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.

Development, Validation and Optimisation of a vHTS Protocol for Identification of Estrogen Receptor Modulators

A thesis submitted to the
University of Dublin
for the degree of
Doctor of Philosophy in Pharmacy & Pharmaceutical Sciences

Presented by

Andrew J.S. Knox, B.Sc.

July 2006

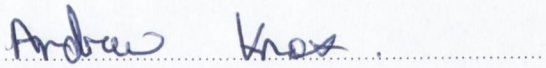
Based on research carried out under the supervision of

Mary J. Meegan, B.Sc., M.A., Ph.D. (N.U.I.), M.R.S.C., C.Chem.

at the
School of Pharmacy & Pharmaceutical Sciences,
Trinity College Dublin.

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.



Andrew J.S. Knox

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: Andrew Knox.

"We cannot direct the wind but we can adjust the sails"

Summary

The discovery of small-molecule drugs has reached a critical point in time where the number of new drugs released to market has become static regardless of the technological advances observed over the last few years. 'Big-pharma' companies had estimated they would release three to five drugs per year annually from 2000 onwards, however the average remains less than one. Methods for drug discovery over the past decades have generally involved the synthesis of compounds selected by an intuitive medicinal chemist, however, a move towards a more rational approach has occurred with the concomitant increase in the number of highly resolved protein structures publicly available.

One approach, Virtual High Throughput Screening (vHTS), takes advantage of the rational process by extracting information about the active site of a receptor or enzyme from a protein database and prioritises molecules most likely to bind to the chosen target from a large database of compounds (<1,000,000). When successfully applied, a subset of active compounds may be retrieved by numerous methods in the context of vHTS, namely, 2D/3D descriptors, 2D/3D pharmacophores, QSAR and receptor-based docking.

The work presented in this thesis describes the development, refinement and validation of a vHTS protocol against a target of therapeutic importance, Estrogen Receptor alpha (ERα), converging towards the discovery of novel scaffolds and inhibitors.

Chapter 1 details the biological role of ERα with respect to its history, active site, ligands and molecular interactions. A comprehensive description of the vHTS process is given, designed to highlight the importance of understanding the chemical space in which the drug designer must navigate to retrieve active molecules against numerous cancer targets. A review of all available studies involving the validation of docking/scoring algorithms utilising the ER is presented and illustrates the problems and successes of each approach through analysis of Enrichment rates and False Positive rates. A large number of virtual screening experiments with regard to Nuclear Receptors (NRs) are

discussed with a special emphasis on those involving the ER. It is clear at this stage that the best docking/scoring platform to utilize is highly dependent on the choice of target.

Chapter 2 investigates the notion of compound database preparation prior to docking. The 'hidden' impact of pre-processing on prioritization of actives in the virtual screening process is examined, grouped according to protonated, tautomeric, stereochemical and conformational studies. This section was devised to answer the following questions: Is docking/scoring highly dependent on the physiochemical environment of the active site and how should they be accurately represented in a screen? What computational expense does this have? What level of flexibility is needed to account for receptor-ligand movement? To clearly elucidate these answers, a test set 'seeded' with actives of ERα was used to initially optimize the docking/scoring process and allow each protocol to be examined.

Applying the same pre-processing strategy, Chapter 3 describes a novel automated vHTS method involving docking with a rigid docking algorithm (LIGIN) to ERα. Firstly a complementarity function was used to evaluate the 'buriedness' of a molecule within the active site, followed by evaluation of 14 popular scoring functions in discriminating between actives and inactives in three datasets. Another post-docking filter extracted from Ligand Protein Contacts (LPC) software was applied to distance constraints and H-bonding interactions to be analysed in the docked complexes. The former allows 'scaffold-hopping' and the latter, direct hit identification with proof-of-concept illustrated by discovery of several new scaffolds and a novel ERα antagonist. Several databases were screened using the protocol with binding affinity and antiproliferative data provided. Finally, a method for virtual library enumeration is detailed to direct synthesis towards a more optimized lead candidate with higher binding affinity.

Chapter 4 is dedicated to evaluating the benefits observed by incorporating a treatment of receptor flexibility into the docking algorithm LIGIN. A comparison of vHTS results obtained from docking into multiple receptors versus allowing slight overlap of ligands with one or two residues is given. Finally, a comparison of those methods with a new technique involving rapid receptor conformer generation using FIRST5 software in combination with FRODA is provided.

Finally, Chapter 5 introduces the concept of partitioning and classifying cancer medicinal chemistry space. Cancer medicinal chemistry space in the context of wider chemical space is detailed using filtering and PCA analytical techniques to permit the construction of graphical distributions of each in 3D space. Employing the same techniques, the specific benefits of this approach are illustrated using the ER as an example.

Also, applying other cheminformatic techniques, the presence of an alternative binding site within ERα has been suggested, and corroborating this, detailed computational work is provided using cavity analysis combined with modeling and docking studies. It is suggested that this may be the mechanism whereby non-genomic ER elicits rapid effects.

The research detailed in this thesis has brought together analysis of the main aspects involved in the process of Virtual Screening (VS), resulting in the introduction of several new techniques to enhance and aid in the discovery of new ER antagonists.

I wish to express my sincere gratitude and thanks to Dr. Mary J. Meegan for giving me the opportunity to carry out this research, for her expertise, support and guidance throughout the years. Also to Dr. David G. Lloyd for his supervision, encouragement and ambitious nature which has rubbed off on us all!

A special thanks to Dr. Vladimir Sobolev for providing much of the software used in this Thesis and also for his help and expertise. Thank you to Dermot and Bob for their help with coding. I am still bewildered by their abilities to write code as quickly as I speak English!

I would really like to thank the staff and postgrads in the School of Biochemistry and Immunology (especially Seamus, Ed and Gav) for their help and for all the crack! To the staff in the School of Pharmacy and Pharmaceutical Sciences (Rhona, Ray) for all their assistance.

Thanks to the Health Research Board and The Institute for Information Technology and Advanced Computing for their financial contributions to this project.

A big thanks to the lads and lassies in Pharmacy: Gerry, Juan, Jason, Niall, Irene, and especially, Cormac, Helena, and Miriam for all the laughs and help with Biochemistry! I would also like to thank Dr. Stephen (never leave a birdie putt short) Butler and Eanna (Boolander) for all of their help in dragging me away from this to play golf! Daly and Emer & all the lads, cheers for everything. Its great to have mates like you. To the folks in our lab (in no particular order!): Giorgio, Georgia, Yidong, Valeria, Paul and Darren. Thank you so much for all the crack and 'taastic' days and nights out and making work so easy! To all the Sligo crew! (Barnes & Mary, Barnes, Ultan, Cormac, and all) - thank you.

To my amazing Mother (Hilary), my unbelievable Stepfather (Peter) and my brilliant sister Lisa - theres not enough pages to thank you. I am so grateful for everything. Tara I never would have got this far without you. Thank you so much.

Abbreviations

BLEEP Biomolecular Ligand Energy Evaluation Protocol

CPU Computer Processing Unit

DES Diethylstibesterol

DHFR Dihydrofolate Reductase

E Enrichment

ER Estrogen Receptor

FIRST Floppy Inclusions and Rigid Substructure Topography

FLOG Flexible Ligands Oriented on Grid

FP False Positive

FRED Fast Rigid Exhaustive Docking

FRODA Framework Rigidity Optimised Dynamic Algorithm

GA Genetic Algorithms

HAC Heavy Atom Count

HrERa Human ER-alpha

HRT Hormone Replacement Therapy

HTS High Throughput Screening

IBAC Interaction Based Accuracy Classification

ITC Isothermal Titration Calorimetry

LDH Lactate dehydrogenase

LPC Ligand Protein Contacts

MC Monte Carlo

MD Molecular Dynamics

mER membrane Estrogen Receptor

MOE Molecular Operating Environment

MTT 3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyltetrazolium bromide

NC Normalised Complementarity

NCI National Cancer Institute

nER nuclear Estrogen Receptor

NR Nuclear Receptors

OFTO On the fly Optimisation

OHT 4-Hydroxytamoxifen

PCA Principal Component Analysis

PDB Protein Data Bank

PLP Piecewise Linear Potential

PMF Potential of Mean Force

RMSD Root Mean Square Deviation

SARM Selective Androgen Receptor Modulator

SATIS Simple Atom Type Information System

SBDD Structure Based Drug Design

SBVS Structure Based Virtual Screening

SERM Selective Estrogen Receptor Modulator

SLF Scaffold Linker Functional

SMILES Simplified Molecular Input Line Entry System

SMoG Small molecule Growth

STAR Study of Raloxifene and Tamoxifen

vdW van der Waals

vHTS Virtual High Throughput Screening

VS Virtual Screening

VSA van der Waals Surface Area

WDI World Drug Index

Table of Contents

Chapter 3:
Development of a screening platform for Scaffold Hopping & Hit Identification 96

Table of Figures

Table of Tables

Chapter 1

Estrogen Receptors: Molecular Interactions, Docking/Scoring & Virtual Screening.*

Comprising

* Estrogen Receptors: Molecular Interactions, Virtual Screening and Future Prospects; *Curr. Top. Med. Chem.* 2006; 6(2): 211-237.

Andrew J. S. Knox, Mary J. Meegan, David G. Lloyd.

Identification of the Estrogen Receptor (ER) as a key mediator of the proliferation of breast cancer, and its involvement in pathways leading to osteoporosis and coronary heart disease, has resulted in a surge to discover and design compounds with the ability to modulate its actions, namely Selective Estrogen Receptor Modulators (SERMs). Concurrently, a dramatic increase in the number of crystal structures of the ER has led to a more in depth understanding of the governing mechanisms involved in ER modulation. Entwining computational techniques with the availability of 3D structural data has allowed not only the rational design of potent inhibitors of the ER, but also its incorporation in Virtual Screening (VS) in the search for novel chemotypes that can modulate the ER.

1.1 HRT and Breast Cancer

Hormone replacement therapy (HRT) in the form of estrogen or estrogen/progesterone is widely used to provide effective relief of menopausal symptoms, and also in the long-term prevention of osteoporosis and fractures through inhibition of bone resorption [1]. A concurrent benefit of HRT is that it also reduces the risk of colorectal cancer [2]. Despite the inherent gains obtained by the use of HRT, definitive links between its use and an increase in the incidence of breast cancer have been shown in two recent studies from the Women's Health Initiative trial[1] and the Million Women study [3] . The major findings concur with previous studies where an increase in the risk of breast cancer (RR = 1.30 versus no-use) with estrogen-only products was observed and combined use of estrogen plus progesterone elevated the risk by 50% (RR = 2.00 versus no-use). Further association of Estrogen Receptor (ER)-positive invasive breast carcinomas and HRT was strengthened recently in a study concerning the biological effects of continuing hormone replacement therapy (HRT) after a diagnosis of breast carcinoma [4]. It was shown that a significant decrease in proliferation of ER positive breast tumors was observed in women who stopped HRT. The prevalence of breast cancer in Ireland is evident, according to figures from the National Cancer Registry of Ireland (NCRI) showing a rise in the number of new cases of breast cancer from 1,752 in 1999 to 1,890 in 2000 respectively [5].

All of these findings reflect the need to design and develop new molecular therapeutics that interfere with this proliferative process and consequently reduce the risks associated with HRT and breast cancer.

1.2 Historical Perspective of Estrogen and the Development of Antiestrogens

Exposure to estrogen has been associated for many years with an increase in the risk and incidence of breast cancer and its action through the ER has been widely documented. A connection between the ER and estrogenic action was clarified by Jensen and Jacobson (1960) with the identification of the ER as the target for estrogen action using radiolabelled estradiol [6]. Centrifugation experiments carried out by Toft allowed further characterization of the estradiol-ER complex extracted from the cytosol, and subsequent '*in vitro*' experiments confirmed the classification [7, 8]. At this stage it was postulated that antagonizing the actions of the ER might be efficacious in the prevention of breast cancer. Harper and Walpole discovered the first non-steroidal antiestrogen (Tamoxifen) and encouraged its inclusion in antitumor studies [9, 10]. In the early 1970s Jordan demonstrated the efficacy of antagonizing the 8S ER through the inhibition of the growth of DMBA-induced rat mammary carcinogenesis by administration of Tamoxifen (ICI 46,474). Tamoxifen was also shown to inhibit the growth of estrogen receptor positive MCF-7 cells in culture [11]. An experimental basis for the use of Tamoxifen as a chemo-preventative agent was now provided and this discovery prompted immense research efforts by many scientists to develop more potent inhibitors of estrogen action. Several inhibitors were developed in the 1980s with excellent antiestrogenic properties [12, 13], but the recognition of the effects of clomiphene on maintaining bone density spawned a new era of selective estrogen receptor modulation (SERMs) [14]. It is now clear that an in-depth knowledge of the specific interactions involved with the binding of a ligand to the ER, and the specific pathways induced by each is required to facilitate the design of compounds with a range of agonist and antagonist activities in different tissues. Auxiliary to this is the more recent discovery that many environmental chemicals (PCBs, phytochemicals, pesticides etc.) can act as endocrine disruptors and exert estrogenic

responses in breast tissue acting through the ER [15-17]. Research into the pathways activated by these environmental estrogens is widespread and may result in a natural compound(s) being utilized to reduce the occurrence of breast cancer and promote the positive clinical effects desired [18]. Crystallisation of the ER bound with several ligands has and will in the future help in this process [19-27].

1.3 Tamoxifen And Raloxifene

Evaluation of the ability of both Tamoxifen and Raloxifene to reduce breast cancer risk is currently being undertaken in the Study of Tamoxifen and Raloxifene (STAR) clinical trials [28]. Established differences between the two are that Raloxifene has antagonist properties in the uterus thereby reducing the risk of endometrial cancer [29]. Tamoxifen however, acts as an agonist in the uterus [30]. Benefits with respect to osteoporosis are seen with Raloxifene also as it is currently used as a treatment for it. Raloxifene is also a good choice of therapy for women with a history of breast cancer [31]. Both compounds do not alleviate vasomotor problems and so have their respective drawbacks [32]. There are obvious benefits and drawbacks to the use of these SERMs [33-36], however it is necessary to maintain research in this area to discover compounds that will assuage current clinical problems.

1.4 Overview of the process of Virtual Screening (VS)

An overview of the concepts involved in VS is necessary here to highlight the current methods enabling researchers to identify new candidate leads for a specific disease state such as ER positive breast cancer.

Approaches to discovering novel chemotypes that can exert a desired therapeutic effect have been initially explores using large compound database screening tools such as *in vitro* High Throughput Screens (HTS). Although a very useful process in the early

phase of drug discovery, testing against all disease states with every available compound using High Throughput Screening techniques (HTS) is highly inefficient. The recent incorporation of computational methods that estimate the binding affinity of molecules in a compound collection into the overall process has meant that larger amounts of chemical space can be navigated in order to increase the potential number of active compounds. This '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 [37]. VS not only complements HTS by enriching the compound set to be screened [38], but can also be employed to screen for new compounds that are more 'drug-like' or 'lead-like' [39]. A developing area in VS is the use of computational methods that filter a molecular library prior to docking towards compounds with favourable pharmacokinetics, optimum oral bioavailability, compatibility with certain types of metabolisms, and consequently low toxicity [40].

However, for VS to be viable, computational methods must be reliable, reproducible, fast and economical. There remain many limitations with the tools and techniques currently available and careful consideration of the choice of algorithms to utilise needs to be undertaken by the researcher because of the intrinsic variance associated with each target. Each docking program has its advantages and disadvantages, but these are only highlighted when they are applied to a plethora of targets that encompass a range of features such as receptor flexibility and active sites with different characteristics.

1.5 Rational Drug Design

In the process of VS, two main techniques can be employed, namely, ligand-based or receptor-based (structure-based). Ligand-based approaches use computational information generated using another compound bound to a particular target as a template to search for appropriate lead compounds [41]. Pharmacophore searching or shape complementarity methods are common examples of the ligand-based approach. For the purpose of this review however, we focus on those methods that involve Structure Based

Drug Design (SBDD). For illustrative purposes a flow of the SBDD process is outlined below in Figure 1.

Workflow of structure-based drug design process:

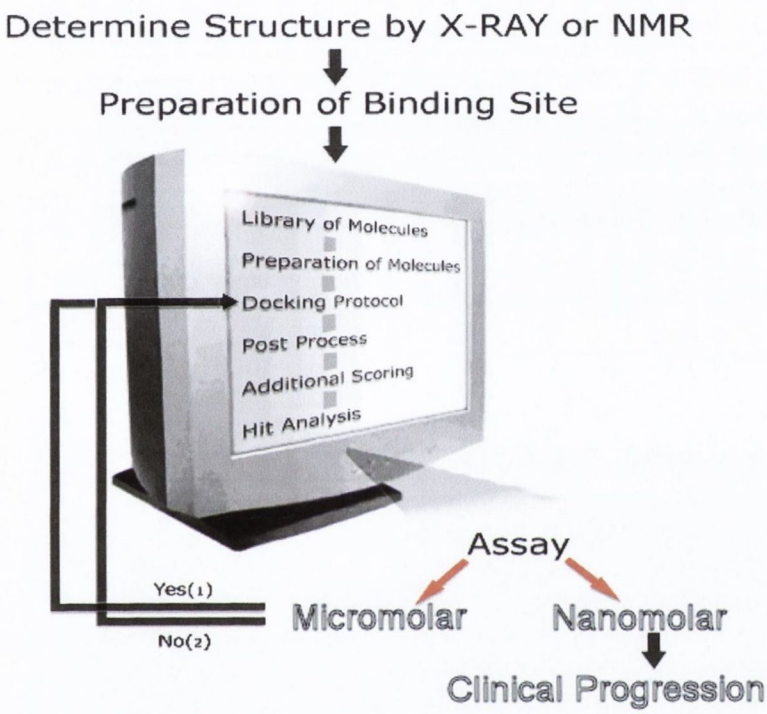


Fig **(1)** ***Yes (1)*** follows that a virtual library of compounds should be generated from this 'hit' and re-docked. *No (2)* Gives the option to optimize the existing compound or choose the next 'hit' from the ranked hitlist.

1.5.1 Target Determination and Preparation

SBDD, whose structure generally involves the computational docking of libraries of molecules into a target that has been resolved through NMR, X-RAY diffraction studies [42] or homology modelling [43]. Typically, crystallization is the method of choice due to the high resolution that can be achieved. One disadvantage of the crystallization process is that the extent of backbone flexibility is sometimes not represented fully as it really is a snapshot technique and a more complete picture may be observed using NMR where a more dynamic representation can be assembled from the many different conformations

available to the protein in solution. (eg. Bertini et al demonstrated recently the complete extent of flexibility over loop regions using NMR of matrix metalloproteinase-12 [44]). Due to the 'frozen' nature of this protein when resolved by crystallography, the different conformations observed by NMR remained undetectable. Nonetheless, the ease of access to crystallographic data through the X-ray Protein Data Bank [45] (a worldwide repository for the processing and distribution of 3-D biological macromolecular structure data) has offered researchers worldwide the opportunity to screen against a wide range of targets. To date the data bank holds above 35,000 structures (February 2006). To ensure the quality of data is high programs such as Whatcheck [46] (Centre for Molecular and Biomolecular Informatics) are available to check for errors in crystallographic structures prior to protein target selection.

Upon target determination, several common procedures are required to be carried out such as addition of hydrogens to the protein followed by optimization of their respective positions by energy minimization. In order to prepare the target for molecular docking, the active site needs to be identified and modified depending on the software used. If the binding site is unknown, identification of possible sites is possible through the use of software such as alpha site binder (MOE), SiteID (Tripos Associates Inc.) and Q-Sitefinder [47] . For the purpose of this review the ER is an excellent choice of target to illustrate VS as a large amount of crystallographic data available [21, 23, 26, 27, 48-50] and the properties and caveats for binding of a ligand to the active site are quite well understood [51, 52].

1.5.2 Specifics of ER binding site

It was thought that initiation of the transcriptional events that led to broad physiological effects in different tissues occurred through a single ER, namely ERα. However, the discovery of a second receptor isoform, ERβ prompted a revision of the nature and action of the ER. ERα has predominantly been shown to be present in the breast and uterus, whereas ERβ is more abundant in its distribution, being expressed in the breast, CNS,

gastrointestinal tract and kidneys [53-55]. Importantly, ERα is also known to be more predominant than ERβ in breast tumor tissue.

Critical differences between the agonist and antagonist ligand-binding domain (LBD) utilizing the crystal structures of ERα complexed with estradiol and raloxifene are shown in Figure 2(a)-(c). The positions of LBD residues in close proximity to the ligands are revealed, and for comparative purposes the crystal structure of ERβ co-crystallised with raloxifene is also shown.

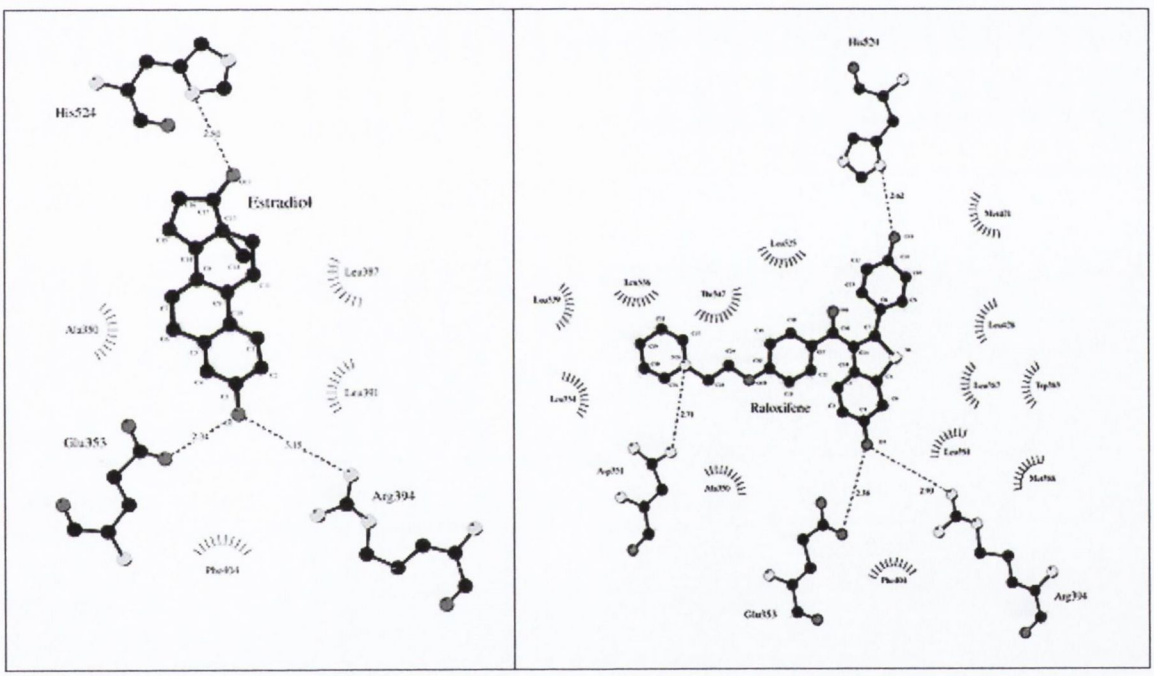


Fig **(2a)** Ligplot of ERα complexed with Estradiol (PDB ID: 1ERE). **(b)** Antagonist raloxifene is illustrated for ERα (PDB ID: 1ERR).

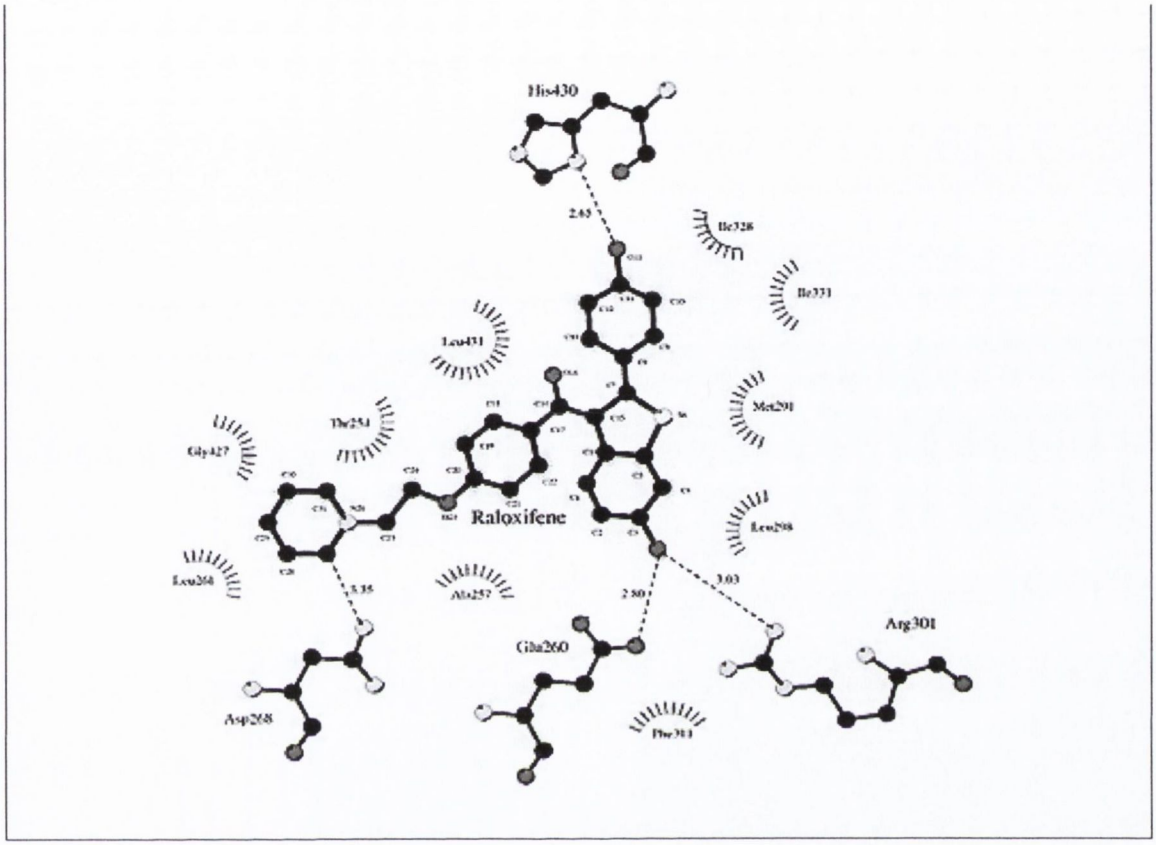


Fig **(2c)** Ligplot of Raloxifene co-crystallised in ERβ (PDB ID: 1QKN).

ERα and ERβ share only 59% homology within the ligand-binding domain. Nonetheless the essential amino acids involved in the binding process of raloxifene appear to be identical. A-ring phenolic hydroxyls interact with Glu 353 (Glu 260 in ERβ) and Arg 394 (Arg 301 in ERβ) making a direct hydrogen bond. His 524 (His 430 in ERβ) provides an additional hydrogen bond for those ligands possessing a D-ring phenolic hydroxyl group [20]. Residues lining the cavity interact through hydrophobic bonds, helping to maintain the position of the ligand and make ligand recognition achievable. The fact that ERα activity is usually present in the uterus and breast, whereas ERβ is mostly distributed in the CNS, cardiovascular system, immune system, gastrointestinal/urogenital tract, kidney and lung also provides additional scope for selective modulation.

All of these features are common to agonist and antagonist binding, however the most prominent difference is observed with the addition of the basic side-chain of raloxifene and other SERMs. The piperazine ring of the side-chain forms a direct hydrogen bond with Asp 351 (Asp 268 in ERβ) securing it in place. Upon hormone binding Helix-12 adjusts its position to enclose the ligand and seal the hydrophobic cavity within. It is this helix re-positioning that allows recruitment of co-activators to the newly formed AF-2 site and initiation of transcription. However, upon binding of raloxifene, rather than Helix-12 enclosing the ligand, the side-chain interaction with Asp 351 prevents it. This restricts the formation of the co-activator binding site and the orientation of this conformation is the basis of antagonism of the ER [56]. It has been previously shown that ER modulators induce distinct conformations of ERα/β [57] as depicted in Figure 3a.

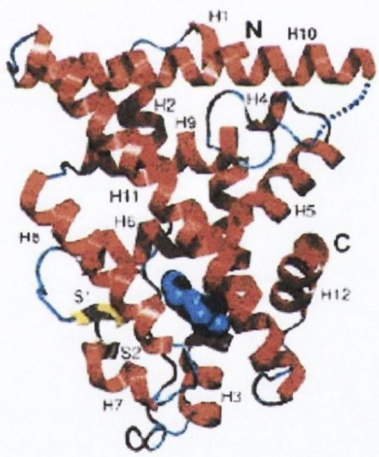


Fig **(3a)** Ribbon representation of ERα-LBD. Alpha helices (H) are coloured red, extended regions (S) are yellow, and coil regions in blue [20].

The degree to which the helix movement is hampered appears to reflect the level of antagonism induced by each modulator. The recent solution of the crystal structure of a partial antagonist (GW5638) bound to ERα supports this hypothesis. Clarification of an alternative position for helix-12 was possible with this bound partial antagonist, as illustrated in Figure 3b below.

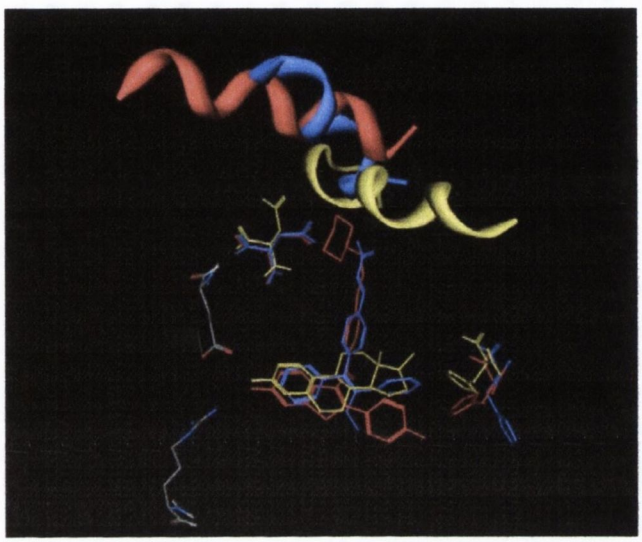


Fig **(3b)** Alternative formation of Helix-12 due to bound partial antagonist GW5638 (Blue). Raloxifene is shown in red and estradiol shown in yellow.

Another form of antagonism is apparent and suggested by the crystal structure of the R, R enantiomer of 5,11-cis-diethyl-5, 6,11,12-tetrahydrochrysene-2, 8-diol (THC) bound to ERβ [27]. Activity of this ligand is derived from relocation of helix-12 via re-positioning helix-11 rather than from directly interacting with helix-12 [58]. This type of antagonism is termed 'passive' antagonism.

It is clear at this stage that there is a plethora of possibilities for modulating the ER and each ligand induces distinct conformations that will mediate different effects in different tissues accordingly. Although the core scaffold of these SERMs illustrated in Figure 3b appears similar, their differential actions at gene level are surprising. An example of this was revealed in a recent study where only 27% of genes regulated by raloxifene were also regulated by tamoxifen [59]. This level of diversity and tissue specific effects has warranted the design and discovery of novel compounds that exhibit different gene expression.

1.5.3 Rational Design of ER modulators

Recent computational advances have permitted the use of molecular docking algorithms as effective drug binding prediction platforms. Application of these algorithms has allowed the rationalization of the biological activity observed from ligands synthetically designed (with prior knowledge of the binding mechanism) to inhibit the actions of ERα [60-63].

The tolerance of the ER to a series of flexible antiestrogens was recently examined in our laboratory and computational studies were initially undertaken to rationalize the choice of synthesis. Post-computational analyses of the binding modes of each ligand allowed us to correlate the set of ligands with their respective potencies and differentiate between 'good' and 'bad' binding modes [61]. With this knowledge, subsequent design of a set of modified novel flexible antiestrogens involved the use of similar molecular docking methods to justify and rationalize their synthesis [62]. As a means of determining the full effect of flexibility of ligands that bind the ER, a Structure Activity Relationship (SAR) investigation, and further computational simulation was carried out in our laboratory to identify compounds with conformational constraint [60]. To avoid the complications observed *in vivo* with E/Z isomerisation of tamoxifen, a benzoxepin ring system was employed as a scaffold for an antiestrogen. The computational studies demonstrated a common binding mode as observed with other SERMs and helped assist our choice of synthesis.

An example of an alternative binding-mode within the same binding site of the ER was brought forward by Kekenes-Huskey et al recently [64]. The group postulated through several docking experiments, using a course-grained model approach and the docking program DOCK4.0 [65], an alternative-binding mode for these ligands involving hydrogen bonding through Thr347 rather than His524 as seen with the endogenous ligand estradiol. Several 2,3-diaryl-imidazolines and 2,3-diaryl-piperazine compounds were docked into the binding site of ERα and due to the angular nature of these compounds as shown in Figure 4, an atypical binding pattern was observed.

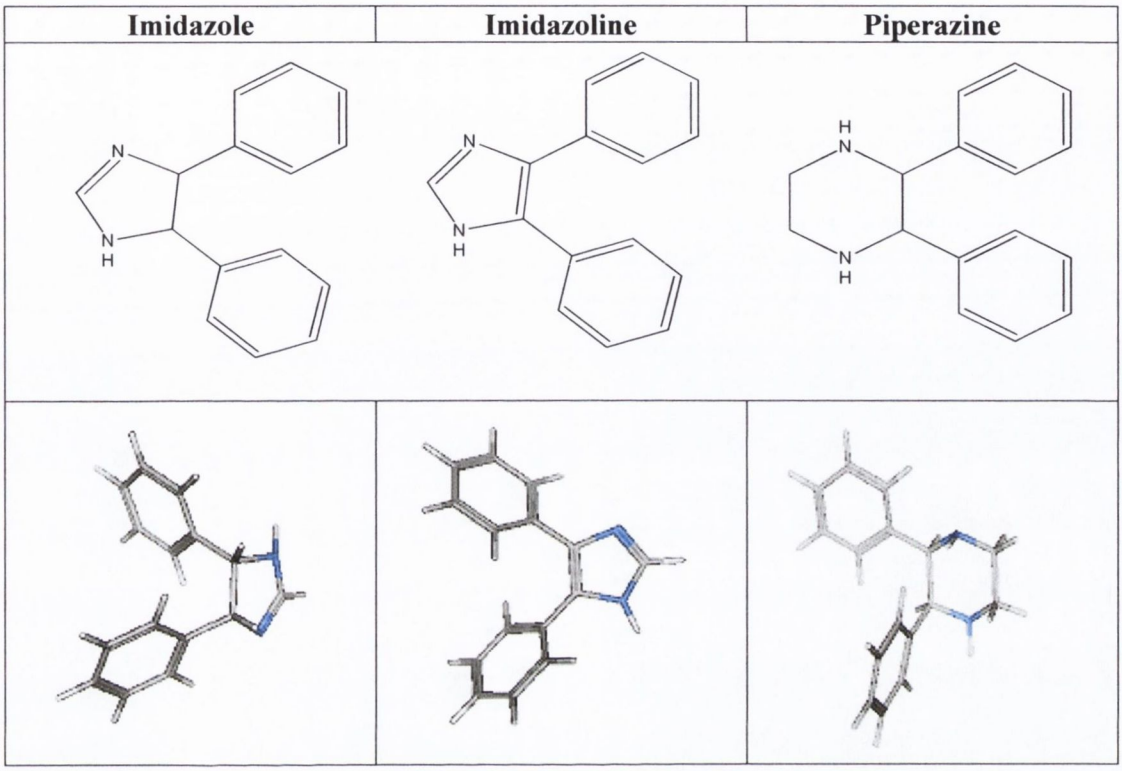


Fig **(4)** 2D and 3D structures of three 2,3-diaryl-imidazolines and 2,3-diaryl-piperazines

Although the binding cavities of ERα and ERβ share common ligand H-bonding residues, exploitation of differences in individual residues within the active site can allow design of selectivity between the two. Manas et al used multiple x-ray structures of both ERα and ERβ co-crystallised with several ligands, combined with docking calculations, to discern the critical differences for obtaining selectivity of ERβ over ERα [22, 50]. More specifically, the group identified a key difference in selectivity between ERα and ERβ by addition of functional groups to a core scaffold that interacted more favourably with Ile373 in ERβ than with Met421 in ERα. Compounds with >100-fold ERβ selectivity over ERα were generated. It is clear from these studies that taking advantage of the crystal structures available for the ER allows validation of not only binding modes but also the visual design of selective inhibitors.

1.6 Receptor & Ligand representation

Carrying out a virtual screen requires all of the above criteria to be fulfilled, but also extremely important is the question of how to represent the system to be studied. For the purposes of docking, a small molecule must have its degrees of flexibility represented as discussed in section 1.5.5. Equally important, however, is the choice of representation of the target receptor (protein). The changes that a receptor undergoes upon ligand binding can be described by three main events [66]. Firstly, small-scale fast motions whereby mainly side-chains move and/or minimal backbone movement is observed. Large-scale slow domain motions are observed secondly, such as hinge-bending caused by flexible regions moving about rigid areas. The third motion is depicted as that observed upon ligand binding with stabilisation of the receptor (order) in contrast to a partially unfolded state (disorder). The docking procedures currently available can be broadly categorised according to their ability to represent these systems. They can be classified as (1) Rigid-Body docking; where both ligand and receptor are treated as rigid bodies (2) Semi-flexible; either ligand or receptor is treated flexibly (3) Fully-flexible; both ligand and receptor have inherent flexibility accounted for within the algorithm. These procedures are now discussed below. A more detailed account of how these procedures are incorporated in docking algorithms is provided in section 1.7.

1.6.1 Ligand flexibility

Inclusion of ligand flexibility into the docking process can be computationally expensive as the degree of flexibility maintained by a ligand is dependent on the number of rotatable bonds it possesses [67]. Three main approaches exist for account for this flexibility **(1) Conformer ensemble generation** [68]; where numerous low-energy conformers are produced prior to docking and stored. Each conformer is then docked rigidly to a receptor active site that can be treated as rigid also or flexible. Lomber et al [67] described an extension of this method, whereby the conformers were docked as an ensemble rather than individually. This allowed identification of fragments that were rigid and they were docked only once with flexible fragments incrementally added. They have also recently

detailed a hierarchical docking method that allows for thousands of conformations to be represented far more concisely [69].

Finally, a method given by Joseph-McCarthy et al [70] describes the program PhDOCK which overlays conformers of the same and different molecules based on their largest pharmacophore and matches the pharmacophore points to a pre-defined set in the active site. Once docked, the interactions for each individual conformer is assessed and scored. This serves to dramatically reduce the time needed to dock single conformers.

(2) Incremental construction [71]; Rigid anchor (DOCK) or base (FlexX) fragments are identified first and placed via hashing technique (FlexX) or sphere-matching (DOCK) into the active site. In a step by step procedure, the rest of the ligand is reconstructed by docking different orientations of fragment sequentially and assigning a score of which the best-scored fragment remains docked until the whole molecule is associated in the active site.

(3) Genetic algorithm [72]; Another strategy for docking is to use a genetic algorithm, as in the program GOLD. A 'chromosome' is used to reflect the conformational state of a ligand. The chromosome is in fact a binary string in which every bit encodes a torsional angle. An island model is used in the algorithm whereby several subpopulations of chromosomes are created, and individual chromosomes may migrate to all other subpopulations by 'mutation' or 'crossover' as observed in Darwinian evolution. A niching technique, in the case of GOLD, is then employed to ensure that when a chromosome is added to a subpopulation, the number of individuals occupying the same niche is calculated. Those individuals deemed as occupying the same niche have an rmsd of <1.0Å between their respective donor and acceptor coordinates. If the niche to which a new individual to be added is too large the new individual replaces the worst member of the niche. A scoring function comprising hydrogen-bonding and van der Waals interaction energies is used to calculate the 'fitness' of new individuals.

1.6.2 Protein flexibility

Proteins are dynamic systems, and can undergo a broad range of conformational changes depending on their environment and also to accommodate their endogenous ligand, thus increasing the potential interactions between the active site and the ligand. These motions span from only a few side-chain movements to larger, hinge-bending movements. Two separate groups, Najmanovich et al [66] and Zavodsky et al [73] have shown that generally upon ligand binding, only a small number of residues will undergo conformational change in the receptor and also they do to a minimal extent. There are a number of cases however that this does not account for, but they are relatively limited. Thus, most computational approaches will either employ an algorithm that keeps the receptor fully rigid or minimally flexible in the docking process. Importantly also is the fact that to computationally represent the proteins full degree of motion would be highly inefficient and unfeasible.

A simple technique to incorporate partial flexibility involved soft docking was introduced by Jiang and Kim [74]. It was implemented by reducing the van der Waals contributions between the ligand atoms and atoms in the binding site. This however failed to account for the degree of 'softness' to be applied and could therefore allow unrealistic conformations to occur. However, it was easily implemented and very fast, as the overall algorithm remained rigid. Ferrari et al also detailed a recent 'soft' scoring function with an attenuated Lennard-Jones potential that allowed a closer approach between the ligand and receptor binding site [75]. Interestingly, the soft potential could better identify known ligands than the hard scoring function when only a single receptor conformation was utilised. On the other hand, when multiple receptor conformations were used the soft function preformed worse than a hard function also implemented.

Docking of ligands into multiple receptor conformations can be carried out by selecting all X-ray structures of a receptor with different bound ligands, or alternatively by utilising NMR conformers. Alternatively an ensemble of structures can be generated by molecular dynamics (MD) or Monte Carlo (MC) simulation [76]. Kuntz and co-workers presented the first study utilizing multiple protein structures. Using both X-ray and NMR structures as

sources, interaction grids were generated for each. Two different methods were then used to average these structures and a composite grid produced for docking with DOCK. A dramatic improvement in docking accuracy was observed. Methods like these attempt to overcome the problem of computational time needed to dock against multiple conformers by reducing the representation. A recent report has also described a multi-conformation approach using FlexE [77]. Multiple structures are generated and those that closely resemble one another are averaged, but those disordered regions are retained to give a degree of flexibility.

A minimal set of receptor conformations selected by firstly docking an active set of ligands against an ensemble of conformations has been reported [78]. Secondly, population weights assigned to each are optimised to yield a minimal set of conformations that correlate well to binding affinities observed for the actives set. This subset is then used to dock the remainder of the compound library and substantially reduce docking speed. Corroborating these studies Erickson showed using CDOCKER that docking to an 'average' structure of a receptor reduced the accuracy of the binding orientation and that this drop in accuracy mirrored the degree to which the protein moved upon binding [79]. In a recent meeting of the MGMS (2005) it was noted however that although enumerating receptor conformers and docking ligands improved docking accuracy, the hit rate was significantly reduced [80]. Noteworthy, Broughton et al [81] introduced a method using molecular simulations of the Cox-2 receptor to produce ensembles and an 'average' model was chosen as the docking target. FLOG [82] was then used to dock a set of ligands. The results showed an improvement over using a single crystal structure in the number of known ligands found in the top 10% of the screened database.

Numerous other groups follow different approaches where regions of flexibility are identified rather than taking flexibility of the whole active site into account. Anderson et al [83] presented an algorithm SOFTSPOTS, whereby residues most likely to change conformation are defined, and a second algorithm PLASTIC selects a set of possible scaffolds based on rotamer libraries. The method was found to reduce the pitfall that a single receptor conformer may have on the docking process.

Principal Component Analysis (PCA) was employed recently to represent flexible regions of a protein by reducing dimensionality [84], and in turn, reducing computational time. This method basically discerns the most significant degrees of freedom in a protein collectively using Singular Value Decomposition instead of taking into account all of the possible degrees of freedom.

SLIDE [73] can also incorporate protein flexibility using a graph theory technique where sets of positions are generated for the backbone through random sampling. It follows the induced fit theory, using template points generated from hydrophobic/ hydrogen bonding regions to allow complementarity matching.

Finally, an attractive method of binding prediction using a combination of multiple simulated annealing and pseudo-crystallographic refinement was undertaken recently [85] to determine the binding mode of major histocompatibility complex (MHC) class I H-2K^b complex with a viral peptide derived from vesicular stomatitis virus nucleoprotein. The procedure reduces ligand bond lengths to 0.3 A, and gradually heats them when the receptor side-chain interactions are removed. Simulated Annealing is carried out to 'grow out' the side chains and regenerate the ligand receptor interactions in its global energy minimum form. This enables a pseudo electron density map to be generated from a set of annealed structures, manifesting the conformational search space that a ligand bind to a receptor should have. The structure is refined using crystallographic simulated annealing refinement techniques. This is a very effective method for binding mode prediction but is not yet applicable to virtual screening being computationally inefficient.

1.7 Molecular Docking

The recent surge in the determination of target protein structures has fuelled implementation of molecular docking algorithms that harness computational power to dock a compound into a receptor binding site and predict the correct binding mode [86].

Docking involves defining the correct positioning of a ligand in the binding site by searching the conformational space available, and applying a scoring function that defines the "best fit" that the ligand has in the active site. The most complex problem that an algorithm has to overcome is to inherently allow for the large degrees of freedom that both ligand and receptor will have. Several approaches have been implemented that are currently incorporated in the docking.

Docking algorithms currently available can be separate into five main sets, namely, Descriptor matching, Fragment-based, Monte Carlo, Genetic and Tabu search methods. We provide a description of each method and the programs that fall into each category in the next section.

1.7.1 Descriptor Matching methods

This class of docking algorithm is generally defined by one that generates numerous ligand orientations within an active site by atom matching with specific points pre-defined in the site. Mostly these algorithms are rigid-body procedures but can be adjusted to take into account a small degree of flexibility. Chemical information can be passed to the docking such as atom typing according to arbitrary sets of rules as utilized in LIGIN [87, 88] and DOCK [89] to enhance the docking process.

FTDOCK [90] is a rigid body algorithm based on surface recognition. Surface complementarity with an electrostatics model using Fourier correlation theory is used to score the docked poses. By computationally placing a grid on each molecule and assigning each node (l,m,n) a value in 3D, a score is calculated based on whether the ligand penetrates the surface of the binding site. If the value is 0 the ligand and receptor do not interact whereas if the value is negative there is overlap.

LIGIN [87] utilizes an input file with a list of the atoms of the ligand to be docked. Each atom is denoted a number which correlates to an arbitrary rule (eg. 1= Hydrophilic: - N and O atoms that can donate and accept H-bonds). This chemical information is carried through the docking process so that the randomly generated starting conformations can then be optimized by surface complementarity and H-bond geometry. LIGIN also surmounts the flexibility obstacle by allowing ~20° rotation about single bonds during the optimization process without any penalty.

Similarly, SANDOCK [91] uses a rigid-body shape complementarity approach with chemical information incorporated to assess the 'fit' of a ligand in a receptor binding site. The active site is represented by dots, which encode chemical properties and accessibilities of essential residue atoms and are mirrored with the atoms of the ligand using a distance-matching algorithm. The poses are evaluated by hydrogen bonding, hydrophobicity and geometric parameters.

FLOG [82] introduces the concept of ligand flexibility through conformation ensembles by generating a set of up to 25 conformers per ligand. A clique-finding algorithm [92] is used to match distances between the ligand atom pairs and favourable sites in the protein. The chosen ligands are superimposed to the favourable sites and optimized with a simplex rigid body optimizer. Scoring involves the use of electrostatic, hydrogen bonding, hydrophobic and van der Waals potentials.

FRED [93] is a rigid exhaustive and systematic docking program that predicts binding modes in a reproducible manner due to its non-stochastic 'engine'. Shape complementarity is the primary method for evaluating poses, however pharmacophoric constraints can also be imposed. Schulz-Gasch et al [49] deemed FRED to be especially attractive as a docking tool because it docks at a high speed compared with other methods.

1.7.2 Fragment-Based methods

Matching algorithms such as those implemented in DOCK4.0 [65] account for shape complementarity in the pose selection process using a clique-search based approach, where the volume of the active site is set in terms of spheres. DOCK4.0 introduces the concept of fragment docking whereby the ligand is separated into fragments and a core fragment is anchored in the cavity by steric complementarity. The other fragments are docked in different orientations sequentially and assigned a score of which the best-scored fragment remains docked until the whole molecule is associated in the active site. A pruning algorithm reduces the number of poses by deleting 'bad' conformations.

FlexX [71] also performs fragment based docking using an incremental construction algorithm but mainly differs from DOCK in its initial placement of the core fragment, where receptor group properties define the interactions with the fragments. Hydrogen bond and hydrophobic interactions mostly influence the placements of the fragments. A pose-clustering algorithm [94] dictates the positions of the fragments and the ligand is built incrementally from there producing a number of solutions. These docked solutions are then subjected to a further round of clustering to determine the best docked pose.

Hammerhead [95] also reduces the ligand into fragments but differs from the other methods described previously in that the initial fragment placements are scored and then the ligand is built incrementally with some minimization occurring concurrently. SURFLEX is the next generation Hammerhead with more efficient incremental construction of the ligand in the binding site [96]. A similarity module incorporated in SURFLEX was modified and used as ligand based screening system in distinguishing actives from inactives [97]. True positive rates of 60% were observed illustrating the efficacy of this additional function.

Finally SLIDE [98] represents another fragment-based approach whereby the binding site is made up of hydrophobic and hydrogen bonding points. An anchor fragment of the ligand containing hydrophobic and hydrogen bonding points is matched to the binding site template by a multi-level hashing algorithm. The remainder of the ligand is added to the core fragment using the input ligand coordinates. SLIDE takes the docking process a step further by allowing both ligand and receptor flexibility through

side-chain rotamers and movement of ligand dihedrals. This side-chain rotation or movement of dihedrals of the ligand serves to remove any overlap during the construction of the remainder of the ligand.

1.7.3 Monte-Carlo methods

These methods often allow incorporation of flexibility into the ligand or receptor or both. ICM [99] uses a Monte Carlo algorithm to minimize an energy function in torsional space. The docking process involves 'freezing' the positions of bond lengths and angles but torsional movement of the ligand and side-chains is allowed. This serves to reduce the degrees of freedom required to model and speeds the process up significantly. Using the Monte Carlo process, a number of low energy conformations are generated and duplication of any conformer's forces the algorithm to double the simulation temperature to overcome the multiple-minima problem [100]. Application of the ECEPP/3 force field ensures realistic conformations are produced.

QXP [101] uses an algorithm derived from a Monte Carlo search also, but a fast step within it produces approximate low-energy structures, which are likely to minimize to a low energy state instead. It then uses a superposition force field that automatically assigns short-range attractive forces to similar atoms in different molecules. This alignment of ligand and residue atoms now initiates the Monte Carlo search again only allowing rigid rotation and translations of the ligand. Sequential minimization of the ligand torsions and initiation of a Monte Carlo search again occurs and the final optimized poses are scored.

Exhaustive enumeration of ligand conformations is performed in GLIDE [102] with a systematic search. A grid representation of the properties of the receptor is required and a pre-screening step of ligand poses is carried out using this grid to select good poses. This allows implementation of a more CPU costly Monte Carlo step to examine those selected poses minima.

1.7.4 Genetic algorithms

Another strategy for docking is to use a genetic algorithm, as implemented in the program GOLD [72]. These algorithms represent the solution as a 'chromosome' reflecting Darwinian evolution. Genetic algorithms (GA) are an iterative process where the best solutions of a population consequently have the best chance of evolving by processes such as 'crossover' and 'mutation'. The selection process is tailored towards hydrogen bonding between ligand and receptor and a scoring function evaluates the 'fitness' of each solution.

Autodock3.0 [103] also utilises a GA, but is combined with a local search procedure of minimization to ensure local minima are found. This type of algorithm combination is termed Lamarckian. The fitness function differs to that of GOLD having 5 rather than 3 terms, sum of independent contributions from a van der Waals term, a Hydrogen-bond term, a screened Columbic electrostatic term, an entropic term based on torsional strain, and a solvation term.

GEMDOCK [104] generates a random population of ligand solutions from the center of the receptor. Each solution accounts for three n-dimensional vectors with the first (x^i) representing the location of the ligand, the rotational angles and the rotable bond angles. The second (σ^i) and third (ψ^i) represent the vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. Each step results in a population of solutions that are produced from the previous set through mutation.

1.7.5 Tabu methods

PRO_LEADS [105] is comprised of a tabu search algorithm and a semi-emperical scoring function is applied to estimate the binding affinities of the docked conformations. The tabu search algorithm parallels simulated annealing techniques where often a solution moves to a worse solution with the expectation that it will eventually lead to a better solution (ie. A stochastic process). If the current solution is better than the best solution so far, store it as the new best solution in a tabu list and remove the oldest item on the tabu list if it contains too many items.

The docking process is always followed by an evaluation of the docked poses generated and subsequent prioritization by 'fit' is possible using a plethora of scoring functions currently available [106]. We outline the scoring functions available separately as they can generally be applied to all docking algorithms.

1.8 Scoring Functions

Scoring functions are used to estimate the strength of binding between a molecule bound in a particular pose to a macromolecule. Generally, most scoring functions estimate the free energy of binding for a receptor-ligand complex in aqueous solution. Scoring functions generally fall into four categories, which are Force Field, Empirical, Knowledge-based and Consensus. A discussion of the merits of each will be presented in section 1.8.2. It is necessary to provide some background at this stage into the thermodynamics of the binding process to contextualise what each function is attempting to do.

1.8.1 Thermodynamic Parameters involved in ligand binding

Binding of a ligand to a receptor is achieved through a series of complex non-bonded interactions such as hydrogen bonding, lipophilic aromatic or aliphatic contacts. As long as equilibrium is maintained, the affinity can be directly related to the binding constant ΔG (Gibbs free energy). ΔG comprises the enthalpic and entropic contributions involved in the process. Strong enthalpic interactions for example would result in a loss of flexibility within the system and a concurrently more ordered state or more specifically a reduction in entropy. This playoff is a difficult to assess mathematically as it is a dynamic process and is therefore constantly changing. The movement of a ligand from solvent, where it possesses most of its conformational degrees of freedom and interacts strongly with some water molecules but not all, to a receptor cavity that is also filled with water molecules results in a loss of water from both. This loss of water from the binding site will contribute to some gain in entropy and loss in enthalpy. However, the

immobilization of the ligand in the active site will most likely attribute with it a loss in entropy of the system also. Conversely a gain in entropy by occur from the re-ordering of local water molecules initially bound to the ligand but now free. The ligand and receptor are now in complex. The delicate balance of this process is difficult to measure but a number of scoring functions are available to estimate the ΔG and in turn the affinity of the ligand for the receptor and are outlined next.

1.8.2 Force-field scoring functions

Typically, force fields are sums of non-bonded interactions corresponding to stretching, bending, torsion, van der Waals and electrostatic interaction energies as functions of conformations. Only estimations of enthalpic contributions are made (ΔH) rather than ΔG (Gibbs free energy). Representation of solvation and entropy terms is also possible as in DOCK. Several scoring functions based on force fields are available in the FlexX module (Tripos Inc.) and are detailed here.

G-score focuses on hydrogen bonding interactions and is based on the program GOLD. This scoring function can be broken down into three main segments: a pairwise energy, a hydrogen bonding and internal energy term. Pairwise energy represents the steric energy between the ligand and receptor. The equation follows the below form where Lennard-Jones 8-4 potential calculates the pairwise energy of protein-ligand complexation [72]. The Lennard-Jones potential describes dispersion forces at long ranges (van der Waals) and repulsive forces resulting from overlapping of electron orbitals (Pauli repulsion).

Eqtn (1): $E_{ij} = \frac{A}{d_{ij}^{8}} - \frac{B}{d_{ij}^{4}}$

The hydrogen bonding term correlates to the sum of all individual bond energies from donor and acceptor combinations from protein and ligand. Finally the internal energy of the ligand is represented by:

Eqtn (2): $E_{ij} = \left(\frac{C}{d_{ij}^{12}} - \frac{D}{d_{ij}^{6}} \right) + \frac{1}{2} V \left[1 + \frac{n}{|n|} \cos(|n|\omega) \right]$

where E_{ij} is the torsional energy, steric energy is determined using a Lennard-Jones 6-12 potential, V is the barrier to rotation, n is the periodicity, ω is the torsional angle.

Again a function based on DOCK, D-score [89] falls into this category considering electrostatic and hydrophobic interactions as contributing to the binding energy only. This algorithm is divided into two parts that consist of a Coulombic term and a van der Waals term:

Eqtn (3): $E_{\text{non-bonded}} = \sum_{i=1}^{lig} \sum_{j=1}^{prot} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332 \frac{q_i q_j}{\in r_{ij}}$

Where A_{ij} and B_{ij} are van der Waals terms of ligand atom i, and protein j. r is the distance between i and j and $\in$ is the dielectric constant. Again the Lennard Jones potential 12-6 is used to calculate the van der Waals energy. Force fields as scoring functions are accurate but lack a biological correlation with sets of ligand bound complexes.

1.8.3 Empirical scoring functions

Empirical scoring functions were introduced in the 1990s [107] to predict the binding affinity of small ligands to proteins by linear combination of physicochemical properties involved in the binding process. They are calibrated with receptor-ligand complexes using multivariate regression analysis and are very computationally efficient. The advantages over Force Field scoring functions are clear with regard to speed, however as each is validated through different training sets that may be small there is the potential for variance between methods and sometimes inaccurate predictions. Nonetheless, these scoring functions are widely used nowadays in screening processes because of their general efficacy. Basically they estimate binding free energy by splitting interactions between receptor and ligand into hydrogen bonding, van der Waals, hydrophobic and

entropic changes etc. Current examples of these scoring functions are, Ligscore (Cerius2) [108], PLP [109], LUDI [110], F-Score [71], Chemscore [111], X-Score [112], Validate [113].

The Ligscore algorithm [108] utilizes three terms, van der Waals interaction using a softened Lennard-Jones 6-9 potential, influence of the buried polar surface area between a protein and ligand involving protein-ligand attractions, minus the influence of the buried polar surface area between a protein and ligand involving both attractive and repulsive protein-ligand interactions and is represented by:

Eqtn (4): $pK_i = VdW + C_{+pol} - Totpol^2$

where C_{+pol} and $Totpol^2$ are the surface descriptors.

PLP (Piecewise Linear Potential) [109] is an empirical scoring function that describes steric and hydrogen-bonding interactions. In the PLP model atoms are classified into four types: H-Bond donor, acceptor, donor/acceptor and nonpolar. Each interaction between these atoms is then assigned one of three interaction types; donor/acceptor H-bonding, donor-donor/acceptor-acceptor repulsion, and dispersion. The energy is then the summation of the of the ineraction energies between the ligand atoms and the receptor heavy atoms. Pairs of interacting atoms are expressed as below:

Eqtn (5): $E_{total} = E_{ligand\text{-}protein}(E_{H\text{-}bond} + E_{repulsion} + E_{contact}) + E_{ligand}$

Bohm developed an empirical scoring function [110] that accounts for neutral and ionic hydrogen bonding, hydrophobicity and torsional entropic changes.

Eqtn (6):

$$\Delta G_{bind} = \Delta G_{H-bond} \sum_{neutral H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic} f(\Delta R, \Delta\alpha) + \Delta G_{hydrophobic} \sum_{hydrophobic} \left| A_{hydrophobic} \right| +$$

$$\Delta G_{rotor} N_{rotor} + \Delta G_0$$

where ΔR and $\Delta\alpha$ is the deviation in distance and angle respectively and f is a scaling function to penalize these deviations from ideal geometries. Parameter values for each ΔG were extracted from a set of 82 complexes.

F-score [71] was introduced in FlexX (Tripos Inc.) to estimate the binding free energy of a bound ligand complex. It is comparable to the previously developed scoring function by Bohm.

Eqtn (7):

$$\Delta G_{bind} = \Delta G_{H-bond} \sum_{neutral H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic} f(\Delta R, \Delta\alpha) + \Delta G_{aromatic} \sum_{aromatic} f(\Delta R, \Delta\alpha) +$$

$$\Delta G_{lipophilic} \sum_{lipophilic} f^{*}(\Delta R)$$

$$\Delta G_{rotor} N_{rotor} + \Delta G_{0}$$

However, the third and fourth terms represent calculations of aromatic and lipophilic contacts.

Chemscore [111] also gives a prediction of the free energy of binding as above but includes an additional term that accounts for metal interactions in the docked complex:

Eqtn (8): $\Delta G_{metal} \sum_{metal} f(\Delta R, \Delta\alpha)$

Wang et al [112] accounted for Van der Waals interactions, hydrogen bonding, and hydrophobic effects using their scoring function X-Score that was derived from training set of 200 complexes. The energy of binding is represented as:

Eqtn (9): $\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{H\text{-}Bond} + \Delta G_{deformation} + \Delta G_{hydrophobic} + \Delta G_{0}$

where ΔG_{vdw} accounts for van der Waals interactions, $\Delta G_{H\text{-}Bond}$ is H-bonding between ligand and protein, $\Delta G_{deformation}$ accounts for the deformation effect, $\Delta G_{hydrophobic}$ accounts for the hydrophobic effect, ΔG_0 describes translational and rotational loss of system.

As in previous empirical formulae, Lennard-Jones is used to assess van der Waals (vdW) potential, using a softened 8-4 potential.

Finally, Validate [113] uses a hybrid approach combining molecular mechanics and heuristic approaches to estimate the free energy of binding. Its function can be classified as follows. Firstly it determines the entropy changes of the ligand and receptor upon binding. It then uses a partition coefficient to estimate the ligands affinity for the receptor by calculating the lipophilic/hydrophilic inclination of the active site. Steric complementarity and vdW are assessed. Lipophilic, hydrophilic, polar and unsuitable hydrophilic contact surface area are then determined. Ligand strain energy which is the energy required for the ligand to adopt the conformation of the active site is calculated using IE = $|E_{bs} - E_{solv}|$, where E_{bs} is the energy of bound ligand and E_{solv} is energy of ligand in solvent.

These methods attempt to illustrate the physics of the binding process, by attributing it as separate contributions from hydrogen bonding, ionic and lipophilic interactions, clashes and entropy. They have been widely used in academic and pharmaceutical drug discovery processes.

1.8.4 Knowledge-Based Methods

The next method used to score ligands bound in a receptor, describes the occurrence of favourable and unfavourable atomic interactions resulting from statistical analysis of observed interatomic distances and/or frequencies. Knowledge based methods evolved from a technique originally used to analyse protein folds, and are applied in this case to the structures of protein-ligand complexes like those in the PDB [114].

BLEEP (Biomolecular Ligand Energy Evaluation Protocol), was developed by Mitchell et al [115]. The function was developed using a set of very diverse, high resolution

(<2.0A^0) PDB complexes, to give it a broader application than most. It initially uses an algorithm called SATIS [116] (Simple Atom Type Information System), to generate the atom types of the complex and give it a ten-digit code representing it bonds and connectivity. BLEEP then converts the distance distributions of the atom types into pair potentials, and from this the PMF (Potential of Mean Force) for each complex is calculated. BLEEP performed excellently in identifying the binding modes of 90 complexes.

Muegge et al [117] described a method based on the 3D structure of a complex (PMF). By analyzing a training set of 697 complexes they derived the following equation:

Eqtn (10): $$\text{PMF\_score} = \sum_{kl}^{ij} \sum_{r<rcut-off} A_{ij}(r) ,$$

where $r^{ij}_{cut\text{-}off}$ is the cut-off radius for the atom type pair *ij* and *kl* the ligand-receptor atom pair interactions available to be added. A comparison was also done showing that PMF score, when applied to a training set of eight protein-ligand complexes shows a correlation with five to the actual binding constants observed.

SMoG 2001 (Small molecule Growth), a *de novo* tool with a scoring function developed by Ishchenko et al [118], as a successor to SMoG 96, constituting a knowledge based approach from statistical analysis of 725 receptor-ligand complexes. Again the method used is to sum all of pairwise interactions between protein and ligand. SMoG 2001 outperforms PMF score in estimating the binding affinity of 77 complexes. It also performs comparably to Drugscore [119]. This scoring function was developed from analysis of structural information gained from the PDB using ReliBase [120]. The information is converted into a combination of favorable potentials for each atom pair and the solvent accessible surface dependent singlet preferences for receptor-ligand atom pairs:

Eqtn (11):

$$\Delta W = \gamma \sum_{ki} \sum_{lj} \Delta W_{ij}(r) + (1-\gamma) \times \left[\sum_{ki} \Delta W_i(SAS, SAS_0) + \sum_{lj} \Delta W_j(SAS, SAS_0) \right]$$

ie. By summing all individual contributions of k_i ligand atoms and l_j protein atoms. γ is usually held to the value of 0.5 but can be altered.

Fresno [121] is a scoring function based on the work of Bohm also, but differs in that it attempts to predict the absolute binding free energy of a receptor-ligand complex. It is split up into measuring H-bonding, lipophilic, rotational entropy and two novel terms are introduced, namely, buried-polar and desolvation terms. The advantage of using this scoring function is that it has been optimized for use against particular complexes only but can be optimized for any receptor-ligand complex.

Finally, shape based methods such as complementarity functions can be used to estimate binding affinity of ligands in a random set. Sobolov et al produced a function capable of ranking a set of ligands by analysis of interatomic ligand –protein contacts and therefore estimation of complementarity. This approach uses a surface complementarity function (CF), previously defined [122]:

Eqtn (12): $CF = S_1 - S_i - E$

Where $|S_1|$ is the sum of all the surface areas of legitimate atomic contacts and $|S_i|$ is the sum of all illegitimate contacts between ligand and receptor. E is a repulsive term similar to that used in force fields. This function is discussed in more detail in Chapter 3 & 4.

1.8.5 Consensus scoring functions

Improvement of hit rates has been observed with implementation of a combination of multiple scoring functions to estimate binding affinity [123]. Charifson et al show that a combination of Chemscore, PLP, Dock produces 'hit' rates of approximately 5-10% for enzymes possessing buried binding sites.

A concept for selecting the 'best' docked ligand (multiselect) in a quantitative statistical manner using PCA (principal component analysis) to select the single ligand conformation closest to the bioactive conformation was developed recently [124]. This pose is then subjected to partial least squares analysis of eight combined scoring functions to

estimate binding affinity. Results obtained from the work of Wang et al [125], also conclude that a combination of only three or four scoring functions are needed to serve as a consensus scheme and improve hit rates.

Paul et al more recently introduced the idea of consensus docking using a combination of current docking tools, DOCK, FlexX, and Gold [126]. The program Consdock uses a clustering technique to converge realistic docking poses and rank them accordingly. Using a test set of 100 complexes from the PDB, Consdock significantly surpasses any singly utilised docking protocol. It is apparent then, that the use of a consensus-scoring scheme dramatically enhances the ability of the virtual screen to identify candidate leads.

Muryshev et al describe a procedure for scoring with a combination of a knowledge-based function and an empirical function [127]. Using four different docking algorithms the group observed that Algodock generally outperformed FlexX, DOCK and GOLD in the docking of a set of 19 crystal structures. The general form of the consensus scoring function is represented by,

Eqtn (13): $\Delta G = \sum_{ij} F_{A,B}(r_{i,j})$

where A and B denote the atom types of ligand *i* and protein *j*, and $F_{A,B}$ $(r_{i,j})$ is derived from the Boltzmann equation below where the probablility ($P_{A,B}$ (r)) of finding a ligand atom A from protein atom B at a distance is proportional to,

Eqtn (14): $P_{A,B}(r) \propto \exp\left(\frac{-F_{A,B}(r)}{T}\right)$,

where T=300K.

Finally a linear combination of two knowledge-based and three empirical scoring functions was recently shown to outperform rankings produced by any of the functions singly [128]. Marsden et al applied BLEEP, PMF, GOLD, DOCK and Chemscore to score a set of 205 different protein-ligand complexes from the PDB. A comparison was drawn between the predicted binding affinity of each and the actual observed experimental binding affinity. The consensus scoring function was shown to predict the affinity more

accurately than the individual scoring function. The group also combined the average ranks of each scoring function to produce another consensus function defined as,

$$AR(c) = \sum_{i=1}^{5} - R_i^c / 5$$

where each complex c is ranked according to scoring function i over all scoring functions. This is very useful when many aspects of a scoring function are required to describe the binding in an active site.

1.9 Validation of docking/scoring algorithm performance in the ER

Incorporating information from the active site into a virtual screen gives us the ability to screen for compounds that can inhibit or activate a target such as the ER. Several molecular docking tools (Glide [129], FRED [93], Pro_Leads [130], Dock [65], FlexX [71], Surflex [96] and Gold [72]) have been evaluated as virtual screening platforms using the crystal structure of the ERα as a target. Reproduction of the pose observed in the crystal structure is achieved by generation of multiple ligand poses in the active site of a receptor using these docking tools, and identification of the optimally docked pose using an appropriate scoring function [131]. Prior to application of a molecular docking algorithm in a virtual screening strategy it is necessary firstly to evaluate the ability of the docking program to correctly predict binding modes as observed in the crystal structure. The evaluation process usually involves the comparison of the root mean square deviation (rmsd) of the docked ligand in the active site of a receptor versus the actual co-crystallized version of the protein. This method has been implemented in numerous studies, however, it has also been pointed out that the quality of a docked pose often does not correlate well with the rank of the ligand upon application of a scoring function [132]. For this reason we review all available studies involving the ER and VS in the context of the ability of each docking

and scoring technique to differentiate between actives and inactives in a compound database.

The potential of these programs in identifying a set of known active ligands from a set of drug-like 'decoys' is typically measured using the metric of Enrichment (E). Enrichment is a measure of the proportion of hits retrieved in a subset of compounds compared with the proportion of hits expected from a random sample of compounds,

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

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

Secondly, the False Positive (FP) rate of a ranked dataset gives an excellent indication of the number of molecules needed to look at in order to attain a certain percentage of true positives ('hits'),

$$FP = \frac{Decoy_{sampled}}{Decoy_{Total}} \text{ in } \%\ Hits_{sampled}$$

Thus, if for example the ranks of 5 hits were 1, 4, 6-8 and the total number of random molecules was 995, an FP rate of 0.3% would be expected if one was looking at an 80% true positive rate. To gain a better understanding into the performance of several docking algorithms, we reconcile the differences in false positive rates and also enrichment rates observed with GOLD, FlexX, Glide, Dock and Surflex in four separate

studies involving the use of the same set of actives and 'decoys' screened against the ERα.

Bissantz and co-workers evaluated the proficiency of GOLD1.1, FlexX1.8 and DOCK4.01 in discriminating between actives and inactives using a validation set of 990 compounds selected from the ACD and enriched with 10 known ERα actives [132]. Seven scoring functions including ChemScore [111], Dock score [89], FlexX [71], Fresno [121], GOLD [72], PMF [117] and Score [133] were assessed. 9 out of 10 hits were recovered in the top 2% of the ranked hitlist using a docking/scoring combination of GOLD/Dock [132]. This combination furnished the optimal enrichment rate of 45, but also a false positive (FP) rate of 1.2% for 80% of the true positives. Other docking and scoring combinations were tested against the ER, however, none performed better than GOLD/Dock. For example, using a Dock/Dock approach, ~ 20% false positives remained in 100% of the true positives. Implementing a consensus scoring [134] approach using two or three scoring functions collectively yielded remarkable 'hit' rates as high as 70%. As outlined in the previous section, applying a consensus scoring function generally enhances the prioritization of actives in a dataset. Most importantly however, Bissantz et al concluded that all docking methods allowed a clear differentiation between true hits and random ligands, illustrating that the ER is very suited to VS methods.

Halgren et al [135] recently applied Glide to identify the same 10 low nanomolar ERα antagonists from a set of 990 'decoys' , making a direct comparison possible with results from the Bissantz study. The dockings were also based on utilizing the same procedures for receptor and ligand preparations. For the purpose of this study, Glide was compared to only the scoring functions that are distributed in conjunction with the docking engine (eg. GOLD-docking/GOLD-scoring). In this case Glide gives superior enrichment rates. A look at the first 2% of the database reveals that 7 out of 10 hits were retrieved in the top 20. This converts to an Enrichment rate of 35 and equivalent FP of 1.32% using GlideScore, a modified version of ChemScore, as the scoring function.

However, correlating this to the GOLD-docking/DOCK-scoring combination, as examined by Bissantz et al [132], an enrichment rate of 45 and a False Positive (FP) rate of 1.5% for 90% of the true positives was observed which is comparable. As is often the

case, Halgren et al [135] maintain that in a pharmaceutical setting there may not be the resources to mix docking and scoring functions, however it is nonetheless interesting to note the enhancement of *E* rates when different docking scoring combinations are used.

Jain also validated the performance of Surflex against ERα using the Bissantz dataset [136] and so the same comparisons can be made. Surflex employs its own scoring function taking into account in order of significance, hydrophobic, polar, entropic and solvation terms. A single setting with a protein penetration penalty threshold of –6.0, allows 9/10 actives to be docked correctly, in the top 15 ranked compounds. This translates to an E rate of 60 and an FP rate of 0.7% for a true positive rate of 90%. This is significantly better than the GOLD-docking/DOCK-scoring combination shown by Bissantz. The value of FP is made clear here as one can tell that the rankings of most of the actives are higher using Surflex regardless of the fact that the same E rate is observed. Surflex was seen to yield ~ 2-fold higher FP rate than with the GOLD-docking/DOCK-scoring combination.

The final study involves the validation of a pharmacophore-based evolutionary algorithm (GEMDOCK) by Yang et al [137] using the same ER set. GEMDOCK has the ability to guide the docking by pharmacophore preference extracted from a set of known actives. Utilizing the set proposed by Bissantz et al, the performance of GEMDOCK was assessed and compared with Surflex, DOCK, FlexX and GOLD as above. Analyzing at a true positive rate of 80% where all 10 actives were docked, the FP rates were 1.3% for Surflex, 13.3% for DOCK, 57.8% FlexX and 5.3% for GOLD respectively. Where GEMDOCK was used without pharmacological preferences an FP rate of 1.5% was apparent. With pharmacological preferences turned on, an FP rate of 0% resulted. GEMDOCK appears to be superior to the other methods in retrieving actives from inactives in this case. The benefit of guided docking is highlighted here through consideration of both ligand preferences and binding site Pharmacophore weights. For a clear comparison of FP rates a tabular format is presented in Table 1:

Table **(1)** False positive rates for several docking algorithms

True Positive %	GEMDOCK	Surflex	DOCK	FlexX	GOLD	GOLD/DOCK
80	0	1.3	13.3	57.8	5.3	1.2
90	0.4	1.6	17.4	70.9	8.3	1.5
100	0.9	2.9	18.9	------	23.4	12.1

Table 1 shows clearly that GEMDOCK is a superior docking method for use with the ER compared with the others. Surflex and GOLD-docking/DOCK-scoring perform equivalently in finding 9/10 actives. FlexX is a poor docking algorithm with regards to the ER producing FP rates only slightly above random.

Subsequent evaluation of several docking tools has been carried out using larger decoy sets of ~ 10,000. Stahl and Rarey presented a docking study [138], testing the performance of FlexX in combination with scoring functions FlexX, PLP, PMF, DrugScore [119]. Fifty-five estrogen receptor actives were added to a database of 10,000 drug-like decoys selected from the WDI. A maximum of 50 for enrichment could be achieved in this case for 2% of the ranked database. FlexX-docking/Screenscore-scoring combination gave the best results with an enrichment of ~ 22 for 2% of the database screened. In a follow-up study by Schulz-Gasch and Stahl, two current docking program's, FRED and Glide, were evaluated and compared with previous results obtained using FlexX [49]. A variety of scoring functions were also used in the study, namely FlexX, ScreenScore, Glidescore, glidecomp, and Chemscore, to assess the best strategy for screening. Glide produces an Enrichment factor of ~ 25 with GlideScore, whereas FRED combined with ScreenScore gives an Enrichment of ~ 28.75 with 31 of 55 actives retrieved in the top 2%. FRED is superior to FlexX possibly due to FRED's exhaustive docking approach covering more conformational space than FlexX. In the case of the ER the docking and scoring is reliant on shape and hydrophobicity more so than H-bonding interactions. For this reason FRED in combination with most scoring functions produces a reasonable enrichment over random screening. Corroborating this, our lab has also shown FRED in combination with a chemically aware Gaussian scoring function (Chemgauss) combined with PLP to produce excellent enrichment rates [139] (see Chapter 3 for discussion).

A decrease in Enrichment is observed in this study when compared with the previous studies, because the dataset tested was ~ 7.5 times larger with 55 actives, and the possibility of finding false actives in the top 2% is thus propagated. PLP [140] is seen to be the least effective scoring function in obtaining good enrichment rates in this case.

Baxter et al have also validated a molecular docking method (Pro_Leads) against the ER [141]. ChemScore, was modified to make it more applicable to this docking function. 66,877 molecules were selected from the Chembridge [142] Prime database. Drug-like filters were applied to remove compounds with poor drug-like profiles and leave compounds paralleling lead-like ones. A set of 6 agonists and 12 antagonists were docked in the antagonist conformation of ERα (PDB ID: 1ERR). As expected, the antagonists ranked more highly than agonists. All antagonists were found in the top 1% of the ranked database.

It has been shown in this section the inherent suitability of the ER to the realm of VS with all docking and scoring combinations producing some enrichment. It must be borne in mind that there are some pitfalls associated with enrichment calculations.

1.9.1 Problems associated with enrichment calculations

It is important to note at this stage some issues associated with enrichment calculations. Optimization and validation of docking algorithms for use in the virtual screening process must draw on a decoy set with characteristics and properties reflecting the nature of the actives. Verdonk et al addressed the importance of this in a study based on virtual screening against four targets of therapeutic importance, including ERβ, using GOLD [143]. Calculation of the Heavy Atom Count (HAC) of 20 agonists and 17 antagonists demonstrated a bimodal distribution and therefore should be split into separate groups according to HAC. More importantly, it is concluded that using ATLAS (Astex Technology Ltd.) as the decoy set, where on average the compounds contained in the set are smaller than estrogen antagonists, the enrichments obtained using a random library are significantly higher than those used for a focused library where 1D properties are similar. This illustrates the significance of using a decoy set with similar properties to

that of the active set and shows that the differences seen in the Enrichments of the above studies may be attributed to the quality of the larger (10,000) decoy sets.

Muegge and Enyedy showed that enrichment depends on the decoy set also using kinase as a target and employing Dock, Glide, Ligandfit [144]. To assess this the group generated three separate databases of 10,000 compounds. The first from the MDDR [145], was filtered using drug-like criteria outlined in a previous publication [146, 147]. The second from a diverse collection of compounds that were either found active against kinase targets screened by HTS or, using a neural network to produce a kinase-like library with a higher hit rate for kinases than a random subset. The third set was extracted from the ACD [148]. It would be assumed that it is hardest to discriminate between actives and inactives in the set containing other kinases. However, the lowest enrichment rates were observed with the MDDR set as it contains potentially very diverse structures. Distinguishing actives from inactives using ACD gave the highest hit rates not surprisingly. Most importantly though, this shows the need also for applying a decoy set that is diverse but also with 'drug-like' rather than 'lead-like' characteristics in order to attain valid enrichment rates [149-151].

1.10 Post-filtering Procedures

Post-processing protocols are incorporated in many vHTS procedures to implement a degree of target bias that is often mandated by the medicinal chemist. Requirements such as hydrogen bonding to specific residues of the active site that are known to be important in the ligand binding process can be incorporated. During the docking procedure using FRED2.11 [152] a pharmacophore can be selected by SMARTS string that codes for the presence of certain functional group or atom types. Another program, Magnet [153] allows the user to home-in on important features of the docked structure such as H-bonding, van der Waals and the % of ligand surface area buried. Importantly, features can also be weighed to bias the score towards selection of correct poses and orientations. FlexPharm [154] also allows the user to include constraints such as those observed by a certain pharmacophore including specific interactions and occupancy volume of the ligand.

Overall the inclusion of a post-filter appears to enhance the screening process as scoring functions alone do not account for specific interaction but rather a combination of all.

1.11 Virtual Library Generation

To better exploit chemical space, combinatorial chemistry has been used to synthesise large and diverse libraries. The combinatorial approach works on two levels with respect to VS. Firstly large and diverse libraries can be generated and synthesis focussed on smaller subsets to prevent combinatorial explosion and screening all molecules. This method allows design of diverse, drug-like subsets or those focussed against a particular target for screening. Secondly it can be usefully applied to generate alternatives to a lead compound derived from a core scaffold.

Measuring molecular diversity generally involves calculation of molecular descriptors such as ClogP, molecular weight, free energy of solvation, BCUT descriptors, structural motifs, toplogical indices. pharmacophores etc. Brown and Martin have shown that the use of substructure keys can be as powerful as molecular descriptors in describing biological activity [155]. The quantification of diversity based on the method chosen to assign or weight a selection of compounds can be a number of techniques of which clustering [156], maximum dissimilarity search algorithm [157], and factor analysis [158] generally fulfill.

Wang et al describe a technique involving the prediction of drug feasibility of compounds [159]. The method utilizes the concept of the multilevel chemical compatibility (MLCC) between a compound and a drug library as a measure of the drug-like character of a compound. The method suggested that ~80% of all viable types of drug are contained within the drug set used (MDDR & CMC). It also predicted that no known problematic compound was drug-like and that the tool was efficacious enough to be applied to large combinatorial libaries. A multi-objective genetic algorithm (MOGA) was recently implemented to reduce combinatorial library size and converge towards those molecules possessing drug-like physiochemical property profiles [160].

To derive a subset with characteristics and properties similar to those of a known active that exerts a therapeutic effect against a known target, a number of strategies have

been implemented. A program MoSELECT developed by Gillet et al [161], does this by searching the product-space of a virtual combinatorial library to generate a subset of solutions where each represents a combinatorial subset of the virtual library. In this manner, large combinatorial libraries can be reduced to produce smaller sets of between 2K-20k molecules. Bravi et al also detailed an algorithm PLUMS [162], that reduces the size of a virtual combinatorial library by an iterative process until a targeted subset is produced that reflects a balance between effectiveness (ratio between the number of virtual hits in the sub-library and the total number of virtual hits in the full library) and efficiency (ratio between the number of virtual hits in the sub-library and the size of the sub-library). An elegant study by Jamois et al describes an approach termed as on the fly optimization (OTFO) where descriptors can be computed as needed within the subset optimization cycle [163]. The method is extremely fast, robust and allows focused subsets to be generated while sampling only a fraction of a virtual library.

Virtual combinatorial chemistry also plays a large part in the design of libraries with a range of substituents and functional groups attached at user-defined points on a core scaffold. Algorithms such as PRO_SELECT [164] and CombiDock [165] work by incrementally building up a set of molecules from a core. Functional groups and substituents are attached to the scaffold and scored independently until a library is built. PRO_SELECT, however, can move a step higher by not only scoring each substituent but also discriminating by 2-D similarity and ease of synthesis. PRO_SELECT has been applied successfully to the development of a series of highly potent factor Xa inhibitors [166].

More minimalistic approaches have been detailed in software such as COREGEN [167]and SMILIB [168] with both using a scaffold-linker-functional group concept. The user defines a core scaffold and the rest of the library is enumerated accordingly by adding of fragments, which in the case of SMILIB are encoded though SMILES [169] notation. An application of such program, SLF_Libmaker, was very recently reported by Krier et al, to the structure based optimization of a Phosphodiesterase 4 Inhibitor [170]. The program SLF_Libmaker differed slightly to the previously mentioned two with regard to it choice of fragments. A second round of optimization can be undertaken to ensure that the building blocks utilized explore sufficient areas of chemical space.

1.12 Virtual Screening for Ligands of Nuclear Receptors

The main body of research carried out in this thesis concerns studies regarding the vHTS of the Estrogen Receptor with the aim of identifying new modulators. The ER has been shown to be an excellent choice for VS. Characteristics of the ER such as a relatively rigid binding site, specific H-bonding residues and Helix-12 movement are common to the more broad family of Nuclear Receptors. For this reason we review to date VS of the NR family and describe the successes and associated novel ligands discovered. Table 2 below illustrates the recent VS applications to the NR family and those yet to be screened with crystal structures available in the PDB. For this section we only review those structures that have a ligand co-crystallised. Several NR orphan receptors lack a known endogenous binding ligand and have therefore not been included.

Table **(2)** Crystal structures of Nuclear Receptors and their co-crystallised ligands indicating whether VS has been carried out or not.

TARGET	NAME *	ISOFORM	PDB ID	VS (Ref)
ER	4-HYDROXYTAMOXIFEN	α	3ERT	Yes (149)
	GENISTEIN	α	1X7R	No
	(2S,3R)-2-(4-{2-[(3R,4R)-3,4-DIMETHYLPYRROLIDIN-1-YL]ETHOXY}PHENYL)-3-(4-HYDROXYPHENYL)-2,3-DIHYDRO-1,4 BENZOXATHIIN-6-OL	α	1XP1	No
	(2S,3R)-2-(4-{2-[(3S,4S)-3,4-DIMETHYLPYRROLIDIN-1-YL]ETHOXY}PHENYL)-3-(4-HYDROXYPHENYL)-2,3-DIHYDRO-1,4 BENZOXATHIIN-6-OL	α	1XP6	No
	(2S,3R)-3-(4-HYDROXYPHENYL)-2-(4-{[(2S)-2-PYRROLIDIN-1-YLPROPYL]OXY}PHENYL)-2,3-DIHYDRO-1,4-BENZOXATHIIN-6 OL	α	1XP9	No
	(2S,3R)-3-(4-HYDROXYPHENYL)-2-(4-{[(2R)-2-PYRROLIDIN-1-YLPROPYL]OXY}PHENYL)-2,3-DIHYDRO-1,4-BENZOXATHIIN-6 OL	α	1XPC	No
	(R,R)-5,11-CIS-DIETHYL-5,6,11,12-TETRAHYDROCHRYSENE-2, 8-DIOL	α	1L2I	No
	(2S,3R)-2-(4-(2-(PIPERIDIN-1-YL)ETHOXY)PHENYL)-2,3-DIHYDRO-3-(4-HYDROXYPHENYL)BENZO[B][1,4]OXATHIIN-6-OL	α	1SJO	No
	(1S)-1-{4-[(9AR)-OCTAHYDRO-2H-PYRIDO[1,2-A]PYRAZIN-2-YL]PHENYL}-2-PHENYL-1,2,3,4-TETRAHYDROISOQUINOLIN-6-O	α	1XQC	No
	DIETHYLSTILBESTROL	α	3ERD	Yes (150)
	GENISTEIN	β	1QKM	Yes (151)
	5-HYDROXY-2-(4-HYDROXYPHENYL)-1-BENZOFURAN-7-CARBONITRILE	β	1X76	No
	DIETHYLSTILBESTROL	γ	1SP9	No
	4-HYDROXYTAMOXIFEN	γ	1S9Q	No
RAR	4-[(1E)-2-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRONAPHTHALEN-2-YL)PROP-1-ENYL]BENZOIC ACID	β	1XAP	No
	R-3-FLUORO-4-[2-HYDROXY-2-(5,5,8,8-TETRAMETHYL-5,6,7,8,-TETRAHYDRO-NAPHTALEN-2- YL)-ACETYLAMINO]-BENZOIC ACID	γ	1EXA	No
	S-3-FLUORO-4-[2-HYDROXY-2-(5,5,8,8-TETRAMETHYL-5,6,7,8,-TETRAHYDRO-NAPHTALEN-2- YL)-ACETYLAMINO]-BENZOIC ACID	γ	1EXX	No

	6-[HYDROXY-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRO-NAPHTALEN-2-YL)-METHYL]-NAPHTALENE-2-CARBOXYLIC ACID	γ	1FCX	No
	6-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRO-NAPHTALENE-2-CARBONYL)-NAPHTALENE-2-CARBOXYLIC ACID	γ	1FCY	No
	4-[3-OXO-3-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRO-NAPHTHALEN-2-YL)-PROPENYL]-BENZOIC ACID	γ	1FCZ	No
	6-[HYDROXYIMINO-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRO-NAPHTALEN-2-YL)-METHYL]-NAPHTALENE-2-CARBOXYLIC ACID	γ	1FDO	No
RXR	RETINOIC ACID	γ	2LBD	No
	2,4-THIAZOLIDIINEDIONE, 5-[[4-[2-(METHYL-2-PYRIDINYLAMINO)ETHOXY]PHENYL]METHYL]-(9CL)	α	1FM6	No
	2,4-THIAZOLIDIINEDIONE, 5-[[4-[2-(METHYL-2-PYRIDINYLAMINO)ETHOXY]PHENYL]METHYL]-(9CL)	α	1FM9	No
	RETINOIC ACID	α	1G5Y	No
	DOCOSA-4,7,10,13,16,19-HEXAENOIC ACID	α	1MV9	No
PPAR	4-[2-(5,5,8,8-TETRAMETHYL-5,6,7,8-TETRAHYDRO-NAPHTHALEN-2-YL)-[1,3]DIOXOLAN-2-YL]-BENZOIC ACID	α	1MVC	No
	(2S)-2-ETHOXY-3-[4-(2-{4-[(METHYLSULFONYL) OXY]PHENYL}ETHOXY)PHENYL]PROPANOIC ACID	α	1I7G	No
VDR	APO-STRUCTURE	γ	1PRG	No
	5-{2-[1-(5-HYDROXY-1,5-DIMETHYL-HEXYL)-7A-METHYL-OCTAHYDRO-INDEN-4-YLIDENE]- ETHYLIDENE}-4-METHYLENE-CYCLOHEXANE-1,3-DIOL	-	1DB1	No
	5-(2-{1-[1-(4-ETHYL-4-HYDROXY-HEXYLOXY)-ETHYL]-7A-METHYL-OCTAHYDRO-INDEN-4-YLIDENE}-ETHYLIDENE)-4-METHYLENE-CYCLOHEXANE-1,3-DIOL	-	1IE8	No
	5-{2-[1-(5-HYDROXY-1,5-DIMETHYL-HEXYL)-7A-METHYL-OCTAHYDRO-INDEN-4-YLIDENE]-ETHYLIDENE}-4-METHYLENE-CYCLOHEXANE-1,3-DIOL	-	1IE9	No
PR	5-{2-[1-(5-HYDROXY-1,5-DIMETHYL-HEXYL)-7A-METHYL-OCTAHYDRO-INDEN-4-YLIDENE]-ETHYLIDENE}-2-METHYLENE-CYCLOHEXANE-1,3-DIOL	-	1RJK	No
	PROGESTERONE	-	1A28	No
	NORETHINDRONE	-	1SQN	No
	MOMETASONE FUROATE	-	1SR7	No
	5-(4,4-DIMETHYL-2-THIOXO-1,4-DIHYDRO-2H-3,1-BENZOXAZIN-6-YL)-1-METHYL-1H-PYRROLE-2-CARBONITRILE	-	1ZUC	No
AR	DIHYDROTESTOSTERONE	-	1I37	No
	DIHYDROTESTOSTERONE	-	1T63	No
	DIHYDROTESTOSTERONE	-	1T65	No
	DIHYDROTESTOSTERONE	-	1T7T	No

<2.2 Å X-ray structures selected for possible VS application.

Abbreviations: ER, Estrogen Receptor; RAR, Retinoic Acid Receptor; RXR, Retinoid X Receptor; VDR, Vitamin D Receptor; PPAR, Peroxisome proliferator activated receptor; PR, Progesterone Receptor; AR, Androgen Receptor.

* Name as it appears in the Protein Data Bank (www.rcsb.org)

Schapira et al demonstrated the efficacy and feasibility of receptor based virtual screening in identifying inhibitors of the Nuclear Receptor family [171]. Several antagonists of the human alpha Retinoic Acid Receptor (RAR) were recently discovered in a separate study [171]. The initial structure of RXRα was derived by homology modelling from a model of RARγ complexed with a known antagonist (AGN193109) developed from a docking study. Information such as the positioning of Helix-12 the LBD in the antagonist conformation of estrogen receptor alpha was extracted, as the molecular conformation induced in the ER antagonist form is similar. Molsoft ICM 2.7 [172] was utilised and both receptor and ligand flexibility were introduced for the docking. Docking of a library of 153,000 compounds led to the discovery of two novel antagonists and a novel agonist specific for RAR. The structures of two novel anatagonists of the RARα are outlined below in Figure 5.

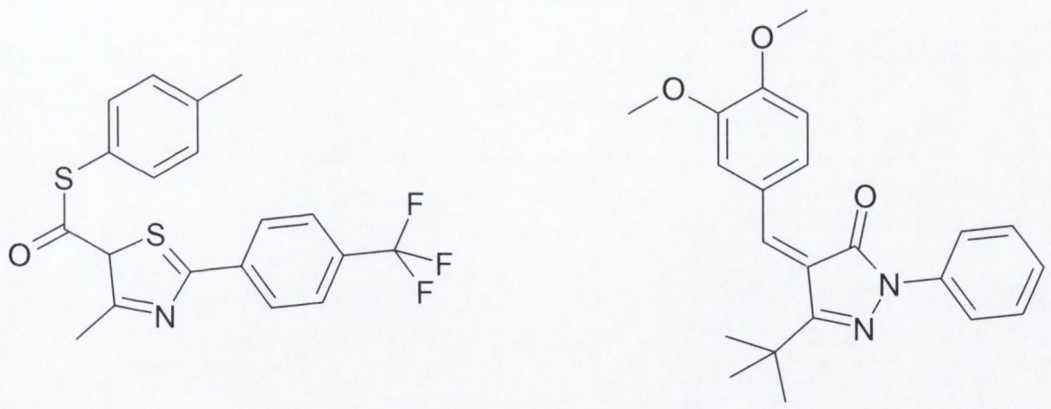


Fig **(5)** Two antagonists of RAR discovered by VS using ICM.

This study demonstrates the two crucial aspects of computational drug design where rational design against receptor model was undertaken and then a library of compounds screened against it.

In the second study also by Schapira, thyroid hormone receptor antagonists were sought after also using the ICM virtual screening module. Again a computational model of the antagonist bound ligand to the thyroid receptor was built through structural

homology, derived from the crystal structure of ERα with tamoxifen bound. A set of fourteen novel and diverse ligands, prioritised from a database of 250,000 compounds, were found to antagonise the TR with an IC50 value of 1.5 to 30μm [173]. The highest affinity antagonist was selected as depicted in Figure 6 and a small virtual library of compounds based on the hits was generated. The next generation of inhibitors, exhibited higher affinity than those previously tested.

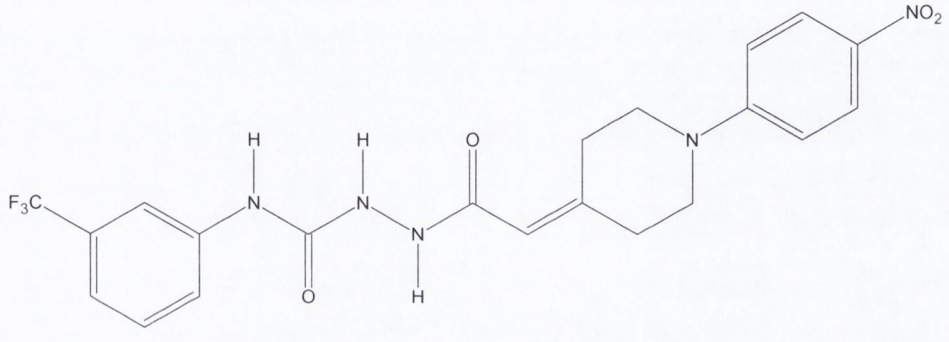


Fig **(6)** Antagonist identified by VS and developed using Virtual Library.

The DockCrunch project [174] was undertaken to analyse whether virtual screening has value in identifying leads for the estrogen receptor in its agonist and antagonist form. A library of 1.5 million commercially available compounds was seeded with 20 known agonists and antagonists. The Brookhaven codes 1ERE (Estradiol) and 3ERT (Raloxifene) were downloaded from the PDB and used as targets fro the screen. A set of filters was applied to the library to remove compounds displaying nondrug-like physiochemical properties. 1.1 million structures were yielded. The PRO_LEADS [175] docking algorithm was then applied to flexibly dock the library of compounds. Using ChemScore [111], the PRO_LEADS implementation of the scoring function, good separation between those known agonists/antagonists and the random set was found, and a novel set of 37 ligands were obtained from the screening process with high binding affinity to the receptor.

Table **(3)** Binding Affinities for 37 compounds retrieved by VS using PRO_LEADS

Binding Affinity (Ki)	Number of Compounds
<10nM	2
<100nM	14
<300nM	21

Two compounds in particular displayed <10nM activity when assayed using a competitive radio-ligand binding assay. Waszkowycz et al [176] clearly demonstrate the possibility of finding novel chemotypes in this study through application of VS.

Identification of novel ERα antagonists was assessed utilizing a cell-based assay and also virtual screening [177]. Both methods have been shown to complement each other previously [178, 179]. The mode of action of these antagonists however differs from the normal mode of action. They are designed to disrupt the agonist action of the ERα LBD, which usually recruits the co-activator peptide SRC-3 to initiate transcription.

SRC-3 has been shown to be overexpressed in 60% of breast cancers and inhibition of the interaction between it and the ER would prove beneficial [180]. The x-ray structure of ERα bound with Diethylstibesterol (DES) and co-crystallised with an LxxLL peptide was utilised. The binding site was derived from site points identified using MCSS2SPTS [181]. DOCK4.01 was used as the docking engine and more specifically a recently published method, PharmDOCK [181], was implemented. Screening of the Available Chemicals Directory against the target was carried out, and a set of 36 ligands was identified from the virtual screen. 14 possessed IC50s ranging from 0.79μM to 31μM. Two of these new classes of ERα inhibitors are shown in Figure 7. To corroborate this particular mode of inhibition, a screening assay was carried out with no estradiol displacement observed. Consequently, a different set was retrieved using the cell-based assay with equivalent potency.

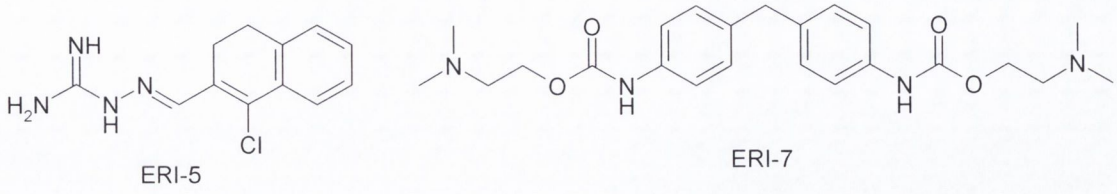


Fig **(7)** Two novel SRC-3 inhibitors identified from screen

Finally, VS of ERβ was undertaken by Zhao et al to find plant-based selective ligands [182]. The crystallographic structure of ERβ LBD with genistein co-crystallised, was utilised to screen an 'in-house' natural source chemical collection. GOLD2.0 was employed as the docking engine. Prior to database screening an initial validation was carried out using a test set with genistein incorporated. Application of the GoldScore scoring function led to close reproduction of the actual X-ray structure (rmsd = 0.3566). VS of ERβ using the whole database led to 500 well ranked molecules. Visual analysis of the ranked hitlist allowed further selection of 100 molecules to undergo a more restrictive analysis with Affinity (Accelrys Inc.[183]). Those molecules portraying a 'favourable' H-bond interaction with His475, complementarity, proper binding mode and finally structural diversity were passed through a Lipinski filter and had their respective Blood-Brain Barrier properties predicted. A set of 12 molecules passed this process and an FP assay was applied to determine their binding affinities experimentally. Three out of twelve molecules displayed >100-fold selectivity toward ERβ over ERα and an example is illustrated below.

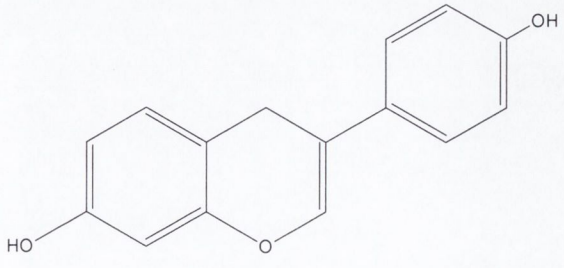


Fig **(8)** A novel ER-beta plant-based molecule identified from the screen with >100-fold selectivity.

The benefits of employing virtual screening techniques and rational design methods to a target such as the ER are multifold, as we have shown case studies where molecular docking algorithms can identify superior binding modes over unsatisfactory ones and distinguish between actives and inactives from a database comprised of both. Also, it is clear that molecular docking can be applied to the ER to discover diverse and novel molecules with the ability to modulate the actions of the ER.

1.13 Conclusions

We have provided a detailed review of the structure based drug design process and its subsequent application to the ER in a number of studies. The successes of vHTS of the ER and Nuclear Receptors more generally by numerous groups have also been detailed. It is clear that the vHTS process requires knowledge of a target and more specifically its active site. There are numerous ways in which one can represent ligand and receptor flexibility and the cheminformatician must discern to what degree each needs to be accounted for in the process. Also a plethora of methods exist that allow docking of molecules into the active site of interest and each bears its own advantages and disadvantages. To permit actives to be retrieved from a docking run one must also decide which scoring function to use that best describes that nature of the binding process of the protein of interest. Finally, it has been outlined in this chapter pitfalls associated with the vHTS process and some solutions published that allow these hurdles to be overcome. The applicability of the ER and NR's in general to the vHTS process has been described and chapters 2-5 will provide additional studies using the ER as a target.

1.14 References

1. Cauley, J. A.; Robbins, J.; Chen, Z.; Cummings, S. R.; Jackson, R. D.; LaCroix, A. Z.; LeBoff, M.; Lewis, C. E.; McGowan, J.; Neuner, J.; Pettinger, M.; Stefanick, M. L.; Wactawski-Wende, J.; Watts, N. B., Effects of estrogen plus progestin on risk of fracture and bone mineral density: the Women's Health Initiative randomized trial. *Jama* **2003,** 290, (13), 1729-38.
2. Chlebowski, R. T.; Wactawski-Wende, J.; Ritenbaugh, C.; Hubbell, F. A.; Ascensao, J.; Rodabough, R. J.; Rosenberg, C. A.; Taylor, V. M.; Harris, R.; Chen, C.; Adams-Campbell, L. L.; White, E., Estrogen plus progestin and colorectal cancer in postmenopausal women. *N Engl J Med* **2004,** 350, (10), 991-1004.
3. Beral, V., Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet* **2003,** 362, (9382), 419-27.
4. Prasad, R.; Boland, G. P.; Cramer, A.; Anderson, E.; Knox, W. F.; Bundred, N. J., Short-term biologic response to withdrawal of hormone replacement therapy in patients with invasive breast carcinoma. *Cancer* **2003,** 98, (12), 2539-46.
5. Ireland, T. N. C. R., http://www.ncri.ie/pubs/pubfiles/report-2000.pdf.
6. Jensen EV, J. H., Basic guides to the mechanism of estrogen action. *Recent Prog Horm Res* **1962,** 18, 387-414.
7. Toft, D.; Gorski, J., A receptor molecule for estrogens: isolation from the rat uterus and preliminary characterization. *Proc Natl Acad Sci U S A* **1966,** 55, (6), 1574-81.
8. Toft, D.; Shyamala, G.; Gorski, J., A receptor molecule for estrogens: studies using a cell-free system. *Proc Natl Acad Sci U S A* **1967,** 57, (6), 1740-3.
9. Harper, M. J.; Walpole, A. L., A new derivative of triphenylethylene: effect on implantation and mode of action in rats. *J Reprod Fertil* **1967,** 13, (1), 101-19.
10. Harper, M. J.; Walpole, A. L., Mode of action of I.C.I. 46,474 in preventing implantation in rats. *J Endocrinol* **1967,** 37, (1), 83-92.
11. Lippman, M. E.; Bolan, G., Oestrogen-responsive human breast cancer in long term tissue culture. *Nature* **1975,** 256, (5518), 592-3.
12. Black, L. J.; Goode, R. L., Uterine bioassay of tamoxifen, trioxifene and a new estrogen antagonist (LY117018) in rats and mice. *Life Sci* **1980,** 26, (17), 1453-8.
13. Jones, C. D.; Jevnikar, M. G.; Pike, A. J.; Peters, M. K.; Black, L. J.; Thompson, A. R.; Falcone, J. F.; Clemens, J. A., Antiestrogens. 2. Structure-activity studies in a series of 3-aroyl-2-arylbenzo[b]thiophene derivatives leading to [6-hydroxy-2-(4-hydroxyphenyl)benzo[b]thien-3-yl] [4-[2-(1-piperidinyl)ethoxy]-phenyl]methanone hydrochloride (LY156758), a remarkably effective estrogen antagonist with only minimal intrinsic estrogenicity. *J Med Chem* **1984,** 27, (8), 1057-66.
14. Beall, P. T.; Misra, L. K.; Young, R. L.; Spjut, H. J.; Evans, H. J.; LeBlanc, A., Clomiphene protects against osteoporosis in the mature ovariectomized rat. *Calcif Tissue Int* **1984,** 36, (1), 123-5.
15. Schmitt, E.; Dekant, W.; Stopper, H., Assaying the estrogenicity of phytoestrogens in cells of different estrogen sensitive tissues. *Toxicol In Vitro* **2001,** 15, (4-5), 433-9.
16. Ohno, K.; Suzuki, S.; Fukushima, T.; Maeda, M.; Santa, T.; Imai, K., Study on interactions of endocrine disruptors with estrogen receptor using fluorescence polarization. *Analyst* **2003,** 128, (8), 1091-6.
17. Mueller, S. O.; Simon, S.; Chae, K.; Metzler, M.; Korach, K. S., Phytoestrogens and their human metabolites show distinct agonistic and antagonistic properties on estrogen receptor alpha (ERalpha) and ERbeta in human cells. *Toxicol Sci* **2004,** 80, (1), 14-25.
18. Beck, V.; Rohr, U.; Jungbauer, A., Phytoestrogens derived from red clover: An alternative to estrogen replacement therapy? *J Steroid Biochem Mol Biol* **2005,** 94, (5), 499-518.
19. Blizzard, T. A.; Dininno, F.; Morgan, J. D., 2nd; Chen, H. Y.; Wu, J. Y.; Kim, S.; Chan, W.; Birzin, E. T.; Yang, Y. T.; Pai, L. Y.; Fitzgerald, P. M.; Sharma, N.; Li, Y.; Zhang, Z.; Hayes, E. C.; Dasilva, C. A.; Tang, W.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. Part 9: Dihydrobenzoxathiin SERAMs with alkyl substituted pyrrolidine side chains and linkers. *Bioorg Med Chem Lett* **2005,** 15, (1), 107-13.
20. Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M., Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997,** 389, (6652), 753-8.

21. Kim, S.; Wu, J. Y.; Birzin, E. T.; Frisch, K.; Chan, W.; Pai, L. Y.; Yang, Y. T.; Mosley, R. T.; Fitzgerald, P. M.; Sharma, N.; Dahllund, J.; Thorsell, A. G.; DiNinno, F.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. II. Discovery of benzoxathiins as potent, selective estrogen receptor alpha modulators. *J Med Chem* **2004,** 47, (9), 2171-5.
22. Manas, E. S.; Xu, Z. B.; Unwalla, R. J.; Somers, W. S., Understanding the selectivity of genistein for human estrogen receptor-beta using X-ray crystallography and computational methods. *Structure (Camb)* **2004,** 12, (12), 2197-207.
23. Pike, A. C.; Brzozowski, A. M.; Walton, J.; Hubbard, R. E.; Bonn, T.; Gustafsson, J. A.; Carlquist, M., Structural aspects of agonism and antagonism in the oestrogen receptor. *Biochem Soc Trans* **2000,** 28, (4), 396-400.
24. Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, P.; Fournier, B.; Geiser, M.; Halleux, C.; Kallen, J.; Keller, H.; Ramage, P., Selective estrogen receptor modulators with conformationally restricted side chains. Synthesis and structure-activity relationship of ERalpha-selective tetrahydroisoquinoline ligands. *J Med Chem* **2005,** 48, (2), 364-79.
25. Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, P.; Fournier, B.; Halleux, C.; Kallen, J.; Keller, H.; Schlaeppi, J. M.; Stark, W., Estrogen receptor modulators: identification and structure-activity relationships of potent ERalpha-selective tetrahydroisoquinoline ligands. *J Med Chem* **2003,** 46, (14), 2945-57.
26. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L., The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998,** 95, (7), 927-37.
27. Shiau, A. K.; Barstad, D.; Radek, J. T.; Meyers, M. J.; Nettles, K. W.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A.; Agard, D. A.; Greene, G. L., Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nat Struct Biol* **2002,** 9, (5), 359-64.
28. Kelminski, A., The Study of Tamoxifen and Raloxifene (STAR trial) for the prevention of breast cancer. *Hawaii Med J* **2002,** 61, (9), 209-10.
29. Cohen, F. J.; Watts, S.; Shah, A.; Akers, R.; Plouffe, L., Jr., Uterine effects of 3-year raloxifene therapy in postmenopausal women younger than age 60. *Obstet Gynecol* **2000,** 95, (1), 104-10.
30. Fisher, B.; Costantino, J. P.; Wickerham, D. L.; Redmond, C. K.; Kavanah, M.; Cronin, W. M.; Vogel, V.; Robidoux, A.; Dimitrov, N.; Atkins, J.; Daly, M.; Wieand, S.; Tan-Chiu, E.; Ford, L.; Wolmark, N., Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* **1998,** 90, (18), 1371-88.
31. Plouffe, L., Jr., Selective estrogen receptor modulators (SERMs) in clinical practice. *J Soc Gynecol Investig* **2000,** 7, (1 Suppl), S38-46.
32. Dutertre, M.; Smith, C. L., Molecular mechanisms of selective estrogen receptor modulator (SERM) action. *J Pharmacol Exp Ther* **2000,** 295, (2), 431-7.
33. Miller, C. P., SERMs: evolutionary chemistry, revolutionary biology. *Curr Pharm Des* **2002,** 8, (23), 2089-111.
34. Henke, B. R.; Heyer, D., Recent advances in estrogen receptor modulators. *Curr Opin Drug Discov Devel* **2005,** 8, (4), 437-48.
35. Gasco, M.; Argusti, A.; Bonanni, B.; Decensi, A., SERMs in chemoprevention of breast cancer. *Eur J Cancer* **2005,** 41, (13), 1980-9.
36. Nilsson, S.; Koehler, K. F., Oestrogen receptors and selective oestrogen receptor modulators: molecular and cellular pharmacology. *Basic Clin Pharmacol Toxicol* **2005,** 96, (1), 15-25.
37. Mestres, J., Virtual screening: a real screening complement to high-throughput screening. *Biochem Soc Trans* **2002,** 30, (4), 797-9.
38. Jenkins, J. L.; Kao, R. Y.; Shapiro, R., Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. *Proteins* **2003,** 50, (1), 81-93.
39. Hann, M. M.; Oprea, T. I., Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **2004,** 8, (3), 255-63.
40. Hou, T.; Xu, X., Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* **2004,** 10, (9), 1011-33.
41. Stahura, F. L.; Bajorath, J., New methodologies for ligand-based virtual screening. *Curr Pharm Des* **2005,** 11, (9), 1189-202.

42. Tickle, I.; Sharff, A.; Vinkovic, M.; Yon, J.; Jhoti, H., High-throughput protein crystallography and drug discovery. *Chem Soc Rev* **2004,** 33, (8), 558-65.
43. Wieman, H.; Tondel, K.; Anderssen, E.; Drablos, F., Homology-based modelling of targets for rational drug design. *Mini Rev Med Chem* **2004,** 4, (7), 793-804.
44. Bertini, I.; Calderone, V.; Cosenza, M.; Fragai, M.; Lee, Y. M.; Luchinat, C.; Mangani, S.; Terni, B.; Turano, P., Conformational variability of matrix metalloproteinases: beyond a single 3D structure. *Proc Natl Acad Sci U S A* **2005,** 102, (15), 5334-9.
45. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** 28, (1), 235-42.
46. Hooft, R. W.; Vriend, G.; Sander, C.; Abola, E. E., Errors in protein structures. *Nature* **1996,** 381, (6580), 272.
47. 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.
48. Razandi, M.; Pedram, A.; Greene, G. L.; Levin, E. R., Cell membrane and nuclear estrogen receptors (ERs) originate from a single transcript: studies of ERalpha and ERbeta expressed in Chinese hamster ovary cells. *Mol Endocrinol* **1999,** 13, (2), 307-19.
49. 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.
50. Manas, E. S.; Unwalla, R. J.; Xu, Z. B.; Malamas, M. S.; Miller, C. P.; Harris, H. A.; Hsiao, C.; Akopian, T.; Hum, W. T.; Malakian, K.; Wolfrom, S.; Bapat, A.; Bhat, R. A.; Stahl, M. L.; Somers, W. S.; Alvarez, J. C., Structure-based design of estrogen receptor-beta selective ligands. *J Am Chem Soc* **2004,** 126, (46), 15106-19.
51. Bourguet, W.; Germain, P.; Gronemeyer, H., Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol Sci* **2000,** 21, (10), 381-8.
52. Folkertsma, S.; van Noort, P. I.; Brandt, R. F.; Bettler, E.; Vriend, G.; de Vlieg, J., The nuclear receptor ligand-binding domain: a family-based structure analysis. *Curr Med Chem* **2005,** 12, (9), 1001-16.
53. Shaw, J. A.; Udokang, K.; Mosquera, J. M.; Chauhan, H.; Jones, J. L.; Walker, R. A., Oestrogen receptors alpha and beta differ in normal human breast and breast carcinomas. *J Pathol* **2002,** 198, (4), 450-7.
54. Matthews, J.; Gustafsson, J. A., Estrogen signaling: a subtle balance between ER alpha and ER beta. *Mol Interv* **2003,** 3, (5), 281-92.
55. Gustafsson, J. A., Estrogen receptor beta--a new dimension in estrogen mechanism of action. *J Endocrinol* **1999,** 163, (3), 379-83.
56. Levenson, A. S.; Jordan, V. C., The key to the antiestrogenic mechanism of raloxifene is amino acid 351 (aspartate) in the estrogen receptor. *Cancer Res* **1998,** 58, (9), 1872-5.
57. Paige, L. A.; Christensen, D. J.; Gron, H.; Norris, J. D.; Gottlin, E. B.; Padilla, K. M.; Chang, C. Y.; Ballas, L. M.; Hamilton, P. T.; McDonnell, D. P.; Fowlkes, D. M., Estrogen receptor (ER) modulators each induce distinct conformational changes in ER alpha and ER beta. *Proc Natl Acad Sci U S A* **1999,** 96, (7), 3999-4004.
58. Nettles, K. W.; Greene, G. L., Ligand control of coregulator recruitment to nuclear receptors. *Annu Rev Physiol* **2005,** 67, 309-33.
59. Kian Tee, M.; Rogatsky, I.; Tzagarakis-Foster, C.; Cvoro, A.; An, J.; Christy, R. J.; Yamamoto, K. R.; Leitman, D. C., Estradiol and selective estrogen receptor modulators differentially regulate target genes with estrogen receptors alpha and beta. *Mol Biol Cell* **2004,** 15, (3), 1262-72.
60. Lloyd, D. G.; Hughes, R. B.; Zisterer, D. M.; Williams, D. C.; Fattorusso, C.; Catalanotti, B.; Campiani, G.; Meegan, M. J., Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J Med Chem* **2004,** 47, (23), 5612-5.
61. Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M., Flexible estrogen receptor modulators: design, synthesis, and antagonistic effects in human MCF-7 breast cancer cells. *J Med Chem* **2001,** 44, (7), 1072-84.
62. Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M., Ethyl side-chain modifications in novel flexible antiestrogens--design, synthesis and biological efficacy in assay against the MCF-7 breast tumor cell line. *Anticancer Drug Des* **2001,** 16, (1), 57-69.

63. Minutolo, F.; Antonello, M.; Bertini, S.; Ortore, G.; Placanica, G.; Rapposelli, S.; Sheng, S.; Carlson, K. E.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A.; Macchia, M., Novel estrogen receptor ligands based on an anthranylaldoxime structure: role of the phenol-type pseudocycle in the binding process. *J Med Chem* **2003,** 46, (19), 4032-42.
64. Kekenes-Huskey, P. M.; Muegge, I.; von Rauch, M.; Gust, R.; Knapp, E. W., A molecular docking study of estrogenically active compounds with 1,2-diarylethane and 1,2-diarylethene pharmacophores. *Bioorg Med Chem* **2004,** 12, (24), 6527-37.
65. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **2001,** 15, (5), 411-28.
66. Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M., Side-chain flexibility in proteins upon ligand binding. *Proteins* **2000,** 39, (3), 261-8.
67. Lorber, D. M.; Shoichet, B. K., Flexible ligand docking using conformational ensembles. *Protein Sci* **1998,** 7, (4), 938-50.
68. OMEGA 1.8.1, distributed by Openeye Scientific Software.
69. Lorber, D. M.; Shoichet, B. K., Hierarchical docking of databases of multiple ligand conformations. *Curr Top Med Chem* **2005,** 5, (8), 739-49.
70. Joseph-McCarthy, D.; Thomas, B. E. t.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C., Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins* **2003,** 51, (2), 172-88.
71. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** 261, (3), 470-89.
72. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, (3), 727-48.
73. Zavodszky, M. I.; Kuhn, L. A., Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci* **2005,** 14, (4), 1104-14.
74. Jiang, F.; Kim, S. H., "Soft docking": matching of molecular surface cubes. *J Mol Biol* **1991,** 219, (1), 79-102.
75. Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K., Soft docking and multiple receptor conformations in virtual screening. *J Med Chem* **2004,** 47, (21), 5076-84.
76. Carlson, H. A.; Masukawa, J.; McCammon, J. A., A method for including the dynamic fluctuations of a protein in computer-aided drug design. *J. Phys. Chem* **1999,** 103, 10213-19.
77. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T., FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* **2001,** 308, (2), 377-95.
78. Yoon, S.; Welsh, W. J., Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J Chem Inf Comput Sci* **2004,** 44, (1), 88-96.
79. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M., Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **2004,** 47, (1), 45-55.
80. Hubbard, R. E., Theory in practice: Experiences in predictive structure-based drug discovery. *MGMS International Meeting* **2005,** Biomolecular Simulation, Trinity College Dublin.
81. Broughton, H. B., A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *J Mol Graph Model* **2000,** 18, (3), 247-57, 302-4.
82. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P., FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des* **1994,** 8, (2), 153-74.
83. Anderson, A. C.; O'Neil, R. H.; Surti, T. S.; Stroud, R. M., Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem Biol* **2001,** 8, (5), 445-57.
84. Teodoro, M. L.; Phillips, G. N., Jr.; Kavraki, L. E., Understanding protein flexibility through dimensionality reduction. *J Comput Biol* **2003,** 10, (3-4), 617-34.
85. Ota, N.; Agard, D. A., Binding mode prediction for a flexible ligand in a flexible pocket using multi-conformation simulated annealing pseudo crystallographic refinement. *J Mol Biol* **2001,** 314, (3), 607-17.
86. Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L., Docking: successes and challenges. *Curr Pharm Des* **2005,** 11, (3), 323-33.

87. Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M., Molecular docking using surface complementarity. *Proteins* **1996,** 25, (1), 120-9.
88. Sobolev, V.; Moallem, T. M.; Wade, R. C.; Vriend, G.; Edelman, M., CASP2 molecular docking predictions with the LIGIN software. *Proteins* **1997,** Suppl 1, 210-4.
89. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **1982,** 161, (2), 269-88.
90. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J., Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **1997,** 272, (1), 106-20.
91. Burkhard, P.; Taylor, P.; Walkinshaw, M. D., An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 A X-ray structure of a thrombin-ligand complex. *J Mol Biol* **1998,** 277, (2), 449-66.
92. Rhodes, N.; Willett, P.; Calvet, A.; Dunbar, J. B.; Humblet, C., CLIP: similarity searching of 3D databases using clique detection. *J Chem Inf Comput Sci* **2003,** 43, (2), 443-8.
93. FRED (version 2.0.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com)*.
94. Rarey, M.; Wefing, S.; Lengauer, T., Placement of medium-sized molecular fragments into active sites of proteins. *J Comput Aided Mol Des* **1996,** 10, (1), 41-54.
95. Welch, W.; Ruppert, J.; Jain, A. N., Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* **1996,** 3, (6), 449-62.
96. Surflex, developed and distributed by Jain Lab. *http://jainlab.ucsf.edu*.
97. Jain, A. N., Ligand-based structural hypotheses for virtual screening. *J Med Chem* **2004,** 47, (4), 947-61.
98. Schnecke, V.; Kuhn, L. A., Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity. *Proc Int Conf Intell Syst Mol Biol* **1999**, 242-51.
99. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., 3rd, Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* **2003,** 17, (11), 755-63.
100. Trosset, J. Y.; Scheraga, H. A., Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. *Proc Natl Acad Sci U S A* **1998,** 95, (14), 8011-5.
101. McMartin, C.; Bohacek, R. S., QXP: powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des* **1997,** 11, (4), 333-44.
102. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **2004,** 47, (7), 1739-49.
103. Goodsell, D. S.; Morris, G. M.; Olson, A. J., Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit* **1996,** 9, (1), 1-5.
104. Yang, J. M.; Chen, C. C., GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **2004,** 55, (2), 288-304.
105. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D., Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998,** 33, (3), 367-82.
106. Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D., Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Des* **2004,** 18, (5), 333-44.
107. Bohm, H. J., A novel computational tool for automated structure-based drug design. *J Mol Recognit* **1993,** 6, (3), 131-7.
108. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M., LigScore: a novel scoring function for predicting binding affinities. *J Mol Graph Model* **2005,** 23, (5), 395-407.
109. Gehlhaar, D., PLP (Piecewise Linear Potential). *Rational drug design (ACS symposium series 719)* **1999**, 292-311.
110. Bohm, H. J., Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* **1998,** 12, (4), 309-23.
111. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **1997,** 11, (5), 425-45.

112. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **2002,** 16, (1), 11-26.
113. Head, R., M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall, VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc* **1996,** 118, 3959-69.
114. Jones, D. T.; Taylor, W. R.; Thornton, J. M., A new approach to protein fold recognition. *Nature* **1992,** 358, (6381), 86-9.
115. Mitchell JBO, L. R., Alex A, Thornton JM, BLEEP - Potential of Mean Force Describing Protein-Ligand Interactions. *Journal of Computational Chemistry* **1999,** 20, 1165-76.
116. Mitchell JBO, A. A., Snarey M, SATIS: Atom Typing from Chemical Connectivity. *J. Chem. Inf. Comp. Sci* **1999,** 39, 751-7.
117. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **1999,** 42, (5), 791-804.
118. Ishchenko, A. V.; Shakhnovich, E. I., SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. *J Med Chem* **2002,** 45, (13), 2770-80.
119. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **2000,** 295, (2), 337-56.
120. Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G., Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* **2003,** 326, (2), 607-20.
121. Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V., Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* **1999,** 42, (22), 4650-8.
122. Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M., Automated analysis of interatomic contacts in proteins. *Bioinformatics* **1999,** 15, (4), 327-32.
123. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **1999,** 42, (25), 5100-9.
124. Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S., A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities. *J Med Chem* **2001,** 44, (14), 2333-43.
125. Wang, R.; Wang, S., How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci* **2001,** 41, (5), 1422-6.
126. Paul, N.; Rognan, D., ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins* **2002,** 47, (4), 521-33.
127. Muryshev, A. E.; Tarasov, D. N.; Butygin, A. V.; Butygina, O. Y.; Aleksandrov, A. B.; Nikitin, S. M., A novel scoring function for molecular docking. *J Comput Aided Mol Des* **2003,** 17, (9), 597-605.
128. Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B.; Glen, R. C., Predicting protein-ligand binding affinities: a low scoring game? *Org Biomol Chem* **2004,** 2, (22), 3267-73.
129. Glide 3.5, developed and distributed by Schrodinger. *http://www.schrodinger.com/Products/glide.html*.
130. Pro_Leads, developed and distributed by Protherics Inc. *http://www.protherics.com/*.
131. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** 7, (20), 1047-55.
132. Bissantz, C.; Folkers, G.; Rognan, D., Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **2000,** 43, (25), 4759-67.
133. Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *JOURNAL OF MOLECULAR MODELING* **1998,** 4, (12), 379-394.
134. Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B., Consensus scoring for ligand/protein interactions. *J Mol Graph Model* **2002,** 20, (4), 281-95.
135. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **2004,** 47, (7), 1750-9.
136. Jain, A. N., Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **2003,** 46, (4), 499-511.

137. Yang, J. M.; Shen, T. W., A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **2005,** 59, (2), 205-20.
138. Stahl, M.; Rarey, M., Detailed analysis of scoring functions for virtual screening. *J Med Chem* **2001,** 44, (7), 1035-42.
139. 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.
140. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W., Deciphering common failures in molecular docking of ligand-protein complexes. *J Comput Aided Mol Des* **2000,** 14, (8), 731-51.
141. Baxter, C. A.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G.; Perkins, T. D.; Wylie, W., New approach to molecular docking and its application to virtual screening of chemical databases. *J Chem Inf Comput Sci* **2000,** 40, (2), 254-62.
142. ChemBridge Corporation, 16981 Via Tazon, Suite G, San Diego, CA 92127. *http://www.chembridge.com*.
143. 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.
144. Muegge, I.; Enyedy, I. J., Virtual screening for kinase targets. *Curr Med Chem* **2004,** 11, (6), 693-707.
145. Sheridan, R. P.; Shpungin, J., Calculating similarities between biological activities in the MDL Drug Data Report database. *J Chem Inf Comput Sci* **2004,** 44, (2), 727-40.
146. Muegge, I.; Heald, S. L.; Brittelli, D., Simple selection criteria for drug-like chemical matter. *J Med Chem* **2001,** 44, (12), 1841-6.
147. Muegge, I., Selection criteria for drug-like compounds. *Med Res Rev* **2003,** 23, (3), 302-21.
148. Available Chemicals Directory is available from MDL Information Systems Inc., San Leandro, CA, 94577. *http://www.mdli.com*.
149. Proudfoot, J. R., The evolution of synthetic oral drug properties. *Bioorg Med Chem Lett* **2005,** 15, (4), 1087-90.
150. Proudfoot, J. R., Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg Med Chem Lett* **2002,** 12, (12), 1647-50.
151. Fichert, T.; Yazdanian, M.; Proudfoot, J. R., A structure-permeability study of small drug-like molecules. *Bioorg Med Chem Lett* **2003,** 13, (4), 719-22.
152. FRED (version 2.1.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com)*.
153. Magnet, developed by Metaphorics in collaboration with Chiron. *www.metaphorics.com/products/magnet/*.
154. Hindle, S. A.; Rarey, M.; Buning, C.; Lengaue, T., Flexible docking under pharmacophore type constraints. *J Comput Aided Mol Des* **2002,** 16, (2), 129-49.
155. Brown, R. D.; Martin, Y. C., An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ Res* **1998,** 8, (1-2), 23-39.
156. Bocker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G., A hierarchical clustering approach for large compound libraries. *J Chem Inf Model* **2005,** 45, (4), 807-15.
157. Flower, D. R., DISSIM: a program for the analysis of chemical diversity. *J Mol Graph Model* **1998,** 16, (4-6), 239-53, 264.
158. Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M., Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J Chem Inf Comput Sci* **1996,** 36, (4), 750-63.
159. Wang, J.; Ramnarayan, K., Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J Comb Chem* **1999,** 1, (6), 524-33.
160. Wright, T.; Gillet, V. J.; Green, D. V.; Pickett, S. D., Optimizing the size and configuration of combinatorial libraries. *J Chem Inf Comput Sci* **2003,** 43, (2), 381-90.
161. Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V., Designing focused libraries using MoSELECT. *J Mol Graph Model* **2002,** 20, (6), 491-8.
162. Bravi, G.; Green, D. V.; Hann, M. M.; Leach, A. R., PLUMS: a program for the rapid optimization of focused libraries. *J Chem Inf Comput Sci* **2000,** 40, (6), 1441-8.

163. Jamois, E. A.; Lin, C. T.; Waldman, M., Design of focused and restrained subsets from extremely large virtual libraries. *J Mol Graph Model* **2003,** 22, (2), 141-9.
164. Murray, C. W.; Clark, D. E.; Auton, T. R.; Firth, M. A.; Li, J.; Sykes, R. A.; Waszkowycz, B.; Westhead, D. R.; Young, S. C., PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J Comput Aided Mol Des* **1997,** 11, (2), 193-207.
165. Sun, Y.; Ewing, T. J.; Skillman, A. G.; Kuntz, I. D., CombiDOCK: structure-based combinatorial docking and library design. *J Comput Aided Mol Des* **1998,** 12, (6), 597-604.
166. Liebeschuetz, J. W.; Jones, S. D.; Morgan, P. J.; Murray, C. W.; Rimmer, A. D.; Roscoe, J. M.; Waszkowycz, B.; Welsh, P. M.; Wylie, W. A.; Young, S. C.; Martin, H.; Mahler, J.; Brady, L.; Wilkinson, K., PRO_SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective factor Xa inhibitors. *J Med Chem* **2002,** 45, (6), 1221-32.
167. Aronov, A. M.; Bemis, G. W., A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins* **2004,** 57, (1), 36-50.
168. Schuller, A.; Schneider, G.; Byvatov, E., SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation. *QSAR* **2003,** 22, 719-721.
169. Weininger, D., SMILES: A Chemical Language and Information System. *J. Chem. Inf. Comput* **1988,** 28, 31-36.
170. Krier, M.; Araujo-Junior, J. X.; Schmitt, M.; Duranton, J.; Justiano-Basaran, H.; Lugnier, C.; Bourguignon, J. J.; Rognan, D., Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *J Med Chem* **2005,** 48, (11), 3816-22.
171. Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R., Rational discovery of novel nuclear hormone receptor antagonists. *Proc Natl Acad Sci U S A* **2000,** 97, (3), 1008-13.
172. Abagyan, R.; Totrov, M., Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **1994,** 235, (3), 983-1002.
173. Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S. R.; Abagyan, R.; Samuels, H. H., Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc Natl Acad Sci U S A* **2003,** 100, (12), 7354-9.
174. The DockCrunch Project. *http://www.protherics.com/crunch*.
175. Murray, C. W.; Baxter, C. A.; Frenkel, A. D., The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* **1999,** 13, (6), 547-62.
176. Waszkowycz, B.; Perkins, T. D.; Sykes, R. A.; Li, J., Large-scale virtual screening for discovering leads in the post-genomic era. *IBM Systems Journal* **2001,** 40, (2), 360-376.
177. Shao, D.; Berrodin, T. J.; Manas, E.; Hauze, D.; Powers, R.; Bapat, A.; Gonder, D.; Winneker, R. C.; Frail, D. E., Identification of novel estrogen receptor alpha antagonists. *Journal of Steroid Biochemistry & Molecular Biology* **2004,** 88, 351-360.
178. Stahura, F. L.; Bajorath, J., Virtual screening methods that complement HTS. *Comb Chem High Throughput Screen* **2004,** 7, (4), 259-69.
179. Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **2002,** 1, (11), 882-94.
180. De Miguel, F.; Lee, S. O.; Onate, S. A.; Gao, A. C., Stat3 enhances transactivation of steroid hormone receptors. *Nucl Recept* **2003,** 1, (1), 3.
181. McCarthy, D. J.; Thomas, B. E.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C., Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins* **2003,** 51, 172-188.
182. Zhao, L.; Brinton, R. D., Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases. *J Med Chem* **2005,** 48, (10), 3463-6.
183. Affinity, developed and distributed by Accelrys, 9685 North Scanton Road, San Diego, CA 92121, USA. *(URL:http://www.accelrys.com)*.

Chapter 2

Considerations in Compound Database Preparation *

Comprising

* Considerations in compound database preparation – 'hidden' impact on virtual screening results; *J. Chem. Inf. Model.* 2005 Nov-Dec; 45(6): 1908-19.

Andrew J. S. Knox, Mary J. Meegan, Giorgio Carta, David G. Lloyd.

2.1 Abstract

Structure-based virtual screening (SBVS) utilizing docking algorithms has become an essential tool in the drug discovery process, and significant progress has been made in successfully applying the technique to a wide range of receptor targets. *In silico* validation of virtual screening protocols before application to a receptor target using a corporate or commercially available compound collection is key to establishing a successful process. Prior to docking, it is important to introduce the notion of tailoring a molecular library towards compounds that display similar characteristics to known actives or ligands of a particular target. To this end, we show using active ligands of the ER, the importance of filtering towards the type of chemical space that needs to be navigated in order to retrieve compounds similar to known actives.

Ultimately, retrieval of a set of active compounds from a database of inactives is required and the metric of Enrichment *(E)* is habitually used to discern the quality of separation of the two. Numerous reports have addressed the performance of docking algorithms with regard to quality of binding mode prediction and the issue of post-processing 'hitlists' of docked ligands. However, the impact of ligand database pre-processing has yet to be examined in the context of virtual screening and prioritization of compounds for biological evaluation. We provide an insight into the implications of cheminformatic pre-processing of a validation database of compounds where multiple protonated, tautomeric, stereochemical and conformational states have been enumerated. Several commonly used methods for the generation of ligand conformations and conformational ensembles are examined, paired with an exhaustive rigid-body algorithm for the docking of different 'multimeric' compound representations to the ligand binding site of the human estrogen receptor alpha. Chemgauss, a shapegaussian scoring function with intrinsic chemical knowledge, was combined with Piecewise Linear Potential (PLP) as a consensus-scoring scheme to rank output from the docking protocol, and enrichment rates calculated for each screen. The overheads of CPU consumption and effect on relative database size (disk requirement) for each of the protocols employed are considered. Assessment of these parameters indicates that SBVS enrichments are highly

dependent on the initial cheminformatic treatment(s) used in database construction. The interplay of SMILES representations, stereochemical information, protonation state enumeration and ligand conformation ensembles are critical in achieving optimum enrichment rates in such screening.

2.2 Introduction

A typical drug discovery research programme involves the testing of every available compound in a corporate or commercial compound library using high throughput biological screening techniques (HTS). Such an approach inevitably leads to high cost and large timescales [1]. One of the major problems emerging in the pharmaceutical industry is that biologically assaying many thousands of compounds for receptor affinity is unlikely to yield the potential number of drug-like molecules that realistically exist in a random set, because the screening set is not enriched with compounds that have previously been categorised according to specific 'drug-like' parameters.

Evidence for this becomes apparent when we see that R&D production of new drugs has remained constant over the last number of years with major pharmaceutical companies each launching roughly one new drug per year [2]. A recent report also suggested that HTS has generated no lead compounds when used as a sole technique by a large number of major R&D companies [3].

An increase in the number of highly resolved X-ray crystal structures of pharmaceutically relevant biological targets [4], has prompted the use and development of computational techniques to predict ligand affinity to such macromolecules [5, 6]. Structure-Based Virtual Screening (SBVS), the most prominent computational technique employed, involves use of a docking program to generate ligand poses in the active site of a receptor and identification of the optimally docked pose using a scoring function [7]. The scoring function should reflect the complementarity or binding affinity of the ligand for the receptor [8, 9]. In the context of a virtual screen, this method should discriminate compounds that bind to the receptor from non-binders and where possible, subsequently rank the binders according to their potency.

Virtual screening of compound libraries against therapeutic targets for a particular disease state has become integrated in the drug discovery process, and provides a low-cost, rapid and effective method of enriching a random compound library with the possibility of identifying active species directly. This technique has been applied successfully to compound libraries against a number of targets using various docking programs [10-12]. It is still at an early stage in its development and improvement of sampling methods and scoring functions will undoubtedly advance both the reliability and efficacy of the technique. The significance of these factors will manifest themselves in the form of increased differentiation between active and inactive compounds in a corporate compound collection.

Of utmost importance is the driving concept of the quality of the database to be screened. It is often the choice of people in the drug discovery sector to search for molecules that are not only synthetically feasible, but that also exhibit 'drug-like' characteristics. Thus many computational pre-filters have been introduced to select only those molecules from a dataset that are 'drug-like', or more likely to have favourable ADMET properties [13]. Lipinski developed a set of rules (rule of five) to describe drug-like space through analysis of the Derwent Drug Index and found ~90% of orally available drugs were shown to have <5 hydrogen bond donors, <10 hydrogen bond acceptors, mw<500, and LogP<5 [14]. Many other filters such as functional group filters [15], fingerprint screening [16, 17], QSAR/CoMFA analysis and screening [18], compound clustering and partitioning [19] have been subsequently developed. These filters are generally based on a structural similarity theme where a known compound is used as a template to generate a property fingerprint to represent the active.

Several other studies aimed at describing 'drug-like' space have been carried out such as one by Veber [20] at GlaxoSmithKline utilising 1100 drug candidates showed that rather than implementing Lipinski's 'rule of 5' the only criteria necessary to filter for in a database is that compounds have 10 or fewer rotatable bonds and the polar surface area is approximately 140 Å (or max 12 H-bond donors or acceptors). They also rated flexibility as having a negative influence on permeability. Wenlock et al also carried out a study to compare the distributions of physiochemical properties of oral drugs in clinical development with those available on the market [21]. The trend emerging was that

molecular weight and lipophilicity had the most bearing on a drug making it through the stages of clinical development.

It is important to note at this stage that these filters are often rigorously applied to all molecule databases to 'weed' out non drug-like molecules. However, the filters should not be applied without prior analysis of the target to be screened against [22]. It has been previously noted that cancer related chemical space is intrinsically different from general drug-like space [23]. We have previously examined, looking at various classes of small-molecule oncology therapeutics, the regions of medicinal chemical space occupied by each using PCA analysis [22]. The extent to which these compounds move away from medicinal chemistry space is directly related to properties that, most importantly, can be accounted for in commonly applied computational filters applied prior to the drug discovery process. For example, if signficantly lower tox-effects (e.g. moving away from poisons such as alkylating agents) and oral bioavailability are desired (e.g., look at aromatase inhibitors, antiestrogens & antiandrogens), filters can be incorporated in a protocol to converge on molecules that occupy the same portion of chemical space.

Equally important is the concept of ligand database pre-processing prior to SBVS. In this study we seek to examine the effects of pre-processing on the prioritization of known active ligands from a database containing both known actives and inactives, where protonation, tautomeric, stereochemical and conformational states are represented. Research in our group is focused on the identification of novel modulators of human nuclear hormone receptors [24-28]. Subsequently, we have applied a virtual screening protocol to a validation target of therapeutic importance, Estrogen Receptor (ER) alpha, where several different pre-processing techniques were used to generate the database of ligands to be screened.

The estrogen receptor (ER) alpha is a nuclear hormone receptor [29] with a buried lipophilic binding site where liganding is highly dependent on hydrogen bonds as well as lipophilic contacts. The binding site is enclosed upon ligand binding and liganding results in reorganization of the receptor. In particular helix-12 encapsulates the receptor if an agonist (estradiol) is bound but is prevented from attaining this orientation when an antagonist (4-hydroxytamoxifen) is bound [30] as in Figure 1. The main differences between the antagonist 4-hydroxytamoxifen and the endogenous agonist ligand

(estradiol) are that the antagonist lacks a second hydroxyl group, which prevents hydrogen bonding with His524, but has an extended side chain to accommodate additional interaction with Asp351. The large amount of crystallographic data available and our understanding of mechanism of action make the ER a viable and therapeutically important target for virtual screening. More specifically, estrogens are mitogenic for ER positive breast cancer cells and as 50% of primary breast cancers contain ER alpha [31] we deemed this to be an important target for application to the optimization of virtual screening approaches.

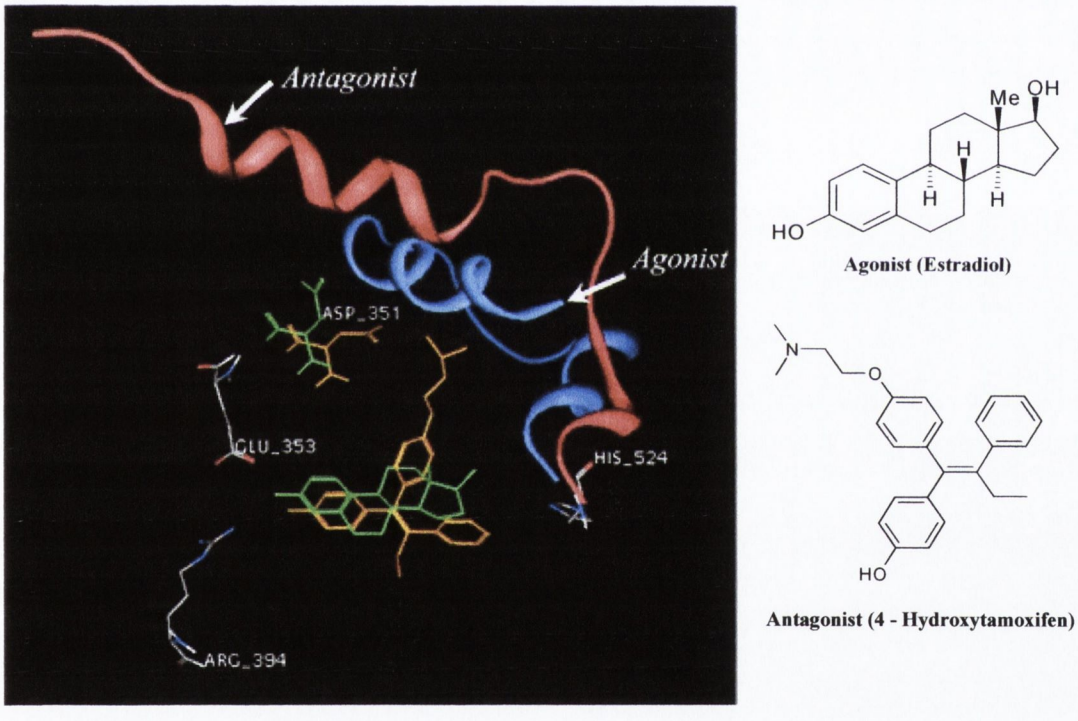


Fig **(1)** Binding of agonist (green) and antagonist (orange) induces different Helix-12 conformations.

This study is split into three main stages where stage one involved the assessment of the effect on the prioritization of the actives from the compound collection where protonation, tautomeric, stereochemical and conformational states are enumerated. To quantify these effects we report enrichment rates (E) for each level of pre-processing where a compound database consisting of 1000 compounds (Haystack) seeded with 40

ER known actives (Needles) was utilized as input. Figure 2 depicts an indicative portion of several of the ER alpha active ligands that were included in the needle set.

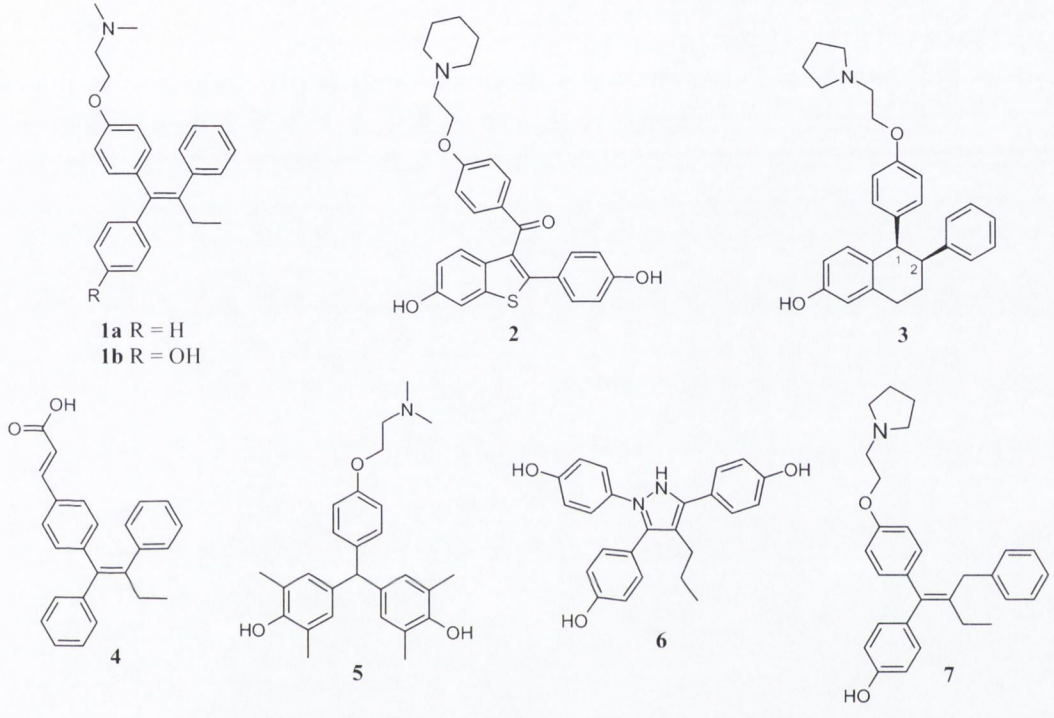


Fig **(2)** (1a) Tamoxifen (1b) 4-Hydroxytamoxifen (2) Raloxifene (3) Lasofoxifene (4) GW5638 (5) Sumimoto Biphenol (6) Pyrazole Antagonist (7) Flexible Antagonist

The second stage consisted of a more rigorous test using 10,000 compounds (claimed inactive, with disclosed biological data) seeded with a single known active, in this case a previously disclosed potent flexible antagonist [32] (Figure 2, Compound 7) illustrated in Figure 2. Using the same pre-processing protocols, the impact on ranking of the known active was assessed. Application of these processes using a training database allows one to effectively calibrate the docking and scoring protocol prior to deployment on a large corporate or commercial compound library. We also determined the quality of the binding poses produced under each protocol compared with the crystal structure 3ERT [33].

Finally, the third stage presents an analysis of the same 10,000 compounds seeded with the set of 40 actives used in stage 1. The impact of pre-processing here was assessed through calculation of False Positive (FP) rates for 50% of true positives.

Why is it necessary to pre-process a database prior to virtual screening? Upon expansion of a database of 2D structures to 3D structures, accounting for hydrogen bond donor and acceptor capabilities is imperative, as changes in the positions of hydrogens in 2D format denotes conformational changes in 3D format [34]. Accurate representation of the correct tautomeric and protonated forms of a compound, depending on the ultimate physiological environment of a compound, is also extremely important in this case. It is computationally expensive and difficult to assign the most probable state of a compound, thus all physiological relevant states are typically enumerated as a representation. Similarly, it is necessary to generate all physiologically relevant stereoisomers of a compound arising from ambiguous stereo center descriptions.

Conformational changes must also be accounted for as ligands rarely bind to a receptor in their lowest energy form and usually experience some strain upon binding that induces an increase in the energy of the ligand [35]. Computational methodologies that account for ligand flexibility can be divided into several types. Firstly, generation of conformational ensembles prior to a docking experiment, or 'on the fly' [36] can account for conformational changes upon ligand binding (FRED). Secondly, an evolutionary algorithmic approach is taken where core fragments of molecules are 'grown' in the binding site of a receptor (GOLD) [37]. Thirdly incremental build-up (DOCK and FlexX) [37, 38] or molecular dynamics simulated annealing (CDOCKER) [39] can be used to account for ligand flexibility within a docking algorithm. It has been previously noted that in order to reproduce the binding mode of a ligand in a receptor it is beneficial for the docking algorithm to provide a number of docked poses on the basis of rmsd [40]. For this reason we have chosen to use ensemble generation in this study to assist in the process. Conformer ensemble generation is also shown to be imperative in achieving optimum enrichment for the ER.

For the docking algorithm this is CPU costly, as multiple forms of the same ligand are docked, however this overhead is usually only a fraction of the time needed to predict the individual most dominant and relevant state of a molecule using software

modules currently available. To further reduce docking CPU time we chose to use a docking tool with extremely rapid docking times [41]. FRED (Fast Rigid Exhaustive Docking) [42] from OpenEye Scientific was used in combination with a consensus scoring scheme, Chemgauss and PLP, to prioritize the compounds in our docking studies in the ER. The use of FRED integrated with several scoring functions has been previously reviewed in screening experiments for seven targets with different active site characteristics, e.g. Lipophilic buried cavities, intermediate polarity, and very polar solvent exposed binding sites. Chemscore was found to be the most applicable general scoring function for SBVS [43]. We take this scoring a step further by implementing a consensus scoring scheme to rank the database post-docking, where a shape based function with terms accounting for chemical potentials, most significantly hydrogen bonding interactions, is combined with PLP, an empirical scoring function shown to correlate well with protein-ligand binding affinities. Preliminary docking and scoring studies in our laboratory have highlighted the efficacy of this method combined with FRED over other scoring methods when the ER is used as a target. This study also highlights the unanticipated impact on enrichment rates. Very different *in silico* enrichments can be achieved depending on the initial SMILES string representation used.

2.3 Computational Methods

See Appendix A for summary of commands.

2.3.1 Preparation of Estrogen Receptor (ER) Alpha

Protein target coordinates were extracted from the PDB entry corresponding to the crystal structure of the estrogen receptor complexed with 4-hydroxytamoxifen (3ERT) [33]. Structural waters were removed from the monomer and Macromodel 6.5 [44] was used to subsequently re-establish the correct connections in the PDB file. MOE [45] was used to add hydrogens to the protein and a minimization protocol using MMFF94 force-field was implemented to adjust the positions of hydrogens and keep the heavy atoms fixed at their respective crystallographic positions. The complexed ligand was extracted and used to define a 5Å search box for docking in the receptor. Protein and ligand files were saved as mol2 format using MOE.

2.3.2 Preparation of Validation Set

Verdonk et al recently described 'virtual enrichments' achieved using decoy sets that are dissimilar from the active set [46]. Here we implement the same strategy of a 'focused' decoy set with similar properties to those of the active set using a validation set (haystack) of 1,000 compounds. The haystack was built as follows:

A subset of the Derwent World Drug Index (WDI) [47] was selected using Lipinski's rules [14], by removing compounds with intrinsically non-drug like properties such as those with molecular weight <200 or >550, number of hydrogen bond donors $0<x<6$ and acceptors $0<x<10$, calculated logP <7, using an MCL script implemented in the Daylight toolkit [48]. Additional compound filtering was carried out with FILTER [49]. To remove reactive species, known toxics, carcinogens etc, an 'in-house' Perl script was used to select a random subset of compounds from this filtered dataset. Over half of all known marketed drugs contain chiral centers and it was deemed of importance to represent this in the dataset. To this end, 500 molecules with their respective specific active chiral and isomeric data were taken from the World Drug Index using the Daylight toolkit.

Subsequently 460 molecules whose active chirality and isomers were unspecified or 'ambiguous' were also selected and added to the dataset. This prevented any sources of imbalanced results where the decoy compounds are not representative of active species. The dataset was retained in SMILES format.

A set comprising 40 active ligands (needles) for the ER alpha was selected from literature where binding and/or anti-proliferative data were experimentally determined, with activities ranging from nanomolar to low micromolar potency. This active ligand set was then added to the validation set to make the total number of compounds up to 1000.

2.3.3 Preparation of Stages 2 and 3 Decoy Sets (10 000)

To more comprehensively test the protocols a decoy database of 9,999 inactive compounds was built from the WDI and CHEMBANK [50] and seeded with a single potent flexible estrogen alpha antagonist [51] (shown in Figure 2) to bring the total database size to 10,000 ligands for stage 2. In generation of this dataset all known estrogen actives were excluded, and as with preparation of the 1000 ligand dataset, all compounds with non drug-like properties were removed using MCL scripting in Daylight and application of FILTER. As before, a Perl script was used to randomly select the final 9,999 compounds from a larger filtered population before seeding with the active ligand. The dataset was stored in SMILES format.

Finally, a database of 10,000 compounds combining the decoy set from stage 2 and the 40 active ligands (needles) for the ER alpha used in stage 1 was prepared for stage 3. Similarly, the procedure was as above for stage 2. Stage 3 was carried out to make sure that the results obtained from stage 2 were not biased in any way due to the incorporation of only a single active antiestrogen.

2.3.4 Preprocessing of Validation Set

Generation of SMILES:

The Daylight toolkit [48] was used to export a dataset of actives in tab format, allowing retention of the correct assignment of isomeric where specified and 'ambiguous'

SMILES where stereochemistry was not defined in the compound records. All structures were stored as SMILES format using two methods. MOL2SMI (Daylight Toolkit) and CONVERT (Molecular Networks GmbH) [52] were used to produce two alternate SMILES representations, A and B respectively.

Generation of Tautomeric and Protonated states:

The utility TAUTOMER (Molecular Networks GmBH) [53] was used to generate relevant tautomeric states of each molecule in the database. Conversion of SMILES strings to SDF format using UNITY [54] was necessary as strings lose their stereochemical information through canonicalisation. To preserve the effects of input SMILES formatting differences, tautomerically processed SDF files were reconverted to SMILES strings using either MOL2SMI or CONVERT as required. This procedure was repeated using TAUTOMER (Openeye Scientific Software) [55] to facilitate a direct comparison of two commercially available and widely used tautomer generators.

The computational utility QUACPAC [56] was used to enumerate physiologically relevant protonation states of the validation set. Again, to preserve stereochemical information SDF files were used as input and reconverted to SMILES strings using either MOL2SMI or CONVERT as required.

Options to limit enumeration to a specified maximum number of protonated and tautomeric states are possible using both QUACPAC and TAUTOMER, however for the purposes of this study all calculable protonation and tautomeric states were enumerated in the pH range 2-14.

Generation of Stereoisomers:

Different conformations exist for enantiomers, and it is necessary to manifest this molecular conformational space in a virtual screen as compound libraries often have inadequate stereochemical information denoted. STERGEN [57] identified ligand stereocenters and generated a set of isomeric structures where none was explicitly specified. This step enumerates only the multiple possible stereocenters for 'ambiguous' SMILES strings in the validation set as the actives and approximately 50% of the selected haystack strings had explicit (correct) stereochemistry defined. As before, MOL2SMI and

CONVERT were used to produce the alternate sets of SMILES strings following stereoisomer generation. This procedure was repeated using FLIPPER [58] as an alternative stereochemistry tool to facilitate a direct comparison of two widely used stereochemical generators.

Conformer generation:

To account for the fact that a molecule can adopt several 3-D conformations by rotation about single and acyclic bonds, four 2-D to 3-D conformer generators were considered in this study, CORINA [59], OMEGA [60], RUBICON [61] and CATALYST [62].

In all cases with the exception of CORINA, a single conformer database and a multiple conformer database (10 conformers) was generated. In the case of CORINA, generation of only one true conformer (when one discounts ring-flipping variants) is possible and thus, OMEGA was subsequently used to expand the data to 10 conformers from the original input conformer passed by CORINA.

CORINA uses monocentric fragments with standard bond lengths, angles and dihedral angles to form a 3-D representation of a molecule. Sadowski et al have shown that CORINA reproduced the correct conformation of bound ligands for almost half of a dataset of 639 X-ray structures [63]. OMEGA uses a torsion-driving beam rule-based method to generate conformational ensembles. A SMILES string is reduced to fragments with rotatable bonds and rules are then applied to regenerate the ensembles. Application of the MMFF force field to refine input geometries allows any high energy constructs to be minimized. RUBICON uses distance-geometry methods to randomly sample conformations. A rule-based method for establishing geometric constraints based on SMARTS [64] is utilized. CATALYST employs two methods of conformer sampling using a poling algorithm, FAST and BEST. CPU time is a contributing factor to the choice of pre-processing protocol in SBVS so for the purpose of this study the FAST option was chosen.

2.3.5 Structure-Based Virtual Screening Protocol (SBVS)

FRED2.01 [42] was utilized in this study to dock all pre-processed compound sets. FRED2.01 uses a systematic, non-stochastic algorithm to ensure reproducible results are attained. FRED rigidly and exhaustively examines all poses in an active site and filters by shape complementarity then ranks by 'fitness' prior to scoring using gaussian functions which have chemical awareness incorporated (eg. Chemgauss, Shapegauss). The final poses can be scored simultaneously utilizing a number of scoring functions such as, Shapegauss, PLP, Chemgauss, Chemscore, Screenscore, Zapbind.

For this study, default operational values were applied and the docking of separately generated input conformers was enabled. Following rigid-body optimization of the ligands in the docking, ranking of the ligand poses using several scoring functions is possible. In internal validation studies using rigid-body docking algorithms and scoring with several scoring functions - either separately or as a consensus [65] scoring function, we have found Chemgauss and PLP [66] used as a consensus score to be the most efficacious scoring method for ranking the docked poses of a lipophilic binding site such as that of ER alpha [28]. The Chemgauss scoring function accounts most significantly for hydrogen bond interactions. PLP scoring accounts for both simple and steric hydrogen bond interactions. A recent report reviewed a set of screening experiments for seven targets with different active site characteristics, eg. lipophilic buried cavities, intermediate polarity, and very polar solvent exposed binding sites. Chemscore emerged as the most applicable general scoring function for SBVS [41]. In this work we have chosen to utilize the most recent code release - FRED2.01 - where shape docking with a chemical knowledge function (Chemgauss) in combination with PLP deliver superior enrichment results using the same datasets in comparative trials.

2.3.6 Computational Overheads – CPU Time Consumption and Database Size

Despite steady drops in the cost of computational equipment, in parallel with increases in processing power, all computational SBVS experiments have associated overheads in terms of the time required for processing and the resultant physical database size

produced. We therefore examine the time involved in producing the various 'multimeric' databases examined, and their relative sizes.

2.3.7 Stage 1. Impact of Preprocessing Levels on Enrichment Rate (1000 Compounds)

The two different representations of SMILES string generated according to section 1 of 'Pre-processing of Validation Set' above were used as the input for all subsequent enumeration of protonation, stereochemical, tautomeric states shown as A (MOL2SMI) or B (CONVERT). In presentation of the data, a qualifier 'X' denotes which of four 2D-3D toolkits (CORINA, OMEGA, RUBICON and CATALYST) was used for the validation set. Table 1 outlines the various pre-processing levels considered in this study. Each level was repeated for each of the SMILES generated – ie LEVEL1-8_X_A/B is run for both MOL2SMI and CONVERT SMILES string representations, to furnish 64 individual protocols in total, when all four conversion tools are employed.

Table 1. Classification of database pre-processing protocols applied

LEVEL	**PRE-PROCESSING PROTOCOL**
LEVEL1_X_A/B	SMILES – 1 CONFORMER
LEVEL2_X_A/B	SMILES – 10 CONFORMERS
LEVEL3_X_A/B	SMILES – PROTONATION – 1 CONFORMER
LEVEL4_X_A/B	SMILES – PROTONATION – 10 CONFORMERS
LEVEL5_X_A/B	SMILES – STEREOISOMERS – 1 CONFORMER
LEVEL6_X_A/B	SMILES – STEREOISOMERS –10 CONFORMERS
LEVEL7_X_A/B	SMILES – TAUTOMERS –1 CONFORMER
LEVEL8_X_A/B	SMILES – TAUTOMERS –10 CONFORMER

Following SBVS in each of the pre-processed databases outlined, enrichment rates for the first 0.5%, 1%, 1.5%, 2% and 4% of the screen population were calculated. Enrichment (E) indicates the ratio of the yield of actives in the hit list (post-screen ranked database population) relative to the random yield of actives as distributed throughout the unranked database and is calculated as in section 1.9.

2.3.8 Stage 2. Ranking of a Single Potent ER-alpha Antagonist in a 10 000-Decoy Set

This set of 10,000 compounds was created to more stringently test the docking and scoring procedure utilizing the optimum levels of pre-processing identified from the above protocols as determined by their respective enrichment values in the 1,000 ligand validation set. The efficacy of each protocol was determined according to the ability of each to prioritize the single active ligand contained in the dataset. In each case the protocols that achieved the highest enrichment rates for each level in experiments using the database of 1,000 structures were employed, and designated as levels 9-17. For example, if in Level 1 processing, SMILES generated using MOL2SMI and subsequent 3D conformers produced the optimum enrichment post-docking, this protocol would then be used for single conformer generation (denoted Level 9) for the set of 10,000 compounds. A level of conformer generation producing 100 conformers of each compound in the dataset (Level 11) was also added to assess the effect of increased conformer generation. The classification of each of the database pre-processing protocols is outlined in Table 2.

Table 2. Classification of 10,000 compound database pre-processing protocols applied

Level	Protocol
9	Single Conformer
10	10 Conformers
11	100 Conformers
12	Protonation + Single Conformer
13	Protonation + 10 Conformers
14	Stereoisomers + Single Conformer
15	Stereoisomers + 10 Conformers
16	Tautomers + Single Conformer
17	Tautomers + 10 Conformers

2.3.9 Stage 3. Ranking of a Diverse Set of ER-Alpha Antagonists in a 10,000-Decoy Compound Set

To ensure that the results obtained from stage 2 reveal the full potential for variation of E rates, a decoy dataset of 9960 compounds 'spiked' with the diverse set of 40 estrogen antagonists was utilized. This would account for any discrepancies that may be observed through diversity, as some of the antagonists may not intrinsically have different protonation, tautomeric, and stereochemical states. The classification of each of the database pre-processing protocols is outlined as per table 3. The efficacy of each protocol was measured by assessing False Positive (FP) rates for 50% of the true positives.

Table 3. Classification of 10,000 compound database pre-processing protocols applied

Level	Protocol
18	Single Conformer
19	10 Conformers
20	100 Conformers
21	Protonation + Single Conformer
22	Protonation + 10 Conformers
23	Stereoisomers + Single Conformer
24	Stereoisomers + 10 Conformers
25	Tautomers + Single Conformer
26	Tautomers + 10 Conformers

2.4 Results and Discussion

2.4.1 Computational Overheads

The dependence of 2D to 3D conversion rates on the nature of input SMILES string passed to the conversion tools is illustrated in Figures 3(A) and (B). An overview of CPU time used and conversion rate achieved using each conformer generation program where a single conformer is constructed from the SMILES strings in the validation set is provided. All programs were run on 32-bit Linux (Fedora) architecture with Intel(R) Xeon(TM) CPU 3.00GHz, 2Gb RAM.

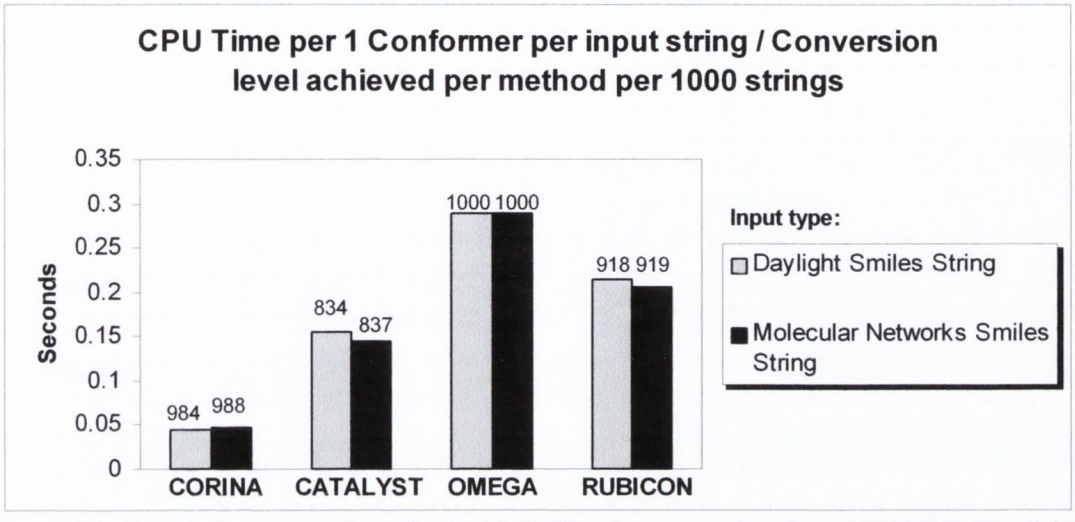


Fig **(3)(A)** Graphical representation of 2-D SMILES string conversion from MOL2SMI (Daylight) SMILES string and CONVERT (Molecular Networks GmbH) SMILES string to 3-D molecules. Data labels over each column are equivalent to the number of molecules converted in the validation set. (A): Using four conformer generation methods, one conformer was generated per molecule.

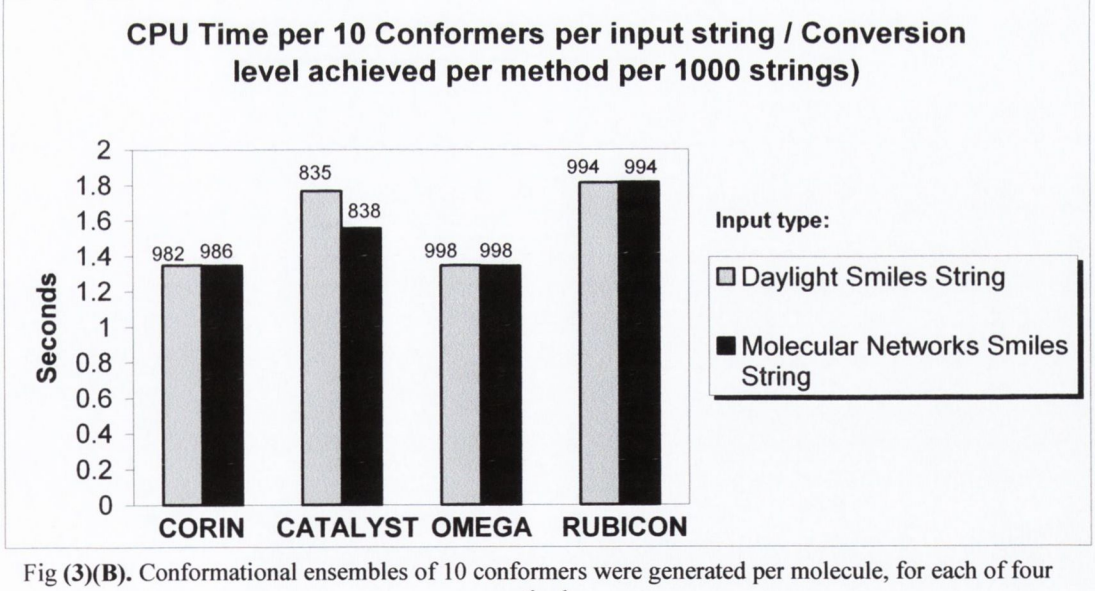


Fig **(3)(B).** Conformational ensembles of 10 conformers were generated per molecule, for each of four methods.

On examination of the data presented in Figure 3A we find that CORINA converts the dataset with a wall clock time of < 0.05 sec/molecule, expending the least CPU time when processing MOL2SMI SMILES representations. However ~ 1.6% of the dataset remained unprocessed through conversion errors. The significance of this is apparent when the dataset is scaled to a more representative 1 million compounds - a database size employed for typical discovery screening. In such a set, up to sixteen thousand molecules could remain unprocessed, leading to potential 'hits' being overlooked through this initial attrition. Similarly, Catalyst and Rubicon exhibit string-parsing errors, which could cumulatively impact on the quality of the database produced. OMEGA converts 100% of the dataset in < 0.29 sec/molecule when input SMILES were generated using either MOL2SMI or CONVERT strings.

Good et al have previously demonstrated that with increasing sampling increases the chance of producing conformers closer to the bioactive conformer in a crystal structure [67]. Figure 3(B) shows conformational ensemble generation where 10 conformers of each molecule are produced. For CORINA and OMEGA the conversion rates remain relatively unchanged, scaling as expected. With RUBICON, increased conformer sampling increases the performance of both conversion time and rate. CATALYST is seen to produce a single conformer per molecule in ~ 0.15 seconds from a SMILES string but when the production of 10 conformers per molecule is undertaken we observe an increase in processing time to 1.75 seconds per molecule. This is in line with all of the conversion tool rates. However, the low conversion rate observed when using CATALYST could mitigate against its incorporation as a large-scale pre-processing tool for database construction in SBVS.

2.4.2 Stage 1. Effect of Preprocessing Levels on Enrichment Rate

Table 4 depicts LEVEL1-8 of pre-processing using MOL2SMI and CONVERT SMILES string representations of the validation set of compounds. These data illustrate clearly the impact on the enrichment rate when protonated, tautomeric, stereochemical and multiple

conformations of the validation set are docked and scored, and also the clear significance of using alternate SMILES strings as input for such SBVS protocols.

Table 4 Enrichment results obtained for each LEVEL1-8 of pre-processing [1].

	Subset Size%					
LEVEL	**0.5**	**1**	**1.5**	**2**	**4**	**Database Size**
LEVEL1_CATALYST_A	20.85	20.85	19.46	16.68	9.38	834
LEVEL1_CATALYST_B	20.93	20.93	19.46	16.68	9.94	837
LEVEL1_CORINA_A	19.68	22.14	21.32	19.68	17.22	984
LEVEL1_CORINA_B	19.76	22.23	21.40	19.76	16.67	988
LEVEL1_OMEGA_A	**25.00**	**25.00**	**25.00**	**21.25**	**15.63**	**1000**
LEVEL1_OMEGA_B	25.00	22.50	21.66	18.75	11.88	1000
LEVEL1_RUBICON_A	18.36	18.36	19.89	17.21	11.48	918
LEVEL1_RUBICON_B	13.79	18.38	15.31	16.08	13.21	919
LEVEL2_CATALYST_A	20.88	20.88	19.48	18.78	14.09	835
LEVEL2_CATALYST_B	20.95	20.95	19.55	17.80	14.14	838
LEVEL2_CORINA_A	24.55	24.55	24.55	24.55	17.19	982
LEVEL2_CORINA_B	24.65	22.09	23.00	23.42	16.64	986
LEVEL2_OMEGA_A	**24.95**	**24.95**	**24.95**	**24.95**	**19.34**	**998**
LEVEL2_OMEGA_B	24.95	24.95	24.95	23.70	17.47	998
LEVEL2_RUBICON_A	24.85	22.37	21.54	21.12	16.77	994
LEVEL2_RUBICON_B	24.85	22.37	21.54	21.13	16.77	994
LEVEL3_CATALYST_A	20.65	18.59	15.14	14.46	9.80	826
LEVEL3_CATALYST_B	19.66	22.12	18.02	14.75	11.06	983
LEVEL3_CORINA_A	24.60	22.14	22.96	22.14	18.45	984
LEVEL3_CORINA_B	24.48	22.03	22.84	20.80	17.74	979
LEVEL3_OMEGA_A	**25.00**	**25.00**	**25.00**	**22.50**	**18.13**	**1000**
LEVEL3_OMEGA_B	25.00	25.00	23.33	23.75	14.38	1000
LEVEL3_RUBICON_A	19.40	19.40	21.02	20.61	15.76	970
LEVEL3_RUBICON_B	23.85	23.85	22.26	20.27	16.10	954
LEVEL4_CATALYST_A	20.65	20.65	19.27	17.55	13.42	826
LEVEL4_CATALYST_B	24.33	24.33	22.70	19.46	16.42	973
LEVEL4_CORINA_A	24.53	24.53	24.53	24.53	14.10	981
LEVEL4_CORINA_B	24.40	24.40	24.40	24.40	16.47	976
LEVEL4_OMEGA_A	24.93	24.93	24.93	23.68	15.58	997
LEVEL4_OMEGA_B	**24.93**	**24.93**	**24.93**	**23.68**	**16.20**	**997**
LEVEL4_RUBICON_A	24.85	22.37	23.19	22.37	16.77	994
LEVEL4_RUBICON_B	24.85	24.85	24.85	22.37	17.40	994

LEVEL5_CATALYST_A	20.65	20.65	19.27	17.55	11.36	826
LEVEL5_CATALYST_B	24.33	21.89	21.08	18.24	13.99	973
LEVEL5_CORINA_A	24.45	22.00	21.19	20.78	16.50	978
LEVEL5_CORINA_B	24.43	24.43	21.17	20.76	15.27	977
LEVEL5_OMEGA_A	**25.00**	**22.50**	**21.66**	**16.25**	**13.75**	**1000**
LEVEL5_OMEGA_B	20.00	22.50	23.33	22.50	15.63	1000
LEVEL5_RUBICON_A	18.44	18.44	18.44	16.14	13.25	922
LEVEL5_RUBICON_B	23.00	23.00	23.00	19.55	13.80	920
LEVEL6_CATALYST_A	20.65	20.65	20.65	19.62	12.90	826
LEVEL6_CATALYST_B	24.33	24.33	24.33	23.10	13.98	973
LEVEL6_CORINA_A	24.38	24.38	24.38	21.94	17.06	975
LEVEL6_CORINA_B	24.33	21.89	21.08	20.67	15.20	973
LEVEL6_OMEGA_A	24.95	24.95	23.29	23.70	17.47	998
LEVEL6_OMEGA_B	**24.95**	**24.95**	**24.95**	**23.70**	**16.84**	**998**
LEVEL6A_OMEGA_B	24.93	24.93	23.26	23.68	14.33	997
LEVEL6_RUBICON_A	24.60	22.14	22.96	20.91	15.38	984
LEVEL6_RUBICON_B	24.58	24.58	24.58	20.88	15.36	983
LEVEL7_CATALYST_A	22.08	22.08	19.13	16.56	10.49	883
LEVEL7_CATALYST_B	19.78	22.25	21.43	19.78	12.36	989
LEVEL7_CORINA_A	24.73	22.25	23.07	22.25	16.07	989
LEVEL7_CORINA_B	24.68	22.20	21.39	20.97	14.81	987
LEVEL7_OMEGA_A	**25.00**	**25.00**	**25.00**	**22.50**	**15.00**	**1000**
LEVEL7A_OMEGA_A	25.00	25.00	25.00	23.75	17.50	1000
LEVEL7_OMEGA_B	25.00	25.00	25.00	21.25	14.38	1000
LEVEL7_RUBICON_A	19.00	20.58	20.58	17.81	12.47	950
LEVEL7_RUBICON_B	23.40	21.06	18.72	16.38	10.53	936
LEVEL8_CATALYST_A	22.10	19.89	20.67	18.79	13.26	884
LEVEL8_CATALYST_B	24.75	24.75	24.75	21.04	14.23	990
LEVEL8_CORINA_A	24.95	24.95	24.95	23.70	15.59	998
LEVEL8_CORINA_B	24.95	24.95	24.95	22.46	14.35	998
LEVEL8_OMEGA_A	24.95	24.95	24.95	23.70	15.59	998
LEVEL8_OMEGA_B	**24.98**	**24.98**	**24.98**	**22.48**	**14.36**	**999**
LEVEL8_RUBICON_A	24.90	22.41	21.58	19.92	14.32	996
LEVEL8_RUBICON_B	24.93	24.93	23.26	21.19	13.71	997
Theoretical Optimal Value	**25.00**	**25.00**	**25.00**	**25.00**	**25.00**	**1000**

[1] Optimum enrichments of each level obtained are highlighted in bold.

LEVEL 1 & 2 depict 'entry level' preprocessing, where consideration of either one (LEVEL1) or ten (LEVEL2) conformers per ligand in the SBVS protocol is applied, in this instance using both MOL2SMI (X_A) and CONVERT (X_B) generated SMILES input. In all cases the enrichment is calculated using the number of molecules converted by each 2D-3D conversion program as input. We observe LEVEL1_OMEGA_A and LEVEL1_OMEGA_B outperforming other levels in LEVEL1. A large difference in enrichment rates can also be seen between the two alternate representations of SMILES strings. Screening a database prepared with LEVEL1_OMEGA_A for instance achieves the maximum possible enrichment until 1.5% and then reduces. LEVEL1_OMEGA_B however exhibits poorer enrichment rates in the same section of database. The same trend is equally observed using RUBICON. Both RUBICON and CATALYST failed to convert ~ 17% and ~ 8.2% of the database respectively, with lowered enrichment rates resulting. RUBICON conformer generation results in the lowest enrichment rate in the first 1% of the dataset when only a single conformer is presented in the SBVS, but increasing the sampling rate, as with OMEGA and CORINA, considerably improves the outcome. Also the conversion rates increase dramatically with RUBICON converting 99.4% of the database. LEVEL2_CORINA_A and LEVEL2_OMEGA_A perform almost equally well with enrichment rates up to 2% of 24.55 and 24.95 respectively.

LEVEL3 illustrates the impact on virtual screening arising from the introduction of a step to incorporate protonation states in the processing of the screening database. Single conformers generated by OMEGA contribute to achieving higher enrichment values and are enhanced by the addition of protonation as seen in table 4. Initial observation clearly shows LEVEL3_OMEGA_A achieves a superior enrichment rate over the 4% of the database when compared with the others. LEVEL3_RUBICON_A produces the lowest (*E*). Both CATALYST and RUBICON exhibit a large deviation in (E) between using two alternate SMILES representations as input. RUBICON generates higher (E) using CONVERT SMILES strings.

A remarkable increase is observed using RUBICON where multiple conformers with a treatment of protonation are considered. LEVEL4_RUBICON_B accomplishes an *(E)* of 24.85, just slightly lower than that of LEVEL4_OMEGA_B. CORINA performs well also with MOL2SMI input giving an (E) of 24.53 across the first 2% of the ranked

database. However the opposite is seen with CATALYST at this stage in the process. CATALYST converts 97.3% of the database from CONVERT SMILES strings, however only 82.6% is converted using MOL2SMI. Again the effect of alternating SMILES string representations is highlighted here.

In general, protonation has a considerable influence on the orientation of a docked ligand, as the conformers produced by each 2D-3D tool will vary according to the positions of the hydrogens on a molecule. OMEGA has been previously shown to achieve significantly better results when the input structure is supplied from CORINA [68]. Additional refinement and minimization of the CORINA input structure under the MMFF force field before ensemble generation using OMEGA also enhances the performance. These concepts were integrated in the current version of OMEGA used in this study and so in general the performance using OMEGA appears to be superior.

In the absence of specification of chemical structure the addition of arbitrary stereochemical information may introduce stereoisomers and regioisomers that may not actually exist in reality and impact on the enrichment rate achieved. STERGEN was used in the context of preserving SMILES strings with assigned stereochemical information, and to assign multiple stereochemical representations to those ligands with partial or ambiguous information. The value of this is immediately apparent as typical commercial compound libraries often contain incomplete or non-specific stereochemical information for a percentage of compound entries. LEVEL 5 denotes how consideration of stereochemical information assists Chemgauss and PLP in prioritizing the actives in the validation set only in the case of LEVEL5_RUBICON_B. Interestingly, this effect was more pronounced using FLIPPER (LEVEL6A_OMEGA_B), which we used to enumerate *all* possible stereoisomers for *all* ligands - without preserving the information of those with defined chirality. Accordingly, a slightly lower *(E)* is achieved as decoys were introduced to the docking and scoring procedure. There is clearly a fine balance which needs to be achieved in generating realistic stereochemical information for 'ambiguous' structures in a dataset and preserving the known structural information down through the processing stages.

The introduction of a tautomer treatment to database pre-processing demonstrated that in combination with a single conformer consideration, processed using

LEVEL7_OMEGA_A/B, a higher enrichment could be achieved than when processing multiple conformers (LEVEL8_OMEGA_A/B). However, for all other tools a benefit is observed on addition of multiple conformers to the database treatment. All possible tautomeric forms were enumerated, not solely the one considered to be most prevalent in solution at physiological pH's. The addition of tautomeric representations, for consideration in the SBVS protocol, makes the scoring function work harder but helps in more finely discriminating actives from inactives, as is demonstrated in LEVEL7_OMEGA_A/B. To facilitate a direct comparison of tools, LEVEL7A_OMEGA_A made use of an alternate processing utility, *Tautomer*, (Openeye Scientific Software). A slight enrichment increase is observed over the 4% of the ranked dataset using this tool. Tautomeric variation of a ligand also impacts on conformer orientation and thus the poses generated in a binding site during docking and prior to scoring. Although an additive effect on enrichment is observed by addition of tautomers as with protonation, it is difficult to immediately establish if the highest scoring tautomer ranked is actually representative of the most prevalent species *in vivo*.

Finally, to ensure the values in table 4 for each level were statistically significant we conducted an ANOVA two-way analysis of variance using Minitab14.20. At a confidence level of 99% a statistical difference was observed between programs within each level ($p < 0.0001$). A Friedman two-way analysis of variance was also used to decide the optimum levels of pre-processing using the sum of ranks from enrichments observed in each level. We chose LEVEL5_OMEGA_A rather than LEVEL5_OMEGA_B as being superior because enrichment in the first 0.5% was 25 compared with 20. This is more important as when searching a larger database it is preferable to search the smallest ranked hitlist possible (ie. 0.5-1%).

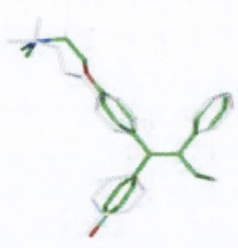
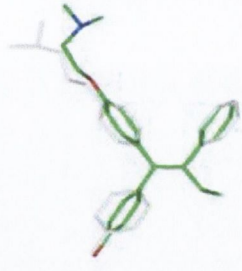
The impact of alternate SMILES representations

If alternate representations of SMILES strings are used as input 2-D structures, such as those produced by MOL2SMI and CONVERT, radically different effects on the enrichment rates are achieved when the same pre-processing protocols are utlilized in advance of rigid docking experiments. For example, LEVEL1_RUBICON_A and LEVEL1_RUBICON_B use the same pre-processing protocol of generating a single

conformer, yet enrichment rates in the first 0.5% of the database were 18.36 and 13.79 respectively. Major enrichment differences can be seen throughout table 4 arising from variation in initial SMILES depiction. This effect is not restricted to any individual 2D-3D conformer generator studied, as is evident from LEVEL5_OMEGA_A/B where enrichments are 25 and 20 respectively. These differences are possibly caused by the way in which each conformer generator parses a SMILES string. For example, programs that use a library of SMARTS strings to generate segments of a compound from a SMILES string would produce different conformers depending on the initial representation.

To emphasize the significance of this observation we have endeavored to test six alternate SMILES string representations of the known ER active modulator hydroxytamoxifen (OHT) and compare the generated conformers with those of a set of six conformers produced from a *single* SMILES string representation of the ligand. The RMSD of each conformer generated was compared with the co-crystal structure of bound hydroxytamoxifen in the ER active site and each conformer was also docked using FRED in the binding site of 3ERT to determine the best ranking conformer score. Table 5 shows the results:

Table 5. Comparison of RMSD of alternate SMILES generated conformers versus conformer generation taken from a single SMILES string.

6 SMILES permutations	RMSD	TOP SCORE PLP = -55.90
CC\C(c1ccccc1)=C(/c2ccc(O)cc2)c3ccc(OCCN(C)C)cc3	0.75	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3	0.76	
CC\C(=C(/c1ccc(cc1)O)c1ccc(cc1)OCCN(C)C)c1ccccc1	1.07	
CN(C)CCO(c1ccc(cc1)C(\c2ccc(O)cc2)=C(c3ccccc3)\CC)	1.25	
CN(C)CCO(c1ccc(cc1)C(=C(c2ccccc2)\CC)\c3ccc(O)cc3)	0.77	
CN(C)CCOc(ccc(c1)C(\c(ccc(c2)O)c2)=C(c(cccc2)c2)\CC)c1	1.23	
6 SMILES of equivalent permutation	**RMSD**	**TOP SCORE PLP = -54.91**
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3	0.6	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc4	0.76	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc5	0.99	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc6	1.25	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc7	1.13	
CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc8	0.89	

RMSD values were calculated using the OpenEye OEChem RMSD calculator, which fully accounts for automorphisms (self symmetry of the molecules being compared). Crystal structure of hydroxytamoxifen (white) superimposed with conformer of hydroxytamoxifen (coloured by atom) generated from SMILES string with the lowest RMSD value. PLP values shown correspond to the lowest RMSD structure docked in active site of 3ERT.

It is clear that different SMILES representations of the same molecule produce different conformers and as we have shown in Table 3 this has a clear impact on enrichment values obtained also. Alternating the initial SMILES string produces a range conformers, the best of which exhibits an RMSD of 0.75 Angstroms when compared to the crystal structure of hydroxytamoxifen, and results in the highest scoring docked pose (–55.90) using FRED2.01. The immediate conclusion to draw from this data is that nominally, while each individual molecule is constant, not all SMILES representations are equal, and a hitherto unexplored link exists between this simplest cheminformatic treatment of

molecular representations and the results obtained in three-dimensional molecular recognition studies involving SBVS of this nature.

Computational Overheads – how big is too big?

As previously discussed, each level of 'multimeric' ligand treatment increases the physical size of the database under consideration. This size limit begins to impact on the feasibility and practicality of the SBVS when the base number of compounds is large. Figure 4 illustrates how the size of the dataset using each protocol differs. This poses a significant quandary with respect to virtual screening. Database size must be reduced to a minimum to keep disk space at an affordable level, and also to prune the overall number of molecules to be screened from a performance cost / benefit perspective. If we consider the relative scaling for LEVEL8 – 1,000 compounds (a 312k SMILES file) expands to a 3D sdf file containing 27,212 multimeric forms (73 Mb). Scaling this up to a database of 1,000,000 screening compounds will accordingly generate 25 to 30 million ligand representations for consideration in both the pre-processing steps and in the actual docking procedure. While the ability of newer software to utilize and read compressed data files may alleviate this difficulty, the issues of how large a dataset one can afford to use, and correspondingly which pre-processing method one will employ will ultimately be determined by available resources.

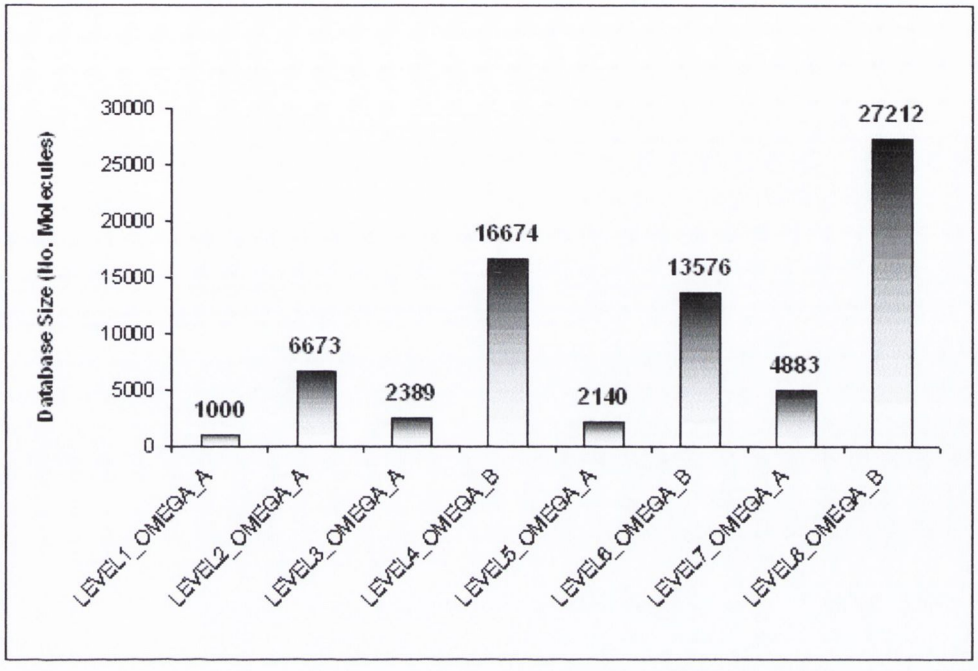


Fig **(4)** Column data labels show number of molecules produced from 1000 input SMILES strings using the optimum pre-processing techniques as previously evaluated.

2.4.3 Stage 2. Ranking of a Single Potent ER-alpha Antagonist in a 10 000-Molecule Compound Set

Tautomerism and protonation states have little synthetic consequence, although their consideration may be crucial in enhancing the enrichment level of a virtual screening protocol. A successful virtual screening protocol should not only prioritize compounds for biological testing but also provide useful information about the stereochemistry of the compound, should synthesis be required. We therefore sought to utilize from the previous steps, protocols that provide the highest possible enrichment but also give the most information about the exact 3D or isomeric nature of the compounds identified. The importance of this with respect to the estrogen receptor is clear when one considers that the *E*-isomer of the ligand tamoxifen exhibits estrogenic activity, while the *Z*-isomer exhibits antiestrogenic potency. From our initial individual screening runs, it appears that addition of stereochemical information with ensemble conformer generation provides

optimum benefit pre- and post-screening where the compound collection lacks defined stereochemical information.

To assess the utility of the various screening levels in a 'real world' application, our dataset of 10,000 ligands containing only a single known active was used. The impact of the various protocols (Levels 9-17) on the ability of SBVS to identify the active is given in Table 6.

Table 6. Ranking of a single active by FRED2.01 from a screening database totalling 10,000 drug-like compounds using pre-processing protocol Levels 9-17.

Level	Rank	Protocol Details
9	146	Single Conformer
10	**1**	10 Conformers
11	**1**	100 Conformers
12	399	Protonation + Single Conformer
13	**6**	Protonation + 10 Conformers
14	254	Stereoisomers + Single Conformer
15	**6**	Stereoisomers + 10 Conformers
16	163	Tautomers + Single Conformer
17	**3**	Tautomers + 10 Conformers

This comparative study highlights the wide impact on enrichment and variance possible when moving from single to multiple conformer treatments (Level 9→10), and where protonation (Level 9→12, Level 10→13, Level 12→13), chirality (Level 9→14, Level 10→15, Level 14→15) and tautomer treatments (Level 9→16, Level 10→17, Level 16→17) are explored in SBVS.

Optimal ranking is achieved using only 10 conformers, with no additional benefit seen when expanded to a treatment of 100 conformers per ligand. In these cases all levels involved generation of conformers using OMEGA. Ligand conformer sampling is highly important in SBVS when the active or binding site is deemed to be flexible. The estrogen receptor exhibits a relatively rigid binding site upon examination. Nonetheless, slight variations in binding site residue positions occur and a greater treatment of ligand flexibility is expected to improve the enrichment rate. Bostrom et al [68] deemed 1000 conformations per molecules to be adequate sampling in the context of a virtual screen, while we agree that such a treatment is highly important when the target is flexible, no apparent benefit is seen for the estrogen receptor above 10 conformers. This emphasizes the requirement for SBVS protocol optimization on a target-by-target basis.

As a final comparison of the impact of pre-processing, we sought to elucidate the RMSD between docked solutions for hydroxytamoxifen (coloured by atom) generated conformers generated using each of the above pre-processing protocols and the ligand co-crystal structure pose (white) taken from 3ERT. All of the docked structures are shown to be close to the crystal structure of hydroxytamoxifen (Figure 5). The docked structure of the single potent antiestrogen used to seed the database of 10,000 structures is also overlayed (coloured by atom) to show its equivalent docked solutions for each processing level. Interestingly, although a minimal difference is observed in the RMSD values, a large difference is seen in the relative rankings of the seed antiestrogen.

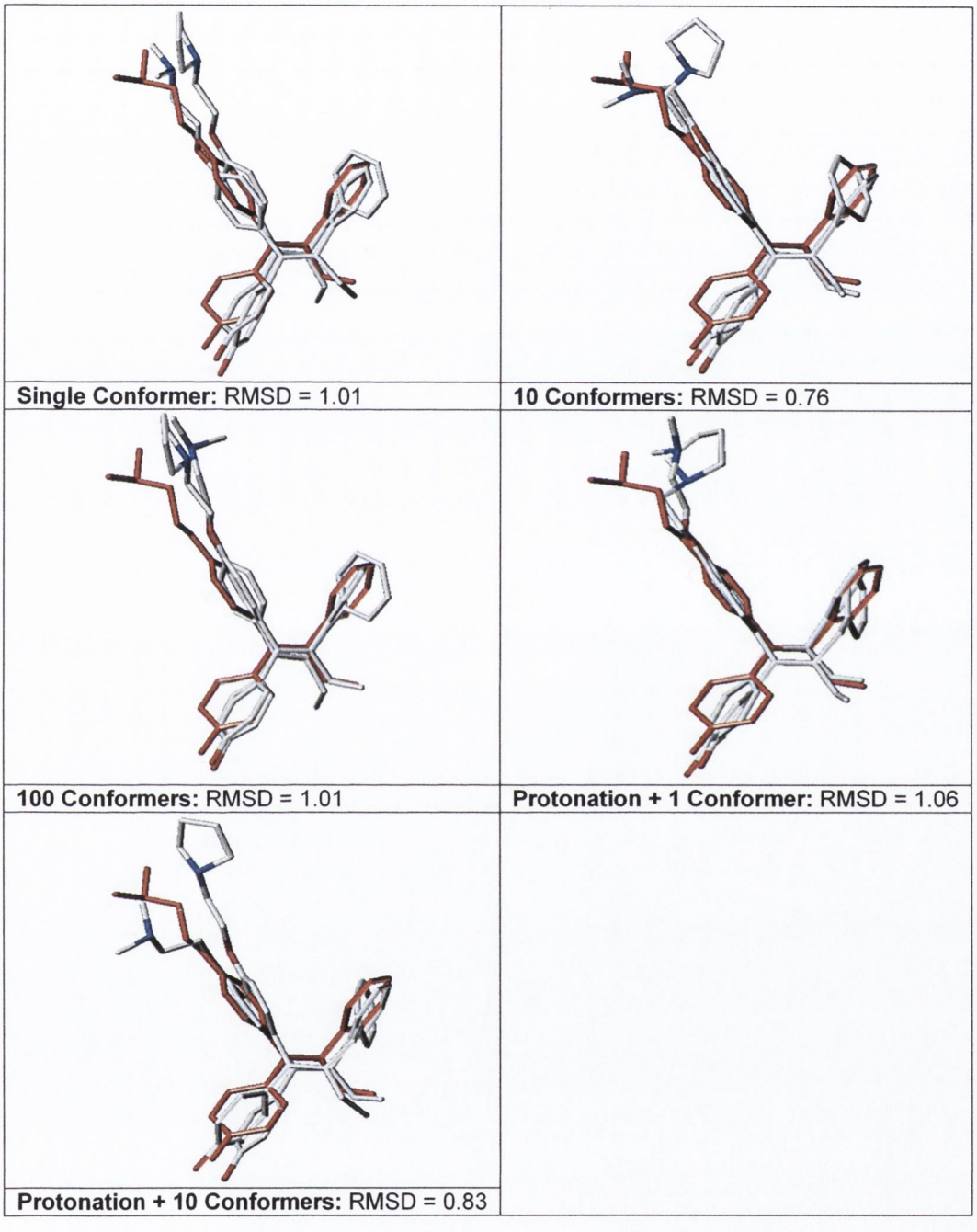


Fig **(5)** RMSD difference between docked conformers generated using each of the above pre-processing protocols and the ligand crystal structure taken from 3ERT [33]

2.4.4 Stage 3. FP rate for 40 ER-alpha Antagonists in a 10 000-Molecule Compound Set

False Positive rates for the ER-alpha antagonist set are outlined below in Table 7. Levels 18-26 use the same pre-processing procedures as those utilized with Levels 9-17, however a broader range of antagonists, namely forty, are used to rigorously test each procedure.

It is evident that the use of 10 conformers again gives the lowest FP rate at 0.35% for a true positive rate of 50% corroborating results from the previous section. The use of a more broad and diverse set of antagonists (40) spiked within a large decoy set of 9960 enables us to definitively show the influence of each level of pre-processing on the database. We observe that the database conversion rate is also optimal using multiple conformers only. Addition of multiple conformers with protonation or stereochemical generation results in a higher FP rate and thus an increase in the number of random molecules among the actives. Addition of tautomers also increases the FP rate.

Table 7 False Positive rates for recovery of 50% of true positives

LEVEL	FP 50%	Protocol Details
LEVEL18	0.44	Single Conformer
LEVEL19	0.35	10 Conformers
LEVEL20	1.10	100 Conformers
LEVEL21	0.58	Protonation + Single Conformer
LEVEL22	2.30	Protonation + 10 Conformers
LEVEL23	0.65	Stereoisomers + Single Conformer
LEVEL24	1.44	Stereoisomers + 10 Conformers
LEVEL25	1.87	Tautomers + Single Conformer
LEVEL26	1.50	Tautomers + 10 Conformers

2.5 Conclusions

We have endeavored to elucidate the optimal pre-processing protocols and determine a method generic for the optimisation of enrichment in SBVS. Establishing a successful virtual screening process also requires that the techniques used must be CPU and resource 'friendly', while keeping database physical size to a minimum.

This study provided a functional comparison of some of the most popular available techniques used in database pre-processing. All of the programs used were utilized with their default settings and adjustment of certain parameters may further enhance the enrichment rates achieved. The results presented here have important implications for those embarking on structure-based virtual screening experimentation. With increasing commercial and academic code available to researchers, the choice of pre-processing technique used to expand and represent a screening compound collection can and will have significant impact on the performance of the virtual screen. Compound libraries are often represented in 2-D format as SMILES strings and are converted to 3-D format for the purpose of pharmacophore searching or structure-based virtual screening. SMILES strings can be constructed in a number of ways and we have demonstrated clearly in this study that different representations have markedly different effects, not only on virtual screening enrichment rates achieved, but also the quality of the docked structures. If speed is a concern, then different tools are available with associated performance benefits – but not all methods will convert all input. A clear pattern is observed when using Daylight's MOL2SMI SMILES strings, where optimum enrichment rates are achieved over those found using CONVERT string representations when compounds are enumerated as single conformations. Protonation, tautomerisation and assignment of correct stereochemistry appears to have little benefit in this test case, but with respect to single conformers an optimal enrichment can be achieved starting from a MOL2SMI string. SMILES strings generated from CONVERT require enumeration of multiple conformations to produce a high enrichment regardless of whether they are protonated, tautomerised or stereochemistry is assigned.

Importantly, enrichment rates observed when using a smaller dataset of 1000 compounds seeded with 40 actives show a marked difference to the ranking of a single

active in 10,000. The best enrichment is achieved using OMEGA in combination with propagation of 10 conformers per compound. No additional benefit is observed when using 100 conformers per compound with this particular receptor (ER), but this may not be the case when working with more flexible target systems. To allow a sufficient amount of synthetic information to be retrieved about each compound, a slight decrease in the ranked position of the active must be accepted where stereochemical information is added. However a balance between the introductions of false positive results and exact structural information needs to be achieved – and a cost-benefit assessment made for each target studied. A marked difference in the ranking ability of SBVS imbued by alternate pre-processing protocols is observed when using the larger test set. We therefore suggest the adoption of larger ('real scale') validation and training datasets as more beneficial to those involved training a docking procedure for the identification of active species in virtual screening. We are currently applying these findings to other virtual screening protocols optimized for the identification of novel modulators of the human estrogen receptors.

2.6 References

1. DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G., The price of innovation: new estimates of drug development costs. *J Health Econ* **2003,** 22, (2), 151-85.
2. Smith, A., Screening for drug discovery: The leading question. *Nature* **2002,** 418, 453-459.
3. Lahana, R., How many leads from HTS? *Drug Discov Today* **1999,** 4, (10), 447-448.
4. Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J., The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol* **2000,** 7 Suppl, 957-9.
5. Jorgensen, W. L., The many roles of computation in drug discovery. *Science* **2004,** 303, (5665), 1813-8.
6. Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004,** 432, (7019), 862-5.
7. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** 7, (20), 1047-55.
8. Bissantz, C.; Folkers, G.; Rognan, D., Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **2000,** 43, (25), 4759-67.
9. Stahl, M.; Rarey, M., Detailed analysis of scoring functions for virtual screening. *J Med Chem* **2001,** 44, (7), 1035-42.
10. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **2002,** 45, (11), 2213-21.
11. Schapira, M.; Abagyan, R.; Totrov, M., Nuclear hormone receptor targeted virtual screening. *J Med Chem* **2003,** 46, (14), 3045-59.
12. Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P., Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J Med Chem* **2000,** 43, (3), 401-8.
13. Wishart, D. S., Bioinformatics in drug development and assessment. *Drug Metab Rev* **2005,** 37, (2), 279-310.
14. 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.
15. Rishton, G. M., Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov Today* **2003,** 8, (2), 86-96.
16. Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **2002,** 1, (11), 882-94.
17. 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.
18. Muegge, I.; Heald, S. L.; Brittelli, D., Simple selection criteria for drug-like chemical matter. *J Med Chem* **2001,** 44, (12), 1841-6.
19. Ekins, S.; Berbaum, J.; Harrison, R. K., Generation and validation of rapid computational filters for cyp2d6 and cyp3a4. *Drug Metab Dispos* **2003,** 31, (9), 1077-80.
20. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002,** 45, (12), 2615-23.
21. Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D., A comparison of physiochemical property profiles of development and marketed oral drugs. *J Med Chem* **2003,** 46, (7), 1250-6.
22. Lloyd, D. G.; Golfis, G.; Knox, A. J. S.; Oprea, T. I.; Meegan, M. J., Oncology Exploration: Charting Cancer Medicinal Chemistry Space. *Drug Discov Today* **2006,** 11, (3/4), 149-159.
23. Wang, J.; Ramnarayan, K., Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J Comb Chem* **1999,** 1, (6), 524-33.
24. Lloyd, D. G.; Hughes, R. B.; Zisterer, D. M.; Williams, D. C.; Fattorusso, C.; Catalanotti, B.; Campiani, G.; Meegan, M. J., Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J Med Chem* **2004,** 47, (23), 5612-5.

25. Lloyd, D. G.; Buenemann, C. L.; Todorov, N. P.; Manallack, D. T.; Dean, P. M., Scaffold hopping in de novo design. Ligand generation in the absence of receptor information. *J Med Chem* **2004,** 47, (3), 493-6.
26. Meegan, M. J.; Lloyd, D. G., Advances in the science of estrogen receptor modulation. *Curr Med Chem* **2003,** 10, (3), 181-210.
27. Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M., Ethyl side-chain modifications in novel flexible antiestrogens--design, synthesis and biological efficacy in assay against the MCF-7 breast tumor cell line. *Anticancer Drug Des* **2001,** 16, (1), 57-69.
28. Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M., Flexible estrogen receptor modulators: design, synthesis, and antagonistic effects in human MCF-7 breast cancer cells. *J Med Chem* **2001,** 44, (7), 1072-84.
29. MacGregor, J. I.; Jordan, V. C., Basic guide to the mechanisms of antiestrogen action. *Pharmacol Rev* **1998,** 50, (2), 151-96.
30. Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M., Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997,** 389, (6652), 753-8.
31. Scott, J., McGuire, WL, Endocrine-Dependent Tumors. **1991**, 179-196.
32. Lloyd, D., Smith, HM, O' Sullivan, TP, Zisterer, DM, Meegan, MJ, Antiestrogenically active 2-benzyl-1,1-diarylbut-2-enes: Synthesis, Structure-Activity Relationships and Molecular Modelling Study for Flexible Estrogen Receptor Antagonists. *Med Chem* **2006,** In press.
33. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L., The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998,** 95, (7), 927-37.
34. Miller, M. A., Chemical database techniques in drug discovery. *Nat Rev Drug Discov* **2002,** 1, (3), 220-7.
35. Perola, E.; Charifson, P. S., Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **2004,** 47, (10), 2499-510.
36. Bostrom, J., Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J Comput Aided Mol Des* **2001,** 15, (12), 1137-52.
37. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** 261, (3), 470-89.
38. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **2001,** 15, (5), 411-28.
39. Wu, G.; Robertson, D. H.; Brooks, C. L., 3rd; Vieth, M., Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J Comput Chem* **2003,** 24, (13), 1549-62.
40. Kallblad, P.; Mancera, R. L.; Todorov, N. P., Assessment of multiple binding modes in ligand-protein docking. *J Med Chem* **2004,** 47, (13), 3334-7.
41. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D., Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004,** 57, (2), 225-42.
42. FRED (version 2.0.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com).*
43. 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.
44. MACROMODEL 6.5, developed and distributed by Schrodinger Inc. *(URL: http://www.schrodinger.com).*
45. Molecular Operating Environment (MOE), developed and distributed by Chemical Computing Group. *(http://www.chemcomp.com).*
46. 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.
47. Derwent World Drug Index, (URL: http://thomsonderwent.com/products/lr/wdi).
48. Daylight Chemical Informations Systems Inc, (URL: http://www.daylight.com).
49. FILTER, distributed by Openeye Scientific Software.

50. Strausberg, R. L.; Schreiber, S. L., From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **2003,** 300, (5617), 294-5.
51. Lloyd, D., Smith, H, O' Sullivan, TP, Zisterer, DM, Meegan, MJ, *Med Chem* **2005,** In press.
52. CONVERT, developed and distributed by Molecular Networks GmbH. *(URL: http://www.mol-net.de)*.
53. TAUTOMER, distributed by Molecular Networks GmbH.
54. UNITY, distributed by Tripos Inc.
55. TAUTOMER, distributed by Openeye Scientific Software.
56. QUACPAC 1.1, distributed by Openeye Scientific Software.
57. STERGEN, distributed by Molecular Networks GmbH.
58. FLIPPER, distributed by Openeye Scientific Software.
59. CORINA 3.6, distributed by Molecular Networks GmbH.
60. OMEGA 1.8.1, distributed by Openeye Scientific Software.
61. RUBICON, distributed by Daylight Chemical Informations Systems Inc.
62. CATALYST, developed and distributed by Accelrys, 9685 North Scanton Road, San Diego, CA 92121, USA. *(URL:http://www.accelrys.com)*.
63. Sadowski J, *J. Chem. Inf. Comput. Sci* **1994,** 34, 1000.
64. SMARTS, (URL: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html).
65. Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B., Consensus scoring for ligand/protein interactions. *J Mol Graph Model* **2002,** 20, (4), 281-95.
66. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W., Deciphering common failures in molecular docking of ligand-protein complexes. *J Comput Aided Mol Des* **2000,** 14, (8), 731-51.
67. 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.
68. Bostrom, J.; Greenwood, J. R.; Gottfries, J., Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **2003,** 21, (5), 449-62.

Appendix A (List of pre-processing commands used)

Generation of SMILES:

Mol2smi -output_format ISM < infile.sdf > outfile.smi (Daylight Inc.)

Convert –outfile outfile.smi –format smileswithname infile.sdf (Molecular Networks GmBH)

Generation of Tautomeric and Protonated states:

Tautomer –all –outfile outfile.sdf infile.sdf (Molecular Networks GmBH)

Tautomer –all –in infile.sdf –out outfile.sdf (Openeye Scientific Software)

Pkatyper –in infile.sdf –out outfile.sdf (Openeye Scientific Software)

Generation of Stereoisomers:

Stergen –it=sdf –ot=sdf –d preserve infile.sdf outfile.sdf

Flipper –in infile.sdf –out outfile.sdf

Conformer Generation:

Corina –it=smiles –ot=sdf infile.smi outfile.sdf

Omega –maxconfs 10 –finalopt true –verbose true –in infile.smi –out outfile.sdf –warts

cat infile.smi | Rubicon –RUBE_NCONFS 10 –RUBE_HYDROGENS ALL –RUBE_DEBUG TERSE –RUBE_OUTPUT_FORMAT PDB > outfile.pdb

catconf –sd infile.sd –outsd outfile.sd 10 –fast

Structure-Based Virtual Screening Protocol:

```
#Interface settings
-param  fred_setup.txt
#-pvmconf (Not set, no default)

#Input_Ligands :
   -dbase  LEVELX.sdf
```

```
  -scdbase true
  #-molnames (Not set, no default)

#Active_Site :
  -pro 3ert_protein.mol2
  -box 3ert_reference_lig.mol2
  -addbox 5.000000

#Constraints :
  #-pharm (Not set, no default)

#Docking :
  -no_dock false

  #Exhaustive_Search :
    -exhaustive_scoring chemgauss
    -rstep 1.500000
    -tstep 1.000000

  #Negative_Image :
    -clash_checking 0.750000
    -neg_img_size normal

  #Number_of_poses :
    -num_poses 10
    -sqrt_poses false

#Refinement :
  #-num_refined_poses_retained (Not set, no default)
  -refine no_refinement

#Scoring :

  #MASC_Corrected_Scoring_Functions :
    -shapegauss_masc false
    -plp_masc false
    -chemgauss_masc false
    -chemscore_masc false
    -screenscore_masc false
    -zapbind_masc false

  #Standard_Scoring_Functions :
    -shapegauss false
    -plp true
    -chemgauss true
    -chemscore false
```

```
      -screenscore false
      -zapbind false

#Output :
   -prefix fred
   -oformat mol2
   -output_alt_scores false
   -output_alt_structs false
   -output_scores true
   -output_structs true

   #Hitlists :
      -serial false

      #Cutoff : Cutoff values for ligand scores
         #-shapegauss_cut (Not set, no default)
         #-shapegauss_masc_cut (Not set, no default)
         #-plp_cut (Not set, no default)
         #-plp_masc_cut (Not set, no default)
         #-chemgauss_cut (Not set, no default)
         #-chemgauss_masc_cut (Not set, no default)
         #-chemscore_cut (Not set, no default)
         #-chemscore_masc_cut (Not set, no default)
         #-screenscore_cut (Not set, no default)
         #-screenscore_masc_cut (Not set, no default)
         #-zapbind_cut (Not set, no default)
         #-zapbind_masc_cut (Not set, no default)

      #List_size : Maximum size of hitlists
         -shapegauss_masc_size 1000
         -shapegauss_size 1000
         -plp_masc_size 1000
         -plp_size 1000
         -chemgauss_masc_size 1000
         -chemgauss_size 1000
         -chemscore_masc_size 1000
         -chemscore_size 1000
         -screenscore_masc_size 1000
         -screenscore_size 1000
         -zapbind_masc_size 1000
         -zapbind_size 1000
```

Chapter 3

Development of a screening platform for Scaffold Hopping & Hit Identification.*

Comprising

* Development of a screening platform for Scaffold Hopping & Hit Identification; *J. Med. Chem (Ready for submission)*

Andrew J.S. Knox*, Mary J. Meegan, Vladimir Sobolev, Dermot Frost, Daniella Zisterer, David G. Lloyd

3.1 Introduction

Chapter 3 describes several studies that intertwine to give an initial validation and proof of concept of a novel vHTS protocol. The purpose of these studies was to initially determine the applicability of an 'in-house' tailored version of the non-commercial docking algorithm (LIGIN) to the virtual screening process followed by establishing the most efficacious scoring functions to utilise in conjunction with LIGIN that would permit identification of new 'hits'. Several key questions were addressed:

- Could LIGIN reproduce the binding mode of known co-crystallised ligands of the Estrogen Receptor (ER)?
- How does the overall screening procedure compare with well-known and commercial docking utilities?
- Could LIGIN assist in finding novel scaffolds for the ER?
- Could LIGIN be used to identify new actives (agonists/antagonists) for the ER?

To demonstrate the efficacy of the final computational protocols, and to corroborate the findings, biochemical validation of all top ranked compounds was undertaken. Figure 1 illustrates the general computational screening procedure utilised in the vHTS process.

3.2 Methodological Validation

This section details work carried out for the development of the parameters and techniques needed to carry out a successful virtual screen for ligands that bind to the ERα. We describe the use of an 'in-house' protocol employing the rigid-body docking algorithm LIGIN, and its application to the task of vHTS. Details of the initial validative processes that prompted our choice of method for docking and post-docking are provided, including conformer generation, comparison of the docking program LIGIN [1] (a description of which is provided in section 3.3) with other docking protocols, and finally an analysis of hit and enrichment rates (E) using a haystack of 1000 compounds

seeded with 35 antiestrogens extracted from literature. Firstly, two conformer generation tools were examined in their ability to reproduce the bioactive conformation of 4-hydroxytamoxifen where a single 3D structure was created. 4-Hydroxytamoxifen was built in Macromodel v6.5 [2] and several minimization steps (Steepest-Descent, Full-Matrix Newton-Raphson, Polak-Ribiere) executed to produce a low energy structure.

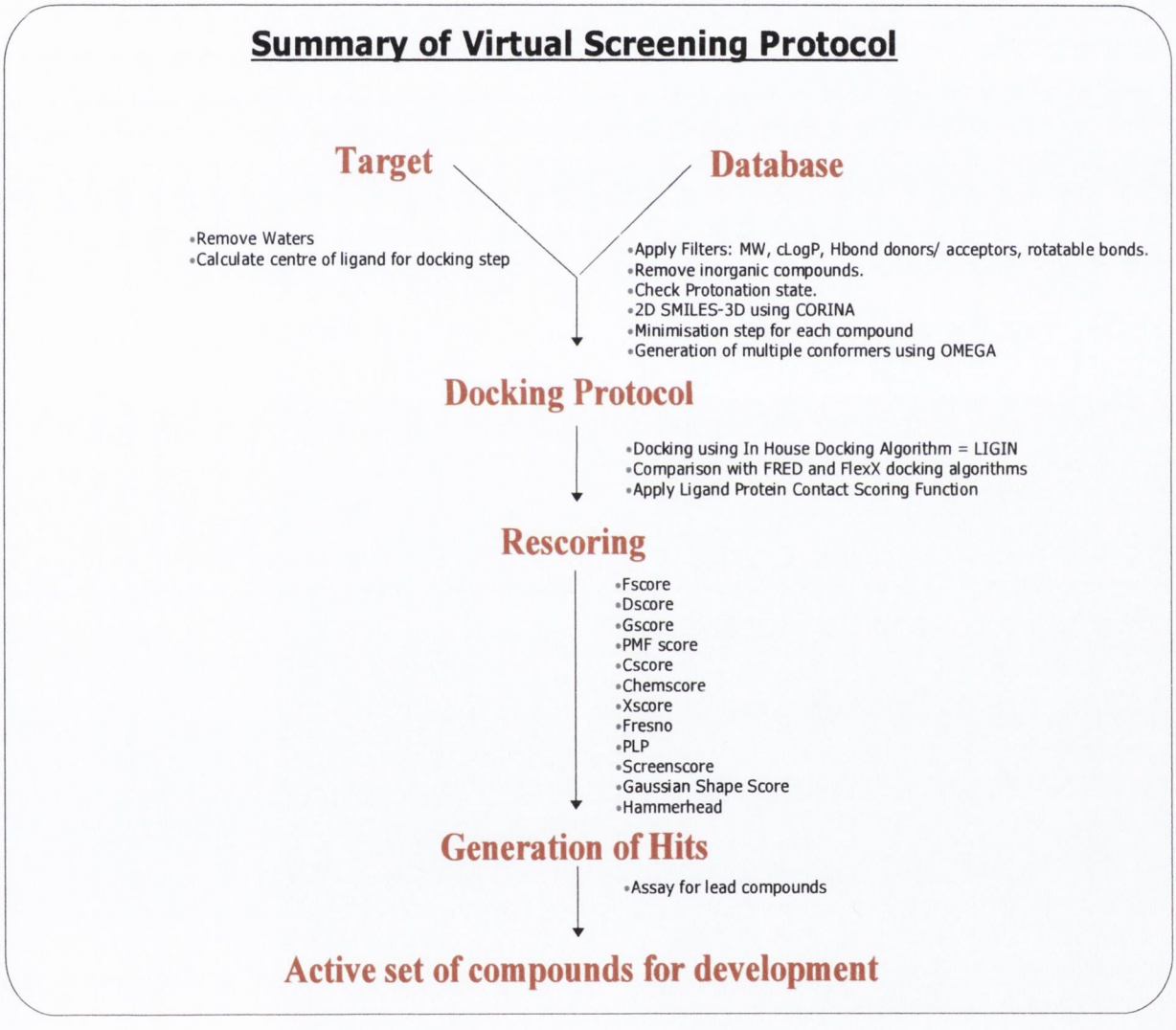


Fig **(1)** Overview of the virtual High Throughput Screening (vHTS) process

Finally a Monte Carlo minimization was carried out to ensure the global minimum was found. As a comparison, a rapid rule-based conformation generator Corina 2.64 [3], was used to generate a 3D structure from a SMILES [4] string representation of the compound. To demonstrate the advantage of using ligand flexibility to converge more accurately on the bioactive conformation, Omega v1.8 [5] was utilized to build 10 low-energy conformers from the 3D structure . All conformers were compared by RMSD to the crystal structure of 4-hydroxytamoxifen extracted from PDB entry 3ERT [6].

The utility of the docking algorithm (LIGIN) was next examined by comparing with two widely used commercial programs, Flexidock v6.9 [7] and InsightII v2000 [8], to determine the degree of receptor flexibility necessary to incorporate within future docking protocols. A training set of 8 estrogen receptor antagonists and partial antagonists, outlined in section 3.5.2, was built in Macromodel v6.5, chosen from literature where either RBA values or antiproliferative data was available. Ligand Protein Contacts (LPC) [9] output of each docked complex was generated and interactions with key residues examined. A full description of LPC is given in section 3.4 below. Thirdly, a validation of the protocol utilizing a database of 1000 ligands seeded with 35 antagonists extracted from literature was carried out.

3.3 LIGIN

LIGIN docks ligands based on maximisation of surface complementarity of atoms of the ligand with those of the receptor. Atoms are assigned a chemical type (8 classes defined) according to their chemical properties and they participate in non-bonding interactions with residues of the active site. Interactions are quantified by 'legitimate' (complimentary) and 'illegitimate' (uncomplimentary) contact assignments designated by the two atoms involved. The basic presumption is that two atoms will be in contact if they share a common surface area with a distance between them smaller than $R_a + R_b + 2R_w$, where R_a and R_b are van der Waals radii of the atoms and R_w is that of the solvent molecule. A final evaluation of the fit of a molecule in the active site is given by calculation of a complementarity function (CF):

$$CF = S_l - S_i - E$$

where S_l and S_i are the sum of all 'legitimate' (complimentary) and 'illegitimate' (uncomplimentary) contact surface areas respectively between ligand and residues of receptor. E is a repulsion term (wall term) similar to that used in energy force fields. A 'wall' term is also incorporated, similar to the repulsive term used in the Lennard-Jones potential to account for intermolecular clashes. As the CF value would be ultimately dependent on the size of the ligand, it is normalised by dividing by the solvent accessible surface of the uncomplexed ligand, producing the Normalised Complementarity (NC).

Another feature of LIGIN is that a certain degree of receptor flexibility can be taken into account through neglecting the contribution of one or more side-chains of residues lining the binding pocket in the docking process, a so-called 'Soft-Docking' approach. This permits a ligand to occupy the same place as side-chains in 3D space. For the purpose of this chapter, however, we examined docking, treating the receptor as a rigid body.

3.4 LPC

Ligand Protein Contacts (LPC) software [9] outputs a Normalised Complementarity value that describes the level to which a molecule interacts with the residues of the binding site (complementarity) and thus a measure of its solvent accessible surface. A docked molecule producing a score of ~1 is one that is 100% in the binding site and the solvent accessible surface is 0 and all contacts are legitimate. LPC outputs an NC value also, however a 'wall term' is included in the calculation to prevent atomic bumping. The software can also calculate all the receptor residues in contact with the ligand and their respective types of contact. Table 1 below illustrates the LPC output for all residues in contact with the ligand and their respective types.

Table **(1)** Residues in contact with the ligand OHT600 (4-hydroxytamoxifen) in PDB entry 3ERT.

Residue	Dist	Surf	Specific contacts HB	Arom	Phob	DC
343A MET*	3.8	26.0	-	-	+	-
346A LEU*	3.6	42.2	-	-	+	-
347A THR*	3.7	40.5	+	-	-	+
349A LEU*	4.1	13.9	-	-	+	-
350A ALA*	3.3	32.8	-	-	+	-
351A ASP*	3.2	29.0	+	-	-	+
353A GLU*	2.4	34.2	+	-	-	-
354A LEU*	6.5	1.3	-	-	-	-
383A TRP*	3.7	32.8	-	+	-	-
384A LEU*	4.0	25.1	-	-	+	-
387A LEU*	3.7	40.1	+	-	+	+
388A MET*	4.4	10.3	-	-	+	-
391A LEU*	4.1	19.7	-	-	+	-
394A ARG*	3.0	22.2	+	-	-	-
404A PHE*	3.8	21.5	-	+	+	-
419A GLU	3.9	2.0	-	-	-	-
420A GLY	3.8	15.9	-	-	-	-
421A MET*	3.5	50.3	-	-	+	-
424A ILE*	4.0	12.1	-	-	+	-
428A LEU*	3.7	17.9	-	-	+	-
521A GLY*	3.6	34.8	-	-	-	-
524A HIS*	4.0	14.4	-	-	+	-
525A LEU*	3.8	47.3	-	-	+	+
528A MET*	5.3	13.0	-	-	-	-
530A CYS*	6.1	4.9	-	-	-	-
536A LEU*	6.3	3.1	-	-	-	-
539A LEU*	6.3	2.9	-	-	-	-

Dist, nearest distance (Å) between atoms of the ligand and the residue; Surf, contact surface area (Å^2) between the ligand and the residue; HB, hydrophilic-hydrophilic contact (hydrogen bond); Arom, aromatic-aromatic contact; Phob, hydrophobic-hydrophobic contact; DC, hydrophobic-hydrophilic contact (destabilizing contact); or +/- indicates presence/absence of a specific contacts between ligand and residue.
*, indicates residues contacting ligand by their side chain (including CA atoms).

From Table 1, it is evident that Thr347, Asp351, Glu353, Leu387 and Arg394 all form H-bonds with atoms on the ligand. Utilising a Perl script that scans this section of the LPC output for every docked ligand, a set of filters that ensure only molecules that interact in a certain way are retained may be applied. (e.g. the residues Glu353/Arg394 must interact within a distance of <3Å).

3.5 Experimental Section - Computational

3.5.1 Conformation Validation

Macromodel v6.5

The hydroxytamoxifen structure was built in Macromodel v6.5. Initial energy minimization was carried out through sequential minimization steps using Steepest Descent (SD), Polak-Ribiere Conjugate Gradient (PRCG) and Full Matrix Newton Raphson (FMNR) techniques. Subsequently, a global energy minimization using a Monte Carlo conformational search technique under the PRCG method for 1000 iterations was carried out on this structure. The Macromodel force field MM3* was used in all cases, as it is an excellent force field for simple monofunctional organic molecules [10].

CORINA v2.64

CORINA v2.64 was used to convert information on atoms and bonds from SMILES format to three-dimensional atomic coordinates. CORINA uses monocentric fragments with standard bond lengths, angles and dihedral angles to form a 3-D representation of a molecule. Sadowski et al have shown that CORINA reproduced the correct conformation of almost half of a dataset of 639 X-ray structures and outperformed five automatic 3D structure generators (CONCORD, ALCOGEN, Chem-X, MOLGEO, COBRA) [3].

OMEGA v1.8

Omega v1.8 uses a torsion-driving beam rule-based method to generate conformational ensembles. A SMILES string is reduced to fragments with rotatable bonds and rules are then applied to regenerate the ensembles. Application of the MMFF force field [10] to refine input geometries allows any high energy constructs to be minimized. However at the time of this experiment no MMFF refinement was available as an option within Omega. The maximum number of conformers of each molecule generated was set to 10 in the omega.com input file. The INPUT structures were those generated from CORINA.

3.5.2 Docking Protocol Validation

An initial training set of 8 ligands was built with Corina initially and followed by 10 conformers produced by Omega. Each conformer of a ligand was docked using LIGIN. The SMILES strings for the ligands are provided below in table 2.

Table **(2)** SMILES strings for a set of 8 antiestrogens with RBA values.

SMILES representation	Compound Name	RBA	Type
C(\c1ccccc1)(CC)=C(\c1ccc(cc1)OCC[NH+](C)C)c1ccc(cc1)O	4-Hydroxytamoxifen	175.24	Rat cytosol
C(\c1ccccc1)(CC)=C(\c1ccc(cc1)OCC[NH+]1CCCC1)c1ccccc1	Idoxifene	12.00	Rat cytosol
C[NH+](C)CCOc1ccc(cc1)/C(c2ccccc2)=C(/CC)c3ccccc3	Tamoxifen	1.00	Human ER
C(\CCC)(c1ccccc1)=C(\c1ccccc1)c1ccc(cc1)OCC[NH+](C)C	Toremifene	1.38	Rat cytosol
OC(=O)/C=C/c1ccc(cc1)\C(=C(\CC)c2ccccc2)c3ccccc3	GW5638	4.30	Human ER
Oc1cccc(c1)C(\c2ccc(OCC[NH+](C)C)cc2)=C(/CC)c3ccccc3	Droloxifene	15.24	Rat cytosol
c1c2sc(c(c2ccc1O)C(c1ccc(cc1)OCC[NH+]1CCCCC1)=O)c1ccc(cc1)O	Raloxifene	25.00	Human ER
c12cc(ccc2[C@H]([C@@H](C(O1)(C)C)c1ccccc1)c1ccc(cc1)OCC[NH+]1CCCC1)OC	Levermeloxifene	1.54	Human ER

As a direct comparison, and to account for ligand and receptor flexibility, the same set of ligands were docked using Flexidock vSybyl6.9 and InsightII v2000 respectively. For all dockings the Normalised Complementarity values were calculated and compared.

LIGIN

Each ligand was superimposed over the endogenous ligand of the crystal structure of 3ERT and the endogenous ligand deleted. The receptor with new ligand initially docked is saved in PDB format. The complex is then read in and receptor and ligand are separately saved as PDB files PROT and LIG respectively. All the above steps were carried out using Macromodel 6.5. An INPUT file was generated according to the arbitrary rules set and LIGIN was run on each set of ligands accordingly. The PROT and CR1 files were merged together and these rigid docked were considered to be the best-docked structure for each ligand. LPC was executed for each docked complex and the highest Normalised Complementarity produced by each docked ligand retained per conformer set.

Flexidock

The rigid docked structures from the previous docking using LIGIN were imported to Sybyl6.9 and put through a flexidock docking routine (default parameters used in all cases except the iterations were set to 30000), allowing only the ligand to be fully flexible and the protein kept rigid. This generated new docked structures saved in PDB format again.

InsightII

The Flexidock docked structure for eight random structures of the training set was then carried through a fully flexible docking routine in InsightII and the final docked structure retained in PDB format also. Three scripts were needed to run the docking automatically: predock.log, godock.log, dock.log. Predock.log reads in the crystal structure, removes waters, extracts the endogenous ligand and fixes unresolved amino acids. Godock.log deletes the endogenous ligand, creates a docking assembly for the substrate ligand in the estrogen receptor, fixes charges for the complex, displays the active site and sets up the docking grids. Dock.log finally runs affinity [11] docking of the complex.

3.5.3 Assessment of binding mode

Finally, to ensure that LIGIN could position antiestrogens in the correct binding orientation, three ERα antagonists from crystal structures 3ERT, 1ERR, and 1UOM, representing structurally diverse ER ligands were docked. These docked structures were then compared by RMSD (Root Mean Square Deviation) to the crystal structures. The docking protocol was as above in the LIGIN section. For a valid commercial comparison Flexidock was used to dock structures produced by Corina initially as above. Default parameters were used in all cases but iterations were set to 10,000. Flexidock was chosen because of its inherent ability to incorporate ligand flexibility, which is somewhat analogous to the generation of a number of rigid conformers to incorporate flexibility.

3.5.4 Screening Validation

Preparation of Target PDB entries

Crystallographic structures were downloaded from the Protein Data Bank. The chosen targets were 3ERT (Hydroxytamoxifen), 1ERR (Raloxifene), and 1UOM (2-phenyl-1- [4-(2-piperidin-1-yl-ethoxy)-phenyl]-1,2,3,4-tetrahydroisoquinolin-6-ol). Crystallographic waters were removed. The structures were read into Macromodel 6.5 and re-saved in PDB format to re-connect the bonds correctly in this format.

Preparation of Training set

A diverse set of 1000 compounds was downloaded from the Maybridge plc HiTS Kits. In a study carried out selecting 15 commercially or freely available chemical libraries listed in Figure 2 below, Maybridge was shown to be a drug-like and diverse library of compounds [12]. Also all of the compounds in the databases follow Lipinski's rule of five and are drug-like. Physiochemical properties such as cLogP, numbers of H-bond donors/acceptors, numbers of rotatable bonds, were calculated. 2-D filters were then applied that allowed removal of inorganic compounds and also compounds that were not drug-like, as shown in a recent review by Veber et al (e.g. compounds having 10 or fewer rotatable bonds and a maximum of 12 H-bond donors or acceptors) [13].

Finally, the remaining compounds were retained in a database in SMILES format, and converted to 3-D format using CORINA. Omega was utilized to generate 10 3D conformations of each molecule using its rule-based torsion-driving algorithm. Accounting for ligand flexibility is thought to provide a more realistic assumption of the actual nature of the bioactive conformation of a bound ligand. All molecules were subsequently protonated at physiological pH (7.4). The dataset remaining contained 922 molecules. ≅ 0.1% of the database was removed after these processes (removal of inorganics).

To assess whether the screening protocol was effective, a set of 35 antagonists and estradiol were added to the database, comprising well-known antagonists, and recent

literature compounds. The full dataset was stored in an SQL database containing information on the compound ID, the SMILES string of each compound, the PDB format of each compound, a unique conformation identifier. The use of an SQL database at this step allows easy manipulation of data and ensures continuous access to stored datasets. Another benefit is that any compound databases to be screened are converted into a usable format for docking with LIGIN and other docking platforms. Docking results were extracted from an SQL database using a perl script. All docked structures with a value over 0.80 for the complementarity value and putative hydrogen bonds with at least two of the following residues: Asp351, Arg394, Thr347, Glu353, His524, Leu387 were outputted. Following docking and application of LPC to the docked poses, ≅ 300 of the 957 docked structures for each crystal structure were removed.

Company	**Library name**	**Web address**
AcbBlocks	ACB	http://www.acbblocks.com
Asinex	Asinex	http://www.asinex.com
Key Organics	Bionet	http://www.keyorganics.ltd.uk
ChemBridge	ChemBridge	http://www.chembridge.com
ChemDiv	ChemDiv	http://www.chemdiv.com/main.phtml
ChemStar	Chemstar	http://www.chemstar.ru
InterBioScreen	IBS	http://www.ibscreen.com
MayBridge	MaybBridge	http://www.maybridge.com
Molecular Diversity Preservation International	MDPI	http://www.mdpi.org
Micro Source Discovery Systems	MsDiscovery	http://www.msdiscovery.com
Nanoscale Combinatorial Synthesis	NanoSyn	http://www.nanosyn.com

NCI/NIH Developmental Therapeutics Program	NCI	http://dtp.nci.nih.gov/index.html
Timtec	Timtec	http://www.timtec.net
Tripos	Tripos	http://www.tripos.com
Institut de Chimie Organique et Analytique	ICOA	Corporate database

Fig **(2)** Origin of chemical libraries

Re-scoring of docked poses

Re-scoring of docked poses was undertaken in order to achieve greater separation of poorly docked poses from well-docked poses. The merits of each were discussed in an earlier section (Section 1.8) and all were applied using their default parameters. The scoring functions applied are listed below:

a. F-score [14]
b. D-score [15]
c. G-score [16]
d. PMF score [17]
e. Chemscore [18]
f. Xscore [19]
g. Fresno [20]
h. PLP [21]
i. Screenscore [22]
j. Hammerhead [23]

Comparison with FRED, FlexX

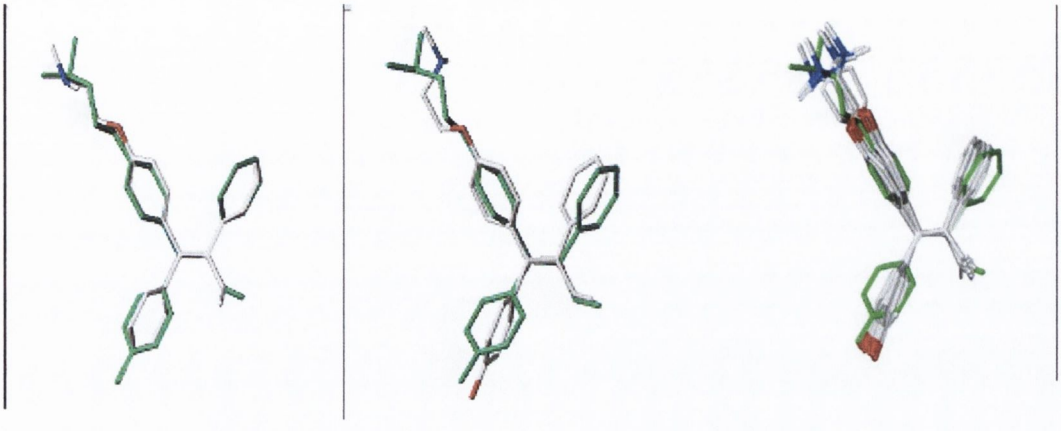
The filtered dataset with 922 ligands and 35 antagonists seeded within it was carried through a docking protocol initially with FRED and then with FlexX to provide feasible comparators for our optimized docking protocol. Again default parameter settings were used in both, and the standard scoring function of each (Gaussian Shape Scoring and FlexX respectively), were used to dock the set. Finally the enrichment rate for each was calculated and compared with our screening protocol.

3.6 Results and Discussion

Conformation Validation

Optimization of the procedure for generation of conformers to be docked using LIGIN was assessed intially. RMSD values were used to indicate which of the conformations was closest to the bioactive conformation of hydroxytamoxifen as it is found in the crystal structure of PDB entry 3ERT. This step is of the utmost importance when using a rigid docking protocol, as no ligand flexibility is incorporated within the docking algorithm usually. Figure 3 illustrates the superimposition of hydroxytamoxifen conformers generated from a number of protocols outlined in the materials and methods section. It is clear from the figure, that Macromodel6.5 produces a slightly closer conformation to the crystal structure than Corina, but however for the purposes of Virtual Screening it would not be applicable due to time constraints in the process.

Omega produced a set of low energy conformers, of which one produced an RMSD closer to the bioactive conformation than previous minimisation methods of 0.9Å. Bostrom et al show that pre-optimizing the OMEGA input structures produced from Corina, using the MMFF94 force field, dramatically improves the output conformers RMSD. This is because CORINA constructs the 3-D co-ordinates with sometimes-unfavorable bond angles causing vdW clashes, which will reduce the chances of achieving the bioactive conformation of the ligand.



Macromodel6.5
Minimisation:RMSD=1.2A

Initial energy minimization was carried out through sequential minimization steps using Steepest Descent (SD), Polak-Ribier Conjugate Gradient (PRCG) and Full Matrix Newton Raphson (FMNR) techniques.
Following, a global energy minimization using a Monte Carlo conformational search technique under the PRCG method for 1000 iterations was carried out on this structure.

CORINA:RMSD=1.22A

CORINA was used to convert atoms and bonds to three-dimensional atomic coordinates from SMILES format. A comparison of available six automatic 3D structure generators (CONCORD, ALCOGEN, Chem-X, MOLGEO, COBRA, and CORINA) using 639 X-ray structures has been published recently (3), and shows CORINA to be the best choice and for this reason was the method chosen for this study.

OMEGA:RMSD=0.90A

Omega, or Optimized Molecular Ensemble Generation Application, is a program for conformer enumeration. A maximum number of ten conformers of each molecule were generated. The INPUT structure was generated from CORINA.

Fig **(3)** Comparison of conformer generation of Hydroxytamoxifen using several techniques.

Docking Protocol Validation

It can be seen clearly from Table 3 that adding full flexibility to both the ligand and receptor, does not augment the docking process with respect to the Normalised Complementarity value. This is a measure of the 'buriedness' of a molecule and how well contacts are made, and it appears that generation of an ensemble of conformers prior to the docking process adequately accounts for the level of flexibility observed with the ER and having the added advantage of being fast.

Table **(3)** Normalised Complementarity values for set of 8 ligands docked with LIGIN, Flexidock, InsightII.

Compound Name	NC[a] LIGIN (rigid)	NC Flexidock (ligand flexible)	NC Insightll (both ligand & receptor flexibility)
4-Hydroxytamoxifen	1	0.88	0.9
Idoxifene	0.99	0.88	0.92
Tamoxifen	1	0.92	0.93
Toremifene	0.99	0.9	0.89
GW5638	0.91	0.92	0.87
Droloxifene	0.99	0.84	0.86
Raloxifene	0.85	0.76	0.71
Levermeloxifene	0.94	0.73	0.69

a = Normalised Complementarity value.

Assessment of binding mode

Figure 4 shows the docked poses arising from the three methods for 4-Hydroxytamoxifen, superimposed by heavy atoms of the protein backbone so as to retain a true pictorial representation of the exact binding mode of each. The rmsd observed between the endogenous ligand, Flexidock and LIGIN is 3.76, 3.67 respectively.

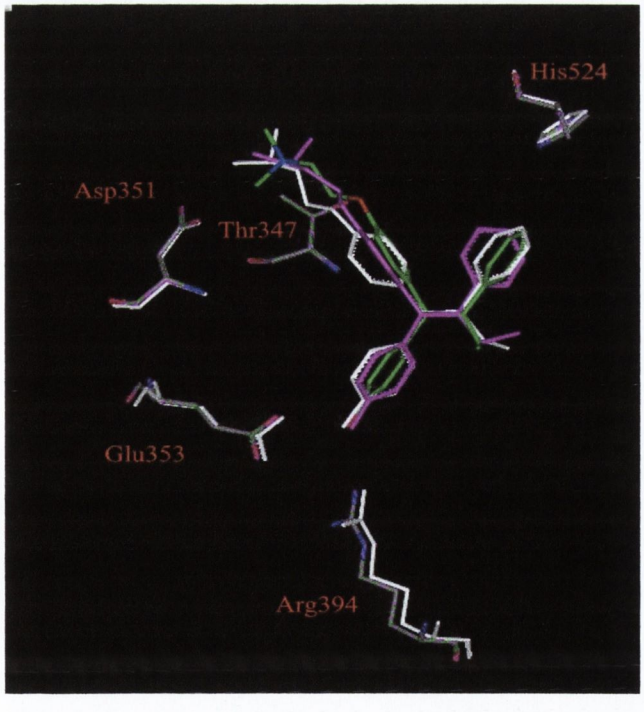


Fig **(4)** Superposition of three active sites of 3ERT. Endogenous Ligand (White), Flexidock solution (Pink), LIGIN solution (by atom).

Again using Raloxifene the rmsd between the crystal structure, Flexidock and LIGIN are 5.61Å, 2.61Å respectively. LIGIN outperforms Flexidock again in this case. Docking the Tetrahydroisoquinoline compound from 1UOM rmsds of 5.27Å and 5.34Å are observed for Flexidock and LIGIN respectively.

Screening Validation

In order to validate the screening process as a whole, the ability of a set of 10 well known scoring functions to discriminate between 35 antagonists (See Appendix A) of ERα and 957 'drug-like' inactives was measured. Moreover this validation allowed selection of an appropriate scoring function for the ER using LIGIN.

Figure 5 shows the hits retrieved using each scoring function post-docking using our protocol. It becomes immediately clear that applying Chemscore or D_Score provides us with the best-hit rates.

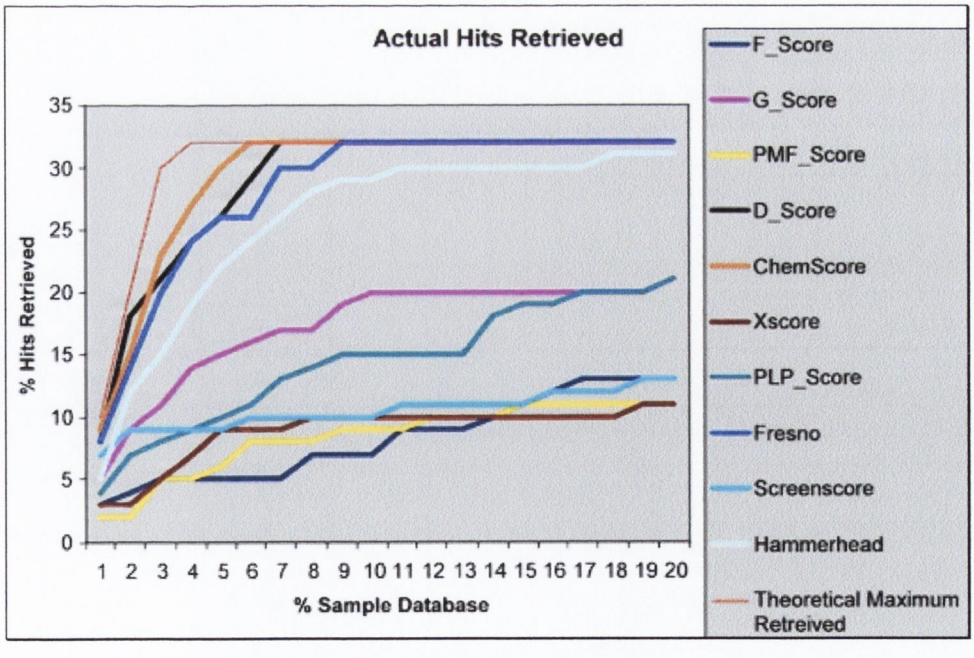


Fig **(5)** Actual hits retrieved from database using several scoring functions.

Table 4 shows the best possible enrichment rates that can be achieved by any screening protocol with a database of 957 ligands and 35 antagonists seeded within it.

Table **(4)** vHTS Performance Measures (Enrichment *).

Subset Size %	1	5	10	15	20
Ligands	10	48	96	144	192
Maximum Actives	10	35	35	35	35
Best Possible Value	27	19.95	10.08	6.7	5

* calculated using equation given in section 1.9

To provide a realistic comparison as to the level of enrichment that might be expected we docked and scored with two commercially used programs, FlexX and FRED. Table 5 shows the actual enrichment rates observed using our system versus FlexX and FRED. Divisions of 5 are highlighted in yellow and plotted on the graph below in Figure 6.

Table **(5)** Comparison of Enrichment rates using different protocols.

Subset Size %	In House Procedure	FlexX	FRED	Best Value
1	22.85	11.43	8.57	27.00
2	24.29	8.57	14.29	27.00
3	20.00	5.71	11.43	27.00
4	17.00	6.43	10.71	25.00
5	14.86	5.14	9.71	19.95
6	13.80	4.28	9.52	16.80
7	13.48	3.67	8.99	14.29
8	11.88	3.21	8.28	12.60
9	10.50	2.86	7.96	11.14
10	9.43	2.57	7.78	10.08
11	8.57	2.34	7.56	9.12
12	7.85	2.38	7.14	8.33
13	7.23	2.40	6.79	7.66
14	6.74	2.25	6.33	7.15
15	6.29	2.10	6.37	6.70
16	5.90	1.97	5.90	6.26
17	5.54	1.85	5.54	5.87
18	5.24	1.75	5.25	5.57
19	4.96	1.65	4.96	5.26
20	4.73	2.00	4.73	5.00
21	4.47	1.90	4.47	4.74
22	4.28	1.82	4.41	4.54
23	4.11	1.74	4.23	4.35
24	3.93	1.66	4.00	4.17
25	3.76	1.60	3.87	3.99

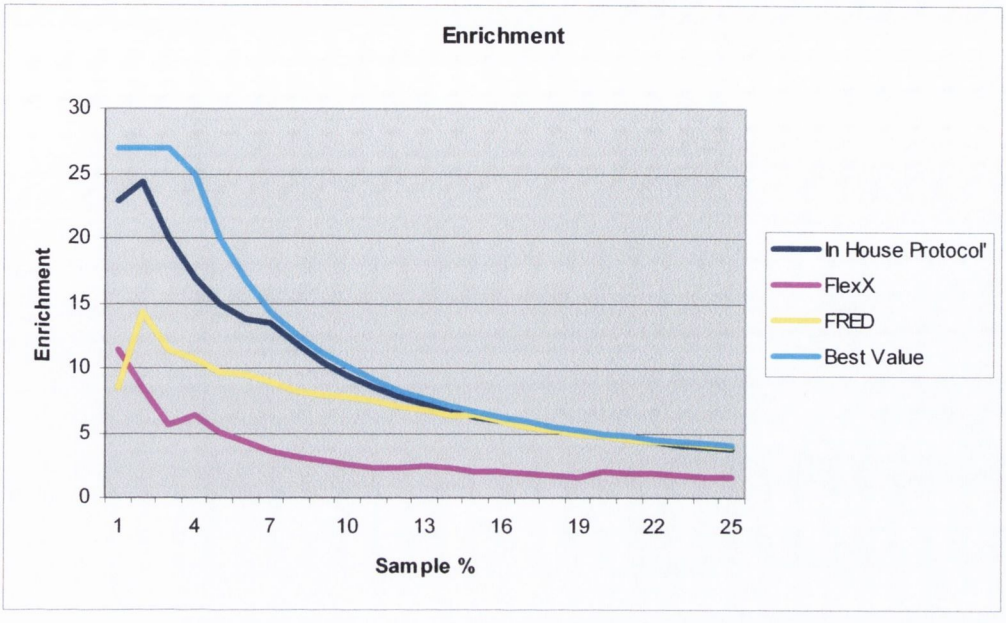


Fig **(6)** Enrichment rates using each protocol as outlined in the key.

An (E) rate of 22.85 in the first 1% of the ranked database is observed using our protocol scored with chemscore. By comparison, (E) rates of 11.43 and 8.57 are seen using FlexX and FRED respectively. Considering the optimal value could be 27 in 1% of the ranked hitlist, our protocol ranks in the top 10 compounds, 8 actives. With an (E) rate of 17 for 4% of the ranked hitlist, which allows all 35 ligands to be contained, we observe 24 active compounds within the set using our protocol. Using FlexX or FRED only 9 and 15 active ligands are ranked respectively in the top 40.

3.7 Conclusion

We have shown the applicability of LIGIN to the docking process and also the degree of flexibility needed to produce reasonably good binding predictions. Generally the production of 10 conformers prior to the docking process is sufficient to represent ligand flexibility. Adding receptor flexibility appears to be advantageous, however, this will be fully examined in Chapter 4 later. Having shown LIGIN to be efficacious as a docking utility, we assessed the ability of LIGIN to discriminate between actives and inactives in a virtual screening context. An (E) rate of 22.85 was observed for our 'in-house' protocol,

which outperformed two widely used commercial programs FlexX and FRED. To conclude, pre-processing a database of compounds beginning from SMILES format using Corina followed by Omega and docking using LIGIN allows accurate prediction of binding modes. Finally, scoring with Chemscore permits separation of actives from inactives in a dataset. It is clear that additional restrictions might augment the docking and scoring process to reduce a compound collection and reveal ERα binders. Utilising some of these methods and also combining findings from Chapter 2, we describe in the next section a further virtual screen involving a more refined version of our protocol.

3.8 Methodological Validation and vHTS

This section details the optimisation of the '*in silico*' validated protocol and the successful application against the ER of the SPECS compound database (Release: Aug2005, 202054 compounds in total), leading to identification of three novel compounds with ERα binding values (IC_{50}) of 1.4μM, 57nM and 53nM. In MTT assays, antiproliferative IC_{50} values of 15μM, 11.4μM and 7μM respectively were also measured for these compounds when evaluated in the MCF-7 breast cancer cell line.

3.9 Abstract

To extract compounds with novel chemotypes from large compound collections, we describe here the development and optimization of a fully automated Virtual High Throughput Screening (vHTS) protocol and its application to a target of therapeutic importance, Estrogen Receptor alpha (ERα). The vHTS platform encompasses a docking algorithm (LIGIN) automated by a proprietary C routine to sequentially dock conformers, followed by scoring using a two-tiered scoring scheme.

Firstly, Ligand Protein Contacts (LPC) software is used to calculate a Normalised Complementarity (NC) value in a pre-screen stage to pre-score and filter compounds unable to fit in the active site. Re-scoring is then carried out using a second scoring function that was chosen from an evaluation of 15 diverse scoring functions.

The incorporation of our post-docking filter permits two alternative paths to be taken. Path 1 when selected, gives the user the ability to 'scaffold-hop' to identify novel scaffolds. Path 2, however, allows one to directly identify 'hits' from a screen. We demonstrate the efficacy of these methods whereby a virtual screen of the SPECS database (Release: Aug2005, 202054 compounds in total), revealed both previously known and new scaffolds following Path 1. Following Path 2 for the same screen allowed identification of three novel active compounds with ERα binding values (IC_{50}) of 1.4μM, 57nM and 53nM and also antiproliferative IC_{50} values of 15μM, 11.4μM and 7μM respectively for the MCF-7 breast cancer cell line. Finally, we present an optimized structure from subsequent enumeration of a virtual library using the core substructure of one of the 'hits'. Figure 7 depicts the overall procedure.

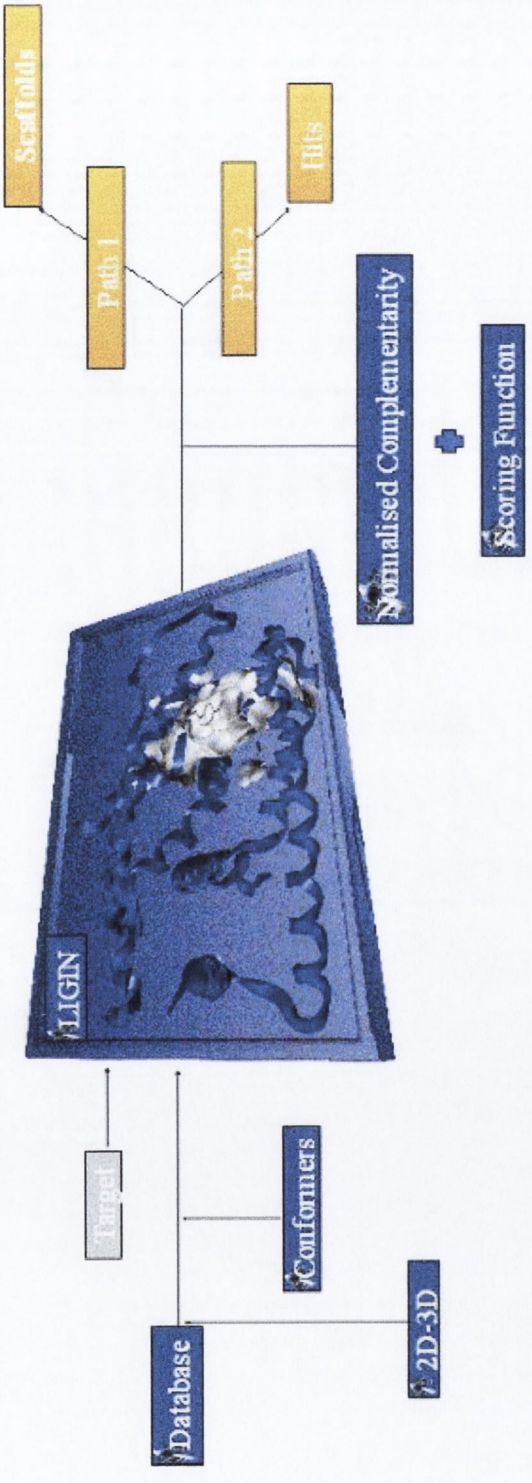


Fig (7) Overview of Screening Protocol

3.10 Introduction

A brief survey of the literature currently available (http://www.ncbi.nlm.nih.gov/) in the area of Virtual High Throughput Screening (vHTS) reveals over 250 related entries since the late 90's. This field has emerged in the last decade as a key player for both the pharmaceutical industry and academia in the discovery of new lead compounds that possess specific therapeutic properties. Acceleration of the drug discovery process can now be envisaged, with the ever-increasing harnessing of computational power, concomitant with an increase in both the number of small molecule databases available, and also the number of 3D target structures solved by X-ray diffraction, NMR or Homology modelling.

The vHTS process is typically performed by docking a molecule into a receptor active site and determining the optimal orientation by conformational, translational and rotational movement [24-26]. Subsequent scoring of these complexes is undertaken to assess the correct binding modes of the complexes, allowing ranking by affinity [27]. This ranking allows prioritisation and selection of compounds for biological testing.

Several studies have examined the ability of docking and scoring combinations to retrieve a set of known actives from databases of decoys [20, 28-31]. Evaluation of their efficacy has been determined through analyses of Enrichment (E) rates [32] or their ability to correctly reproduce binding modes observed in crystal structures as measured by RMSD [29]. Noteworthy, a common associated pitfall of these processes was addressed recently by Verdonk et al [33], who demonstrated the importance of utilising a validation set containing compounds similar to the actives rather than 'drug-like' or 'random' in order to prevent artificial enrichment. Cole et al [34] also importantly point out that RMSD calculations can be flawed because docked solutions can exhibit a low RMSD, but they can also have substituents oriented incorrectly with respect to residues of the active site. RMSD calculations do not account for different atom types that may be involved in key interactions and thus a system termed IBAC (interactions-based accuracy classification) has been proposed by Kroemer et al to overcome these problems [35].

However, enrichment calculations are very useful in the context of virtual screening and have been widely used as success criteria. For this reason, we utilise the

same metric to evaluate the success of our vHTS platform. Halgren et al [36] also point out that the common definition of Enrichment does account for the actual rank of each active in a scored hitlist, and for this reason we also calculate False Positive (FP) rates for our program in the validation process as another indicator of success.

In this section an automated vHTS platform consisting of a rigid-body docking algorithm (LIGIN) and a two-tiered scoring scheme is evaluated in its ability to retrieve both known scaffolds and also novel modulators of ERα. LIGIN has been previously tested in CASP2 experiments involving binding pocket identification [37], modelling the quinone binding site in the D1 protein of photosystem-2 reaction centre [38], and the inhibitory/stimulatory binding sites for tentoxin within chloroplast F0F1-ATPase [39]. However, in the context of virtual screening, LIGIN has not yet been examined or challenged as a tool. The LIGIN methodology is fully described elsewhere [1], and the main features of LIGIN are described in section 3.1.1.

Post-docking, we apply a filter to remove 'poorly' docked structures. In this case, Ligand Protein Contacts (LPC) software [9] was applied, to the best of our knowledge, for the first time as an integrated post-docking pre-score utility. LPC also calculates an NC value similar to LIGIN, and we calculated these values for a training dataset of 19 'drug-like' ERα actives extracted from literature, to establish the range of NC values for these antiestrogens allowing a threshold to be set. This threshold value, acted as an initial filter to remove any 'poorly' posed structures from a database of 1000 structures with the 19 actives seeded.

For the remaining compounds we evaluated the ability of 15 popular scoring functions, ie., X-Score, Drugscore, five scoring functions implemented in Sybyl6.91 (D-Score, PMF-Score, G-Score, ChemScore, and F-Score), four implemented in FRED2.11 (Chemscore, Chemgauss, Chemgauss2, Plp, Screenscore) and one from Surflex (Hammerhead), to discriminate between the actives and inactives in the set.

At this stage in the protocol two paths are available to the user. For Path 1, a set of distance constraints between specific residues (Glu353, Arg394, Asp351, His524, Leu384/387, Met343) and the nearest atom of the ligand can be applied, to allow retrieval of compounds whose scaffolds may be adequately oriented for ER binding but may not be revealed in a focussed virtual screen due to the presence of inappropriate substituents.

Figure 8 illustrates the key interactions involved in the binding process to allow contextualisation of the imposed distance constraints.

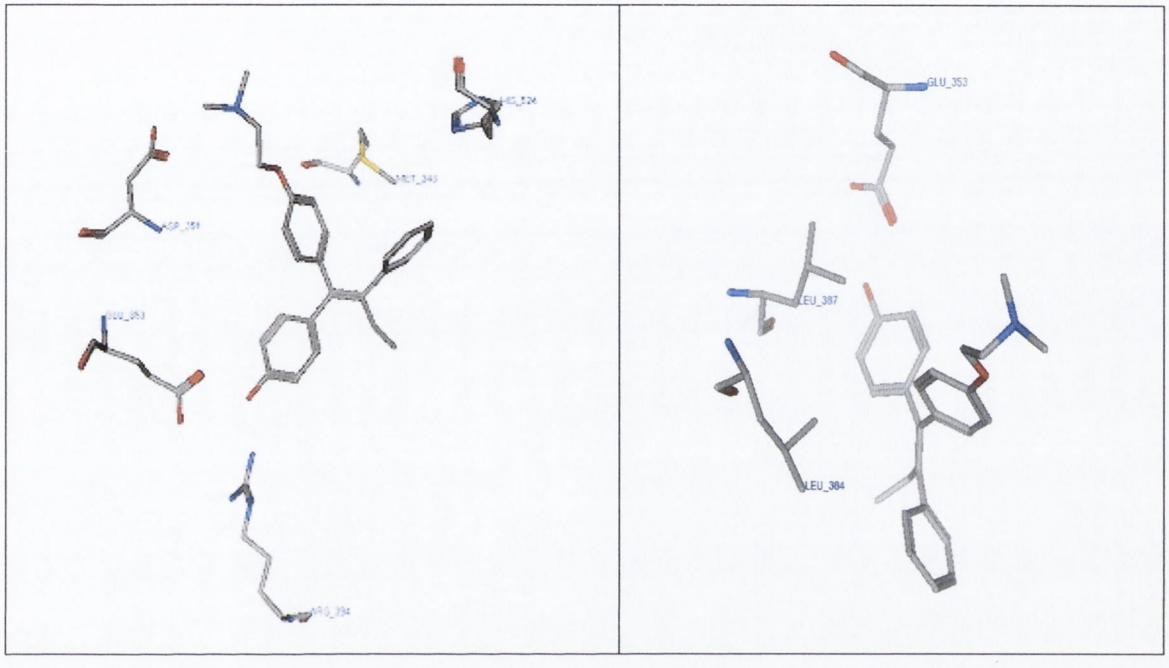


Fig **(8)** Residue-Ligand interactions of 4-Hydroxytamoxifen in crystal structure 3ERT.

To illustrate this process, we show utilising one of the scaffolds retrieved, how an online substructure search could be undertaken on a large database of commercial compounds (ZINC [40]) to enable compounds containing this scaffold with more appropriate substituents likely to bind in an antestrogenic manner to the ER to be identified.

Path 2 involved implementing a more focused approach, whereby LPC was re-applied to the docked complexes of the virtual screen from Path 1 and only those molecules bearing substituents that could interact through H-bonding to specific residues (Thr347, Glu353, Leu387, Arg394) are retained. This method although wholly reliant on the quality of the input database, is nonetheless very effective and numerous examples in literature describe docking/scoring successes [25, 41]. Corroborating this, we reveal a 53nM ERα binding compound with 7μm antiproliferative activity against the MCF-7 breast cancer cell line, retrieved from the top 7 compounds of the same virtual screen used to identify novel scaffolds. Importantly, Asp351 was not included in the filter list of

essential H-bonding residues as we wanted to identify a core modulating scaffold whose activity could be optimised towards agonism or antagonism by enumeration of a virtual library.

Finally, we generated a virtual library *de novo* based on the identified 'hit' by enumerating the same core with alternative antiestrogenic side-chains attached. The procedure utilised MoSS miner [42-44] to find substructures and discriminative fragments similar to those of known antiestrogenic side-chains. To ensure a reasonable set of substructures were found, Omega combined with ROCS was used to generate conformers of each and evaluate the Tanimoto similarity co-efficient with respect to the dimethylaminoethyl side-chain of tamoxifen. Applying the recently reported "Scaffold-Linker-Functional Group" (SLF) approach [45], SMILIB [46] was utilized to enumerate a virtual library of these compounds and the set re-processed through Path 2 to prioritize the compounds for synthesis and to suggest compounds that should possess increased inhibitory effects.

As previously discussed, we deemed the ER to be a feasible target to illustrate vHTS using our protocol, as there is a large amount of crystallographic data available [6, 30, 47-51] and the properties and requirements for ligand binding to the active site are reasonably well understood [52, 53]. Although there are now known to be two isoforms of the ER, namely ERα and ERβ, we chose to concentrate on the former as it has been implicated in the proliferation of breast cancer, especially through the prolonged use of Hormone Replacement Therapy (HRT) [54]. Consequently, Structure-Based Drug Design (SBDD) of novel inhibitors through exploitation of the ER is facilitated by the ligand-binding specificity and differential tissue distribution of each isoform [55]. In the case of both isoforms, the following is common, however for the purpose of this article we focus on these properties relative to ERα. Upon hormone binding Helix-12 orients itself in such a manner as to encapsulate the ligand in the hydrophobic cavity within. This re-positioning of Helix-12 allows co-activator recruitment to the AF-2 site and initiation of transcription. Conversely, binding of Raloxifene, as shown in Figure 9 disturbs the motion of the Helix and prevents this encapsulation and subsequent formation of the AF-2 site through interaction with Asp351 and the side-chain piperazine ring nitrogen of Raloxifene [56]. A few studies have reported the selective design of inhibitors by exploiting

these key differences between agonist and antagonist binding through vHTS methods [57-59].

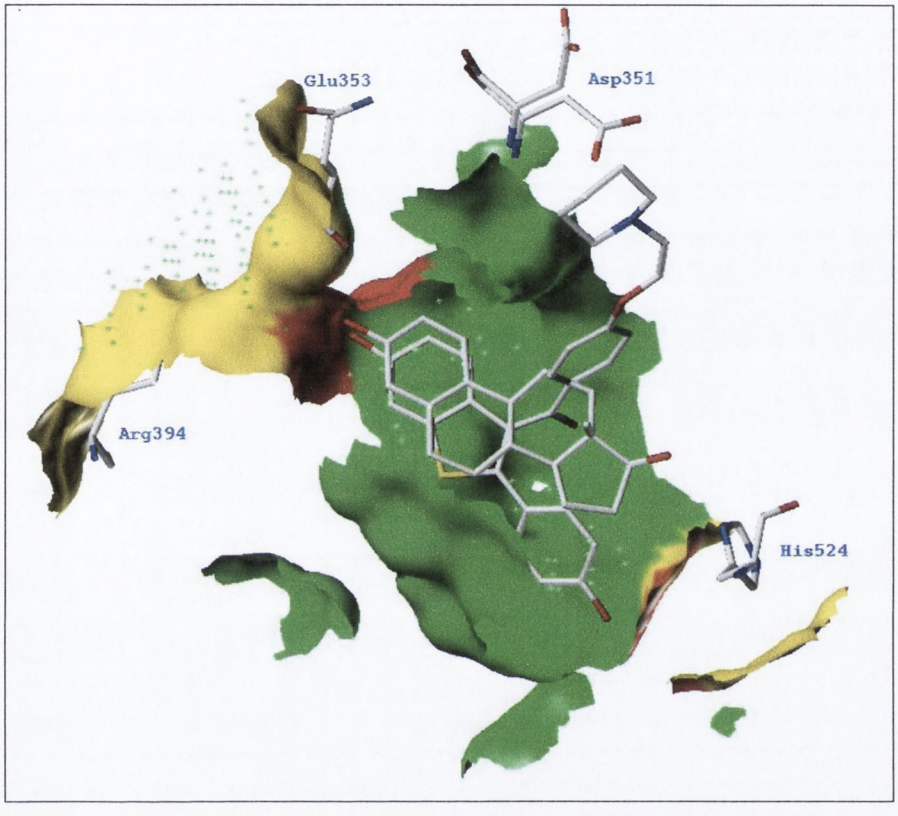


Fig **(9)** Residue-Ligand interactions of the agonist Estradiol ERα (PDB ID: 1ERE) and antagonist Raloxifene ERα (PDB ID: 1ERR).

To illustrate the full potential of vHTS, we present a study that draws on the main aspects involved in the process, utilizing ERα as a target. We show importantly the impact that addition of distance constraints in combination with an appropriate scoring function improves discrimination of actives from a set of inactives. Application of these 'in silico' methods with iterative wet-lab validation has allowed us to optimize our suite of algorithms and discover lead compounds of the ERα.

3.11 Experimental Section - Computational

3.11.1 Conformer Generation and Storage

Cheminformatic pre-processing of databases of molecules has been assessed by our group in relation to ERα in a previous study [60]. We have demonstrated the impact it has in the context of virtual screening and prioritization of compounds for biological evaluation using FRED 2.01 [61] as a rigid-exhaustive docking algorithm. Multiple protonated, tautomeric, stereochemical and conformational states were enumerated and their associated effects on Enrichment (E) rates and False Positive (FP) rates were examined using datasets of 1000 and 10,000 compounds respectively. Unexpectedly, the initial SMILES [4] representation of a compound prior to pre-processing had a significant impact on the Enrichment obtained. It is concluded that only generation of 10 conformers of each compound using OMEGA 1.81 [62] is needed to produce optimum results when docking in the ERα. Conversely, addition of multiple protonation, tautomeric and stereochemical states does not provide additional benefit. As a result we have chosen the same method of space sampling using OMEGA 1.81 in the current protocol.

To begin with, OpenBabel 1.100.2 [63] was utilized to convert the databases with full stereochemical information denoted. OMEGA 1.81 converts this database into a multi-conformer sdf database and again OpenBabel 1.100.2 is utilized to convert the multi-mol2 database into a multi-PDB file. A set of C subroutines automates these processes and also splitting of the multi-conformer file in to separate conformers. An open source database, MYSQL [64], is used as the core information storage system for the conformers. Fields containing conformer id, INPUT files, PROTEIN id, SMILES id, and job status ensure an elegant structure to the database is maintained.

3.11.2 Protein Preparation

The crystal structure 3ERT was downloaded from the Protein Data Bank and crystallographic waters were removed. The subsequent structure was read into Macromodel 6.5 [2] and re-written in PDB format to ensure bonds were represented

correctly in this format. LIGIN does not take hydrogen atoms into account in the docking process and so no addition or minimisation of them was needed.

3.11.3 LIGIN-based Docking Protocol

A description of the LIGIN docking program employed at this stage is provided. LIGIN is executed when three main files are present, INPUT, PROT and LIG. The LIG file consists of each conformation of a ligand in the database in standard PDB format. The PROT file is generated from the crystal structure (3ERT [6]), and contains the coordinates of the protein atoms and other atoms in the target and but does not include information about the ligand chosen. The input file is then generated from a set of arbitrary rules that classify and assign a number to particular atom types numbered 1-8:

1) Hydrophilic: - N and O atoms that can donate and accept H-bonds (e.g., oxygen of hydroxyl group of Ser or Tyr).

2) Acceptor: N or O atoms that can only accept H-bond.

3) Donor: N atom that can only donate H-bond.

4) Hydrophobic: Cl, Br, I and all C atoms that are not in aromatic rings and do not have a covalent bond to a hydrophilic atom.

5) Aromatic: C atoms in aromatic rings.

6) Neutral: C atoms that have a covalent bond to at least one atom of class 1, or two or more atoms from class 2 or 3; N atom if it has covalent bonds with 3 carbon atoms; S and F atoms in all cases.

7) Neutral-donor: C atom that has a covalent bond with only one atom of class 3.

8) Neutral-acceptor: C atom that has covalent bond with only one atom of class 2.

In order to reduce the sampling time, the co-ordinates of the LIG files are translated to those of the co-crystallised ligand (4-Hydroxytamoxifen), to ensure the docking begins in the binding site. LIGIN begins by generating a number of ligand positions in 6-dimensions in the binding site of the receptor. Each of the ligand positions has their respective binding modes assessed according to the complementarity function as given in equation (1). After generating random position of the ligand within the binding site, the

program then optimizes this position with simplex optimization method and using NC as scoring function. The docked positions obtained have their respective hydrogen bond lengths optimized to allow for refinement of the final structure. After searching for the global maximum of the complementarity function, the programme creates ≤20 files (CR1, CR2, CR3...) containing the coordinates of the ligand in PDB format that correspond to the 'global' (CR1) and 'local' maxima (CR2, CR3…). Merging of the PROT file and each CR file is carried out to produce the final docked complexes. Each step in the process, namely, extraction of ligand information from MYSQL database, generation of each INPUT file from the associated LIG files, translating the co-ordinates of the LIG files to the endogenous ligand, execution of LIGIN, and merging of the output CR files with PROT file are carried out by a series of C subroutines that produce a fully automated suite.

3.11.4 Tiered Scoring and Validation

A two-fold scoring scheme was adopted using Ligand Protein Contacts (LPC) software [9]. Firstly, the Normalized Complementarity (NC) according to LPC is calculated for each of the docked complexes. To train the scoring process a set of 19 active ERα inhibitors (Figure 10) extracted from literature with potencies ranging from nM to μM were introduced to the docking process and the lowest NC value was set as the threshold value for follow-up docking studies.

CC\C(=C(/c1ccc(O)cc1)c2ccc(OCCN(C)C)cc2)c3ccccc3
O=C(c2ccc(OCCN1CCCCC1)cc2)c4c(c3ccc(O)cc3)sc5cc(O)ccc45
COc4ccc(C3=C(c1ccc(OCCN(C)C)cc1)c2ccccc2OCC3)cc4
Cc4c(c1ccc(O)cc1)n(Cc3ccc(OCCN2CCCCC2)cc3)c5ccc(O)cc45
CN(C)CCOc5ccc(output 3CC1(C)[C@H](O)CCC1C4CCc2cc(O)ccc2C34)cc5
Oc1ccc3c(c1)CC[C@H](c2ccccc2)[C@@H]3c5ccc(OCCN4CCCC4)cc5
Oc1ccc3c(c1)CCN(c2ccccc2)C3c5ccc(OCCN4CCCCC4)cc5
CCCC3C(c1ccc(O)cc1)NN(c2ccc(O)cc2)C3c4ccc(O)cc4
CC\C(=C(\c1ccc(OCCN(C)C)cc1)c2cccc(O)c2)c3ccccc3
CC4=C(c1ccc(O)cc1)[C@H](c3ccc(OCCN2CCCCC2)cc3)Cc5cc(O)ccc45
CCCC(=C(c1ccc(O)cc1)c2ccc(O)cc2)c3ccc(O)cc3
CC\C(=C(\Cc1ccccc1)c3ccc(OCCN2CCOCC2)cc3)c4ccccc4
Oc5ccc([C@@H]2Sc1cc(O)ccc1O[C@@H]2c4ccc(OCCN3CCCCC3)cc4)cc5
Oc5ccc(C2CCc1cc(O)ccc1N2Cc4ccc(OCCN3CCCC3)cc4)cc5
Oc6ccc(C2=Cc1cc(O)ccc1C25Cc4ccc(OCCN3CCCCC3)cc4C5)cc6
Oc5ccc(c2sc1cc(O)ccc1c2c4ccc(OCCN3CCCCC3)cc4)cc5
O=c4oc1cc(O)ccc1c(Cc3ccc(OCCN2CCCCC2)cc3)c4c5ccccc5
Oc1ccc4c(c1)O[C@@H](c3ccc(OCCN2CCCCC2)cc3)c5c4ccc6cc(O)ccc56
COc5ccc(C4=C(C(=O)c2ccc(OCCN1CCCC1)cc2)c3ccccc3CC4)cc5

Fig **(10)** SMILES strings representation of 19 actives (See **Appendix B** for Compound structures)

A decoy set of 1000 compounds seeded with the same 19 actives was subsequently docked according to the above procedure with the threshold set. The resultant docked complexes were re-scored using the following scoring functions: F-Score, D-Score, PMF-Score, G-Score, Chemscore, and Drugscore as implemented in Sybyl6.91, Chemscore, Chemgauss, Chemgauss2, Shapegauss, plp, and Screenscore as implemented in FRED 2.11, Hammerhead as implemented in Surflex, and a standalone scoring function XScore.

Upon establishing E and FP rates for each scoring function in the process, the optimal scoring function was selected from a set of 14 diverse scoring functions. The 19 actives were docked and the LPC output for each solution was examined to determine the residues in contact with ligand and their respective distances. From a calculation of the interatomic contacts using LPC (output shown in figure 2) on the crystal structure 3ERT, the putative H-bonds were deemed to be Glu353, Arg394, Leu387, Thr347 and Asp351. Distance constraints were set for each of these residues according to the range of distances observed for the entire 19 actives. Several rounds of docking using the decoy set of 1000 compounds with re-scoring were undertaken with the distance constraints being adjusted each time accordingly, until the E and FP rates were maximized.

A final validative process was carried out using a decoy set of 9,999 molecules reflecting the same characteristics as the smaller set. We chose to seed the set with a single potent antiestrogen (3E) [65] to fully test the ability of the protocol to retrieve the hit from the decoy set.

3.11.5 Active and Decoy sets

Forty antiestrogens were selected from literature with activities ranging from nanomolar to low micromolar potency and converted to SMILES format using ACD/ChemSketch 8.17. The set was passed through FILTER [66] to remove those antiestrogens that were not considered to be 'drug-like' leaving only 19 remaining. Our laboratory and others have highlighted the importance of incorporating a set of actives in a decoy set that reflect the properties of the rest of the decoy set when validating a vHTS protocol [33, 67]. We sought to optimize the protocol towards discovery of inhibitors of ER that would also possess more 'drug-like' properties and our choices of filter parameters reflected this. A subset of the Derwent World Drug Index (WDI) [68] was then extracted and passed through FILTER using the same filtering properties, such as molecular weight <200 or >550, number of hydrogen bond donors $0<x<6$ and acceptors $0<x<10$, calculated logP <7. The set remaining totaled 10343 compounds. From this, five hundred molecules with stereochemical information denoted and 481 without were randomly selected using a proprietary Perl script. This was done to reflect the portion of marketed drugs that contain chiral centers or not. The two sets were merged with the 19 actives to produce a set of 1000 compounds with similar characteristics.

A larger database comprising 9,999 compounds was formed using the WDI and CHEMBANK. A single potent antiestrogen was added to the set to make up the 10,000. This set was previously used in our study of database pre-processing [69]. This set was considered here by us to validate the vHTS protocol more thoroughly.

3.11.6 Success Criteria

The success of out protocol was evaluated by two metrics. Firstly, Enrichment (E) rates were calculated for the top 0.6%, 1.2%, 1.8% and 2.4% as the number of actives was 19. Secondly, the efficacy of each protocol was also measured by assessing False Positive (FP) rates for 80% of the true positives.

3.11.7 Virtual Screen – Path 1

A virtual screen of the SPECS database screening collection (Release: Aug2005, 202054 compounds in total) was carried out using our optimised protocol to identify new scaffolds of ERα. Ten conformers of each molecule were generated using Omega 1.81 [62] and docked and scored according to the protocol detailed in the Tiered Scoring and Validation section. A visual inspection of the compounds that passed was undertaken and a number of scaffolds selected.

3.11.8 Virtual Screen – Path 2

From the screen in Path 1, the same docked complexes were re-filtered using LPC with additional constraints imposed to guarantee H-bonding of ligand atoms occurred with residues Thr347, Glu353, Leu387 and Arg394. Asp351 was not selected as a H-bonding residue to allow both agonist and antagonist cores to be discovered as agonists can readily be converted to antagonists through addition of an antiestrogenic side-chain.

3.11.9 Virtual Library Enumeration

Refraining from including Asp351 as a H-bonding residue in the filtering process provided additional scope for introduction of side-chains to 'hits' discovered through generation of a virtual library. A "Scaffold-Linker-Functional Group" (SLF) approach approach was adopted using SMILIB [46] with a complete reaction scheme for all possible combinations containing 1 R-group on each scaffold applied. Firstly, to identify molecular substructures and discriminative fragments that could be used as side-chains in

the enumeration process, MoSS miner was employed [42-44]. Using the 19 ERα actives, MoSS was run with a minimum support of 100. Selecting (O-C1:C:C:C:C:C:1) as the common core of all the antagonist side-chains, both Asinex (Release: July 2004, 121857 compounds) and SPECS (Release: Aug2005, 202054 compounds) were mined for substructures containing this core. To extract only reasonably sized fragments, the minimum support was set to 1 and minimum/maximum substructure size set to 10/25 respectively. Ten conformers of each fragment and the side-chain of Tamoxifen (SMILES = CN(C)CCOc1ccccc1) were generated using Omega 1.81. Finally, ROCS with the conformers of the side-chain of Tamoxifen as a query molecule and a Tanimoto cutoff of 0.85, retrieved similar structures from the fragment database. The resultant structures were split into linkers and functional groups as input for SMILIB. The set underwent screening with distance and H-bonding constraints imposed.

3.12 Experimental Section -Biological

3.12.1 Receptor Binding Assay

Competitive binding affinity experiments were carried out using purified baculovirus-expressed human ER-alpha (HrERα) and applying Fluoromone (ES2), a fluorescein-labeled estrogen ligand. Both were contained in the Estrogen Receptor (ER) Competitor Assay Kits, Green obtained from Invitrogen Corporation. HrERα was stored at -80°C, and not subjected to any vortexing.

Methodology

Upon binding of a fluorescent molecule to the HrERα and formation of the receptor-ES2 complex, it tumbles slowly resulting in a high polarization value. Polarization, measured in mP units, is directly related to the molecular volume of the tumbling molecule. Thus, if a competing compound displaces ES2, it causes the molecule to tumble rapidly in solution and results in a low polarization value being obtained. This change in mP value allows determination of the binding affinity of the compound. The concentration of the

competitor that results in a half-maximum shift in polarization is equivalent to the IC_{50} value, and a curve can be plotted using:

(2) $Y = mP_{100\%} + (mP_{0\%} - mP_{100\%})/1 + 10^{((LogIC50-X) * Hillslope)}$

where; Y = mP, X = Log performing , $mP_{100\%}$ = 100% inhibition, and $mP_{0\%}$ = 0% inhibition.

HrERα titration

HrERα was serially diluted from 400nM to 0.391 nM in screening buffer (40 mM Tris-HCl, pH 7.5; 50 mM KCl; 5% glycerol; 10% dimethylformamide; 0.02% sodium azide; 50 µg/ml bovine gamma globulin) to a final volume of 100 µl in borosilicate test tubes. ES2 was added to each tube at a concentration of 1nM and the tubes were mixed by shaking lightly. After incubation for 1 hour at room temperature, the FP assays were carried out using a Beacon 2000 fluorescence polarization instrument (PanVera Corporation) with 360 nm excitation filter and 530 nm emission filter. Fluorescence anisotropy was measured for each solution and the amount of ER that gives 80% of the maximal shift in mP was selected as the concentration to use for competitive binding studies.

Competitive Binding Assay

Competing compounds were prepared at a standard concentration of 10 mM in DMSO. HrERα and ES2 were combined on ice (4°C) in a glass vial to produce the receptor-fluoromone complex. The vial was gently inverted 2-3 times, again ensuring no vortexing of the mixture. In Duplicate, the compounds were serially diluted in ethanol to ensure the final concentration of DMSO and ethanol was below 1% in solution. 1µl of each solution was diluted in 49µl of buffer and added to 50µl of the receptor-fluoromone complex in borosilicate test tubes. Following 45min incubation, the samples underwent FP measurement. E_2 was used as a negative control and 50µl of the receptor-fluoromone complex in 50µl buffer as the positive control.

3.12.2 Antiproliferative studies

The ER(+) breast cancer MCF-7 cell line was maintained in 75-cm^2 culture flasks (Greiner) containing (Dulbecco) Eagles minimum essential medium in a 5% CO2 atmosphere with 10% fetal calf serum. The medium was also supplemented with 1% nonessential amino acids.

Methodology

The MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assay is based on the premise that mitochondrial dehydrogenase enzymes from viable cells will cleave tetrazolium rings of the pale yellow MTT resulting in the formation of intracellular purple formazan, which can be solubilized and quantified spectrophotometrically at 570nm.

MTT Assay

Cells were trypsinized and seeded at a density of 1.5 x 10^4 in a 96-well plate and incubated at 37 °C, 5% CO_2 atmosphere for 24 h. All compounds were prepared at a standard concentration of 10 mM in DMSO and serially diluted to produce a range of concentrations spanning, 1ηm-100μm. 2μl of each compound solution were added to the cells and reincubated for an additional 72h. Control wells contained 2μl of vehicle (DMSO) in all cases. At the end of the incubation period, culture medium was removed and all cells were washed with 100μl PBS. 50μl of MTT solution was added to each well and the plates were incubated in darkness for ~2hrs at 37°C. The converted dye was solubilised with 200μl DMSO and pipetted up and down several times to ensure the dye dissolves completely. Absorbance of the converted dye was measured at 570nm with control cells set to 100% cell viability.

3.12.3 Cytotoxicity Studies

Lactate dehydrogenase (LDH) assay was used to measure cellular toxicity effects of the various doses of each compound and was examined using a colorimetric determination kit (Promega).

Methodology

Lactate dehydrogenase (LDH) is usually released upon cell lysis and converts the tetrazolium substrate provided into a red formazan product as shown below in Equation 2.1/2.2. This can also be solubilized and quantified spectrophotometrically at 490nm.

(2.1) NAD^+ + Lactate $\xrightarrow{LDH}$ Pyruvate + NADH

(2.2) NADH + INT $\xrightarrow{Diaphorase}$ NAD^+ + Formazan (Red).

LDH Assay

The assay was carried out concurrently with the MTT assay following dosing of the compounds and incubation for 72 hrs as above. Prior to removal of the culture medium in the MTT assay, 50µl aliquots of medium were removed to a fresh 96-well plate. 50µl of LDH solution was added to each well and the plate was left in darkness for ~20-30 minutes at room temperature. 50µl stop solution was then added to each well and the absorbance read at 490nm on a micro-plate reader. Control of 100% lysis was measured by addition of 20µl lysing solution 45 minutes prior to harvesting.

3.13 Results and Discussion

A protocol was set out that ultimately allows the retrieval of both scaffolds and novel molecules that can modulate the actions of the ERα. Another goal was to demonstrate that VS of this target is not entirely reliant on the choice of an optimal scoring function, and that certain steps may be incorporated in a protocol that can substantially enhance

both Enrichment and False Positive rates. Finally, a method for enumeration of a virtual library based on one of the 'hits' retrieved from a virtual screen was introduced.

Normalised Complementarity Threshold

As previously mentioned, the first step in the process was to introduce an initial scoring step to remove compounds that would otherwise not dock with a reasonable orientation or binding mode, and thus score poorly. This step was carried out by implementing a function taken from LPC software [9], known as Normalised Complementarity, and defined as previously (eqtn 12, section 1.8.4)

Table 6 shows the list of 19 active ligands extracted from literature with known antiproliferative activies ranging from nM-μM for the ER. The lowest LPC value produced by these compounds was selected as the minimum threshold value (0.81) that must be overcome allowing a molecule to move to the next stage in the process.

Table **(6)** 19 active ligands extracted from literature with ER binding activity

Name (ref)	**Complementarity Value**	**ERα Binding IC_{50}** (nm)/ (ref)
4-Hydroxytamoxifen [70]	1	8.5 [71]
Raloxifene [72]	0.89	1.4 [71]
(2-{4-[4-(4-methoxy-phenyl)-2,3-dihydro-benzo[b]oxepin-5-yl]-phenoxy}-ethyl)-dimethyl-amine [73]	0.89	11100 [73]
2-(4-hydroxy-phenyl)-3-methyl-1-[4-(2-piperidin-1-yl-ethoxy)-benzyl]-1H indol-5-ol hydrochloride [74]	0.98	0.2 [74]
11-[4-(2-dimethylamino-ethoxy)-phenyl]-13-methyl-7,8,9,11,12,13,14,15,16,17-decahydro-6H-cyclopenta[a]phenanthrene-3,17-diol [75]	0.94	1 [76]
Lasofoxifene [77]	0.97	2.3 [71]
2-phenyl-1-[4-(2-piperidin-1-yl-ethoxy)-phenyl]-1,2,3,4-tetrahydro-isoquinolin-6-ol [71]	0.93	29 [71]
1,3,5-tris (4-hydroxyphenyl)-4-propyl-1H-pyrazole [78]	0.99	18 [78]
Droloxifene [79]	1	10000 [80]
EM652 [81]	0.94	0.146 [82]
1,1,2-tris(4-hydroxyphenyl)pent-1-ene [83]	1	20000 [83]
4-{2-[4-(1-benzyl-2-phenyl-but-1-enyl)-phenoxy]-ethyl}-	0.95	2530 [65]

morpholine [65]		
3-(4-hydroxy-phenyl)-2-[4-(2-piperidin-1-yl-ethoxy)-phenyl]-2,3-dihydro-benzo[1,4]oxathiin-6-ol [51]	0.86	3 [84]
2-(4-hydroxy-phenyl)-1-[4-(2-pyrrolidin-1-yl-ethoxy)-benzyl]-1,2,3,4-tetrahydro-quinolin-6-ol [85]	0.91	68 [85]
Table **(6) (cont.)**		
2-Phenylspiroindene ([86])	0.87	1([86])
Desketoraloxifene ([87])	0.94	-
7-hydroxy-3-phenyl-4-[4-(2-piperidin-1-yl-ethoxy)-benzyl]-chromen-2-one ([88])	0.83	7.7 ([89])
LY357489 ([90])	0.84	0.4 ([90])
Trioxifene ([91])	0.81	203.49([92])

It is also clear however that there is no direct correlation between the complementarity value and the antiproliferative values. What is clear is that the antiestrogens require a substituent on the A-ring that can form a Hydrogen bond with Glu353 and Arg394 to compete significantly with estradiol. Applying this methodology to a validation set of 1000 compounds seeded with 19 actives, a total of 140 compounds could not dock appropriately.

Rescored Docked Complexes

We next sought to examine the effect that re-scoring docked complexes would have on Enrichment. Table 7 depicts the Enrichment calculated for 0.6%, 1.2%, 1.8% and 2.4% of the ranked hitlist. The increasing order of merit of each is D_Score=PMF_Score=F_Score=G_Score=Drugscore=FRED_Chemscore=HammerHead <Fresno<Chemscore<FRED_plp<Xscore<FRED_Screenscore<FRED_Shapegauss<FRED_Chemgauss<FRED_Chemgauss2. FRED_Chemgauss2 integrates the Shapegaussian scoring function with potentials based on smooth gaussian functions also for groups that are chemically complimentary (mainly H-bonding interactions). From Table 7 is undoubtedly the best performing scoring function. Noteworthy, a wide disparity is observed in enrichments using different implementations of the Chemscore scoring function (Sybyl6.91/FRED2.11). Interestingly also, seven of the fifteen scoring functions provide no enrichment at all.

Table **(7)** Enrichment of actives for ERα using 15 different scoring functions.

Database Size	Chemscore	D_Score	PMF_Score	G_Score	Drugscore	F_Score	FRED_Chemgauss	FRED_Chemgauss2
0.6%	0	0	0	0	0	0	36.21	45.26
1.2%	4.526	0	0	0	0	0	31.68	31.68
1.8%	12.07	0	0	0	0	0	27.16	27.16
2.4%	11.315	0	0	0	0	0	22.63	22.63

Database Size	FRED_Chemscore	FRED_plp	FRED_Screenscore	FRED_Shapegauss	Fresno	HammerHead	Xscore
0.6%	0	18.11	27.16	36.21	0	0	18.11
1.2%	0	18.11	22.63	27.16	9.05	0	18.11
1.8%	0	12.07	15.09	24.14	6.04	0	18.11
2.4%	0	11.32	13.58	22.63	11.32	0	13.58

Distance Constraints

LPC (Ligand Protein Contacts) software is used to calculate all the receptor residues in contact with the ligand and their respective types of contact. To highlight the residue distances that need to be adjusted in the constraining process, contacts were initially calculated for the PDB structure 3ERT as shown in Table 8.

Table **(8)** Residues in contact with the ligand OHT600 in PDB entry 3ERT.

Residue	Dist	Surf	Specific contacts			
			HB	Arom	Phob	DC
343A MET*	3.8	26.0	-	-	+	-
346A LEU*	3.6	42.2	-	-	+	-
347A THR*	3.7	40.5	+	-	-	+
349A LEU*	4.1	13.9	-	-	+	-
350A ALA*	3.3	32.8	-	-	+	-
351A ASP*	3.2	29.0	+	-	-	+
353A GLU*	2.4	34.2	+	-	-	-
354A LEU*	6.5	1.3	-	-	-	-
383A TRP*	3.7	32.8	-	+	-	-
384A LEU*	4.0	25.1	-	-	+	-
387A LEU*	3.7	40.1	+	-	+	+
388A MET*	4.4	10.3	-	-	+	-
391A LEU*	4.1	19.7	-	-	+	-
394A ARG*	3.0	22.2	+	-	-	-
404A PHE*	3.8	21.5	-	+	+	-
419A GLU	3.9	2.0	-	-	-	-
420A GLY	3.8	15.9	-	-	-	-
421A MET*	3.5	50.3	-	-	+	-
424A ILE*	4.0	12.1	-	-	+	-
428A LEU*	3.7	17.9	-	-	+	-
521A GLY*	3.6	34.8	-	-	-	-
524A HIS*	4.0	14.4	-	-	+	-
525A LEU*	3.8	47.3	-	-	+	+
528A MET*	5.3	13.0	-	-	-	-

```
    530A  CYS*      6.1      4.9     -       -       -       -
    536A  LEU*      6.3      3.1     -       -       -       -
    539A  LEU*      6.3      2.9     -       -       -       -

-------------------------------------------------------------
Dist, nearest distance (Å) between atoms of the ligand and the residue; Surf, contact
surface area (Å²) between the ligand and the residue; HB, hydrophilic-hydrophilic contact
(hydrogen bond); Arom, aromatic-aromatic contact; Phob, hydrophobic-hydrophobic contact;
DC, hydrophobic-hydrophilic contact (destabilizing contact); or +/- indicates
presence/absence of a specific contacts between ligand and residue.
*, indicates residues contacting ligand by their side chain (including CA atoms).
```

From Table 8, it is evident that Thr347, Asp351, Glu353, Leu387 and Arg394 all form H-bonds with atoms on the ligand. Utilising a Perl script that scans this section of the LPC output for every docked ligand we introduced a set of filters that permitted either 'scaffold-hopping' or direct hit retrieval.

Beginning by allowing only those ligands that have a hydrogen bonding distance of $2.5 \leq x \leq 3.5$ between the negatively charged carboxyl groups of Asp351 and the nearest atom of the ligand, all of the actives could be docked correctly and numerous inactives removed. Several iterative adjustments were made to the other residue distance constraints in the Perl script to ensure that the actives produced a good binding pose and those inactives that did not adhere to these distances were removed. Re-applying Chemgauss2 at this stage allowed one to examine the actual ranking of the hitlist. Figure 11 illustrates the section of the Perl script that was tailored.

```
if( (2.6 <= $asp351 && $asp351 <= 3.2) &&
    (2.4 <= $glu353 && $glu353 <= 4) &&
    (2.5 <= $leu387 && $leu387 <= 4) &&
    (2.8 <= $arg394 && $arg394 <= 5.4) &&
    (2.8 <= $thr347 && $thr347 <= 3.7) &&
    (2.9 <= $his524 && $his524 <= 5.1) &&
    (4.8 >= $leu384) &&
    (5.2 >= $leu349) &&
    (2.5 <= $met343) &&
    ($lpcval > $threshold) ) {

$trash=0;
print "Keeping conformer with value of $lpcval\n";
```

Fig **(11)** Distance constraints implemented in Perl script

His524, Leu384, Leu349 and Met343 were added to the constraints, as they appear to provide additional important interactions in the binding process for antiestrogens [50]. A drastic reduction the number of docked complexes needed to be re-scored resulted using this process, which ensured that all actives remained, and only 52 inactives passed. What is important to note at this stage is that no specific H-bonding constraints have been applied from the HB column of Table 6. Only interatomic distance constraints have been imposed. We wanted to examine if these distance constraints were sufficient to allow discrimination between the actives and inactives in the dataset.

Evaluation of Screening Accuracy

The criteria of Enrichment $\{(H_A/H_T)/(A/D)\}$ and False Positive $\{(H_T-H_A)/(D-A)\}$ calculation was utilised to assess the accuracy of our procedure, where H_t and H_a are the total number of compounds and total number of active ligands in the hit list respectively. A is the total number of active ligands in the database and D is the total number of compounds in the database.

Table **(9)** Comparison of E rates for ChemScore before (wo) and after addition (w) of distance constraints.

Database Size	Chemgauss2 (wo)	Chemgauss2 (w)	Theoretical Max
0.60%	36.21	45.26	45.26
1.2%	31.68	45.26	45.26
1.8%	27.16	42.24	45.26
2.4%	22.63	31.68	45.26

All Enrichment calculations are adjusted to 860 compounds post NC filtering.

Table **(10)** Comparison of FP rates for ChemScore before and after addition of distance constraints.

True Positive (%)	Chemgauss2 (wo)	Chemgauss2 (w)	Theoretical Max
80	19.62	0.95	0
90	40.55	2.49	0
100	99.29	5.59	0

All False Positive calculations are adjusted to 860 compounds post NC filtering.

As observed in Table 9 and 10, the optimised protocol dramatically improves both Enrichment and False Positive rates for the dataset. Prior to incorporation of distance constraints with the normalized complementarity threshold of 0.8, the E rates observed using Chemgauss2 were 36.21 in the first 0.6% and 31.68 for the first 1.2% respectively.

The addition of these constraints to focus docking, furnished maximum enrichment of 45.26 for both 0.5% and 1%. Interestingly, the least active compound was also ranked second last in the hit list of actives. The calculation of FP rates for both protocols serves to clearly distinguish the advantage of adding distance constraints in a screening process.

Comparison with Bissantz Dataset and 10,000 set

To provide a direct comparison of False Positive rates of our protocol with other screening platforms currently available, we utilized a data set of 1000 compounds as proposed by Bissantz et al [20]. Several programs with their associated scoring functions have been evaluated by their ability to retrieve actives from a set of inactives using this dataset and the crystal structure of the ERα as a target (Glide [93], FRED [61], Pro_Leads [94], Dock [15], FlexX [14], Surflex [95] and Gold [16]).

Table **(11)** False positive rates for several docking algorithms

True Positive %	GEMDOCK	LIGIN	GOLD/DOCK	Surflex	GOLD	DOCK	FlexX
80	0	0.9	1.2	1.3	5.3	13.3	57.8
90	0.4	------	1.5	1.6	8.3	17.4	70.9
100	0.9	------	12.1	2.9	23.4	18.9	------

Table 11 gives a clear indication of the advantages of this type of protocol in generating low false positive rates. LIGIN outperforms all docking/scoring combinations except GEMDOCK whose false positive rates were 0.9. Noteworthy, LIGIN (our protocol) could not provide a docked structure of either of the compounds ICI-164384 or RU-58668 using the constraints we have imposed on the docking process. It is reasonable to expect this, as the actives used in our initial validation procedure specifically excluded any steroidal compounds with large extended side-chains and for this reason was not

optimized towards retrieving this type of compound. Thus, the distance boundaries may have been too tight to sufficiently allow them to dock well. Importantly, however, we were encouraged to see that among the actives retrieved in the Bissantz hitlist, one of the random compounds ranked 6th was actually a known antiestrogen tamoxifen (MFCD00010454). We also note that re-scoring with Chemgauss (FRED2.01) gives an FP rate of 0.1 for a true positive rate of 80%.

Finally, to rigorously test the procedure, we screened a set of 10,000 compounds with similar properties to antiestrogens seeded with a single known antiestrogen used previously by us to show the importance of pre-processing a database prior to docking [69]. From the ranked database of 10,000 compounds our procedure managed to select the single antiestrogen in 14th place. Generally, following a successful virtual screen, the top 1% are ordered for further biological testing. In this case, this would translate to the top 100 compounds selected. Thus the single antiestrogen would have been retrieved from the set successfully.

Virtual Screen – Path 1

The procedure at this stage although specific in discriminating between actives and inactives is still 'fuzzy' enough to allow a molecule that docks in the correct orientation with appropriate distances from the required residues to pass the LPC filter. This effectively means that new scaffolds can be derived from a virtual screen of a compound database that may not possess the appropriate H-bonding substituents to be immediate binders, but that could be tailored to do so. To highlight this we carried out a virtual screen of the SPECS database (Release: Aug2005, 202054 compounds in total) utilizing this method. A selection of some the scaffolds obtained is illustrated in Figure 12.

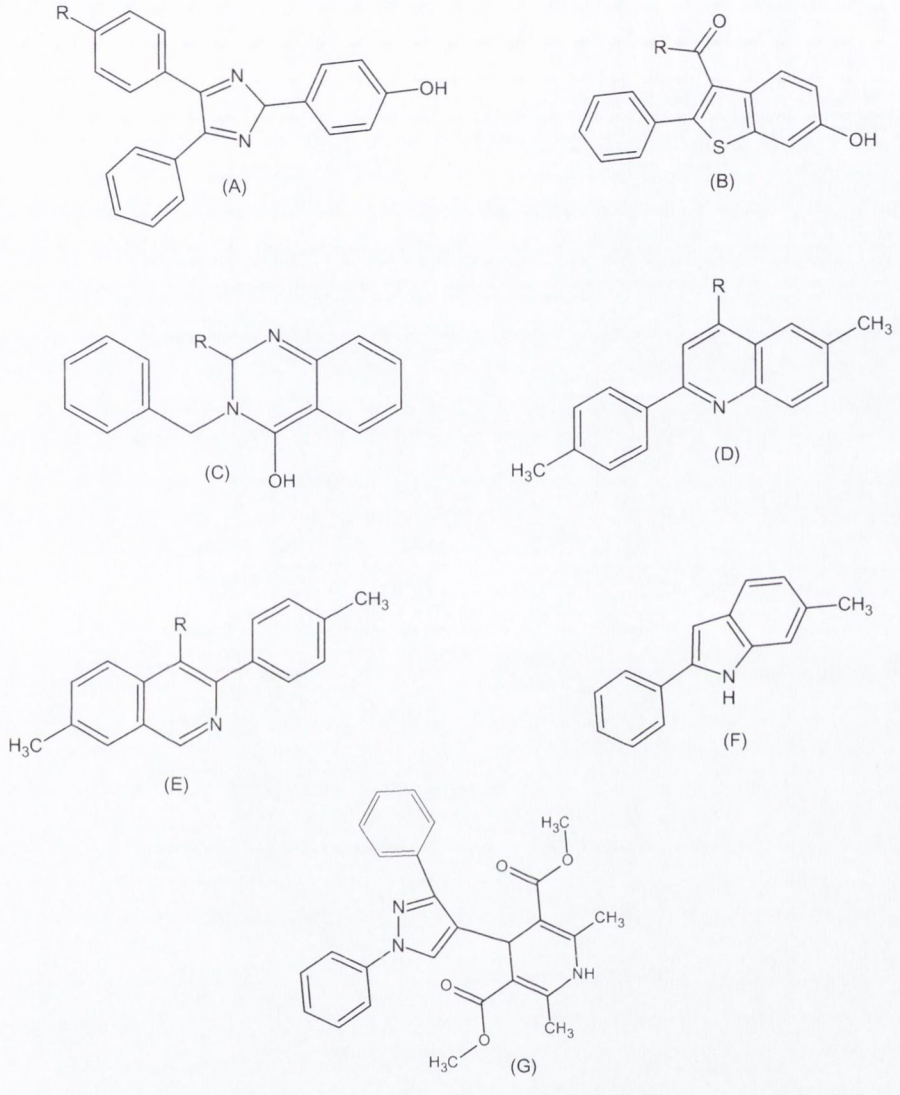


Fig **(12)** Several selected scaffolds identified by vHTS protocol outlined.

It is apparent that many of the chemotypes illustrated represent not only known scaffolds but also new one's. Scaffolds (B) & (F) represent known scaffolds present in the raloxifene moiety and ZK-119010 moiety respectively. Scaffold (A) represents a triarylimidazole-type scaffold, which has been previously investigated as an ER binder. However, Fink et al [96] have previously demonstrated the high affinity binding of 1,3,5-triaryl-alkyl-pyrazoles and importantly, Stauffer et al [97] have detailed the differences in

binding affinity when the core (diazoles, imidazoles, pyrazoles) are replaced by one another with the rings containing the same substituents in the same positions. Although the imidazole core still permitted binding to the ER it was less efficacious than the pyrazole core (Scaffold G). Scaffolds (C), (D) and (E) all contain a quinoline core that has been previously shown to be effective when incorporated in an antiestrogenic moiety as illustrated in Figure 13 below.

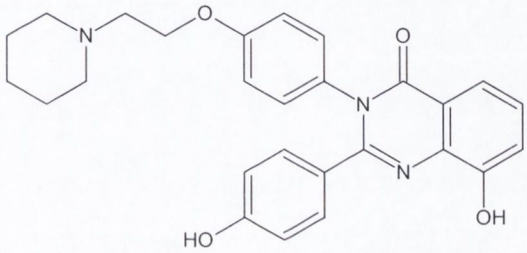


Fig **(13)** Quinoline structure developed by American Home Products.

Using an example scaffold from Figure 10, a substructure search was carried out using the ZINC database [40] to identify compounds that would have more antiestrogenic-like characteristics. The set were re-docked and scored using the same procedure, however, to add an additional degree of specificity towards antiestrogenic-like molecules, the LPC filter had a HB requirement of H-bonding by substituents of the ligand to Glu353/Arg394 and Asp351 of the receptor. The top-ranking compound is depicted in Figure 14.

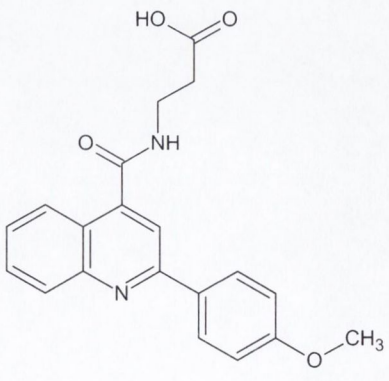


Fig **(14)** Antiestrogen-like compound identified by substructure search of scaffold identified by vHTS.

Virtual Screening – Path 2

Continuing with this methodology where specific H-bonding interactions must be obeyed along with distance criteria, we sought to directly identify a compound that would modulate the ER from the same screen of the SPECS database carried out to identify novel scaffolds. The set of ranked compounds from the vHTS of SPECS were filtered again using LPC with H-bond constraints set (Arg394, Glu353, Thr347, Leu387). Following visual inspection of 13 compounds remaining (Figure 15a – 6 compounds not selected for biological testing), and as a proof of concept a set of 7 compounds (Figure 15b) with ER-like characteristics were purchased and examined for their ability to bind to the ER by fluorescence binding assay.

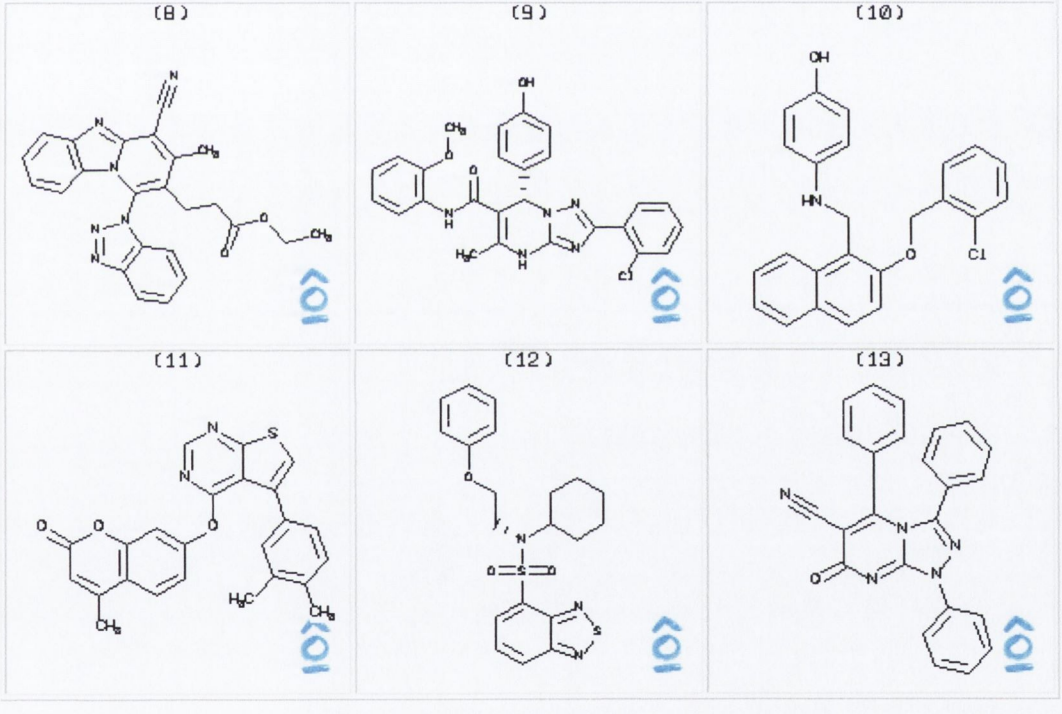


Fig **(15a)** Hits identified by vHTS but rejected from biochemical testing

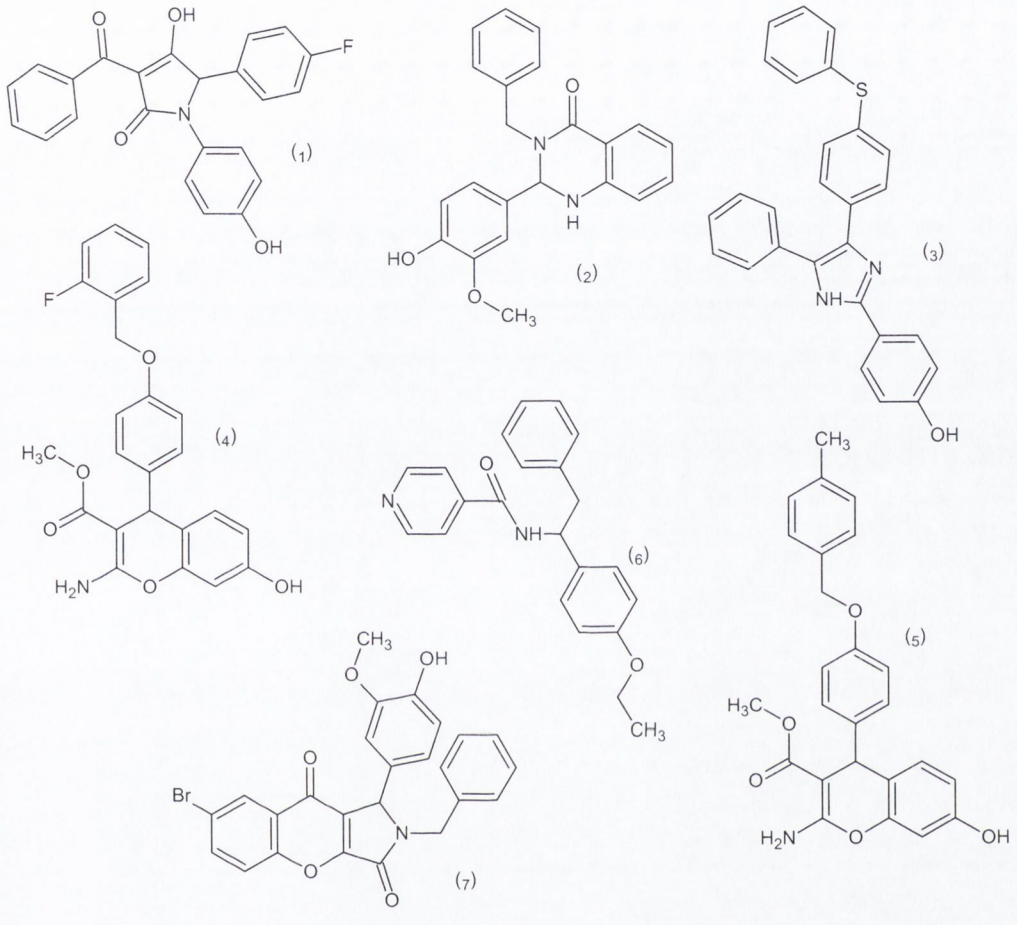
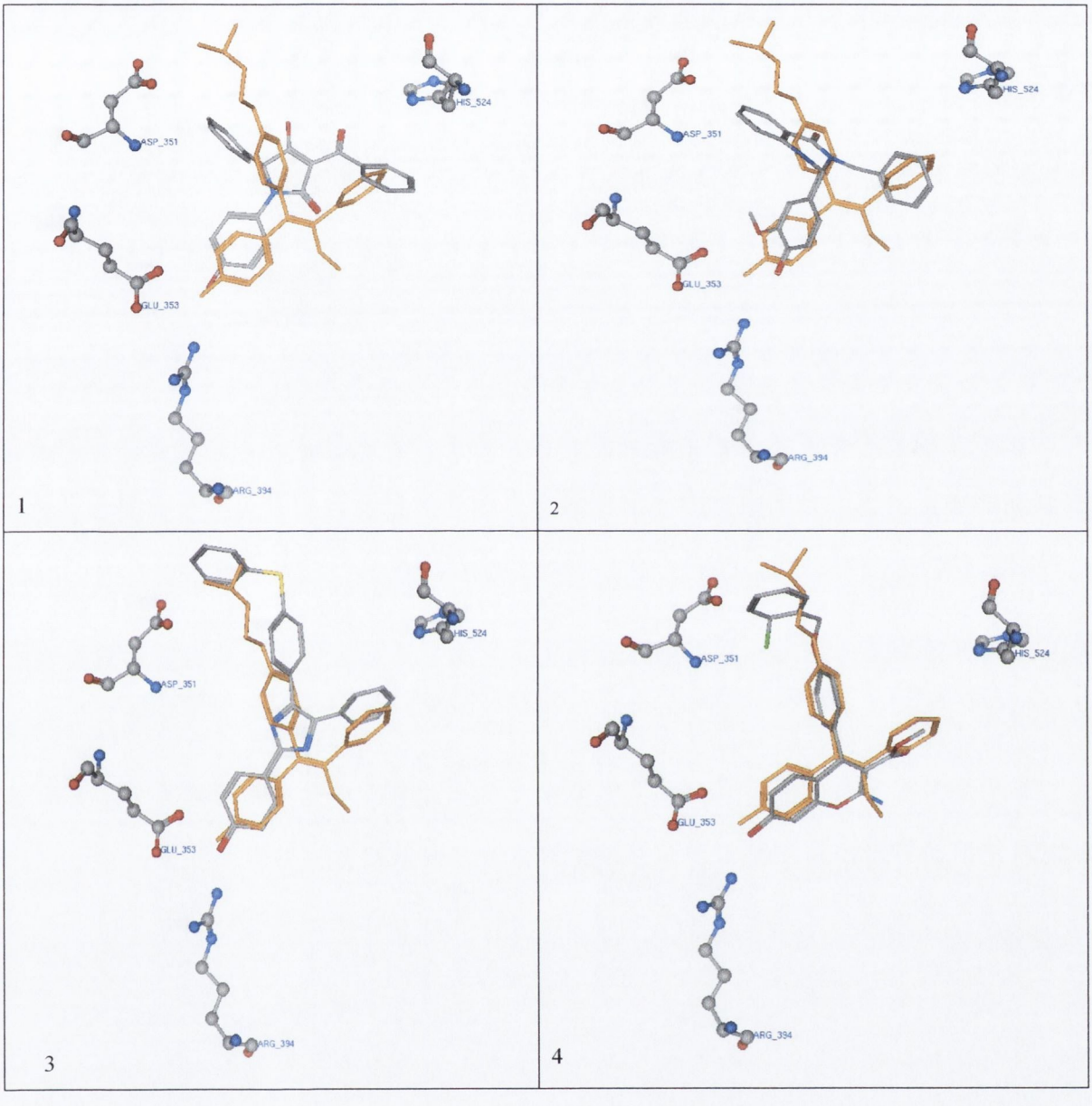


Fig **(15b)** Hits identified by vHTS and chosen for biochemical testing

Figure 16 illustrates predicted binding modes of the 7 compounds and gives an indication as to why on assaying, three compounds displayed IC_{50} values of (2) 1.1μM, (4) 53nM and (5) 56nM for binding to ERα (See Appendix C).


HIS_524
ASP_351
GLU_353
ARG_394
1
HIS_524
ASP_351
GLU_353
ARG_394
2
HIS_524
ASP_351
GLU_353
ARG_394
3
HIS_524
ASP_351
GLU_353
ARG_394
4

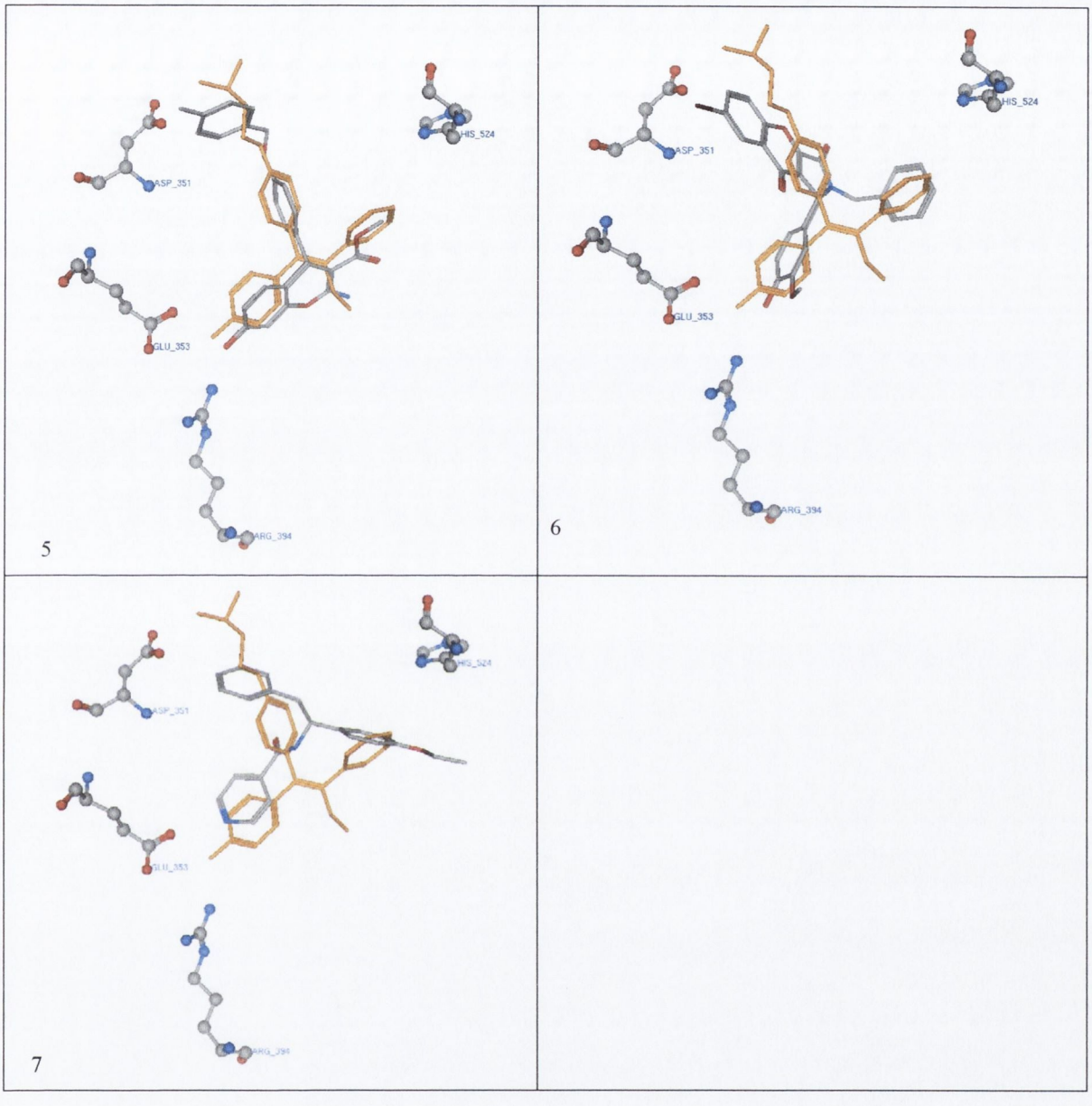


Fig **(16)** X-ray of 4-Hydroxytamoxifen (orange) in active site of ERα (3ERT) with docked structure of hits 1-7 overlayed.

From Figure 16 it is clear that compounds 4 and 5 adopt an orientation very close to that of 4-hydroxytamoxifen (orange). The presence of free hydroxy substituents allows interaction with Glu353 and Arg394 and also His524. These compounds were shown to

possess the best binding affinity corroborating the computational analysis. Compound 4 possesses a fluorine on the ortho position which appears to reduce the binding affinity because of steric interactions with Asp351, compared with compound 5 which possesses a methyl group on the para position of the side-chain ring. The affinity of compound 2 for ERα is probably due to the presence of a para hydroxy group near Glu353 and Arg394 which permits a strong H-bond interaction to occur.

LPC output of the three active docked complexes is depicted below in Table 12 and shows how each substituent interacts through H-bonding to different residues of the binding site.

Table **(12)** H-bonding distances from active site residues compared with 4-hydroxytamoxifen

Compound 2								
11	O	II	THR	347	N	III	2.9	17.3
11	O	II	THR	347	OG1	I	3.1	6.2
18	O	I	LEU	387	O	II	3.5	13.7
18	O	I	ARG	394	NH2	III	3.7	16.8
18	O	I	GLU	353	OE2	II	3.9	1.9
Compound 4								
18	O	II	THR	347	OG1	I	3.0	7.1
26	O	I	ARG	394	NH2	III	2.8	24.3
26	O	I	GLU	353	OE2	II	3.2	8.5
26	O	I	LEU	387	O	II	3.4	8.5
26	O	I	GLU	353	OE1	II	3.8	2.4
Compound 5								
11	O	II	THR	347	N	III	2.9	17.3
11	O	II	THR	347	OG1	I	3.1	6.2
18	O	I	LEU	387	O	II	3.5	13.7
18	O	I	ARG	394	NH2	III	3.7	16.8
18	O	I	GLU	353	OE2	II	3.9	1.9
4-hydroxytamoxifen								
9	O20	II	THR	347	OG1	I	4.0	4.3
12	N24	I	ASP	351	OD1	II	3.8	1.6
21	O4	I	GLU	353	OE2	II	2.4	16.5
21	O4	I	ARG	394	NH2	III	3.0	18.4
21	O4	I	GLU	353	OE1	II	3.3	1.0
21	O4	I	LEU	387	O	II	3.9	4.5

Table 12 illustrates the distances observed between atoms of the ligand and those of the active site residues. It is clear that all ligands do not interact the same way as 4-hydroxytamoxifen with Asp351, however, all other interactions are similar.

To determine whether our screen was specific enough to select hits that would preferentially bind ERα over ERβ, we examined binding of these compounds to ERβ by the same method. Compounds 2, 4 and 5 exhibited binding IC_{50} values of 6.2uM, 780nM and 915nM respectively, showing a 4.4, 13.7 and 17-fold selectivity ERα over ERβ and demonstrating the efficacy of our protocol (See Appendix D).

To evaluate the ability of these compounds to inhibit proliferation of MCF-7 breast cancer cells, an MTT assay was also carried out. The compounds exhibited (2) 15μm, (4) 11.4μm and (5) 7μm inhibitory activity reflecting the binding results obtained (See Appendix E). As discussed elsewhere [67], the key to turning an estrogenic substance into an antiestrogen is by addition of a side-chain such as that of a dimethylaminoethyl of tamoxifen. It is interesting to note that none of the compounds possessed the ability to interact with what is usually considered to be the key antiestrogenic residue, Asp351, but yet they all exhibited inhibitory activity close to that of Tamoxifen (4.6uM).

Noteworthy, a literature search of compounds 2, 4, and 5 revealed that they are novel with respect to ER binding. Compounds 4 and 5 have been previously synthesized by Elagamey et al [98], and compound 2 synthesised by Chernobrovin et al [99] for mechanistic purposes. The 4H-Chromene substructure has recently been shown to induce apoptosis through caspase high throughput screening but differs from compounds 4 and 5. Tetrahydroquinozoline derivatives have previously been shown to exhibit fungicidal activity but as before are not structurally similar to those obtained by our screening protocol.

To further investigate the value of these compounds, the core of compound 4/5 was extracted and served as a basic scaffold, which underwent side-chain inter-conversion by generation of a virtual library and re-docking to maximize the compounds interactions with Helix-12 and produce a more potent active.

Virtual Library Enumeration

A database of antagonist side-chains was constructed using MoSS miner from substructures or discriminative fragments within both Asinex and Specs databases. ROCS was subsequently utilized to ensure the side-chains obtained were similar to those of known antiestrogens such as tamoxifen. Upon splitting of the substructures into linkers and functional groups, SMILIB generated all possible combinations (456 molecules) at a specific attached point shown in Figure 17. Figure 17 also depicts the top-ranking compound after docking, scoring and filtering as in Path 2 was carried out.

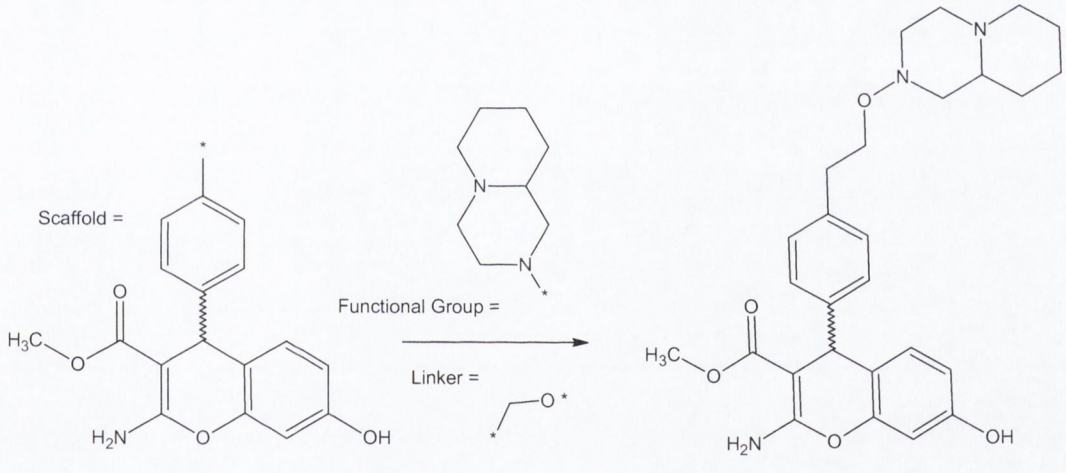


Fig **(17)** Top-ranked molecule from virtual library post-docking, scoring and filtering.

The binding orientation of this virtual 'hit' is shown in Figure 18 below with LPC output given in Table 13.

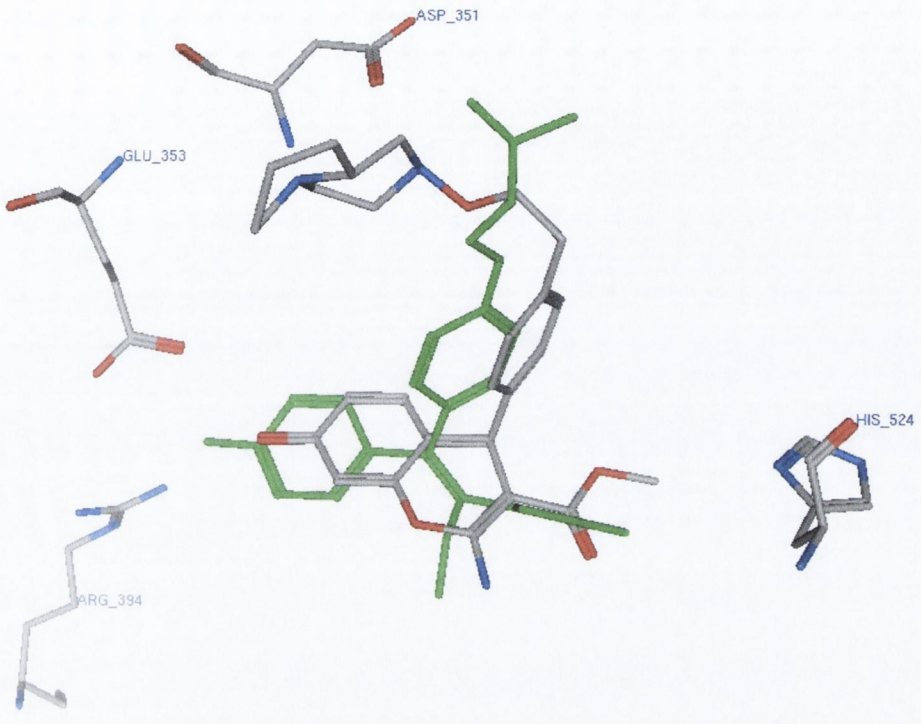


Fig **(18)** Virtual 'Hit' docked in active site of 3ERT.

It is clear from Figure 18 that the optimized hit is predicted to adopt a similar binding mode to that of 4-hydroxytamoxifen with importantly a significantly shorter interaction distance observed between the side chain nitrogen and Asp351 when compared with 3ERT as in Tables 13 & 14.

Table **(13)** Putative H-bond between virtual 'hit' and residues of active site.

Ligand atom			Protein atom					
N	Name	Class	Residue		Name	Class	Dist	Surf
16	O1	I	GLU	353	OE2	II	3.2	18.7
16	O1	I	LEU	387	O	II	4.2	4.2
16	O1	I	GLU	353	OE1	II	4.6	0.2
16	O1	I	ARG	394	NH2	III	4.6	7.1
26	N1	I	ASP	351	OD1	II	2.6	5.3

Table **(14)** Putative H-bond between 4-hydroxytamoxifen and residues of active site.

Ligand atom			Protein atom			Dist	Surf
N	Name	Class	Residue	Name	Class		
9	O20	II	THR 347	OG1	I	4.0	4.3
12	N24	I	ASP 351	OD1	II	3.8	1.6
21	O4	I	GLU 353	OE2	II	2.4	16.5
21	O4	I	ARG 394	NH2	III	3.0	18.4
21	O4	I	GLU 353	OE1	II	3.3	1.0
21	O4	I	LEU 387	O	II	3.9	4.5

At this stage the value and efficacy of the protocol have been shown and it is apparent that not only can hits be identified using the procedure but they can also be optimized to produce a more 'lead-like' compound. Table 15 summarizes the biological results obtained for the three active compounds outlined previously.

Table **(15)** Summary of results obtained for biologically active compounds

Compound No.	ERα	ERβ	α/β	MTT
2	1.4	6.2	4.4	15
4	56.8	780	13.7	11.4
5	53.3	915	17.2	7

3.14 Conclusion

The virtual screening protocol outlined in this study has been optimized to carry out in-silico screening of the ER protein combined with a universally utilized scoring function. By incorporation of distance constraint filters, both novel scaffolds and lead compounds can be directly obtained for the ERα protein. We have also shown that this method can be used in conjunction with any scoring function because the final deciding factor involved in the hit retrieval process is that of the distance constraints imposed. Therefore as long as the threshold used as a cutoff to select compounds from the ranked hitlist is not too stringent, most scoring functions should generate the same final hitlist post LPC scoring. This is highly advantageous, as limited budgets may not always allow access to a large number of these scoring functions.

As a validation, we have positively identified 1 micromolar (compound 2 = 1.35uM) and two novel nanomolar (compound 4 = 56.87nM, compound 5 = 53.25nM) ligands of ERα by virtual screening of 202054 compounds, of which only 7 were selected for biological testing. The compounds also exhibit low micromolar inhibition of MCF-7 proliferation and were also shown to be selective in targeting ERα over ERβ (eg. compound 5 = 17-fold selective). An optimized version of the 'hit' as a result of virtual library generation followed by re-docking, scoring and filtering as in Path 2 was also designed using novel methods.

The procedure is fully automated and access to a mid-sized 130 Intel Xeon 3.06GHz processor cluster [100] allows us to carry out vHTS via these methods in a short time. This procedure is currently being extended to carry out virtual screening to identify compounds selective for ERβ.

3.16 References

1. Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M., Molecular docking using surface complementarity. *Proteins* **1996,** 25, (1), 120-9.
2. Macromodel v6.5, Schrodinger Inc.: Portland, OR 97201. *(http://www.schrodinger.com/Products/macromodel. html).*
3. Sadowski J, *J. Chem. Inf. Comput. Sci* **1994,** 34, 1000.
4. Weininger, D., SMILES: A Chemical Language and Information System. *J. Chem. Inf. Comput* **1988,** 28, 31-36.
5. OMEGA 1.8, distributed by Openeye Scientific Software.
6. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L., The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998,** 95, (7), 927-37.
7. Flexidock, Tripos Inc. *http://www.tripos.com*.
8. InsightII, v2000. *www.accelrys.com*.
9. Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M., Automated analysis of interatomic contacts in proteins. *Bioinformatics* **1999,** 15, (4), 327-32.
10. Thomas, A. H., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** 17, (5-6), 490-519.
11. affinity, v2000. *www.accelrys.com*.
12. Mozziconacci, J. C. A., E.; Baurin, N.; Marot, C.; Morin-Allory, L., Preparation of a molecular database from a set of 2 million compounds for virtual screening applications: gathering, structural analysis and filtering. *9th Electronic Computational Chemistry Conference (ECCC9)* **2003**.
13. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002,** 45, (12), 2615-23.
14. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** 261, (3), 470-89.
15. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **2001,** 15, (5), 411-28.
16. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, (3), 727-48.
17. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **1999,** 42, (5), 791-804.
18. Bohm, H. J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **1994,** 8, (3), 243-56.
19. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **2002,** 16, (1), 11-26.
20. Bissantz, C.; Folkers, G.; Rognan, D., Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **2000,** 43, (25), 4759-67.
21. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T., Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol* **1995,** 2, (5), 317-24.
22. Stahl, M.; Rarey, M., Detailed analysis of scoring functions for virtual screening. *J Med Chem* **2001,** 44, (7), 1035-42.
23. Jain, A. N., Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **2003,** 46, (4), 499-511.
24. Bajorath, J., Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **2002,** 1, (11), 882-94.
25. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **2004,** 3, (11), 935-49.
26. Schneider, G.; Bohm, H. J., Virtual screening and fast automated docking methods. *Drug Discov Today* **2002,** 7, (1), 64-70.
27. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** 7, (20), 1047-55.

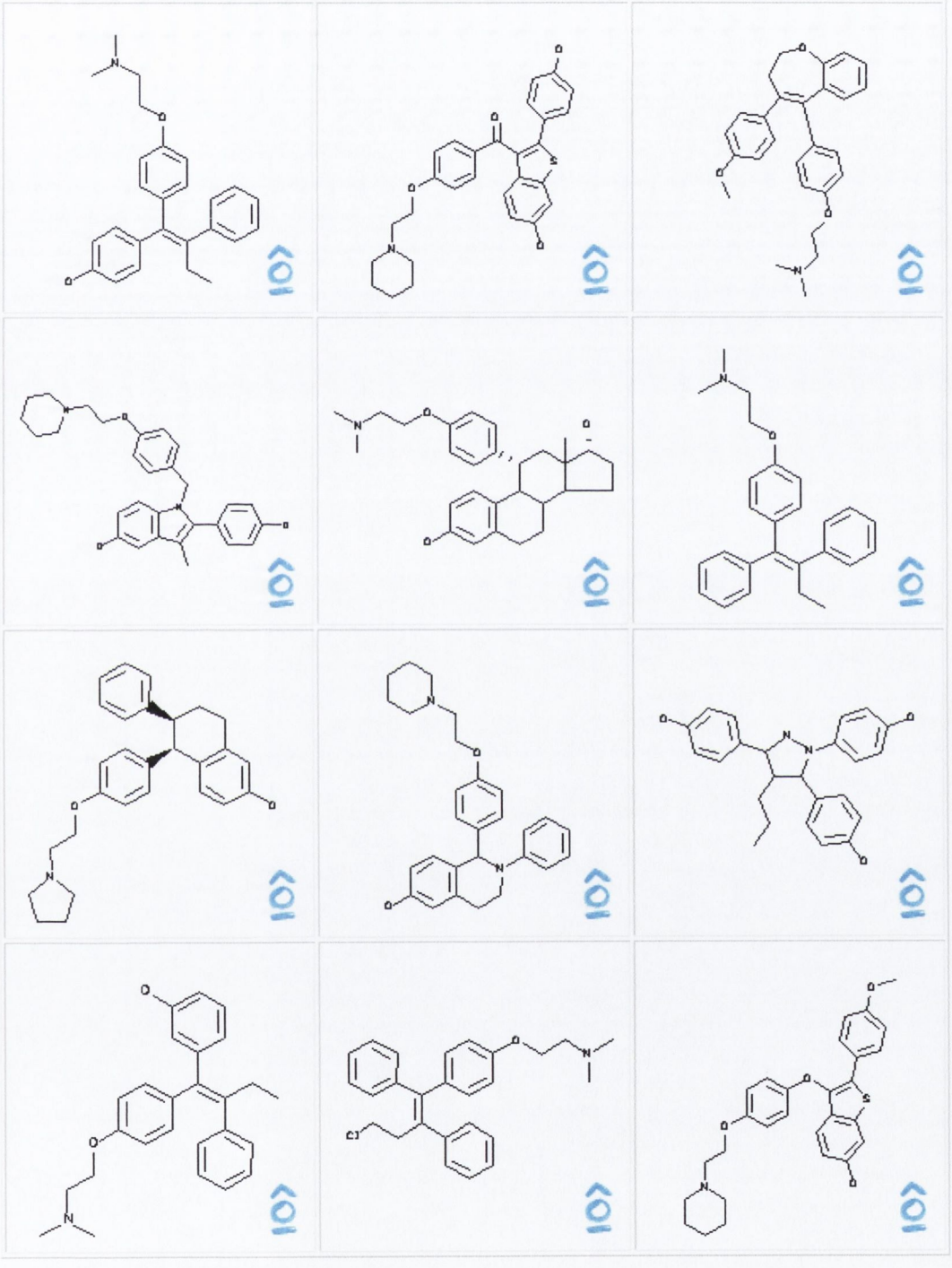
28. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D., Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004,** 57, (2), 225-42.
29. Perola, E.; Walters, W. P.; Charifson, P. S., A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004,** 56, (2), 235-49.
30. 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.
31. Warren, G. L., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **2005**.
32. Krovat, E. M.; Langer, T., Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J Chem Inf Comput Sci* **2004,** 44, (3), 1123-9.
33. 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.
34. Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R., Comparing protein-ligand docking programs is difficult. *Proteins* **2005,** 60, (3), 325-32.
35. Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F., Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J Chem Inf Comput Sci* **2004,** 44, (3), 871-81.
36. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **2004,** 47, (7), 1750-9.
37. Sobolev, V.; Moallem, T. M.; Wade, R. C.; Vriend, G.; Edelman, M., CASP2 molecular docking predictions with the LIGIN software. *Proteins* **1997,** Suppl 1, 210-4.
38. Sobolev, V.; Edelman, M., Modeling the quinone-B binding site of the photosystem-II reaction center using notions of complementarity and contact-surface between atoms. *Proteins* **1995,** 21, (3), 214-25.
39. Sobolev, V.; Niztayev, A.; Pick, U.; Avni, A.; Edelman, M., A proteomic approach to resolving the binding sites for tentoxin in plastid CF1-ATPase. *Proceedings of the 12th International Congress on Photosynthesis*
2001.
40. 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.
41. Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L., Docking: successes and challenges. *Curr Pharm Des* **2005,** 11, (3), 323-33.
42. Borgelt, C.; Meinl, T.; Berthold, M. R., MoSS: A Program for Molecular Substructure Mining. *Workshop Open Software for Data Mining (OSDM'05, Chicago, IL)* **2004**.
43. Borgelt, C.; Hofer, H.; Berthold, M. R., Finding Discriminative Molecular Fragments. *German Conference on Artificial Intelligence, Hamburg, Germany 2003* **2003**.
44. Borgelt, C.; Berthold, M. R., Mining Molecular Fragments: Finding Relevant Substructures of Molecules. *IEEE International Conference on Data Mining (ICDM 2002, Maebashi, Japan)* **2002**, 51-58.
45. Krier, M.; Araujo-Junior, J. X.; Schmitt, M.; Duranton, J.; Justiano-Basaran, H.; Lugnier, C.; Bourguignon, J. J.; Rognan, D., Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *J Med Chem* **2005,** 48, (11), 3816-22.
46. Schuller, A.; Schneider, G.; Byvatov, E., SMILIB: Rapid assembly of combinatorial libraries in SMILES notation. *QSAR Comb Sci* **2003,** 22, 719-721.
47. Shiau, A. K.; Barstad, D.; Radek, J. T.; Meyers, M. J.; Nettles, K. W.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A.; Agard, D. A.; Greene, G. L., Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nat Struct Biol* **2002,** 9, (5), 359-64.
48. Razandi, M.; Pedram, A.; Greene, G. L.; Levin, E. R., Cell membrane and nuclear estrogen receptors (ERs) originate from a single transcript: studies of ERalpha and ERbeta expressed in Chinese hamster ovary cells. *Mol Endocrinol* **1999,** 13, (2), 307-19.
49. Pike, A. C.; Brzozowski, A. M.; Walton, J.; Hubbard, R. E.; Bonn, T.; Gustafsson, J. A.; Carlquist, M., Structural aspects of agonism and antagonism in the oestrogen receptor. *Biochem Soc Trans* **2000,** 28, (4), 396-400.

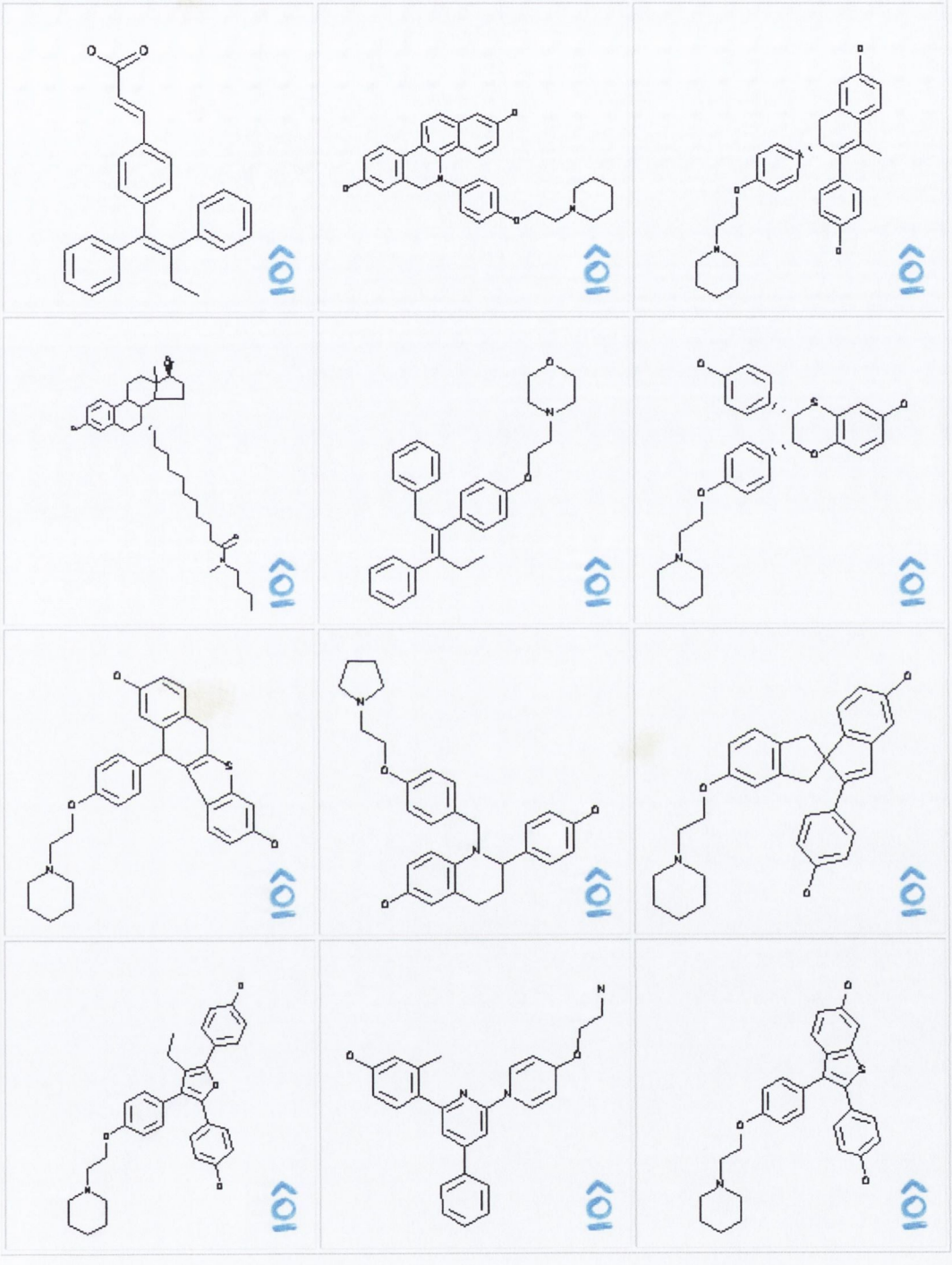
50. Manas, E. S.; Unwalla, R. J.; Xu, Z. B.; Malamas, M. S.; Miller, C. P.; Harris, H. A.; Hsiao, C.; Akopian, T.; Hum, W. T.; Malakian, K.; Wolfrom, S.; Bapat, A.; Bhat, R. A.; Stahl, M. L.; Somers, W. S.; Alvarez, J. C., Structure-based design of estrogen receptor-beta selective ligands. *J Am Chem Soc* **2004,** 126, (46), 15106-19.
51. Kim, S.; Wu, J. Y.; Birzin, E. T.; Frisch, K.; Chan, W.; Pai, L. Y.; Yang, Y. T.; Mosley, R. T.; Fitzgerald, P. M.; Sharma, N.; Dahllund, J.; Thorsell, A. G.; DiNinno, F.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. II. Discovery of benzoxathiins as potent, selective estrogen receptor alpha modulators. *J Med Chem* **2004,** 47, (9), 2171-5.
52. Bourguet, W.; Germain, P.; Gronemeyer, H., Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol Sci* **2000,** 21, (10), 381-8.
53. Folkertsma, S.; van Noort, P. I.; Brandt, R. F.; Bettler, E.; Vriend, G.; de Vlieg, J., The nuclear receptor ligand-binding domain: a family-based structure analysis. *Curr Med Chem* **2005,** 12, (9), 1001-16.
54. Imamov, O.; Shim, G. J.; Warner, M.; Gustafsson, J. A., Estrogen Receptor beta in Health and Disease. *Biol Reprod* **2005**, 73, (5), 866-71.
55. Koehler, K. F.; Helguero, L. A.; Haldosen, L. A.; Warner, M.; Gustafsson, J. A., Reflections on the discovery and significance of estrogen receptor beta. *Endocr Rev* **2005,** 26, (3), 465-78.
56. Levenson, A. S.; Jordan, V. C., The key to the antiestrogenic mechanism of raloxifene is amino acid 351 (aspartate) in the estrogen receptor. *Cancer Res* **1998,** 58, (9), 1872-5.
57. Shao, D.; Berrodin, T. J.; Manas, E.; Hauze, D.; Powers, R.; Bapat, A.; Gonder, D.; Winneker, R. C.; Frail, D. E., Identification of novel estrogen receptor alpha antagonists. *J Steroid Biochem Mol Biol* **2004,** 88, (4-5), 351-60.
58. Waszkowycz, B.; Perkins, T. D.; Sykes, R. A.; Li, J., Large-scale virtual screening for discovering leads in the post-genomic era. *IBM Systems Journal* **2001,** 40, (2), 360-376.
59. Zhao, L.; Brinton, R. D., Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases. *J Med Chem* **2005,** 48, (10), 3463-6.
60. 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.
61. FRED (version 2.0.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com)*.
62. OMEGA 1.8.1, distributed by Openeye Scientific Software.
63. OpenBabel, *http://openbabel.sourceforge.net/*.
64. MYSQL, *(URL:http://www.mysql.com)*.
65. Meegan, M. J.; Hughes, R. B.; Lloyd, D. G.; Williams, D. C.; Zisterer, D. M., Flexible estrogen receptor modulators: design, synthesis, and antagonistic effects in human MCF-7 breast cancer cells. *J Med Chem* **2001,** 44, (7), 1072-84.
66. FILTER, distributed by Openeye Scientific Software.
67. Knox, A. J. S., Meegan M.J., Lloyd D.G, Estrogen Receptors: Molecular interactions, virtual screening and future prospects. *Current Topics in Medicinal Chemistry* **2005**, (In Press).
68. Derwent World Drug Index, (URL: http://thomsonderwent.com/products/lr/wdi).
69. Knox, A. J. S., Meegan M.J., Carta G., Lloyd D.G., Considerations in compound database preparation - 'hidden' impact on virtual screening results. *J Chem Inf Model* **2005,** accepted for publication.
70. Harper, M. J.; Walpole, A. L., A new derivative of triphenylethylene: effect on implantation and mode of action in rats. *J Reprod Fertil* **1967,** 13, (1), 101-19.
71. Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, P.; Fournier, B.; Halleux, C.; Kallen, J.; Keller, H.; Schlaeppi, J. M.; Stark, W., Estrogen receptor modulators: identification and structure-activity relationships of potent ERalpha-selective tetrahydroisoquinoline ligands. *J Med Chem* **2003,** 46, (14), 2945-57.
72. Jones, C. D.; Jevnikar, M. G.; Pike, A. J.; Peters, M. K.; Black, L. J.; Thompson, A. R.; Falcone, J. F.; Clemens, J. A., Antiestrogens. 2. Structure-activity studies in a series of 3-aroyl-2-arylbenzo[b]thiophene derivatives leading to [6-hydroxy-2-(4-hydroxyphenyl)benzo[b]thien-3-yl] [4-[2-(1-piperidinyl)ethoxy]-phenyl]methanone hydrochloride (LY156758), a remarkably effective estrogen antagonist with only minimal intrinsic estrogenicity. *J Med Chem* **1984,** 27, (8), 1057-66.

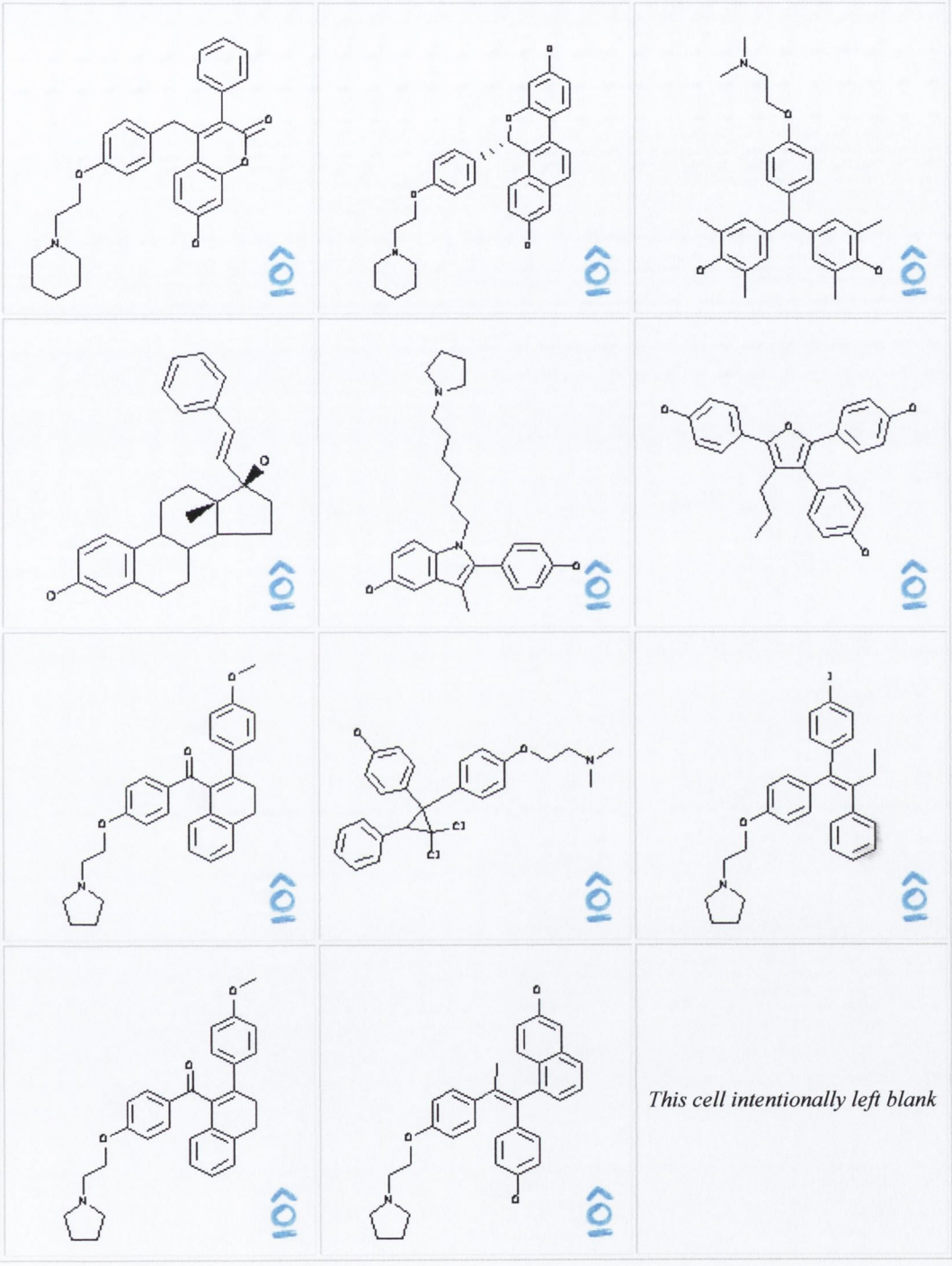
73. Lloyd, D. G.; Hughes, R. B.; Zisterer, D. M.; Williams, D. C.; Fattorusso, C.; Catalanotti, B.; Campiani, G.; Meegan, M. J., Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J Med Chem* **2004,** 47, (23), 5612-5.
74. Greenberger, L. M.; Annable, T.; Collins, K. I.; Komm, B. S.; Lyttle, C. R.; Miller, C. P.; Satyaswaroop, P. G.; Zhang, Y.; Frost, P., A new antiestrogen, 2-(4-hydroxy-phenyl)-3-methyl-1-[4-(2-piperidin-1-yl-ethoxy)-benzyl]-1H-in dol-5-ol hydrochloride (ERA-923), inhibits the growth of tamoxifen-sensitive and -resistant tumors and is devoid of uterotropic effects in mice and rats. *Clin Cancer Res* **2001,** 7, (10), 3166-77.
75. Gottardis, M. M.; Jiang, S. Y.; Jeng, M. H.; Jordan, V. C., Inhibition of tamoxifen-stimulated growth of an MCF-7 tumor variant in athymic mice by novel steroidal antiestrogens. *Cancer Res* **1989,** 49, (15), 4090-3.
76. Jin, L.; Borras, M.; Lacroix, M.; Legros, N.; Leclercq, G., Antiestrogenic activity of two 11 beta-estradiol derivatives on MCF-7 breast cancer cells. *Steroids* **1995,** 60, (8), 512-8.
77. Yang, X.; Reinhold, A. R.; Rosati, R. L.; Liu, K. K., Enzyme-catalyzed asymmetric deacylation for the preparation of lasofoxifene (CP-336156), a selective estrogen receptor modulator. *Org Lett* **2000,** 2, (25), 4025-7.
78. Stauffer, S. R.; Coletta, C. J.; Tedesco, R.; Nishiguchi, G.; Carlson, K.; Sun, J.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A., Pyrazole ligands: structure-affinity/activity relationships and estrogen receptor-alpha-selective agonists. *J Med Chem* **2000,** 43, (26), 4934-47.
79. Ke, H. Z.; Simmons, H. A.; Pirie, C. M.; Crawford, D. T.; Thompson, D. D., Droloxifene, a new estrogen antagonist/agonist, prevents bone loss in ovariectomized rats. *Endocrinology* **1995,** 136, (6), 2435-41.
80. Gao, H. D.; Sun, J. Z.; Bi, D. S.; Ma, R., [Status of estrogen receptor affects the drug sensitivity of drug-resistant MCF-7/Adr human breast cancer cells to droloxifene and Adriamycin]. *Ai Zheng* **2003,** 22, (4), 376-9.
81. Gauthier, S.; Caron, B.; Cloutier, J.; Dory, Y. L.; Favre, A.; Larouche, D.; Mailhot, J.; Ouellet, C.; Schwerdtfeger, A.; Leblanc, G.; Martel, C.; Simard, J.; Merand, Y.; Belanger, A.; Labrie, C.; Labrie, F., (S)-(+)-4-[7-(2,2-dimethyl-1-oxopropoxy)-4-methyl-2-[4-[2-(1-piperidinyl)- ethoxy]phenyl]-2H-1-benzopyran-3-yl]-phenyl 2,2-dimethylpropanoate (EM-800): a highly potent, specific, and orally active nonsteroidal antiestrogen. *J Med Chem* **1997,** 40, (14), 2117-22.
82. Simard, J.; Labrie, C.; Belanger, A.; Gauthier, S.; Singh, S. M.; Merand, Y.; Labrie, F., Characterization of the effects of the novel non-steroidal antiestrogen EM-800 on basal and estrogen-induced proliferation of T-47D, ZR-75-1 and MCF-7 human breast cancer cells in vitro. *Int J Cancer* **1997,** 73, (1), 104-12.
83. Lubczyk, V.; Bachmann, H.; Gust, R., Antiestrogenically active 1,1,2-tris(4-hydroxyphenyl)alkenes without basic side chain: synthesis and biological activity. *J Med Chem* **2003,** 46, (8), 1484-91.
84. Tan, Q.; Birzin, E. T.; Chan, W.; Yang, Y. T.; Pai, L. Y.; Hayes, E. C.; DaSilva, C. A.; DiNinno, F.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. Part 6: Synthesis and binding affinity of dihydrobenzodithiins. *Bioorg Med Chem Lett* **2004,** 14, (14), 3753-5.
85. Wallace, O. B.; Lauwers, K. S.; Jones, S. A.; Dodge, J. A., Tetrahydroquinoline-based selective estrogen receptor modulators (SERMs). *Bioorg Med Chem Lett* **2003,** 13, (11), 1907-10.
86. Blizzard, T. A.; Morgan, J. D., 2nd; Mosley, R. T.; Birzin, E. T.; Frisch, K.; Rohrer, S. P.; Hammond, M. L., 2-Phenylspiroindenes: a novel class of selective estrogen receptor modulators (SERMs). *Bioorg Med Chem Lett* **2003,** 13, (3), 479-83.
87. Weatherman, R. V.; Carroll, D. C.; Scanlan, T. S., Activity of a tamoxifen-raloxifene hybrid ligand for estrogen receptors at an AP-1 site. *Bioorg Med Chem Lett* **2001,** 11, (24), 3129-31.
88. Brady, H.; Doubleday, M.; Gayo-Fung, L. M.; Hickman, M.; Khammungkhune, S.; Kois, A.; Lipps, S.; Pierce, S.; Richard, N.; Shevlin, G.; Sutherland, M. K.; Anderson, D. W.; Bhagwat, S. S.; Stein, B., Differential response of estrogen receptors alpha and beta to SP500263, a novel potent selective estrogen receptor modulator. *Mol Pharmacol* **2002,** 61, (3), 562-8.
89. McKie, J. A.; Bhagwat, S. S.; Brady, H.; Doubleday, M.; Gayo, L.; Hickman, M.; Jalluri, R. K.; Khammungkhune, S.; Kois, A.; Mortensen, D.; Richard, N.; Sapienza, J.; Shevlin, G.; Stein, B.; Sutherland, M., Lead identification of a potent benzopyranone selective estrogen receptor modulator. *Bioorg Med Chem Lett* **2004,** 14, (13), 3407-10.

90. Grese, T. A.; Pennington, L. D.; Sluka, J. P.; Adrian, M. D.; Cole, H. W.; Fuson, T. R.; Magee, D. E.; Phillips, D. L.; Rowley, E. R.; Shetler, P. K.; Short, L. L.; Venugopalan, M.; Yang, N. N.; Sato, M.; Glasebrook, A. L.; Bryant, H. U., Synthesis and pharmacology of conformationally restricted raloxifene analogues: highly potent selective estrogen receptor modulators. *J Med Chem* **1998,** 41, (8), 1272-83.
91. Sharma, A. P.; Saeed, A.; Durani, S.; Kapil, R. S., Structure-activity relationship of antiestrogens. Phenolic analogues of 2,3-diaryl-2H-1-benzopyrans. *J Med Chem* **1990,** 33, (12), 3222-9.
92. Neubauer, B. L.; McNulty, A. M.; Chedid, M.; Chen, K.; Goode, R. L.; Johnson, M. A.; Jones, C. D.; Krishnan, V.; Lynch, R.; Osborne, H. E.; Graff, J. R., The selective estrogen receptor modulator trioxifene (LY133314) inhibits metastasis and extends survival in the PAIII rat prostatic carcinoma model. *Cancer Res* **2003,** 63, (18), 6056-62.
93. Glide 3.5, developed and distributed by Schrodinger. *http://www.schrodinger.com/Products/glide.html*.
94. Pro_Leads, developed and distributed by Protherics Inc. *http://www.protherics.com/*.
95. Surflex, developed and distributed by Jain Lab. *http://jainlab.ucsf.edu*.
96. Fink, B. E.; Mortensen, D. S.; Stauffer, S. R.; Aron, Z. D.; Katzenellenbogen, J. A., Novel structural templates for estrogen-receptor ligands and prospects for combinatorial synthesis of estrogens. *Chem Biol* **1999,** 6, (4), 205-19.
97. Stauffer, S. R.; Huang, Y.; Coletta, C. J.; Tedesco, R.; Katzenellenbogen, J. A., Estrogen pyrazoles: defining the pyrazole core structure and the orientation of substituents in the ligand binding pocket of the estrogen receptor. *Bioorg Med Chem* **2001,** 9, (1), 141-50.
98. Elagamey, A. G. A.; El-Taweel, F. M. A. A., Nitriles in heterocyclic synthesis: Synthesis of condensed pyrans. *Indian J. Chem. Sect. B* **1990,** 29, (9), 885-886.
99. Chernobrovin, N. I. K., Yu. V.; Bobrovskaya, O. V.; Syropyatov, B. Ya., SYNTHESIS AND BIOLOGICAL ACTIVITY OF 1-ACETYL-2,3-DIARYL-1,2,3,4-TETRAHYDROQUINAZOLINE-4-ONES. *Khim. Farm. Zh.* **1991,** 25, (5), 37-39.
100. IITAC, High Performance Computing Facility. *http://www.iitac.tchpc.tcd.ie/*.
101. Pozo-Guisado, E.; Merino, J. M.; Mulero-Navarro, S.; Lorenzo-Benayas, M. J.; Centeno, F.; Alvarez-Barrientos, A.; Salguero, P. M., Resveratrol-induced apoptosis in MCF-7 human breast cancer cells involves a caspase-independent mechanism with downregulation of Bcl-2 and NF-kappaB. *Int J Cancer* **2005,** 115, (1), 74-84.

Appendix A – 35 antagonists

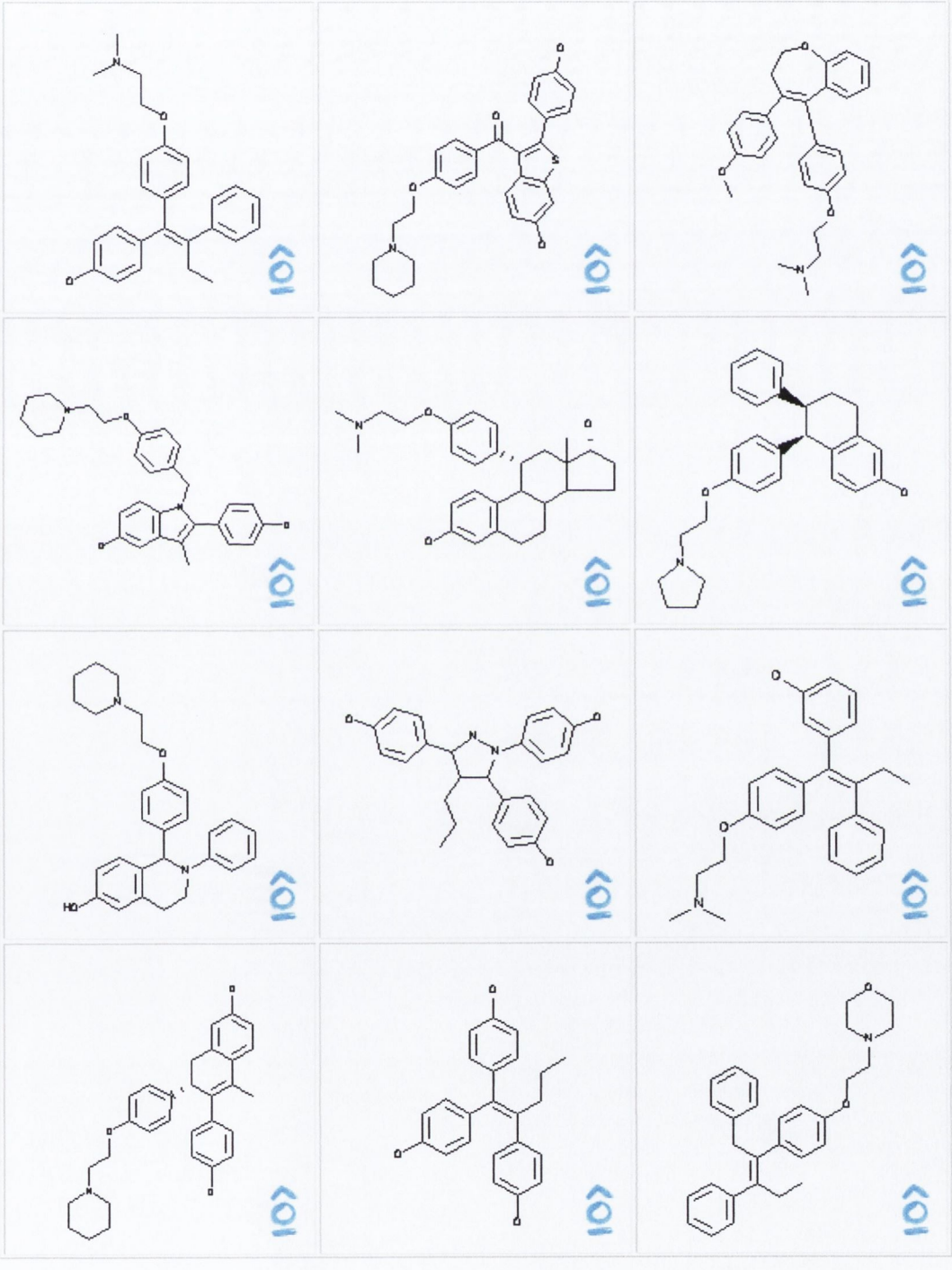


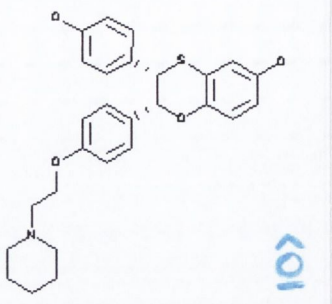
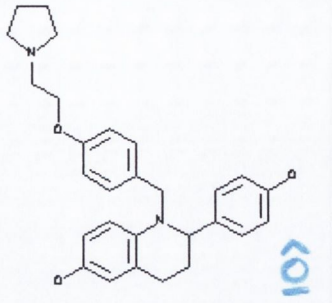
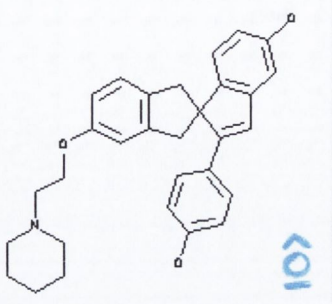
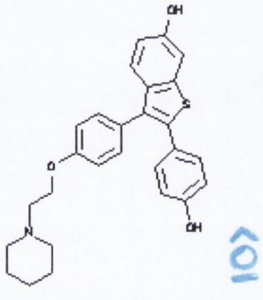
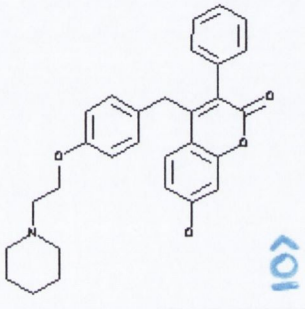
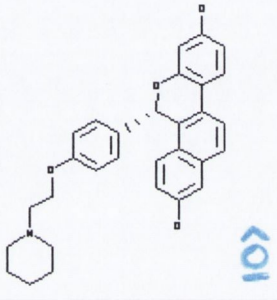
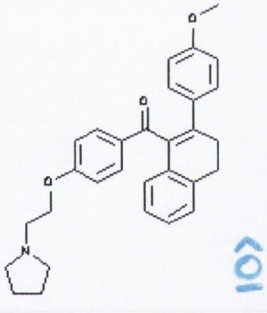




This cell intentionally left blank

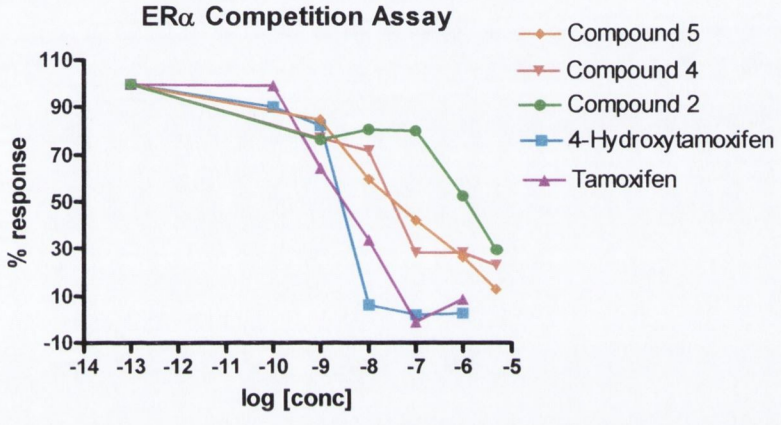
Appendix B – 19 drug-like actives



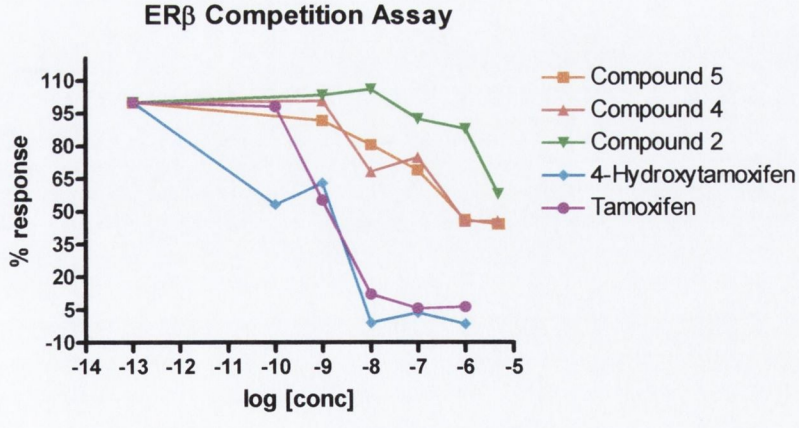
		
		
	*This cell intentionally left blank.*	*This cell intentionally left blank.*

Appendix C - Plots

Competition assay plots for recombinant ERα and ERβ and a fluorescent estrogen. Displacement of the fluorescent estrogen with increasing concentrations of competitor that results in a half maximum shift polarization equals the IC_{50} of the competitor. This is a measure of the relative binding affinity of each competitor for both receptors.

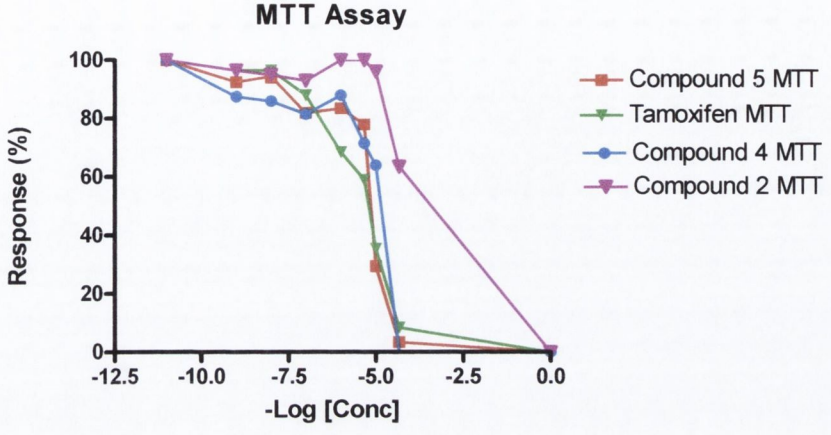


	4-Hydroxytamoxifen	Tamoxifen	Compound 4	Compound 5	Compound 2
Std. Error					
LOGEC50	0.08723	0.1366	0.2702	0.09496	0.3260
HILLSLOPE	0.3752	0.1583	0.07952	0.03307	0.09732
95% Confidence Intervals					
LOGEC50	-8.877 to -8.393	-8.898 to -8.139	-8.107 to -6.607	-7.613 to -7.086	-6.897 to -5.087
HILLSLOPE	-2.842 to -0.7589	-1.197 to -0.3186	-0.5583 to -0.1168	-0.4711 to -0.2875	-0.5963 to -0.05591
EC50	1.327e-009 to 4.047e-009	1.264e-009 to 7.253e-009	7.822e-009 to 2.474e-007	2.438e-008 to 8.207e-008	1.268e-007 to 8.184e-006



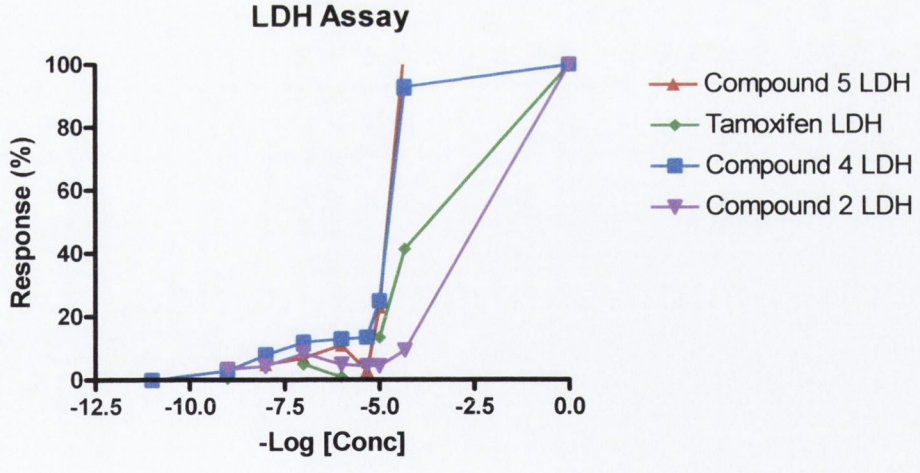
	Compound 5	Compound 4	Compound 2	4-Hydroxytamoxifen	Tamoxifen
Std. Error					
LOGEC50	0.1264	0.3266	0.1044	0.3704	0.06500
HILLSLOPE	0.03241	0.08111	0.2192	0.2217	0.1560
95% Confidence Intervals					
LOGEC50	-6.246 to -5.596	-6.719 to -5.039	-5.415 to -4.878	-10.34 to -8.433	-9.045 to -8.710
HILLSLOPE	-0.3849 to -0.2183	-0.5055 to -0.08846	-1.469 to -0.3425	-1.073 to 0.06703	-1.431 to -0.6290
EC50	5.681e-007 to 2.536e-006	1.911e-007 to 9.138e-006	3.848e-006 to 1.324e-005	4.598e-011 to 3.693e-009	9.021e-010 to 1.948e-009

Plot of antiproliferative effects of active compounds with Tamoxifen used as a control.



Std. Error	Compound 5 MTT	Tamoxifen MTT	Compound 4 MTT	Compound 2 MTT
LOGEC50	0.03467	0.07716	0.0679	0.06759
HILLSLOPE	0.6551	0.08911	0.4085	0.5689
95% Confidence Intervals				
LOGEC50	-5.194 to -5.050	-5.507 to -5.190	-5.082 to -4.802	-4.299 to -4.018
HILLSLOPE	-4.284 to -1.574	-0.8475 to -0.4804	-2.446 to -0.7629	-2.847 to -0.4876
EC50	6.402e-006 to 8.908e-006	3.109e-006 to 6.464e-006	8.286e-006 to 1.578e-005	5.027e-005 to 9.586e-005

Plot of cytotoxic effects of active compounds with Tamoxifen used as a control.



	Compound 5 LDH	Tamoxifen LDH	Compound 4 LDH	Compound 2 LDH
LOGEC50	0.2946	0.09352	0.1174	0.7312
HILLSLOPE	14.48	0.4947	0.7299	0.2148
95% Confidence Intervals				
LOGEC50	-5.499 to -4.282	-4.396 to -4.011	-5.037 to -4.552	-4.228 to -1.210
HILLSLOPE	-25.21 to 34.57	0.3705 to 2.409	0.4421 to 3.455	0.1229 to 1.010
EC50	3.173e-006 to 5.220e-005	4.017e-005 to 9.755e-005	9.191e-006 to 2.806e-005	5.909e-005 to 0.06162

Appendix D

Code for InsightII docking and automation of LIGIN

Included in this section are the code for predock.log, godock.log and dock.log needed to automate the docking process with InsightII v2000. Secondly, two Perl scripts (Insert_Confs.pl and Docking.pl) written to automate the docking utilizing LIGIN and post-filtering phases with LPC are given. Captions are provided beside the code to describe each section.

Predock.log

```
m:Get Molecule PDB User 3ERT.pdb P_3ERT Heteroatom -Keep_Alternates -Use_Segids -
Keep_All_Frames -Reference_Object
m:Color Molecule Atoms P_3ERT:*H Specified Specification 255,255,0
m:Biopolymer
m:Unmerge P_3ERT:*H:O WAT
r:Unmerge P_3ERT:600H:C22 TAMOXIFEN
m:Rename Object WAT WATER
m:Hydrogens P_3ERT set_pH 7 -Lone_Pairs Charged -Assign_IUPAC_Names
m:Browse Molecule
m:Delete Object WATER
m:Color Molecule Atoms P_3ERT Specified By_Atom
m:Replace Residue P_3ERT:A529 GLY L
r:Replace Residue P_3ERT:A335 ARG L
r:Replace Residue P_3ERT:A481 LYS L
r:Replace Residue P_3ERT:A529 LYS L
r:Replace Residue P_3ERT:A531 LYS L
m:clip auto
c:Blank Object On P_3ERT
m:Get Molecule PDB User substrate.pdb SUBSTRATE Heteroatom -Keep_Alternates -Use_Segids -
Keep_All_Frames -Reference_Object
c:Selection Subset "" Sel_End
c:Selection Subset 4244 Sel_End
c:Selection Subset 4244 Sel_End
c:Selection Subset "" Sel_End
m:clip auto
```

Godock.log

```
c:delete object TAMOXIFEN
m:Associate Assembly P_3ERT SUBSTRATE COMPLEX
m:Select Forcefield Clear_Potentials -Clear_Charges cvff.frc -Make_Copy
m:Potentials Forcefield Fix -Print_Potentials Fix -Print_Part_Chargs Fix -Print_Form_Chargs P_3ERT
c:Blank Object Off P_3ERT
m:Interface Subset BIND SUBSTRATE P_3ERT Static Monomer/Residue 7 Color_Subset 255,0,0
m:Display Molecule Only Atoms Specified BIND
m:Color Molecule Atoms p_3ert:a353 Specified By_Atom
r:Color Molecule Atoms p_3ert:a351 Specified By_Atom
r:Color Molecule Atoms p_3ert:a394 Specified By_Atom
```

r:Color Molecule Atoms p_3ert:a524 Specified By_Atom
r:Label Molecule Properties Monomer/Residue On P_3ERT Type_and_number
c:Selection Subset 1496 Sel_End
m:Potentials Forcefield Fix -Print_Potentials Fix -Print_Part_Chargs Fix -Print_Form_Chargs COMPLEX
m:Docking
m:Setup_System COMPLEX P_3ERT SUBSTRATE BIND -solvation_grid -use_hbond -apply_tethering -confine_ligand

Dock.log

m:SA_Docking job1 De_Novo 5 -Acceptance_Filter -Command_file_only Cell_multipole dist_dep_diel 1 1 1 1 1 Energy_range -Flexible_Ligand 10 1 180 1e+06 1 5000 -Apply_SA
m:Color Molecule Atoms SUBSTRATE Specified By_Atom

Insert_Confs.pl

```
#!/usr/bin/perl

#use File::stat;
#use strict;
use Getopt::Std;
use DBI;

getopts('b:i:j:p:o', \%opts);
$inputfile = $opts{'i'};
$jobno = $opts{'j'};
$protein_id = $opts{'p'};
$babel = "/usr/bin/babel";
$workdir = "/tmp/setup.$$";

usage() unless $inputfile;
usage() unless $protein_id;
usage() unless $jobno;

system("/bin/rm -rf $workdir");
mkdir($workdir);

open(INFILE, $inputfile) or die "Can't open $inputfile: $!\n";

$i=0;
$contents="";
while(my $line = <INFILE>) {
      if($contents =~ /^$/) {
            ($sdfsmile, $junk) = split / /,$line, 2;
      }
      $contents .= $line;
      if( $line =~ /\$\$\$\$/ ) {
            # $contents now has a complete SDF entry.

            # Create input file
            $input = create_input($workdir, $contents, $i);

            # Now create LIG file
            $lig = create_lig($workdir, $i);

            # Insert into database
            create_dbentry($jobno, $protein_id, $input, $lig);
```

```
            $contents = "";
            $i=$i+1;
            if($i%10 == 0) {
                  print ".";
            }
            system("/bin/rm -rf $workdir");
            mkdir($workdir);
      }
}
print "\n";
system("/bin/rm -rf $workdir");


# Sub routines to make the code easier to manage!

sub create_dbentry() {

      my ($jobno, $protein_id, $input, $lig) = @_;
      my $dsn = 'DBI:mysql:HRB2:localhost';
      my $db_user_name = 'hrb';
      my $db_password = 'Seafiy7x';
      my ($dbh, $db_query, $db_sel);

      $dbh = DBI->connect($dsn, $db_user_name, $db_password);
      $db_query = "insert into `parallel` (`jobid` , `protein` , `lig`
, `input`) "
                  . "values ('$jobno', '$protein_id', '$lig',
'$input')";

      # print "$db_query\n";
      $dbh->do($db_query);
}

sub create_lig() {
      my ($workdir, $i) = @_;
      my $basename = sprintf "$workdir/%6.6d", $i;
      my $nohfile = $basename . "-without-h.pdb";
      my $contents = "";

      # sdf file is already there. Make PDB with no hydrogens
      push(@cmd, "$babel", "-d", "-isdf", "$basename.sdf", "-opdb",
"$nohfile");
      # print join (' ', @cmd) . "\n";
      system join (' ', @cmd);
      undef(@cmd);

      # Slurp in the contents of $nohfile
      open NOHFILE, "$nohfile" or die "create_lig: Can't open $nohfile:
$!\n";
      while(<NOHFILE>) {
            $contents .= $_;
      }
      close NOHFILE;

      return $contents;
```

```
}

sub create_input() {
      my $gen_input = "/home/trhpc/dfrost/bin/gen_input";
      my ($workdir, $sdfcontents, $i) = @_;
      my $basename = sprintf "$workdir/%6.6d", $i;
      my $hfile = $basename . "-with-h.pdb";
      my $contents = "";

      # Dump sdf contents to a file
      open OUTFILE, ">$basename.sdf" or die "create_input: Can't open
$basename: $!\n";
      print OUTFILE $sdfcontents;
      close OUTFILE;

      # Get a pdb with hydrogens from the sdf
      push(@cmd, "$babel", "-h", "-isdf", "$basename.sdf", "-opdb",
"$hfile");
      # print join (' ', @cmd) . "\n";
      system join (' ', @cmd);
      undef(@cmd);

      # Now use the hfile to generate the input file
      push(@cmd, "$gen_input", "$hfile", "$basename.input");
      # print join (' ', @cmd) . "\n";
      system join (' ', @cmd);
      undef(@cmd);

      # Slurp in the contents of $nohfile
      open INPUTFILE, "$basename.input" or die "create_input: Can't
open $basename.input: $!\n";
      while(<INPUTFILE>) {
            $contents .= $_;
      }

      # Return the contents of the INPUT file
      return $contents;
}

sub usage() {
      print "Incorrect Input\n";
      print "$0 -p protein_id -i sdf_file -j jobno\n";

      exit;
}

# vim:ts=4:sw=4
```

Docking.pl

```
#!/usr/bin/perl

#use strict;
```

```
use Digest::MD5 qw(md5_hex);
use Getopt::Std;
use File::Copy;
use DBI;
use Cwd;

# Database related variables
my $dsn = 'DBI:mysql:HRB2:134.226.112.32';
my $db_user_name = 'hrb';
my $db_password = 'Seafiy7x';

my $basedir = "/scratch/ligin.$$";
#$basedir = "/scratch/ligin";
my $centre_executable = "/home/trhpc/dfrost/bin/centre";
my $translate_executable = "/home/trhpc/dfrost/bin/translate";
my $ligin_executable = "/home/trhpc/dfrost/bin/ligin";
my $lpc_executable = "/home/trhpc/dfrost/bin/lpcEx";
my $hetfile = $basedir . "/hetfile";
my $protfile = $basedir . "/PROT";
my $infile = $basedir . "/INPUT";
my $templig = $basedir . "/TEMPLIG";
my $ligfile = $basedir . "/LIG";

my $linecount;
my $trash = 1;


my ($tx, $ty, $tz);

if($#ARGV < 1) {
      &usage();
}

# -c conf_id -i iterations -t threshold -v -n
getopts('c:i:t:vnr', \%opts);
$conf_id = $opts{'c'};
$iterations = $opts{'i'};
$threshold = $opts{'t'};
$verbose = 1 if $opts{'v'};
$cleanup = 1 if $opts{'n'};
$trash=0 if $opts{'r'};
&usage unless $conf_id;

# $verbose = 1;

$ligin_executable = $ligin_executable . " 2>&1 >/dev/null" unless
$verbose;
$iterations = 50 unless $iterations;
$threshold = 0.8 unless $threshold;
$threshold = 0.20;

mkdir $basedir || die "Can't make dir $basedir: $!\n";

$dbh = DBI->connect($dsn, $db_user_name, $db_password);

# Get the data for this job
$db_query = "select jobid, lig, input, protein, " .
```

```
                    "status from parallel where conf_id = '$conf_id'";
$db_sel = $dbh->prepare($db_query);
$db_sel->execute();
($jobid, $conformation, $inputfile, $protein_id, $status) = $db_sel-
>fetchrow_array();
printf("From parallel: %.70s %.70s\n", $lig, $input) if $verbose;
$db_sel->finish();

if($status == 2) {
      print "Conf $conf_id already examined\n";
      exit;
}


$db_query = "select hetatm, data from proteins where protein_id =
$protein_id";
print "$db_query\n" if $verbose;
$db_sel = $dbh->prepare($db_query);
$db_sel->execute();
($hetatm, $protdata) = $db_sel->fetchrow_array();
printf("From proteins: %s %.70s\n", $hetatm, $protdata) if $verbose;
$db_sel->finish();

# In case there is no hetatm
$tx = $ty = $tz = 0;

if($hetatm == 1) {
      print "Extracting HETATM and determining centre\n" if $verbose;
      open OUTFILE, ">$hetfile" || die "Can't open $hetfile for
writing: $!\n";
      print OUTFILE $protdata;
      close OUTFILE;
      open PIPEFILE, "$centre_executable $hetfile|" || die "Can't run
$centre_executable: $!\n";
      if( !($tx = <PIPEFILE>)) {
            print STDERR "Error reading from pipe\n";
            die;
      }
      # close PIPEFILE;
      ($tx, $ty, $tz) = split (/\s+/, $tx);
      print "Translating by $tx $ty $tz\n" if $verbose;

      ($receptor, $het) = split (/HETATM/, $protdata, 2);
}

print "Saving PROT file: $protfile\n" if $verbose;
open OUTFILE, ">$protfile" || die "Can't open $ligfile for writing:
$!\n";
print OUTFILE $receptor . "END\n";
close OUTFILE;

print "Saving INPUT file: $infile\n" if $verbose;
open OUTFILE, ">$infile" || die "Can't open $inputfile for writing:
$!\n";
print OUTFILE "PROT\nLIG\n".$iterations."\n0\n0\n0\n";
print OUTFILE $inputfile;
```

```
print OUTFILE "    5.000    5.000    5.000  !!  DIMENSIONS OF BOX OF
RANDOM\n";
print OUTFILE "                              !!  STARTING POINTS FOR
SEARCHING\n";
close OUTFILE;


print "Saving TEMPLIG: $templig\n", if $verbose;
open OUTFILE, ">$templig" || die "Can't open $templig for writing:
$!\n";
print OUTFILE $conformation;
close OUTFILE;

$cmd = join( ' ', $translate_executable, $tx, $ty, $tz, $templig,
$ligfile);
print "Translation Executing: $cmd\n" if $verbose;
system($cmd);
$owd = cwd();
chdir($basedir);
print "Running LIGIN: $ligin_executable\n" if $verbose;
system($ligin_executable);


open INFILE, "PROT" || die "Cannot open $workdir/PROT: $!\n";
while(<INFILE>) {
      if (/^TER/) {
            ($junk, $lastid, $junk) = split( /\s+/, $_, 3 );
      }
}
print "Last id = $lastid\n" if $verbose;
$lastid = $lastid-1;

seek(INFILE, 0, 0);
$protein = "";
while(<INFILE>) {
      last if /^END/;
      $protein .= $_;
}
close INFILE;
chomp($protein);

opendir DIRHANDLE, $basedir or die "Can't open $basedir for reading:
$!\n";
@filenames = grep /CR\d+/, readdir DIRHANDLE;
closedir DIRHANDLE;
print "@filenames\n" if $verbose;

#
# Run LPC on each CR file that I've made
#
foreach $crfile (@filenames) {
      mkdir("lpc-$crfile") || die "Cannot make directory lpc-
$crfile:$!\n";
      chdir("lpc-$crfile");

      open(OUTFILE, ">lpcinput") || die "Can't open lpcinput for
writing: $!\n";
```

```
      print OUTFILE "$protein\n";

      open(INFILE, "../$crfile") || die "Can't open ../$crfile for
reading: $!\n";
      $linecount = 0;
      while($line = <INFILE>) {
            $linecount++;
            next if ( ($line =~ /^ATOM/) and ($linecount == 1) );
            next if $line =~ /^HEADER/;
            next if $line =~ /^COMPND/;
            next if $line =~ /^END/;
            next if $line =~ /^REMARK/;
            $line =~ s/^ATOM  /HETATM/;
            chomp($line);
            $_ = $line;
            /.*?(30?\d+\.\d+)\s*(30?\d+\.\d+)\s*(30?\d+\.\d+)/;
            $x = $1; $y = $2; $z = $3;

            /HETATM\s*(\d+)\s*(\w+)/;
            $id = $lastid + $1; $symb = $2;

            #          1         2         3         4         5
6         7         8
            #
12345678901234567890123456789012345678901234567890123456789012345678901
234567890
            # HETATM 1933  C   UNK A 1      34.076  -3.086  16.427

            # print "=>$symb<=>$x<=>$y<=>$z<=\n";
            # print "$line\n";

            printf OUTFILE "%6s%5d %2s   %3s %1s %-4d
%8.3f%8.3f%8.3f\n", "HETATM", $id, $symb, "UNK", "A", 1, $x, $y, $z;
      }
      close INFILE;

      open(INFILE, "../LIG") || die "Can't open ../LIG for reading:
$!\n";
      while(<INFILE>) {
            chomp($_);
            next unless /^CONECT/;
            /^(CONECT)\s+(\d+)\s+(\d+)\s*(\d*)\s*(\d*)/;
            print OUTFILE "$1";
            print OUTFILE " " . ($2+$lastid) if $2;
            print OUTFILE " " . ($3+$lastid) if $3;
            print OUTFILE " " . ($4+$lastid) if $4;
            print OUTFILE " " . ($5+$lastid) if $5;
            print OUTFILE "\n";
      }
      close INFILE;
      print OUTFILE "END\n";
      close OUTFILE;

      $cmd = join( ' ', $lpc_executable, "1", "lpcinput");
      print "Executing $cmd\n" if $verbose;
      system($cmd);
```

```
      open(INFILE, "RES1") || die "Can't open RES1: $!\n";
      while(<INFILE>) {
            next unless /Normalised complementarity\s+(\d+\.\d+)/;
            close(INFILE);
            if ($1 > 1 or $1 < $threshold) {
                  # Value not within allowable threshold
                  last;
            }
            $lpcval = $1;
            open(INFILE, "RES1") || die "Can't open RES1: $!\n";
            while(<INFILE>) {
                  last if /2. Residues in contact with ligand/;
            }


            $residue_count = 0;
            $lys529 = $leu384 = $leu349 = $leu387 = $leu525 = $met343 =
$his524 = $arg394 = $glu353 = $thr347 = $asp351 = $ala350 = $met421 =
$leu539 = 10;

            while(<INFILE>) {
                  last if /3. List of putative hydrogen bonds between/;

                  if( /351A  ASP\*/) {
                        ($a, $b, $c, $val, $rest) = split /\s+/;
                        $asp351 = $val if $val < $asp351;
                  } elsif( /353A  GLU\*/) {
                  ($a, $b, $c, $val, $rest) = split /\s+/;
                        $glu353 = $val if $val < $glu353;
                  } elsif( /387A  LEU\*/) {
                     ($a, $b, $c, $val, $rest) = split /\s+/;
                        $leu387 = $val if $val < $leu387;
                  } elsif( /394A  ARG\*/) {
                        ($a, $b, $c, $val, $rest) = split /\s+/;
                        $arg394 = $val if $val < $arg394;

                  } elsif( /347A  THR\*/) {
                        ($a, $b, $c, $val, $rest) = split /\s+/;
                        $thr347 = $val if $val < $thr347;
                  } elsif( /524A  HIS\*/) {
                        ($a, $b, $c, $val, $rest) = split /\s+/;
                        $his524 = $val if $val < $his524;
                  } elsif( /387A  LEU\*/) {
                  ($a, $b, $c, $val, $rest) = split /\s+/;
                        $leu387 = $val if $val < $leu387;
            } elsif( /525A  LEU\*/) {
                  ($a, $b, $c, $val, $rest) = split /\s+/;
                        $leu525 = $val if $val < $leu525;
                  } elsif( /343A  MET\*/) {
                     ($a, $b, $c, $val, $rest) = split /\s+/;
                        $met343 = $val if $val < $met343;
            } elsif( /529A  LYS\*/) {
                  ($a, $b, $c, $val, $rest) = split /\s+/;
                        $lys529 = $val if $val < $lys529;
            } elsif( /353A  GLU\*/) {
                  ($a, $b, $c, $val, $rest) = split /\s+/;
                  $glu353 = $val if $val < $glu353;
```

```
                } elsif( /350A  ALA\*/) {
                    ($a, $b, $c, $val, $rest) = split /\s+/;
                      $ala350 = $val if $val < $ala350;
            } elsif( /421A  MET\*/) {
                    ($a, $b, $c, $val, $rest) = split /\s+/;
                      $met421 = $val if $val < $met421;
                } elsif( /349A  LEU\*/) {
                      ($a, $b, $c, $val, $rest) = split /\s+/;
                  $leu349 = $val if $val < $leu349;
                } elsif( /384A  LEU\*/) {
                      ($a, $b, $c, $val, $rest) = split /\s+/;
                      $leu384 = $val if $val < $leu384;
                }
            }
            close(INFILE);

            # print "$arg\t$thr\t$asp\t$glu\t$arg\t$his\t$lpcval\n";
            # if( (($arg < 4) || (($glu < 3) && ($asp < 3)) ) && (($ala
< 3.2) || ($thr < 3.2)) && ($lpcval > $threshold) ) {


            if( #(2.6 <= $asp351 && $asp351 <= 3.2) &&
                  #(2.4 <= $glu353 && $glu353 <= 4) &&
                  #(2.5 <= $leu387 && $leu387 <= 4) &&
                  #(2.8 <= $arg394 && $arg394 <= 5.4) &&
                  #(2.8 <= $thr347 && $thr347 <= 3.7) &&
                  #(2.9 <= $his524 && $his524 <= 5.1) &&
                  #(4.8 >= $leu384) &&
                  #(5.2 >= $leu349) &&
                  #(2.5 <= $met343) &&
                  #(1 > $lys529) &&
                  ($lpcval > $threshold) ) {

                  $trash=0;
                  print "Keeping value of $lpcval\n";

                  # Get LPC output file
                  open(INFILE, "RES1") || die "Can't open RES1: $!\n";
                  $lpcfile = "";
                  while(<INFILE>) {
                        $lpcfile .= $_;
                  }
                  close INFILE;

                  open(INFILE, "../$crfile");
                  $crcontents = "";
                  while(<INFILE>) {
                        $crcontents .= $_;
                  }
                  close INFILE;

                  $lpchash = md5_hex($lpcfile);
                  $crhash = md5_hex($crcontents);
                  $db_query = "insert into `parallel-results` " .
                                    "(conf_id, jobno, protein_id,
lpcval, residues, crhash, lpchash, crfile, lpcfile) " .
                                    "values " .
```

```
                                        "($conf_id, $jobid, $protein_id,
\'$lpcval\'," .
                                        "$residue_count, \'$crhash\',
\'$lpchash\'," .
                                        " \'$crcontents\', \'$lpcfile\')";
                    printf("%.130s\n",$db_query) if $verbose;
                    $dbh->do($db_query);

            }
            last;
      }
      chdir("..");
}

$cmd = join( " ", "/bin/rm", "-rf", $basedir);
print "$cmd\n" if $verbose;
system($cmd);

if($trash) {
      print "Zapping $conf_id from confs table\n" if $verbose;
      $db_query = "delete from parallel where conf_id = '$conf_id'";
      # $dbh->do($db_query);
}

$db_query = "update parallel set status='2' where conf_id =
'$conf_id'";
$dbh->do($db_query);


sub usage() {
      print "Incorrect arguments\n";
      print "-c conf_id [-i iterations] [-t threshold] [-v] [101]\n";
      exit(-1);
}
#
# vim:ts=4:sw=4
#
```

Chapter 4

Receptor Flexibility in Virtual Screening of Estrogen Receptor modulators *

Comprising

* Receptor Flexibility in Virtual Screening for Estrogen Receptor modulators; *J. Com. Aid. Mol. Des. (Ready for Submission)*

Andrew J. S. Knox, Mary J. Meegan, Vladimir Sobolev, Dermot Frost, David G. Lloyd.

*Antiestrogenically active 2-benzyl-1,1-diarylbut-2-enes: Synthesis, Structure-Activity Relationships and Molecular Modelling Study for Flexible Estrogen Receptor Antagonists; *Medicinal Chemistry,* 2006, 2(2): 147-168.

David G. Lloyd, Helena M. Smith, Timothy O' Sullivan, **Andrew J. S. Knox**, Daniela M. Zisterer and Mary J. Meegan.

4.1 Abstract

A key question in the area of computational drug design today is that of how to efficiently incorporate receptor flexibility in a virtual screening process. The effect of the ligand binding process has been widely studied and discussed with theories relating to induced fit or the pre-existence of receptor conformational ensembles being furnished to account for protein rearrangements. Subsequently, this has provoked numerous efforts to improve the ability of the computational chemist to model these motions through docking algorithm enhancements. In this work, three different studies are presented that illustrate how to effectively account for such residue movements in the docking process. Initially, using the ER as a target, the impact of docking into multiple X-ray receptor conformations was investigated. Secondly, a new method not yet utilized in drug design, whereby a Framework Rigidity Optimised Dynamic Algorithm (FRODA) was employed to generate receptor conformers starting from a single X-ray structure and docking examined. In the third study, results obtained from our vHTS protocol whereby the docking algorithm has been tailored to allow overlapping of ligands with certain flexible residues of the ER are presented. A measure of the degree of overlapping with residues is calculated to allow a threshold to be set. It is also shown how this can be used to correctly reproduce the binding mode of a ligand docked in another crystal structure of the ER other than its own. Having established a set of protocols for incorporating receptor flexibility, we finally endeavoured to determine a rationale behind the experimental binding selectivity observed for ERα over ERβ using one of the flexible procedures.

4.2 Introduction

Inclusion of protein flexibility in the initial phases of the drug discovery process using receptor-based virtual screening is of utmost importance when probing the full range of binding modes available to a series of ligands [1]. In screening a large database of compounds against a therapeutic target such as the Estrogen Receptor (ER) alpha, typically a single receptor conformation is utilised for ease and to negate the inevitable combinatorial explosion if all residues were allowed full motion [2]. However, this generally means that novel ligands extracted from the database have intrinsically similar binding modes, as they cannot fully explore all clefts of the binding site materialising from residue rearrangement [3]. In turn, identification of new and enhanced binding modes may be overlooked which could provide additional ligand selectivity through their inherent diversity [4].

To reduce all accessible conformational states of a receptor, two main schemes have been adopted to incorporate flexibility in a docking protocol [5-7]. Firstly, docking against a number of rigid receptor conformations using X-ray structures with different co-crystallised ligands, NMR spectroscopy or geometric simulation through dynamics may be employed [8-11].

Knegtel et al [12], introduced two different methods whereby an 'energy-weighted average' and a 'geometric-weighted average' of NMR or X-ray structures were used to create composite grids for docking and ranking a series of active ligands 'seeded' in a small decoy database. All known ligands docked within the top 21% of the database for the 'energy-weighted average' method while the 'geometric-weighted average' method find all inhibitors within the first 33%.

Barril and Morley [8] more recently have shown that utilising multiple X-ray receptor conformations of HSP90 permitted better binding mode predictions, but conversely lowered Enrichment (*E*) rates because of an increase in the presence of false positives. They also demonstrate that addition of simple pharmacophoric constraints remedies the *E* rate problems. Corroborating this, Murray et al [13] analysed the sensitivity of side-chain flexibility in binding mode prediction to thrombin, thermolysin and influenza virus neuraminidase with at least 6 different ligands co-crystallised to each using the PRO_LEADS algorithm. Importantly, only 49% of ligands cross-docked

correctly to non-native crystal structures, and the authors also point out that small movements in the receptor structure can cause up to 14 kJ/mol differences in binding energy from the correct binding energy.

FlexE [14] has been recently introduced as a method of sampling discrete receptor conformations starting from a single conformation using a united protein description, which was created by superimposing the structures of the ensemble. It has been evaluated using 105 crystal structures from the PDB with at least 3 ligands co-crystallised to each set. Starting from a single crystal structure, FlexE could generate an ensemble and dock each ligand to within 2Å of the experimental pose in 83% of the cases. FlexX [14] was used as a comparison and performed equally well in docking all ligands in 'cross-docking' experiments. However, in the case of aldose reductase, which has a highly flexible binding site, docking three potent inhibitors using FlexE and FlexX, only FlexE could reproduce binding modes accurately and FlexX could only dock them correctly in their native receptors respectively. Thus, it is important to note that these potent inhibitors would not be discovered via a single receptor conformation.

The second type of scheme adopted by the drug discovery community uses 'on-the-fly' methods that permit dynamic movement of identified flexible side-chains through rotamer library use or by exploring optimal torsional angles [15-17]. To date, several examples of these techniques have been published with evidence that flexible receptor flexibility docking outperforms rigid docking in binding predictions and also in the diversity of 'hits' retrieved in vHTS studies [18, 19].

Rockey et al [20], describe a new method, SCR, to compute side-chain flexibility in kinases through the use of rotamer libraries sampled by Monte Carlo methods. For five of the seven inhibitors the method could identify the correct receptor targets within the kinase family in >87% of the cases.

In agreement with the static receptor conformations produced by FlexE, Alberts et al [6] observed with their algorithm, which represented side-chain conformers by rotamers or χ dihedral angles dynamically, that the flexible version of MMP-1 allowed identification of known potent binders that would not have docked into a rigid site.

In this work we carried out three separate studies to show a variety of methods for incorporating side-chain movements. The first study demonstrates the variation in

docking results that can be attained through separately docking a set of 1000 compounds seeded with 36 known active ligands of the Estrogen Receptor (ER) to ten X-ray structures of ER alpha. Upon combining the highest docked ligand ranks from separate dockings the impact on *E* and FP was assessed to determine if utilising multiple receptors could account for some side-chain flexibility.

Beginning by selecting the X-ray structure that performs worst in single receptor docking experiments from above, a new algorithm never before applied to the realm of drug design developed by Thorpe and co-workers, FRODA (Framework Rigidity Optimised Dynamic Algorithm) [21] was used to explore its internal mobility in the second study. The FRODA method works in conjunction with another algorithm FIRST (Floppy Inclusions and Rigid Substructure Topography) [22] for rigidity analysis to discern which dihedral angles are locked or rotatable. Templates known as ghost templates are generated that contain information about rigid/flexible portions of the protein and are used to guide the motions of the flexible regions of the receptor during the simulation. We sought to determine if the conformational space explored by FRODA would permit docking of a greater number of ligands or enhance binding mode predictions when compared with the inputted structure.

In the third study, we introduce a new extension to the LIGIN [23-26] docking algorithm, that enables a treatment of flexible receptor docking through residue overlapping. A wall term that grows rapidly with distance decreases between atoms of the ligand and those of the nearest atoms of a residue is normally applied in the LIGIN docking procedure, but has been adjusted to allow a small degree of overlapping. In the rigid case, normally atomic bumping is calculated by;

$$E = \sum_{a} \sum_{b} E_{ab}$$

where,

$$E_{ab} = \begin{cases} 0 & \text{if } R_{ab} \geq R_0 \\ K(1/R_{ab}^{12} - 1/R_0^{12}) & \text{if } R_{ab} < R_0 \end{cases}$$

and $R_0 = 0.9 (R_a + R_b)$ and $K = 10^6$ Å 14. R_α and R_β are the van der Waals radii of contacting atoms, with R_{ab} the distance between them. In the case of inclusion of side-chain flexibility in the docking process, clashing side-chains are omitted from the overall binding prediction. In the docked solution files (CR1, CR2, CR3…) normally outputted immediately after docking, the degree of overlap between atoms of the ligand and the particular residue involved is output. Coupling this procedure with one that identifies those residues more susceptible to movement, through superposition of available well-resolved (<2.3Å) crystal structures and calculating the rsmd for each residue, one can ensure that overlapping thresholds can be set. This is very useful as the flexible procedure can be extended to allow a specific overlap of a number of flexible residues with the ligand and does not significantly hinder the speed of the docking process. We show how this flexible version of LIGIN can be utilized to dock co-crystallised ligands in their non-native forms and reproduce the same interactions observed in their native forms. Subsequently, the set of 1000 compounds used in the first two studies, were re-docked into the non-native receptors with both rigid and flexible settings, and E and FP rates compared.

In a further experiment, the interactions of a series of compounds [27] in both ERα and β through docking with FREDv2.11 are assessed. Differences of up to 17-fold selectivity were observed experimentally for binding to ERα than ERβ. As no ER active has been crystallised in human ERα or ERβ to date, it is not possible to determine the reasons why the same ligand would preferentially interact better in one isoform binding site over another utilising rigid docking. For this reason, FRODA was used to generate conformers of a well resolved alpha receptor structure and beta receptor structure and docking of each of the series of ligands to all receptor conformers was initiated.

4.3 Experimental Section - Computational

4.3.1 Receptor preparation –Study 1 & 3

Ten crystallographic structures displaying a resolution < 2.28 Å with a bound antagonist were downloaded from the Protein Data Bank (PDB_ID: 1SJ0 [28], 1UOM [29], 1XP1, 1XP6 , 1XP9, 1XPC [30], 1XQC [31], 1YIM, 1YIN [32], 3ERT [33]) and crystallographic waters were removed manually. The subsequent structures were read into Macromodel 6.5 [34] and re-written in PDB format to ensure bonds were represented correctly in this format. As two separate docking algorithms were employed in the docking to all receptors, hydrogens were added or negated accordingly. LIGIN does not take hydrogen atoms into account in the docking process and so no addition or minimisation of them was needed. However, FRED2.11 [35, 36] required hydrogen addition and so addition and optimisation of hydrogen positions was carried out using MOE.2005.06 [37] ensuring all other atom positions remained fixed.

4.3.2 Receptor preparation – Study 2

Initially, a subcomponent of the FRODA [21] package, FIRST5.2 (Floppy Inclusions and Rigid Substructure Topography) [22] was used to analyse rigid or flexible sections of the receptor and this information was carried into the geometric simulative process explored by FRODA (Framework Rigidity Optimised Dynamic Algorithm). Ten conformers of the receptor were produced with body-led movement rather than atom-led and an energy cut-off of –1.0KJ/cal. One thousand conformers were generated with every 100th saved as a PDB file. Sybyl6.91 [38] was utilised to convert all PDB structures to mol2.

4.3.3 Active and Decoy sets – Study 1-3

Thirty-six estrogen actives were selected from literature with activities ranging from nanomolar to low micromolar potency and converted to SMILES [39] format using ACD/ChemSketch 8.17 [40]. A subset of the Derwent World Drug Index (WDI) [41] was then extracted and passed through FILTER [42] with filtering properties, such as molecular weight <200 or >550, number of hydrogen bond donors 0<x<6 and acceptors 0<x<10, calculated logP <7. 500 molecules with specific stereochemical information denoted and 464 without were randomly selected using a Perl script. This was done to reflect the portion of marketed drugs that contain chiral centers or not. The two sets were merged with the 36 actives to produce a final set of 1000 compounds with similar characteristics.

4.3.4 Docking & Scoring – study 1& 2

FRED2.11 [35] was applied with default values and in this study to dock each ligand in all estrogen receptor structures. Rigid-body optimisation of each ligand pose using Chemgauss2 was also applied. In internal validation studies Chemgauss2 was deemed to be an excellent method for correct prediction of binding modes of ligand conformers in a lipophilic binding site such as the ER (see Chapter 3). All conformers were docked sequentially and scored using the Chemgauss2 scoring function.

4.3.5 Docking & Scoring – study 3

The LIGIN docking algorithm methodology is described elsewhere [24] but for the purpose of this study was extended to incorporate side-chain rearrangement through facilitating a degree of overlapping of atoms of the ligand with those of the residues. As an initial test of the efficacy of the procedure, docking to a non-native receptor using 4-Hydroxytamoxifen was carried out both rigidly and flexibly where overlap with one or two residues was allowed. The residues in contact with the ligands of each of the 10 crystal structures were assessed and the minimum/maximum distance thresholds set as constraints for use as a post-filter using a Perl script. Superimposition of the 10 crystal structures was executed and an svl script (MOE.2005.06) was used to analyse the rmsd of

each residue versus the same residue as it exists in each of the other nine conformational forms. Only those residues of the binding site that were deemed to be rigid from analysis of all the rmsd movements of all residues for each crystal structure (Glu353, Arg394, Leu387, Thr347) were constrained in this process. In both rigid and flexible dockings a two tiered scoring scheme involving a Normalised Complementarity threshold followed by rescoring with ChemScore (Sybylv6.91) was applied. However for the flexible docking, prior to rescoring with ChemScore, an additional energy term measuring the degree of overlap was included in the procedure to ensure reasonable and not excessive overlapping was achieved. All docked ligand poses with an overlap < nE + 5 (where n is a numerical value) with any of residues (those deemed to be flexible through analysis of a set of crystal structures) Met343, Phe404, Met421, Ile424, His524, Phe425 or Leu525 were retained. After all dockings, LPC (Ligand Protein Contacts) [43] was employed to assess if the correct interactions were reproduced as observed for 4-hydroxytamoxifen in PDB entry 3ERT [33]. Finally, docking of the decoy set of 1000 compounds containing 36 actives was undertaken both rigidly and flexibly utilising the same settings as above.

4.3.6 Success Criteria

The success of all protocols was evaluated by two metrics. Enrichment (E) rates were calculated for the top 0.5%, 1%, 1.5%, 2%, 2.5%, 3% and 5%. False Positive (FP) rates were also calculated for 60% (22 actives), 80% (29 actives) and 90% (32 actives) of the true positives. The calculation of E and FP are described elsewhere [44] in chapter 2.

4.3.7 ER selectivity studies

4.3.7.1 Ligand Preparation: Structures for compounds 17, 30, 32 and 50 were built using ACD/Chemsketch 8.17 and SMILES [39] strings generated for each. Marvinview 4.0.1 [45] was utilised to determine the protonation states of each ligand at pH 7.4 with each adjusted accordingly in the SMILES string. One hundred conformers of each compound were produced using Omega 1.8.1 with all conformers receiving a final MMFF optimisation step. All conformers were saved in mol2 format.

4.3.7.2 Receptor Preparation: PDB entries 3ERT and 1QKN were downloaded from the Protein Data Bank (PDB) and all crystallographic waters removed. Addition and optimisation of hydrogen positions was carried out using MOE.2005.06 ensuring all other atom positions remained fixed. FIRST5 (Floppy Inclusions and Rigid Substructure Topography) [46] in combination with FRODA (Framework Rigidity Optimised Dynamic Algorithm) [47] was utilised to firstly establish flexible regions of both proteins and subsequently, to generate conformers of the receptor. To ensure receptor conformational space was fully explored, a step size of 1.0 was used to displace every mobile atom randomly by a distance of up to 1Å with an energy cut-off of –1.0 KJ/cal. Four hundred conformers were generated with every 20th saved as a PDB. Macromodel 6.5 was utilised to convert all PDB structures to mol2.

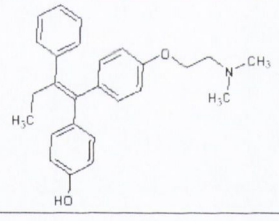
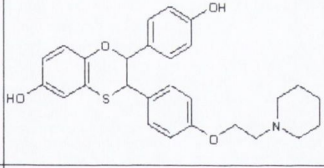
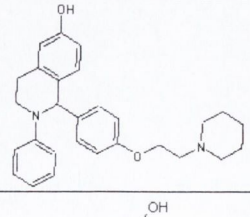
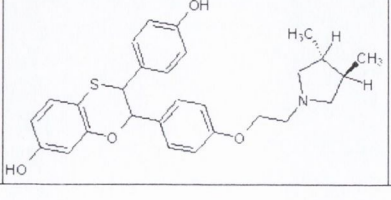
4.3.7.3 Docking: FRED2.11 [48] was utilized in this study to dock each ligand in both estrogen receptor isoforms. All default values were applied with rigid-body optimisation of each ligand pose using Chemgauss2. In internal validation studies we have found Chemgauss2 to be an efficacious method of evaluating the binding affinity of docked poses in a lipophilic binding site such as the ER. Sequential docking of all ligand and receptor conformers was carried out and the optimally docked solutions established by top score. Ligand Protein Contacts (LPC) software was used to calculate all interatomic contacts between ligand and receptor and furnish Normalised Complementarity (NC) values for each docked complex.

4.4 Results and Discussion

4.4.1 Study 1

This section firstly describes the impact that docking in 10 individual cavities of Estrogen Receptor alpha (ERα) has on the Enrichment (*E*) and False Positive (FP) rates. To prevent bias in the results, and to reflect the portions of a library typically assessed, we report the *E* rates at 0.5%, 1%, 1.5%, 2%, 2.5%, 3% and 5% for docking in each cavity. The theoretical maximum *E* rates for this procedure for each subset are 27.77 up to 3% of the ranked hitlist and 20 for 5%, as the decoy set comprises 36 actives and 964 decoys to make up 1000 compounds. Table 1 shows the performance of docking into each cavity separately, and portrays dramatically different results.

Table **(1)** Enrichment Factor (EF) for set of 1000 compounds docked in 10 cavities.

PDB	Co-crystallised Ligand	Resolution (Å)	EF 0.5%	EF 1%	EF 1.5%	EF 2%	EF 2.5%	EF 3%	EF 5%
3ERT		1.9	27.77	27.77	27.77	27.77	25.55	23.15	16.11
1SJ0		1.9	27.77	27.77	25.92	25	22.22	22.22	15.55
1UOM		2.28	27.77	27.77	25.92	25	22.22	22.22	15.55
1XP1		1.8	27.77	27.77	25.93	22.22	20	19.44	15

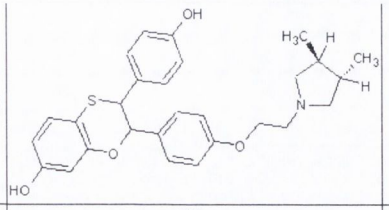
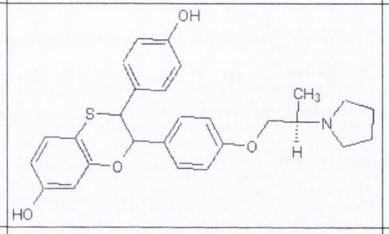
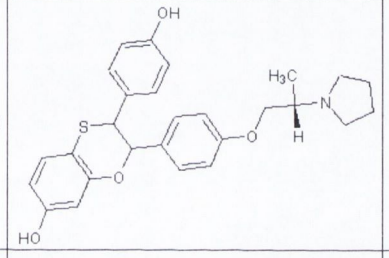
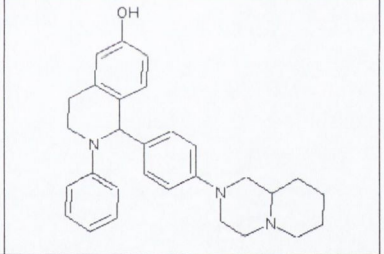
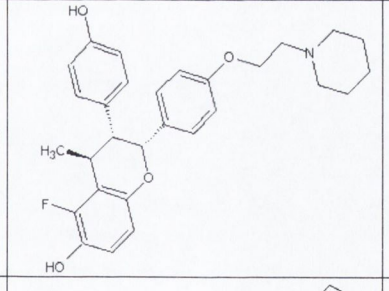
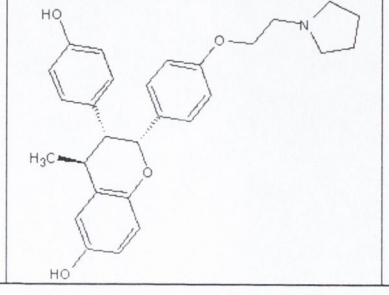
1XP6		1.7	27.77	27.77	25.93	22.22	20	19.44	14.44
1XP9		1.8	27.77	27.77	25.93	22.22	18.88	18.51	14.44
1XPC		1.6	27.77	27.77	25.93	23.61	22.22	19.44	15.55
1XQC		2	27.77	27.77	25.93	23.61	22.22	22.22	16.11
1YIM		1.9	27.77	25	24.07	20.83	18.88	18.51	13.33
1YIN		2.2	27.77	25	25.93	25	24.44	21.3	14.44

Table **(2)** False Positive rates for set of 1000 compounds docked in 10 cavities.

PDB	*FP 60%*	*FP 80%*	*FP 90%*	No. Actives Docked
3ERT	0.1	1.66	4.77	34 Docked
1SJ0	0.73	3.73	8.82	35 Docked
1UOM	0	2.8	9.65	34 Docked
1XP1	0.93	3.11	17.84	34 Docked
1XP6	0.93	4.88	6.74	34 Docked
1XP9	1.24	4.05	14.73	33 Docked
1XPC	1.04	3.32	11.62	34 Docked
1XQC	0.62	2.07	14.21	34 Docked
1YIM	1.45	10.68	19.19	34 Docked
1YIN	1.45	10.68	19.19	35 Docked

The PDB entry 3ERT is shown to deliver the best *E* rates throughout the first 5% of the ranked hitlist. In fact as much as a 25% difference between the *E* rates of 3ERT and the worst performing structure, 1YIM, for 2/2.5% of the ranked hitlist. Unique differences not observed using *E* rate calculations are observed when FP rates are calculated. For retrieval of 60%, 80% and 90% of true positives, false positive rates vary from 0 -1.45, 1.66 – 10.68, and 4.77 – 19.19 respectively. Importantly, it is clear again that the crystal structure 3ERT provides optimal discrimination between actives and inactives from the docked solutions obtained. Equally important however is the number of actives that are actually docked in each receptor. Taking only a hitlist containing the top 1000 docked conformers, it was observed that 1XP9 allowed docking and scoring of 33/36 actives, compared with 1SJ0 or 1YIN which allowed 35 to be docked and scored. This would have a direct impact on the diversity of hits retrieved in the screen if a PDB entry was chosen randomly.

To determine a rationale as to why some crystal structures allow more actives to be docked correctly and well ranked, we superimposed the ten receptors and calculated the rmsd of each residue as a unit between 1SJ0, 1XP9 and the remaining nine crystal structures. Figure 1 depicts the residues of the binding site (5Å radius) that are susceptible to movement calculated by superposition with MOE.2005.06.

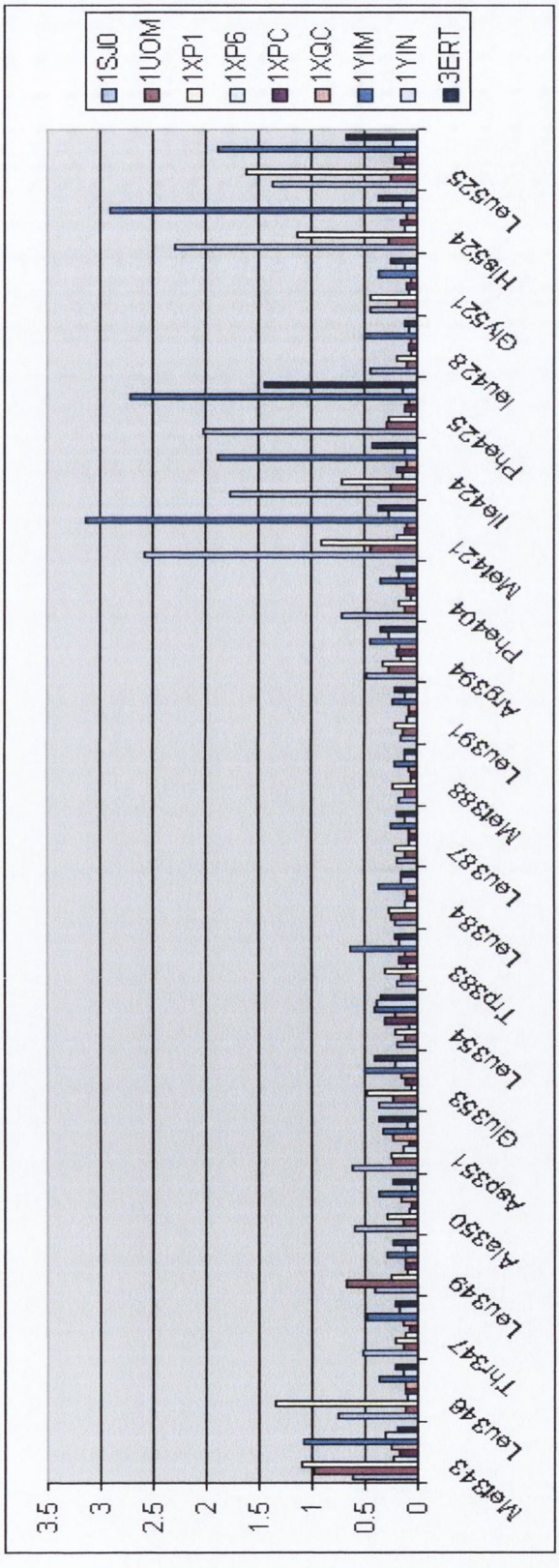


Figure 1. RMSD difference between residues of 1XP9 and nine other antagonist receptors.

It is apparent from Figure 1 that 1SJ0 differs from 1XP9 in docking 35/36 actives compared with 33/36 actives due to the movement of Met343, Leu346, Phe404, Met421, Ile424, Phe425, His524 and Leu525. The main difference between 1SJ0 and all other crystal structures appears to be the movement of Phe404. As a consequence of these residue movements, it is clear that only certain receptor x-rays will allow the docking of a full diverse set of actives. To overcome the effect that choosing a random receptor conformation may have, we sought to examine if docking to multiple receptors would facilitate a solution. Combining the ranked hitlists from dockings to two random receptors of the ER and selecting the highest rank for each compound, a new hitlist was produced.

Table **(3)** Enrichment Factor for set of 1000 compounds docked in multiple cavities.

PDB combination	***EF 0.50%***	***EF 1%***	***EF 1.50%***	***EF 2%***	***EF 2.5%***	***EF 3%***	***EF 5%***
1SJ0 + 1UOM	27.77	27.77	27.77	25	24.44	22.22	15.55
1UOM + 1YIN	27.77	27.77	27.77	26.39	25.56	25	15
1XP9 + 1XPC	27.77	27.77	25.93	22.22	20	18.51	14.44
1SJ0 + 3ERT	27.77	27.77	27.77	26.39	24.44	22.22	16.11
3ERT_1YIN	27.77	27.77	27.77	27.77	25.55	23.15	16.11
3ERT +1XP9	27.77	27.77	27.77	26.39	23.33	21.3	16.11

Table **(4)** False Positive rates for set of 1000 compounds docked in multiple cavities.

PDB	***FP 60%***	***FP 80%***	***FP 90%***	**No. Actives Docked**
1SJ0 + 1UOM	0.2	3.94	9.75	35
1UOM + 1YIN	0.2	4.98	11.62	35
1XP9 + 1XPC	1.14	3.53	18.98	34
1SJ0 + 3ERT	0.1	1.97	5.19	35
3ERT_1YIN	0.1	1.87	5.29	35
3ERT +1XP9	0.62	2.07	5.6	35

When two receptors are combined as depicted in Table 3 the *E* rates are significantly better with full enrichment observed generally for the first 1.5% with any combination of receptors (with the exception of 1XP9 + 1XPC). Combining 3ERT and 1YIN, the E rates do not change when compared with those obtained using the single cavity from 3ERT, however, the number of actives docked in the top 1000 of the ranked hitlist moves from 34 to 35. This would equate to more diversity in the hits retrieved from a virtual screen of

a large compound database without any loss of *E*. The gains observed at 2.5%, 3% and 5% on average are less but still represent increases of 8%, 5.1% and 1.8% respectively. These fractions are not important with regards to screening a large dataset as it would mean a very large portion of molecules would need to be assayed in order to find hits in this range. In the case of recovering 60% - 90% of true positives from the decoys, receptors that produce mediocre FP rates will exhibit better results when combined with another receptor whose docking results provided good FP rates on their own as observed in Table 4. The opposite is also true in that the receptor that provides good FP rates in its own will demonstrate lower FP rates when combined with a single worse performing receptor. Importantly if two receptors with equal performance in single docking runs, but showing different amounts of actives docked are then combined, the FP rate will remain the same but a higher number of actives can be docked in the hitlist of 1000 conformers.

4.4.2 Study 2

This section examines the use of a new computational method, FRODA (Framework Rigidity Optimized Dynamic Algorithm) and its ability to explore the conformational space of a receptor structure of the ER taken from the previous experiment. Selecting the single worst performing receptor from the docking studies previous, we sought to determine if FRODA could be used to generate 10 new conformations, some of which might either provide additional scope for docking more actives than the 33 observed with 1XP9 or alternatively enhance the *E* and FP rates. Table 5/6 below illustrates the E and FP values of each receptor compared with the inputted receptor structure and shows a wide variation in the number of actives retrieved.

Table **(5)** Enrichment Factor for set of 1000 compounds docked in multiple receptor conformations.

Receptor Conf	*0.50%*	*1%*	*1.50%*	*2%*	*2.5%*	*3%*	*5%*
1XP9	27.77	27.77	25.93	22.22	18.88	18.51	14.44
100	27.77	27.77	24.07	23.61	22.22	20.37	15
200	27.77	27.77	25.93	26.38	25.55	22.22	15
300	27.77	19.44	16.66	16.66	16.66	16.66	13.33
400	16.66	16.66	12.96	12.5	14.44	12.96	8.88
500	22.22	22.22	16.66	15.28	16.66	15.74	11.66
600	27.77	27.77	24.07	20.83	20	18.51	13.88
700	27.77	27.77	24.07	22.22	20	17.59	12.77
800	27.77	22.22	24.07	20.83	18.88	15.74	12.77
900	22.22	16.66	14.81	13.88	13.33	12.03	8.33
1000	11.11	11.11	9.26	9.72	8.88	8.33	5.55

Table **(6)** False Positive rates for set of 1000 compounds docked in multiple receptor conformations.

Receptor Conf	*60%*	*80%*	*90%*	No. Actives Docked
1XP9	1.24	4.05	14.73	33 Docked
100	0.83	4.77	20.44	32 Docked
200	0.1	4.36	7.88	33 Docked
300	1.45	15.56	24.59	34 Docked
400	4.98	14.63	12.47	34 Docked
500	3.11	21.16	23.96	34 Docked
600	1.03	7.26	11.51	33 Docked
700	1.76	5.5	12.03	34 Docked
800	2.39	6.74	23.34	33 Docked
900	9.23	20.02	22.1	33 Docked
1000	16.7	25.73	28.32	32 Docked

One thousand iteration were performed with every 100th being saved. It is evident that receptor conformation 200 provides a receptor conformation with significantly better E and FP values than its precursors with the same test database. Receptor conformation 700 shows a slight decrease in E and FP values but permits the retrieval of another active from the top 1000 ranked dataset. Figure 2 shows the rmsd's of each residue between conformation 200 produced by FRODA and all 10 crystal structures. Comparing with Figure 1, we see that receptor conformation 200 moves closer in conformation towards a conformation similar to 1SJ0 when compared with its original crystal structure 1XP9. This would account for the observed increase in the number of docked actives.

It is clear at this stage that FRODA samples conformational space sufficiently to allow new receptor conformations to be suggested that encompass structures closely related to those observed by crystallography. FRODA works in combination with the algorithm FIRST ((Floppy Inclusions and Rigid Substructure Topography) to analyse rigid or flexible portions of the receptor and this information is then transferred to the simulative process executed by FRODA (Framework Rigidity Optimised Dynamic Algorithm). FRODA has been successfully employed by Thorpe's group to explore the internal motions of receptor Barnase [21] and it was shown that the program captured the same main features of mobility when compared with those of the NMR data.

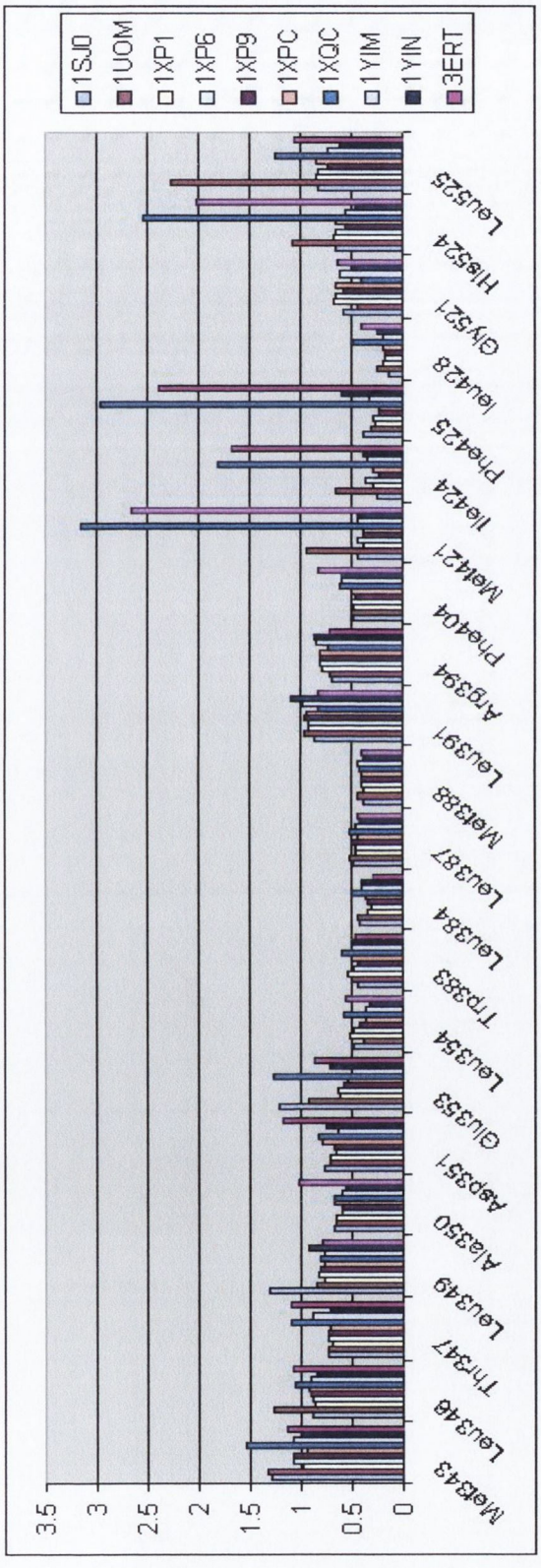


Figure 2. RMSD difference between residues of receptor conf 200 and 10 antagonist crystal structures

4.4.3 Study 3

This section examines the ability of a modified version of the docking program LIGIN [24]. In this version, coding had been extended to deliver a treatment of receptor flexibility, to dock firstly 4-hydroxytamoxifen in its non-native receptor form 1XP9, and secondly compares rigid docking of the same test set of 1000 compounds with the flexible docking procedure in terms of *E* and FP rates as before. Initially, 4-hydroxytamoxifen (OHT) was superposed over the bound ligand of 1XP9 (2s,3r)-3-(4-hydroxyphenyl)-2-(4-{[(2s)-2-pyrrolidin-1-ylpropyl]oxy}phenyl)-2,3-dihydro- 1,4-benzoxathiin-6-ol . LPC was employed to determine the interaction distances between OHT and the residues of 1XP9 as illustrated in Table 7.

Table **(7)** Putative H-bond interactions and distances between residues of OHT and active site residues of 1XP9

Ligand atom			Protein atom				Dist	Surf
N	Name	Class	Residue		Name	Class		
15	O4	I	GLU	353	OE2	II	2.3	23.6
15	O4	I	ARG	394	NH2	III	3.0	16.1
15	O4	I	GLU	353	OE1	II	3.5	0.7
15	O4	I	LEU	387	O	II	4.1	4.0
25	N24	I	ASP	351	OD1	II	3.4	2.2

When compared with the actual native structure 3ERT, no H-Bonding interaction is observed between the ethoxyaminoalkyl side-chain oxygen and the oxygen of Thr347 as highlighted below in Table 8.

Table **(8)** Putative H-bond interactions and distances between residues of OHT and active site residues of 3ERT

Ligand atom			Protein atom				Dist	Surf
N	Name	Class	Residue		Name	Class		
15	O4	I	GLU	353A	OE2	II	2.4	16.5
15	O4	I	ARG	394A	NH2	III	3.0	18.4
15	O4	I	GLU	353A	OE1	II	3.3	1.0
15	O4	I	LEU	387A	O	II	3.9	4.5
22	O20	II	THR	347A	OG1	I	4.0	4.3
25	N24	I	ASP	351A	OD1	II	3.8	1.6

Three separate dockings were carried out, one rigidly, one allowing overlap of a single residue and finally one allowing overlap of two residues. Taking only those with predicted interactions with the relevant residues, Table 9 depicts the interaction distances for each.

Table **(9)** Putative H-bond interactions and distances between residues of OHT and active site residues for both rigid and flexible docking runs.

Residue	3ERT	Superposed	RG1	RG2	RG3	RG4	FL1_1	FL1_2	FL1_3	FL1_4	FL2_1	FL2_2	FL2_3	FL2_4
THR* 347	3.7	3.5	4.1	2.8	3.8	3.8	3.4	3.8	5	**4**	4	5	4	3.1
ASP* 351	3.2	2.9	2.8	2.8	2.9	3.2	2.9	2.9	5	**2.9**	2.3	3.2	0.9	2.3
GLU* 353	2.4	2.3	3	2.5	3.1	2.8	3.5	3	3.3	**2.9**	4.8	5.8	3	3
LEU* 387	3.7	3.8	3.8	3.2	2.8	2.4	3.9	2.7	2.4	**3.4**	2.8	3.9	3	2.6
ARG* 394	3	3	2.9	4.4	2.9	3.4	2.6	2.8	5.4	**2.9**	4.5	5.5	4.6	4
HIS* 524	4	2.8	3.4	3	3.6	2.4	3	3.8	3.9	**3.5**	2.5	1	2.9	3.4

It can be seen from Table 9 that docking allowing overlapping of a single residue produces the closest binding mode to native observed in 3ERT. An overlap of 0.4E + 03 with Leu525 is produced in this case and corroborates both Figures 1 and 2 which show greater flexibility in this region among all receptor forms.

With this in mind, flexible docking permitting overlapping of a single residue of 1XP9 for each molecule was undertaken. As a comparison, and to exemplify the advantages of this flexible treatment, we undertook to dock the same set rigidly using 1XP9 as the receptor. Assessing the binding modes of the 36 actives from both docking methods exemplified the problem associated with receptor rigidity, as one antiestrogen could not find a docking mode and another docked in an inverted binding orientation as shown in Figure 3. With the flexible parameters, all actives docked correctly in stereotypical estrogenic/antiestrogenic modes. Merging the docked ligands from the rigid docking together as a set, and those from the flexible docking as another set also, the total van der Waals surface area (VSA) was calculated for both using MOE.2005.06. The VSA calculated for the rigid set was 907.45 and for the flexible set was 867.5. This gives a clear idea as to the compactness, relative to the native pose of the docked actives for the rigid and flexible docking procedures. The flexible procedure appears to allow 'tighter' docking because of more accurate positioning of the ligands due to the allocation of certain side-chain movements.

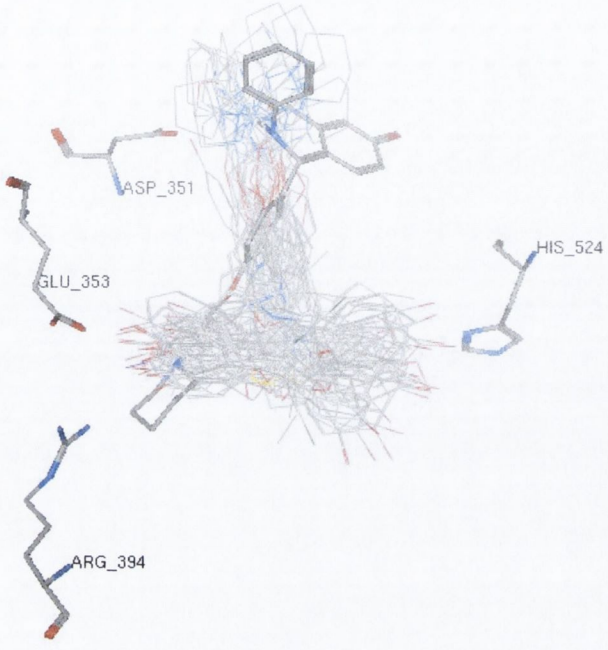


Fig **(3)** Overlay of 34/36 docked actives using rigid docking protocol.

Table 10 shows the difference in E and FP rates achieved by each using the whole test set of 1000 molecules. Dramatically different results are obtained clearly indicating the advantages of this extension to LIGIN.

Table **(10)** Enrichment Factors & False Positive rates for rigid and Flexible docking runs with LIGIN

	EF 0.50%	***EF 1%***	***EF 1.50%***	***EF 2%***	***EF 2.5%***	***EF 3%***	***EF 5%***
RIGID	0	11.11	12.96	15.27	15.55	16.66	12.77
FLEXIBLE	16.66	19.44	16.66	19.44	17.77	17.59	13.88

	FP 60%	***FP 80%***	***FP 90%***
RIGID	2.49	6.02	10.99
FLEXIBLE	1.35	6.85	8.82

4.4.4 ER receptor selectivity studies

We report the development of a series of hydroxylated 2-benzyl-1,1-diarylbut-2-enes containing a flexible core scaffold structure differing from the 1,1,2-triarylethylene typical of tamoxifen analogues. To rationalize the observed biological activity of the series of hydroxylated 2-benzyl-1,1-diarylbut-2-enes containing a flexible core scaffold as shown in Figure 4, a thorough computational investigation was undertaken [49]. Compounds A-D were shown to have 10, 17, 7.2, 11.5 times ERα over ERβ selectivity respectively. The compounds also exhibited antiproliferative activities (IC_{50}) 9.36uM, 0.697uM, 1.09uM, 0.873uM respectively when evaluated using the MCF-7 breast cancer cell line. Noteworthy, compound A was found to be the most active with pyrrolidin-2-one side-chain assisting in binding and antiproliferative activity. The compounds B & C contained pivaloyloxy esters on their C-ring at ortho and para positions respectively which were designed as prodrugs for the release of free phenolic compounds as they would be slowly hydrolysed *in vivo*.

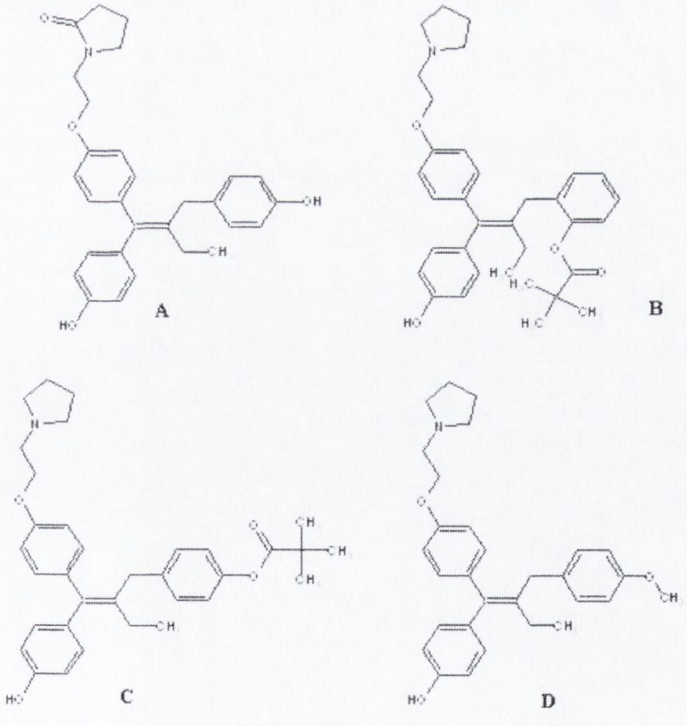


Fig **(4)** Compound series A (Top-left), B (Top-right), C (Bottom-left) and D (Bottom-right).

To quantitatively examine the interactions established by compounds in both isoforms of the ER (α and β) is a challenging problem as currently there are no crystal structures of human ERα/β with the same antagonist co-crystallized. To address this issue, we employed a technique involving receptor conformer generation using FIRST5 software [46] in combination with FRODA as recently developed by Thorpe and co-workers [47]. FIRST determines flexible regions of the protein by assigning topological bars that describe the nature of all the bonds present in the protein (covalent, H-bond, salt bridge, hydrophobic tethers). The network of bars are then analysed via graph-theoretical algorithm to provide input for FRODA on whether a bond will participate in a dihedral angle rotation or not. Ghost templates extracted from information about the rigid sections of the protein guide the movement of the flexible areas within FRODA. Noteworthy, to reduce the possibility of steric clashes being present, it is critical that the initial crystal structures used are highly resolved (~ ≤ 2Å).

For this reason, the crystal structure of ERα (3ERT[33]) with 4-Hydroxytamoxifen co-crystallised and ERβ (1QKN [50]) with Raloxifene were utilised. FRED2.11 [48] was used to dock conformers of the series of ligands generated by Omega1.8.1 [51] into each conformer of the receptor and scored with Chemgauss2. The top ranked pose over all of the receptor and ligand conformers for each ligand were selected and atomic interactions were analysed by Ligand Protein Contacts (LPC) software [43]. The residues depicted are those that have been previously shown to be crucial in the binding process: Asp351 (interacts with the basic side-chain nitrogen), Glu353 and Arg394 (anchor the ligand in the active site), His524 (additionally important in ligand binding process). Table 11 illustrates the interactions made by each ligand with both receptor isoforms.

Table **(11)** Summary of key Ligand-Protein contacts[a]

Compd	Isoform	Asp 351 (Asp 258)	Glu 353 (Glu 260)	Arg 394 (Arg 301)	His 524 (His 430)	NC	Chemgauss2
A	α	3.9	3.2	4.1	3.2	0.89	-52.84
A	β	3	3.5	5.5	2.4	0.69	-51.22
B	α	2.8	2.4	3.2	3	0.91	-50.48
B	β	3.2	3.2	5.1	4.1	0.72	-47.65
C	α	3.1	2.3	1.8	5.5	0.76	-54.91
C	β	4.3	5.3	---	---	0.61	-51.22
D	α	3.3	3.9	1.8	5.1	0.9	-56.21
D	β	4.1	3.1	4.9	6.4	0.8	-43.08
OHT	α	3.2	2.4	3	4	0.89	---
RAL	β	3.3	2.6	3	2.6	0.69	---

[a]Data provided as nearest distance (Å) between atoms of ligand and the residue. Residues named are those present in crystal structure 3ERT, those in parenthesis denote residues of crystal structure 1QKN. OHT, 4-Hydroxytamoxifen from PDB entry 3ERT; RAL, Raloxifene from PDB entry 1QKN; NC, Normalised Complementarity; Chemgauss2, score attributed to the top ranked solution.

As detailed in Table 11 compounds A-D dock in a typical antiestrogenic manner when compared with OHT and RAL. Figure 5 (A)/(B) clearly illustrate a similar binding mode for compound B in both receptor isoforms, with only the benzyl pivaloxy moiety differing in position. In a previous study involving the synthesis and biological testing of a series of flexible antiesrogens, it was noted that addition of substituents on the C-ring might provide additional binding through His524 of ERα [52]. In agreement with this, Grese et al [53] noted that replacement of the para-hydroxy atom on the C-ring of Raloxifene with a hydrogen atom decreased its binding affinity 5-fold. Compound A confirms this hypothesis, showing interaction distances from the para-hydroxy to His524 (His430) of 3.2 (2.4) respectively in Table 11.

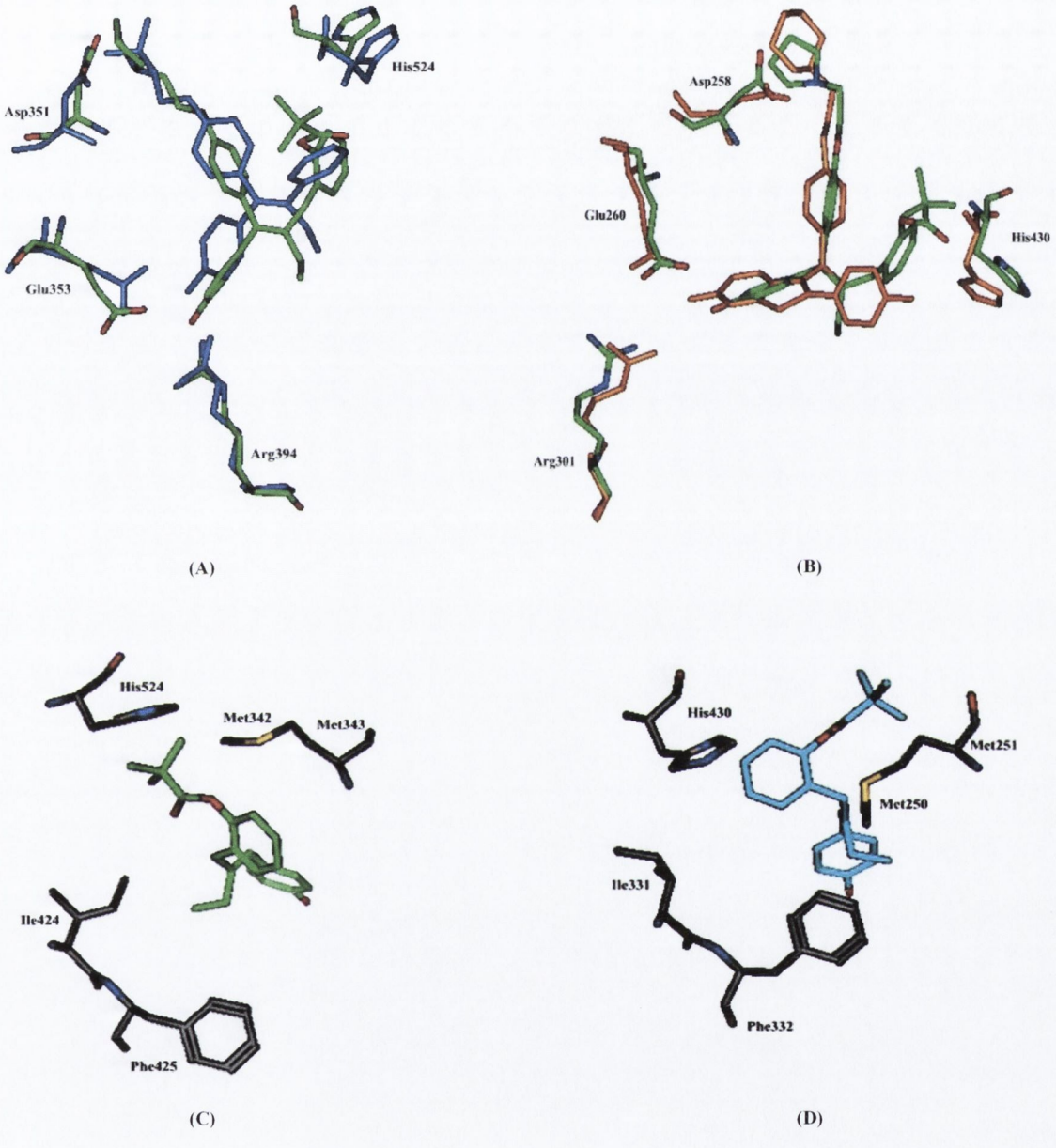


Fig **(5)** Top ranked docking solution of compound B (coloured by atom) superimposed by backbone on 3ERT (blue) (A) and 1QKN (orange) (B). (C) Depicts residues of ERα that form a small lipophilic cavity preventing rotation of the benzyllic pivaloxy group away from the active site. (D) Phenyl ring of Phe332 is projected inwards into active site with Met250 further enclosing the section.

Another feature of the series is their inherent isoform selectivity with compound B for example, possessing 17-fold preference in binding to ERα over ERβ. Table 11 also depicts the calculated Normalised Complementarity (NC) for each top docked solution, which illustrates the 'buriedness' of a molecule within an active site of a protein and chemical compatibilities of contacting atoms. What is evident from the docking studies is that the NC value is always lower for ERβ than ERα. Corroborating this, the Chemgauss2 scores obtained were consistently more negative in the alpha active site versus the beta site.

To further investigate the reasons for the selectivity, the active site of the top ranked solution for compound B docked in both isoforms was visually examined. Figure 5 shows the two positions of the pivaloxy group substituent attached at the ortho position of the C-ring of compound B when docked in ERα and β. The movement of flexible residues Ile424, Phe425, His524 and mainly Met342 of ERα compared with Ile331, Phe332, His430 and Met250 of ERβ prevent the pivaloxy substituent from adopting a position facing away from the active site. This small lipophilic cavity allows a dramatic difference in interactions with His524 and Arg394 to be achieved if occupied as observed in Table 11 and corroborated by binding affinity data. Taking the twenty receptor conformers of 3ERT and 1QKN, we determined by an svl script written for MOE.2005.06 [37] the predicted degree of motion by rmsd of Ile, Phe and Met for each receptor. This showed that Ile424, Phe425 and Met342 could move 1.18, 0.86 and 1.57 angstroms respectively in ERα. However, in ERβ, Ile331, Phe332 and Met250 could move 3.80, 3.16 and 1.88 angstroms. This indicated to us that this particular lipophilic binding cleft was far more restricted in ERα, allowing the benzylic moiety to adopt a different position and holding it in a more favourable position. In agreement with this, Amari et al also recently recognized the presence of a small lipophilic cavity maintained by Ile424 (Ile376) and His524 (His475) in both receptor isoforms that could accommodate a small substituent [54]. However, the involvement of Met342 (Met250) in the process was not described.

Importantly, this cavity difference would not be recognised without incorporating receptor flexibility into the procedure and exemplifies the necessity to apply both ligand and receptor flexibility to accurately assess binding modes of newly synthesized ligands.

4.5 Conclusion

We have investigated the quality of results obtained by three different flexible docking methods compared with rigid receptor docking utilising ERα as a target. Flexibility of the receptor was considered in the first study using multiple crystallographic receptors of the ER and was shown to assist in overcoming the potential problems observed with randomly selecting a single crystal structure of ERα for rigid docking. In some cases combining two receptor ranked hitlists and choosing each compounds highest rank improved the enrichment by up to 25% over the first 2-2.5%. Concurrently an increase in the number of actives docked occurred with the use of multiple receptors. Barril and Morley have also shown that combination of two receptor cavities leads to a 12% increase in E but combining more than that degrades this effect [8]. We also suggest that only two cavities should be combined if they exhibit equivalently low E rates in single dockings in order to ensure a minimum of false positives is observed.

We next elucidated whether generating 10 new conformers of a receptor that performed poorly in single docking experiments would sufficiently mimic side-chain movement to positions that could allow enhanced binding predictions or an increase in the number of docked actives. Of the 10 conformations, one exhibited higher E and FP rates than the inputted PDB entry 1XP9. Additionally, with 1XP9 in the single docking, only 33 out of 36 actives docked within the top 1000 conformers and four of the conformation sampled by FRODA allowed 34 to dock. The FRODA method has never been used in this respect before and here we show how it can effectively enhance both enrichment and diversity beginning form a single X-ray representation.

Finally, we introduced a new extension to the docking algorithm, LIGIN, which accounts for side-chain rearrangements through a wall term that now tolerates some side-chain overlapping. The new protocol allows reproduction of the correct binding orientation and distances observed in 3ERT when 4-hydroxytamoxifen is docked into its non-native receptor form 1XP9. Comparing rigid with flexible docking using LIGIN we see that 1 of 36 actives docks incorrectly with another not docking at all in 1XP9. Flexible docking produce more realistic docking poses that overlap more closely together as calculated from the van der Waals surface area of both sets (34 rigid merged, 36

flexible merged). More specifically, we show that E and FP rates are significantly enhanced (3-fold better for the first 0.5%) utilising the flexible docking procedure over the rigid one.

In a final experiment, we demonstrate the efficacy of FRODA in rationalizing the selective binding of a series of flexible antiestrogens binding to ERα over ERβ. It is clear also that without introducing a level of receptor flexibility there would be no way to determine the actual interactions between atoms of the ligand and those of the residues. These results will allow a refinement in the selection of candidate compounds for synthesis of selective ERα or ERβ ligands.

Ultimately, it is clear that flexible docking provides an improved representation of binding modes which impacts positively on hit retrieval. It is also important to note that some receptor representations will have a negative effect on this process and should be 'weeded' out prior to docking against large compound databases. Efforts concentrated on enhancing protein flexibility will inevitably assist in the drug discovery process in the future.

4.6 References

1. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M., Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **2004,** 47, (1), 45-55.
2. Hou, T.; Xu, X., Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* **2004,** 10, (9), 1011-33.
3. Davis, A. M.; Teague, S. J.; Kleywegt, G. J., Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* **2003,** 42, (24), 2718-36.
4. Teague, S. J., Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* **2003,** 2, (7), 527-41.
5. Teodoro, M. L.; Kavraki, L. E., Conformational flexibility models for the receptor in structure based drug design. *Curr Pharm Des* **2003,** 9, (20), 1635-48.
6. Alberts, I. L.; Todorov, N. P.; Dean, P. M., Receptor flexibility in de novo ligand design and docking. *J Med Chem* **2005,** 48, (21), 6585-96.
7. Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A., Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc* **2005,** 127, (26), 9632-40.
8. Barril, X.; Morley, S. D., Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J Med Chem* **2005,** 48, (13), 4432-43.
9. Cavasotto, C. N.; Abagyan, R. A., Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol* **2004,** 337, (1), 209-25.
10. Yoon, S.; Welsh, W. J., Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J Chem Inf Comput Sci* **2004,** 44, (1), 88-96.
11. Vigers, G. P.; Rizzi, J. P., Multiple active site corrections for docking and virtual screening. *J Med Chem* **2004,** 47, (1), 80-9.
12. Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M., Molecular docking to ensembles of protein structures. *J Mol Biol* **1997,** 266, (2), 424-40.
13. Murray, C. W.; Baxter, C. A.; Frenkel, A. D., The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* **1999,** 13, (6), 547-62.
14. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T., FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* **2001,** 308, (2), 377-95.
15. Leach, A. R.; Lemon, A. P., Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* **1998,** 33, (2), 227-39.
16. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, (3), 727-48.
17. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **1997,** Suppl 1, 215-20.
18. Frimurer, T. M.; Peters, G. H.; Iversen, L. F.; Andersen, H. S.; Moller, N. P.; Olsen, O. H., Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys J* **2003,** 84, (4), 2273-81.
19. Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; Kuhn, L. A., Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **1998,** 33, (1), 74-87.
20. Rockey, W. M.; Elcock, A. H., Rapid computational identification of the targets of protein kinase inhibitors. *J Med Chem* **2005,** 48, (12), 4138-52.
21. Wells, S.; Menor, S.; Hespenheide, B.; Thorpe, M. F., Constrained geometric simulation of diffusive motion in proteins. *Phys Biol* **2005,** 2, (4), S127-36.
22. Zavodszky, M. I.; Lei, M.; Thorpe, M. F.; Day, A. R.; Kuhn, L. A., Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins* **2004,** 57, (2), 243-61.
23. Sobolev, V.; Eyal, E.; Gerzon, S.; Potapov, V.; Babor, M.; Prilusky, J.; Edelman, M., SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. *Nucleic Acids Res* **2005,** 33, (Web Server issue), W39-43.
24. Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M., Molecular docking using surface complementarity. *Proteins* **1996,** 25, (1), 120-9.

25. Sobolev, V.; Moallem, T. M.; Wade, R. C.; Vriend, G.; Edelman, M., CASP2 molecular docking predictions with the LIGIN software. *Proteins* **1997,** Suppl 1, 210-4.
26. Sobolev, V.; Edelman, M., Modeling the quinone-B binding site of the photosystem-II reaction center using notions of complementarity and contact-surface between atoms. *Proteins* **1995,** 21, (3), 214-25.
27. Lloyd, D. G.; Smith, H. M.; O' Sullivan, T.; Knox, A. J. S.; Zisterer, D. M.; Meegan, M. J., Antiestrogenically active 2-benzyl-1,1-diarylbut-2-enes: Synthesis, Structure-Activity Relationships and Molecular Modelling Study for Flexible Estrogen Receptor Antagonists. *Medicinal Chemistry* **2006,** In Press.
28. Kim, S.; Wu, J. Y.; Birzin, E. T.; Frisch, K.; Chan, W.; Pai, L. Y.; Yang, Y. T.; Mosley, R. T.; Fitzgerald, P. M. D.; Sharma, N.; Dahllund, J.; Thorsell, A. G.; DiNinno, F.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen Receptor Ligands. II. Discovery of Benzoxathiins as Potent, Selective Estrogen Receptor <img src="http://pubs.acs.org/images/gifchars/alpha.gif" border="0"> Modulators. *J. Med. Chem.* **2004,** 47, (9), 2171-2175.
29. Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, P.; Fournier, B.; Halleux, C.; Kallen, J.; Keller, H.; Schlaeppi, J. M.; Stark, W., Estrogen receptor modulators: identification and structure-activity relationships of potent ERalpha-selective tetrahydroisoquinoline ligands. *J Med Chem* **2003,** 46, (14), 2945-57.
30. Blizzard, T. A.; Dininno, F.; Morgan, J. D., 2nd; Chen, H. Y.; Wu, J. Y.; Kim, S.; Chan, W.; Birzin, E. T.; Yang, Y. T.; Pai, L. Y.; Fitzgerald, P. M.; Sharma, N.; Li, Y.; Zhang, Z.; Hayes, E. C.; Dasilva, C. A.; Tang, W.; Rohrer, S. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. Part 9: Dihydrobenzoxathiin SERAMs with alkyl substituted pyrrolidine side chains and linkers. *Bioorg Med Chem Lett* **2005,** 15, (1), 107-13.
31. Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, P.; Fournier, B.; Geiser, M.; Halleux, C.; Kallen, J.; Keller, H.; Ramage, P., Selective estrogen receptor modulators with conformationally restricted side chains. Synthesis and structure-activity relationship of ERalpha-selective tetrahydroisoquinoline ligands. *J Med Chem* **2005,** 48, (2), 364-79.
32. Tan, Q.; Blizzard, T. A.; Morgan, J. D., 2nd; Birzin, E. T.; Chan, W.; Yang, Y. T.; Pai, L. Y.; Hayes, E. C.; DaSilva, C. A.; Warrier, S.; Yudkovitz, J.; Wilkinson, H. A.; Sharma, N.; Fitzgerald, P. M.; Li, S.; Colwell, L.; Fisher, J. E.; Adamski, S.; Reszka, A. A.; Kimmel, D.; DiNinno, F.; Rohrer, S. P.; Freedman, L. P.; Schaeffer, J. M.; Hammond, M. L., Estrogen receptor ligands. Part 10: Chromanes: old scaffolds for new SERAMs. *Bioorg Med Chem Lett* **2005,** 15, (6), 1675-81.
33. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L., The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998,** 95, (7), 927-37.
34. MACROMODEL 6.5, developed and distributed by Schrodinger Inc. *(URL: http://www.schrodinger.com).*
35. FRED (version 2.1.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com).*
36. 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.
37. Molecular Operating Environment (MOE), developed and distributed by Chemical Computing Group. *(http://www.chemcomp.com).*
38. Sybyl6.91, distributed by Tripos Inc.
39. Weininger, D., SMILES: A Chemical Language and Information System. *J. Chem. Inf. Comput* **1988,** 28, 31-36.
40. Chemsketch. v8.17, *www.acdlabs.com*.
41. Derwent World Drug Index, (URL: http://thomsonderwent.com/products/lr/wdi).
42. FILTER, distributed by Openeye Scientific Software.
43. Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M., Automated analysis of interatomic contacts in proteins. *Bioinformatics* **1999,** 15, (4), 327-32.
44. 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.
45. MarvinView, distributed by Chemaxon Ltd. *(URL:http://www.chemaxon.com/marvin).*
46. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F., Protein flexibility predictions using graph theory. *Proteins* **2001,** 44, (2), 150-65.

47. Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M., Protein Flexibility using constraints from molecular dynamics simulations. *Physical Biology* **2005,** Accepted for Publication.
48. FRED (version 1.1), developed and distributed by Openeye Scientific Software. *(URL:http://www.eyesopen.com)*.
49. Lloyd, D. G.; Smith, H. M.; O' Sullivan, T.; Knox, A. J. S.; Zisterer, D. M.; Meegan, M. J., Antiestrogenically active 2-benzyl-1,1-diarylbut-2-enes: Synthesis, Structure-Activity Relationships and Molecular Modelling Study for Flexible Estrogen Receptor Antagonists. *Medicinal Chemistry* **2005,** Accepted for Publication.
50. Pike, A. C.; Brzozowski, A. M.; Hubbard, R. E.; Bonn, T.; Thorsell, A. G.; Engstrom, O.; Ljunggren, J.; Gustafsson, J. A.; Carlquist, M., Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *Embo J* **1999,** 18, (17), 4608-18.
51. OMEGA 1.8.1, distributed by Openeye Scientific Software.
52. Lloyd, D. G.; Smith, H. M.; O' Sullivan, T.; Zisterer, D. M.; M.J., M., Synthesis, Structure-Activity Realtionships and Antagonistic Effects in Human MCF-7 Breast Cancer Cells of Flexible Estrogen Receptor Modulators. *Medicinal Chemistry* **2005,** 1, (4), 335-353.
53. Grese, T. A.; Cho, S.; Finley, D. R.; Godfrey, A. G.; Jones, C. D.; Lugar, C. W., 3rd; Martin, M. J.; Matsumoto, K.; Pennington, L. D.; Winter, M. A.; Adrian, M. D.; Cole, H. W.; Magee, D. E.; Phillips, D. L.; Rowley, E. R.; Short, L. L.; Glasebrook, A. L.; Bryant, H. U., Structure-activity relationships of selective estrogen receptor modulators: modifications to the 2-arylbenzothiophene core of raloxifene. *J Med Chem* **1997,** 40, (2), 146-67.
54. Amari, G.; Armani, E.; Ghirardi, S.; Delcanale, M.; Civelli, M.; Caruso, P. L.; Galbiati, E.; Lipreri, M.; Rivara, S.; Lodola, A.; Mor, M., Synthesis, pharmacological evaluation, and structure-activity relationships of benzopyran derivatives with potent SERM activity. *Bioorg Med Chem* **2004,** 12, (14), 3763-82.

Chapter 5

Cheminformatic treatments of the Estrogen Receptor *

Comprising

*Oncology Exploration: Charting Cancer Medicinal Chemistry Space; *Drug Discovery Today,* 2006; 11(3/4): 149-159

David G. Lloyd, Georgia Golfis, **Andrew J. S. Knox**, Darren Fayne, Mary J. Meegan, Tudor I. Oprea.

* Estrogen Receptors: Molecular Interactions, Virtual Screening and Future Prospects; *Curr. Top. Med. Chem.* 2006; 6(2): 211-237.

Andrew J. S. Knox, Mary J. Meegan, David G. Lloyd.

5.1 Oncology Exploration – Abstract

Approaches to the experimental determination of protein–ligand molecular interactions are reliant on the chemical appropriateness of the compounds being tested. The application of large, randomly designed combinatorial libraries has allowed the creation of more-focused 'drug-like' libraries. Prior to construction of a library of small-molecule compounds for chemical synthesis, it is important to screen the potential compounds to remove undesired chemical moieties within defined limits of physiochemical properties. A Principal Component Analysis (PCA) approach was utilized to analyze the 3D descriptor space of active and inactive (drug-like) cancer medicinal chemistry compounds. The computational analysis indicates that cancer-active compounds exist in different regions in space and occupy a much larger volume than a generic drug-like space. Additionally, most of them fail the commonly applied filters for orally bioavailable drugs. This is of great significance when designing orally bioavailable cancer drugs. Moreover, this generic treatment is shown to have a specific application, illustrated using active ligands of the ER. The importance of pre-filtering or biasing towards an appropriate area of chemical space to facilitate retrieval of compounds similar to known validated actives for oncology targets is thus demonstrated.

5.2 Introduction

The successful application of the processes of virtual and physical screening for active ligands is wholly reliant on the structural suitability 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. When dealing with oncology however, the tried and trusted rules of engagement do not always apply.

Cancer chemotherapeutics generally utilise the same approach in targeting the cellular processes that cancers use to grow, divide and multiply. The dawn of rational drug design has initiated a move towards drugs that can selectively modulate specific biochemical processes and their machinery, thus reducing toxic effects. The main mechanistic groups falling under the arsenal of chemotherapeutics are outlined next.

Antimetabolites interfere with enzymatic steps in nucleotide biosynthesis in tumor cells. A well-known target, the Dihydrofolate Reductase (DHFR), limits the formation of nucleotides and thus prevents DNA replication when inhibited by a drug such as methotrexate. Other mechanisms that are utilised in this approach involve blocking the production of purines or pyrimidines nucleosides by compounds such as 6-mercaptopurine and gemcitibine respectively.

Genotoxic drugs such as alkylating agents, and platinum-containing compounds, actually modify DNA directly either through alkylating and crosslinking guanine bases or by intra-strand crosslinking of guanine bases. A third class of drugs namely, intercalating drugs, inserts themselves in the minor groove between nitrogen base pairs in the DNA helix and prevent replication.

Kinase inhibitors use the phosphorylation pathway of the kinases to regulate cell proliferation. Gleevec (imatinib) for example is a tyrosine kinase inhibitor that prevents ATP binding and concurrent kinase activity by binding to the BCR-ABL receptor, often damaged in chronic myeloid leukaemia and gastrointestinal tumors. Iressa, another

tyrosine kinase inhibitor, functions by binding to the EGF receptor and inhibiting phosphorylation of its tyrosine residues.

Proteasome inhibitors block the protein degradation pathway inducing cell death. For example, Bortezomib interacts with a threonine residue at the catalytic site of the proteasome, and induces apoptosis.

Tubulin inhibitors, Vincristine and Taxol (albeit through polymerisation or depolymerisation), act by inhibiting the formation or stabilise microtubules, which results in interference in the normal process of cell division.

Finally, hormonal agents such as Selective Estrogen Receptor Modulators (SERMs), Selective Androgen Receptor Modulators (SARMs) and aromatase inhibitors all are effective by blocking the activity of a hormone or the biosynthetic production of it. In many cancers such as breast, ovarian, and prostate it is known that hormone over-production leads to cell proliferation and in turn increases the size of the tumor.

Partitioning and classifying cancer medicinal chemistry space is not straightforward. Even the relatively small number of validated molecular targets makes for a multitude of molecular scaffolds identified as cancer actives. The last two decades have witnessed a tremendous increase in our understanding of the pathology and molecular biology of human cancers. [1] While 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 available for cancers clearly 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, which have been identified in the development of treatments for cancers with improved specificity, toxicology profiles and efficacy. [3]

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 Rule of 5 compliance: molecular weight (MW)$\leq$500; (LogP)$\leq$5; no. H-bond donors$\leq$5; and acceptors$\leq$10).

5.3 Computational Analysis

5.3.1 Cancer Space

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 [4]. These operations facilitated the creation of graphical representations of the 3D space spanned by the compound set.

5.3.2 Antiestrogenic Space

Thirty-five known antiestrogens were extracted from literature and converted to SMILES strings using ACD/Chemsketch v8.17 [5]. OMEGA [6] was employed to generate 3D representations of each. The ZINC dataset [7] was downloaded as SMILES format and filtered using FILTER (filter_light) [8] to reduce the set to a subset containing 'drug-like' molecules. Compounds containing toxic and reactive functionalities, as well as those with poor bioavailability were removed accordingly. Of the remainder, 35,000 molecules were randomly selected using a Perl script and converted to 3D format using OMEGA. The full set including ERα actives were imported into MOE [9] and 145 2D molecular descriptors describing atomic nature, molecular size, polarity, fleixibility, lipophilicity etc., were calculated for all. The Molinspiration [10] implementation of XLogP, called miLogP, was utilized to calculate the LogP of the set as it was deemed to be more accurate. As above in section 5.2.1 PCA analysis was carried out on the set.

5.4 Results and Discussion

5.4.1 Cancer Space

Table 1 illustrates the nature of cancer compound space considered in this study. In a total of 12,714 verified-active anticancer compounds extracted 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), and the NCI activity database (GI_{50}>6), 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, eg. the nitrogen mustards that make up a large proportion of the available 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 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 1).

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).

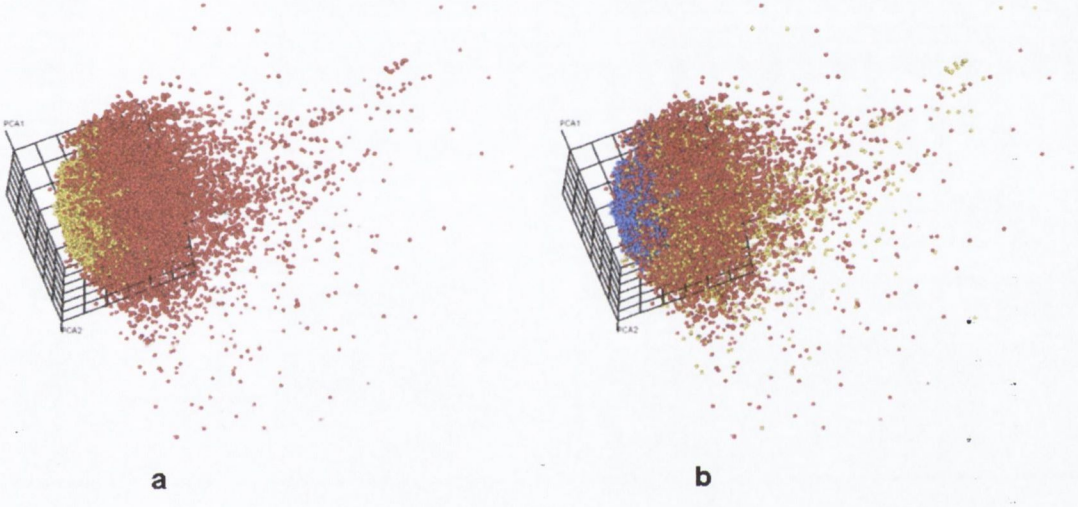


Fig **(1)** Charting cancer medicinal chemistry space. (a) Key: yellow sphere, generic hit space; red spheres, medicinal chemistry space. (b) Key: blue spheres, generic hit space; red spheres, cancer-inactive medicinal chemistry space; yellow spheres, cancer-active medicinal chemistry space.

Figure 1a 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 1b 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 [11], the breadth of cancer space spanned by actives is considerable (Figure 2). 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.

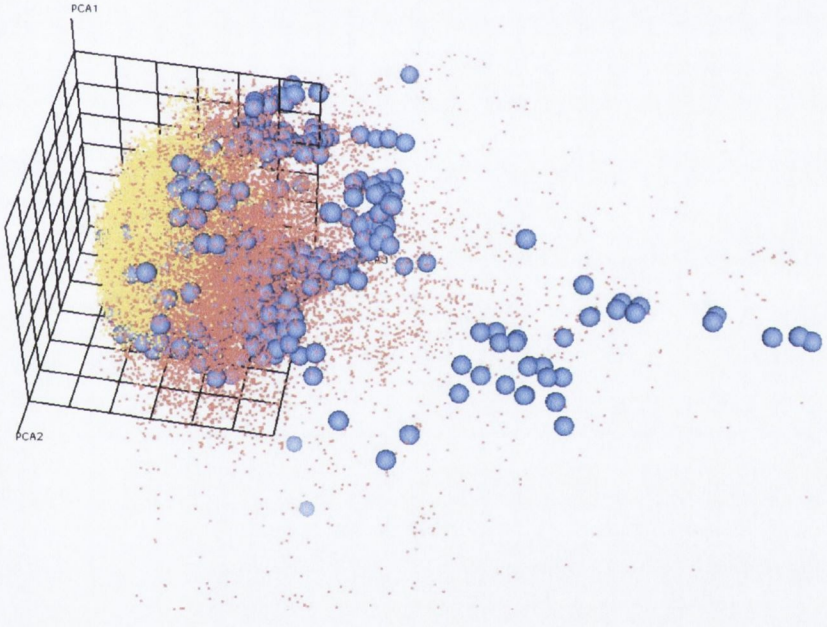


Fig **(2)** Anticancer kinase-targeted space. Key: blue spheres, active ligands; yellow cloud, generic hit space; red cloud, non-kinase targeted medicinal chemistry space.

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, but 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 which explore target relevant chemical space – the caveat

here is the need for unambiguous target information to drive selection of those regions of space that are to be explored.

In Figure 3 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.

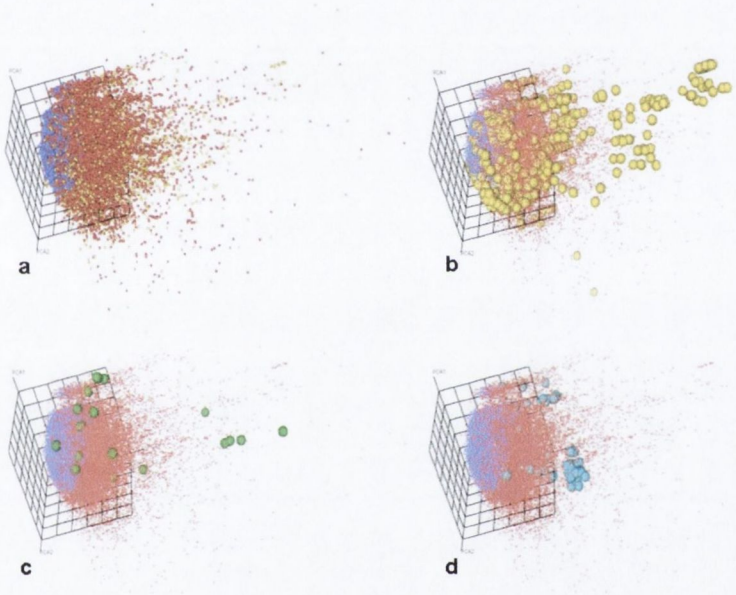


Fig **(3)** Charting cancer medicinal chemistry space. (a) Key: yellow spheres, all cancer actives; red spheres, cancer inactives; blue spheres, traditional hit-like space (indeterminate activity – passes FILTER). (b) Key: yellow spheres, cancer actives known to target kinases; red cloud, cancer medicinal 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.

5.4.2 Antiestrogenic Space

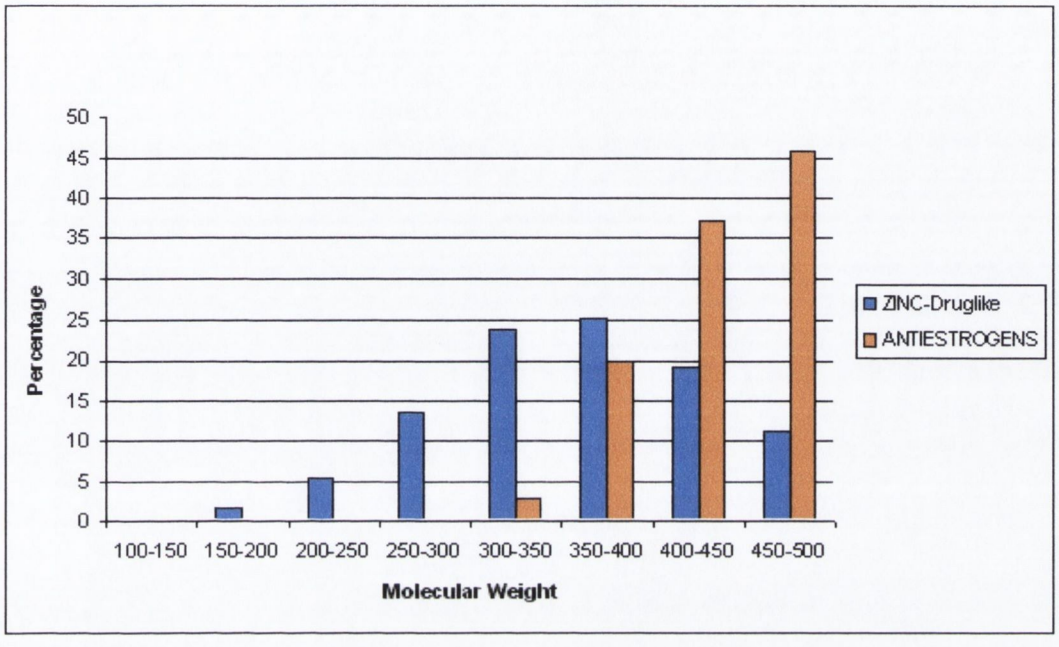


Fig **(4)** Histogram of Molecular weight calculated for Zinc 'drug-like' set and antiestrogen active set.

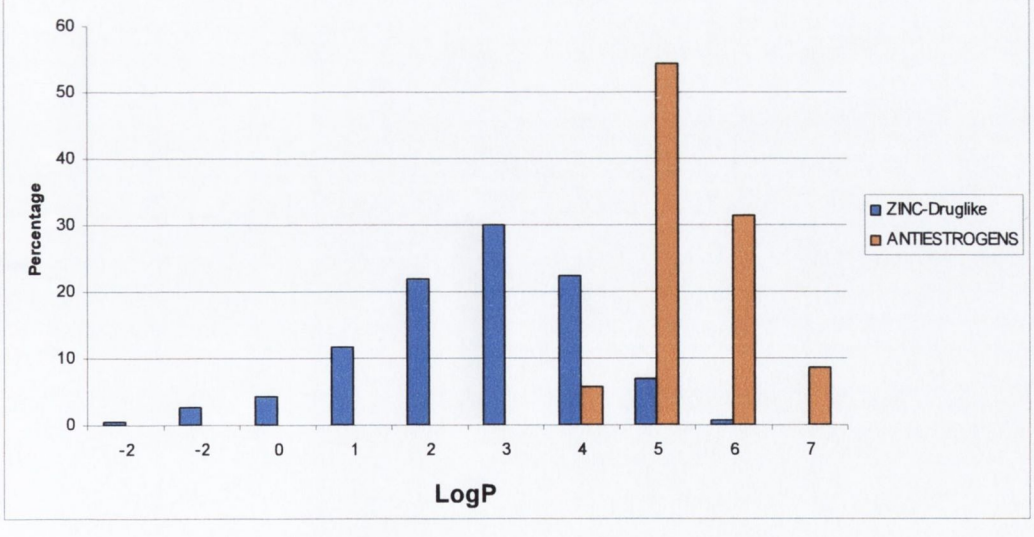


Fig **(5)** Histogram of LogP calculated for Zinc 'drug-like' set and antiestrogen active set.

It is clear from the molecular weight histogram for the two sets that most antiestrogens occupy a heavier weight range than the Gaussian distribution observed with the drug-like set (Figure 4 & 5). Similarly, it is shown that the LogP of antiestrogens differs largely from that of the drug-like set. The LogP of the antiestrogen and Zinc drug-like set was calculated using the Molinspiration implementation of XLogP, called miLogP [10]. Of the 35 actives, the LogP range was ~ 4-9 with only two of the actives having a LogP lower than 5. This equates to ~95% of the dataset not passing Lipinski's rules. On the contrary, the distribution of hydrogen bond acceptors and donors from Figures 6 and 7 appear to be manifestly equivalent to the drug-like set.

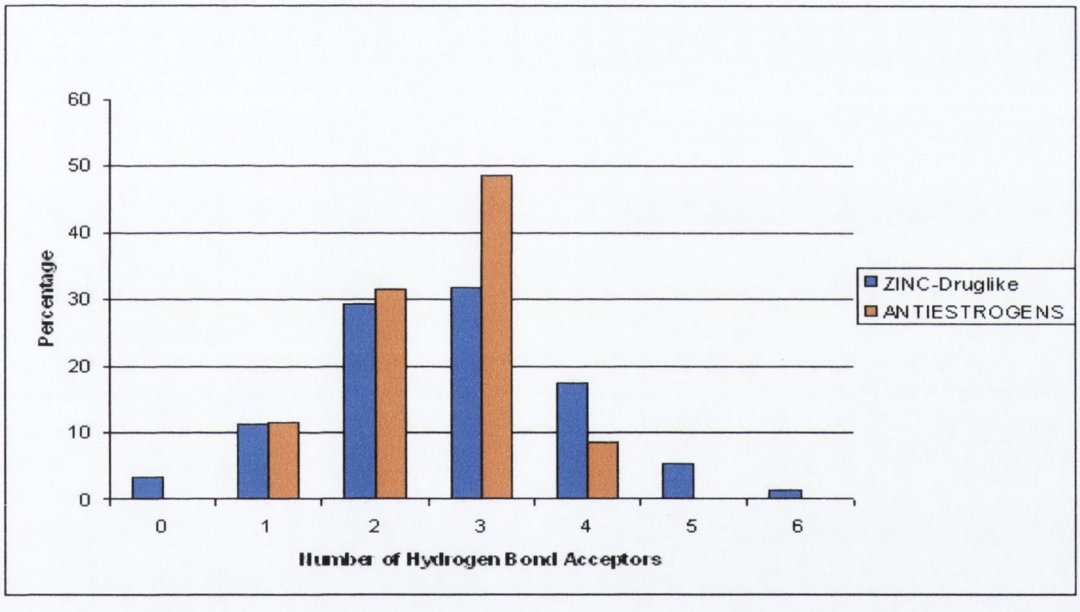


Fig **(6)** Histogram of Number of H-bond acceptors calculated for Zinc 'drug-like' set and antiestrogen active set.

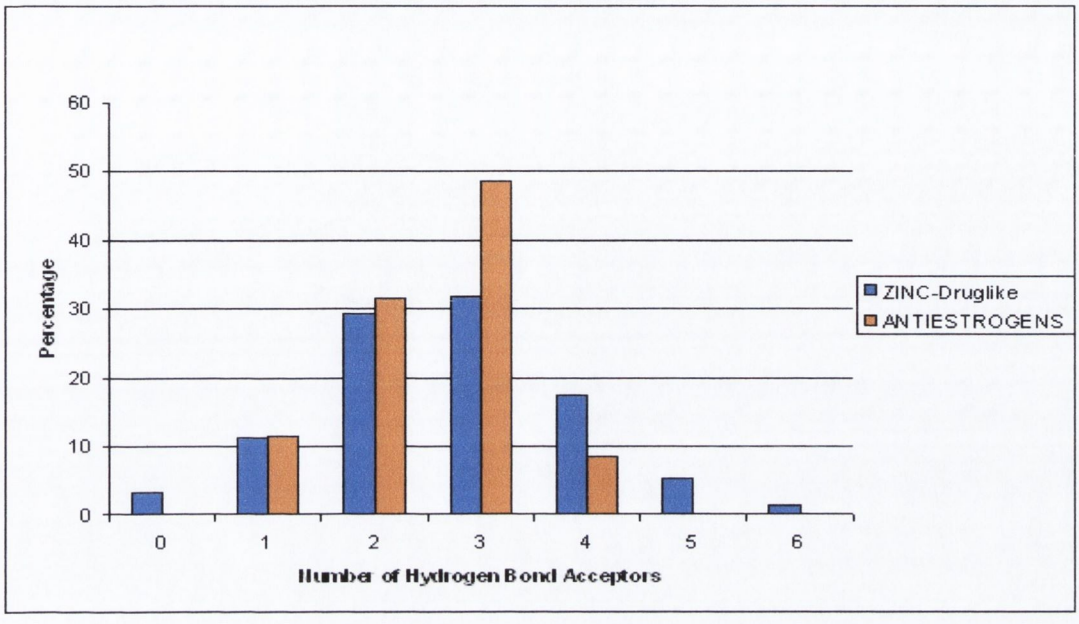


Fig (7) Histogram of Number of H-bond Donors calculated for Zinc 'drug-like' set and antiestrogen active set.

Using Principal Component Analysis (PCA) analysis of 145 2D calculated structural and physiochemical descriptors (See Appendix A), the chemical space occupied by the set of 35 antiestrogens in relation to the drug-like subset extracted from ZINC database [7] was examined as illustrated in Figure 8. This analysis was carried out to allow a clear visual differentiation between the occupancies of chemical space of the 35 actives and the drug-like subset. Corroborating Figures 4-7 it is clear that only two of the thirty-five antiestrogens occupy drug-like space. This demonstrates the need to understand properties of any known actives or 'binders' of a target in order to determine the correct filters to apply prior to screening and potentially increase the 'hit' rate.

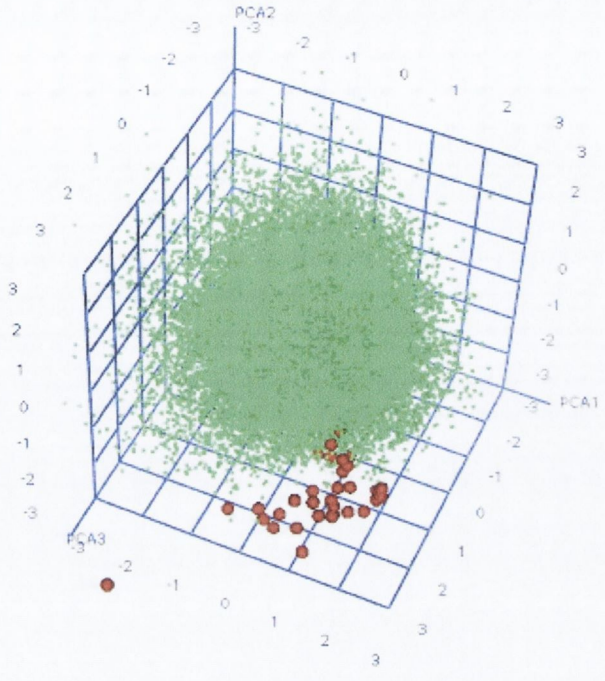


Fig **(8)** PCA analysis of 145 descriptors calculated using MOE. Zinc 'drug-like' set is shown in green and antiestrogens are shown in red.

5.5 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. Moreover, we have shown that known antiestrogens generally do not occupy 'drug-like' space due to their increased molecular weight and lipophilicity. It is highly important to incorporate this in the filtering process and not generically apply a 'drug-like' filter with default settings.

Thus, 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 that don't always need to be applied.

5.6 ER Second Binding Site Hypothesis - Abstract

The recent discovery of non-genomic mechanisms of the ER and publication of a model involving a recently identified alternative binding-pocket of the ER are related through cavity analysis suggesting how the same receptor can invoke these 'classical' and rapid responses concurrently.

5.7 Introduction

Existence of a membrane ER

Several recent studies have converged on the idea that in addition to their role as direct regulators of gene transcription mediated by 'classical' nuclear hormone receptors, numerous non-genomic pathways are also mediated through steroid hormone receptors [12-17]. A description of the numerous virtual screening possibilities associated with the design of new ligands to assist in the prevention of breast cancer through interaction with membrane ERs is provided.

Recently a more complete picture is beginning to emerge of the full extent of ER signalling, with the recent association of rapid non-genomic signalling events and the existence of a membrane ER that binds E_2 (Estradiol) [18, 19]. Breast cancer cells (eg. MCF-7) are now known to express both membrane ER (mER) and nuclear ER (nER) [20-22] and both been shown to exist as dimers [20]. The extent of plasma bound ER and mechanism of binding is controversial, however, interactions with other membrane bound proteins such as Caveolin-1 [23], IGF-1 [24], MNAR [25, 26] and Shc [27] have been well established. Berthois et al (1986) demonstrated the presence of E_2 membrane binding sites in MCF-7 cells using E_2-BSA conjugates. Levin points out in recent review that E_2-BSA conjugates are unstable in many instances allowing dissociation into free E_2 and BSA [28] and although it has been used to show the existence of plasma ER, BSA can also bind on its own to caveolae. This suggests that E_2-BSA may not be suitable for ER membrane studies [19], however results obtained recently by Monje et al indicated the presence of a population of ERα localized at the plasma membrane using fluorescent estrogen-BSA derivatives and it was noted that the existence of free E_2 could not account for the results obtained [19, 29]. These receptors were shown to face the extracellular media in MCF-7 cells. Marquez

et al [30] also confirmed the presence of a significant amount of specific high affinity E_2 binding mER, using controlled cell homogenisation and fractionation of MCF-7 cells. These studies have allowed us to verify that mER exists, and that they elicit rapid cellular effects. Both membrane forms of ER have been identified to originate from a single transcript [31] and can elicit nuclear function in certain cases [32], however further characterization of mER isoforms will be required to elucidate how to separately modulate these processes. Identification of compounds, both synthetic and natural, that can bind either the mER or nER, may be key to addressing issues such as those observed with SERMs. For example, Kousteni et al revealed a compound of this nature (4-estren-3-alpha, 17beta-diol) that could restore and maintain bone density without exerting any 'classical' estrogenic effects [33] indicating a new pathway whereby no interaction with the genomic pocket (classical binding site) of the ER has occurred.

Requirements for ER membrane localisation

Full-length mERα is not required to mediate the range of actions produced at the membrane and downstream. When only the E-domain of mERα is expressed and targeted to the membrane, full signaling is still maintained. There are several structural requirements that have been recently studied that allow innate mERα signaling. Razandi et al found that serine 522 was a critical AA required for full localization at the membrane in MCF-7 cells. Mutation of S522 to alanine (S522A) not only reduced membrane translocation, but also consequently inhibited ERK, cyclin D1, cdk 4 signaling and G_1/S phase progression [34]. Acconcia et al demonstrated that ERα undergoes palmitoylation at Cys447, and mutation of this AA to alanine prevents this palmitoylation [35]. Moreover, this inhibition prevents association with Caveolin-1 at the membrane and its respective non-genomic activities. Also, E_2 stimulation results in a reduction in palmitoylation and Caveolin-1 association with mER [36]. It has been shown that ligand bound ER LBD protects Cys447 from derivatization either due to the direct interaction of the ligand or from the conformation that the protein adopts upon ligand binding [37].

Alternative binding pocket of the ER

Crystallographic data of the ER over the past few years has allowed identification of single hormone-binding site for either agonists or antagonists. The antagonists' tamoxifen, and its metabolite 4-hydroxy-tamoxifen, have shown unusual characteristics in that they behave as agonists in both uterus and breast cancer (MCF-7) cells initially at low concentrations but revert to antagonists at higher doses. Hedden et al proposed a model of the ER that incorporates two recognition sites to explain the atypical behaviour [38].

Hedden indicates that the actual mechanism of antagonism of the ER is due to further interaction with a second site that is not recognised by the endogenous ligand. Two separate enzyme immunoassays found that antiestrogens, both steroid and non-steroid, caused exposure of an epitope for the marker antibody, H222 [38, 39]. This additional epitope was exposed only upon addition of antiestrogen, but more importantly only occurred at a site of the ER not recognised by the endogenous ligand. This suggests an alternative site for binding of antiestrogens and may explain the unusual nature of action of tamoxifen and 4-hydroxytamoxifen. These studies also showed that with ER pre-incubated with estradiol, addition of low antiestrogen concentrations significantly increased H222 epitopes. The concentrations of antiestrogens were too low to compete with estradiol for the primary binding site and indicate that another region of the ER must be occupied by antiestrogen. The two-site model proposed by both Hedden and Martin depicts agonist activity occurring when antiestrogens bind at the estradiol site (P) and antagonist action is mediated by occupation of alternative site (A). Comparing the binding of OHT and RU 58668 with estradiol to the ER at different concentrations, it is observed that estradiol binding eventually saturates when all the primary sites are filled, but antiestrogenic binding continues until approximately twice the amount of ligand is bound [38]. This is confirmed by sedimentation patterns of estradiol, OHT, and RU 58668 bound to ER with MCF-7 cytosol showing almost twice as much antiestrogen bound over endogenous ligand. Katzenellenbogen et al [40] showed previously that OHT behaves as an agonist at low concentrations and antagonist at higher ones. The conversion from agonism to antagonism occurs at approximately the same concentration that additional antiestrogen binding becomes evident.

Several other studies have addressed the probability of a second binding site of the ER to account for the puzzling mode of action of some ER ligands. Trilostane (3beta-hydroxysteroid dehydrogenase inhibitor) has been shown recently to act synergistically but also non-competitively with estradiol in ERβ only, suggesting the presence of a second binding site in ERβ that is distinct from ERα [41]. Addition of trilostane enhances estradiol binding to ERβ possibly by stabilizing the receptor in a conformation that promotes estradiol binding. A second binding site for OHT has been identified in the co-activator region of ERβ and may account for synergistic trilostane binding [42]. It has also been shown that OHT can act synergistically with E_2 in a yeast-transcription assay [43]. Kohler et al observed this also using a genetic, non-transcriptional binding assay and their results suggested that it is only seen at the level of ligand binding [44]. Tyulmenkov and Klinge reported that tetrahydrochrysene Ketone binds to both E_2-liganded and unliganded ERα and ERβ [45]. This indicated the interaction of the chrysenes with an alternative pocket in the ER, and van Hoorn later reconciled this theory using computational methods [46]. Docking studies revealed the primary site is more favourable for binding of all chrysene derivatives. The author concludes that in the presence of both E_2 and chrysene, E_2 is predicted to bind only at the primary site (P) with chrysene competing for this site. The balance of the chrysene will bind to the second site (A) and accounts for the observation of both competitive and uncompetitive binding. Van Hoorn concludes that the (A)-site is an evolutionary remnant with no apparent function by comparing the ER with other Nuclear Receptors (NR) such as RARγ.

5.8 Results

In the present work, to corroborate these findings a set of computational studies using number of crystal structures of the ER was examined, designed to explore the properties of the second ligand binding site. An alternative pocket (A) was identified in the ER for all available (23) crystal structures of both ERα and ERβ with a bound ligand, using Q-SiteFinder [47]. Figure 9(a)-(d) depicts the alternative pocket using ERα only.

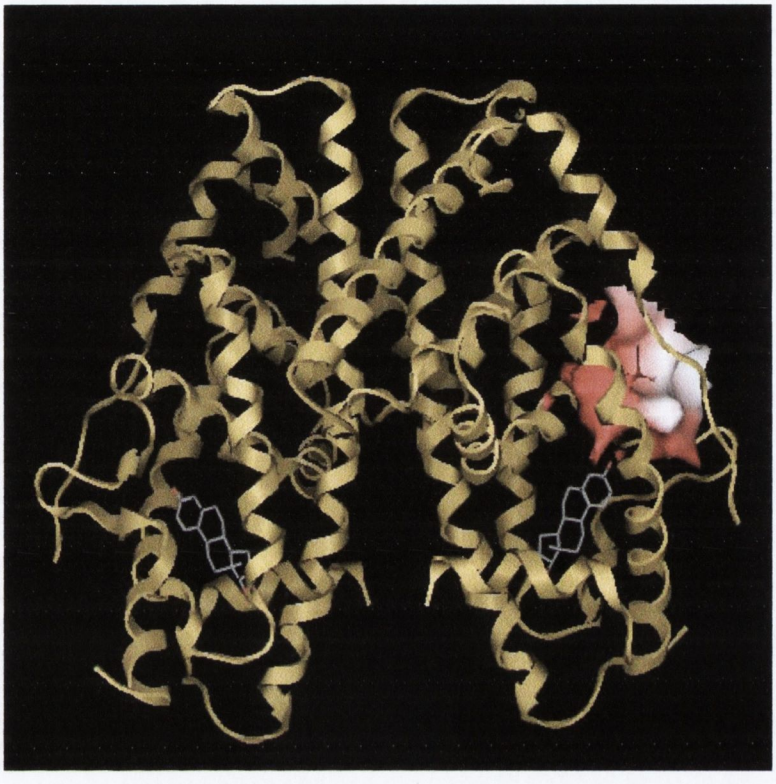


Fig **(9a)** ERα Dimer with pocket (A) illustrated (red)

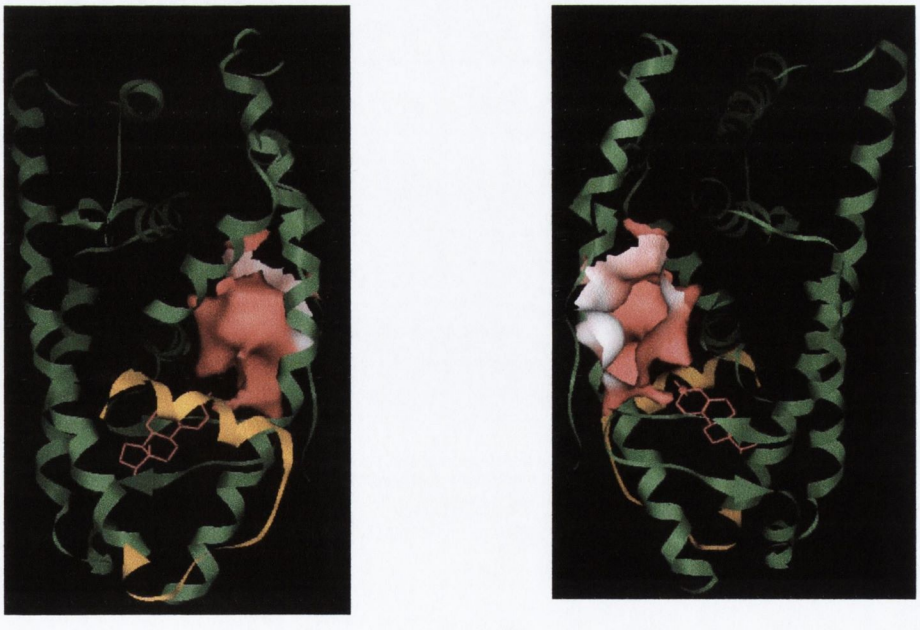


Fig **(9b)** Helix-12 is shown in orange with the alternative pocket (A) exposed in red (1ERE)

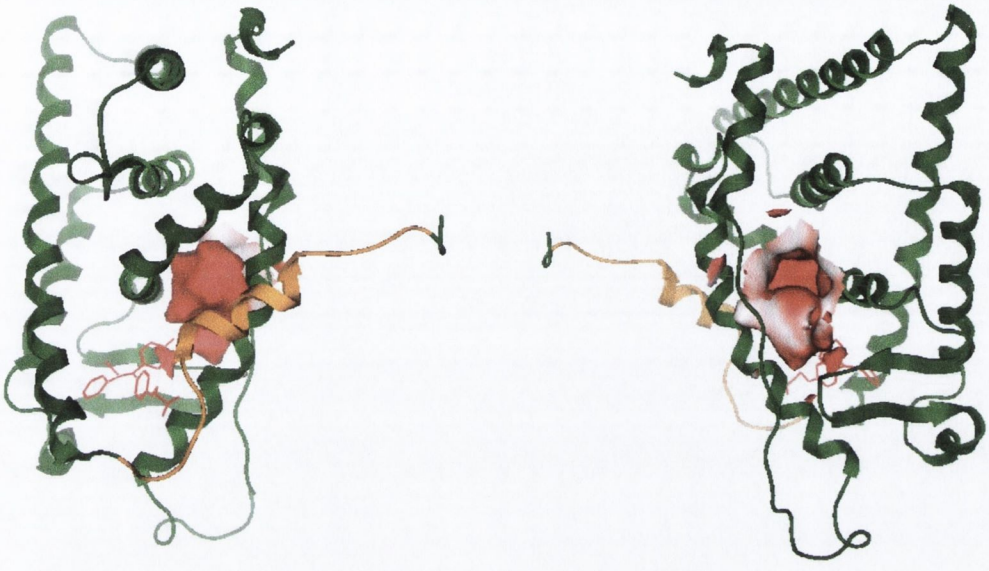


Fig **(9c)** Pocket (A) observed in X-ray 3ERT (stereo view)

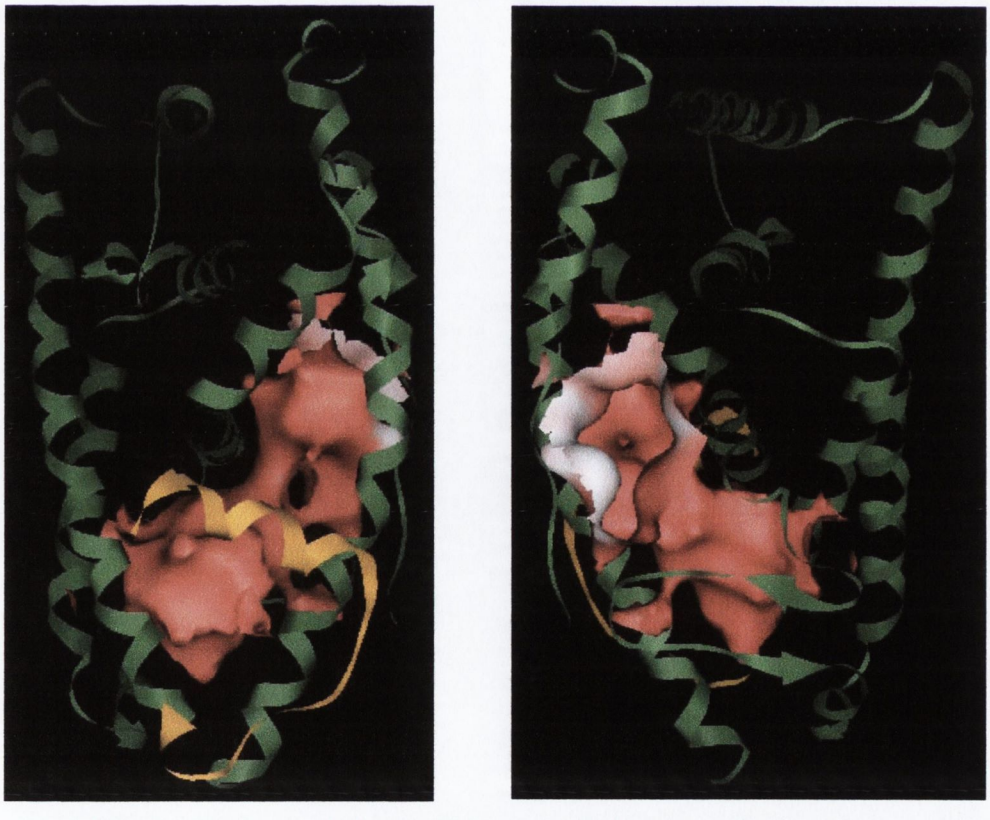
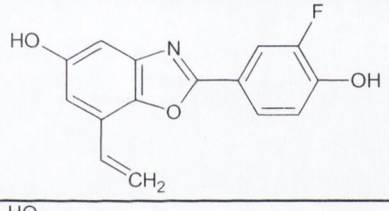
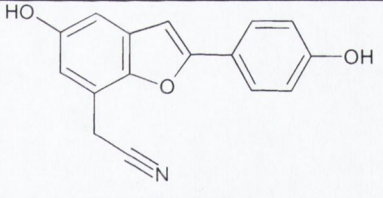
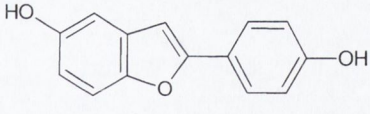


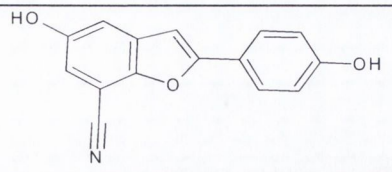
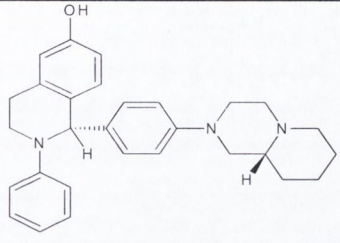
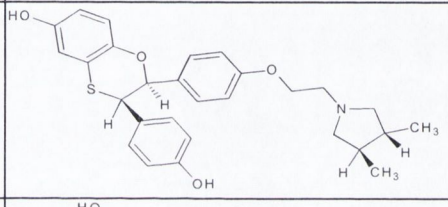
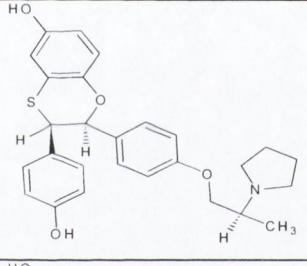
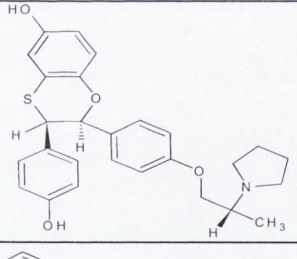
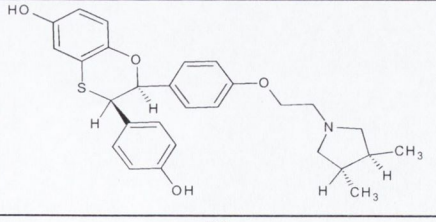
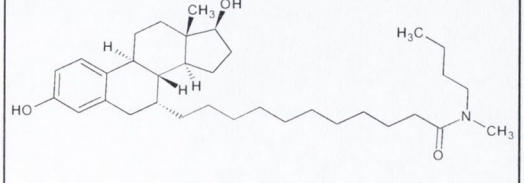
Fig **(9d)** Pocket (A) observed in X-ray 1ERE with classical pocket depicted also (stereo view)

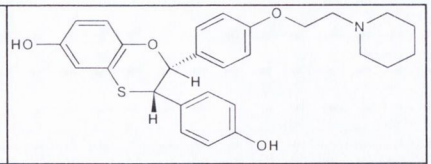
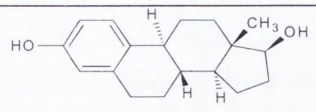
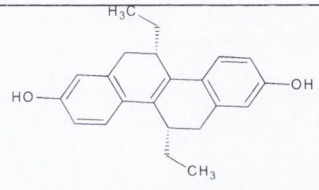
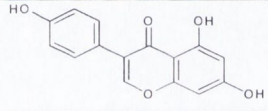
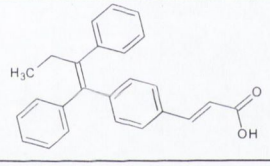
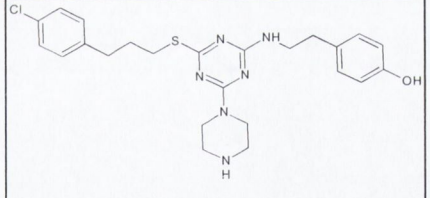
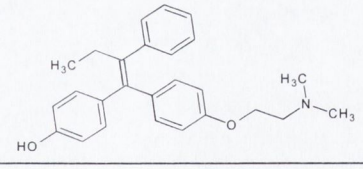
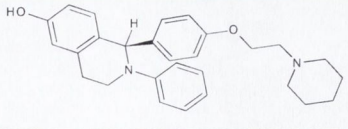
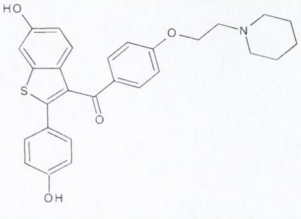
It is clear from Figures 9(a)-(d) that the non-classical pocket A exists in both forms of ER (antagonists & agonist). It is also evident that there are two separate openings to allow a molecule into the two different pockets. It is suggested that one molecule may not migrate from one pocket to another within the ER but rather enters from either side to occupy them separately.

Table 2 depicts all of the PDB entries utilized in the analysis. The set used represent a wide degree of diversity in the ligands co-crystallised and only those structures that were well resolved (≤3Å) were chosen.

Table **(2)** PDB entries of ERα/β with bound ligands illustrated.

PDB_ID	LIGAND	LIG_VOL	STRUCTURE
1X7B	41	228	
1X78	244	239	
1U9E	397	206	

1X76	697	224	
1XQC	AEJ	399	
1XP1	AIH	416	
1XP9	AIJ	410	
1XPC	AIT	398	
1XP6	AIU	418	
1HJ1	AOE	503	

1SJO	E4D	402	
1ERE	EST	242	
1L2J	ETC	308	
1L2I	ETC	295	As Above
1X7J	GEN	229	
1X7R	GEN	229	As Above
1QKM	GEN	224	As Above
1R5K	GW5	337	
1NDE	MON	421	
3ERT	OHT	368	
1UOM	PTI	419	
1QKN	RAL	414	
1ERR	RAL	412	As Above

The effect of ligand size binding to the primary site on the formation of the second binding site was then examined. The volume of (A)-sites for all available crystal structures of ERα with a bound ligand, and the volume of the ligands co-crystallised with each receptor in the (P)-sites were plotted. An inverse relationship was shown to exist as illustrated in Figure 10.

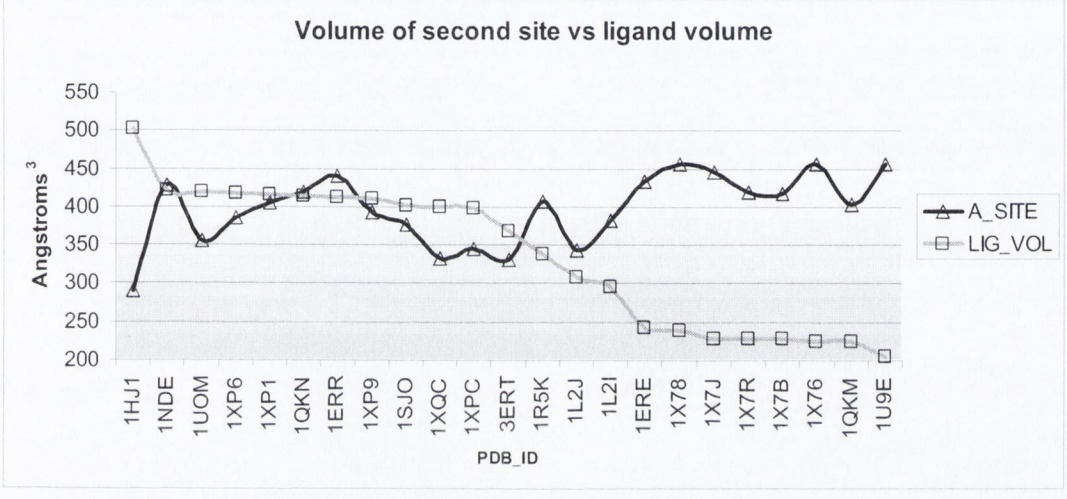


Fig **(10)** Inverse relationship observed between size of the ligand binding to the primary site and volume of (A)-site.

This suggested to us that the likelihood of the (A)-site allowing additional binding as seen with Jensen's studies, would only be possible if the ligand in the (P)-site is relatively small, as with E_2. It is also interesting that the binding pockets of both (A)/(P)-sites of pdb entries 1L2J and 1L2I are similar volumes and may explain the possibility of exchange between sites as seen by Tyulmenkov and Klinge [45]. It is also clear that 1HJ1 with bound ligand ICI 164384 exhibits the smallest volume for the (A)-site due to the sheer size of the ligand. The size and shape of the (A)-site in this case would not allow any additional binding and may be part of the rationale for this ligand being the most potent. The main conclusion that can be suggested from this is that in order to create the A-site, the receptor Helix-12 needs to be in an apo or agonist conformation.

Prompted by a previous successful docking study that resulted in excellent hit rates against the ER, FRED2.0.1 was selected as the docking algorithm to dock 4-Hydroxytamoxifen into the (A)-site of the ER (PDB_ID: 1ERE). The docked position of 4-OHT in the (A)-site of 1ERE was used as a reference position for docking of a validation set of 1000 ligands previously used to characterise the different levels of pre-processing needed to achieve high enrichment rates [48]. The validation set comprised of 19 antagonists of the ERα and 981 'decoys' with pharmacological activity and possessing properties that pass stringent drug-like filters. Enrichment rates of 36.84 in 1% (equivalent to 7/10 'hits' retrieved for ranked database) were observed, illustrating not only the possibility of added binding to the (A)-site when E_2 is bound in the steroid site, but also the shape of the (A)-site is driven towards the binding of estrogen-like ligands. Figure 11 illustrates docking of the top ranked ligand in the second binding site. H-bonding interactions with Glu323, Leu327, Trp393 and Phe404 occur to stabilise the ligands in the alternative-binding pocket.

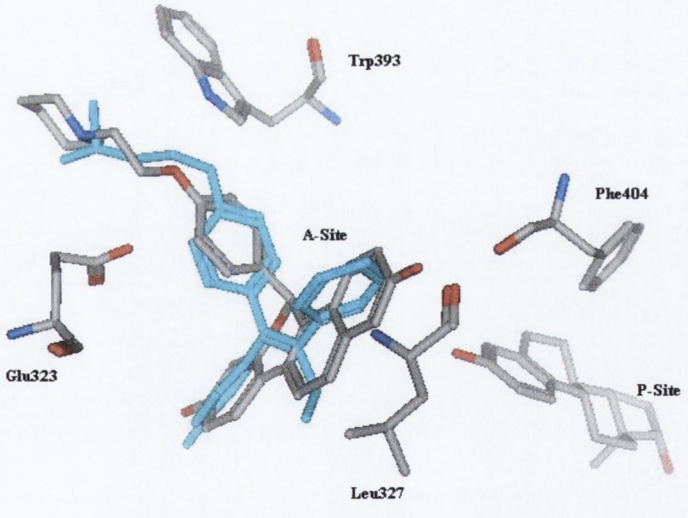


Fig **(11)** Docking of 4-hydroxytamoxifen (cyan) and another ER-antagonist in A-site.

The E rates observed from the ranked database are also very comparable to other methods used to validate docking algorithms as shown in section 1.4. The same method when used only at the (P)-site of 3ERT gives an enrichment of 52.63 in 1% (equivalent to 10/10 'hits' retrieved for ranked database), however the order of the ranked hitlist is different and may offer a partial explanation for the diverse affinities observed with different ligands if they have the opportunity to also bind to another site in the ER.

In conclusion the possibility of additional binding to the ER and steroid hormone receptors is suggested when the (P)-site is occupied by its endogenous ligand. This is in agreement with several studies that show almost double the binding of antagonist over E_2. A more 'apo-like' conformation is implicated in the binding of endogenous ligand at the membrane.

At low tamoxifen concentrations, the ligand may act as an agonist by occupying (A)-sites and this subsequently enhances E_2 binding at the primary site, in agreement with Fishmans' hypothesis [49]. On increasing antagonist concentration OHT behaves as an antagonist by competing with E_2 for primary sites as outlined in Figure 12. Occupation of the (A)-site should not lead to a conformational change that would result in antagonism, and so is in agreement with all crystallographic data published to date on the ER. Design of ligands for the alternative binding site of ERα offers potential for further modulation of the ER.

P-Site	A-Site	Agonism/Antagonism
Estradiol		Agonism
Estradiol	Tamoxifen	Agonism
Tamoxifen		Antagonism

Fig **(12)** Sequence of binding to P-Site and A-Site of the ER.

5.9 Conclusion

A model for the ability of the ER to exert both 'classical' and 'non-classical' effects has been proposed through interaction of ligands with an alternative binding pocket located in the ER. It has also been shown that active known antiestrogens can successfully dock to this site identified by Q-Sitefinder even when the primary site is occupied by its endogenous ligand. Finally the cycle of events leading to occupation of the second alternative site (A) is provided.

The presence of this alternative binding site and sequence of binding to each could be validated experimentally using Isothermal Titration Calorimetry (ITC). This is a thermodynamic technique that measures the heat generated or absorbed by the species involved. By measuring the interaction heats, binding constants (K_i), reaction stoichiometry (n), enthalpy (ΔH) and entropy (ΔS) can be determined. I propose future validative studies involving these techniques to experimentally clarify and unambiguously determine the presence of pocket A in the ER. The experimental confirmation of this second binding pocket would be extremely useful in the design of ligands that could not only bind to this pocket to elicit novel responses but also to ensure that those ligands already known could be tailored to be have enhanced selectivity towards the known 'classical' pocket.

5.10 References

1. Desany, B.; Zhang, Z., Bioinformatics and cancer target discovery. *Drug Discov Today* **2004,** 9, (18), 795-802.
2. 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.
3. Cozzi, P.; Mongelli, N.; Suarato, A., Recent anticancer cytotoxic agents. *Curr Med Chem Anti-Canc Agents* **2004,** 4, (2), 93-121.
4. 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.
5. Chemsketch. v8.17, *www.acdlabs.com*.
6. OMEGA 1.8.1, distributed by Openeye Scientific Software.
7. 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.
8. FILTER, distributed by Openeye Scientific Software.
9. Molecular Operating Environment (MOE), developed and distributed by Chemical Computing Group. *(http://www.chemcomp.com)*.
10. Molinspiration, http://www.molinspiration.com/.
11. Druker, B. J., Imatinib as a paradigm of targeted therapies. *Adv Cancer Res* **2004,** 91, 1-30.
12. Lu, M. L.; Schneider, M. C.; Zheng, Y.; Zhang, X.; Richie, J. P., Caveolin-1 interacts with androgen receptor. A positive modulator of androgen receptor mediated transactivation. *J Biol Chem* **2001,** 276, (16), 13442-51.
13. Boonyaratanakornkit, V.; Edwards, D. P., Receptor mechanisms of rapid extranuclear signalling initiated by steroid hormones. *Essays Biochem* **2004,** 40, 105-20.
14. Bertelsen, E. L.; Endresen, P. C.; Orbo, A.; Sager, G., Non-genomic cell growth inhibition by progesterone. cell cycle retardation and induction of cell death. *Anticancer Res* **2004,** 24, (6), 3749-55.
15. Bassett, J. H.; Harvey, C. B.; Williams, G. R., Mechanisms of thyroid hormone receptor-specific nuclear and extra nuclear actions. *Mol Cell Endocrinol* **2003,** 213, (1), 1-11.
16. Lipworth, B. J., Therapeutic implications of non-genomic glucocorticoid activity. *Lancet* **2000,** 356, (9224), 87-9.
17. Losel, R.; Schultz, A.; Wehling, M., A quick glance at rapid aldosterone action. *Mol Cell Endocrinol* **2004,** 217, (1-2), 137-41.
18. Losel, R. M.; Falkenstein, E.; Feuring, M.; Schultz, A.; Tillmann, H. C.; Rossol-Haseroth, K.; Wehling, M., Nongenomic steroid action: controversies, questions, and answers. *Physiol Rev* **2003,** 83, (3), 965-1016.
19. Kelly, M. J.; Levin, E. R., Rapid actions of plasma membrane estrogen receptors. *Trends Endocrinol Metab* **2001,** 12, (4), 152-6.
20. Razandi, M.; Pedram, A.; Merchenthaler, I.; Greene, G. L.; Levin, E. R., Plasma membrane estrogen receptors exist and functions as dimers. *Mol Endocrinol* **2004,** 18, (12), 2854-65.
21. Aronica, S. M.; Kraus, W. L.; Katzenellenbogen, B. S., Estrogen action via the cAMP signaling pathway: stimulation of adenylate cyclase and cAMP-regulated gene transcription. *Proc Natl Acad Sci U S A* **1994,** 91, (18), 8517-21.
22. Le Mellay, V.; Grosse, B.; Lieberherr, M., Phospholipase C beta and membrane action of calcitriol and estradiol. *J Biol Chem* **1997,** 272, (18), 11902-7.
23. Razandi, M.; Oh, P.; Pedram, A.; Schnitzer, J.; Levin, E. R., ERs associate with and regulate the production of caveolin: implications for signaling and cellular actions. *Mol Endocrinol* **2002,** 16, (1), 100-15.
24. Song, R. X.; Barnes, C. J.; Zhang, Z.; Bao, Y.; Kumar, R.; Santen, R. J., The role of Shc and insulin-like growth factor 1 receptor in mediating the translocation of estrogen receptor alpha to the plasma membrane. *Proc Natl Acad Sci U S A* **2004,** 101, (7), 2076-81.
25. Edwards, D. P.; Boonyaratanakornkit, V., Rapid extranuclear signaling by the estrogen receptor (ER): MNAR couples ER and Src to the MAP kinase signaling pathway. *Mol Interv* **2003,** 3, (1), 12-5.

26. Barletta, F.; Wong, C. W.; McNally, C.; Komm, B. S.; Katzenellenbogen, B.; Cheskis, B. J., Characterization of the interactions of estrogen receptor and MNAR in the activation of cSrc. *Mol Endocrinol* **2004,** 18, (5), 1096-108.
27. Zhang, Z.; Kumar, R.; Santen, R. J.; Song, R. X., The role of adapter protein Shc in estrogen non-genomic action. *Steroids* **2004,** 69, (8-9), 523-9.
28. Stevis, P. E.; Deecher, D. C.; Suhadolnik, L.; Mallis, L. M.; Frail, D. E., Differential effects of estradiol and estradiol-BSA conjugates. *Endocrinology* **1999,** 140, (11), 5455-8.
29. Monje, P.; Zanello, S.; Holick, M.; Boland, R., Differential cellular localization of estrogen receptor alpha in uterine and mammary cells. *Mol Cell Endocrinol* **2001,** 181, (1-2), 117-29.
30. Marquez, D. C.; Pietras, R. J., Membrane-associated binding sites for estrogen contribute to growth regulation of human breast cancer cells. *Oncogene* **2001,** 20, (39), 5420-30.
31. Razandi, M.; Pedram, A.; Greene, G. L.; Levin, E. R., Cell membrane and nuclear estrogen receptors (ERs) originate from a single transcript: studies of ERalpha and ERbeta expressed in Chinese hamster ovary cells. *Mol Endocrinol* **1999,** 13, (2), 307-19.
32. Acconcia, F.; Totta, P.; Ogawa, S.; Cardillo, I.; Inoue, S.; Leone, S.; Trentalance, A.; Muramatsu, M.; Marino, M., Survival versus apoptotic 17beta-estradiol effect: role of ER alpha and ER beta activated non-genomic signaling. *J Cell Physiol* **2005,** 203, (1), 193-201.
33. Kousteni, S.; Chen, J. R.; Bellido, T.; Han, L.; Ali, A. A.; O'Brien, C. A.; Plotkin, L.; Fu, Q.; Mancino, A. T.; Wen, Y.; Vertino, A. M.; Powers, C. C.; Stewart, S. A.; Ebert, R.; Parfitt, A. M.; Weinstein, R. S.; Jilka, R. L.; Manolagas, S. C., Reversal of bone loss in mice by nongenotropic signaling of sex steroids. *Science* **2002,** 298, (5594), 843-6.
34. Razandi, M.; Alton, G.; Pedram, A.; Ghonshani, S.; Webb, P.; Levin, E. R., Identification of a structural determinant necessary for the localization and function of estrogen receptor alpha at the plasma membrane. *Mol Cell Biol* **2003,** 23, (5), 1633-46.
35. Acconcia, F.; Ascenzi, P.; Fabozzi, G.; Visca, P.; Marino, M., S-palmitoylation modulates human estrogen receptor-alpha functions. *Biochem Biophys Res Commun* **2004,** 316, (3), 878-83.
36. Acconcia, F.; Ascenzi, P.; Bocedi, A.; Spisni, E.; Tomasi, V.; Trentalance, A.; Visca, P.; Marino, M., Palmitoylation-dependent estrogen receptor alpha membrane localization: regulation by 17beta-estradiol. *Mol Biol Cell* **2005,** 16, (1), 231-7.
37. Hegy, G. B.; Shackleton, C. H.; Carlquist, M.; Bonn, T.; Engstrom, O.; Sjoholm, P.; Witkowska, H. E., Carboxymethylation of the human estrogen receptor ligand-binding domain-estradiol complex: HPLC/ESMS peptide mapping shows that cysteine 447 does not react with iodoacetic acid. *Steroids* **1996,** 61, (6), 367-73.
38. Hedden, A.; Muller, V.; Jensen, E. V., A new interpretation of antiestrogen action. *Ann N Y Acad Sci* **1995,** 761, 109-20.
39. Martin, P. M.; Berthois, Y.; Jensen, E. V., Binding of antiestrogens exposes an occult antigenic determinant in the human estrogen receptor. *Proc Natl Acad Sci U S A* **1988,** 85, (8), 2533-7.
40. Katzenellenbogen, B. S.; Kendra, K. L.; Norman, M. J.; Berthois, Y., Proliferation, hormonal responsiveness, and estrogen receptor content of MCF-7 human breast cancer cells grown in the short-term and long-term absence of estrogens. *Cancer Res* **1987,** 47, (16), 4355-60.
41. Puddefoot, J. R.; Barker, S.; Glover, H. R.; Malouitre, S. D.; Vinson, G. P., Non-competitive steroid inhibition of oestrogen receptor functions. *Int J Cancer* **2002,** 101, (1), 17-22.
42. Wang, W., Chirgadze, N.Y., Briggs, S.L., Khan S., Jensen, E.V., Burris, T.P., A second binding site for hydroxy-tamoxifen within the co-activator binding groove of estrogen receptor beta. *Proc Natl Acad Sci U S A* **submitted for publication**.
43. Graumann, K.; Jungbauer, A., Agonistic and synergistic activity of tamoxifen in a yeast model system. *Biochem Pharmacol* **2000,** 59, (2), 177-85.
44. Kohler, F.; Zimmermann, A.; Hager, M.; Sippel, A. E., A genetic, non-transcriptional assay for nuclear receptor ligand binding in yeast. *Gene* **2004,** 337, 113-9.
45. Tyulmenkov, V. V.; Klinge, C. M., Interaction of tetrahydrocrysene ketone with estrogen receptors alpha and beta indicates conformational differences in the receptor subtypes. *Arch Biochem Biophys* **2000,** 381, (1), 135-42.
46. van Hoorn, W. P., Identification of a second binding site in the estrogen receptor. *J Med Chem* **2002,** 45, (3), 584-9.
47. 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.

48. 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.
49. Fishman, J. H., Estradiol and tamoxifen interaction at receptor sites at 37 C. *Endocrinology* **1983,** 113, (3), 1164-6.