44.
7. Jordan, M. A.; Wilson, L., Microtubules as a target for anticancer drugs. *Nat Rev Cancer* 2004, 4, (4), 253-65.
8. Nogales, E.; Wolf, S. G.; Downing, K. H., Structure of the alpha beta tubulin dimer by electron crystallography. *Nature* 1998, 391, (6663), 199-203.
9. Ravelli, R. B.; Gigant, B.; Curmi, P. A.; Jourdain, I.; Lachkar, S.; Sobel, A.; Knossow, M., Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* 2004, 428, (6979), 198-202.
10. Gigant, B.; Wang, C.; Ravelli, R. B.; Roussi, F.; Steinmetz, M. O.; Curmi, P. A.; Sobel, A.; Knossow, M., Structural basis for the regulation of tubulin by vinblastine. *Nature* 2005, 435, (7041), 519-22.
11. Hirokawa, N., Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science* 1998, 279, (5350), 519-26.
12. Vale, R. D., The molecular motor toolbox for intracellular transport. *Cell* 2003, 112, (4), 467-80.
13. Mitchison, T.; Kirschner, M., Dynamic instability of microtubule growth. *Nature* 1984, 312, (5991), 237-42.
14. Margolis, R. L.; Wilson, L., Microtubule treadmills--possible molecular machinery. *Nature* 1981, 293, (5835), 705-11.
15. Watts, N. R.; Cheng, N.; West, W.; Steven, A. C.; Sackett, D. L., The cryptophycin-tubulin ring structure indicates two points of curvature in the tubulin dimer. *Biochemistry* 2002, 41, (42), 12662-9.
16. Nogales, E.; Wang, H. W., Structural intermediates in microtubule assembly and disassembly: how and why? *Curr Opin Cell Biol* 2006, 18, (2), 179-84.
17. Prigozhina, N. L.; Oakley, C. E.; Lewis, A. M.; Nayak, T.; Osmani, S. A.; Oakley, B. R., gamma-tubulin plays an essential role in the coordination of mitotic events. *Mol Biol Cell* 2004, 15, (3), 1374-86.
18. Lu, Q.; Luduena, R. F., In vitro analysis of microtubule assembly of isotypically pure tubulin dimers. Intrinsic differences in the assembly properties of alpha beta II, alpha beta III, and alpha beta IV tubulin dimers in the absence of microtubule-associated proteins. *J Biol Chem* 1994, 269, (3), 2041-7.
19. Roach, M. C.; Boucher, V. L.; Walss, C.; Ravdin, P. M.; Luduena, R. F., Preparation of a monoclonal antibody specific for the class I isotype of beta-tubulin: the beta isotypes of tubulin differ in their cellular distributions within human tissues. *Cell Motil Cytoskeleton* 1998, 39, (4), 273-85.
20. Luduena, R. F., Multiple forms of tubulin: different gene products and covalent modifications. *Int Rev Cytol* 1998, 178, 207-75.
21. Luduena, R. F.; Banerjee, A.; Khan, I. A., Tubulin structure and biochemistry. *Curr Opin Cell Biol* 1992, 4, (1), 53-7.
22. Oakley, B. R.; Oakley, C. E.; Yoon, Y.; Jung, M. K., Gamma-tubulin is a component of the spindle pole body that is essential for microtubule function in Aspergillus nidulans. *Cell* 1990, 61, (7), 1289-301.
23. Inclan, Y. F.; Nogales, E., Structural models for the self-assembly and microtubule interactions of gamma-, delta- and epsilon-tubulin. *J Cell Sci* 2001, 114, (Pt 2), 413-22.
24. Tavosanis, G.; Llamazares, S.; Goulielmos, G.; Gonzalez, C., Essential role for gamma-tubulin in the acentriolar female meiotic spindle of Drosophila. *Embo J* 1997, 16, (8), 1809-19.
25. Panda, D.; Miller, H. P.; Banerjee, A.; Luduena, R. F.; Wilson, L., Microtubule dynamics in vitro are regulated by the tubulin isotype composition. *Proc Natl Acad Sci U S A* 1994, 91, (24), 11358-62.

26. Savage, C.; Hamelin, M.; Culotti, J. G.; Coulson, A.; Albertson, D. G.; Chalfie, M., mec-7 is a beta-tubulin gene required for the production of 15-protofilament microtubules in Caenorhabditis elegans. *Genes Dev* 1989, 3, (6), 870-81.
27. Fukushige, T.; Siddiqui, Z. K.; Chou, M.; Culotti, J. G.; Gogonea, C. B.; Siddiqui, S. S.; Hamelin, M., MEC-12, an alpha-tubulin required for touch sensitivity in C. elegans. *J Cell Sci* 1999, 112 (Pt 3), 395-403.
28. Lowe, J.; Amos, L. A., Crystal structure of the bacterial cell-division protein FtsZ. *Nature* 1998, 391, (6663), 203-6.
29. Erickson, H. P., FtsZ, a prokaryotic homolog of tubulin? *Cell* 1995, 80, (3), 367-70.
30. RayChaudhuri, D.; Park, J. T., Escherichia coli cell-division gene ftsZ encodes a novel GTP-binding protein. *Nature* 1992, 359, (6392), 251-4.
31. Schlieper, D.; Oliva, M. A.; Andreu, J. M.; Lowe, J., Structure of bacterial tubulin BtubA/B: evidence for horizontal gene transfer. *Proc Natl Acad Sci U S A* 2005, 102, (26), 9170-5.
32. Carpenter, E. J.; Huzil, J. T.; Luduena, R. F.; Tuszynski, J. A., Homology modeling of tubulin: influence predictions for microtubule's biophysical properties. *Eur Biophys J* 2006, 36, (1), 35-43.
33. Sept, D.; Baker, N. A.; McCammon, J. A., The physical basis of microtubule structure and stability. *Protein Sci* 2003, 12, (10), 2257-61.
34. Burkhart, C. A.; Kavallaris, M.; Band Horwitz, S., The role of beta-tubulin isotypes in resistance to antimitotic drugs. *Biochim Biophys Acta* 2001, 1471, (2), O1-9.
35. Kavallaris, M.; Tait, A. S.; Walsh, B. J.; He, L.; Horwitz, S. B.; Norris, M. D.; Haber, M., Multiple microtubule alterations are associated with Vinca alkaloid resistance in human leukemia cells. *Cancer Res* 2001, 61, (15), 5803-9.
36. Zhang, C. C.; Yang, J. M.; White, E.; Murphy, M.; Levine, A.; Hait, W. N., The role of MAP4 expression in the sensitivity to paclitaxel and resistance to vinca alkaloids in p53 mutant cells. *Oncogene* 1998, 16, (12), 1617-24.
37. Gan, P. P.; Pasquier, E.; Kavallaris, M., Class III beta-tubulin mediates sensitivity to chemotherapeutic drugs in non small cell lung cancer. *Cancer Res* 2007, 67, (19), 9356-63.
38. Tommasi, S.; Mangia, A.; Lacalamita, R.; Bellizzi, A.; Fedele, V.; Chiriatti, A.; Thomssen, C.; Kendzierski, N.; Latorre, A.; Lorusso, V.; Schittulli, F.; Zito, F.; Kavallaris, M.; Paradiso, A., Cytoskeleton and paclitaxel sensitivity in breast cancer: the role of beta-tubulins. *Int J Cancer* 2007, 120, (10), 2078-85.
39. Gokmen-Polar, Y.; Escuin, D.; Walls, C. D.; Soule, S. E.; Wang, Y.; Sanders, K. L.; Lavallee, T. M.; Wang, M.; Guenther, B. D.; Giannakakou, P.; Sledge, G. W., Jr., beta-Tubulin mutations are associated with resistance to 2-methoxyestradiol in MDA-MB-435 cancer cells. *Cancer Res* 2005, 65, (20), 9406-14.
40. Giannakakou, P.; Sackett, D. L.; Kang, Y. K.; Zhan, Z.; Buters, J. T.; Fojo, T.; Poruchynsky, M. S., Paclitaxel-resistant human ovarian cancer cells have mutant beta-tubulins that exhibit impaired paclitaxel-driven polymerization. *J Biol Chem* 1997, 272, (27), 17118-25.
41. Lowe, J.; Li, H.; Downing, K. H.; Nogales, E., Refined structure of alpha beta-tubulin at 3.5 A resolution. *J Mol Biol* 2001, 313, (5), 1045-57.
42. Gigant, B.; Curmi, P. A.; Martin-Barbey, C.; Charbaut, E.; Lachkar, S.; Lebeau, L.; Siavoshian, S.; Sobel, A.; Knossow, M., The 4 A X-ray structure of a tubulin:stathmin-like domain complex. *Cell* 2000, 102, (6), 809-16.
43. Nettles, J. H.; Li, H.; Cornett, B.; Krahn, J. M.; Snyder, J. P.; Downing, K. H., The binding mode of epothilone A on alpha,beta-tubulin by electron crystallography. *Science* 2004, 305, (5685), 866-9.
44. Kirschner, M. W.; Honig, L. S.; Williams, R. C., Quantitative electron microscopy of microtubule assembly in vitro. *J Mol Biol* 1975, 99, (2), 263-76.
45. Scheele, R. B.; Borisy, G. G., Electron microscopy of metal-shadowed and negatively stained microtubule protein. Structure of the 30 S oligomer. *J Biol Chem* 1978, 253, (8), 2846-51.
46. Mandelkow, E. M.; Rapp, R.; Mandelkow, E., Microtubule structure studied by quick freezing: cryo-electron microscopy and freeze fracture. *J Microsc* 1986, 141, (Pt 3), 361-73.
47. Beese, L.; Stubbs, G.; Cohen, C., Microtubule structure at 18 A resolution. *J Mol Biol* 1987, 194, (2), 257-64.
48. Nogales, E.; Wolf, S. G.; Khan, I. A.; Luduena, R. F.; Downing, K. H., Structure of tubulin at 6.5 A and location of the taxol-binding site. *Nature* 1995, 375, (6530), 424-7.
49. Aldaz, H.; Rice, L. M.; Stearns, T.; Agard, D. A., Insights into microtubule nucleation from the crystal structure of human gamma-tubulin. *Nature* 2005, 435, (7041), 523-7.
50. Kikkawa, M.; Hirokawa, N., High-resolution cryo-EM maps show the nucleotide binding pocket of KIF1A in open and closed conformations. *Embo J* 2006, 25, (18), 4187-94.
51. Burns, R. G.; Farrell, K. W., Getting to the heart of beta-tubulin. *Trends Cell Biol* 1996, 6, (8), 297-303.

52. Molecular Operating Environment, Software developed by Chemical Computing Group Inc. www.chemcomp.com.
53. Stewart, R. J.; Farrell, K. W.; Wilson, L., Role of GTP hydrolysis in microtubule polymerization: evidence for a coupled hydrolysis mechanism. *Biochemistry* 1990, 29, (27), 6489-98.
54. Richards, K. L.; Anders, K. R.; Nogales, E.; Schwartz, K.; Downing, K. H.; Botstein, D., Structure-function relationships in yeast tubulins. *Mol Biol Cell* 2000, 11, (5), 1887-903.
55. Hyman, A. A.; Chretien, D.; Arnal, I.; Wade, R. H., Structural changes accompanying GTP hydrolysis in microtubules: information from a slowly hydrolyzable analogue guanylyl-(alpha,beta)-methylene-diphosphonate. *J Cell Biol* 1995, 128, (1-2), 117-25.
56. Mandelkow, E. M.; Mandelkow, E.; Milligan, R. A., Microtubule dynamics and microtubule caps: a time-resolved cryo-electron microscopy study. *J Cell Biol* 1991, 114, (5), 977-91.
57. Nogales, E.; Whittaker, M.; Milligan, R. A.; Downing, K. H., High-resolution model of the microtubule. *Cell* 1999, 96, (1), 79-88.
58. Nogales, E.; Wang, H. W.; Niederstrasser, H., Tubulin rings: which way do they curve? *Curr Opin Struct Biol* 2003, 13, (2), 256-61.
59. Muller-Reichert, T.; Chretien, D.; Severin, F.; Hyman, A. A., Structural changes at microtubule ends accompanying GTP hydrolysis: information from a slowly hydrolyzable analogue of GTP, guanylyl (alpha,beta)methylenediphosphonate. *Proc Natl Acad Sci U S A* 1998, 95, (7), 3661-6.
60. Baas, P. W.; Buster, D. W., Slow axonal transport and the genesis of neuronal morphology. *J Neurobiol* 2004, 58, (1), 3-17.
61. Honore, S.; Pasquier, E.; Braguer, D., Understanding microtubule dynamics for improved cancer therapy. *Cell Mol Life Sci* 2005, 62, (24), 3039-56.
62. Amos, L. A.; Schlieper, D., Microtubules and maps. *Adv Protein Chem* 2005, 71, 257-98.
63. Cassimeris, L.; Spittle, C., Regulation of microtubule-associated proteins. *Int Rev Cytol* 2001, 210, 163-226.
64. Marx, A.; Muller, J.; Mandelkow, E. M.; Hoenger, A.; Mandelkow, E., Interaction of kinesin motors, microtubules, and MAPs. *J Muscle Res Cell Motil* 2006, 27, (2), 125-37.
65. Moores, C. A.; Milligan, R. A., Lucky 13-microtubule depolymerisation by kinesin-13 motors. *J Cell Sci* 2006, 119, (Pt 19), 3905-13.
66. Chen, J.; Kanai, Y.; Cowan, N. J.; Hirokawa, N., Projection domains of MAP2 and tau determine spacings between microtubules in dendrites and axons. *Nature* 1992, 360, (6405), 674-7.
67. Lin, Y. T.; Cheng, J. T.; Liang, L. C.; Ko, C. Y.; Lo, Y. K.; Lu, P. J., The binding and phosphorylation of Thr231 is critical for Tau's hyperphosphorylation and functional regulation by glycogen synthase kinase 3beta. *J Neurochem* 2007, 103, (2), 802-13.
68. Kar, S.; Fan, J.; Smith, M. J.; Goedert, M.; Amos, L. A., Repeat motifs of tau bind to the insides of microtubules in the absence of taxol. *Embo J* 2003, 22, (1), 70-7.
69. Dye, R. B.; Fink, S. P.; Williams, R. C., Jr., Taxol-induced flexibility of microtubules and its reversal by MAP-2 and Tau. *J Biol Chem* 1993, 268, (10), 6847-50.
70. Felgner, H.; Frank, R.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E.; Ludin, B.; Matus, A.; Schliwa, M., Domains of neuronal microtubule-associated proteins and flexural rigidity of microtubules. *J Cell Biol* 1997, 138, (5), 1067-75.
71. Bosc, C.; Andrieux, A.; Job, D., STOP proteins. *Biochemistry* 2003, 42, (42), 12125-32.
72. Sobel, A.; Boutterin, M. C.; Beretta, L.; Chneiweiss, H.; Doye, V.; Peyro-Saint-Paul, H., Intracellular substrates for extracellular signaling. Characterization of a ubiquitous, neuron-enriched phosphoprotein (stathmin). *J Biol Chem* 1989, 264, (7), 3765-72.
73. Hailat, N.; Strahler, J.; Melhem, R.; Zhu, X. X.; Brodeur, G.; Seeger, R. C.; Reynolds, C. P.; Hanash, S., N-myc gene amplification in neuroblastoma is associated with altered phosphorylation of a proliferation related polypeptide (Op18). *Oncogene* 1990, 5, (11), 1615-8.
74. Howell, B.; Larsson, N.; Gullberg, M.; Cassimeris, L., Dissociation of the tubulin-sequestering and microtubule catastrophe-promoting activities of oncoprotein 18/stathmin. *Mol Biol Cell* 1999, 10, (1), 105-18.
75. Steinmetz, M. O.; Kammerer, R. A.; Jahnke, W.; Goldie, K. N.; Lustig, A.; van Oostrum, J., Op18/stathmin caps a kinked protofilament-like tubulin tetramer. *Embo J* 2000, 19, (4), 572-80.
76. Hoenger, A.; Sack, S.; Thormahlen, M.; Marx, A.; Muller, J.; Gross, H.; Mandelkow, E., Image reconstructions of microtubules decorated with monomeric and dimeric kinesins: comparison with x-ray structure and implications for motility. *J Cell Biol* 1998, 141, (2), 419-30.

77. Sosa, H.; Dias, D. P.; Hoenger, A.; Whittaker, M.; Wilson-Kubalek, E.; Sablin, E.; Fletterick, R. J.; Vale, R. D.; Milligan, R. A., A model for the microtubule-Ncd motor protein complex obtained by cryo-electron microscopy and image analysis. *Cell* 1997, 90, (2), 217-24.
78. Arnal, I.; Metoz, F.; DeBonis, S.; Wade, R. H., Three-dimensional structure of functional motor proteins on microtubules. *Curr Biol* 1996, 6, (10), 1265-70.
79. Warner, S. L.; Gray, P. J.; Von Hoff, D. D., Tubulin-associated drug targets: Aurora kinases, Polo-like kinases, and others. *Semin Oncol* 2006, 33, (4), 436-48.
80. Hirose, K.; Lockhart, A.; Cross, R. A.; Amos, L. A., Three-dimensional cryoelectron microscopy of dimeric kinesin and ncd motor domains on microtubules. *Proc Natl Acad Sci U S A* 1996, 93, (18), 9539-44.
81. Sosa, H.; Hoenger, A.; Milligan, R. A., Three different approaches for calculating the three-dimensional structure of microtubules decorated with kinesin motor domains. *J Struct Biol* 1997, 118, (2), 149-58.
82. Mandelkow, E.; Hoenger, A., Structures of kinesin and kinesin-microtubule interactions. *Curr Opin Cell Biol* 1999, 11, (1), 34-44.
83. Schwartz, K.; Richards, K.; Botstein, D., BIM1 encodes a microtubule-binding protein in yeast. *Mol Biol Cell* 1997, 8, (12), 2677-91.
84. Tirnauer, J. S.; O'Toole, E.; Berrueta, L.; Bierer, B. E.; Pellman, D., Yeast Bim1p promotes the G1-specific dynamics of microtubules. *J Cell Biol* 1999, 145, (5), 993-1007.
85. Okada, Y.; Hirokawa, N., Mechanism of the single-headed processivity: diffusional anchoring between the K-loop of kinesin and the C terminus of tubulin. *Proc Natl Acad Sci U S A* 2000, 97, (2), 640-5.
86. Tucker, C.; Goldstein, L. S., Probing the kinesin-microtubule interaction. *J Biol Chem* 1997, 272, (14), 9481-8.
87. Walker, R. A., ncd and kinesin motor domains interact with both alpha- and beta-tubulin. *Proc Natl Acad Sci U S A* 1995, 92, (13), 5960-4.
88. Kim, H.; Jensen, C. G.; Rebhun, L. I., The binding of MAP-2 and tau on brain microtubules in vitro: implications for microtubule structure. *Ann N Y Acad Sci* 1986, 466, 218-39.
89. White, E. A.; Burton, P. R.; Himes, R. H., Polymorphic assembly of subtilisin-cleaved tubulin. *Cell Motil Cytoskeleton* 1987, 7, (1), 31-8.
90. Giannakakou, P.; Gussio, R.; Nogales, E.; Downing, K. H.; Zaharevitz, D.; Bollbuck, B.; Poy, G.; Sackett, D.; Nicolaou, K. C.; Fojo, T., A common pharmacophore for epothilone and taxanes: molecular basis for drug resistance conferred by tubulin mutations in human cancer cells. *Proc Natl Acad Sci U S A* 2000, 97, (6), 2904-9.
91. Gaitanos, T. N.; Buey, R. M.; Diaz, J. F.; Northcote, P. T.; Teesdale-Spittle, P.; Andreu, J. M.; Miller, J. H., Peloruside A does not bind to the taxoid site on beta-tubulin and retains its activity in multidrug-resistant cell lines. *Cancer Res* 2004, 64, (15), 5063-7.
92. Kowalski, R. J.; Giannakakou, P.; Gunasekera, S. P.; Longley, R. E.; Day, B. W.; Hamel, E., The microtubule-stabilizing agent discodermolide competitively inhibits the binding of paclitaxel (Taxol) to tubulin polymers, enhances tubulin nucleation reactions more potently than paclitaxel, and inhibits the growth of paclitaxel-resistant cells. *Mol Pharmacol* 1997, 52, (4), 613-22.
93. Pryor, D. E.; O'Brate, A.; Bilcer, G.; Diaz, J. F.; Wang, Y.; Wang, Y.; Kabaki, M.; Jung, M. K.; Andreu, J. M.; Ghosh, A. K.; Giannakakou, P.; Hamel, E., The microtubule stabilizing agent laulimalide does not bind in the taxoid site, kills cells resistant to paclitaxel and epothilones, and may not require its epoxide moiety for activity. *Biochemistry* 2002, 41, (29), 9109-15.
94. Aneja, R.; Lopus, M.; Zhou, J.; Vangapandu, S. N.; Ghaleb, A.; Yao, J.; Nettles, J. H.; Zhou, B.; Gupta, M.; Panda, D.; Chandra, R.; Joshi, H. C., Rational design of the microtubule-targeting anti-breast cancer drug EM015. *Cancer Res* 2006, 66, (7), 3782-91.
95. Wong, V. K.; Chiu, P.; Chung, S. S.; Chow, L. M.; Zhao, Y. Z.; Yang, B. B.; Ko, B. C., Pseudolaric acid B, a novel microtubule-destabilizing agent that circumvents multidrug resistance phenotype and exhibits antitumor activity in vivo. *Clin Cancer Res* 2005, 11, (16), 6002-11.
96. Sengupta, S.; Thomas, S. A., Drug target interaction of tubulin-binding drugs in cancer therapy. *Expert Rev Anticancer Ther* 2006, 6, (10), 1433-47.
97. Mahindroo, N.; Liou, J.-P.; Chang, J.-Y.; Hsieh, H.-P., Antitubulin agents for the treatment of cancer – a medicinal chemistry update *Exp Opin on Ther Pat* 2006, 16, (5), 647-691.
98. Rowinsky, E. K.; Calvo, E., Novel agents that target tublin and related elements. *Semin Oncol* 2006, 33, (4), 421-35.
99. Zhou, J.; Giannakakou, P., Targeting microtubules for cancer chemotherapy. *Curr Med Chem Anticancer Agents* 2005, 5, (1), 65-71.

100. Hait, W. N.; Rubin, E.; Goodin, S., Tubulin-targeting agents. *Cancer Chemother Biol Response Modif* 2005, 22, 35-59.
101. Attard, G.; Greystoke, A.; Kaye, S.; De Bono, J., Update on tubulin-binding agents. *Pathol Biol (Paris)* 2006, 54, (2), 72-84.
102. Simmons, T. L.; Andrianasolo, E.; McPhail, K.; Flatt, P.; Gerwick, W. H., Marine natural products as anticancer drugs. *Mol Cancer Ther* 2005, 4, (2), 333-42.
103. Wallace, S. L., Colchicine. *Semin Arthritis Rheum* 1974, 3, (4), 369-81.
104. Kumar, A.; Seghal, N.; Naidu, P. S.; Padi, S. S.; Goyal, R., Colchicines-induced neurotoxicity as an animal model of sporadic dementia of Alzheimer's type. *Pharmacol Rep* 2007, 59, (3), 274-83.
105. Schlesinger, N., Reassessing the safety of intravenous and compounded injectable colchicine in acute gout treatment. *Expert Opin Drug Saf* 2007, 6, (6), 625-9.
106. Tozer, G. M.; Kanthou, C.; Baguley, B. C., Disrupting tumour blood vessels. *Nat Rev Cancer* 2005, 5, (6), 423-35.
107. Fankhauser, G.; Humphrey, R. R., The Rare Occurrence of Mitosis Without Spindle Apparatus ("Colchicine Mitosis") Producing Endopolyploidy in Embryos of the Axolotl. *Proc Natl Acad Sci U S A* 1952, 38, (12), 1073-82.
108. Panda, D.; Daijo, J. E.; Jordan, M. A.; Wilson, L., Kinetic stabilization of microtubule dynamics at steady state in vitro by substoichiometric concentrations of tubulin-colchicine complex. *Biochemistry* 1995, 34, (31), 9921-9.
109. Hastie, S. B., Interactions of colchicine with tubulin. *Pharmacol Ther* 1991, 51, (3), 377-401.
110. Andreu, J. M.; Timasheff, S. N., Conformational states of tubulin liganded to colchicine, tropolone methyl ether, and podophyllotoxin. *Biochemistry* 1982, 21, (25), 6465-76.
111. Bhattacharyya, B.; Panda, D.; Gupta, S.; Banerjee, M., Anti-mitotic activity of colchicine and the structural basis for its interaction with tubulin. *Med Res Rev* 2008, 28, (1), 155-83.
112. Nguyen, T. L.; McGrath, C.; Hermone, A. R.; Burnett, J. C.; Zaharevitz, D. W.; Day, B. W.; Wipf, P.; Hamel, E.; Gussio, R., A common pharmacophore for a diverse set of colchicine site inhibitors using a structure-based approach. *J Med Chem* 2005, 48, (19), 6107-16.
113. Farce, A.; Loge, C.; Gallet, S.; Lebegue, N.; Carato, P.; Chavatte, P.; Berthelot, P.; Lesieur, D., Docking study of ligands into the colchicine binding site of tubulin. *J Enzyme Inhib Med Chem* 2004, 19, (6), 541-7.
114. Tripathi, A.; Fornabaio, M.; Kellogg, G. E.; Gupton, J. T.; Gewirtz, D. A.; Yeudall, W. A.; Vega, N. E.; Mooberry, S. L., Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. *Bioorg Med Chem* 2007.
115. Medrano, F. J.; Andreu, J. M.; Gorbunoff, M. J.; Timasheff, S. N., Roles of ring C oxygens in the binding of colchicine to tubulin. *Biochemistry* 1991, 30, (15), 3770-7.
116. Shivdasani, R. A., Molecular and transcriptional regulation of megakaryocyte differentiation. *Stem Cells* 2001, 19, (5), 397-407.
117. Chen, G.; Zheng, S.; Luo, X.; Shen, J.; Zhu, W.; Liu, H.; Gui, C.; Zhang, J.; Zheng, M.; Puah, C. M.; Chen, K.; Jiang, H., Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J Comb Chem* 2005, 7, (3), 398-406.
118. Friedman, P. A.; Platzer, E. G., Interaction of anthelmintic benzimidazoles and benzimidazole derivatives with bovine brain tubulin. *Biochim Biophys Acta* 1978, 544, (3), 605-14.
119. Pipeline Pilot, http://accelrys.com/products/scitegic/ Software developed by Scietegic, a product from Accelrys., Pieline client version 6.0.2.0 , Pipeline Pilot Server version 6.5.1.0.
120. Lee, J. C.; Field, D. J.; Lee, L. L., Effects of nocodazole on structures of calf brain tubulin. *Biochemistry* 1980, 19, (26), 6209-15.
121. Head, J.; Lee, L. L.; Field, D. J.; Lee, J. C., Equilibrium and rapid kinetic studies on nocodazole-tubulin interaction. *J Biol Chem* 1985, 260, (20), 11060-6.
122. Rappl, C.; Barbier, P.; Bourgarel-Rey, V.; Gregoire, C.; Gilli, R.; Carre, M.; Combes, S.; Finet, J. P.; Peyrot, V., Interaction of 4-arylcoumarin analogues of combretastatins with microtubule network of HBL100 cells and binding to tubulin. *Biochemistry* 2006, 45, (30), 9210-8.
123. Bellina, F.; Cauteruccio, S.; Monti, S.; Rossi, R., Novel imidazole-based combretastatin A-4 analogues: evaluation of their in vitro antitumor activity and molecular modeling study of their binding to the colchicine site of tubulin. *Bioorg Med Chem Lett* 2006, 16, (22), 5757-62.
124. Kong, Y.; Grembecka, J.; Edler, M. C.; Hamel, E.; Mooberry, S. L.; Sabat, M.; Rieger, J.; Brown, M. L., Structure-based discovery of a boronic acid bioisostere of combretastatin A-4. *Chem Biol* 2005, 12, (9), 1007-14.
125. Griggs, J.; Metcalfe, J. C.; Hesketh, R., Targeting tumour vasculature: the development of combretastatin A4. *Lancet Oncol* 2001, 2, (2), 82-7.

126. Cirla, A.; Mann, J., Combretastatins: from natural products to drug discovery. *Nat Prod Rep* 2003, 20, (6), 558-64.
127. Shan, B.; Medina, J. C.; Santha, E.; Frankmoelle, W. P.; Chou, T. C.; Learned, R. M.; Narbut, M. R.; Stott, D.; Wu, P.; Jaen, J. C.; Rosen, T.; Timmermans, P. B.; Beckmann, H., Selective, covalent modification of beta-tubulin residue Cys-239 by T138067, an antitumor agent with in vivo efficacy against multidrug-resistant tumors. *Proc Natl Acad Sci U S A* 1999, 96, (10), 5686-91.
128. Brancale, A.; Silvestri, R., Indole, a core nucleus for potent inhibitors of tubulin polymerization. *Med Res Rev* 2007, 27, (2), 209-38.
129. Leoni, L. M.; Hamel, E.; Genini, D.; Shih, H.; Carrera, C. J.; Cottam, H. B.; Carson, D. A., Indanocine, a microtubule-binding indanone and a selective inducer of apoptosis in multidrug-resistant cancer cells. *J Natl Cancer Inst* 2000, 92, (3), 217-24.
130. Beckers, T.; Reissmann, T.; Schmidt, M.; Burger, A. M.; Fiebig, H. H.; Vanhoefer, U.; Pongratz, H.; Hufsky, H.; Hockemeyer, J.; Frieser, M.; Mahboobi, S., 2-aroylindoles, a novel class of potent, orally active small molecule tubulin inhibitors. *Cancer Res* 2002, 62, (11), 3113-9.
131. Kim do, Y.; Kim, K. H.; Kim, N. D.; Lee, K. Y.; Han, C. K.; Yoon, J. H.; Moon, S. K.; Lee, S. S.; Seong, B. L., Design and biological evaluation of novel tubulin inhibitors as antimitotic agents using a pharmacophore binding model with tubulin. *J Med Chem* 2006, 49, (19), 5664-70.
132. Prinz, H.; Ishii, Y.; Hirano, T.; Stoiber, T.; Camacho Gomez, J. A.; Schmidt, P.; Dussmann, H.; Burger, A. M.; Prehn, J. H.; Gunther, E. G.; Unger, E.; Umezawa, K., Novel benzylidene-9(10H)-anthracenones as highly active antimicrotubule agents. Synthesis, antiproliferative activity, and inhibition of tubulin polymerization. *J Med Chem* 2003, 46, (15), 3382-94.
133. Edsall, A. B.; Mohanakrishnan, A. K.; Yang, D.; Fanwick, P. E.; Hamel, E.; Hanson, A. D.; Agoston, G. E.; Cushman, M., Effects of altering the electronics of 2-methoxyestradiol on cell proliferation, on cytotoxicity in human cancer cell cultures, and on tubulin polymerization. *J Med Chem* 2004, 47, (21), 5126-39.
134. Sengupta, S.; Smitha, S. L.; Thomas, N. E.; Santhoshkumar, T. R.; Devi, S. K.; Sreejalekshmi, K. G.; Rajasekharan, K. N., 4-Amino-5-benzoyl-2-(4-methoxyphenylamino)thiazole (DAT1): a cytotoxic agent towards cancer cells and a probe for tubulin-microtubule system. *Br J Pharmacol* 2005, 145, (8), 1076-83.
135. Liou, J. P.; Chang, J. Y.; Chang, C. W.; Chang, C. Y.; Mahindroo, N.; Kuo, F. M.; Hsieh, H. P., Synthesis and structure-activity relationships of 3-aminobenzophenones as antimitotic agents. *J Med Chem* 2004, 47, (11), 2897-905.
136. Romagnoli, R.; Baraldi, P. G.; Pavani, M. G.; Tabrizi, M. A.; Preti, D.; Fruttarolo, F.; Piccagli, L.; Jung, M. K.; Hamel, E.; Borgatti, M.; Gambari, R., Synthesis and biological evaluation of 2-amino-3-(3',4',5'-trimethoxybenzoyl)-5-aryl thiophenes as a new class of potent antitubulin agents. *J Med Chem* 2006, 49, (13), 3906-15.
137. Kasibhatla, S.; Gourdeau, H.; Meerovitch, K.; Drewe, J.; Reddy, S.; Qiu, L.; Zhang, H.; Bergeron, F.; Bouffard, D.; Yang, Q.; Herich, J.; Lamothe, S.; Cai, S. X.; Tseng, B., Discovery and mechanism of action of a novel series of apoptosis inducers with potential vascular targeting activity. *Mol Cancer Ther* 2004, 3, (11), 1365-74.
138. Zhang, Q.; Peng, Y.; Wang, X. I.; Keenan, S. M.; Arora, S.; Welsh, W. J., Highly potent triazole-based tubulin polymerization inhibitors. *J Med Chem* 2007, 50, (4), 749-54.
139. Ohki, H.; Hirotani, K.; Naito, H.; Ohsuki, S.; Minami, M.; Ejima, A.; Koiso, Y.; Hashimoto, Y., Synthesis and mechanism of action of novel pyrimidinyl pyrazole derivatives possessing antiproliferative activity. *Bioorg Med Chem Lett* 2002, 12, (21), 3191-3.
140. Cao, T. M.; Durrant, D.; Tripathi, A.; Liu, J.; Tsai, S.; Kellogg, G. E.; Simoni, D.; Lee, R. M., Stilbene derivatives that are colchicine site microtubule inhibitors have antileukemic activity and minimal systemic toxicity. *Am J Hematol* 2008.
141. Kim, S.; Min, S. Y.; Lee, S. K.; Cho, W. J., Comparative molecular field analysis study of stilbene derivatives active against A549 lung carcinoma. *Chem Pharm Bull (Tokyo)* 2003, 51, (5), 516-21.
142. Correia, J. J.; Lobert, S., Physiochemical aspects of tubulin-interacting antimitotic drugs. *Curr Pharm Des* 2001, 7, (13), 1213-28.
143. Erickson, H. P., Negatively stained vinblastine aggregates. *Ann N Y Acad Sci* 1975, 253, 51-2.
144. Rai, S. S.; Wolff, J., Localization of the vinblastine-binding site on beta-tubulin. *J Biol Chem* 1996, 271, (25), 14707-11.
145. Downing, K. H., Structural basis for the interaction of tubulin with proteins and drugs that affect microtubule dynamics. *Annu Rev Cell Dev Biol* 2000, 16, 89-111.

146. Li, Y.; Kobayashi, H.; Tokiwa, Y.; Hashimoto, Y.; Iwasaki, S., Interaction of phomopsin A with porcine brain tubulin. Inhibition of tubulin polymerization and binding at a rhizoxin binding site. *Biochem Pharmacol* 1992, 43, (2), 219-24.
147. Mitra, A.; Sept, D., Localization of the antimitotic peptide and depsipeptide binding site on beta-tubulin. *Biochemistry* 2004, 43, (44), 13955-62.
148. Aherne, G. W.; Hardcastle, A.; Valenti, M.; Bryant, A.; Rogers, P.; Pettit, G. R.; Srirangam, J. K.; Kelland, L. R., Antitumour evaluation of dolastatins 10 and 15 and their measurement in plasma by radioimmunoassay. *Cancer Chemother Pharmacol* 1996, 38, (3), 225-32.
149. Hamel, E., Antimitotic natural products and their interactions with tubulin. *Med Res Rev* 1996, 16, (2), 207-31.
150. Anderson, H. J.; Coleman, J. E.; Andersen, R. J.; Roberge, M., Cytotoxic peptides hemiasterlin, hemiasterlin A and hemiasterlin B induce mitotic arrest and abnormal spindle formation. *Cancer Chemother Pharmacol* 1997, 39, (3), 223-6.
151. Nunes, M.; Kaplan, J.; Wooters, J.; Hari, M.; Minnick, A. A., Jr.; May, M. K.; Shi, C.; Musto, S.; Beyer, C.; Krishnamurthy, G.; Qiu, Y.; Loganzo, F.; Ayral-Kaloustian, S.; Zask, A.; Greenberger, L. M., Two photoaffinity analogues of the tripeptide, hemiasterlin, exclusively label alpha-tubulin. *Biochemistry* 2005, 44, (18), 6844-57.
152. Ravi, M.; Zask, A.; Rush, T. S., 3rd, Structure-based identification of the binding site for the hemiasterlin analogue HTI-286 on tubulin. *Biochemistry* 2005, 44, (48), 15871-9.
153. Wang, Z.; McPherson, P. A.; Raccor, B. S.; Balachandran, R.; Zhu, G.; Day, B. W.; Vogt, A.; Wipf, P., Structure-activity and high-content imaging analyses of novel tubulysins. *Chem Biol Drug Des* 2007, 70, (2), 75-86.
154. Khalil, M. W.; Sasse, F.; Lunsdorf, H.; Elnakady, Y. A.; Reichenbach, H., Mechanism of action of tubulysin, an antimitotic peptide from myxobacteria. *Chembiochem* 2006, 7, (4), 678-83.
155. Bai, R.; Taylor, G. F.; Cichacz, Z. A.; Herald, C. L.; Kepler, J. A.; Pettit, G. R.; Hamel, E., The spongistatins, potently cytotoxic inhibitors of tubulin polymerization, bind in a distinct region of the vinca domain. *Biochemistry* 1995, 34, (30), 9714-21.
156. Uckun, F. M.; Sudbeck, E. A.; Mao, C.; Ghosh, S.; Liu, X. P.; Vassilev, A. O.; Navara, C. S.; Narla, R. K., Structure-based design of novel anticancer agents. *Curr Cancer Drug Targets* 2001, 1, (1), 59-71.
157. Schmidt, E. W.; Raventos-Suarez, C.; Bifano, M.; Menendez, A. T.; Fairchild, C. R.; Faulkner, D. J., Scleritodermin A, a cytotoxic cyclic peptide from the lithistid sponge Scleritoderma nodosum. *J Nat Prod* 2004, 67, (3), 475-8.
158. Elnakady, Y. A.; Sasse, F.; Lunsdorf, H.; Reichenbach, H., Disorazol A1, a highly effective antimitotic agent acting on tubulin polymerization and inducing apoptosis in mammalian cells. *Biochem Pharmacol* 2004, 67, (5), 927-35.
159. Zhang, N.; Ayral-Kaloustian, S.; Nguyen, T.; Afragola, J.; Hernandez, R.; Lucas, J.; Gibbons, J.; Beyer, C., Synthesis and SAR of [1,2,4]triazolo[1,5-a]pyrimidines, a class of anticancer agents with a unique mechanism of tubulin inhibition. *J Med Chem* 2007, 50, (2), 319-27.
160. Fodstad, O.; Breistol, K.; Pettit, G. R.; Shoemaker, R. H.; Boyd, M. R., Comparative antitumor activities of halichondrins and vinblastine against human tumor xenografts. *J Exp Ther Oncol* 1996, 1, (2), 119-25.
161. Dabydeen, D. A.; Burnett, J. C.; Bai, R.; Verdier-Pinard, P.; Hickford, S. J.; Pettit, G. R.; Blunt, J. W.; Munro, M. H.; Gussio, R.; Hamel, E., Comparison of the activities of the truncated halichondrin B analog NSC 707389 (E7389) with those of the parent compound and a proposed binding site on tubulin. *Mol Pharmacol* 2006, 70, (6), 1866-75.
162. Jan, S. T.; Mao, C.; Vassilev, A. O.; Navara, C. S.; Uckun, F. M., COBRA-1, a rationally-designed epoxy-THF containing compound with potent tubulin depolymerizing activity as a novel anticancer agent. *Bioorg Med Chem Lett* 2000, 10, (11), 1193-7.
163. Uckun, F. M., Rationally designed anti-mitotic agents with pro-apoptotic activity. *Curr Pharm Des* 2001, 7, (16), 1627-39.
164. Lazo, J. S.; Tamura, K.; Vogt, A.; Jung, J. K.; Rodriguez, S.; Balachandran, R.; Day, B. W.; Wipf, P., Antimitotic actions of a novel analog of the fungal metabolite palmarumycin CP1. *J Pharmacol Exp Ther* 2001, 296, (2), 364-71.
165. Koizumi, Y.; Arai, M.; Tomoda, H.; Omura, S., Oxaline, a fungal alkaloid, arrests the cell cycle in M phase by inhibition of tubulin polymerization. *Biochim Biophys Acta* 2004, 1693, (1), 47-55.
166. Chang, J. Y.; Chang, C. Y.; Kuo, C. C.; Chen, L. T.; Wein, Y. S.; Kuo, Y. H., Salvinal, a novel microtubule inhibitor isolated from Salvia miltiorrhizae Bunge (Danshen), with antimitotic activity in multidrug-sensitive and -resistant human tumor cells. *Mol Pharmacol* 2004, 65, (1), 77-84.

167. Edler, M. C.; Fernandez, A. M.; Lassota, P.; Ireland, C. M.; Barrows, L. R., Inhibition of tubulin polymerization by vitilevuamide, a bicyclic marine peptide, at a site distinct from colchicine, the vinca alkaloids, and dolastatin 10. *Biochem Pharmacol* 2002, 63, (4), 707-15.
168. Kingston, D. G.; Newman, D. J., Taxoids: cancer-fighting compounds from nature. *Curr Opin Drug Discov Devel* 2007, 10, (2), 130-44.
169. Galletti, E.; Magnani, M.; Renzulli, M. L.; Botta, M., Paclitaxel And Docetaxel Resistance: Molecular Mechanisms and Development of New Generation Taxanes. *ChemMedChem* 2007, 2, (7), 920-942.
170. Karliga, B.; Schilling, J. K.; Kingston, D. G.; Bane, S.; Ravindra, R.; Talinli, N., Synthesis and biological evaluation of N-(arylsulfanyl)carbonyl analogues of paclitaxel (taxol). *Chem Biodivers* 2006, 3, (4), 396-404.
171. Choy, H., Taxanes in combined modality therapy for solid tumors. *Crit Rev Oncol Hematol* 2001, 37, (3), 237-47.
172. Vici, P.; Belli, F.; Di Lauro, L.; Amodio, A.; Conti, F.; Foggi, P.; Gionfra, T.; Morelli, M. F.; Botti, C.; Ferraironi, A.; Lopez, M., Docetaxel in patients with anthracycline-resistant advanced breast cancer. *Oncology* 2001, 60, (1), 60-5.
173. Jourdain, L.; Curmi, P.; Sobel, A.; Pantaloni, D.; Carlier, M. F., Stathmin: a tubulin-sequestering protein which forms a ternary T2S complex with two tubulin molecules. *Biochemistry* 1997, 36, (36), 10817-21.
174. Amos, L. A.; Lowe, J., How Taxol stabilises microtubule structure. *Chem Biol* 1999, 6, (3), R65-9.
175. Bollag, D. M.; McQueney, P. A.; Zhu, J.; Hensens, O.; Koupal, L.; Liesch, J.; Goetz, M.; Lazarides, E.; Woods, C. M., Epothilones, a new class of microtubule-stabilizing agents with a taxol-like mechanism of action. *Cancer Res* 1995, 55, (11), 2325-33.
176. He, L.; Jagtap, P. G.; Kingston, D. G.; Shen, H. J.; Orr, G. A.; Horwitz, S. B., A common pharmacophore for Taxol and the epothilones based on the biological activity of a taxane molecule lacking a C-13 side chain. *Biochemistry* 2000, 39, (14), 3972-8.
177. Mani, S.; Macapinlac, M., Jr.; Goel, S.; Verdier-Pinard, D.; Fojo, T.; Rothenberg, M.; Colevas, D., The clinical development of new mitotic inhibitors that stabilize the microtubule. *Anticancer Drugs* 2004, 15, (6), 553-8.
178. ter Haar, E.; Kowalski, R. J.; Hamel, E.; Lin, C. M.; Longley, R. E.; Gunasekera, S. P.; Rosenkranz, H. S.; Day, B. W., Discodermolide, a cytotoxic marine agent that stabilizes microtubules more potently than taxol. *Biochemistry* 1996, 35, (1), 243-50.
179. Minguez, J. M.; Giuliano, K. A.; Balachandran, R.; Madiraju, C.; Curran, D. P.; Day, B. W., Synthesis and high content cell-based profiling of simplified analogues of the microtubule stabilizer (+)-discodermolide. *Mol Cancer Ther* 2002, 1, (14), 1305-13.
180. Martello, L. A.; LaMarche, M. J.; He, L.; Beauchamp, T. J.; Smith, A. B.; Horwitz, S. B., The relationship between Taxol and (+)-discodermolide: synthetic analogs and modeling studies. *Chem Biol* 2001, 8, (9), 843-55.
181. Sanchez-Pedregal, V. M.; Kubicek, K.; Meiler, J.; Lyothier, I.; Paterson, I.; Carlomagno, T., The tubulin-bound conformation of discodermolide derived by NMR studies in solution supports a common pharmacophore model for epothilone and discodermolide. *Angew Chem Int Ed Engl* 2006, 45, (44), 7388-94.
182. Buey, R. M.; Barasoain, I.; Jackson, E.; Meyer, A.; Giannakakou, P.; Paterson, I.; Mooberry, S.; Andreu, J. M.; Diaz, J. F., Microtubule interactions with chemically diverse stabilizing agents: thermodynamics of binding to the paclitaxel site predicts cytotoxicity. *Chem Biol* 2005, 12, (12), 1269-79.
183. Raccor, B. S.; Vogt, A.; Sikorski, R. P.; Madiraju, C.; Balachandran, R.; Montgomery, K.; Shin, Y.; Fukui, Y.; Jung, W. H.; Curran, D. P.; Day, B. W., Cell-Based and Biochemical Structure-Activity Analyses of Analogues of the Microtubule Stabilizer Dictyostatin. *Mol Pharmacol* 2007.
184. Hamel, E.; Day, B. W.; Miller, J. H.; Jung, M. K.; Northcote, P. T.; Ghosh, A. K.; Curran, D. P.; Cushman, M.; Nicolaou, K. C.; Paterson, I.; Sorensen, E. J., Synergistic effects of peloruside A and laulimalide with taxoid site drugs, but not with each other, on tubulin assembly. *Mol Pharmacol* 2006, 70, (5), 1555-64.
185. Pineda, O.; Farras, J.; Maccari, L.; Manetti, F.; Botta, M.; Vilarrasa, J., Computational comparison of microtubule-stabilising agents laulimalide and peloruside with taxol and colchicine. *Bioorg Med Chem Lett* 2004, 14, (19), 4825-9.
186. Lai, Y. Y.; Huang, L. J.; Lee, K. H.; Xiao, Z.; Bastow, K. F.; Yamori, T.; Kuo, S. C., Synthesis and biological relationships of 3',6-substituted 2-phenyl-4-quinolone-3-carboxylic acid derivatives as antimitotic agents. *Bioorg Med Chem* 2005, 13, (1), 265-75.
187. Chen, Z.; Merta, P. J.; Lin, N. H.; Tahir, S. K.; Kovar, P.; Sham, H. L.; Zhang, H., A-432411, a novel indolinone compound that disrupts spindle pole formation and inhibits human cancer cell growth. *Mol Cancer Ther* 2005, 4, (4), 562-8.

188. Reijo, R. A.; Cooper, E. M.; Beagle, G. J.; Huffaker, T. C., Systematic mutational analysis of the yeast beta-tubulin gene. *Mol Biol Cell* 1994, 5, (1), 29-43.
189. Karjala, G.; Chan, Q.; Manzo, E.; Andersen, R. J.; Roberge, M., Ceratamines, structurally simple microtubule-stabilizing antimitotic agents with unusual cellular effects. *Cancer Res* 2005, 65, (8), 3040-3.
190. Campiani, G.; Fiorini, I.; De Filippis, M. P.; Ciani, S. M.; Garofalo, A.; Nacci, V.; Giorgi, G.; Sega, A.; Botta, M.; Chiarini, A.; Budriesi, R.; Bruni, G.; Romeo, M. R.; Manzoni, C.; Mennini, T., Cardiovascular characterization of pyrrolo[2,1-d][1,5]benzothiazepine derivatives binding selectively to the peripheral-type benzodiazepine receptor (PBR): from dual PBR affinity and calcium antagonist activity to novel and selective calcium entry blockers. *J Med Chem* 1996, 39, (15), 2922-38.
191. Braestrup, C.; Squires, R. F., Specific benzodiazepine receptors in rat brain characterized by high-affinity (3H)diazepam binding. *Proc Natl Acad Sci U S A* 1977, 74, (9), 3805-9.
192. Patel, D. S.; Bharatam, P. V., New leads for selective GSK-3 inhibition: pharmacophore mapping and virtual screening studies. *J Comput Aided Mol Des* 2006, 20, (1), 55-66.
193. Kim, M. K.; Min, J.; Choi, B. Y.; Lim, H.; Cho, Y. H.; Lee, C. H., Discovery of cyclin-dependent kinase inhibitor, CR229, using structurebased drug screening. *J Microbiol Biotechnol* 2007, 17, (10), 1712-6.
194. Limongelli, V.; Marinelli, L.; Cosconati, S.; Braun, H. A.; Schmidt, B.; Novellino, E., Ensemble-Docking Approach on BACE-1: Pharmacophore Perception and Guidelines for Drug Design. *ChemMedChem* 2007, 2, (6), 740.
195. Singh, N.; Cheve, G.; Ferguson, D. M.; McCurdy, C. R., A combined ligand-based and target-based drug design approach for G-protein coupled receptors: application to salvinorin A, a selective kappa opioid receptor agonist. *J Comput Aided Mol Des* 2006, 20, (7-8), 471-93.
196. Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T., Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol Divers* 2006, 10, (1), 39-79.
197. Bohacek, R. S.; McMartin, C., Modern computational chemistry and drug discovery: structure generating programs. *Curr Opin Chem Biol* 1997, 1, (2), 157-61.
198. Goodford, P. J., A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985, 28, (7), 849-57.
199. Wade, R. C.; Goodford, P. J., The role of hydrogen-bonds in drug binding. *Prog Clin Biol Res* 1989, 289, 433-44.
200. Wade, R. C.; Clark, K. J.; Goodford, P. J., Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J Med Chem* 1993, 36, (1), 140-7.
201. Miranker, A.; Karplus, M., Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* 1991, 11, (1), 29-34.
202. Silberstein, M.; Dennis, S.; Brown, L.; Kortvelyesi, T.; Clodfelter, K.; Vajda, S., Identification of substrate binding sites in enzymes by computational solvent mapping. *J Mol Biol* 2003, 332, (5), 1095-113.
203. Laurie, A. T.; Jackson, R. M., Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 2005, 21, (9), 1908-16.
204. Bate, P.; Warwicker, J., Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J Mol Biol* 2004, 340, (2), 263-76.
205. Peters, K. P.; Fauck, J.; Frommel, C., The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol* 1996, 256, (1), 201-13.
206. Ho, C. M.; Marshall, G. R., Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J Comput Aided Mol Des* 1990, 4, (4), 337-54.
207. Liang, J.; Edelsbrunner, H.; Woodward, C., Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 1998, 7, (9), 1884-97.
208. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model* 2003, 21, (4), 289-307.
209. Brady, G. P., Jr.; Stouten, P. F., Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 2000, 14, (4), 383-401.
210. Levitt, D. G.; Banaszak, L. J., POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph* 1992, 10, (4), 229-34.
211. Laskowski, R. A., SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995, 13, (5), 323-30, 307-8.
212. Kleywegt, G. J.; Jones, T. A., Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr D Biol Crystallogr* 1994, 50, (Pt 2), 178-85.

213. Del Carpio, C. A.; Takahashi, Y.; Sasaki, S., A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J Mol Graph* 1993, 11, (1), 23-9, 42.
214. Hendlich, M.; Rippmann, F.; Barnickel, G., LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 1997, 15, (6), 359-63, 389.
215. Edelsbrunner, H., Weighted Alpha Shapes. *Technical Paper of the Department of Computer Science of the University of Illinois at Urbana-Champaign;* Urbana, Illinois 61810.
216. Abdel-Rahman, H. M.; Morsy, M. A., Novel benzothiazolyl urea and thiourea derivatives with potential cytotoxic and antimicrobial activities. *J Enzyme Inhib Med Chem* 2007, 22, (1), 57-64.
217. Rapid Overlay of Chemical Structures, http://www.eyesopen.com. developed by Openeye scientific software,.
218. Hawkins, P. C.; Skillman, A. G.; Nicholls, A., Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* 2007, 50, (1), 74-82.
219. OMEGA2, http://www.eyesopen.com. developed by Openeye scientific software,.
220. 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, (6), 1908-19.
221. Fast Exhaustive Rigid Docking, http://www.eyesopen.com. developed by Openeye scientific software,.
222. Yakaiah, T.; Lingaiah, B. P.; Narsaiah, B.; Shireesha, B.; Ashok Kumar, B.; Gururaj, S.; Parthasarathy, T.; Sridhar, B., Synthesis and structure-activity relationships of novel pyrimido[1,2-b]indazoles as potential anticancer agents against A-549 cell lines. *Bioorg Med Chem Lett* 2007, 17, (12), 3445-53.
223. 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), 337-44.
224. 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.
225. Labute, P., The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J Comput Chem* 2008.
226. ACD/ChemSketch, Software developed by Advance Chemistry Development. *www.acdlabs.com* ACD/Labs Release 10, Product version 10.02 (Build 14890, 09 Nov 2006).
227. FRED (version 2.0.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com).*
228. Schulz-Gasch, T.; Stahl, M., Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J Mol Model (Online)* 2003, 9, (1), 47-57.
229. Labute, P., Probabilistic Receptor Potentials. *Journal of the Chemical Computing Group* 2001, http://www.chemcomp.com/journal/cstat.htm.
230. Halgren, T. A., The Merck Force Field. *J. Comp. Chem.* 1996, 17, 490-512, 520-552, 553-586, 587-615, 616-641.
231. Grant, J. A., Pickup, B.T., Nicholls, A,, A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J. Comput. Chem.* 2001, 22(6), 608-640.
232. Labute, P., Atomic Contact Radii in the Protein Data Bank. *Not published* 2000.

[a] All pictures were generated using MOE from CCG, http://www.chemcomp.com/

[+] In the figures marked, geometries are not appropriate, in particular for the phenol (Tyr) geometries. The causes are being investigated thus the conclusion about energies of binding should be taken with a pinch of salt.

Chapter V:

Coupling High Content Screening to Virtual Screening

Introduction

Objective

The present Chapter describes the *in vitro* biological work that was performed on a series of PBOX compounds in an attempt to define a Structure Activity Relationship (SAR) for this group of compounds. This SAR data was then exploited to guide the modelling work previously presented in Chapter 3. In addition, the Virtual Screening (VS) work carried out, in chapter 3 produced numerous potential "hits" (denoted MDG compounds). For validation purposes, it was necessary for the new MDG compounds to be tested *in vitro* and their potencies evaluated. The use of High Throughput Screening (HTS) methods applied to cell testing *i.e.* High Content Screening method (HCS) appeared the most relevant choice of technique for analysing both PBOX compounds and the new MDG compounds. On the one hand, the assays presented in this chapter, aimed to evaluate the cell cycle effects of all compounds on cancer cell lines, and on the other hand, to determine and/or quantify MDG potential effects on the microtubule (MTs) network. As MDG compounds were expected to induce similar effects to PBOX compounds and in order to develop a suitable HTS assay, it was essential to better understand the molecules mechanisms underlying the apoptosis induced by PBOX compounds.

PBOX effects

Pyrrolo-1,5-benzoxazepine (PBOX) have been suggested as potential novel antineoplastic due to their ability to induce programmed cell death (apoptosis) and to elicit antiproliferative effects on cancerous cell lines. Indeed, some members of this series appeared to potently induce apoptosis in diverse cancer cell lines including breast cancer cell lines and drug resistant leukaemia cell lines. [1-3]After drug treatment, specific apoptotic characteristics were observed[4, 5] such as cell shrinkage, chromatin condensation, DNA fragmentation, and poly(ADP-ribose) polymerase (PARP) cleavage. Recently it has been reported that PBOX compounds can be devided into two groups with distinct mechanisms of action; those that exert antiproliferative effects by arresting cancerous cell in the G1 phase of the cell cycle without eliciting cytotoxic effects *e.g* PBOX21 and those that exhibit pro-apoptotic ability by binding to tubulin and acting as microtubule targeting agents (MTAs) *e.g.* PBOX6. This second group of MTAs have been shown to arrest sells in the G2/M phase of the cell cycle and then to potently induce apoptosis in a wide range of solid tumours and cell lines derived from the hematopoietic system and also *in vivo*[1, 5, 6]. Further investigations on their anti-proliferative effect and on the apoptotic mechanism of action involved, revealed cell cycle incidences and no cytotoxicity[1]. The analysis carried out through Fluorescence Activated Cell Sorting (FACS), upon numerous cancer cell lines, at

nanomolar and micromolar concentrations, showed two different arrests in specific phases of the cell cycle inducing apoptosis in different pathways[2, 7]. These observations distinguish therefore PBOX compounds into two subsets which will be referred to as 'G1 group' for compounds inducing a G1 arrest and 'G2 group' for compounds inducing G2/M arrest. Finally, tubulin was confirmed as one of the molecular targeted proteins of the PBOX compounds leading to apoptotic cell death[8]. Therefore, the detection of a G2 peak using HCS would be as suitable assay to track as it is common to all PBOX actives which target tubulin. A study of the biology associated with the PBOX compounds at the cellular level was essential for the successful realisation of a large scale assay aiming to measure changes in cell cycle profiles.

5.1 Biological background

5.1.1. The cell cycle

Precise coordination of the progression through cell cycle phases is critical not only for normal cell division but also for effective growth arrest under conditions of stress or after DNA damage. As a consequence, a disruption in the cell cycle machinery may contribute to uncontrolled cell growth, which is the principal feature of cancer. Conversly, disturbing the cell cycle of cancer cells could initiate cell death through an apoptotic process.

The cell cycle can be considered as a collection of highly ordered processes that results in the duplication of the cell, thus underlying the growth and development of all living organisms. The molecular basis of this control is highly conserved from simple unicellular eukaryotes such as yeast, to complex metazoans such as humans. The precision with which cell cycle events are executed ensures the survival of living organisms, while loss of this precision increases genomic instability, an important factor in the formation of cancer. Figure 5.1 illustrates the eukaryotic cell cycle which consists of four distinct phases:

G1 phase, the first gap or growth phase and the phase in which the cell spends the majority of its time;
S phase, the period during which the cell synthesises or duplicates its DNA content;
G2 phase, the second gap or growth phase;
M phase, the cell undergoes mitosis/meiosis and divides into two daughter cells;

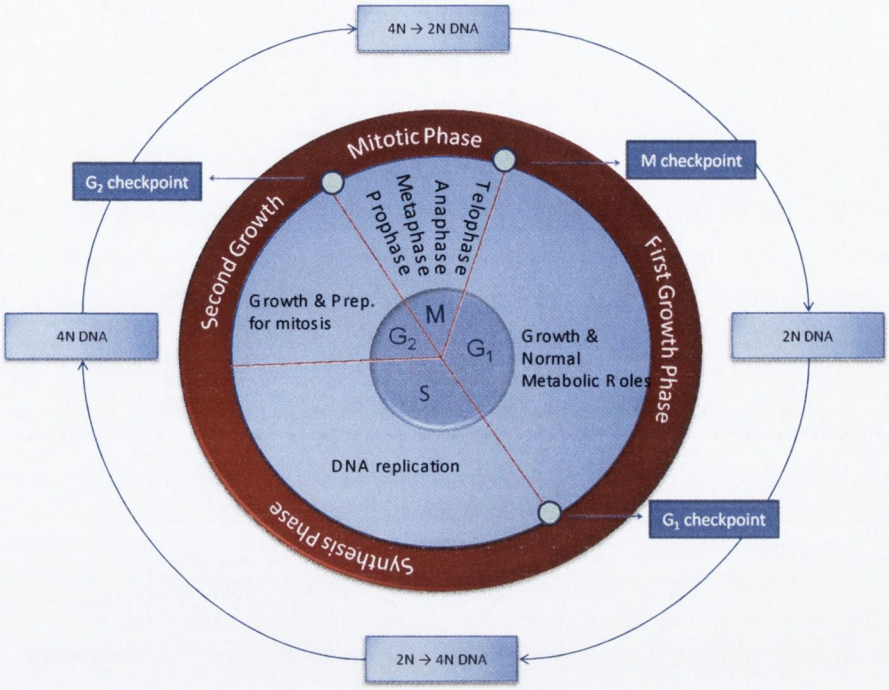


Figure 5.1: The Phases of the Cell Cycle

During mitosis, genetic information (DNA) is duplicated and distributed evenly between two daughter cells. The bipolar spindle is formed between two microtubule organising centres during the prophase and the condensed sister chromatids are attached to the spindle via their kinetochores in pro-metaphase and metaphase to finally be pulled to opposite poles during anaphase (see Figure 4.1 Phases of Mitosis).

The succession of events in mitosis is controlled by three checkpoints[9-11] among which the spindle assembly checkpoint controls the metaphase-anaphase transition. This checkpoint prevents cells from entering anaphase by inhibiting the separation of sister chromatids until all kinetochores are attached to microtubules[10, 12, 13] and chromosomes properly aligned. If chromosomes are not properly oriented on the spindle and are not arranged in the metaphase network shape, eukaryotic cells arrest[14].

Microtubules are the key components of the cytoskeleton and were largely presented and discussed in Chapter 4. They are highly dynamic and adjust their network shape according to the cell cycle phase. Perturbing the MTs network involves alteration of the mitotic checkpoint leading to the failure of mitotic spindle formation and thus failure of spindle attachment to the kinetochores of chromosomes engaging the spindle assembly checkpoint[15].

As Cancer is recognised as a disease of the cell cycle, antimitotic drugs including antitubulin drugs generate a tremendous interest in oncology research. Indeed, as cancer disease can translate by an

uncontrolled and amplified cell growth, targeting signalling pathways checkpoints aim to steer cancer cells on to a programmed death process which they had initially avoided. Mitotic arrest induced by antimitotic agents is associated with an upregulation disfunctionality which prevents exit from mitosis[16]. The majority of clinically effective anticancer drugs, inhibit some aspect of the machinery responsible for DNA replication and chromosome segregation. This gives rise to cell cycle arrest via induction of specific cell cycle checkpoints. These checkpoints ensure that subsequent cell cycle events are inhibited until the inflicted damage is repaired. The fidelity of checkpoint controls, and susceptibility of cells to die through initiation of apoptotic pathways[17] while repair is underway, may be important factors in the success of chemotherapy[18].

5.1.2 Apoptosis

The term 'apoptosis' is derived from the Greek word for 'falling off', a reference to leaves falling off a tree in autumn in response to the impending threat of freezing and damage in winter. An understanding of apoptosis finds its roots, when Weigert and Cohnheim described the microscopic appearance of cell death in necrotic tissue as coagulation necrosis. Later, Kerr described the electron microscopic appearance of single cell death in the livers of animals treated with toxins and ischaemia and called it 'shrinkage necrosis'. Subsequently, Wyllie[19] suggested a common pathway for cell death initiated by hormone regulation and carcinogen-induced injury. Later, Kerr *et al.* in 1972[20] described the characteristic sequential changes occurring in cell structure during the death process in healthy tissue, normal development, tumour regression and atrophy. Finally, the specific cell death during the development of the nematode was reported by Horvitz with the cloning of the death genes (ced genes) responsible. The nematode anti-apoptotic gene ced-9 shared homology with the mammalian proto-oncogene, Bcl-2 and was shown to preserve B lymphocytes.

Apoptosis specifically refers to a genetically controlled and energy dependant process by which unnecessary or damaged single cells self destruct when the apoptosis genes are activated. Prominent features of necrosis are swelling of cells, disruption of membranes, and lysis of the nuclear chromatin. Because cellular contents are lost into the extra cellular space, necrotic tissue evokes an inflammatory response. Apoptosis is distinguished from necrosis by a characteristic set of features including membrane blebbing, shrinkage of the cell, nuclear fragmentation and chromatin condensation.

There are three distinct phases of apoptosis. During the first phase, the cell shrinks and detaches from its substratum and adjacent cells with a loss of microvilli and junctional complexes or desmosomes. Later, the nuclear membrane becomes convoluted, whilst the nucleus becomes enlarged and abnormally granular. Chromatin condenses (pyknotsic chromatin) and forms aggregates near the nuclear membrane while endonucleases cleave the DNA into fragments or strand breakage

(karyorhexis). Whereas organelles are preserved almost intact, the endoplasmic reticulum and mitochondria swell and exocytoses their contents. The cell becomes denser as the cytoplasm shrinks and involutes. In the second phase, several cell surface molecules change to ensure that apoptotic cells will be immediately recognised and engulfed by neighbouring cells or phagocytes leading to little or no inflammation. The cell produces pseudopodia (budding) which contain organelles or nuclear fragments, and these break off into multiple membrane-bound vesicles. These membrane-bound vesicles are ingested by surrounding cells. The remaining cell becomes a round, smooth membrane-bound remnant (apoptotic body). In the third phase, the cell membrane becomes permeable to dyes such as Trypan Blue. The apoptotic body and membrane-bound buds may then be phagocytosed by macrophages, epithelial cells, vascular endothelium or tumour cells. The entire process may take only 15 min, and therefore may be undetectable on tissue sections.

During the last few decades the mechanisms involved in disordered apoptosis were revealed, suggesting that alterations in control of cell death or survival is implicated in pathogenesis of a variety of human diseases including cancer and many other chronic diseases. Studies in both normal and transformed human cells, treated with members of either the polymerising or depolymerising groups of anti-microtubule inhibitors, have shown that apoptosis can be initiated rapidly and directly from mitosis. In light of this, intensive research into the discovery of novel anti-mitotic compounds have been carried out leading to the discovery of new anti-cancer drugs largely presented in Chapter 4. The great advances made in the HTS and HCS fields, open new opportunities for the rapid and efficient screening of a vast library of compounds to assess their potential bioactivities such as apoptosis, cell cycle modulations or MTs alterations which are presented in this work.

5.1.3 High-throughput fluorescence microscopy

In the recent past few years, fluorescence microscopy for large scale assay has been enhanced and attracts more and more interest due to the constant advances in HCS technology development. Several platforms for HCS and High Content images Analysis (HCA) are nowadays available offering diverse scanning methods and data management software. Most of these platforms provide multi-channel detection enabling several spectrally different fluorophore recognitions simultaneously. They can be classified into two types depending on their focusing mechanism. The first type consists of image-based autofocus routines where a number of images along the z axis are analysed to adjust the next autofocus iteration. This type provides high quality images overcoming the variability observed within cells attached, for example, to a culture dish. The second type is hardware autofocus based where the bottom of the cell, *i.e.* edge of the plate, is measured and the focus height stays constant during the full

scanning process. This type provides high speed reading and therefore can avoid autobleaching effects due to overexposure to light.

Live cell assays are now made possible thanks to environmental control parameters such as temperature or humidity CO2 levels. Nonetheless, assays using fixed cells remains quicker and more informative as the data read-out is straightforward. Novel HCS technology also enables quantitative tracking of changes in the localisation, texture and shape at both cellular and sub-cellular levels. Thus, changes consequent to drug treatment can be automatically processed.

For instance, in the context of toxicology assessment, HCS has been proved to efficiently detect and measure diverse phenotypes such as apoptotic or necrotic cells as a result of drug exposure[21]. Using cultured mammalian cells, it was possible to successfully cluster diverse phenotypes according to their nuclei states as DNA damage and formation of micronuclei enables genotoxicity assessment[22].

Furthermore, HCS was used in the context of novel molecular target discovery, where fluorescence microscopy enabled the determination of new pathways. For instance, downregulation of a specific tumour suppressor[23] by RNAi was shown to cause an asymmetric microtubule distribution[24] . Using combinatorial screens, this phenotype could be overcome and new drugs developed[25, 26].

5.2 Strategy for screening of PBOX and MDG compounds

Screening of 47 PBOX and 50 MDG compounds was required. To accomplish this, a suitable novel reliable HCS method of screening was developed.

Choice of assay endpoint

To determine the apoptotic potencies of all tested compounds, it was decided to measure their effect upon the cell cycle. Despite the need to observe a G2 peak of cells when a PBOX active is tested, this observation on its own is not sufficient to identify subsequent apoptotic effect. Apoptosis involves cell death which would result in the decrease in the total number of cells. In addition, these nuclei of these cells would present specific nuclear changes specific to. Therefore, the antiproliferative effect of the compounds combined with a morphological analysis of cell nuclei was carried out in conjunction with the cell cycle analysis.

Traditionally FACS assays have been used for cell cycle analysis. In the present study in order to extrapolate cell cycle analysis to a compound library screen, an HCS approach was alternatively chosen. Cell cycle analysis by HCS presents the advantages cited below compared to traditional FACS analysis:

- To save time by providing more information per screening run.
- Higher order information provides more reliable insight into drug effects.
- Multiplexed assays are cost-effective

In the last few years, HCS novel technology has been used to develop multiplex high throughput assays such as cell proliferation, toxicology, cell cycle analysis and cell morphology studies[8, 27, 28]. Therefore, HCS appeared as a challenging novel technology in the application of cellular morphological change tracking. Following drug treatment, the cells' morphological and physiological changes were tracked at the cellular and sub-cellular level; qualitative and quantitative modulation was determinated for whole cells shapes, nuclei and MT network.

The percentage of cells observed in the G2 arrested phase was determined and IC50 values estimated. The IC50 value is the most common parameter used to classify and rank compounds' activities. It consisted of the concentration of drug needed to obtain 50% of the maximum measured effect. There are some variants of this parameter among which ones an EC50 value utilised for agonist drugs' testing. According to the Food and Drug Administration (FDA), IC50 represents the concentration of a drug that is required for 50% inhibition *in vitro*[29]. EC50 also represents the plasma concentration required for obtaining 50% of a maximum effect *in vivo*[30]. Several other variants of this parameter exist where their differences generally find their basis in the addition of correcting values as for the GI50 number used by National Cancer Institute (NCI) and discussed in Chapter 2.

Screening layout

Figure 5.3 illustrates the layout of the procedurs followed for all the screens. This layout can be split into four distinct stages corresponding to the group of compounds screened:

- Stage 1: PBOX compound screen. PBOX were screened for the immediate and rapid assessment of their potencies. The percent of cells arrested in G2/M phase at diverse concentrations of drug was measured for the determination of IC50 values.
- Stage 2: MDGs primary screen. MDG new potential hits were first tested at two concentrations and their effect on the different phases of the cell cycle examined. This enabled the filtering out of MDG inactive compounds.
- Stage 3: MDG actives secondary screen. The secondary screen consisted of a measurement of the percent of cells arrested in G2/M phase at varying concentrations of drug for the determination of IC50 values.
- Stage 4: MDG actives compounds tertiary screen. MDGs were further selected for modelling studies with the expectation they would have a similar effect and follow the same mode of

action of PBOX compounds. As a consequence, MDGs active compounds were selected and screened for their effect on the MT network.

Set up and validations of assays

PBOX compounds have been suggested as potential anticancer agents and in particular in the treatment of drug resistant leukaemia. In addition, HCA technology has been shown to be most accurate when using fixed adherent cells. leukaemia cell lines grow in suspension but PBOX activities have also been successfully detected on breast cancer cell lines which was thought to be more suitable for the actual HCS assay undertaken. Therefore, all screens were performed using the MCF-7 breast cancer cell line.

In order to compare the potency of the effect of the PBOX compounds and any novel compounds, positive controls were essential. To validate the IC50 evaluation protocol, nocodazole was chosen as a positive control. Nocodazole elicits an antiproliferative effect on cells by binding to tubulin at the colchicine binding site and by inducing arrest at the G2/M phase of the cell cycle (*c.f.* chapter 4). The IC50 value for Nocodazole based on the percentage of cells arrested in G2 phase, was evaluated during the validation step.

Furthermore, other microtubule targeting drugs such as taxol and combretastatin A-4 were additionally utilized as positive controls for the stage 4 screen. Taxol is a stabilizing agent whilst nocodazole, PBOX6 and combrestatine A-4 are destabilizing agents. These drugs are representative of both the detabilising and stabilising group, and were used to evaluate modulation of morphological features of the MT network according to their particular properties. The effect of the MDG drugs on the MT network was compared to the positive and negative controls (untreated cells) in order to estimate into which group they would fall: stabiliser, destabiliser or inactives.

Finally, to fully validate the whole process, a last validation step was undertaken after stage 4. The most active compounds determined were further tested on target using a tubulin polymerisation assay. As this particular assay is highly expensive and highly variable, only one concentration of the drug was tested.

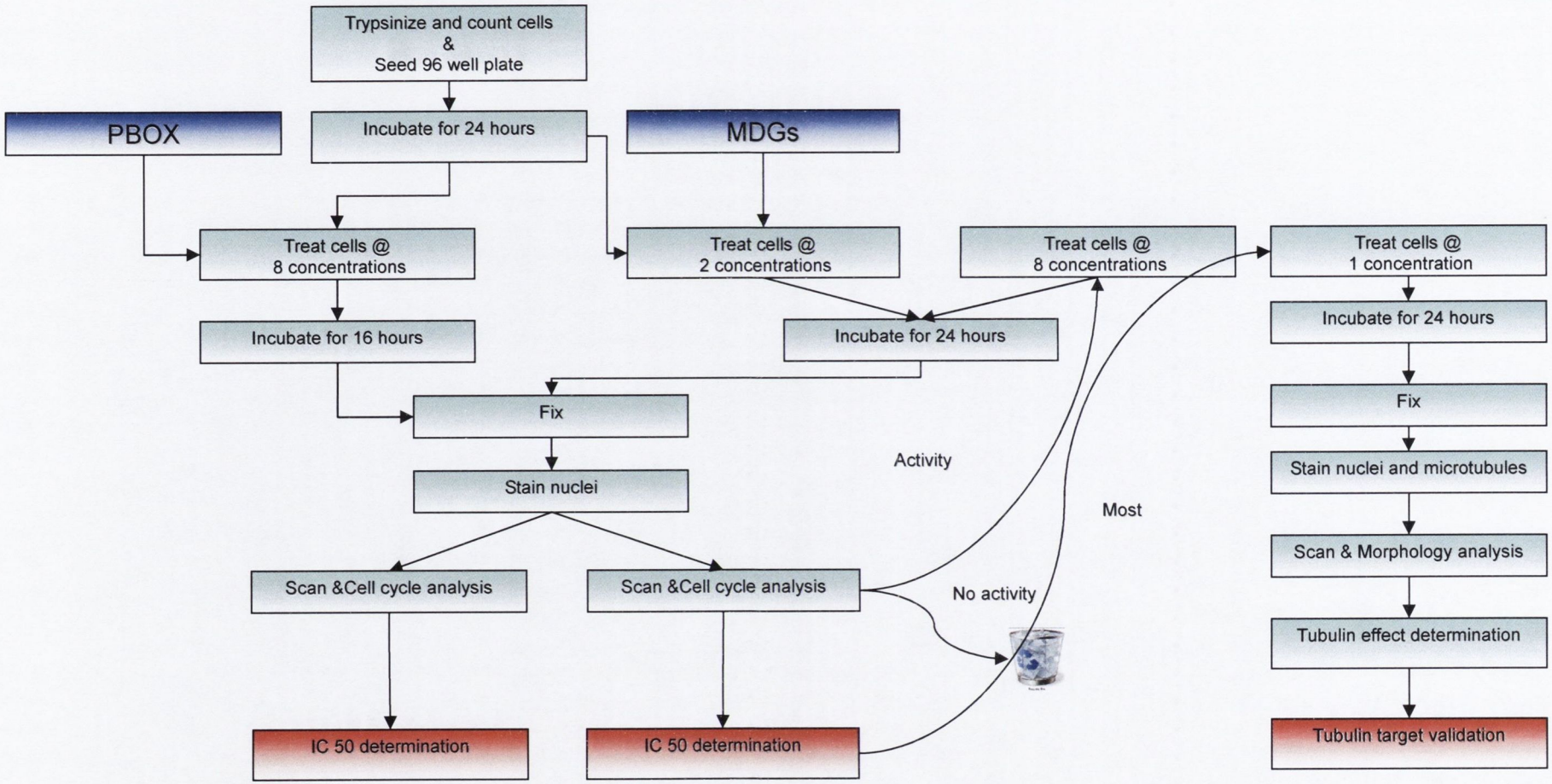


Figure 5.1: Physical screen general workflow.

5.3 Material and methods

5.3.1 Materials

Table 5.1 lists the material used for this experimental work:

MATERIAL	SUPPLIER
Anti-β-Tubulin–Cy3 solution	Sigma
BSA	Sigma
Ethanol	Chemical Store,TCD
FCS	Sigma
Formaldehyde solution	Sigma
Gentamicin	Sigma
Glutamine	Gibco
GTP	Cytoskeleton
Hoechst 33342 fluorochrome	Sigma
MCF-7 cell line	E.C.A.C.C
MEM	Sigma
Nocodazole	Sigma
Non-Essential Amino Acid	Sigma
Paclitaxel	Sigma
PBS tablets	Sigma
Pipettes, sterile (10, 25 mL)	Sterilin
Tubulin Protein	Cytoskeleton
Tissue culture flasks	Greiner
Triton X-100	Sigma
Trypsin	Sigma
Tubulin	Cytoskeleton
Tween20	Sigma
96 Half Area plates	Corning Costar Inc.

Table 5.1: Materials

Addresses of Suppliers:

Cytoskeleton Inc., 1830 S. Acoma St., Denver, CO 80223, USA.

European Collection of Animal Cell Cultures (E.C.A.C.C.), PHLS Centre for Applied Micorbiology and Research, Porton Down, Salisbury SP40JG, UK.

Gibco, Life Technologies Ltd., 3 Fountain Drive, Inchinnan Business Park, Paisley PA4 9RF, UK.

Greiner, GmbH, Maybachstrasse 2, P.O. Box 1162, D-7443 Frickenhause, Germany

Sigma, Chemical Co., Ltd., Fancy Road, Poole, Dorset, UK.

Sterlin, Bibby Sterlin Ltd., Stone, Staffordshire, UK.

5.3.2. HCS theory

The kinetic scan reader (KSR)[31] uses image and hardware based autofocusing methods. It is a wide-field microscope enabling 96 well plate format, suitable for screening large libraries of compounds. The KSR is an automated solution for analysis of cellular and intracellular spatial parameters, over time and can manage living cells. The reader takes high resolution fluorescence images of dye-labelled cells that are located on the bottom of the microplate wells. The images or "fields" can be analyzed by means of image processing software among with C*ellomics* associated software denoted "Bioapplications". After data mining the result of the assay in each well can be determined. The full process involves firstly acquisition of images and then data treatment. The assay developed for the present work consisted of dual cell dye-labelling using two channels. Both nuclei and tubulin were marked using two different dyes with non-overlapping wave-length emission bands (respectively in blue and red). Cells were stained for the two targets, DNA (nucleus) and tubulin (microtubules from the cytoskeleton), either consecutively or simultaneously.

Once the KSR has scanned plates and stored the fields of each well, cell recognition as "objects" is obtained using diverse available tuneable algorithms. *Cellomics* software offers several algorithms depending on the biology to read *i.e* the type of object to be described. A stained more globular nucleus or stained microtubule fibres from the cytoskeleton, obviously differ from their forms, shapes, intensity ranges etc. Therefore, the selection in an image of the object, be it a nucleus or the cytoskeleton, requires the use of a specific algorithm suitable for the particular object. Once an object is defined and selected, a large amount of data can be obtained. Figure 5.4 illustrates the selection of labelled objects *i.e.* nuclei and fibres at the field level.

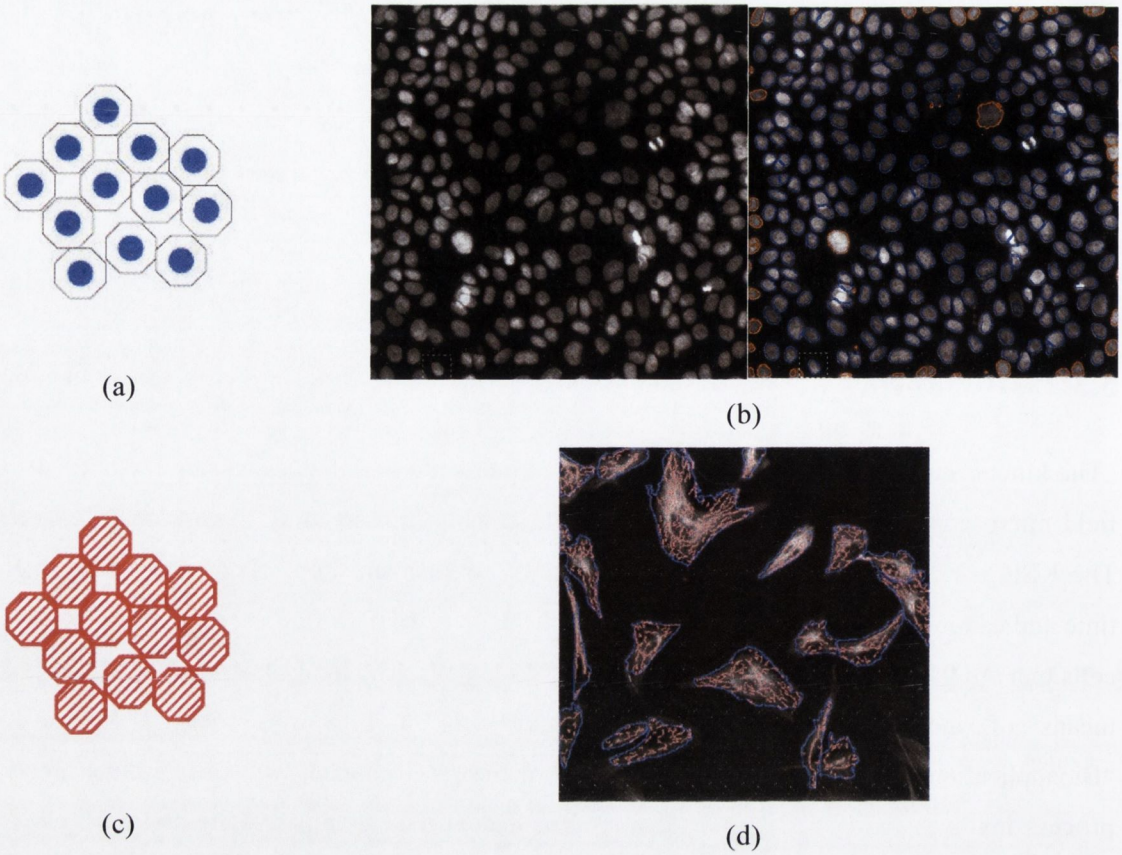


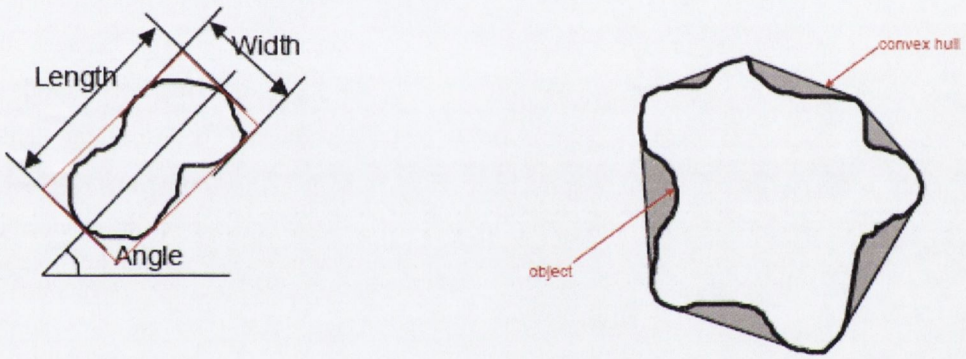
Figure 5.3: Principle of labelled objects selection
(a) nuclei stained scheme, (b) algorithm for nuclei selection, surrounded by blue lines are the retained objects (nuclei), and by orange lines the rejected, image acquired using 10X objective (c) tubulin (from the cytoskeleton fibres) stained scheme, (d) algorithm for fibres selection, surrounded by blue lines are the retained objects (cytoskeleton) and in pink the detected fibres, image acquired using 20X objective

Figure 5.4 (a&c) are simple schemes of stained cell nuclei and tubulin and roughly illustrates in blue the nuclei and in red microtubules. Figure 5.4 (b) shows the results after running the algorithm, in blue are the selected nuclei and in orange the rejected ones. Objects to be rejected from the following data analysis can be multiple nuclei superimposed, sections of nuclei that happened to be at the image edges or dirt. Figure 5.4 (d) illustrates in pink the microtubule fibres selected and in blue the surrounding of the objects selected. The cytoskeleton with non-detected associated nuclei or cut in section due to their position at the edge of the image should be set to be rejected prior to data analysis. In this figure, for pedagogic purposes, the two images shown are not of the same magnification. In practice, only one magnification is used for a simultaneous multi-channel image acquisition. In general, the camera focus is set on Channel 1 which reads the nuclei stain. The two separated channel images are then superimposed giving what is called the composite image. In the Channel 1 image, the nuclei

coordinates determined the object references. These coordinates are used to associate a cell nucleus with its own cytoskeleton from another channel image. To determine how to select, keep or reject an object, several tuneable parameters are available in all bioapplications among which are:

- Background correction: refers to the radius of the area that is sampled around each pixel in the image to determine that pixel's local background.
- Object Identification method: influences the intensities measured for the individual nuclei, affects how much of an identified object's intensity is valid for measurement.
- Intensity ranges: for instance to discard objects which are too bright.
- Morphological ranges: for instance to discard objects with too small or too big areas.
- Segmentation parameter: helps resolve and identify individual nuclei in clumps.

At the well-level, the number, density and location (by the x and y coordinates of its centroid) of nuclei and cytoskeleton are reported. Shape parameters such as the object's roundness, aspect ratio, and dimensions such as its length and width are also reported. Some dimensions are defined by a rectangular bounding box which encompasses the object (see Figure 5.5 a).



(a) (b)

Figure 5.4: (a) Bounding Box, (b) convex hull

The bounding box is the rectangle of minimum area that includes the object. Additional shape and dimension output features are related to an object's convex hull. The convex hull is defined as the smallest convex set containing the points of the original object. The easiest way to think of it is the shape defined by a rubber band placed around the object (Figure 5.5 b). These represent the basic features and more features can be calculated depending on the biology as illustrated in Table 5.2.

Channel	OUTPUT FEATURES
Channel 1, whole object morphology For instance nuclei stained which can be used for cell cycle analysis	• Object counts & field density, • Object location, area, shape, length, breadth, orientation, extent, perimeter. • Object 3-D, metrics • Processes extending from object • Ch 1 label intra-object amount, texture & distribution • Object spacing & proximity
Channel 2, Intra-object compartment properties This could be a cytoplasmic stain or mitochondrial membrane stain etc.	• Ch 2 label amount inside object • Degree of multi-cell assemblage • Location of cells in cell cluster • Intracellular compartment intensities, shape & morphology • Intracellular compartment whole object comparisons
Channel 3 , Intra-object identifiers or fibres For instance, cytoskeleton stain for microtubule network variation detection.	• Ch 3 label amount, location inside object • Spacing & proximity between similar objects • Fibre arrangement inside object • Ch 3 label intensity distribution statistics inside object • Ch 3 label texture & intensity arrangement measurements inside object

Table 5.2: Output features categories

These features in association with biological knowledge permits the description of cell behaviour in each well. Hence, channel 1 can be used to described the effect of drugs upon the cell cycle while channel 3 their effect on the microtubule network.

Cell cycle bioapplication

In the approach taken by the KSR as well as in FACS, a cell's DNA is labelled with a fluorescent dye with intensity proportional to the cell's DNA content. This bioapplication uses only one channel, channel 1. There are three categories of quantitative measurements reported for labelled nuclei:

- Intensity: Used to measure DNA content & identify the cell's cell cycle phase
- Morphology & Location: can be used to identify damaged, apoptotic or necrotic cells

- Cell Number & Density: can be used to identify conditions that stop cell proliferation due to a cell cycle block

To describe the cell cycle variation per well, specific features can be outputed by an appropriate software amongst which is the cell cycle bioapplication output presented in Table 5.3.

Feature category	Cell level feature	Corresponding well-level features
Intensity	Total Nuclear Intensity	Mean, SD, SE, CV
		Nuclear Total Intensity Peak
	Total Nuclear Intensity	Number and % of cells with DNA content in 5 categories: <2N, 2N, 2N-4N, 4N, >4N
		Ratio of cells in 2N/4N (ratio of cells in G1/G2)
	Average Nuclear Intensity	
Morphology & Location	Nucleus Size & Status	Mean, SD, SE, CV, %High, %Low
	Top, Left, Width, Height, X Centroid, Y Centroid	
	Nucleus Area, Nucleus Shape, Nucleus Shape LWR	
Cell Counts & Density	Cell Number	Valid Cell Count, Selected Cell Count, % Selected Cell Count, Valid Field Count, Cell per Field Count

Table 5. 3: *Cellomics* bioapplication cell cycle features

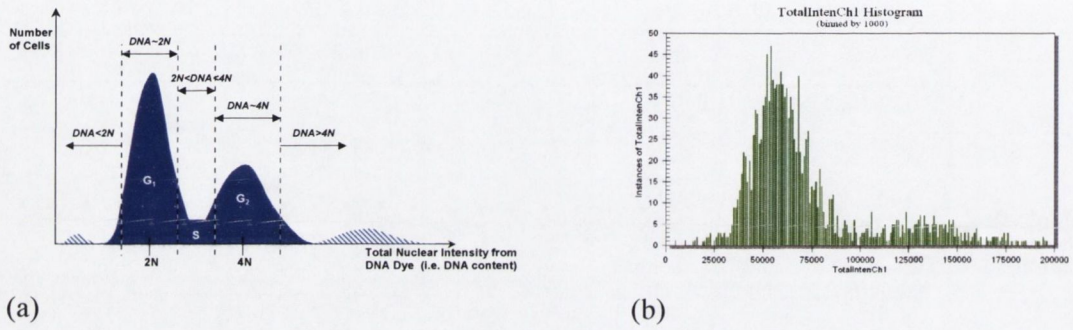
N: one quantity of DNA, SD standard Deviation, SE: standard error, CV: covariance, %High and %Low: respectively percentage of cells whose nuclei intensities are high or low with regard to the user thresholds settings.

Using the Total Nuclear Intensity, cells are binned into 5 groups:

- group 1 cells with a DNA quantity < 2N (pre-G1 phase)
- group 2 cells with a DNA quantity = 2N that (G1 phase)
- group 3 cells with a DNA quantity between 2N and 4N (S phase)
- group 4 cells with a DNA quantity = 4N (G2 phase)
- group 5 cells with a DNA quantity > 4N (cells with more than two times DNA).

Figure 5.5 (a) illustrates the distribution of cell nuclei for a theoretical healthy profile whilst Figure 5.5 (b) is an example of the output obtained experimentally. In practice, the G1 peak is firstly detected and used to detect the second G2 peak therefore describing the entire cell cycle profile. Three main parameters must be set:

- 2N peak intensity: usually set at the average intensity value of the first G1 peak read on untreated cells (*i.e.* negative control). Default is set on the total average intensity of cells.
- 4N peak intensity factor : ideally should be 2 as the G2 peak in theory should contain exactly twice the amount of DNA with respect to the one DNA cells classified in the G1 peak bins. Default value 2.
- The peak width fraction factor: d/L where d is the width of the 2N peak and L the distance from the median of the 2N peak to the median of 4N peak Default value 0.25.
- Minimum number of bins: this is the minimum total number of bins used to plot the cell cycle histogram as shown in Figure 5.6 (b). Default value 10.



(a) (b)

Figure 5. 5: Cell cycle healthy profiles

(a) cell cycle healthy theoretical profile, (b) cell cycle healthy experimental profile

To set the bioapplication parameters correctly, firstly, the untreated wells, or negative controls and positive controls, or wells treated with drugs of known effect, are manually scanned. The analysis of the raw histograms obtained for the control wells per plate determined the most advantageous parameter values needed to obtain the optimum separation between untreated and responding groups. These parameters are then kept unchanged for the whole plate scan.

Morphology explorer bioapplication

Under normal physiological conditions, cell morphology changes can occur such as the entire cell shape or area and the spacing or proximity between different cells. For instance, during mitosis, intracellular location of the nucleus, arrangement and structure of the cytoskeleton varies. Moreover, these changes may also be induced after drug treatment and may have physiological consequences. Quantification of morphological changes over different dimensional scales is a powerful, information-rich method of evaluating compounds or cell responses. The morphology explorer bioapplication software (developed by *Cellomics*) aims to track cellular morphological changes. Morphology explorer was used for the analysis of three channels and the output features obtained are listed in Table 5.4.

Channel 1	
Feature category	**Cell level feature**
Object Counts & Density	Object Number
Object Location	Top, Left, Width, Height, X Centroid, Y Centroid
Object Area & Shape	Area, Shape Factor, Aspect Ratio, Bounding Box Area Fill Ratio, Convex Hull Area Ratio
Object Length & Breadth	Actual Length, Length & Width based on Bounding Box, Breadth
Object Orientation	Angle
Object Perimeter & Extent	Perimeter, Equivalent diameter, and ellipse major/minor diameter ratio, Convex Hull perimeter ratio
Object 3-D Metrics	Equivalent sphere, oblate, prolate volumes, Equivalent sphere surface area
Processes Extending from Object	Number and length of processes (Total, average, length of longest of all processes)
Amount inside object & intra-object arrangement, distribution and texture	Total, average and standard deviation intensity inside object, skewness, kurtosis, intensity entropy and difference surface area density inside object.
Object spacing & proximity	Distance of closest object, Average distance of object, Distance variation (SD) of objects
Channel 2	
Feature category	**Cell level feature**
Amount inside cell	Total and average Channel 2 intensity inside object
Inter/Multi-Cellular Mode:	
Degree of multi-cell assemblage	Number of members (i.e., cells) in object (object = either single cell or a multi-cell assemblage)
Location of cells in multi-cell assemblage	Number of exterior cells in colony, Number of interior cells in colony
Feature category	Cell level feature
Intra/Sub-Cellular Mode:	
Intracellular Compartment's Shape	Shape Factor (P2A), Aspect Ratio (LWR), Box Fill Ratio (BFR)
Intracellular Compartment's Morphology	Area Equivalent diameter, Equivalent ellipse major/minor diameter ratio, Convex Hull area ratio, Convex Hull perimeter ratio
Intracellular Compartment's Intensity	Total intensity per compartment, Average pixel intensity per compartment
Intracellular Compartment- Whole Cell Comparisons	Compartment/Whole Cell Area Ratio, Compartment/Whole Cell Area Difference

Channel 3	
Inter/Multi-Cellular Mode:	
Spacing & proximity within specified area between similar object types (i.e., between objects having either Ch3 or Ch4 label)	Distance of closest similar cell Average distance of similar cells Distance variation (SD) of similar cells
Intra/Sub-Cellular Mode:	
Amount inside cell	Total and average intensity inside cell
Labelled macromolecule's location inside cell (radial intensity distribution statistics)	Mean, standard deviation, skew, kurtosis of radial intensity distribution
Labelled macromolecule's arrangement inside cell	Intensity distribution statistics (Intensity standard deviation, skewness, kurtosis, entropy inside cell and intensity difference surface area density)
Labelled macromolecule's arrangement inside cell	Texture & intensity arrangement measurements from intensity co-occurrence matrix (Maximum probability, contrast entropy, angular Second moment (uniformity) of 2D distribution of intensity co-occurrences of 2D)
Morphology of discrete objects (i.e., spots or fibres)	Spot or fibre, number and areas(Number, total area, average area of spots) Fibre arrangement (Standard deviation of fibre alignment angles, Fibre alignment autocorrelation)

Table 5.4: Morphology explorer bioapplication output features

All the features listed in Table 5.4 are calculated for negative and positive controls. Furthermore, analogously to the previous section, several tuneable parameters must be set of which the most important are:

Channel 1:

- SmoothFactorCh1: Radius of area used to smooth image before the object identification. This is a pixel-based factor used to include an area surrounding the object and therefore it is a graphical approach as opposed to the intensity values factor available for object segmentation as the watershed factor.
- Watershed Factor: Factor applied to intensity threshold to identify object peaks, which are then used to identify objects using a watershed approach. A way to contextualize this method is to imagine the separation between two mountains (two peacks of intensity) by a lake. The lake

level will represent the intensity bottom threshold. This is another way of correcting background noise in the aim of separating two different objects if there are too close.

Channel 2:

- Smooth Factor: Radius of area used to smooth image before member identification.
- Member Segmentation: corresponds to a radius of an area. Positive values are the size of a member segmentation is applied to. Negative values invoke a peak detection approach where the algorithm searches for a peak in an area whose radius corresponds to the parameter value.

Channel 3:

- Spot Fibre Size: Radius of area used to identify spots or fibres.
- Fibre Align2Analysis: Perform fibre alignment analysis from anisotropy of image autocorrelation.

The bioapplication parameters are determined by firstly scanning the untreated wells. These parameters are then kept unchanged for the whole plate scan. As opposed to the cell cycle bioapplication which tracks only few nuclar changes, the explorer bioapplication tracks a multitude of changes which can be linked to several effects upon MTs such as stability, instability, depolymerization, polymerization etc. As known drugs can modulate MT networks in various known ways, feature comparison of random drug treated cells with only the negative control features would provide the most fair representation of the effect of the unknown drug. For instance, using a depolymerising drug as positive control to set the previously listed parameters would have led to an optimal setting in separating features of potential depolymerising drugs and inactive drugs. Drugs having stabilising or polymerising effects could not be detected and would be classified as false negatives.

5.3.3. Cell cycle HCA experimental procedure

Stage 1, plate set up

PBOX compounds were screened in human MCF-7 breast cancer cells. Cells were treated for 16 hours which corresponds to the optimal dose response incubation time according Mulligan et al[32]. A stage 1 screen consisted of three plates treated consisting of three independent experiments. Each plate included one positive control drug, two negative controls and three drugs and each plate was mapped as followed:

- 1 x 3 wells of untreated cells: only medium, negative control 1.
- 1 x 3 wells of solvent treated cells: 0.5 % DMSO or ethanol, negative control 2.
- 1 x 3 x 6 wells of Nocodazole treated cells at 6 concentrations: 1; 5; 10; 15; 30; 50μM, positive control.

- 3 x 3 x 8 wells of 3 different PBOX compounds at 8 concentrations: 1; 5; 10; 15; 20; 25; 30; 50μM, drug test.

Stage 2, plate set up

MDG compound primary screening (stage2) use breast cancer cells, treated for 24 hours to optimize the chances of detecting any effect. The screen consisted of three plates treated at independent times. Each plate included one positive control drug, four negative controls and 13 drugs, each plate was mapped as followed:

- 1 x 3 wells of untreated cells: only medium, negative control 1.
- 1 x 3 wells of solvent treated cells: 0.5 % DMSO, negative control 2.
- 1 x 3 wells of solvent treated cells: 1 % DMSO, negative control 3.
- 1 x 3 wells of solvent treated cells: 5 % ethanol, negative control 4.
- 1 x 3 x 2 wells of nocodazole treated cells at 2 concentrations: 10; 50μM, positive control.
- 2 x 3 x 13 wells of 13 different MDGs at: 2 concentrations: 10; 50μM, drug test.

Stage 3, plate set up

All drugs which were tested positive were then processed to the secondary screen (stage 3). MDG compound secondary screen used of breast cancer cells, treated for 24 hours. The screen consisted of three plates treated at independent times. Each plate included one positive control drug at 10 and 50μM, two negative controls and 13 drugs at two concentrations 10 and 50μM and each plate was mapped as followed:

- 1 x 3 wells of untreated cells: only medium, negative control 1.
- 1 x 3 wells of solvent treated cells: 0.5 % DMSO, negative control 2.
- 1 x 3 wells of solvent treated cells: 0.5 % ethanol, negative control 3.
- 1 x 3 x 5 wells of Nocodazole treated cells at 5 concentrations: ; 6.25; 12.5; 25; 50μM, positive control.
- 3 x 3 x 8 wells of 3 different MDG at 8 concentrations, drug test:
  - 1; 5; 10; 15; 20; 25; 30; 50μM: MDG27, MDG29, MDG30, MDG32, MDG35, MDG36, MDG39, MDG40, MDG41, MDG61.
  - 0.1; 0.5; 1; 4; 5.5; 7; 8.5; 10μM: MDG31, MDG42, MDG49.
  - 0.1; 0.5; 2; 4; 6; 8; 10; 12μM: MDG44.
  - 0.1; 0.25; 0.5; 1; 2; 2.5; 3; 4μM: MDG43.
  - 0.1; 0.5; 1; 5; 10; 15; 20; 25μM: MDG45.

Experimental protocol

The experimental protocol steps followed for cell cycle analysis depicted as follows:

- 96 well plates were seeded with MCF-7 cells at a concentration of 3 x 10^4 cells/mL in a final well volume of 200µL of medium.
- Plates were incubated* for 24 hours at 37°C to allow cells to adhere to the plate surface.
- Wells were treated and incubated* according to the description in the previous section. The medium was renewed and the final volume of 200µL per well conserved.
- Cells were washed[a] once.
- Cells were fixed using 100µL fixing solution[b] and left for 15 min.
- Cells were washed[a] once.
- Cells nuclei were stained using 50µL of stain solution[c] and left in dark for 15 mins.
- Cells were washed[a] twice, wrapped in tin foil and conserved in a cold room at 5°C.

*: incubation is detailed in the "cell growth and maintenance" section.

[a]Wash solution: phosphate buffered saline (PBS) solution made up by dissolution of 10 tablets in 1 L of sterile water to reach a concentration of 0.01 M phosphate buffer, 0.0027 M potassium chloride and 0.137 M sodium chloride, pH 7.4, at 25 °C.

[b]Fixing solution: the initial solution of 37% formaldehyde in water (containing 10-15% methanol as inhibitor to prevent polymerization) was diluted in PBS solution down to 4 % at pH 7.4. Formaldehyde (CH_2O) reacts with water to form methylene hydrate ($HOCH_2OH$), which then reacts with various part of proteins in cells and mainly in membranes to form methylene cross-links. Whereas other larger aldehydes also produce a similar fiixing action, none approaches the completeness of formaldehyde.

[c]Stain solution: Cells were stained with Hoechst 33342 at a concentration of 10µg/mL in PBS. The original stock was either powder or solutions adjusted in each case by successive dilution to 10µg/mL. This dye has a high affinity with AT-rich DNA and binds at stoichiometric level. It is one of the most commonly used dyes for this procedure and it fluoresces in the blue range of the spectrum (Excitation max. = 346 nm, Emission max. = 460 nm).

Scan protocols

Each plate was read using the KSR, the settings used beeing adapted to each series of plate. In general five plates at the time were manually prepared for testing. The parameters of fixed exposure time and of the data mining algorithm were set once for a random plate and kept for the whole plate series. Example of settings are presented:

Assay protocol:

- Camera: Quantix camera32.1
- Acquisition mode: Standard
- Focus mode: Autofocus
- Autofocus freq: 1
- Objective: 10x
- Number of channels: 1
- Focus channel: 1
- Label: Nuclear
- Dye: XF93- Hoechst
- Exposure type: Fixed exposure time
- Exposure time: 0.02 seconds
- Object Identification - method: Isodata threshold, value -0.27
- Scan limits- minimum fields per well: 48
- Scan limits- Minimum object for well: 1500

In this section, there are three important settings to observe. First the exposure time had to be constant for each well as the intensity amplitude read is dependent on this parameter. Secondly, the KSR is set in order to generate enough images per well so that the number of cells used for the following analysis is large enough for statistical significance and constant from well to well. Last, the autofocus was set to use the 4 corner wells of the plate giving the best consensus for the whole plate reading.

Kinetic protocol:

- Kinetics mode: well mode
- Kinetic type: kinetic cycles
- Cycle timing: as fast as possible
- Time point: 1
- Store image set for: every Kinetic Cycle

This section denotes that each well will be processed only once per plate. For instance, the KSR can be used for the analysis of living cells. In this case, the user might need to scan each well several times to track cell changes (such as motility) occurring during a certain period of time.

Plate protocol:

- Selected assay protocol: indicates the name of the assay protocol previously described.
- Selected kinetic protocol: indicates the name of the kinetic protocol previously described.
- Selected plate type: Costar thin 96

This section summarises all the settings to be used during the scanning process. The type of plate is of major importance as different makes of plate will have slightly different well bottom widths which need to be taken into account to improve the quality and speed of the camera focus (along the Z axis).

5.3.4. Screening of compounds for effects on the microtubule network (HCA setting)

Stage 4 plate set up

The last MDG screen consisted of the following selection of compounds: MDG27, MDG29, MDG30, MDG31, MDG32, MDG35, MDG36, MDG40, MDG41, MDG42, MDG43, MDG44, MDG45, MDG49, MDG61. Each compound was tested at 10μM. Negative controls consisted of ethanol-treated, DMSO-treated and untreated cells. Positive controls were also tested at 10μM concentration and compounds selected were: nocodazole, taxol, combretastatin A-4 and PBOX6. Each well was replicated thrice and the assay was repeated three independent times.

Experimental protocol

Cells were seeded, treated and their nuclei were stained as described in the cell cycle section. Next, the following protocol was applied:

- Cells were washed[d] once.
- Cells membranes were permeabilised using 100uL of PERM solution[e] and left for 5 min.
- Cells were using 100uL of BLOCK solution[f] and left for 30 min.
- Cells were washed[d] once.
- Cell microtubules were stained using 100μL of Antibody solution[g] and left in dark in the cold room at 5°C for 18 hours.
- Cells were washed[d] three consecutive times.
- Wells were filled in with 200μL of PBS solution[a], wrapped in tin foil and conserved in a cold room at 5°C.

[d]: Wash solution: immunofluorescence wash buffer, consisted of PBS solution[a] mixed with 0.01% Tween-20.

[e]:PERM solution: permeabilisation buffer consisted of 0.3% w/v Triton X100, 5% w/v bovine serum albumine in PBS solution[a].

[f]:BLOCK solution: blocking buffer consisted of 5% 5% v/v Bovine Serum Albumine in PBS solution[a].

[g]: Antibody solution: consisted of monoclonal Anti-β-Tubulin−Cy3 antibody produced in mouse. It was made up clone TUB 2.1(see appendix of chapter 4 for sequence) purified immunoglobulin and the stock solution was made up 0.01 M phosphate buffered-saline, containing 1% bovine serum albumin and 15 mM sodium azide, pH 7.4. The antibody recognizes all five isoforms of β-tubulin (β1-β5)[33]. It reacts with the β-Lc and β-Sc fragments in the carboxy-terminal part of β−tubulin in immunoblotting.

reacts with the β-Lc and β-Sc fragments in the carboxy-terminal part of β–tubulin in immunoblotting. The stock solution was then diluted 1:200 in BLOCK solution and preserved in the dark at 5°C. This dye fluoresces in the red. (Excitation max. = 545 nm, Emission max. = 565 nm).

Scan protocols

Each plate was read using the KSR and the settings used were adapted for each plate. Example of settings are presented :

Assay protocol:

Camera: Quantix camera32.1
Acquisition mode: Standard
Focus mode: Autofocus
Autofocus freq: 1
Objective: 20x
Number of channels: 3
Focus channel: 3
Exposure type: Fixed exposure time
Object Identification method: SigmaThreshold, Algorithm value: -0.99
Min Object per well: 100.

Analogously to the previously described assay protocol, the exposure time was chosen to be constant for each well along a particular channel (the exposure time can be different for channel1 and channel 3). As the objective used was 20X, the minimum number of objects per well was reduced to 100 providing the acquisition to minimise memory storage. The focus channel was done on channel 3 which corresponded to the tubulin antibody reading to optimize the image for cytoskeletal morphology analysis. Channel 1 was set to read the nuclei to be used as reference for the object determination in the analysis and channel 2 was reading the same wavelength as channel 3. Channel 2 was used as a virtual extra channel used by the morphology bioapplication in order to obtain output features concerning the entire cell information such as the whole cell surface which channel three in the bioapplication does not provide.

The Kinetic and plate protocol settings were adapted for the present biological assessment which required three channels.

5.3.5 IC 50 value determination

DATA sorting

The percentage of cells in the G2 phase of the cell cycle was used to determine IC50 values for the compounds tested. As commonly done in the HTS area [34], to examine the reliability of these

according to their nucleus total intensity. The total intensity read per cell nucleus was restricted to 10 % maximum deviation from the mean of the intensity of the same whole cell nucleus. In addition, for each concentration tested a minimum of 4 points was used up to a maximum of 9 points and at least 8 different concentrations were used to evaluate the IC50.

Dose response fit

Each drug's data was normalised per plate using the corresponding appropriate solvent well values as a baseline (0% drug response) and the positive control response as the maximum of the curve. After normalisation of the responses (Y axis) and transformation of the concentrations (X axis) into a logarithmic scale, the data was fitted to the standard dose response curve with the following equation:

PBOX curves and Nocodazole curve; Sigmoidal dose-response

$$Y = Bottom + (Top-Bottom) / (1+10^{((LogIC50-X))})$$

MDG curves; Sigmoidal dose-response (variable slope)

$$Y = Bottom + (Top-Bottom) / (1+10^{((LogIC50-X)*Slope)})$$

X is the logarithm of concentration. Y is the response; Y starts at the Bottom and goes to the Top with a sigmoid shape.
Regressions and curve fittings were accomplished using GraphPad Prism software[36].

5.3.6 Tubulin polymerisation assay

Microtubules are intrinsically dynamic polymers, but do not have a simple equilibrium. They polymerise and depolymerise by the reversible addition and loss of tubulin dimers at the ends of the mircrotubules. The assembly of purified bovine tubulin was monitored using the CytoDYNAMIX ScreenTM3 [37]. This assay uses a 96-well plate format with 300µg of >99% purified bovine brain tubulin in each well. The tubulin (1mg, in lyophilized format) was resuspended on ice in 330µl of ice-cold G-PEM buffer (80mM Piperazine-N,N'-bis[2-ethanesulfonic acid] sequisodium salt (PIPES) pH 6.9, 0.5mM $MgCl_2$, 1mM EGTA, 1mM guanidine triphosphate (GTP), 5% (w/v) glycerol) and left on ice for 1 minute to allow for complete resuspension. 10µl of 10X strength of each compound tested made up in G-PEM buffer was pipetted into a half area 96-well plate prewarmed to 37°C. A 100µl

ice for 1 minute to allow for complete resuspension. 10µl of 10X strength of each compound tested made up in G-PEM buffer was pipetted into a half area 96-well plate prewarmed to 37°C. A 100µl volume of tubulin was pipetted into the designated wells of the prewarmed plate. It is important for the total pipetting to be less than 30 seconds from start to finish. The assay was conducted at 37°C and tubulin polymerisation was followed turbidimetrically at 340nm in a Spectramax 340PC spectrophotometer (Molecular Devices). The absorbance was measured at 30 second intervals for 1 hour.

5.3.7. Cell culture

Growth and maintenance

MCF-7 (Human Caucasian breast adenocarcinoma cells) were obtained from the European Collection of Cell Cultures (ECACC). MCF-7 cells were grown at 37°C under a humidified atmosphere of 95% O2 and 5% CO2. They were maintained in complete medium (87% Minimum Essential Medium (MEM), 10% (v/v) Foetal Calf Serum (FCS), 1% of 2mM L-glutamine, 1% 100µg/ml gentamycin, 1% (v/v) non-essential amino acids.
They were seeded in 75cm^2 flasks in 20ml of medium or in 175cm^2 flasks in 30ml of medium. Confluent cultures were passaged 1:3, twice weekly to be finally discarded after 30 passages.

Cryopreservation

MCF-7 cells were grown to a state of sub-confluency, harvested, counted and centrifuged. The pellet was resuspended in 1ml of a mixture of FCS and dimethyl sulphoxide (DMSO) (9:1) then transferred into 1.5ml cryotubes which were placed at -20°C for 2 hours and then at -80°C for3 hours. The storage of cells was done in a liquid nitrogen vessel. When required, an aliquot of cells was removed from the liquid nitrogen vessel, rapidly thawed and resuspended in 10ml of complete medium, the cell suspension was centrifuged at 500xg for 5 minutes, the medium containing DMSO was discarded, the pellet was resuspended in complete medium and the cells were seeded in tissue culture flasks

5.3.8 Cell count

Cells were counted for seeding using a haemocytometer as follows:
A clean coverslip was placed over the haemocytometer's surface, a 10-20µl volume of cell suspension was pipetted under the cover slip making sure not to overfill the volume underneath the coverslip so as

the left or top boundary are counted, while those that touch or are outside the lower or right hand boundary are not counted. The volume of cell suspension that will occupy one medium square is:

$0.04\ mm^3$ ($0.02\ mm^2$ x 0.1 mm) or 4.0×10^{-6} ml.

- There are 25 of the medium squares, which totals a volume of 1×10^{-4} ml.
- Total cell concentration in the original suspension (in cells/ml) is then:

cells/ml = total count in the 25 medium squares x 1×10^4.

- Viable cell number can be determined by counting the number of cells that exclude a dye such as Trypan Blue. A viable cell possesses an intact cell membrane and therefore does not allow uptake of Trypan Blue.
- A suspension of cells is diluted 1:5 with Trypan blue and the number of cells excluding and including the dye is counted giving a number of viable and non-viable cells.

5.3.9 Preparation of drug solutions

Prior to testing, all drugs were dissolved in DMSO or ethanol depending on their solubility. The initial stock concentration aimed was of 10μM although due to solubility issues for some compounds it was necessary to reduce the stock concentration through a series of successive dilutions. Sometimes to promote compound dissolution, mixtures were warmed up using a water bath at a maximum temperature of 50°C to minimise compound decomposition. Drugs not going into solution were discarded from further testing.

5.5 Results

5.5 1 HCA cell cycle analysis validation

As the assay developed in the present work was the first attempt undertaken on site in measuring cell cycle arrest using high throughput fluorescence microscopy, it was firstly necessary to perform optimisation and validation steps.

Determination of optimal cell and dye concentrations

The first optimisation step consisted of the determination of the optimal cell seeding density and dye concentrations for nuclei staining. A range of diverse cell and stain concentration were tested to ascertain the best combination. The ratio of the number of cells classified in the G1 phase over the G2 phase of the cell cycle (G1/G2) was used as the decisive parameter. Negative control ratios ($G1/G2_{neg}$,

ascertain the best combination. The ratio of the number of cells classified in the G1 phase over the G2 phase of the cell cycle (G1/G2) was used as the decisive parameter. Negative control ratios ($G1/G2_{neg}$, solvent treated cells) and positive control ratios (($G1/G2_{pos}$), Nocodazole) at various concentrations) were plotted against various combinations of cell and dye concentrations. Subsequently, a concentration of $3x10^4$ cells/mL and 10µg/mL of nucleus stain showed the best separation between all tested controls.

Nocodazole, positive control

Nocodazole was chosen as positive control, therefore in order to validate the complete IC50 evaluation method, it was essential to estimates the nocodazole IC50 value for the same assay under the same conditions. Nocodazole was tested over a range of concentrations and Figure 5.6 illustrates a random set of profiles among the multiple results obtained.

The typical healthy profile of MCF-7 untreated cells was retrieved with the main peak of cells in G1 and a smaller G2 peak. Nonetheless the S phase contained more cells than expected. The analysis of the many hundred profiles obtained in the present work showed that this over-estimation of S cells was constant. This feature is specific to the technology of the present assay. Other work carried out with the KSR showed the same overestimation in S phase classified cells. However, the typical shift of the main peak to G2 phase observed for Nocodazole-treated cells was successfully detected. Nocodazole has the property of binding tubulin in the colchicine site preventing the separation of the two daughter cells after mitosis. This explains the increased number of cells found in G2 and the decreased total number of cells once apoptosis is initiated.

Previous work undertaken with nocodazole at 10µM using MCF-7 cells after 16 hours treatment showed ~50% of cells arrested in G2 compared to 10 to 30 % for vehicle treated cells[38]. This assay was carried out using FACS technology. The present work estimated a similar percentage of cells in G2 following nocodazole treatment. The percentage of cells in the G2 phase was subsequently fitted to determine a nocodazole IC50 values. The average IC50 value was 4.5µM +/- 3.72 was calculated using three independent experiments. This is consistent with previous reports from other independent groups: An IC 50 of 3.20µM[37] for MCF-7 tubulin inhibition was calculated using the ratio (neuronal tubulin/cancer cell tubulin); an IC 50 of 8µM[39] for MCF-7 mp53 cells treated for 8h and in combination with another compounds. However a fluorescence microscopy technique was used to detect mitotic cells and the full effect was observed for 40µM; an IC 50 value between 5-7,5µM[31] for A0549 cells treated for 24h where the IC50 calculation was based on Rb phosphorylation measurement and cells in G1 phase.

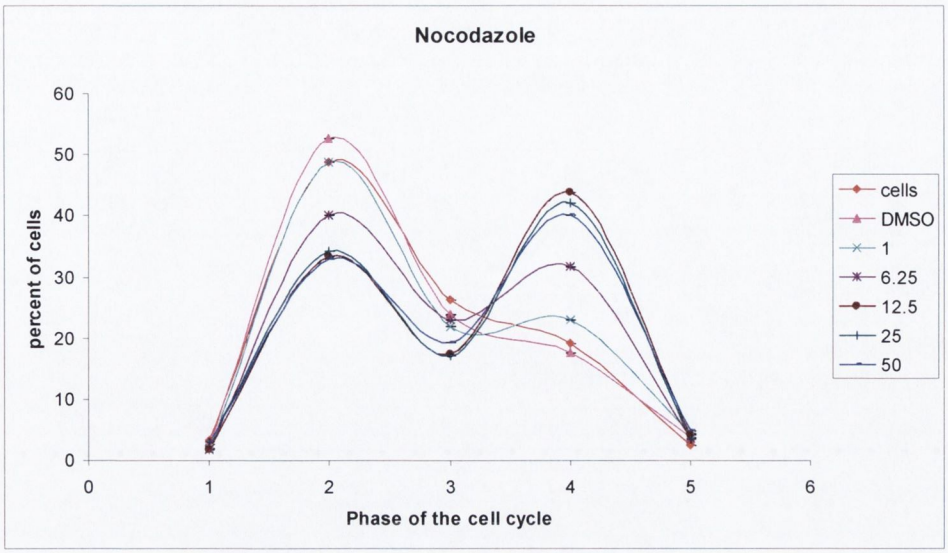


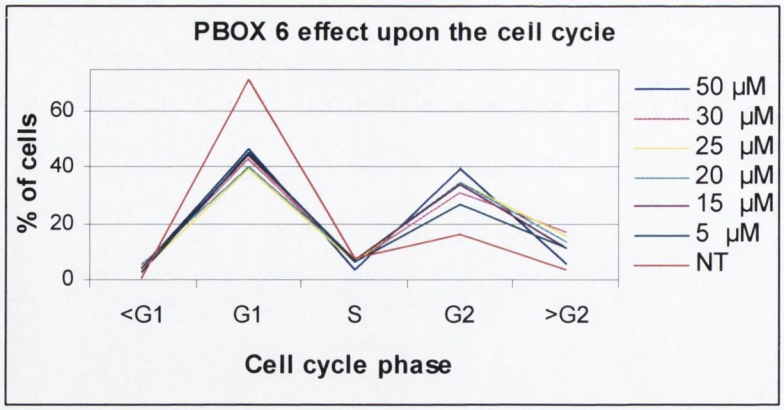
Figure 5.6: Nocodazole cell cycle profiles

On the X axis are the cell cycle phases: (1) = (< G1), (2) = (G1), (3) = (S), (4) = (G2) and (5) = (> G2). MCF-7 cells were treated for 16 hours at a seeding density of $3x10^4$ cells/mL and 10µg/mL of nucleus stain as described in the method section.

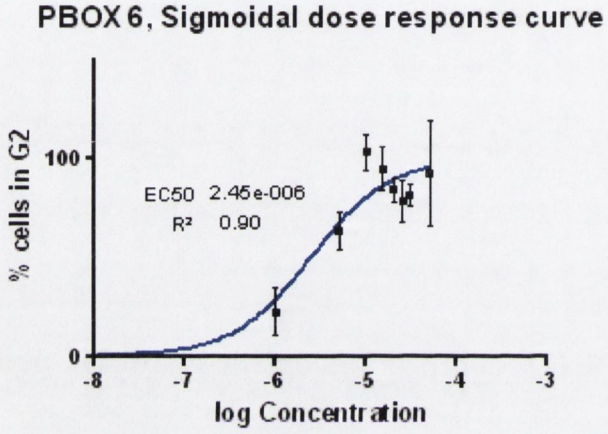
5.5 2 The effect of PBOX compounds on the cell cycle profile of MCF-7 cells

PBOX effect on the cell cycle

Following the previous validation, PBOX compounds were screened under the same conditions and results were fitted to the same curve equation used previously. Figure 5.8.a illustrates a randomly chosen profile among all those generated for the PBOX6 active compound and in Figure 5.7.b, an IC50 curve fitting is shown.



(a)



(b)

Figure 5.7: PBOX 6 effects on the cell cycle

(a) PBOX6 cell cycle profiles for different concentrations of drug and (b) PBOX6 IC50 curve.

MCF-7 cells were treated for 16 hours at a seeding density of $3x10^4$ cells/mL and 10µg/mL of nucleus stain as described in the method section.

Figure 5.7.a profiles slightly underestimated the percentage of cells per phase of the cell cycle when compared to the previous performed with FACS.[32] However, the general trends and separation between diverse concentrations were still observed. Furthermore, every plate was normalized for calculation of the IC 50 value which concurred with a viability assay published where PBOX6 showed an IC 50 value from 1 to 2.3µM[6].

The errors inherent to the HCA assay prevented us from ranking accurately PBOX compound potencies. Alternatively, compounds were categorized in activity classes namely: active (*cf.* Table 5.5), intermediate active (*cf.* Table 5.6) and inactive (*cf.* Table 5.7) PBOX.

PBOX ID	IC50 in μM (test 2)	Error in μM	R^2	G2/M Test (test 2)	Apoptosis test (test 1)
PBOX3	4.78	2.32	0.9	active	YES
PBOX4	4.02	3.84	0.5	active	YES
PBOX5	3.38	1.88	0.9	active	YES
PBOX6	2.98	1.70	0.9	active	YES
PBOX7	4.02	2.43	0.9	active	YES
PBOX8	3.10	2.43	0.7	active	YES
PBOX9	**15.75**	**12.30**	0.5	**active**	**YES**
PBOX15	1.49	ND	1.0	active	YES
PBOX16	2.58	2.57	0.8	active	YES
PBOX68	**80.75**	**59.25**	0.6	**active**	**YES**
PBOX70	2.96	2.80	0.7	active	YES
PBOX71	3.01	2.64	0.8	active	YES

Table 5.5: PBOX active IC50 values.

PBOX ID	IC50 in μM (test 2)	Error in μM	R^2	G2/M Test (test 2)	Apoptosis Test (test 1)
PBOX32	20.25	20.25	< 0.5	**Intermediate**	**NO**
PBOX43	11.60	11.60	< 0.5	**Intermediate**	**NO**
PBOX46	73.20	56.80	< 0.5	**Intermediate**	**NO**
PBOX66	79.05	50.95	< 0.5	**Intermediate**	**NO**
PBOX67	96.50	53.50	< 0.5	**Intermediate**	**NO**
PBOX72	86.20	43.80	< 0.5	Intermediate	ND
PBOX78	71.85	48.15	< 0.5	Intermediate	ND

Table 5.6: PBOX with showing intermediate activity.

PBOX ID	G2/M Test (test 2)	Apoptosis Test (test 1)	PBOX ID	G2/M Test (test 2)	Apoptosis Test (test 1)
PBOX1	inactive	NO	PBOX48	inactive	NO
PBOX2	inactive	NO	**PBOX69**	**inactive**	**YES**
PBOX12	inactive	YES	PBOX73	inactive	ND
PBOX21	inactive	NO	PBOX74	inactive	ND
PBOX23	inactive	NO	PBOX75	inactive	ND
PBOX24	inactive	NO	PBOX79	inactive	ND
PBOX25	inactive	YES	PBOX80	inactive	ND
PBOX26	inactive	NO	PBOX81	inactive	ND
PBOX27	inactive	YES	PBOX82	inactive	ND
PBOX31	inactive	NO	PBOX83	inactive	ND
PBOX33	inactive	NO	PBOX84	inactive	ND
PBOX34	inactive	NO	PBOX85	inactive	ND
PBOX35	inactive	NO	PBOX86	inactive	ND
PBOX38	inactive	NO	PBOX65a	inactive	NO
PBOX42	inactive	NO	**PBOX65b**	**inactive**	**YES**

Table 5.7: PBOX that did not show any activity.

- The IC50 values correspond to the G2/M effect detected after 16 hours treatment on MCF-7 cells. Cells were seeded density of 3x104 cells/mL and 10µg/mL of nucleus stain as described in the method section
- The error column represents the IC50 deviation from the mean of all experiments.
- The R^2 is the correlation factor of the data to the corresponding fitted IC50 curve.
- G2/M test correspond to PBOX classification: active, intermediate or inactive.
- The apoptosis column corresponds to an apoptotic test carried-out by co-workers on HL60 cells with 10µM drug for 16 hours. ND stands for Not Determinated.

All of the tables, denoted Tables 5.5 a, b and c, present a supplementary column related to an apoptosis test carried out by M. Mc Gee. This test, denoted test 1, was a first qualitative evaluation of apoptosis whereas the present work, denoted test 2, was a more quantitative classification of compounds' influences on cell division. Indeed, test 1, measured the quantity of apoptotic cells by

direct observation of drug treated cells under a microscope. In this test, the decision to count a cell as either apoptotic or healthy was entirely dependent on the experimentalist's assessment of subtle morphological features. PBOX 9 (highlighted in bold characters) classified as active in the present test, obtained a large uncertainty compared to the rest of PBOX compounds classified actives. Nonetheless, the principal component analysis (PCA) carried out in Chapter 3, suggested PBOX9 as active. In the case of concerning PBOX68 that also obtained a large error, the PCA suggested it as intermediate. These compounds were classified as actives within some uncertainty. The large errors were probably due to the concentration range used for the screening which was kept unchanged for all PBOX. Therefore, for compounds weakly actives this concentration range was not optimal to cover the full IC50 curve.

PBOX qualitative apoptotic measurement

As, both PBOX compounds and nocodazole can engender cell death through an apoptotic pathway, an attempt to detect apoptotic specific morphological changes on nuclei, was carried out. Figure 5.8 illustrates a set of images obtained with the KSR at 10x magnification of stained nuclei respectively treated with vehicle (solvent), nocodazole and PBOX6 at 10μM.

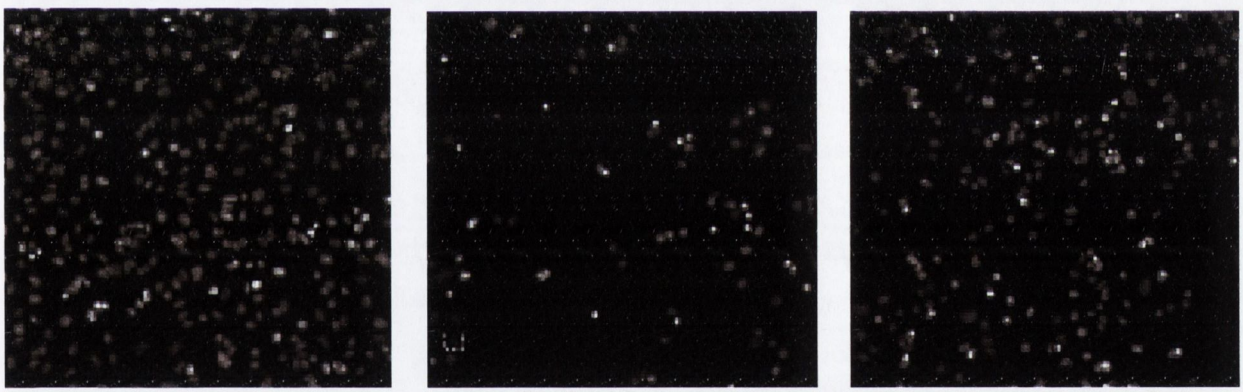


Figure 5.8: Example of images from scanned cell nuclei.

MCF-7 cells were treated for 16 hours and seeded density of $3x10^4$ cells/mL and 10μg/mL of nucleus stain (Hoescht) as described in the Method Section.

The total number of cells was found to decrease with the various treatments and the percentage of more intense spots per image increased. The more intense a nucleus' stain is, the more it fluoresced, the more dye it contained, hence, the more DNA the nucleus contained. Thus, the cell division was stopped by

preventing the cells from separating the duplicated genetic information into two daughter cells. The kineticscan reader coupled to its bioapplication has the ability to assess changes in a large number of morphological features of the nuclei. (more than 20 features calculated in total). These features can represent subtle changes that cannot be detect with the human eye. Some of these features are demonstrated in Figure 5.9 which illustrates the most variable morphological features after drug treatments when compared to untreated cells.

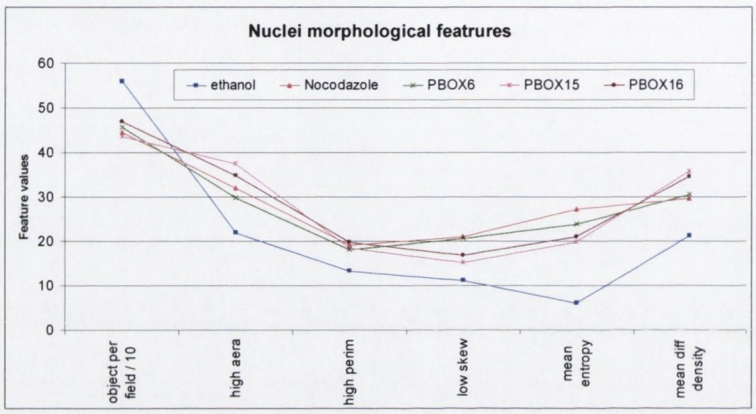
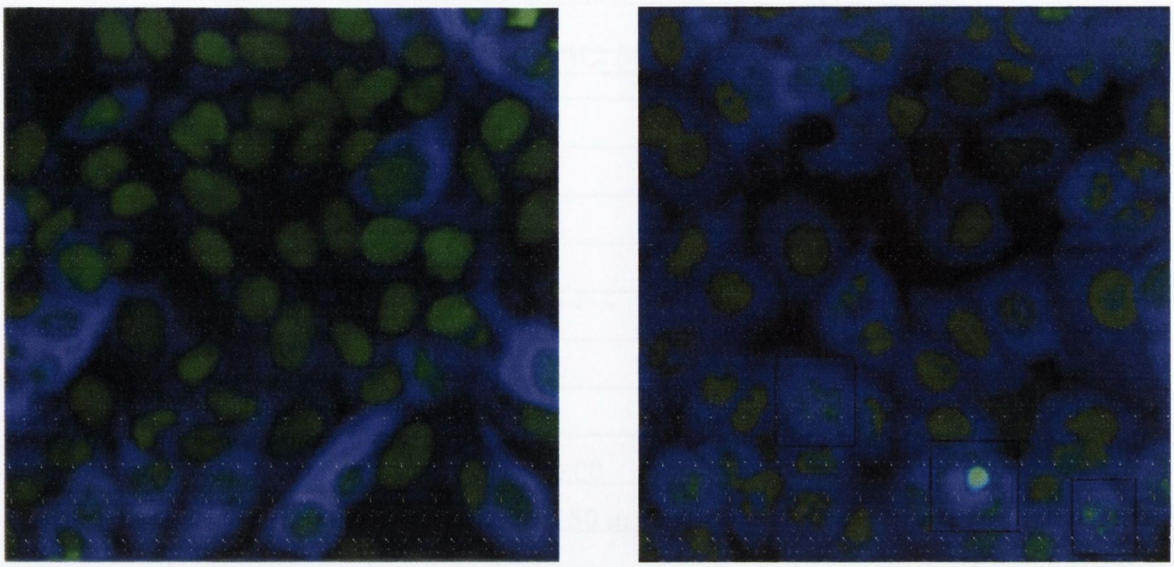


Figure 5.9: Morphological nuclei changes.
MCF-7 cells were treated for 16 hours at 15µM concentration of various drugs, at a seeding density of $3x10^4$ cells/mL and 10µg/mL of nucleus stain

The four drugs tested elicited a similar profile compared to the vehicle (ethanol) profile. As expected the total number of object (object per field /10 in Figure 5.9) and here nucleus, decreases due to the drug treatment's antiproliferative effects but also could be from cytotoxic'. The sub-population of cells having a nucleus mean area higher (high area) than most of the cells increases with drug treatment. The same observation can be made for their perimeters (high perimeter). These could be apoptotic features such as the enlargement of the nucleus and its more convoluted membrane. However, these measurements could also represent polyploid cells. Polyploid cells are cells having more than one nucleus and usually are observed with drugs toxicity. Very similar changes in nuclei morphology have previously been measured by Pfizer for toxicity[34] assessment. Features such as low, skew, mean entropy and mean difference density which also increase with drug treatment could be specific to apoptosis. They all translate the same phenomena, the nuclear material is no longer homogenously distributed inside the nuclei so that the genetic information (stained DNA) is disordered. As explained in this Chapter's Introduction, one of the phases of apoptosis involves DNA digestion into fragments, ultimately packed into vesicles including changes such as karyorhexis and pyknosis. Pyknotic chromatin appears as characteristic crescent-shaped "caps" under light microscopy (densely

heterochromatic regions). Figure 5.10 illustrates at a bigger magnification the DNA changes involved in apoptosis.



(a) Untreated cells (b) Nocodazole 10μM

Figure 5.10: Morphology of apoptotic nuclei.

Composite image of 2 channels at x40 magnification. (a) Solvent treated cells, (b) Nocodazole treated cells. In blue tubulin and in green DNA. MCF-7 cells were treated for 24 hours, nuclei and cytoskeleton were stained as described in the method section.

The squared nuclei are selected as representative of the apoptotic nuclei described earlier.

This HCA assay is consequently promising in detecting antiproliferative effects linked to G2 arrest and apoptosis if all suitable features are to be calculated simultaneously The next step consisted of screening the newly obtained compounds, to assess their effect upon cancer cells.

5.5 3 The effect of MDGs on the cell cycle profile of MCF-7 cells

MDG compounds primary screening or stage 2 , consisted of testing all compounds at 2 concentrations and determining the percentage of cells in the G2 phase. Compounds demonstrated to have arrested cells in the G2/M phase compared to the negative controls were kept for further testing. This was evaluated by calculated all ratios of MDG G2 cells over negative control G2 cells. Table 5.8 summarises the results for all compounds which showed activity and in red is marked the one which were retained for further testing.

MDG ID	RATIO ($G2_{CPD}$ / $G2_{NEG}$) LOW CONCENTRATION	RATIO ($G2_{CPD}$ / $G2_{NEG}$) HIGH CONCENTRATION
Nocodazole	1.2	1.6
MDG49	ND	1.5
MDG45	1.1	1.6
MDG44	1.1	1.6
MDG43	ND	1.4
MDG42	ND	2.6
MDG41	ND	1.5
MDG40	ND	1.7
MDG39	ND	1.1
MDG36	ND	1.7
MDG35	1.2	1.8
MDG32	1.2	1.6
MDG31	ND	2.0
MDG30	1.0	1.1
MDG29	1.0	2.6
MDG27	ND	2.2

Table 5.8: Stage 3, MDG compounds first filter.
Nocodazole high concentration 25μM and low 10μM, MDG high concentration 50μM and low 10μM, ND: Not Determinated.

The subsequent step or stage 3, consisted of the evaluation of actual potencies for the MDGs potent compounds. In the context of the MDG screening, testing only to retrieve G2 cells using the method described in the previous section could be misleading and could lead to the generation of to false positive compounds. Indeed, the previously classified cells as G2 cells, was actually combining both arrested cells and mitotic cells. As a subset of the most potent PBOX compounds had already been investigated by co-workers and demonstrated by diverse parallel techniques to be apoptotic, and so the extrapolation made on the tracked G2 cells was not misleading. The previous screen aimed to classify PBOX analogue potencies and these were more likely when activity was detected to follow the same apoptotic pathways than the thoroughly investigated PBOX15, PBOX16 and PBOX6 actives. In the context of MDG compounds, the screen undertaken was the first activity assessment. Therefore, classifying all G2 cells together was overestimating a still potential anti-mitotic effect. In an attempt to overcome this, IC50 values representing a decrease in nuclei were calculated in parallel. These IC50s

were based on the total number of cells detected independently of their cell cycle phase. A projection of all IC50 values calculated is illustrated in figure 5.11.

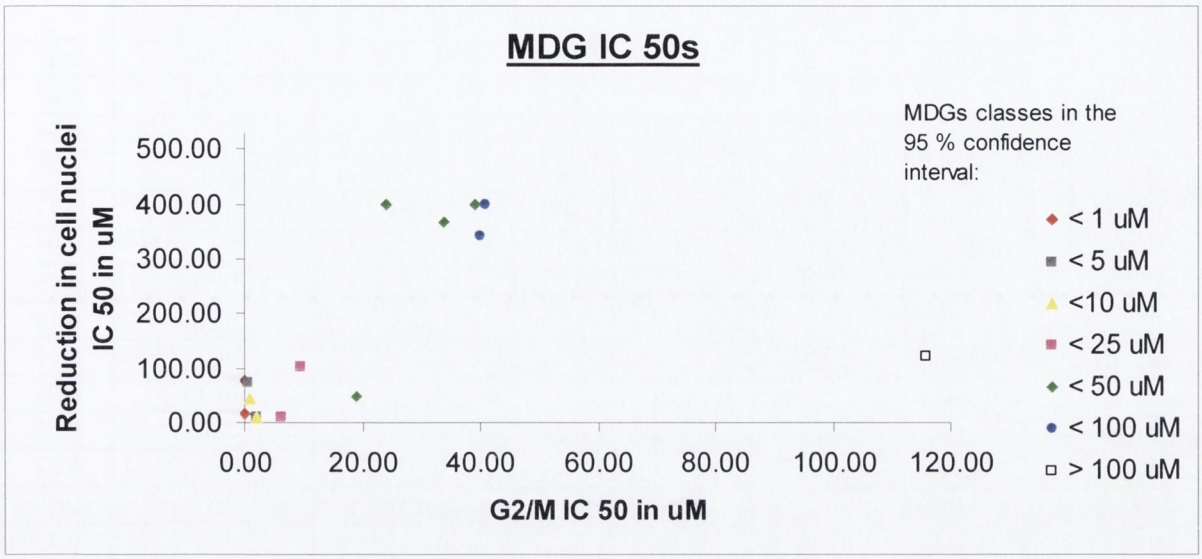


Figure 5.11: MDG_antiproliferative and G2 IC50 values
MCF-7 cells were treated for 24hours, nuclei as described in the method section

In figure 5.11, on the Y axis are the IC50 values for the reduction in cell nuclei and on the X axis G2/M IC50 values are projected. No errors bars were displayed on this graph for more clarity. Compounds which are found close to the origin of the graph (0, 0) present the smaller values for both IC50s. The smaller the IC50 value the more active is a compound. In addition MDG compounds similarly to the previous PBOX compounds were classified per range of activity according to the G2 IC50 maximum error. An example of misleading conclusions would be compounds marked in green which correspond to G2 IC50s < 50μM but do not present a potent anti-proliferative effect, >100μM (*e.g.* Table5.9 MDG27). Combining the data for both effects seemed therefore the best compromise for making conclusions on activity. Table 5.9 lists all the MDG IC50s calculated. MDG61 was added in the following screen as it was not available for the primary screen, (stage 2)..

	G2/M IC 50 in μM				Reduction in nuclei number IC 50 in μM		
MDGs ID	G2 Class	IC50	Error	R^2	IC50	Error	R^2
MDG31	< 1 uM	0.02	0.00	0.68	15.29	7.16	0.87
MDG35	< 1 uM	0.05	0.00	0.67	76.66	25.84	0.87
MDG43	< 5 uM	0.74	0.42	0.81	9.86	5.74	0.95
MDG32	< 5 uM	1.96	1.52	0.87	73.67	27.13	0.93
MDG42	< 5 uM	1.41	ND	0.95	8.94	0.72	0.77
MDG36	<10 uM	4.24	3.47	0.76	43.38	2.65	0.93
MDG44	<10 uM	8.95	0.62	0.91	10.57	0.53	0.96
MDG49	< 25 uM	6.27	6.11	0.79	8.55	0.58	0.78
MDG41	< 25 uM	9.43	4.31	0.85	102.36	79.74	0.69
MDG45	< 50 uM	19.07	15.09	0.62	46.31	20.35	0.96
MDG27	< 50 uM	24.13	10.76	0.84	inactive	ND	0.63
MDG61	< 50 uM	33.86	12.85	0.79	365.60	ND	0.62
MDG29	< 50 uM	39.07	4.28	0.90	inactive	ND	0.10
MDG30	< 100 uM	40.82	30.74	0.61	inactive	ND	0.43
MDG39	< 100 uM	39.94	12.90	0.93	343.22	304.28	0.93
MDG40	> 100 uM	162.71	114.19	0.95	229.84	195.66	0.90

Table 5.9: MDG IC50 values for G2/M and reduction in cell nuclei in μM.
MCF-7 cells were treated for 24hours, nuclei as described in the method section
ND: Not determined.

Compounds with IC50 antiproliferative/cytotoxic < 200μM (error included) and G2/M IC50 values < 50μM (error included) were then tested on two Chronic Myeloid Leukaemia (CML) cell lines (LAMA84 and K562) by S. Bright. Compounds were tested at different concentration for 48h. Results are presented in figure 5.15 were the percentage of apoptosis was the number of cells detected in preG1 phase by FACS.

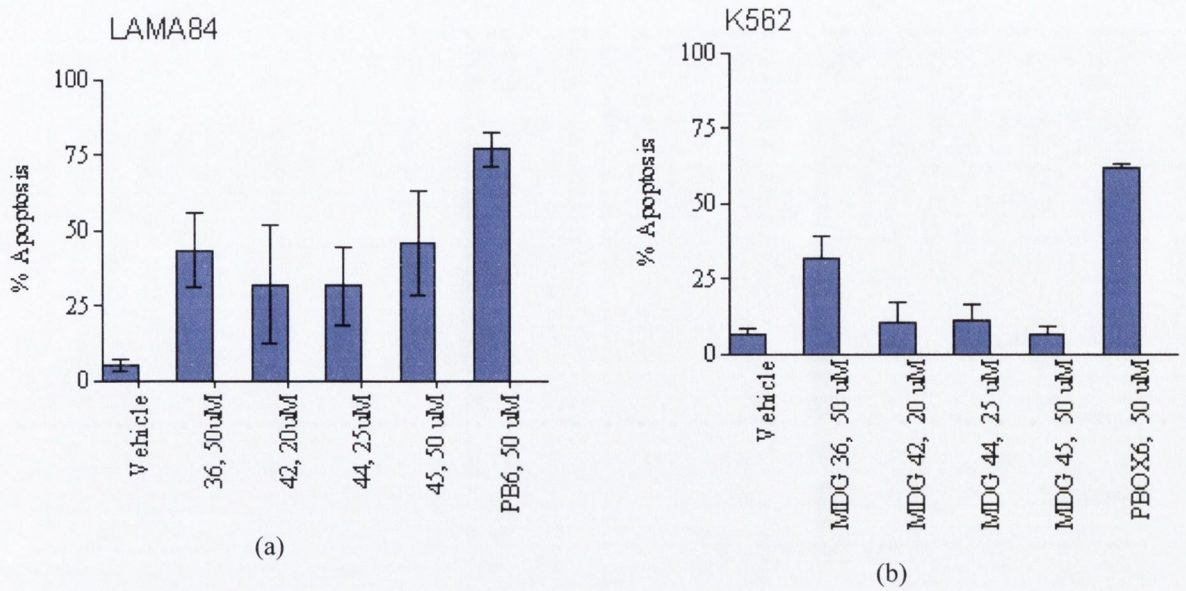


Figure 5.12: Active compounds tested on leukaemia cell lines.(a) LAMA85 cells and (b) K562 cells. Bars are mean of the percentage of cells found in pre-G1 of at least 3 independent experiments and error bars correspond to standard errors. Cells were treated for 48 hours and analysed by FACS.

It was preferable to test all these compounds at the chosen concentration of 50μM for comparisons purposes but due too solubility problems MDG42 and MDG44 attained their maximum concentration respectively at 20μM and 25μM. All the MDG compounds tested demonstrated to induce apoptosis on LAMA84 cells but elicited less potent effect in the respectively highly resistant K562 cells. The apoptotic test carried out demonstrated MDG36 as the most promising of these compounds as it showed activity for both cell lines. Four compounds denoted MDG42, MDG44 and MDG45 showed promising results. All MDG actives were then examined for their effect on the MT network, stage 4.

5.5 4 HCA for tubulin effect detection and validation

In order to assess the compound effect on the MT network, a reliable assay using HCA was developed. The first step as usual consisted in validating the protocol. To do so, several suitable positive control drugs were screened at a concentration of 10μM. Subsequently MCF-7 nuclei and tubulin were stained, scanned and analysed. Nocodazole and combretastatin were chosen as positive controls in order to represent the destabilizing effects on MTs while taxol was used to represent the stabilizing effects. PBOX6 has previously been shown to destabilize MTs and was therefore included in the compound library as a "hidden active" or "decoy". None of the positive control wells were used for the settings of either the scanning process (KSR) or the analysis parameters. Figure 5.16 illustrates images obtained for the control drugs and PBOX6.

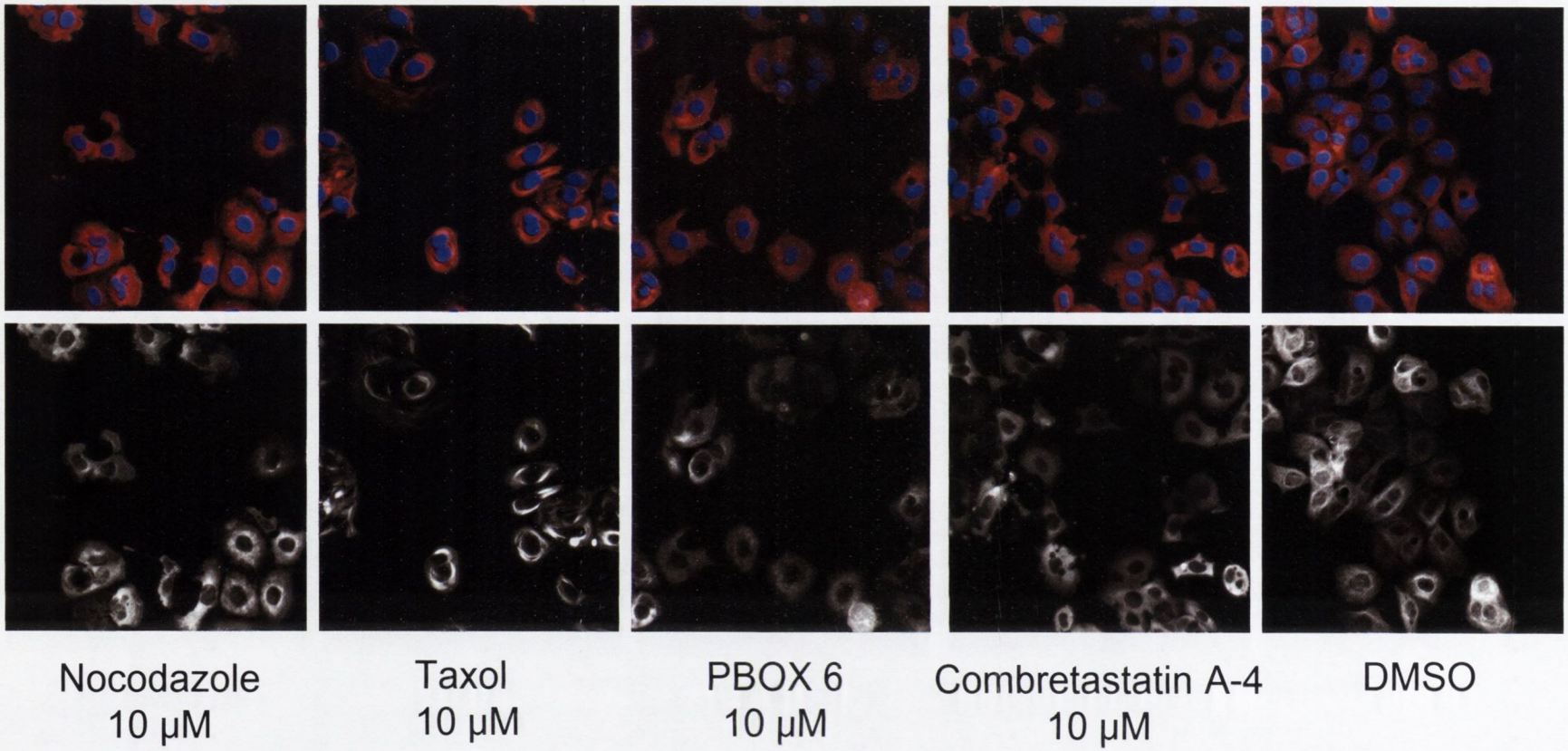


Figure 5.2: Controls images from KSR at x 20 magnificence.

On the composite images, in red tubulin staining and in blue nuclei staining.

All compounds were tested for 24 hours at 10µM concentration as described in the Methods Section.

Figure 5.13 presents two rows of pictures, in the first row is presented the composite images showing in blue the nuclei and in red MTs whilst images in the second row illustrate only tubulin staining. At first sight, some differences can be seen amongst the variously treated cells. Vehicle wells, *i.e.* DMSO images, present cells in a shape recalling cats' eyes shape most likely to correspond to mitotic cells. Indeed, the bipolar spindle is formed between two microtubule organising centres during the prophase and are eventually pulled to opposite poles during anaphase. Thus, the few round cells are most likely to be G1 or S cells. This magnification however does not permit a conclusion to be made on the exact phase of the cell cycle. Nonetheless, a shape comparison among the diverse positive controls is sufficiently informative. Nocodazole treated cells present a more extended general shape showing poorly organized MTs. Nocodazole cells are more dense and on some cells, a star shape can be seen. In Combretastatin treated cells these shapes are more accentuated and the apparition of pseudopodia is more obvious. Cells are more tubulin dense, a characteristic of the complete loss of MT structures. Taxol treated cells result in massive MT bundles with MTs found tight close to the nuclei and the general MT shape is similar to a planet ring. Finally, PBOX6 treated cells present a more dense cytoskeleton with the appearance of some pseudopodia. These features elicited by PBOX6 would correspond to disorganised MTs. The qualitative analysis of PBOX6 treated cells, correctly classify PBOX6 as a destabilising agent. In order to screen considerable number of molecules, it is necessary to be able to quantify these qualitative observations.

To assess MT effects on more than 50 morphological features issuing from the stained MTs and nuclei were measured for all control drugs. These features can be categorised according to the compartment they describe, such as the nucleus, the entire cell or the cytoskeleton. Representative features are illustrated in Figure 5.14, where all features were normalised to untreated cells. No morphological feature in relation with whole cell or the nuclei was able to distinguish the different drugs. However, morphological features in relation to the MT network (read in channel 3) following the tubulin staining were able to differentiate not only drug treated cells from negative controls but also different categories of drugs (*i.e.* polymerising and depolymerising). In particular, the quantity of fibres detected and their distribution within the cell, were capable of separating all drugs from negative controls giving opposite tendency according to the effect upon MTs. PBOX6 nocodazole and combretastatin values are inferior to negative controls, as their MTs are more spread within the cell while taxol value was found superior as MTs are more concentrated around the nuclei

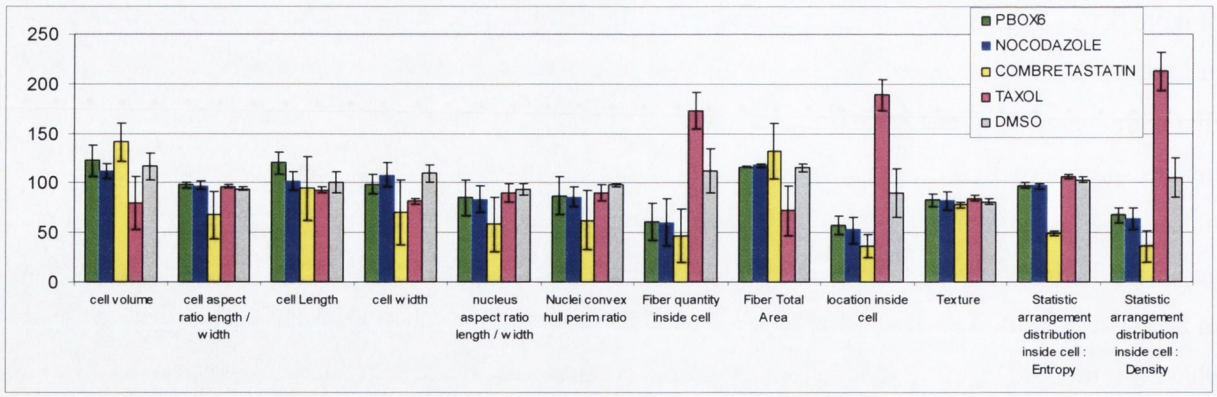


Figure 5.14: Effect of MTA on cell parameters following tubulin staining.
Features were normalized against the untreated cells. All compounds were tested for 24 hours at 10μM concentration as described in the Methods Section.

On the basis of this analysis, MDG compounds were physically screened and images analysed in order to detect potential anti tubulin drugs or anti mitotic agents involving similar effect to the control MTAs.

5.5 5 Screening by HCA of MDG compounds for effect on the MT network

According to the validation study, the most powerful feature for separating active drugs from negative controls and separating drugs with "opposite effect" appeared to be the mean difference density detected in channel 3 (tubulin antibody emission). As a consequence of this caracteristic was chosen to rank all the MDGs as illustrated in Figure 5.15.

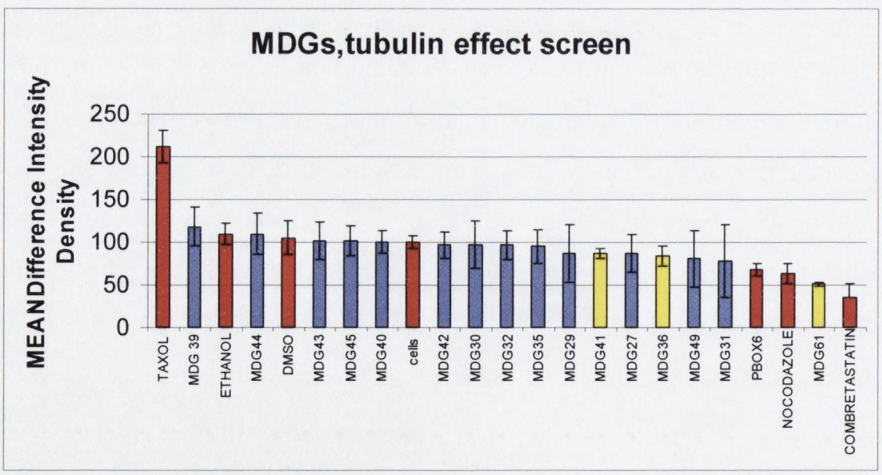


Figure 5.15: MDG Mean difference of tubulin stain intensity density.
All compounds were tested for 24 hours at 10μM concentration as described in the Methods Section.

This analysis was able to retrieve PBOX6 as the hidden active yet it detected only one compound as active: MDG61 (coloured in yellow) which was expected to behave as a destabiliser. Nonetheless, two compounds MDG41 and MDG36 (coloured in yellow) were showing a slight difference when compared to the negative controls yet with overlapping error bars. The full morphological profiles of these three compounds, including all tubulin-calculated features, were therefore plotted for further examination and are presented respectively in figure 5.16, 5.17 and 5.18.

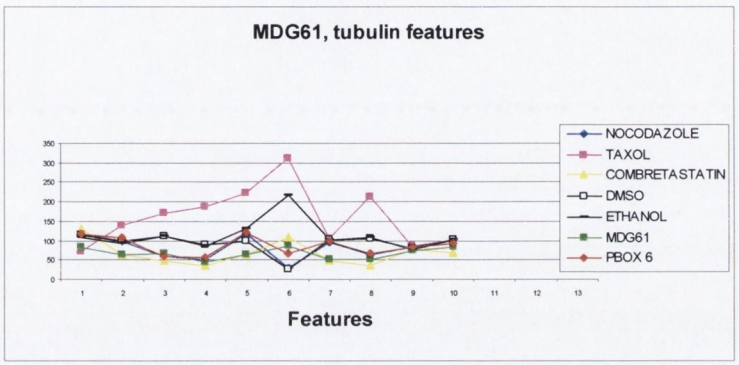


Figure 5.16: MDG61 tubulin morphological profile.

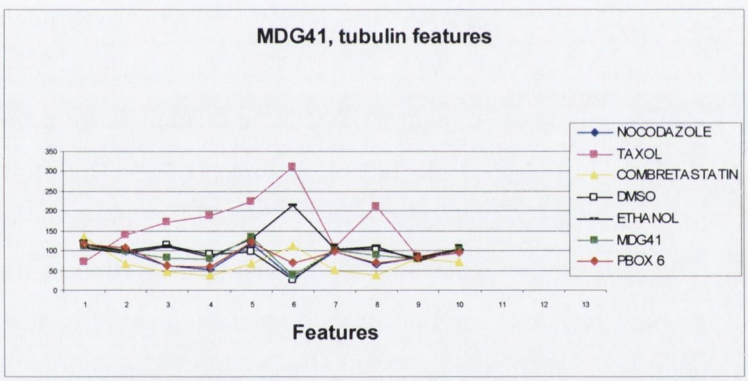


Figure 5.17: MDG41 tubulin morphological profile.

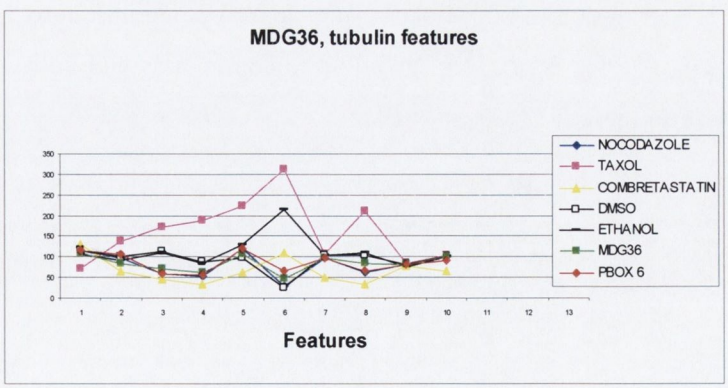


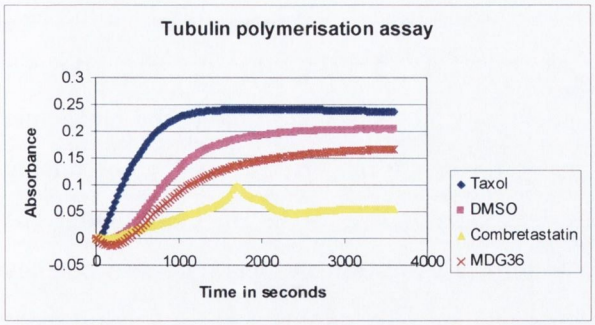
Figure 5.18: MDG36 tubulin morphological profile.

The features plotted in the three presented profiles are numbers as shown in table 5.7.

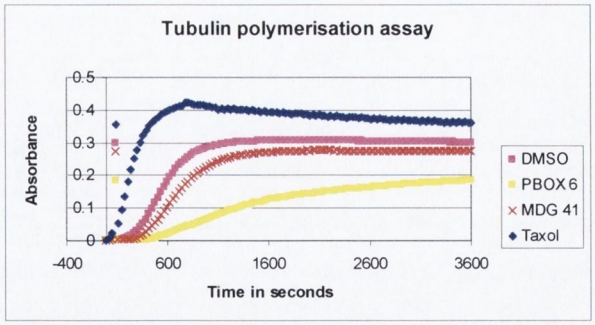
Feature	No
MEAN Fibre Total Area	1
MEAN Fibre Alignment	2
MEAN Average Intensity	3
MEAN Variance Intensity	4
MEAN Skew Intensity	5
MEAN Kurtosis Intensity	6
MEAN Entropy Intensity	7
MEAN Difference Intensity Density	8
MEAN Intensity co-occurrences Max	9
MEAN Average Radial Intensity	10

Table 5.10: Calculated cytoskeleton morphological features

MDG61 follows the combretastatin trend amongst all features with the exception of the first feature, the fibre total area whilst MDG41 and MDG36 show similar effects to the nocodazole and PBOX6 trends. The difference between these two trends is found in one particular feature: the mean skew intensity (feature 5). The two vehicle controls used, DMSO and ethanol, are actually also different to the untreated cells for this same particular feature. Combretastatin, nocodazole and PBOX6 were dissolved in ethanol while Taxol was dissolved in DMSO. The control drugs follow as well the trends taken by their respective vehicle, with Taxol highly accentuating the variation. Therefore, this feature variation was attributed to solvent effects and no further conclusion could be made. In this light, two MDGs were selected for further investigations. The first test undertaken was an "on-target" validation for their suspected destabilizing effect on the MT network. To do so, a tubulin polymerisation assay was perfomed at one concentration of 20μM and results are presented in Figure 5.19. It was decided to choose MDG36 and MDG41 for the final assay as these compounds had both been shown to induce apoptosis in CML cell (*cf.* Figure 5.12).



(a)



(b)

Figure 5.19: Tubulin polymerisation assay.

(a) MDG36 results representative of three separate experiments, (b) MDG41 results representative of two separate experiments. Taxol, combretastatin, PBOX6 were tested at 10μM and MDG compounds at 20μM. Tubulin polymerisation was determined by measuring the increase in absorbance over time at 340 nm.

Tubulin polymerisation assay showed that MDG41 and MDG36 as destabiliser agents in the micromolar range yet these compounds were not as potent as PBOX6. As these compounds were first pass vHTS hits, this is a particularly encouraging result. Even if their potencies do not appear to be as good as expected, these new scaffolds could be used as templates for lead optimisation which should deliver compounds with enhanced potencies.

5.6. Conclusion

This Section discussed the *in vitro* biological work undertaken to assess both known and potentially novel apoptotic agent and MTAs. A novel low cost and fast screening method was successfully developed for these identifications. The HCS using cell cycle analysis was able to classify and give a comprehensive rank of PBOX compound potencies. The new information issued from these *in vitro* studies, permitted an assessment of a structure activity relationship for PBOX compounds which was developed and used in chapter 3. Furthermore, currently, collaborating groups are using these new findings to further investigate apoptotic pathways in the context of cancer cell drug resistance. For all screened compounds, changes of cell cycle profiles were determined which indicated possible apoptotic effects following G2/M arrest pathway. In the future it would be desirable to optimise, both biological and analysis tools, to detect the pre-G1 peak. This peak could give a supplementary indication of apoptosis. It is commonly measured by FACS analysis where the small fragments of DNA which contain less than the normal DNA content of a cell resulting from apoptosis, are detected.

The HCS assay developed for cytoskeletal analysis was also successful to screen libraries of compounds. This assay based on tubulin staining was not only able to discern between active and inactive molecules but was also able to classify the actives according to their specific effect on the MT

network. In total, 17 MDG compounds showed activity on breast cancer cell lines, of which the best were MDG31, MDG36, MDG42, MDG 43, MDG 44, MDG 45, MDG49. In addition, MDG42, MDG44 and MDG45 also demonstrated potency at inducing apoptosis in one of the tested leukaemia cell lines (LAMA84). MDG 61 appeared to be less active on breast cancer cell lines but promising effects should be expected on tubulin binding as it was detected as active by the HCS on tubulin. The identification of tubulin as the target of MDG61 needs further validation. MDG41 showed potency over the breast cancer cell line and destabilising effect on tubulin polymerisation assay but still more work needs to be completed on target. Finally, the method developed here, identified one potent active denoted MDG36 which showed potency in arresting breast cancer cell in G2/M and induced apoptosis on two resistant leukaemia cell lines (LAMA84 and K562). Moreover this compound was suggested to bind tubulin and to destabilise the MT network suggesting a similar apoptotic pathway to PBOX6, one of the templates used to generate the scaffold in Chapter 3.

In a departure from our primary objective, the parasite responsible for sleeping sickness (Trypanosoma) consists of a unique cell with a plasma membrane internally coated by tubulin. This parasite is resistant to several drugs including antitubulin agents. All compounds which were tested as inactives by the diverse set of the experiments were then retained for potency investigation on parasites. Indeed, as these molecules did not kill human cells, they could be useful in the context of parasite-based illnesses. Therefore, as the tested compounds were designed to induce apoptosis and potentially bind tubulin, they were tested for anti-proliferative potency against trypanosomes. This drug recycling step is currently undergoing further examination. Yet, three drugs have shown promising results as anti-parasite drugs from this hypothesis.

References

1. Mc Gee, M. M.; Campiani, G.; Ramunno, A.; Fattorusso, C.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Pyrrolo-1,5-benzoxazepines induce apoptosis in chronic myelogenous leukemia (CML) cells by bypassing the apoptotic suppressor bcr-abl. *J Pharmacol Exp Ther* **2001,** 296, (1), 31-40.
2. Mulligan, J. M.; Campiani, G.; Ramunno, A.; Nacci, V.; Zisterer, D. M., Inhibition of G1 cyclin-dependent kinase activity during growth arrest of human astrocytoma cells by the pyrrolo-1,5-benzoxazepine, PBOX-21. *Biochim Biophys Acta* **2003,** 1639, (1), 43-52.
3. Mc Gee, M. M.; Hyland, E.; Campiani, G.; Ramunno, A.; Nacci, V.; Zisterer, D. M., Caspase-3 is not essential for DNA fragmentation in MCF-7 cells during apoptosis induced by the pyrrolo-1,5-benzoxazepine, PBOX-6. *FEBS Lett* **2002,** 515, (1-3), 66-70.
4. Zisterer, D. M.; Campiani, G.; Nacci, V.; Williams, D. C., Pyrrolo-1,5-benzoxazepines induce apoptosis in HL-60, Jurkat, and Hut-78 cells: a new class of apoptotic agents. *J Pharmacol Exp Ther* **2000,** 293, (1), 48-59.
5. Zisterer, D. M.; McGee, M. M.; Campiani, G.; Ramunno, A.; Fattorusso, C.; Nacci, V.; Lawler, M.; Williams, D. C., Pyrrolo-1,5-benzoxazepines: a new class of apoptotic agents. *Biochem Soc Trans* **2001,** 29, (Pt 6), 704-6.
6. Greene, L. M.; Fleeton, M.; Mulligan, J.; Gowda, C.; Sheahan, B. J.; Atkins, G. J.; Campiani, G.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., The pyrrolo-1,5-benzoxazepine, PBOX-6, inhibits the growth of breast cancer cells in vitro independent of estrogen receptor status and inhibits breast tumour growth in vivo. *Oncol Rep* **2005,** 14, (5), 1357-63.
7. Mc Gee, M. M.; Greene, L. M.; Ledwidge, S.; Campiani, G.; Nacci, V.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Selective induction of apoptosis by the pyrrolo-1,5-benzoxazepine 7-[[dimethylcarbamoyl]oxy]-6-(2-naphthyl)pyrrolo-[2,1-d] (1,5)-benzoxazepine (PBOX-6) in Leukemia cells occurs via the c-Jun NH2-terminal kinase-dependent phosphorylation and inactivation of Bcl-2 and Bcl-XL. *J Pharmacol Exp Ther* **2004,** 310, (3), 1084-95.
8. McGrath, L. B.; Onnis, V.; Campiani, G.; Williams, D. C.; Zisterer, D. M.; Mc Gee, M. M., Caspase-activated DNase (CAD)-independent oligonucleosomal DNA fragmentation in chronic myeloid leukaemia cells; a requirement for serine protease and Mn(2+)-dependent acidic endonuclease activity. *Apoptosis* **2006,** 11, (9), 1473-87.
9. Scolnick, D. M.; Halazonetis, T. D., Chfr defines a mitotic stress checkpoint that delays entry into metaphase. *Nature* **2000,** 406, (6794), 430-5.
10. Burke, D. J., Complexity in the spindle checkpoint. *Curr Opin Genet Dev* **2000,** 10, (1), 26-31.
11. Hoyt, M. A., Exit from mitosis: spindle pole power. *Cell* **2000,** 102, (3), 267-70.
12. Amon, A., The spindle checkpoint. *Curr Opin Genet Dev* **1999,** 9, (1), 69-75.
13. Wassmann, K.; Benezra, R., Mitotic checkpoints: from yeast to cancer. *Curr Opin Genet Dev* **2001,** 11, (1), 83-90.
14. Rieder, C. L.; Schultz, A.; Cole, R.; Sluder, G., Anaphase onset in vertebrate somatic cells is controlled by a checkpoint that monitors sister kinetochore attachment to the spindle. *J Cell Biol* **1994,** 127, (5), 1301-10.
15. Sorger, P. K.; Dobles, M.; Tournebize, R.; Hyman, A. A., Coupling cell division and cell death to microtubule dynamics. *Curr Opin Cell Biol* **1997,** 9, (6), 807-14.
16. Donaldson, K. L.; Goolsby, G. L.; Kiener, P. A.; Wahl, A. F., Activation of p34cdc2 coincident with taxol-induced apoptosis. *Cell Growth Differ* **1994,** 5, (10), 1041-50.
17. Lockshin, R. A.; Zakeri, Z., Programmed cell death and apoptosis: origins of the theory. *Nat Rev Mol Cell Biol* **2001,** 2, (7), 545-50.
18. O'Connor, P. M.; Kohn, K. W., A fundamental role for cell cycle regulation in the chemosensitivity of cancer cells? *Semin Cancer Biol* **1992,** 3, (6), 409-16.
19. Wyllie, A. H.; Kerr, J. F.; Currie, A. R., Cell death: the significance of apoptosis. *Int Rev Cytol* **1980,** 68, 251-306.
20. Kerr, J. F.; Wyllie, A. H.; Currie, A. R., Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer* **1972,** 26, (4), 239-57.
21. Tice, R. R., The cytogenetic evaluation of in vivo genotoxic and cytotoxic activity using rodent somatic cells. *Cell Biol Toxicol* **1988,** 4, (4), 475-86.
22. Fenech, M., In vitro micronucleus technique to predict chemosensitivity. *Methods Mol Med* **2005,** 111, 3-32.

23. Yamada, K. M.; Araki, M., Tumor suppressor PTEN: modulator of cell signaling, growth, migration and apoptosis. *J Cell Sci* **2001,** 114, (Pt 13), 2375-82.
24. Kiger, A. A.; Baum, B.; Jones, S.; Jones, M. R.; Coulson, A.; Echeverri, C.; Perrimon, N., A functional genomic analysis of cell morphology using RNA interference. *J Biol* **2003,** 2, (4), 27.
25. Wheeler, D. B.; Bailey, S. N.; Guertin, D. A.; Carpenter, A. E.; Higgins, C. O.; Sabatini, D. M., RNAi living-cell microarrays for loss-of-function screens in Drosophila melanogaster cells. *Nat Methods* **2004,** 1, (2), 127-32.
26. Kau, T. R.; Schroeder, F.; Ramaswamy, S.; Wojciechowski, C. L.; Zhao, J. J.; Roberts, T. M.; Clardy, J.; Sellers, W. R.; Silver, P. A., A chemical genetic screen identifies inhibitors of regulated nuclear export of a Forkhead transcription factor in PTEN-deficient tumor cells. *Cancer Cell* **2003,** 4, (6), 463-76.
27. Gasparri, F.; Mariani, M.; Sola, F.; Galvani, A., Quantification of the proliferation index of human dermal fibroblast cultures with the ArrayScan high-content screening reader. *J Biomol Screen* **2004,** 9, (3), 232-43.
28. Bjorklund, M.; Taipale, M.; Varjosalo, M.; Saharinen, J.; Lahdenpera, J.; Taipale, J., Identification of pathways regulating cell size and cell-cycle progression by RNAi. *Nature* **2006,** 439, (7079), 1009-13.
29. US Food and Drug Administration, US. department of Health and Human Service,. *http://www.fda.gov/*.
30. US Food and Drug Administration, EC50 *vs.*IC50. *http://www.fda.gov/ohrms/dockets/ac/00/slides/3621s1d/sld036.htm*.
31. Array Scan KSR Reader, www.cellomics.com. *Pittsburgh, PA, USA*.
32. Mulligan, J. M.; Greene, L. M.; Cloonan, S.; Mc Gee, M. M.; Onnis, V.; Campiani, G.; Fattorusso, C.; Lawler, M.; Williams, D. C.; Zisterer, D. M., Identification of tubulin as the molecular target of proapoptotic pyrrolo-1,5-benzoxazepines. *Mol Pharmacol* **2006,** 70, (1), 60-70.
33. Wolff, A.; de Nechaud, B.; Chillet, D.; Gros, F., The aminoterminal domain of alpha-tubulin probed by monoclonal antibodies. Recognition of a rarely exposed epitope by the monoclonal antibody 111 B52 C2. *Biol Cell* **1988,** 63, (3), 319-26.
34. O'Brien P, J.; Irwin, W.; Diaz, D.; Howard-Cofield, E.; Krejsa, C. M.; Slaughter, M. R.; Gao, B.; Kaludercic, N.; Angeline, A.; Bernardi, P.; Brain, P.; Hougham, C., High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. *Arch Toxicol* **2006,** 80, (9), 580-604.
35. Kevorkov, D.; Makarenkov, V., Statistical analysis of systematic errors in high-throughput screening. *J Biomol Screen* **2005,** 10, (6), 557-67.
36. GraphPad Software Inc., GraphPad Prism. **2003,** version4 for windows.
37. Cytoskeleton, *http://www.cytoskeleton.com/*.
38. Mulligan, J., Investigation into the mechanism underlying the antiproliferative and apoptotic effects of a novel series of pyrrolo-1,5-benzoxazepine compounds. *PhD, Trinity College Dublin* **2004**.
39. Curman, D.; Cinel, B.; Williams, D. E.; Rundle, N.; Block, W. D.; Goodarzi, A. A.; Hutchins, J. R.; Clarke, P. R.; Zhou, B. B.; Lees-Miller, S. P.; Andersen, R. J.; Roberge, M., Inhibition of the G2 DNA damage checkpoint and of protein kinases Chk1 and Chk2 by the marine sponge alkaloid debromohymenialdisine. *J Biol Chem* **2001,** 276, (21), 17914-9.

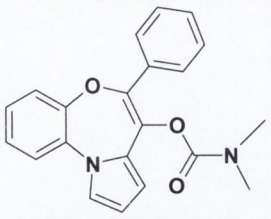
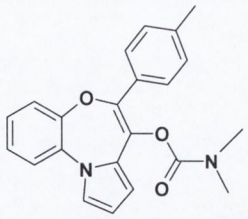
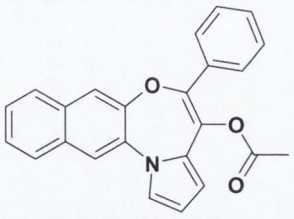
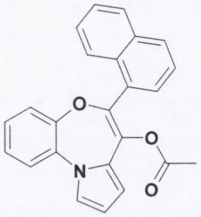
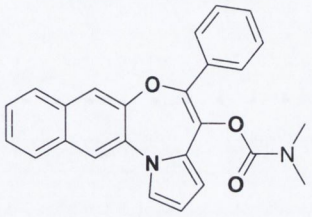
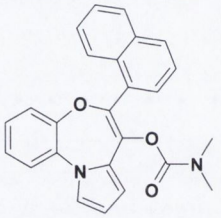
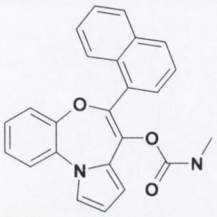
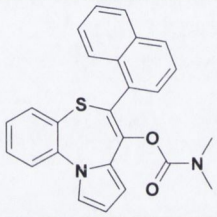
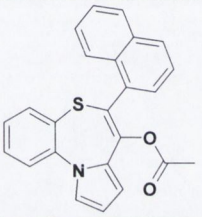
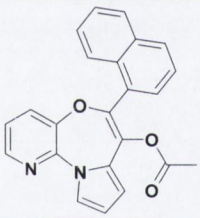
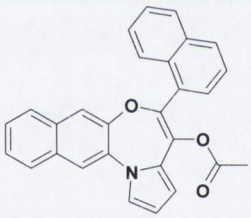
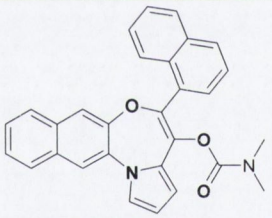
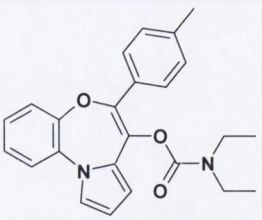
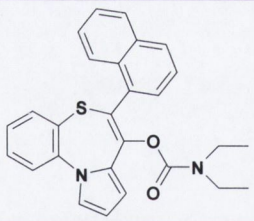
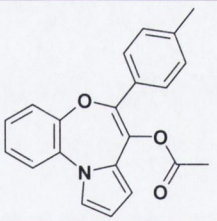
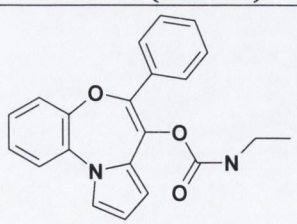
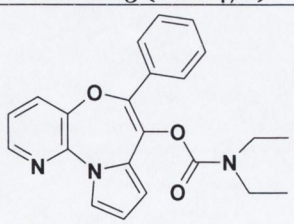
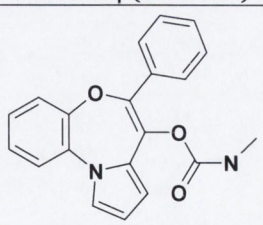
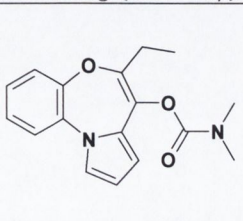
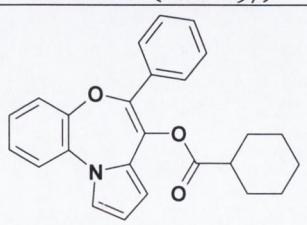
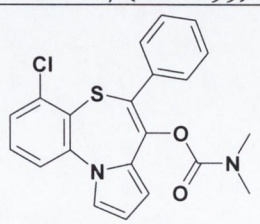
General conclusion

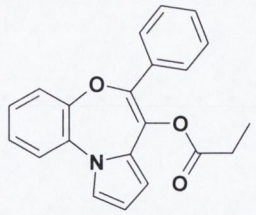
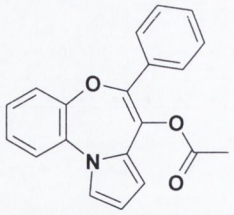
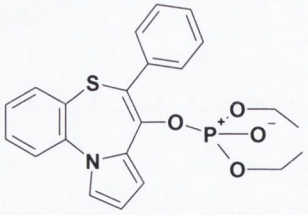
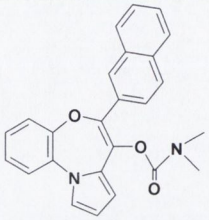
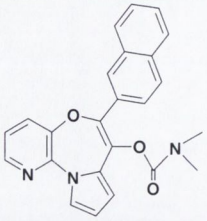
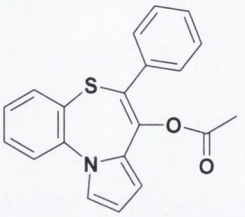
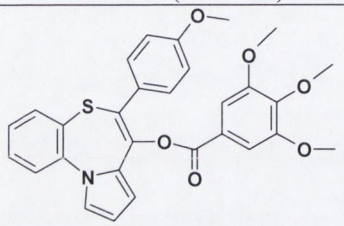
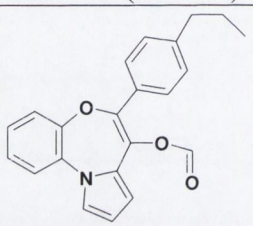
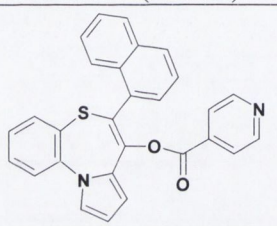
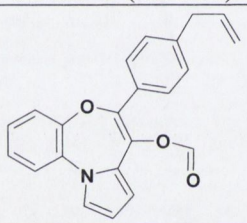
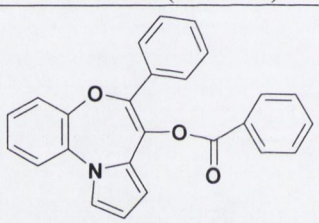
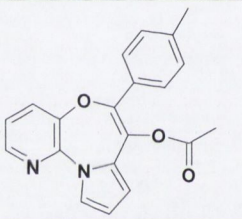
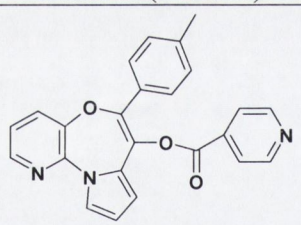
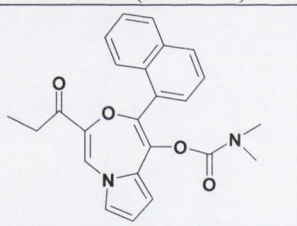
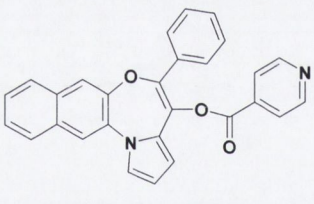
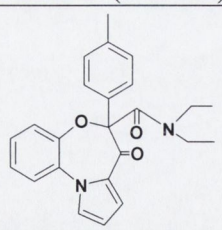
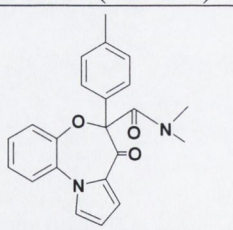
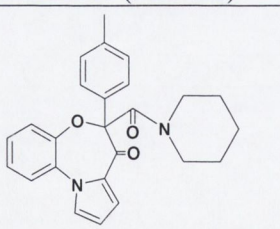
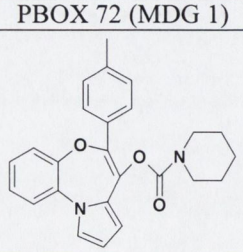
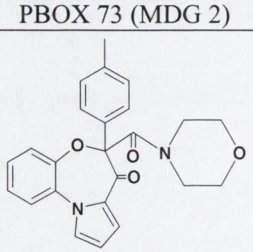
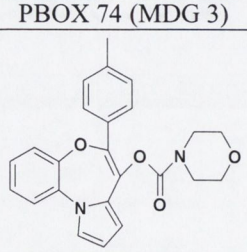
In the presented work, the most successful similarity method search retreived the estrogen receptor β isoform and PDGFR as two potential novel targets for PBOX compounds, suggesting PBOX compounds application for the research in breast cancer and other cancers associated with PDGFR activities. The *in vitro* validation of these two novel potential molecular targets is currently being assessed.

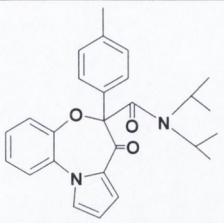
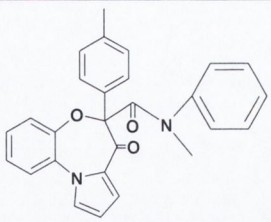
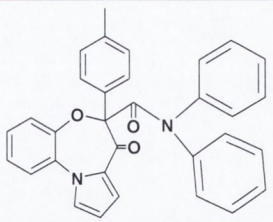
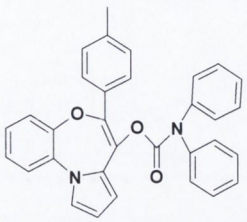
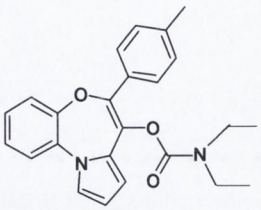
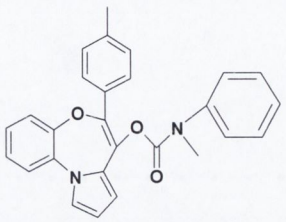
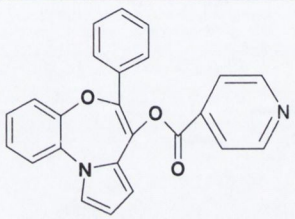
In addition, the novel computational method developed to correlate tubulin binders with potential binding sites on tubulin, suggested that PBOX compounds bind in the nucleotidic site. This binding site has not yet been explored as a potential small drug molecules interaction region and could permit the discovery of new anti-mitotic agent as well as direct the design of optimised PBOX related structures enhancing their potency. This lead optimisation programme will be followed up as a new research project within our team.

The discovery of new leukaemia chemotherapeutic agents has been shown to be more efficient and complete if cancer medicinal chemical space beyond the traditional "drug-like" and "lead-like" space subsets are mapped and explored. Following this rationale, the VS campaign undertaken led to the discovery of novel hits which were tested *in vitro* through a novel low cost and fast HCS method. In total, 17 novel compounds showed activity on breast cancer cell lines and five compounds demonstrated potency at inducing apoptosis in leukaemia cell lines. In addition, two compounds showed a destabilising effect on microtubular networks suggesting a similar mechanistic pathway to PBOX compounds. Compounds tested as inactive on human cancer cell lines were tested for anti-proliferative activity on the trypanosome parasite responsible for sleeping illness. Tree compounds showed promising results and are at the present time forming the basis of an HRB (Health Research Board) grant application.

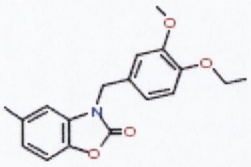
Appendix I: PBOX structures

		
PBOX 1 (MDG 14)	PBOX 2 (NFS 213)	PBOX 3 (NFS 223)
		
PBOX 4 (NFS 221)	PBOX 5 (NFS 224)	PBOX 6 (NFS 225)
		
PBOX 7 (NFS325)	PBOX 8 (468(3))	PBOX 9 (469(4))
		
PBOX 12 (NFS 470)	PBOX 15 (NFS 502)	PBOX 16 (NFS 504)
		
PBOX 21 (MDG 11)	PBOX 23 (NFS 478)	PBOX 24 (NFS 212)
		
PBOX 25 (NFS 267)	PBOX 26 (NFS 297)	PBOX 27 (NFS 299)
		
PBOX 31 (NFS 240)	PBOX 32 (NFS 199)	PBOX 33 (NFS 149)

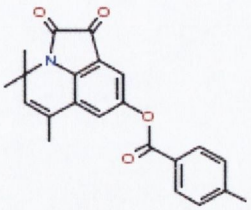
		
PBOX 34 (NFS 181)	PBOX 35 (NFS 260)	PBOX 38 (NFS 198)
		
PBOX 42 (NFS 466)	PBOX 43 (NFS 467)	PBOX 46 (NFS 64)
		
PBOX 48 (NFS 87)	PBOX 65a (NFS 205)	PBOX 65b (NFS 503)
		
PBOX 66 (NFS 206)	PBOX 67 (NFS 576)	PBOX 68 (NFS 539)
		
PBOX 69 (NFS 575)	PBOX 70 (NFS 579)	PBOX 71 (NFS 578)
		
PBOX 72 (MDG 1)	PBOX 73 (MDG 2)	PBOX 74 (MDG 3)
		
PBOX 75 (MDG 4)	PBOX 76 (MDG 5)	PBOX 77 (MDG 6)

		
PBOX 78 (MDG 7)	PBOX 79 (MDG 8)	PBOX 80 (MDG 9)
		
PBOX 81 (MDG 10)	PBOX 82 (MDG 11)	PBOX 83 (MDG 12)
		
PBOX 84 (MDG 13)		

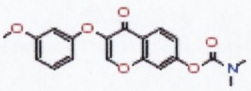
Appendix II: MDG structures



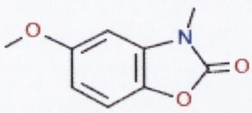
MDG20



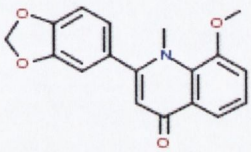
MDG21



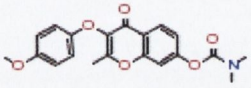
MDG22



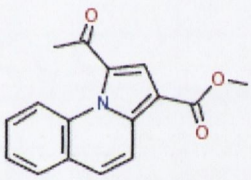
MDG23



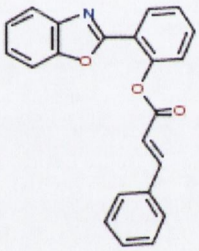
MDG24



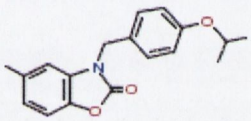
MDG25



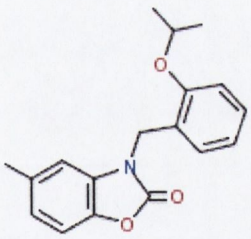
MDG26



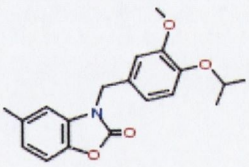
MDG27



MDG28

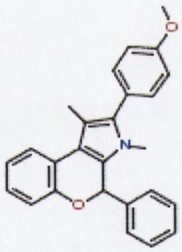


MDG29



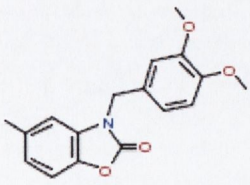
MDG30

Chiral

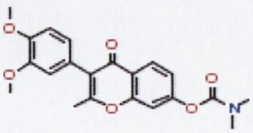


MDG31

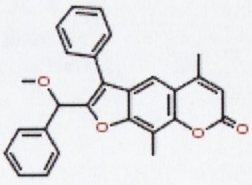
Chiral



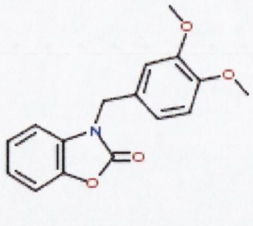
MDG32



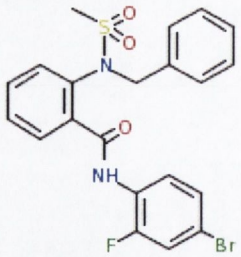
MDG33



MDG34

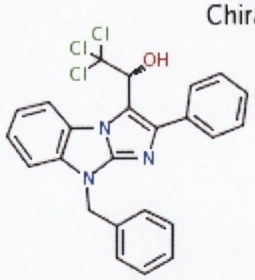


MDG35

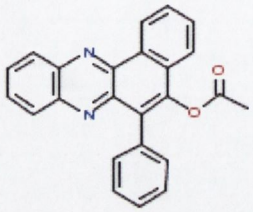


MDG36

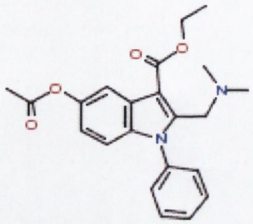
Chiral



MDG37

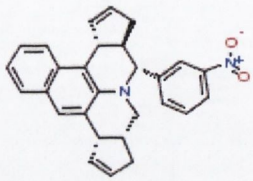


MDG38

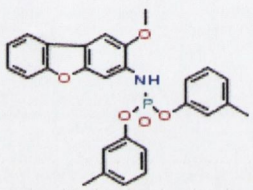


MDG39

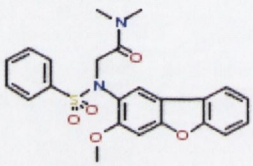
Chiral



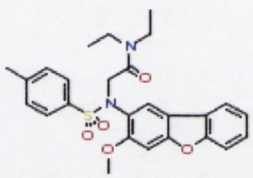
MDG40



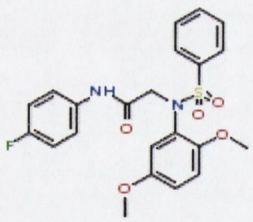
MDG41



MDG42

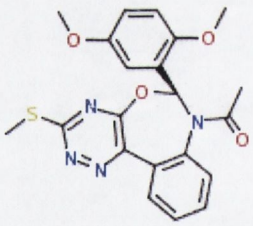


MDG43



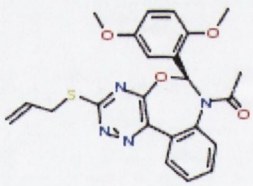
MDG44

Chiral



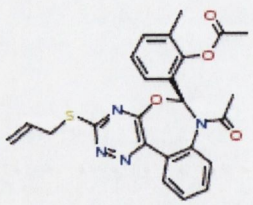
MDG45

Chiral

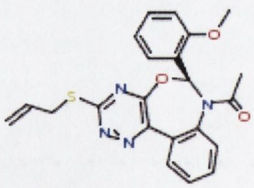


MDG46

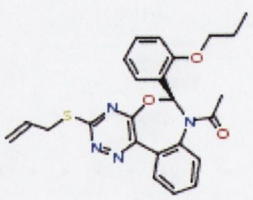
Chiral



MDG47

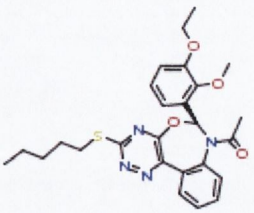


MDG48



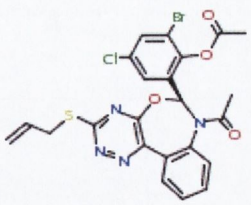
MDG49

Chiral



MDG50

Chiral



MDG61

ELSEVIER

Oncology exploration: charting cancer medicinal chemistry space

David G. Lloyd[1], Georgia Golfis[1], Andrew J.S. Knox[2], Darren Fayne[1], Mary J. Meegan[2] and Tudor I. Oprea[3]

[1]Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland
[2]School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland
[3]Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA

Approaches for the experimental determination of protein–ligand molecular interactions are reliant on the quality of the compounds being tested. The application of large, randomly designed combinatorial libraries has given way to the creation of more-focused 'drug-like' libraries. Prior to synthesis, we wish to screen the potential compounds to remove undesired chemical moieties and to be within a required range of physiochemical properties. We have used a principal-component analysis (PCA) computational approach to analyze the 3D descriptor space of active and non-active (hit-like) cancer medicinal chemistry compounds. We define hit-like those molecules passing the unmodified OpenEye FILTER program. Our analysis indicates that these compounds occupy quite different regions in space. Cancer-active compounds exist in a much greater volume of space than generic hit-like space and most of them fail the commonly applied filters for orally bioavailable drugs. This is of great significance when designing orally bioavailable cancer target drugs.

The successful application of the processes of virtual and physical screening for active ligands is totally reliant on the quality of the molecules being screened. In the simplest terms, if there are no hits in the database or compound library, there is no point in performing the screen. Recent years have seen great advances in our understanding of what makes a molecule drug- or lead-like and cheminformatic treatment of screening collections has focused the attention of discovery research towards drug and lead chemical space. However, when dealing with oncology, the applied and trusted rules of engagement do not always apply.

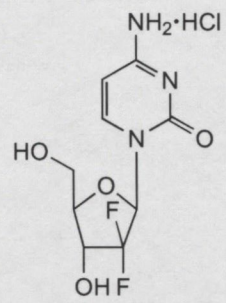
Partitioning and classifying cancer medicinal chemistry space is not straightforward. A multitude of active cancer ligands, containing different molecular scaffolds, have been identified for the relatively small number of cancer targets. The past two decades have witnessed a tremendous increase in our understanding of the pathology and molecular biology of human cancers [1]. Although enormous progress has been made in the development and identification of new molecular medicines and targets in this area, many of the current clinical treatments for cancers have limitations with respect to efficacy, resistance and toxicity in the patient [2]. There is much scope for the exploitation of the many new molecular targets to develop new cancer treatments with improved specificity, toxicology profiles and efficacy [3].

Cancer chemotherapeutics

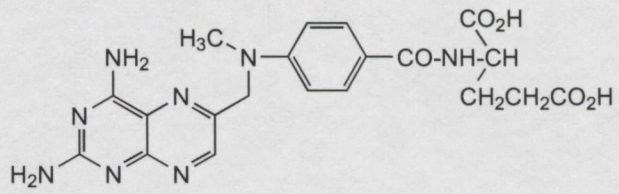
The bulk of existing cancer chemotherapeutic drugs causes cell death by several different mechanisms – the majority by non-selectively targeting the cellular processes that cancers utilize to rapidly grow and divide (i.e. the ability to replicate their DNA) [4]. In general, dividing tumor cells have lost the ability to respond to environmental cues, which in normal cells would control physiological functions such as cell division. Normal cells exhibit a higher basal resistance to chemotherapeutic drugs, whereas rapidly dividing cells, such as bone marrow and intestinal mucosa, are highly susceptible to them and severe toxic side effects are common [5]. The discovery of defects in oncogenes has allowed the development of exciting anticancer therapeutics that can selectively target tumor cells and specific tumor biochemical processes,

Corresponding author: Lloyd, D.G. (lloyddg@tcd.ie).

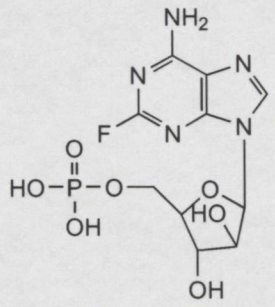
1359-6446/06/$ – see front matter ©2006 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03688-3



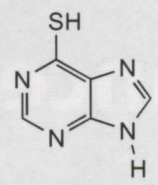
Gemcitabine



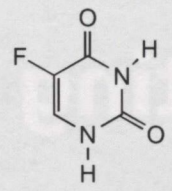
Methotrexate



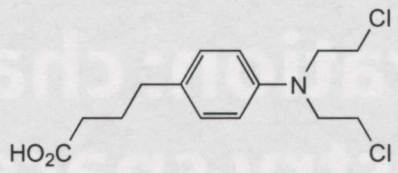
Fludarabine-phosphate



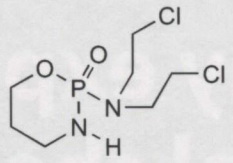
6-Mercaptopurine



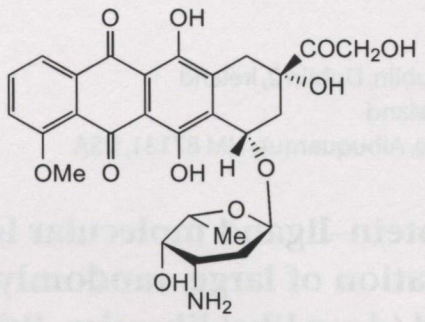
5-Fluorouracil



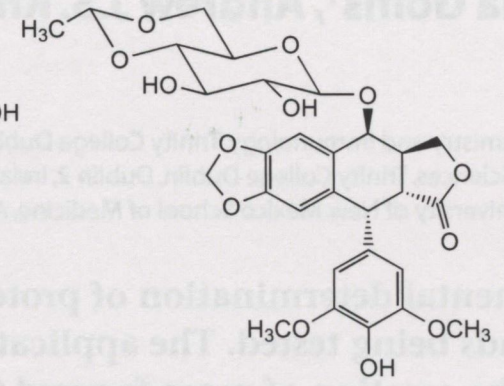
Chlorambucil



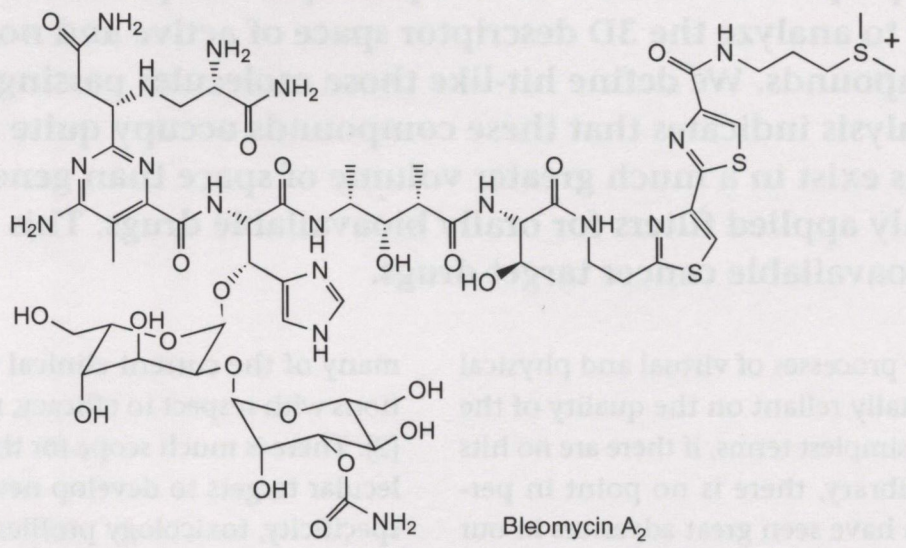
Cyclophosphamide



Doxorubicin



Etoposide



Bleomycin A_2

FIGURE 1

Chemical structures of non-selective chemotherapy drugs acting through different mechanisms. A selection of genotoxic and antimetabolic drugs are depicted, illustrating the great variety in structural form of anticancer ligands.

avoiding the cellular toxicity associated with conventional cancer chemotherapeutic drugs [6]. Knowledge of the specific biochemical differences between normal and cancer cells is growing at an exponential rate and this can be potentially exploited for cancer chemotherapy. The recent development of some FDA-approved treatments that target cancer-specific cellular processes demonstrate the utility of these novel approaches [7].

To appreciate the breadth of cancer-related chemical space, we first examine some of the major mechanistic groupings within this space by looking at various classes of oncology therapeutics, with a view to mapping their relative locations in terms of the medicinal chemical space they occupy.

Antimetabolites

Antimetabolites, such as methotrexate, 6-mercaptopurine and gemcitabine (Figure 1), were developed to interfere with specific enzymatic steps in nucleotide biosynthesis in tumor cells. Methotrexate is an inhibitor of dihydrofolate reductase (DHFR) and, as such, limits the formation of nucleotides and thus inhibits DNA replication. It has a significant role in the treatment of breast

cancer, acute lymphocyte leukemia and lymphomas [8]. Purine antagonists, such as 6-mercaptopurine and fludarabine-phosphate (Figure 1), function by inhibiting DNA production either by blocking the production of the required purines or by incorporating themselves into the growing DNA as false nucleotides through DNA polymerase. 6-Mercaptopurine can be given orally or intravenously and is useful in the treatment of acute lymphocytic and myelogenous leukaemias [9]. Similarly, pyrimidine antimetabolites [e.g. 5-fluorouracil and gemcitabine (Figure 1)] act by preventing the biosynthesis of the required pyrimidine nucleoside, hence inhibiting DNA replication. Gemcitabine is used clinically for the treatment of non-small cell lung cancer, whereas 5-fluorouracil is useful for a range of malignancies in tissues such as breast, colon, liver and skin [10].

Genotoxic drugs

Genotoxic drugs modify DNA and thus prevent accurate nucleic acid replication. Examples of widely used genotoxic chemotherapeutic drugs include alkylating agents [e.g. chlorambucil and cyclophosphamide that is activated by hepatic metabolism (Figure 1)], which alkylate and crosslink guanine bases, thus preventing DNA replication. Chlorambucil is used to treat a variety of cancers, including chronic lymphocytic leukemia, lymphomas and Hodgkin's disease. Cyclophosphamide is useful in the treatment of bladder, bone, cervical and lung cancers [11]. The platinum-containing coordination complexes, cisplatin, carboplatin and ormaplatin, act in a similar manner, by intra-strand crosslinking of the guanine bases of DNA, and are useful alone or as part of a combination regimen to treat a wide variety of neoplastic diseases [12].

Intercalating agents, such as doxorubicin and epirubicin (Figure 1), insert themselves in the minor groove between the nitrogen base pairs in the DNA double helix, causing a distortion in the helix shape and therefore interfering with DNA and RNA replication and transcription processes. Doxorubicin also inhibits topoisomerase type II, reducing the ability of the cell to repair breaks in the DNA. It is clinically useful for the treatment of a wide variety of cancers, such as breast and ovarian cancer, Hodgkin's and non-Hodgkin's lymphomas, testicular, bladder and lung cancers. Like doxorubicin, the natural product etoposide affects the DNA by specifically inhibiting the action of the enzyme topoisomerase II, which allows unwinding of the DNA in the normal process of replication by causing a break in the DNA strands. It is clinically useful for small cell lung cancer [13]. The cytotoxic antibiotic bleomycin A_2 (Figure 1) inhibits DNA synthesis by a mechanism of DNA chain cutting and so prevents cell replication. It is a useful agent for Hodgkin's and non-Hodgkin's lymphomas, together with squamous cell carcinomas.

Mechanism-based cancer therapies

The chemotherapeutic drugs outlined above are effective but relatively non-selective because they target fundamental biochemical processes, such as DNA and protein production. This indiscriminate action is one of the reasons for their severe toxic side effects. In recent years, there has been significant progress in the identification of cancer-specific cellular drug targets and in the design of drugs that selectively target cancer-specific cellular processes. These treatments should exploit the biochemical differences between normal and cancerous cells, resulting in drugs with greater potency and less toxic side effects [14].

Several oncogenes have been identified (e.g. *TP53*, *SRC*, *RAS*, *BCR-ABL*) which affect tumor growth and development, and new methods for oncogene identification are continually progressed [15]. Overexpression of specific gene products, such as epidermal growth factor (EGF) receptors and human epidermal growth factor receptor 2 (HER-2), can also be linked to the progression of some cancers. Significant developments in targeted treatments have been made in the areas of protein kinase inhibitors, farnesyl transferase inhibitors and matrix metalloproteinase inhibitors, as well as for agents targeting nuclear receptors implicated in disease progression.

Kinase inhibitors

The kinases represent a group of enzymes that regulate protein and cellular activity by phosphorylation. Intracellular signaling pathways that stimulate cell proliferation are frequently controlled by kinases [16]. In chronic myeloid leukemia, a specific chromosome (the Philadelphia chromosome) produces a damaged kinase fusion receptor protein BCR-ABL, which signals uncontrolled proliferation. Imatinib (Figure 2) is a tyrosine kinase inhibitor that binds to the BCR-ABL receptor, preventing ATP binding and resultant kinase activity [17]. It is approved by the FDA for the treatment of chronic myeloid leukemia and also for gastrointestinal tumors. EGF receptors are overexpressed on the surface of some lung and colon cancer cells. Activation of these receptors by EGF and TGF-α (transforming growth factor α) is crucial for tumor cell proliferation. Gefitinib (Figure 2) is a tyrosine kinase inhibitor that binds to the EGF receptor and inhibits the phosphorylation of the tyrosine residue by the tyrosine kinase enzyme. It is used clinically in the treatment of refractive non-small cell lung cancer. However, a recent randomized clinical trial, comparing the efficacy of gefitinib versus placebo after chemotherapy and radiation in patients with non-small cell lung cancer, which had spread only to nearby tissues or lymph nodes, was terminated by researchers when an interim data review indicated that gefitinib treatment would not improve survival (http://www.cancer.gov/newscenter/pressreleases/gefitinibNSCLC).

Proteasome inhibitors

Proteasomes function by controlling the processing of cellular proteins; for example they regulate proteins that control cell cycle and are involved in the degradation of damaged proteins. They are composed of several proteolytic enzymes. Inhibition of proteasome function can cause cell-cycle arrest and cell death [18]. Cancer cells are very susceptible to proteasome inhibitors because they divide rapidly and their normal control mechanisms do not function. Bortezomib (Figure 2) is used clinically in the treatment of multiple myeloma. It interacts with a threonine residue at the catalytic site of the proteasome, limiting metastasis and apoptosis, which are both regulated by proteasomal mechanisms.

Antimitotic agents

Antimitotic drugs interfere with the normal process of cell division. In mitosis, tubulin polymerizes to form spindle microtubules, which facilitate the separation of the two replicated chromosomes. Some drugs bind to tubulin monomers and inhibit the formation of microtubules (e.g. vincristine), whereas paclitaxel and epothilones have the opposite effect in that they stabilize the microtubules and

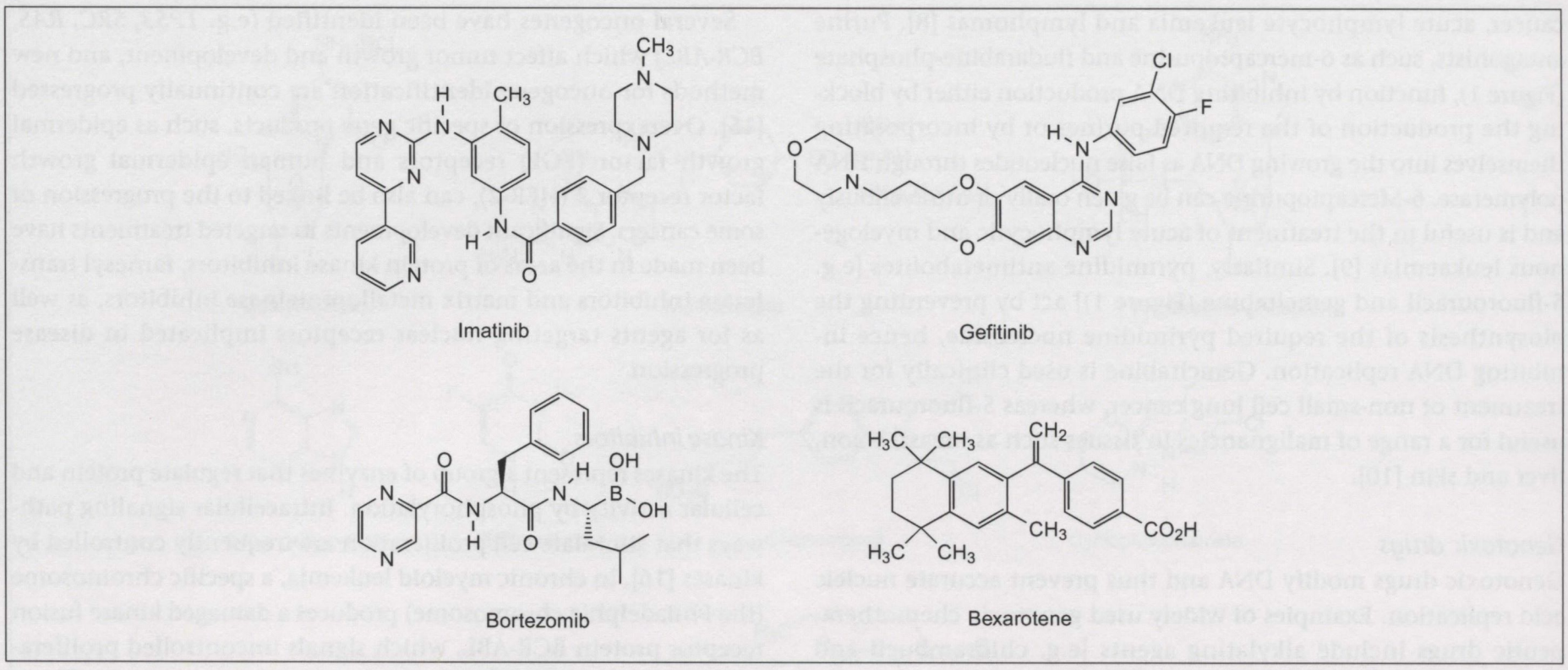


FIGURE 2

Chemical structures of small molecule cancer target selective inhibitors. Kinase and proteasome inhibitors bind to specific protein targets rather than to proteins involved in basic biochemical processes.

prevent the normal cell division process to be completed (Figure 3) [19]. Paclitaxel has several indications for various cancer chemotherapies, including combination regimens with cisplatin for lung and ovarian cancers and with the monoclonal antibody trastuzumab for breast cancer.

Hormonal agents

The proliferation of various breast, ovarian and prostate cancers are known to be dependent on signaling hormones, such as estrogen or testosterone; many of the drugs used in the adjuvant treatments of these cancers are effective by blocking either the biosynthetic production of the hormone or the activity of the hormone in the target cell [20].

Selective estrogen-receptor modulators

Drugs such as tamoxifen and toremiphene (Figure 3) act as estrogen receptor antagonists in breast tumor cells. Binding of these drugs causes a change in the ligand-binding domain conformation, clearly observed in X-ray crystallography data, that effectively blocks estrogen action in the cell, disrupting cell signaling processes and preventing cancer cell proliferation. Tamoxifen is used clinically for the adjuvant treatment of pre- and post-menopausal breast cancer and also for prevention in 'at risk' patients. The related selective estrogen-receptor modulator, raloxifene, is currently indicated for osteoporosis treatment. The compound was originally developed as keoxifene, for the treatment of tamoxifen-resistant breast cancer patients. Although effective in the treatment of estrogen receptor (ER)-dependent breast cancers, keoxifene was unable to meet clinical endpoints for the resistant indication. In contrast to tamoxifen, raloxifene has differential intrinsic actions at its target organs, acting as agonist at bone and as antagonist at the uterus.

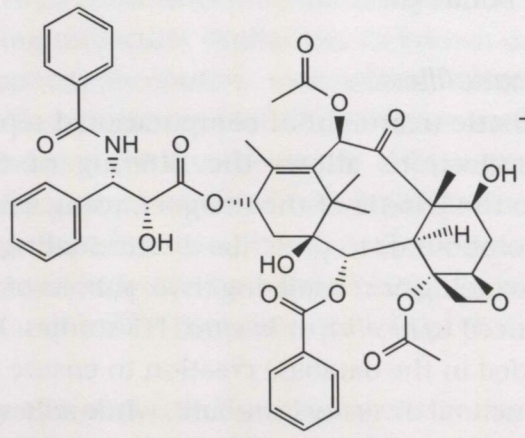
Fulvestrant is an estrogen receptor antagonist (pure antiestrogen) that binds reversibly to the ER and is used clinically to treat ER-positive metastatic postmenopausal breast cancers, with much continued interest in the development of structurally constrained analogues of this drug [21]. Megestrol is a synthetic analogue of progesterone and inhibits proliferation of hormonally sensitive metastatic breast and endometrial cancers (Figure 3). Recent studies have also highlighted the utility of this class of compounds in targeting GPR30, a transmembrane estrogen receptor identified in the endoplasmic reticulum and novel target in estrogen-dependent cancers. Although the physiological and pathological implications of GPR30 remain to be clarified, it is important to note that known antiestrogen agents such as tamoxifen and raloxifene are not estrogen-receptor-selective drugs because they act as agonists on GPR30 [22].

Selective androgen-receptor modulators

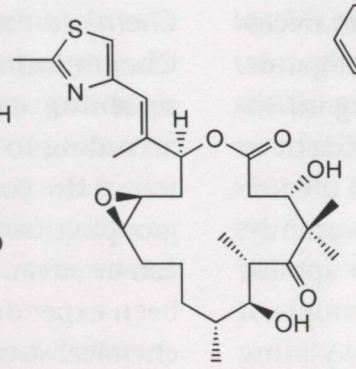
The androgen hormones, testosterone and dihydroxytestosterone, play a signaling role in the normal growth of prostate cells. These androgens also bind to the androgen receptors in prostate cancer cells, resulting in the proliferation of cancer cells. The antiandrogens flutamide, bicalutamide and nilutamide (Figure 3) block the action of testosterone by preventing the binding of testosterone to its receptors and therefore control the growth of prostate cancer cells [23]. Leutenising hormone releasing hormone analogues, such as the peptides goserelin and buserelin, which suppress the production of the steroid hormone testosterone, are routinely used in prostate cancer treatments and also for endometriosis [24].

Aromatase inhibitors

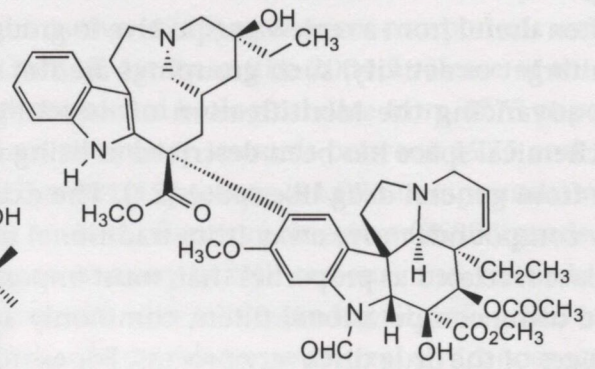
The biosynthesis of estrogen from androgen precursors can be prevented by blocking the action of the enzyme aromatase [25]. There are two types of aromatase inhibitors used clinically for postmenopausal ER-positive breast cancer, the steroidal aromatase inhibitor examestane and the nonsteroidal aromatase inhibitors, anastrazole and letrozole (Figure 3). Recent clinical trials with anastrozole suggest an increased role of this drug in the adjuvant treatment of the disease [26].



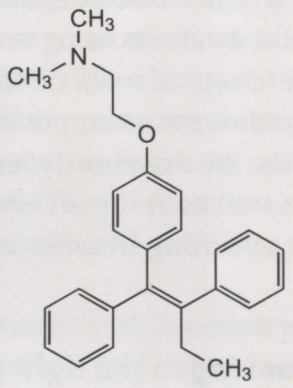
Paclitaxel



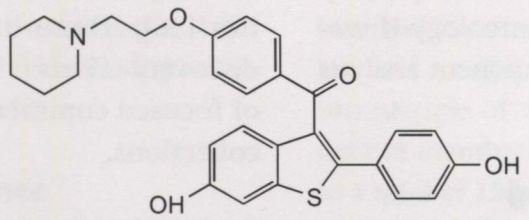
Epothilone A



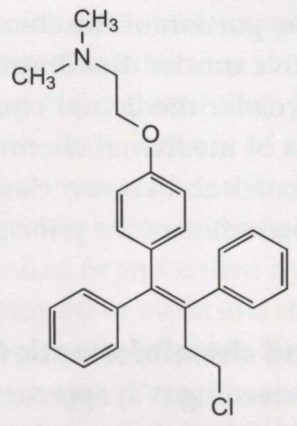
Vincristine



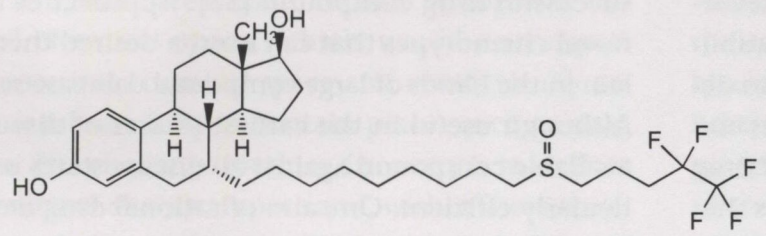
Tamoxifen



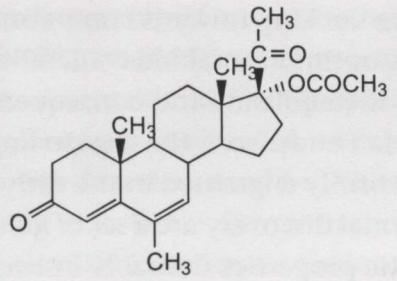
Raloxifene



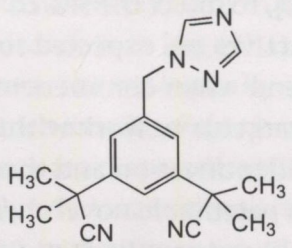
Toremiphene



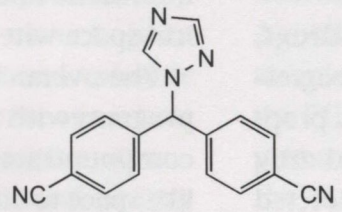
Fulvestrant, ICI 182780



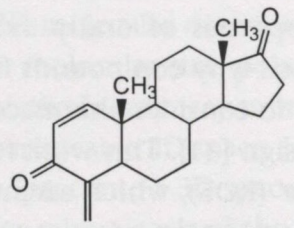
Megestrol



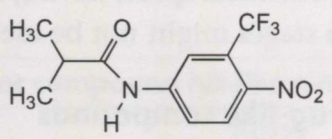
Anastrozole



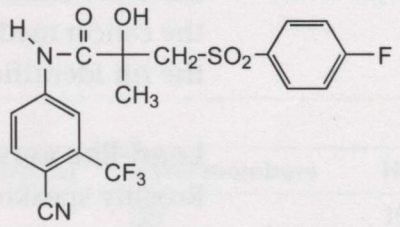
Letrozole



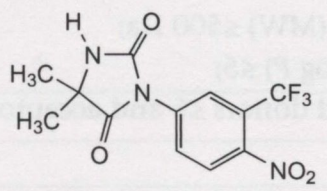
Exemestane



Flutamide



Bicalutamide



Nilutamide

FIGURE 3

Examples of antimitotic, hormonal, SERM, SARM and aromatase targeting ligands. These drugs specifically target different classes of cancer-related proteins.

Exploring chemical space: where are we looking, and what are we looking at?

Although it is often useful from a review perspective to group molecules based on target or activity, such groupings are not necessarily useful in advancing the identification of novel ligands. Cancer-related chemical space has been described as being intrinsically different from general drug-like space [27]. The extent to which oncology compounds move away from traditional medicinal chemistry space is related to properties that, most importantly, can be identified using computational filters, commonly applied at the earliest stages of the drug discovery process. For example, if significantly lower toxicity (e.g. moving away from alkylating agents) or improved oral bioavailability are desired (e.g. aromatase inhibitors, antiestrogens and antiandrogens), filters can be incorporated to a selection protocol to converge on molecules that occupy the appropriate portion of the chemical space. To attain a 'big picture' perspective on the distribution of cancer-targeting compounds in the broader medicinal chemical space, we have examined the regions of medicinal chemical space occupied by cancer-targeted molecules and known classes of oncology therapeutics through the application of the principal-component analysis (PCA) [28].

Virtual screening and cheminformatic filtering

An *in silico* or virtual screening (VS) approach helps to converge on possible active molecules from large molecular libraries and focus physical assaying on a smaller subset of compounds [29]. A developing area in VS is the use of computational methods that filter a molecular library before docking towards compounds with favorable pharmacokinetics, optimum oral bioavailability, compatibility with some types of metabolisms and consequently low toxicity [30]. As with all human endeavors, the urge to impose rules and structures on process is firmly engrained in the early phase of drug discovery. Framing rational discovery are a set of guiding rules that describe cheminformatic properties desirable in lead- or drug-like chemical scaffolds.

The 'rule of five'

Based on a study of the properties of orally available drugs, Lipinski's analysis of the reasons why compounds fail in progression highlighted the necessity to consider pharmacokinetic properties in compound library design [31]. This work furnished drug designers with the 'rule of five' (RO5), which essentially directed library creation and our consideration of screening collection diversity towards orally available drug-like space. Application of the RO5 would enforce the following properties on a screening collection:

(i) molecular weight (MW) ≤500 Da;
(ii) hydrophobicity (log *P*) ≤5;
(iii) number of H-bond donors ≤5 and acceptors ≤10.

The 'rule of three'

This rule set was proposed by Congreve *et al.* [32] and is designed for compliance in fragment-based drug discovery [33]. Application of the rule of three would enforce the following properties on a screening collection:

(i) MW <300 Da;
(ii) log *P* <3;
(iii) number of H-bond donors and acceptors <3;
(iv) flexible bonds <3.

Cheminformatic filters

Cheminformatic treatment of computational representations of screening collections allows the filtering of the collections according to the criteria of the designer, using calculable properties of the compounds to prescribe discriminating parameters for grouping, excluding or considering those subsets of the dataset that can be advanced to *in silico* or *in vitro* HTS studies. Much effort has been expended in the database creation to ensure that maximum chemical structural diversity is in-built, while adhering to the guiding rules of drug- and lead-like chemical properties [34,35].

Popular software utilities, such as FILTER (OpenEye), MOE (Chemical Computing Group) and the Daylight Toolkit (Daylight Chemical Information Systems) readily enable users to apply rapid partitioning of compound databases using tunable cheminformatic parameters based on the 'classical' rules, often enhanced with functionality to recognize and remove compounds with toxic properties [36]. These utilities are frequently employed in internal discovery efforts [37], as well as in the presentation and marketing of focused commercial screening libraries and non-targeted HTS collections.

Rational screening: looking in the right places

In drug discovery, it's fairly well established that a biological HTS campaign, which looks at ~1,000,000 compounds, will deliver viable lead structures that are ultimately progressed to furnish one successful drug compound [38]. Approaches to the discovery of novel chemotypes that can exert a desired therapeutic effect have lain in the hands of large compound database screening campaigns. Although useful in the earliest phases of discovery, testing every available compound against all disease states using HTS is not particularly efficient. One aim of rational drug design is to focus and limit the breadth of screening that must be undertaken to reach this goal. Effective preselection of screening candidates is crucial, usually employing computation means, namely VS [39] and cheminformatic filtering [40], to direct the search towards areas in chemical space where the actives are expected to be.

The overarching trend when commencing a discovery research program with a new target is to work within a so-called drug-like compound set for hit identification and then to work back to lead-like space to advance a patentable novel chemotype. Therefore, the design of compound libraries will often (correctly) take filtering and the design rules into account when creating a diverse screening set for general application in HTS [41]. However, when we consider the challenges associated with active identification within the cancer medicinal chemical space, having too narrow a focus in the hit identification stages might not be ideal or appropriate.

Lead-like versus drug-like compounds

Broadly speaking, compounds sought after in the drug discovery process can be split into two categories, drug-like and lead-like. The concept of lead-likeness implies a physicochemical profile in chemical libraries where the members have reduced complexity (e.g. MW <400) and other more-restricted properties than those deemed drug-like. This leaves room for chemical modification in lead optimization rounds, which subsequently modify the properties

towards drug-likeness [42]. Various authors have proposed different concepts of what constitutes drug-like compounds. These models refer to having molecular similarities to known drugs [43] or acceptable absorption, distribution, metabolism, excretion and toxicology (ADME–Tox) properties [44]. Drug-likeness is often entirely dependent on the mode of administration. Other researchers have utilized the application of hard and fast cheminformatic rule systems to partition chemical space into drug-like and non-drug-like bins. We have previously illustrated that RO5 compliance alone does not necessarily imbue drug-likeness on the post-partitioned cohort [40].

It is generally accepted that the initial active molecules discovered in VS and HTS are often far removed from the drugs into which they might evolve through optimization. In such practices the best we can hope for is the identification of screening hits. For the purpose of this study, we are conferring the loose description of 'hit-likeness' to those small molecule compounds that pass the unmodified OpenEye FILTER (which uses XLOGP [45] as a measure of the hydrophobicity partition coefficient with a maximum cut-off value of six) cheminformatic criteria including a strict application of the RO5. These criteria are now widely adopted in database filtering applied in the early stages of VS.

Describing cancer medicinal chemistry space

Douglas Adams pointed out to would-be intergalactic hitch-hikers that space is mind-bogglingly big [46]. Similarly, chemical space is vast, so vast that it can be conceptually advantageous to notionally partition it into smaller, more-manageable sections. From a drug design perspective, these areas could include biologically relevant or medicinal chemical space, which would encompass smaller areas of theoretical chemical space and synthetically accessible space. These subsets further contain places such as drug space and lead space. The challenge for drug discovery is to explore areas of value in medicinal chemical space and, within these areas, retrieve bioactive, workable molecules. The navigational use of 'chemography' (GPS for drug discovery), mapping compounds onto chemical descriptors, allows us to find our way among the vast number of chemical entities (10^{60}) that exist *in potentia* [47]. Paradoxically, this landscape is so large that we can never hope to explore it fully; however, the biologically relevant and valuable subset of hit space is much smaller than chemical space as a whole but elements of hit space are dispersed within the larger system.

To describe accurately what we term cancer medicinal chemistry space, it was first necessary to amass a set of compounds known to be active against cancer (clinical and preclinical compounds), as well as quality data on compounds purported to be anticancer active but later shown to be inactive in assay. A study set of clinical cancer compounds (34 ligands) and literature-claimed active (4026 ligands tested *in vitro* with a minimum activity of $-\log IC_{50} > 6$) and verified inactive (4285) anticancer compounds were selected from the WOMBAT database (the 2004.2 release of this database contains chemical and biological data from 4773 papers published in medicinal chemistry journals between 1975 and 2005) [48]. The new NCI activity database (41,086 compounds) was also used to identify a subset of anticancer active (8688 ligands with minimum $\log GI_{50} > 6$ in assay in NCI screens) and verified inactive ligands (32,398 ligands with minimum $\log GI_{50} < 6$ in NCI assay) from compounds that had undergone general anticancer screening at NCI.

Finally, to contextualize cancer space in comparison with general non-cancer chemical space, it was necessary to describe and populate generic hit-like space. To achieve this, a pre-filtered subset of the ZINC database [49] was used, containing 109,432 commercially available molecules, with no specific activity claimed. ZINC was our choice of database because it is pre-partitioned into various regions (e.g. drug-like and lead-like). ZINC has been widely adopted by the VS community because of the quality of its content and the availability of compounds therein for validation studies, and for this reason it has been used as a benchmark set. Once uniqueness of the data had been verified by cross-correlating all dataset members, the study set of medicinal chemical space equated to a total of 158,863 compounds, of which 12,714 are known cancer actives, 36,683 are known cancer inactives and the remainder can be described within our study context as hit-like ligands in nature.

Mapping methods

To present the relationship of cancer medicinal chemistry space in the context of wider chemical space in a meaningful and accessible way, it is necessary to construct graphical distributions of the dataset in the 3D space. It was also useful for us to quantify the subsets within our data that would conform to our working definition of hit-like (i.e. those compounds which would pass an application of the FILTER software protocol including RO5 compliance). To partition the dataset on this basis, it was necessary to first calculate 2D descriptors for all members in the MOE and to identify those descriptors which related to adherence to the RO5. A total of 48 2D molecular descriptors were identified, describing atomic nature, molecular size, polarity, lipophilicity and flexibility. PCA is a relatively easy way to transform an *n*-descriptor space into a more-manageable 3D space. In our analysis, we transformed the 48 vectors space into a 3D space described by 3 principal component vectors, where each of the 3 vectors is a combination of the 48 weighted descriptors [50]. These operations facilitated the creation of graphical representations of the 3D space spanned by the compound set.

TABLE 1

Breakdown of cancer compound hit-like nature

Dataset (source)	Active	Inactive	Total members	Hit-like	Rejected[a]
Clinical drugs (WOMBAT-PK)	36	0	36	10	26 (72%)
Designed compounds (WOMBAT)	4026	4285	8311	4150 50% actives, 50% inactives	4161 (50%)
NCI assayed compounds (NCI)	8688	32,398	41,086	13,042 17% actives, 83% inactives	28044 (68%)
Anticancer active (WOMBAT and NCI)	12,714	0	12,714	4248	8466 (66.6%)

[a]Rejected: fails on application of cheminformatic tool FILTER, which takes into account RO5 fails and also the presence of 'toxic' or undesirable reactive functionalities (OpenEye Scientific Software).

Analysis

Table 1 illustrates the nature of cancer compound space considered in this study. In a total of 12,714 verified-active anticancer compounds, only 33.4% pass a cheminformatic hit-like filter, positioning almost two-thirds of cancer actives outside what is accepted as hit-like chemical space. One would expect a modicum of attrition in such a process when utilities such as FILTER are designed to remove not only RO5 fails but also specific molecules containing toxic and reactive functionalities by removing staples such as alkylating agents, nitrogen mustards that make up a large proportion of our anticancer arsenal. In our analysis, however, the vast majority of compound failures stemmed from a lack of MW, log *P* and H-bond acceptor compliance with regard to the RO5. The actual level of attrition in these circumstances is significant when the diversity of the active molecules is considered in the 3D space; this is not simply a matter of 'nasty' groups removing cancer actives

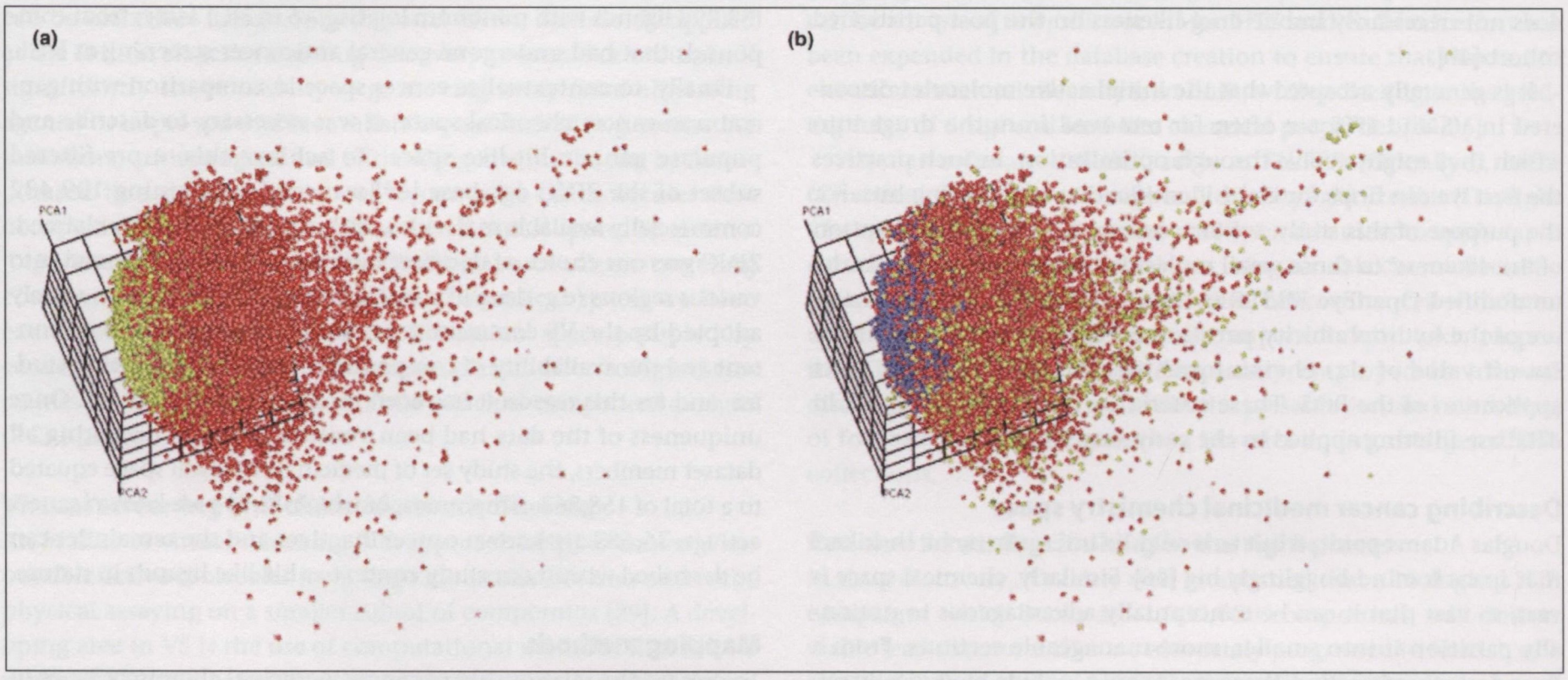


FIGURE 4

Charting cancer medicinal chemistry space. The magnitude of cancer-active medicinal chemistry space, as compared with generic drug-like space, is vast. This illustrates the need to focus in on target related areas of chemical space when seeking new selective drugs. **(a)** Key: yellow sphere, generic hit space; red spheres, medicinal chemistry space. **(b)** Key: blue spheres, generic hit space; red spheres, cancer-inactive medicinal chemistry space; yellow spheres, cancer-active medicinal chemistry space.

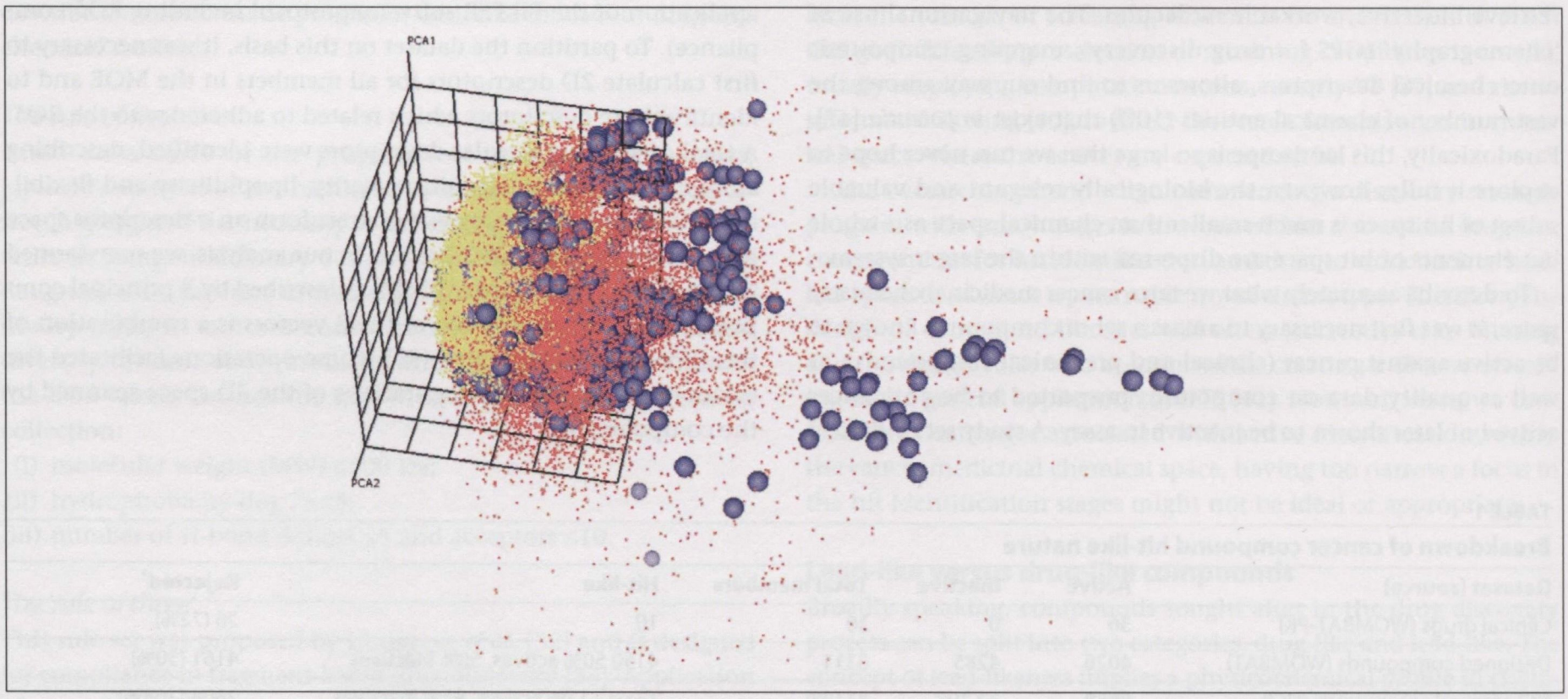


FIGURE 5

Anticancer kinase-targeted space. When a specific cancer target is studied, it is obvious that the majority of its active ligands populate non-generic drug like space. Thus, care needs to be exercised when applying general drug-like filtering criteria to cancer ligand selection procedures. Key: blue spheres, active ligands; yellow cloud, generic hit space; red cloud, non-kinase targeted medicinal chemistry space.

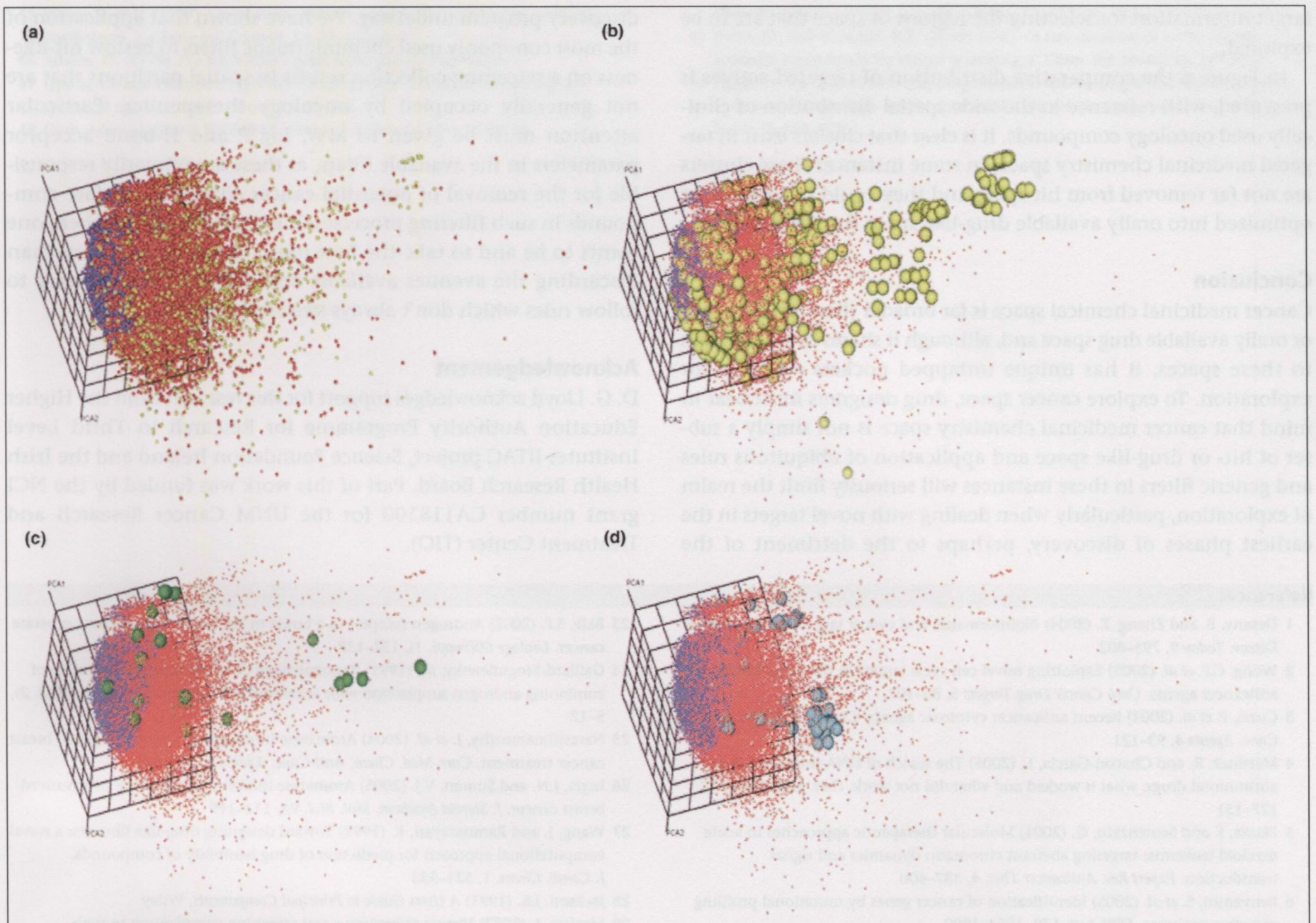


FIGURE 6

Charting cancer medicinal chemistry space. An illustration of the regions of chemical space occupied by different cancer targets shows that filters need to be carefully applied even when considering cancer-active ligands. Drugs binding to different cancer targets do not necessarily reside in closely related areas of chemical space. **(a)** Key: yellow spheres, all cancer actives; red spheres, cancer inactives; blue spheres, traditional hit-like space (indeterminate activity – passes FILTER). **(b)** Key: yellow spheres, cancer actives known to target kinases; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space. **(c)** Key: green spheres, clinically used anticancer compounds; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space. **(d)** Key: cyan spheres, cancer actives known to target tubulin; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space.

from the search space. The implication when looking for cancer actives in prefiltered compound collections is immediately clear.

We can learn a similar lesson when we examine the graphical distribution of the cancer medicinal chemistry space (all compounds assayed *in vitro*) in relation to the wider chemical space (entire study population). The overall distribution of active anticancer compounds spans an area of medicinal chemistry space far beyond that described by our hit-like definition (Figure 4).

Figure 4a contextualizes the nature of medicinal chemistry space considered in this study. The yellow spheres are nonspecific hit-like compounds taken from the filtered ZINC database, illustrating the relatively compact nature of the RO5 space when compared with the wider medicinal chemical space. The red spheres are compounds which were claimed, assayed or demonstrated as anticancer agents from WOMBAT and the NCI databases. These compounds represent charted cancer medicinal chemical space. Figure 4b illustrates the distribution of active anticancer compounds (yellow) in comparison with inactive cancer medicinal chemical space (red) and in relation to generic hit-like compound space (blue).

Even when examining familial distributions, as for the cancer kinome [17], the breadth of cancer space spanned by actives is considerable (Figure 5). Figure 5 shows a view of the distribution of anticancer kinase-targeting compounds in medicinal chemistry space: from a total of 915 active compounds examined (blue spheres), only 156 (15%) lie within our defined hit-like space (yellow cloud), whereas 759 (83%) lie outside (red cloud) and do not pass application of FILTER incorporating RO5 compliance.

Such a general spatial distribution of actives outside the boundaries of the traditional hit-space precludes the creation of all-encompassing cancer-generic filtering rules for database pre-processing. By adopting a class-by-class focus on targeted compound sets (e.g. antitubulin or anti-EGF receptor), compounds could be used to craft tailored cheminformatic filters biased to the target of study for the creation of more rationally focused screening collections. These filters can be utilized in the exploration of target-relevant chemical space. The caveat here is the need for unambiguous

target information for selecting the regions of space that are to be explored.

In Figure 6 the comparative distribution of targeted actives is presented, with reference to the wide spatial distribution of clinically used oncology compounds. It is clear that clusters exist in targeted medicinal chemistry space, in some instances these clusters are not far removed from hit-space and they could potentially be optimized into orally available drug-like space through design.

Conclusion

Cancer medicinal chemical space is far broader than just hit space or orally available drug space and, although it shares common areas to these spaces, it has unique untapped pockets still ripe for exploration. To explore cancer space, drug designers must bear in mind that cancer medicinal chemistry space is not simply a subset of hit- or drug-like space and application of ubiquitous rules and generic filters in these instances will seriously limit the realm of exploration, particularly when dealing with novel targets in the earliest phases of discovery, perhaps to the detriment of the discovery program underway. We have shown that application of the most commonly used cheminformatic filters to bestow hit-likeness on a screening collection results in spatial partitions that are not generally occupied by oncology therapeutics. Particular attention must be given to MW, log *P* and H-bond acceptor parameters in the available filters, as these are primarily responsible for the removal of potential cancer clinical candidate compounds in such filtering processes. It is crucial to think where one wants to be and to take the best route to get there, rather than discarding the avenues available because of a conditioning to follow rules which don't always need to be applied.

Acknowledgement

D. G. Lloyd acknowledges support for this research from the Higher Education Authority Programme for Research in Third Level Institutes IITAC project, Science Foundation Ireland and the Irish Health Research Board. Part of this work was funded by the NCI grant number CA118100 for the UNM Cancer Research and Treatment Center (TIO).

References

1 Desany, B. and Zhang, Z. (2004) Bioinformatics and cancer target discovery. *Drug Discov. Today* 9, 795–802
2 Wong, C.F. *et al.* (2005) Exploiting novel cell cycle targets in the development of anticancer agents. *Curr. Cancer Drug Targets* 5, 85–102
3 Cozzi, P. *et al.* (2004) Recent anticancer cytotoxic agents. *Curr. Med. Chem. Anti-Canc. Agents* 4, 93–121
4 Martinez, R. and Chacon-Garcia, L. (2005) The search of DNA-intercalators as antitumoral drugs: what it worked and what did not work. *Curr. Med. Chem.* 12, 127–151
5 Piazza, F. and Semenzato, G. (2004) Molecular therapeutic approaches to acute myeloid leukemia: targeting aberrant chromatin dynamics and signal transduction. *Expert Rev. Anticancer Ther.* 4, 387–400
6 Benvenuti, S. *et al.* (2005) Identification of cancer genes by mutational profiling of tumor genomes. *FEBS Lett.* 579, 1884–1890
7 Chen, E.X. and Siu, L.L. (2005) Development of molecular targeted anticancer agents: successes, failures and future directions. *Curr. Pharm. Des.* 11, 265–272
8 McGuire, J.J. (2003) Anticancer antifolates: current status and future directions. *Curr. Pharm. Des.* 9, 2593–2613
9 Brathe, A. *et al.* (2003) Cytotoxic activity of 6-alkynyl- and 6-alkenylpurines. *Bioorg. Med. Chem. Lett.* 13, 877–880
10 Malet-Martino, M. *et al.* (2002) The prodrugs of 5-fluorouracil. *Curr. Med. Chem. Anti-Canc. Agents* 2, 267–310
11 Izbicka, E. and Tolcher, A.W. (2004) Development of novel alkylating drugs as anticancer agents. *Curr. Opin. Investig. Drugs* 5, 587–591
12 Brana, M.F. *et al.* (2001) Intercalators as anticancer drugs. *Curr. Pharm. Des.* 7, 1745–1780
13 Larsen, A.K. *et al.* (2003) Catalytic topoisomerase II inhibitors in cancer therapy. *Pharmacol. Ther.* 99, 167–181
14 Saglio, G. *et al.* (2004) Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia. *Ann. N. Y. Acad. Sci.* 1028, 423–431
15 Collier, L.S. *et al.* (2005) Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* 436, 272–276
16 Vieth, M. *et al.* (2005) Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov. Today* 10, 839–846
17 Druker, B.J. (2004) Imatinib as a paradigm of targeted therapies. *Adv. Cancer Res.* 91, 1–30
18 Mitsiades, C.S. *et al.* (2005) Proteasome inhibition as a therapeutic strategy for hematologic malignancies. *Expert Rev. Anticancer Ther.* 5, 465–476
19 Jordan, M.A. (2002) Mechanism of action of antitumor drugs that interact with microtubules and tubulin. *Curr. Med. Chem. Anti-Canc. Agents* 2, 1–17
20 Meegan, M.J. and Lloyd, D.G. (2003) Advances in the science of estrogen receptor modulation. *Curr. Med. Chem.* 10, 181–210
21 Lloyd, D.G. *et al.* (2004) Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J. Med. Chem.* 47, 5612–5615
22 Revankar, C.M. *et al.* (2005) A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* 307, 1625–1630
23 Balk, S.P. (2002) Androgen receptor as a target in androgen-independent prostate cancer. *Urology* 60(*Suppl. 1*), 132–138
24 Gaillard-Moguilewsky, M. (1991) Pharmacology of antiandrogens and value of combining androgen suppression with antiandrogen therapy. *Urology* 37(*Suppl. 2*), 5–12
25 Narashimamurthy, J. *et al.* (2004) Aromatase inhibitors: a new paradigm in breast cancer treatment. *Curr. Med. Chem. Anti-Canc. Agents* 4, 523–534
26 Ingle, J.N. and Suman, V.J. (2005) Aromatase inhibitors for therapy of advanced breast cancer. *J. Steroid Biochem. Mol. Biol.* 95, 113–119
27 Wang, J. and Ramnarayan, K. (1999) Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* 1, 524–533
28 Jackson, J.E. (1991) *A Users Guide to Principal Components*, Wiley
29 Mestres, J. (2002) Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* 30, 797–799
30 Hou, T. and Xu, X. (2004) Recent development and application of virtual screening in drug discovery: an overview. *Curr. Pharm. Des.* 10, 1011–1033
31 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
32 Congreve, M. *et al.* (2003) A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877
33 Fattori, D. (2004) Molecular recognition: The fragment approach in lead generation. *Drug Discov. Today* 9, 229–238
34 Leach, A.R. and Hann, M.M. (2000) The *in silico* world of virtual libraries. *Drug Discov. Today* 5, 326–336
35 Shuttleworth, S.J. *et al.* (2005) Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization. *Curr. Med. Chem.* 12, 1239–1281
36 Olah, M.M. *et al.* (2004) Strategies for compound selection. *Curr. Drug Disc. Tech.* 1, 211–220
37 Chen, G. *et al.* (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J. Comb. Chem.* 7, 398–406
38 Oprea, T.I. (2002) Lead structure searching: are we looking at the appropriate property? *J. Comput. Aided Mol. Des.* 16, 325–334
39 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055
40 Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des.* 14, 251–264
41 Harper, G. *et al.* (2004) Design of a compound screening collection for use in high-throughput screening. *Comb. Chem. High Throughput Screen.* 7, 63–70
42 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263
43 Walters, W.P. and Murcko, M.A. (2002) Prediction of "drug-likeness". *Adv. Drug Deliv. Rev.* 54, 255–271
44 Lipinski, C.A. (2000) Drug-like properties and the cause of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249

45 Wang, R. *et al.* (1997) A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* 37, 615–621
46 Adams, D. (1979) *The hitch-hiker's guide to the galaxy*, Pan Books
47 Oprea, T.I. and Gottfries, J. (2001) Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3, 157–166
48 Olah, M. *et al.* (2004) WOMBAT: world of molecular bioactivity. In *Chemoinformatics in drug discovery* (Oprea, T.I. ed.), pp. 223–239, Wiley-VCH
49 Irwin, J.J. and Shoichet, B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182
50 Haggarty, S.J. *et al.* (2004) Mapping chemical space using molecular descriptors and chemical genetics: deacetylase inhibitors. *Comb. Chem. High Throughput Screen.* 7, 669–676