



# **A new ligand-based approach to virtual screening and profiling of large chemical libraries**

Elisabet Gregori Puigjané

Memòria presentada per optar al grau de Doctor en Biologia per la  
Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del  
Dr. Jordi Mestres al Departament de Ciències Experimentals i de la  
Salut de la Universitat Pompeu Fabra

Jordi Mestres

Elisabet Gregori Puigjané

Barcelona, Maig 2008

The research in this thesis has been carried out at the Chemogenomics Laboratory (CGL) within the Unitat de Recerca en Informàtica Biomèdica (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB).



The research carried out in this thesis has been supported by Chemotargets S. L.







# Table of contents

<b>Acknowledgements</b> .....	<b>3</b>
<b>Abstract</b> .....	<b>5</b>
<b>Objectives</b> .....	<b>7</b>
<b>List of publications</b> .....	<b>9</b>
<b>Part I – INTRODUCTION</b> .....	<b>11</b>
Chapter I.1. Drug discovery .....	13
I.1.1. Obtaining a drug candidate .....	14
<i>I.1.1.1. Hit identification</i> .....	15
<i>I.1.1.2. Hit-to-lead</i> .....	17
<i>I.1.1.3. Lead optimization</i> .....	19
I.1.2. Beyond the one drug – one target paradigm .....	20
Chapter I.2. <i>In silico</i> drug discovery .....	25
I.2.1. Classification schemes .....	26
I.2.2. Annotated chemical libraries .....	28
I.2.3. Molecular descriptors .....	29
I.2.4. Quantitative structure-activity relationships .....	32
I.2.5. Virtual ligand screening .....	34
I.2.6. Virtual target profiling .....	35
I.2.7. Chemical library design .....	36
I.2.8. Network pharmacology .....	37
Chapter I.3. Conclusions and Outlook .....	41
<b>Part II – DISCUSSION</b> .....	<b>43</b>
Chapter II.1. Future directions of research .....	47
II.1. Chemical structure hopping .....	47
II.2. Systems chemical biology .....	48

<b>Part III – PUBLICATIONS .....</b>	<b>51</b>
Chapter III.1. Molecular descriptors and ligand-based virtual screening .	53
Chapter III.2. Virtual target profiling .....	75
Chapter III.3. Targeted library design .....	99
Chapter III.4. Network pharmacology .....	133
Chapter III.5. ViSCA .....	156
<b>Conclusions .....</b>	<b>163</b>
<b>References .....</b>	<b>165</b>

# Acknowledgements

I would like to thank all those who contributed professionally to this thesis, very specially to my supervisor Jordi Mestres who has been a great teacher and very supportive all the way until here. I would also like to thank my mates in the Chemogenomics Lab (Rut, Lulla, Montse, Ricard, Xavi, Ferran, Miguel Ángel, Praveena and Núria) and in Chemotargets (David). I also thank Ramon Aragüés and Fabien Fontaine who were very patient and helped me a lot at the beginning of my thesis. Last, but not least, many thanks to those who helped me at the beginning of my scientific career and that made me realize I wanted to do research: Jordi Villà, Isma Zamora and Manolo Pastor.





The representation of molecules by means of molecular descriptors is the basis of most of the computational tools for drug design. These computational methods are based on the abstraction from the chemical structure to summarize its relevant features while being efficient in the comparison of large molecule libraries. A very important feature of these descriptors is their ability to capture the information relevant for the interaction with any target independently from the scaffold of the compound. This will allow detecting as similar any two compounds with the same features arranged in the same way around essentially different scaffolds, a property referred to as scaffold hopping. With this in mind, a new set of descriptors based on the distribution of atom-centred pharmacophoric feature pairs by means of the information theory concept of Shannon entropy [1], called SHED, have been developed.

These descriptors have been successfully used in a number of applications important in the drug discovery process. After the implementation of novel *in vitro* technologies like high-throughput screening and combinatorial chemistry, the capacity of synthesizing and testing compounds increased exponentially but the need for a rational selection of the compounds arose as well. The prioritisation of compounds in terms of their predicted chances of displaying the targeted activity is thus one of the first applications of the ligand-based virtual ligand screening based on SHED descriptors. This application has shown very good results, both in terms of enrichment of actives in the hit list and in terms of scaffold hopping ability, i.e. the novelty of the scaffolds of the found actives in the top ranked compounds.

Actually, this methodology can be extended to a chemogenomics view of the drug discovery process, using the descriptors to build ligand-based models of all the proteins with any ligand information. This broader approach, the virtual target profiling, is a step towards completing the activity matrix between all possible chemical compounds and all relevant targets. Moreover, a deeper analysis of this complete matrix generated by virtual target profiling can lead us to a network pharmacology perspective of the drug discovery process. This direction can be further followed by adding to ligand-target information the information about pathways and systems approaches, leading to a systems chemical biology approach that could help understanding biological processes as a whole and identifying more rationally novel and promising drug targets



The main objectives of this PhD thesis can be summarized as follows:

1. To develop a new set of topological, feature-based descriptors (SHED).
2. To develop a ligand-based approach to *in silico* chemical screening and target profiling by exploiting pharmacological data extracted from bibliographic sources and stored in annotated chemical libraries.
3. To implement a new approach to design chemical libraries directed to entire protein families.
4. To further exploit ligand – protein information to analyse and understand biological processes in a systems-directed approach.

The first objective was addressed by implementing a set of descriptors based on Shannon entropy called SHED (SHannon Entropy Descriptors) (see **Chapter III.1**). The second objective was pursued by developing a ligand-based method for small molecule – protein activity prediction (see **Chapter III.1** and **Chapter III.2**). The third objective was tackled with the application of SHED-based models to the prediction of the target family profile for large compounds libraries and the selection of those compounds with the most appropriate profile for the project of interest (see **Chapter III.3**). The fourth objective consists on using all the previous generated information to provide a global picture that will help understanding the modulation of biological pathways and processes by small molecules (see **Chapter III.4**).



# List of publications

## Articles:

- Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615–1622.
- Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to in silico pharmacology: nuclear receptor profiling.** *J Chem Inf Model* 2006, **46**:2725–2736.
- Gregori-Puigjané E, Mestres J: **Coverage and bias in chemical library design.** *Curr Opin Chem Biol* 2008, [doi:10.1016/j.cbpa.2008.03.015](https://doi.org/10.1016/j.cbpa.2008.03.015).
- Gregori-Puigjané E, Mestres J: **A ligand-based approach to mining the chemogenomic space of drugs.** *Comb Chem High Throughput Screen* 2008 (*In press*)
- Mestres J, Gregori-Puigjané E, Valverde S, Solé RV: **The effect of data completeness on drug-target interaction networks.** *Nature Biotech* 2008 (*In press*).

## Book chapters:

- Gregori-Puigjané E, Mestres J: **Designing chemical libraries directed to nuclear receptors.** In “*Nuclear Receptors as Drug Targets*”, Ottow E, Weinmann H (Eds.), Wiley-VCH: New York. 2008 (*In press*)

## Posters:

- Gregori-Puigjané E, Pastor M, Mestres J: **SDFm: an open-source molecular database manager.** *EuroQSAR 2004: 15th European symposium on quantitative structure-activity relationships and molecular modelling*, Istanbul, Turkey, September 5-10 2004.
- Mestres J, Gregori-Puigjané E: **SHED: molecular Shannon entropy descriptors from atom-centered feature distributions.** *229th ACS National Meeting*. San Diego, CA, March 13-17 2005.
- Gregori-Puigjané E, Mestres J: **Virtual profiling of 322 drugs on 199 targets.** *New approaches in drug design and discovery: Merging chemical and biological space*. Marburg, Germany, March 26-29 2007.



## Part I - Introduction

This introductory part contains two main chapters. In the first one, an overview on the traditional and new *in vitro* techniques for drug discovery will be provided. The main steps of the drug discovery process will be presented, emphasizing the key points that will be determinant for the overall performance. The main issues of this process identified so far will be discussed, as well as the technological and conceptual alternatives proposed to solve them. Also, the impact of recent technological developments will be assessed and the consequences of their implementation in the overall process.

The second chapter provides an introduction to the diverse *in silico* protocols in use in any of the steps of the drug discovery process and the relationships that are established among these and the *in vitro* techniques. The possible issues highlighted in the first chapter and the proposed *in silico* solutions for these problems will be discussed. On the other hand, the role of computational methods in a more rational approach to drug discovery, their relevance in the compilation of information and the generation of knowledge to understand the process and the help some predictive *in silico* tools can provide in decision making will also be reviewed.

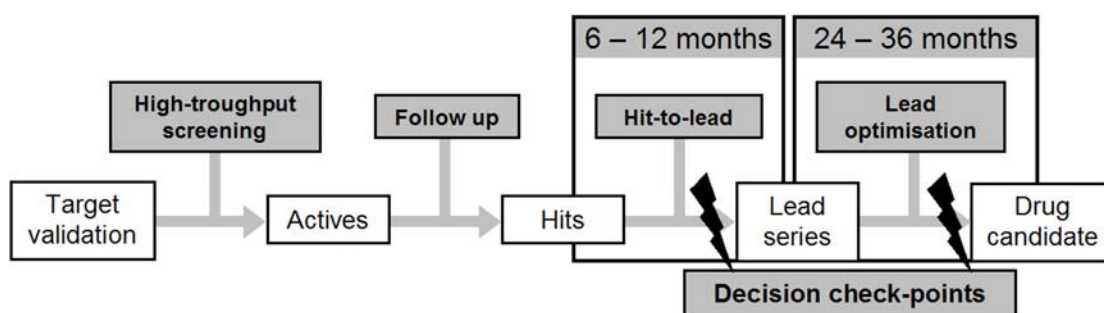




## Chapter I.1 – Drug discovery

Drug discovery is a high-risk crucial process in pharmaceutical industry, as it can result, after many years and big investments, in a successful marketed drug or in total failure. The process of obtaining a drug candidate, which will be discussed in detail in **Section I.1.1**, consists on several steps (**Figure 1**) each of which has different time and technology requirements. After the target has been validated, the first step of this process is obtaining active compounds as defined by fixed criteria in a high-throughput assay. A number of potential issues associated with the high-throughput screening technology make the validation of the obtained actives necessary. This is done in a second step, during the follow-up process. These two steps result in one or several hits, and will be discussed in more detail in **Section I.1.1.1**.

The third step, discussed in **Section I.1.1.2**, is the hit-to-lead process. In this process, structure-activity relationships are established for a number of analogues of the initial hits, and the potency as well as other lead-like important characteristics is optimised. The outcome of this phase is one or more lead series, which have to be selected among all the hit series obtained in the previous steps. The prioritisation according to the likeliness of success of chemical series to be pursued will be key in the final success of the process. This decision-making is an important stepping-stone as the following steps are high time- and resource-consuming. The final step before obtaining of a drug candidate is the lead optimisation, reviewed in **Section I.1.1.3**, in which larger scale assays with emphasis on absorption, distribution, metabolism, excretion and toxicity (ADMET) properties and safety profile are carried out.



**Figure 1.** Drug discovery process scheme.

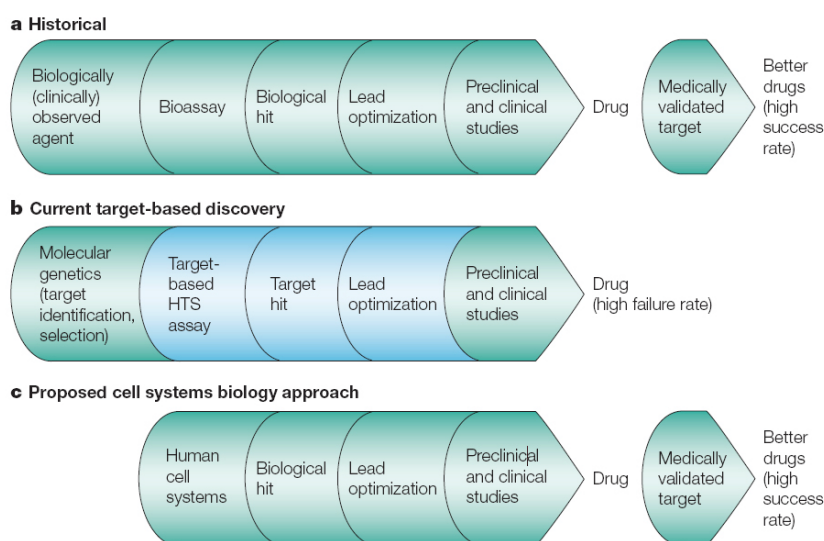
The success rate of this process is around 40%, and only 10% of the resulting drug candidates will finally make it into the market after the clinical testing phase [2]. Moreover, research-based pharmaceutical industry has faced an increase of economic and regulatory pressures with international price controls, rapid appearance of generic products and more stringent regulatory policies [2]. Hence, great efforts towards improving cost-effectiveness and maximizing the probabilities of success have been initiated over the years, leading to a constant

turnover of protocols and paradigms, and to great technological and conceptual advances as will be reviewed in **Section 1.1.2**. Particular focus will be given in how recent technological advances impact the traditional process of drug discovery.

### 1.1.1. Obtaining a drug candidate

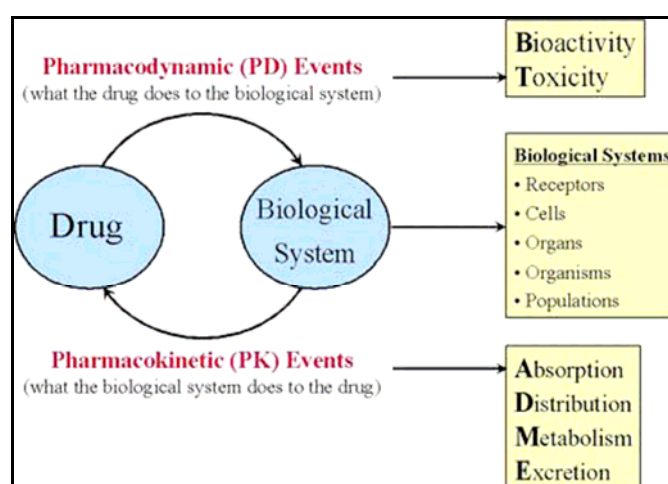
Before the advent of molecular biology, drugs were derived following a long, slow, very much empirical process. *In vivo* models were used as a “black box” to test drug-like or natural compounds for a certain phenotypic effect without knowing its mechanistic basis. After a number of experiences and cases in this so-called forward pharmacology manner, the substance was considered a safe drug. Although this was the first approach to drug discovery, it is still useful when very little or nothing is known about the molecular basis of a process or for complex processes involving whole pathways. Traditionally, these assays were performed on whole organisms, while now it is more common to use cell cultures of unicellular organisms or physiological or pathological tissues.

The implementation of molecular biology into the drug discovery process helped realizing that the effect of these remedies was typically elicited by the binding of an active compound to a single protein, changing completely the focus of the entire drug discovery. In this so-called reverse pharmacology manner, compounds are tested for their activity on a single target in purified protein *in vitro* assays. As illustrated in **Figure 2**, the output of **forward pharmacology, once the protein(s) involved in the phenotype are identified**, can be used in reverse pharmacology to identify new and more potent compounds for the protein target. Similarly, active ligands identified using reverse pharmacology can be biologically validated by examining the phenotypic effect in a forward assay [3].



**Figure 2.** Proposed cell systems biology approach. Extracted from [4].

The final goal of the drug discovery process is to obtain a new chemical entity that has to enter the organism, reach its biological target and elicit the desired effect. In this process, the drug has to reach and pass a number of biological boundaries (oral absorption, membrane permeability, blood-brain barrier) and gets in contact with a large number of proteins other than its theoretical target. Its possible interaction with these off-targets leads to metabolism, when the drug is modified by the interaction with the off-target and can lead to toxic and other undesired side effects, when an off-target path is activated. All these events can be classified in pharmacodynamics (PD), effects on the organism caused by the drug, and pharmacokinetics (PK), effects on the drug caused by the organism, as shown in **Figure 3**. These interactions will determine both the efficacy and the safety of the drug, which are the main reason for which 90% of drug candidates entering clinical development do not reach the market [2,5]. Safety and efficacy have to be taken into account all along the drug discovery pipeline, as will be discussed deeper in **Section I.1.1.2** and **Section I.1.1.3**.



**Figure 3.** Scheme of the interaction between a drug and a biological system and of all the effects triggered by this interaction. Extracted from [6].

### 1.1.1.1. Hit identification

The drug discovery process begins with the identification of one or more active compounds showing activity over a certain threshold in a given assay. Since the advent of reverse pharmacology, these actives are identified on *in vitro* assays in which they are normally tested for a single purified target. This process was dramatically influenced by the development of both combinatorial chemistry (CC) and high-throughput screening (HTS). These two technological advances have appeared to be highly complementary, as CC increased exponentially the capacity for synthesising compounds and HTS increased in the same proportion the ability for testing those compounds on *in vitro* target-based assays.

CC, which takes advantage of miniaturization and parallel synthesis, allows the generation of 100,000s of compounds within several months [7] by assembling building blocks in solution or on solid support [8]. Initially, the focus was set on the number of compounds produced, with little

regard to their quality. Rather, it was assumed that if an active compound could be identified, the possible issues related to its chemical properties could be solved with posterior small chemical modifications. This assumption led to poor initial results as in many cases the activity observed in assays was not reproducible. This was due to a number of reasons, among which the purity of the first compounds synthesized by means of CC is especially remarkable. In some cases, after an active compound was identified, when it was synthesized and characterized by conventional means it turned to be completely inactive. As a result, the early emphasis on the size of mixture-based diversity-driven libraries gave way to parallel synthesis of project-focused libraries of well-characterized discrete compounds, often followed by purification to improve the reliability of the outcome [8,9]. The challenge is now focused on selecting the best compound libraries for synthesis and testing [9], as although the number of compounds in a library is important, the diversity of the chemical structures and the quality of the compounds in the library are even more important.

The implementation of HTS increased dramatically the capacity for testing these large numbers of compounds and thus the ability of exploring the pharmacology of a larger portion of the chemical space. The combination of CC together with HTS implied a revolution in the rate at which these assays were performed and the amount of compounds that could be processed, although the concept of testing each compound for a single target did not change.

HTS technology consists on the testing of large compound collections, which are assayed robotically at a single concentration to one target in 96-, 384- or 1,536-well microtitre plates [10]. A careful design of the HTS assay is key for relevant information to be extracted from it. The nature of the response to be measured, the fact that the response might be stimulus-dependant and the duration of the response have to be taken into account. In the selection of the proper assay format, performance and sensitivity have to be considered and optimised. HTS assays have to be sensitive enough to detect compounds with low potency or efficacy, reproducible and stable among wells and plates, accurate in positive and negative controls and economically feasible [11].

Despite a careful choice of the HTS assay and the improvements in the liabilities formerly associated to CC, the rate of false positives is not neglectable. Some compounds can show activity by acting on mechanisms other than those that are of interest in the project. Other compounds can form aggregates or interfere with the assay signal causing an artifact due to the ability of the compound to mimic the sought effect as a result of its physicochemical properties, mainly fluorescence or absorbance [11]. For this reason, those compounds initially identified as active by HTS need to be confirmed and validated in a follow-up process before being considered hits. The purity and chemical structure of the identified actives have to be checked, as well as the activity through the desired mechanism for hit confirmation. In a second step, hit validation is done by means of lower throughput multipoint activity assays of a selection of diverse representatives of each cluster to confirm their potency. In these assays, half-maximal effective concentrations ( $EC_{50}$ ) for receptor agonists, or half-maximal inhibitory concentrations ( $IC_{50}$ ) for enzyme substrates are determined.

This process ends up, typically, with hundreds of hits from an initial set of thousands actives. The prioritisation of these compounds for hit-to-lead requires a structural clustering of hits into series to provide an easy overview of the different chemical classes that have been identified as being active against the target [12]. Especially in the case of singletons or smaller clusters, there might be a need to search for close analogues to some of them so additional compounds are screened to enable deriving initial structure-activity relationships (SAR) for the corresponding hit series [12].

The resulting most potent validated hits are used for early *in vivo* proof-of-principle studies, in which first pharmacological effects *in vivo* are obtained. Although the targeted activity might be sought through oral administration, these early studies are often realised using intravenous or intraperitoneal administration [13]. Moreover, the formulations used are also often not suitable for later studies or clinical development. Rather than using them solely for target validation, these promising early biological data give such compounds sufficient support for further chemical resourcing, despite the fact that the overall ADMET profile has hardly been investigated so far and might be far from optimal [13].

The integration of CC and HTS allows rapid increases in the size of compound collections and rapid exploration of the structure-activity relationships around chemotypes of interest in medicinal chemistry programs [14]. Despite this, it is clear that the overall performance of the hit identification process will depend highly on the quality of the initial tested compounds. This was already observed in the early phases of CC, leading to a great improvement of the quality of the compounds obtained by this means. Additionally, HTS results are not fully reliable and active compounds identified using this technology have to be confirmed in their chemical structure and validated in their activity. Therefore, the combination of HTS and CC has represented a revolution in the drug discovery process in terms of capacity for testing, but more care needs to be put in the selection of the initial library. As will be discussed in Chapter 1.3, this will be one of the main focuses of the computational methods developed to help in the drug discovery process.

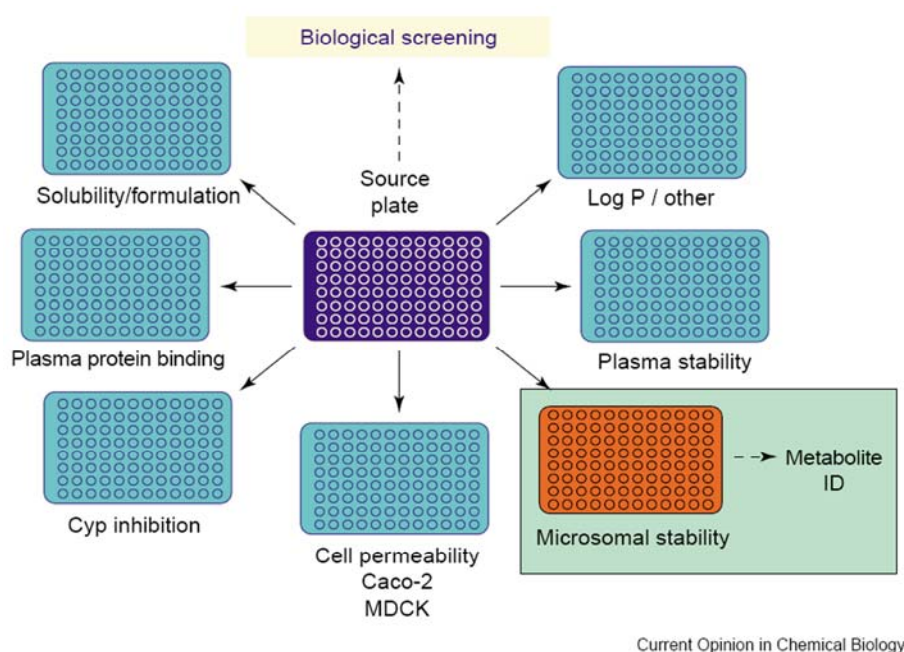
### **1.1.1.2. Hit-to-lead**

Validated hit series selected during hit identification are in this phase chemically modified by iterative synthesis and testing of analogues. All these chemical modifications around a common scaffold are aimed to elucidate SAR to establish consistent correlations of structural features or groups with the biological activity of compounds in a given biological assay. These SAR are aimed to maximize efficacy and potency while keeping adequate ADMET properties and selectivity profile.

Through the late 1980s, medicinal chemistry focused on ligand selectivity and specificity and, although the importance of physicochemical and ADMET properties was acknowledged, they were not given the appropriate relevance in early phases [2]. *In vitro* screening methods for target activity, performed in non-aqueous solvents, often resulted in high affinity compounds that tended to have high molecular weight and lipophilicity and poor water solubility. Later development phases

were then faced with the challenge of fixing the compound's development liabilities, leading to high development time and cost and to high attrition rate due to toxicity, lack of efficacy and poor PK properties [2]. Moreover, possible interactions of the compound with proteins other than the target was not done until a late stage of the drug development process [15]. To solve these drawbacks, the integration of drug discovery and development processes was engaged so that development requirements were taken into account early in the discovery process. In modern drug discovery, potency alone cannot be the driving force behind the prioritisation of hit series to be pursued, as optimisation of ADMET properties is considered to be much more difficult than optimising potency and/or selectivity [13].

Nowadays, the hit-to-lead process can be split in two parts. The first one is based on miniaturised automated *in vitro* assays, achieving the required throughput with minimal compound use [16]. Because the hit set from an HTS campaign often comprises several hundred to several thousand compounds, experimental assessment of the ADMET and physicochemical profile of the entire hit set represents a considerable challenge and for large clusters several representatives based on maximum structural diversity are selected for testing [13]. On the other hand, in order to obtain representative data from these analysis, a purity of at least 80% is mandatory [13].



**Figure 4.** Schematic representation of high-throughput ADME assays. Extracted from [16].

Many ADMET assays can be run in a high-throughput fashion with high sensitivity, selectivity and ease of automation relative to traditional analytical methods, due to the incorporation of liquid chromatography/mass spectrometry (LC/MS) (**Figure 4**). At this point, the most relevant issues for which the representative screening hits are characterised are toxicity and ADMET properties. LC/MS enables the testing for lipophilicity, solubility, metabolic stability, permeability (Caco-2 assay), plasma-protein binding, human ether-a-go-go (hERG) related gene activity, cytochrome P450 inhibition and stability in human liver microsomes [12,16].

The main challenge during hit-to-lead is to make the right decision with the information available at this stage while keeping in mind that these compounds will enter a lead optimisation process in which they need to be further modified. High potency of single hits must not be the main and only issue considered, as increasing the potency of a compound for a target is not considered a major bottleneck in the development process [17]. Instead, there are other issues that are important to be considered at this point. Firstly, hit series containing structurally related compound families are preferred over singletons when available, as they enable to distinguish between problems related to the chemical series and problems only related to some members of the series [12,13]. This is especially relevant in the case of ADMET profile and SAR analysis, where it is important to consider also the inactives. Other important criteria to be considered before a lead series is moved forward into a lead optimisation program are chemical tractability, selectivity, PK properties, demonstrated *in vivo* efficacy and preliminary intellectual property assessment [18]. Finally, ligand efficiency is currently one of the most important aspects in hit-to-lead decision [12]. When efficiency is considered, small, less potent molecules are ranked equally with large more potent compounds.

The outcome of this optimisation process is typically a low number (less than 5) of lead series. Leads are prototypical chemical structures or series of chemical structures that in addition to displaying activity by confirmed specific binding and selectivity in a pharmacological or biochemically relevant screen, do also show emerging SAR for biophysical and ADMET properties [13]. Each distinct lead series has a unique core structure and the ability to be patented separately. In addition, a good lead requires many other properties prior to taking the decision of progressing it through the pipeline: assessing chemical progressability, assessing target selectivity and obtaining *in vivo* proof of principle.

### ***1.1.1.3. Lead optimisation***

With the lead series that result from hit-to-lead, larger scale assays with emphasis on ADMET properties and safety are carried out. This pre-clinical stage of the drug discovery process integrates *in vitro* and *in vivo* pharmacological data to assess undesirable pharmacodynamic effects in humans [19]. In lead optimisation, *in vivo* experiments are essential as its outcome will be a drug candidate that will enter clinical testing and systemic effects have to be addressed. Consequently, the time requirement shifts from around 6 months for hit-to-lead to up to 3 years in the lead optimisation phase.

Early studies using simple assays for selected targets may help eliminate the major causes of a given adverse drug reaction (ADR). On one hand, early testing for the major ADRs saves time and costs by preventing the clinical toxicology pipeline from being truncated with low-quality compounds. On the other hand, a non-selective, large-scale testing for ADRs can slow down the lead optimisation phase. Although traditional toxicology can eliminate the major “zero tolerance” actions of molecules, there could be many other actions that produce minor or even major

“tolerable” side effects [19]. Traditionally, screening for drug safety starts at a relatively late phase of lead optimisation. Any liability discovered at this phase of drug discovery can cause high, late attrition rates associated with escalating costs [19]. However, recent advances have provided pharmaceutical industry with simple, fast and cost-effective *in vitro* screening assays, applicable to the early phases of drug discovery [19].

This type of information can help chemists and biologists to rank and prioritise compounds according to pharmacological profile and to optimise their structures without losing the affinity to the primary target. Once compounds are optimised for an acceptable level of promiscuity in addition to affinity to the primary target and appropriate ADMET properties, teams can make final decisions to promote the best candidate for final risk assessment of safety and toxicology studies, before clinical development.

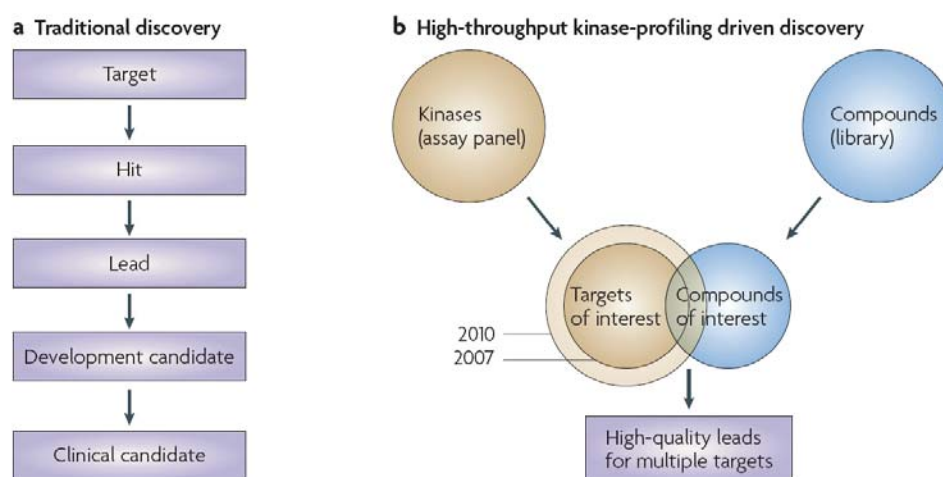
### ***1.1.2. Beyond the one drug – one target paradigm***

Although the “magic bullet” approach seeking compounds binding specifically to a single rationally chosen target has proven to be highly effective in many drug discovery projects, there are certainly a number of other cases for which it presents a number of disadvantages. Recent studies have observed that targeting a single protein can lead to quick resistance in cases like human immunodeficiency virus type 1 therapy [20] or cancer [21]. Furthermore, complex diseases, such as depression [22], inflammation [23] and cancer itself [24], could be modulated more efficiently by targeting several proteins involved in the pathological process [25]. Moreover, it has been proposed that many modern anti-psychotic drugs failed in the clinic because they were too selective for specific targets [26]. Theoretical modelling of biological network structures, also called systems chemical biology [27], predicts that modulating multiple proteins simultaneously is often required to modify phenotype, as some pathways have a certain degree of redundancy and biological systems can often find alternative compensatory routes to single point perturbations [28]. Such network pharmacology approaches are necessary for analysing the consequences of perturbations in physiological pathways by small molecules and for designing multitarget-oriented pharmacological profiles for complex diseases, as will be discussed in **Section 1.2.8** [29].

Central nervous system (CNS) therapeutics has become one of the most profitable sectors of the pharmaceutical market, despite the lack of suitable animal models, disagreements in their biological basis and ineffectiveness of many CNS medications [30]. It has been shown that most of the currently approved atypical antipsychotic drugs have a complex pharmacology, with significant affinities for a variety of aminergic GPCRs [30]. Great efforts have been engaged to elucidate which of these targets are responsible for the therapeutic effects and which are causing the undesired side effects. In this field, the above mentioned systems chemical biology approach can be useful to the mechanistic understanding of all these processes. The discovery and design of the so-called “selectively non-selective” drugs is a challenging issue, as it cannot be faced with the classical *in vitro* medicinal chemistry technology approaches such as HTS. Two ways to face this



problem have been proposed: behaviour- and genomic-based screening [30]. The first one, which can be considered as an *in vivo* forward pharmacology approach, has the drawback of permitting only low throughput rate and the difficulty of having an appropriate animal model, as the polygenic and non-genetic components of CNS diseases make the use of genetically engineered mice difficult. In the second approach, compounds with known functional actions are screened for their effects on coordinated gene expression. These “gene signatures” obtained can contain too many targets, the majority of which may be irrelevant, making posterior rational based filtering based on previous information essential. In both cases, the selection of an appropriate set of compounds to be tested is key, making virtual target profiling approaches, which will be described in **Section I.2.6**, a very useful complementary tool. A third approach based on chemogenomics seems now the most convenient, as, after the desired pharmacological profile is defined thanks to a combination of previous experience and systems chemical biology, the full pharmacological profile of a library of compounds towards as many protein targets as possible could be elucidated and those showing the most promising activities chosen.

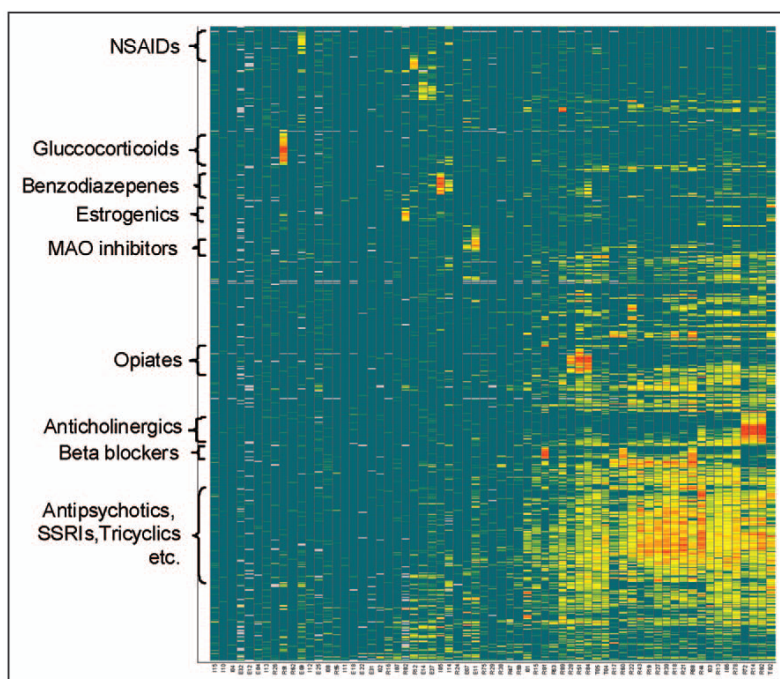


**Figure 5.** Paradigm shift from traditional one drug – one target drug discovery (a) to multitarget-oriented drug discovery (b). Extracted from [31]

Many treatments, such as cancer therapy, are already currently based on multiple medication therapy, which has the drawback of possible drug-drug interaction issues with respect to a single drug with polypharmacological profile. In cancer treatment, polypharmacology is also encouraged by the observation that new protein kinase-targeted drugs have, as in the case of antipsychotic drugs, a higher promiscuity for multiple kinases than initially thought [28]. Recent studies demonstrate that synergistic combined effects of a compound acting on two or more kinases is greater than the additive sum effect of targeting each kinase individually [32].

Consequently, the multitarget profiling of the drug candidates constitutes an important conceptual change in the drug development process, as it extends the available information along the biological axis in the chemistry-biology matrix. This step towards the completeness of the matrix is important both for the identification of the appropriate (single- or multi-target) profile and

for the early assessment of possible toxicity and side effects. The major effort currently available in this direction is BioPrint®. It enables the profiling on a panel of over 630 *in vitro* pharmacological assays for relevant targets and around 130 ADMET assays. These assays are used to identify lead compounds, to define mechanism of action and to identify off-target activities. An example of a subset of the results of this *in vitro* profiling for 2,000 drugs and reference compounds across 70 pharmacological assays can be seen in **Figure 6**.



**Figure 6.** pIC<sub>50</sub> values for approximately 2,000 drugs and reference compounds across 70 pharmacological assays. Extracted from [28]

Recently, tools for early evaluation of mechanism-based toxicity have been introduced but very little has been done to screen pharmacological promiscuity, which can also seriously affect success rate and influence side effects profiles [19]. The relevance of this component of drug discovery was recognized a long time ago but was applied to few compounds only, just before clinical trials. At this stage of the drug discovery process only little or no chemistry capacity was available for corrections. Therefore, in many cases projects were abandoned without establishing whether the undesirable side effect was associated with a particular pharmacophore specific for the scaffold or just an accidental effect of individual molecule [19].

Endogenous ligands are usually promiscuous, example of which is the fact that large receptor families share the same ligand (e.g. serotonin for the 5-HT receptor family). It is thus not surprising that drug-like synthetic molecules act in the same way. However, targeting a particular member of a protein family is often the goal of a therapeutic approach. Traditionally, during the hit or lead optimisation processes, selectivity was considered at most for a handful of targets within the same protein family, most usually only for one or two selectivity targets. Miniaturization and parallel screening have recently enabled the *in vitro* profiling of a handful of compounds to a wide range of

different targets, although the logistics do not enable a high throughput in this process. That makes this multitarget profiling only suitable for late stage leads, helping in the selection of drug candidates among them.

Another step beyond *in vitro* multitarget profiling for safety and toxicity evaluation of drug candidates is cell-based high content screening (HCS). Historically, cellular toxicity screening has relied on the use of single-parameter readouts for toxicity markers such as cell proliferation, mitochondrial activity or membrane permeability. Although useful to an extent, predictability for compound toxicity *in vivo* is poor [33]. HCS is a functional screening in an automated platform based on fluorescence microscopy and quantitative image analysis in the physiologic context of intact cells [34,35]. It allows multiplex analysis, and thus allows two or more discrete responses to be measured in a single assay. While *in vitro* assays on single targets neglect the intracellular structural and functional networks, HCS enables to assess safety within cell context [34].

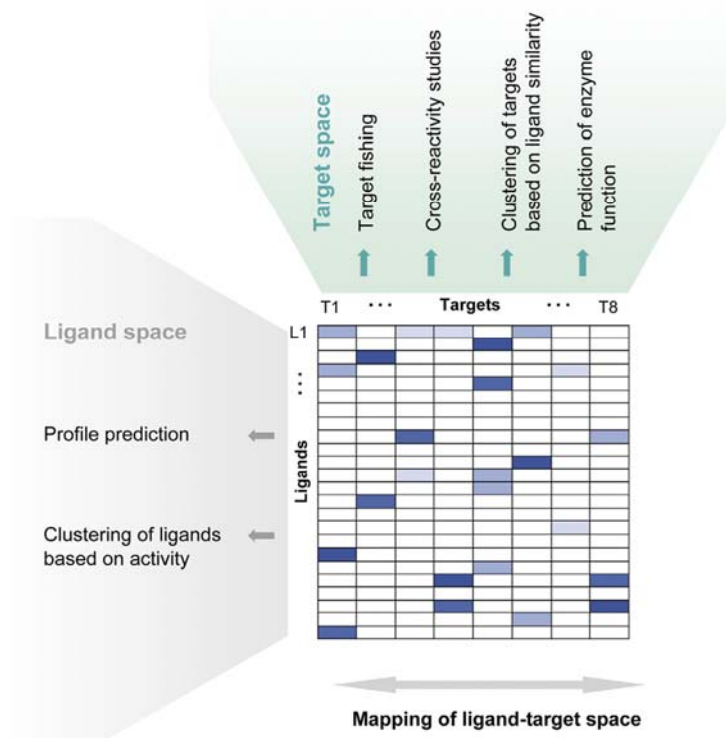
HCS provides with higher biological-content information with respect to HTS for which, while HTS is used as a fast primary screen to identify hits for further testing, HCS can be used in secondary screening in the lead optimisation process [36]. HCS-based cytotoxicity assays analyse the effects of compounds on a number of parameters such as nuclear size, mitochondrial membrane potential, intracellular calcium levels, membrane permeability and cell number, offering increased specificity and selectivity for toxic events, and allow a higher level of predictability for future *in vivo* testing [33].



## Chapter 1.2 – *In silico* drug discovery

The advent of genomics, combinatorial chemistry, HTS and other drug discovery innovations and technologies has resulted in an unprecedented abundance of new targets and potential drug candidates. Despite this, the rate of entry of new chemical entities into the market is still slower than expected [2], suggesting that the quality of drug candidates must be improved. An increase in computational capacity has enabled new, high-throughput *in silico* methodologies that can directly be implemented in the traditional drug discovery process to improve its success rate. The key point addressed by these methodologies is the interface between chemistry and biology for which classification and annotation of data (**Section 1.2.1** and **Section 1.2.2**, respectively) is crucial. Once this has been properly done, the next step for the generation and integration of knowledge and its use for rational and systematic drug design [37] requires the analysis of data relevant to establishing links between proteins and ligands in an attempt to extract knowledge from them. This knowledge is generated by deriving generic rules and models that can be applied to the virtual design and screening of compounds. For these predictions and this knowledge generation, it is key to find the right descriptors that will accomplish the “neighbourhood behaviour” assumption [38] and that will actually enable to extend the biological properties of a compound to its closest neighbours in the n-dimensionality space defined by the descriptors. A broad overview of molecular descriptors will be provided in **Section 1.2.3**.

These descriptors enable to create models that can be used in hit and lead optimisation processes. Quantitative structure-activity relationship analysis (**Section 1.2.4**) can predict ADMET properties of the studied compounds that can help prioritising the series of compounds with the most promising features. Moreover, these rules and models enable also predicting the activity to a certain target for a large set of compounds leading to virtual ligand screening (**Section 1.2.5**). Extending this prediction along the biological space allows for virtual target profiling (**Section 1.2.6**) that attempts for a complete profile of compounds against all relevant targets. Virtual target profiling enables to fill the ligand – target matrix, which, as summarized in **Figure 7**, provides with information useful all along the drug discovery process. Despite the increased throughput of binding assays with miniaturization and parallel screening allows testing thousands of small molecules against hundreds of protein targets, it is still impossible to test all possible lead- or drug-like compounds for all relevant targets. Therefore, virtual target profiling can help predicting the full pharmacological profile of ligands, useful for multitarget-directed drug design and enabling cross-reactivity and safety assessment. This full profile can also be used to cluster ligands depending on the similarity of their target profile and cluster targets in terms of their ligands. Moreover, it can also be used for predicting the protein a given orphan active compound might be acting upon, using the so-called target fishing strategies.



**Figure 7.** Adapted from [39].

Additionally, all these developed *in silico* methodologies can be used in the first step of drug discovery by narrowing down the number of chemical compounds to be tested for hit identification. As discussed in **Section I.1.1.1**, the initial chemical diversity-driven approach to compounds testing relying on random finding of actives did not provide with the expected wellness of interesting hits [40]. Consequently, the development of computational methodologies to predict the full pharmacological profile of compounds led to a more rational approach towards compound selection, selecting smaller subsets of compounds predicted to be relevant for the project of interest, as explained deeper in **Section I.2.7**.

Finally, the most recent use given to the accumulated annotated knowledge for ligands and proteins is network pharmacology (**Section I.2.8**), which takes advantage of this information for a systems approach to analysing the influence of the modulation of one or several targets over whole pathways and all available studied targets.

### ***I.2.1. Classification schemes***

Information on small molecules, proteins and their interactions has to be collected so knowledge can be extracted from it. The usage of both unified nomenclatures (ontologies) and appropriate classification schemes is key for the annotation of all biological and chemical entities and for an integrative and information-rich knowledge generation [41]. Initiatives towards the unified nomenclature [42,43] and classification of proteins have already been successfully engaged. However, from the ligands point of view, no consensus has been reached although several unique

identifiers [44-46] and classification schemes, including one developed in our laboratory [47], have been proposed.

With respect to the functional classification of proteins, targets that share structure activity relationships (SAR) can be thought of a SAR homologous (SARAH). This concept was re-introduced in 1999 by Frye [48], who claimed for the recuperation of this functional classification over the sequence-based classification of protein families. One of the most broadly used target classification scheme is the Enzyme Commission (EC) classification scheme [49]. The EC scheme is a 4-level hierarchical functional classification where enzyme classes are assigned four numbers, one for each level. The first number represents one of the six main classes of the chemical reactions that enzymes catalyze. The second and third numbers describe the subclass and sub-subclass of the overall reaction and the last number reflects the substrate specificity of the reaction [50]. Another largely used protein classification scheme is the one established for nuclear receptors [51], which is a 3-level hierarchical scheme also based in function, with a first number, a letter in the second level and another number in the third level. These different classification schemes for different protein families highlight, however, that no standard has been reached that is applicable throughout all the protein families.

The hierarchical nature of target class similarity has a profound influence on the way novel ligands are discovered. The conservation of the binding site architecture within a target family translates into a conservation of the architectures of ligands which bind to these targets [52], so targets within a gene family will often have similar ligands and properties. A number of compounds active to one family member will be active against other family members, which often have different biological functions [40]. Thus, one way to gain efficiency is to re-use information and know-how among proteins of the same family [40] as knowledge about ligands of one target and the distance between targets in biological space facilitates the prediction as to which molecules are suitable for novel targets. Thus, we can relate targets by the similarity of ligands to which they bind, a central paradigm of chemogenomics [53].

With respect to the structural classification of ligands, while in the case of proteins classification is done in terms of functional hierarchy, in the case of small molecules it is done on the basis of structural hierarchy. In these terms, compounds can be summarized in scaffolds (subset of the molecule where side chains have been removed), frameworks (simplified scaffolds) [54], ring systems or other substructures. Although the classification of proteins has not reached a global consensus for all protein families, different committees for different protein families have reached well-established classifications that have become a standard for the given protein family. For small molecules, however, no such consensus has been reached although a variety of classification schemes have been proposed. One of such chemical graph-based classification schemes is the HierS [55]. This recursive algorithm identifies all possible ring-delimited substructures for each molecule and, once all such subsets for a set of compounds are identified, molecules are grouped by shared ring substructures. The hierarchical structural relationships

between the substructures are established so over-represented structural features can be rapidly identified. Another classification scheme is the Scaffold Tree [56], based on the hierarchical classification of scaffolds, from which rings are iteratively removed. These rings are prioritised following a set of rules by which peripheral, less characteristic rings are removed first.

In our laboratory, a classification scheme, the Chemical Structure Code (CSC) [57], has been developed also based in structural hierarchy. The code consists on a unique hierarchical six-level CSC for each molecule, in which each level encodes for a substructural characteristic, going from the most general ones to the final unique identifier of the molecule. The first, second and third levels are integers specifying, respectively, the number of rings in the largest ring system present in the molecule, the number of bonds in the longest path and the number of branching points in the longest path. The fourth, fifth and sixth levels are unique eight-character hash codes for the molecular framework, scaffold, and the complete molecular structure, respectively.

### ***1.2.2. Annotated chemical libraries***

There are essentially two types of data that are useful to establish a link between proteins and their ligands. These are on one hand structural data on proteins and protein-ligand complexes and on the other hand response data on the interaction of ligands with proteins, including potency, affinity, metabolism and toxicity. These two data types lead, respectively, to structure-based and ligand-based approaches to chemogenomics. Structural data available is growing significantly, even more since the advent of structural genomics [58], and is nowadays centralized in the Protein Data Bank (PDB) [59], a public repository of three-dimensional structures currently containing over 50,000 entries. The information available in it is scarcely classified, and several attempts to extract and organize relevant information from the PDB are currently available. Among them, enzymes structures are organized following the EC classification scheme [49] in the Enzyme Structures Database [60] and protein-ligand complexes from the PDB, together with their binding affinities, are compiled in the PDBbind database [61], currently containing around 3,200 entries [62]. Another effort towards organizing information from the PDB was engaged in our laboratory, where FCP, a publicly accessible, web-based tool to analyse the contents of the PDB, the population of each protein family represented and the trends this protein structures population over the years, has been developed [63].

From the response data perspective, a number of initiatives have recently focused on collecting and storing the structures of small molecules for which pharmacological data is available, giving rise to the so-called annotated chemical libraries (ACL) [64]. Among those, one can find databases relating drugs and targets, such as Therapeutic Target Database [65], DrugBank [66], SuperTarget [67] and Matador [67], or relating chemical compounds to their effect in cell-based assays, as ChemBank [68] and PubChem BioAssay [69]. Another kind of very popular ACLs are those that relate small molecules and their *in vitro* binding affinities to a given protein target as reported in literature. Example of these are the MDL Drug Data Report (MDDR) [70], the WOMBAT



database [71], the AurSCOPE database [72] and the MedChem and Target Inhibitor databases [73]. Under the same spirit, a more modest initiative took place in our laboratory to assemble an annotated chemical library directed to the nuclear receptor family (NRa1) on the basis of public sources of information, mainly reviews and medicinal chemistry journals of the last 13 years [74].

### ***1.2.3. Molecular descriptors***

Recent efforts in collecting, storing, and organizing data on the pharmacology of ligands and on the structure of proteins are facilitating the generation of knowledge on target classes. Comparative studies have suggested that information about a target obtained from known bioactive ligands is as valuable as knowledge of the target structures for identifying novel bioactive scaffolds [6,75,76]. Therefore, the final choice for a method to use will depend on the type and amount of information available without *a priori* having a large impact on performance [6].

Most structure-based approaches are based on docking, which enables visual inspection of the results leading to an intuitive interpretation and understanding of the binding process [75]. Another quality of these methodologies is that they allow for new binding modes fitting within the defined active site. Despite this, it has been reported to be outperformed, in some cases investigated, by purely ligand-based approaches [75] due to the still low accuracy of the scoring functions to predict binding affinity [77] and the need for conformational space exploration. The quality and resolution of the initial structural data are other factors that will have an impact on the performance of these methodologies. Moreover, structure-based approaches can be based on homology models of the structure [78,79] or on very closely related proteins [80] instead of the actual 3D information of the target of interest. Although these approaches have been proven to perform well, more uncertainty is added to the already big amount of variables with which structure-based VS has to deal.

On the other hand, ligand-based approaches rely on the central similarity-property principal which states that similar molecules should exhibit similar properties [81]. Hence, the activity prediction of a compound or a set of compounds will be done based on the similarity or distance to a set of reference ligands with known bioactivity to a protein target [82]. Different types of two- and three-dimensional molecular descriptors, features and substructures in combination with a variety of classification schemes, such as recursive partitioning, Bayesian statistics, neural networks or machine-learning methods have been used for this purpose. Among those, methods aiming at identifying chemical moieties commonly appearing in bioactive ligands have attracted particular attention due to the ease of translation of these privileged structural motifs into compound-library synthesis. However, despite the many evidences of chemical substructures occurring frequently in ligands bioactive across a diverse panel of proteins [55,74,83,84], the true existence of target-family selective privileged substructures continues to be a matter of debate [85].

Pharmacophore-based VS can be considered to be in the intersection between structure-based and ligand-based approaches. On one hand, they use structural information but, on the

other hand, they often use ligands as the references to build the models. One obvious advantage of similarity searching over a pharmacophore-based search is that it does not require a set of structurally related compounds of similar biological activity to derive a model. When using similarity searching, even one active molecule can be used to search a database for related compounds.

This similarity-based model generation has proven very convenient, as it is computationally inexpensive and requires little information [86]. It mostly uses 2D descriptors, also called topological descriptors, which are derived from the connectivity table of the molecule and take into account distances among atoms in terms of number of bonds in the shortest path between them. These methodologies are based on the “neighbourhood behaviour” assumption [38], which states that compounds with high chemical similarity have high probability of sharing biological profile. Thus, the *in silico* annotation of a given ligand as active for a target will be done on the basis of the distance to any ligand with target information [82]. For these distance (or similarity) metrics, numerical representations of chemical structures that provide with relevant information about the compound are necessary [87]. To be effective in a large pharmaceutical environment, an optimal method needs to be fast and sufficiently robust to process millions of similarity calculations in each research project.

The most commonly used descriptors are topological fingerprints [88,89], which encode the presence or absence of substructural fragments in molecules in a binary fingerprint, without taking into account the number of occurrences of the feature. These fingerprints can be precalculated and compared, usually by means of Tanimoto distance, in a very fast and efficient manner to any reference set. The encoded substructures can either be a predefined list common to all sets of molecules analysed or a list that depends on the analysed set, in which all the encountered substructures up to a certain path length are considered. Example of this type of descriptors are MDL MACCS [90], which consist on a 960-bit string in which each position corresponds to a predefined pattern. These lack of generality and high dimensionality have been highlighted as the main handicaps for this method. Conversely, Daylight [91] and UNITY [92] fingerprints do not consider a predefined set of substructures, so the resulting boolean array is dependent on the nature of the chemical structures in the database. Both Daylight and UNITY fingerprints are hashed, meaning that structural patterns are mapped to overlapping bit segments [93]. Consequently, these descriptors are more abstract as single bit positions can no longer be associated with one specific feature and collisions can occur when two different paths cause some of the same bits to be set. The compression of long binary fingerprints using hashing or fingerprint folding algorithms can introduce a systematic error in the similarity metric used, which should be taken into account and corrected [94].

Despite the broad use of topologic fingerprints, some authors have suggested that, based on the increasing knowledge available about ligand-protein interaction, it is reasonable to say that pharmacophoric features may be more important than topology or substructures when discerning if two molecules will bind the same protein [38]. Moreover, as these descriptors are scaffold-

independent due to the abstraction inherent to the data encoding in a set of descriptors, they are especially useful for identifying chemically different but biologically similar compounds, the so-called scaffold hopping ability [5]. These methods are able to detecting as similar molecules with the same pharmacophoric features arranged in the same way around essentially different scaffolds [95]. In this sense, atom-centred feature pairs [96] have been proven to be highly effective [53,97,98]. A set of descriptors based in this concept is CATS [97], in which molecular graph nodes (atoms) are translated to the corresponding pharmacophoric feature (hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), positively charged, negatively charged or lipophilic). Each of the corresponding 15 possible feature pairs is considered for distances of up to 10 bonds, which results on a 150-dimensional vector. Another atom pair-like approach are the Similog keys developed at Novartis [53], which also translate the graph nodes into pharmacophoric features (HBD, HBA, bulkiness and electropositivity) and consider, in this case, triplets of atoms instead of pairs. When used for similarity calculations, the presence or absence of each possible Similog key is encoded in a binary fingerprint that could have as many as 8031 dimensions. Despite this, based on internal data, Novartis researchers have limited this number to 5989 possible Similog keys, and thus descriptors dimensions. The Shannon entropy descriptors (SHED) presented in this thesis are also based in atom pairs, and consider four types of pharmacophoric features (HBD, HBA, aromaticity and hydrophobicity). The main difference with respect to the previously described methods is that they summarize each atom pair distribution into a single value by means of Shannon entropy, leading to a 10-dimensional vector. This will be discussed in detail in **Chapter III.1**.

Another set of descriptors summarizing relevant properties of small molecules are BCUT descriptors [99]. Each BCUT combines physicochemical and structural information, derived from 2D or 3D structure, in a single number. The properties evaluated include atomic polarizability, atomic charge, HBA and HBD. From the total of BCUT descriptors calculated for a set of compounds, the subset that better represents the structural diversity of the analysed dataset is extracted. Consequently, for each reference dataset a new subset of descriptors has to be generated. Using BCUT descriptors, chemistry space is divided into cells so, when looking for a maximally diverse set, molecules will be selected to occupy the maximum number of different chemistry-space cells, whereas for virtual ligand screening molecules will be selected that occupy the same chemistry-space cells as the reference compounds [14].

Finally, other pharmacophoric features-based descriptors are molecular fields, which are ligand-based 3D descriptors. Examples of these are comparative molecular field analysis (CoMFA) and grid-independent descriptors (GRIND). These methods are based on the assumption that shape-dependent descriptors are key to predict the activity of chemical compounds. Still, despite the use of principal component analysis or partial least squares (PLS) to reduce dimensionality, similarity calculations based on molecular fields are computationally expensive due to the large amount of numerical descriptors required for the comparison of the three-dimensional fields around the molecules. In CoMFA [100], a data table is constructed from the field values at lattice

intersections, which reflect steric and electrostatic interaction energies between the compound and a probe atom placed in each lattice point. The analysis of this data table is done using PLS that leads to the generation of a conventional QSAR equation that can be contoured in 3D space. Some of the drawbacks of this approach are that it is highly dependent on the conformation chosen for the analysis and that the comparison among molecules (fields) is done by using a lattice-point-by-lattice-point correlation coefficient, which is very computational- and time-consuming.

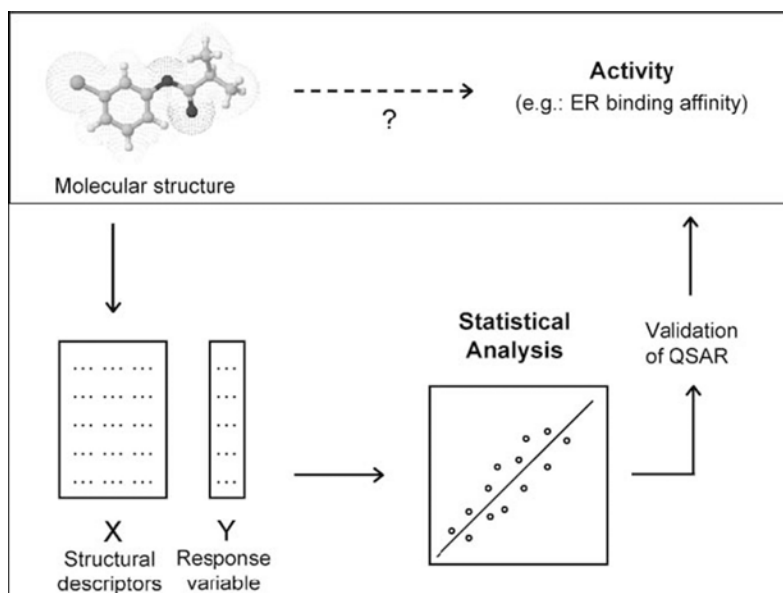
GRIND [101] are alignment-independent molecular fields descriptors calculated in three steps. First, molecular interaction fields are calculated for each molecule. Then, this set of fields is simplified and in a last step the results are encoded into alignment-independent variables using an autocorrelation transform. The results can be then traced back to the original fields and visually represented together with the 3D structures of the molecules, which enables an intuitive interpretation of the results.

#### ***1.2.4. Quantitative structure-activity relationships***

A major goal in pharmaceutical research is to predict the behavior of new molecules using knowledge derived from the analysis of the properties of previously tested molecules. The earliest intuitions and insights in structure-activity relationships (SAR) can be traced back to the nineteenth century. Hansch was among the first authors to use statistics rather than intuition to relate structure and activity, giving birth to quantitative structure-activity relationships (QSAR) in the mid twentieth century [102], assuming that the activity, expressed in the logarithmic form, depended on the substituents' contribution to the parent compound properties [103]. In this sense, QSAR consists on the construction of a mathematical model relating a molecular structure to a chemical property or biological effect by means of statistical techniques [6]. These relationships are derived, as shown schematically in **Figure 8**, by statistical analysis of a data table, filled with numerical property values in one axis and molecules in the other, and usually take the form of a linear equation [100]. For this, QSAR methodologies require a series of structural analogues, at least 4 or 5 compounds having very similar chemical structure per descriptor, that interact in the same way at the same binding site. Based on the assumption that the differences captured by the model are only due to the accommodation of different functional groups to the binding site, QSAR models can be used, in addition to predict the response of the entire molecule, to understand the relevance of particular molecular features for the suited effect [104].

The statistical analysis required for QSAR consists on a first pre-processing step essential to reducing redundant information and selecting the most relevant variables for the building of the prediction model. The increasing number of descriptors commonly calculated has required the introduction of different tools to cope with correlated variables and with matrices constituted by more numerous variables than chemicals in the data set. Tools such as principal component analysis (PCA), a multivariate data exploratory method, have set the stage to deal with the selection of independent and relevant variables based on the use of latent variables generated by a

linear combination of the original set of descriptors. However, since the first implementation of QSAR, the simpler models with few variables based on homogeneous set of chemicals have been replaced with studies using more heterogeneous data sets and a high number of variables [103]. In cases where too many initial variables have to be studied, PCA does not efficiently explore all possible combinations and more sophisticated tools, like genetic algorithms, are required [103].



**Figure 8.** Schematic representation of the steps for developing (Q)SAR models. Extracted from [103].

In the second step, the methodology behind the model derivation itself will strongly depend on the type of response variable studied. For categorical properties classification tools will be used, while for continuous variables regression approaches will be applied. Most of the regression QSAR methods are based on a multiple linear regression or partial least squares analysis. These approaches can only capture linear relationships between molecular characteristics and functional properties. By contrast, artificial neural networks (ANN) can recognize highly non-linear relationships between structural or physicochemical descriptors and biological activities or any other molecular features [105-107]. This inherent feature of non-linearity makes neural networks particularly well suitable to treatments of generally non-linear structure–activity relationships.

Despite the similarity principle in which QSAR is based is often true, experience has shown over the years that two molecules that are chemically very similar can have different activity profiles. On one hand, minor differences in the structure of molecules may result in different mechanism of action or in different binding modes. This can lead to outliers which are unable to fit any QSAR model [108]. On the other hand, the intrinsic noise associated with both the original data and specific methodological aspects involved in the construction of a QSAR model have to be also considered [109]. Another issue are activity cliffs in the so-called activity landscape, defined as compounds with a high ratio of the difference in activity with respect to their chemical difference [110]. These make measures like  $q^2$  unreliable and impossible to separate measurement errors from observations that do not obey the physical assumptions of the model [111].

There are other issues that can affect the accuracy of QSAR models, some of which can be overcome by using ANN. As amply discussed in the literature, overfitting can occur whenever the model has been set to explain and adapt to the peculiarities of the training set at the expense of its predictability for a new test set [112,113]. If the architecture of the ANN and the number and quality of the training examples are adequate, overfitting can be avoided. Another issue to be considered is extrapolation. As any statistical model, QSAR applicability is limited to interpolation within the limits of the data from which they are constructed. Its availability to be applied to any molecule with low similarity to those comprised in the training set has been discussed [104].

### ***1.2.5. Virtual ligand screening***

Beyond this linear relationship of structures and biological effects, chemogenomics has appeared as the attempt of rationally mapping all possible ligands to all possible targets [40]. On one hand, the experimental determination or prediction of binding affinities of small molecules to the desired targets facilitate the rational optimisation of the compounds. On the other hand, information on activity of drug candidates for off-targets enables to predict undesired side effects.

Virtual ligand screening (VS) is the process of scoring and ranking molecules in decreasing probability of activity for a certain target [6,89]. In analogy to HTS, VS is a tool to extend the profiling of compounds against a single target along the chemical space. This technique has shown very good results with respect to HTS alone, enhancing cost-effectiveness, increasing hit rates by a factor of 100 to 1000 [114] and decreasing the high proportion of false negatives provided by HTS alone. Its objective is finding the true actives in the initial database while trying to catch novel chemotypes displaying the sought activity. This means the obtained results can be at most as good as the initial set of compounds.

The main goal of VS is to come up with hits of novel chemical structure that yield a common pharmacological profile. This scaffold hopping differs from the synthesis of analogues in the requirement that, instead of peripheral conservative replacements, the central core of the compound has to be changed. Thus, VS methods have to be fuzzy enough to retrieve structurally different compounds with the same activity without getting too many false positives. The success of a virtual screen should be defined in terms of finding interesting new scaffolds rather than many hits, as low hit rates of interesting scaffolds are clearly preferable over high hit rates of already known scaffolds [115].

Target-based VS approaches rely on docking of libraries of single structures or even multiconformer libraries, and are thus suitable for high-throughput searches [116]. These approaches are limited by the quality of the scoring function, an issue widely discussed in the literature [77]. Another drawback of docking-based VS is that it is computationally intensive, although pre-filtering using pharmacophore models is often applied to reduce non-productive orientations [117]. Docking is generally poorer at selecting actives than most 2D or 3D ligand-based methods, as measured by enrichment factor [76,86,118,119]. Despite this, target-based

approaches generally perform better in terms of scaffold hopping [115]. A review comparing 10 docking algorithms and 37 scoring functions against seven different targets showed that the performance of each program was not consistent across the different targets [77], suggesting that different methodologies and scoring functions are biased towards different binding site environments.

Ligand-based approaches can use either of the descriptors presented in **Section 1.2.3**. In terms of novel lead discovery, pharmacophore searching has perhaps proven the most widely applied VS method with hit rates for selected data sets of 1 to 20% [117]. Conversely, substructure-based fingerprints methods will be those providing with poorer scaffold hopping, as they are based in common substructure searching.

Integrating VS-based subset selection and HTS in an iterative manner leads to the process of sequential screening [120]. VS-selected compounds with increased probability of being active are tested through HTS, and HTS results are used to enrich VS model for a re-selection and re-testing of compounds. Moreover, VS is only feasible when knowledge, either on the target structure or on ligands, is available. When there is no such information, a first HTS screen is useful to provide with initial ligand-based information to derive a VS model.

These iterations might identify series of analogues early on during the sequential screening process and produce initial structure-activity relationships [120]. In order to be suitable for HTS library selection, VS has to perform very well in identifying active compounds within the first 0.1 to 1% of the scored set, as many pharmaceutical companies have screening libraries of up to one million compounds from which, by means of VS, they select a few thousands to be tested [121]. Consequently, it is not only interesting for VS to provide with better enrichment compared to random selection, but to provide with a high enrichment even in this top proportion of the ranked list.

### ***1.2.6. Virtual target profiling***

Beyond the mere improvement of the potency, the generation of safer and more efficacious drugs is one of the main concerns in current pharmaceutical research. *In silico* target profiling methods are emerging as efficient alternatives to the currently unaffordable high-throughput *in vitro* target profiling of compounds [114]. Unlike HTS and VS, virtual target profiling (VP) extends the matrix relating biological and chemical space both along the chemical structures axis, enabling large number of compounds to be tested, and along the biological axis, making predictions for many targets for each molecule.

The uses of this broad virtual profile of compounds cover a wide range of applications. In the first place, it can assist in the drug discovery process by prioritising the hit and lead series in terms of their activity and selectivity profile, both in projects targeting a single protein and in multitarget-oriented projects. In the second place, it can alert of potential secondary effects due to residual

affinities for undesired targets [122], which is useful to establish which *in vitro* selectivity targets would be more relevant for the project of interest. Finally, another promising possibility is in the area of finding new therapeutic uses for approved drugs, an activity often referred to as drug repurposing [123]. Drugs have traditionally been designed to interact with a primary target known to be relevant to the particular disease of corporate interest. During the drug optimisation process, very limited scope was often given to address properly the issue of selectivity, by considering only a handful of additional targets phylogenetically related to the primary target. Through drug repurposing, obtaining a drug candidate and many steps in safety assessment can be skipped leading to a reduction of the 40% of the overall cost of bringing a drug to the market [124]. Moreover, there are less risks of failure, as the systemic effects of the drug have already been proven in its previous use. Thus, the ability of *in silico* target profiling methods to identify new targets for old drugs, as demonstrated recently by Keiser *et al.* [125], has direct implications for using immediately off-patent products in clinical trials [124,126].

As in the case of VS, there have been target-based and ligand-based approaches towards VP. High-throughput docking methods for rapidly computing relative affinity [127] and inverse docking for automatically screen small molecules, in some cases even for a panel of 698 protein structures covering 15 therapeutic areas [128], are representative of target-based VP methodologies, while the extension of ligand-based VS through the biological axis is the basis of ligand-based VP [57,129].

### **1.2.7. Chemical library design**

As discussed in **Section 1.1.1.1**, the implementation of combinatorial chemistry and HTS within the drug discovery process increased the traditional capacity for synthesizing and testing compounds. The early technology-driven diversity-focussed HTS phase showed poorer performance [40] and higher costs [6] than originally anticipated. On one hand, screening purely diverse libraries can lead to missing activity areas, as these areas will be enriched with active compounds but, as very similar compounds often show dramatically different activities, they will also contain a number of inactive compounds [130]. Selecting just 1 or 2 analogues from each cluster of compounds with Tanimoto similarity over 0.85 would result in missing activity within 70% of the clusters, while screening any 10 members of the same cluster would yield a 97% chance of finding an active compound. On the other hand, as corporate collections were compiled before the drug-likeness and lead-likeness filters were broadly used, hits retrieved through HTS often showed poor physicochemical properties [131,132] leading to a higher attrition rate in later stages [132].

As mentioned earlier, the selection of the compounds to be tested *in vitro* is a key step in drug discovery process, as the quality of the initial hits will determine the overall performance of the process. The shift from technology- to knowledge-driven drug discovery has set the stage for efforts towards small and focused compound collections [10] and crucial to this are the many *in silico* tools developed recently that take advantage of the available knowledge rather than trusting



in probabilistic find of interesting compounds in large and diverse databases. Rationally selecting a set of compounds to be tested *in vitro* which are predicted to have desirable characteristics for the project of interest is done in two steps, the filtering of undesired compounds and the ranking and selection of compounds according to their likeliness to have the sought activity. The first step consists on filtering out compounds with undesirable features (*i.e.* reactive groups, poor ADMET prediction) and selecting those that are predicted to have drug-like or lead-like filters based on one-dimensional descriptors [133], such as molecular weight, counts of nitrogen and oxygen or number of rings, which reflect global properties like the size, shape and lipophilicity of molecules [6] or more sophisticated complexity analysis [134]. In many cases, a reactive substructures filter based on two-dimensional topological descriptors is also applied [135,136]. Lead-likeness filters aim basically at retaining compounds with enhanced chemical tractability and development potential, as they have to be modified and analogous molecules have to be synthesized and tested in hit-to-lead, lead generation and lead optimization phases.

In the second step, VS or VP protocols can be used to select the compounds that are more likely to display the desired pharmacological profile, either directed to a particular target or a protein family, respectively. In order to increase the efficacy of the drug discovery process, chemogenomics approaches tend to organize research around target families for effective reuse of chemical and biological information [40]. Target family-oriented approaches rely on the hypothesis that the conservation of the binding site architecture within a target family or subfamily translates into a conservation of the architectures of ligands which bind to these targets [52]. Therefore, similar ligands should bind to similar targets and thus the knowledge obtained from one protein should be transferable to new related proteins. For this to be true, target families should be defined and clustered taking into account the protein structures and particularly the structure of their binding sites, rather than using the traditional sequence-based phylogenetic approach [137].

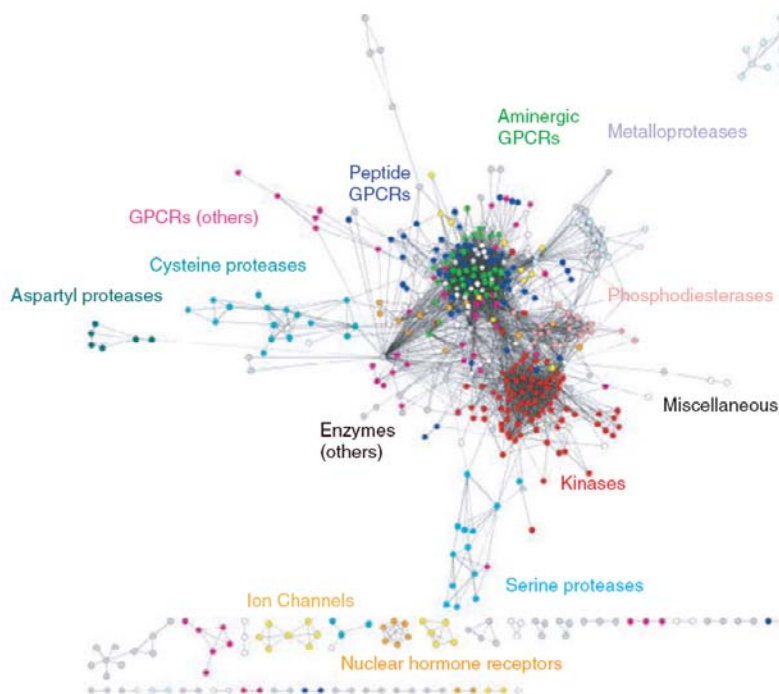
Focused libraries provide a considerably higher hit rate (typically 10 to 100 fold higher than random screening selection [138]) and fewer false positives [138]. Moreover, focused libraries targeting protein families instead of single targets take advantage of the functional similarity for close structural homologues so that the selected set of compounds designed to be active to one of the members of the protein family can be assumed to contain an implicit pharmacophore hypothesis not only valid for single target but also for a whole family of targets [53].

### **1.2.8. Network pharmacology**

Network pharmacology results from the integration of systems biology and the huge amount of response data describing the biological effects of small molecules [27] mostly provided by HTS. Organisms and cells can be considered as complex networks of interacting molecules and, despite all the experimental data available to date, they are still far from being fully understood and properly modelled. Because of the complexity of the cell network, most of the attempts of modelling it have preferred to consider pathways, *i.e.* parts of it acting in concert, instead of the whole

network [15]. In drug discovery process, the modelling of these pathways puts possible drug targets into context and can help to the mechanistic understanding of a disease and the possible consequence of the perturbations on the system caused by the drug action [139]. They can serve as a way to choose the most appropriate protein target and assess possible toxicological effects by analysing the proteins homologous to the target and the consequences the interaction with these may cause [15]. In modern drug discovery, network pharmacology can be used to design a polypharmacology profile for a drug to modulate multiple proteins simultaneously, as discussed in the **Section I.1.2**. Moreover, in a recent study, Yildirim *et al.* [140] have applied network analysis to existing public drug – target information from DrugBank [66], and although this data is far from complete, the authors could observe a network behaviour instead of the isolated bipartite nodes that would be expected if the drugs did actually act selectively on single targets.

To obtain a global picture of their interplay, the increasing amount of data being generated around small molecules, proteins, genes, pathways, and diseases has to be collected, stored, organized, connected and integrated with a variety of existing molecular, cellular and organismal data such as microarray experiments and pathways [141]. By bringing together these heterogeneous data types, it is possible to construct a network that captures many aspects of how small molecules function in a cellular context [142]. This can help studying the relationships among different protein families in terms of shared ligands, as the intra- and inter- protein family promiscuity study performed by Paolini *et al.* illustrated in **Figure 9**.



**Figure 9.** Cross-pharmacology interaction matrix. Extracted from [143]

Despite this, the data currently available relating ligands and proteins are still far from complete. This is due to the fact that, because of limited time and resources, molecules are usually

not screened systematically through a large panel of protein targets for the sake of obtaining the maximum amount of information possible but solely to the target of interest at that point in time. But even if they were screened through multiple targets, habitually only a limited amount of data is made available, since publishing large amounts of negative data is often regarded as not informative. These important, yet often overlooked, aspects lead to a situation of data incompleteness within the interaction matrix, as will be discussed in **Chapter III.4**.



## Chapter I.3 – Conclusions and outlook

During this introduction, a global description of the *in vitro* drug discovery process has been provided. Additionally, the impact and consequences of new technological advances has been discussed. These advances have implied a dramatic increase in the throughput rate at which data can be obtained, leading to a need for computational tools to gathering, storing and classifying this information through annotated chemical libraries, ontologies and classification schemes, which has been stressed as one of the key steps towards the generation of knowledge from this data.

This knowledge, under the form of *in silico* modelling and prediction tools, has led to a conceptual shift in the drug discovery process towards a more rational approach. Relevant for this are the methodologies for representing the chemical structures by means of molecular descriptors able to catch and summarize the pharmacologically relevant features while enabling fast and efficient similarity or distance-based comparisons of large reference and target molecule sets. These predictive models, either for QSAR, VS or VP, can be of use all along the drug discovery pipeline and have so far proven their ability to increase the overall efficiency of the process.

Despite these improvements, the quality of the drug candidates is still far from optimal, as proven by the high attrition rates at late stages of drug discovery process due, mainly, to efficacy and safety issues. These issues have to be considered and properly addressed in early phases and many efforts to optimise all aspects simultaneously have been engaged in pharmaceutical companies. Possible adverse drug reactions are now considered earlier in the drug discovery pipeline by means of hit or lead profiling for several relevant off-targets to avoid late attrition and to improve cost-effectiveness. The greatest challenge, still, remains the simultaneous optimisation of both binding affinity and pharmacokinetic properties [116].

Despite the formal separation of the *in vitro* and *in silico* methodologies for drug discovery in this Introduction, there is actually a great complementarity among them. It is clear that, despite the proven usefulness of the computational algorithms, they can only be regarded as prediction tools, requiring a continuous validation of proposed actives by rapid synthesis and testing to ensure a successful process making an effective integration of *in vitro* and *in silico* approaches necessary [10]. This has already been discussed in **Section I.2.5** with sequential screening, which integrates HTS and VS. Other integration examples are between LC/MS and ADMET models, which are also mutually enriched by iterative cycles of prediction of properties, proposal of compounds and testing.



## Part II - Discussion

The **Part I** of this thesis puts our work into context by providing a general overview of the drug discovery process. A general outline of the main steps towards obtaining a drug candidate is provided, together with an introduction to the new trends in this field. Then, the role of *in silico* methodologies within drug discovery is presented, emphasizing the paradigm shift they have introduced towards a more rational approach. Computational tools are actually integrated with *in vitro* methodologies in most of the pharmaceutical companies' pipelines. These tools take advantage of previous information by organizing it and generating knowledge from it. This knowledge, under the form of computational models, can then be used in almost every step in the drug discovery process.

In **Part III** the work developed during this thesis by means of published research articles will be presented. A new set of computational tools for assisting in drug discovery process will be presented. All these methodologies are based upon a novel set of topological descriptors called Shannon Entropy descriptors (SHED). These descriptors are introduced in **Chapter III.1**, where their capabilities are also proven. These descriptors enable to summarize the pharmacophoric features of molecules in a topological, fuzzy, scaffold-independent manner. In this chapter, it is shown how these characteristics allow SHED to be especially suitable for scaffold hopping. Additionally, a first introduction of the potential use of SHED for ligand-based virtual ligand screening is provided. The combination of these two features makes these descriptors especially useful in the drug discovery process, as they enable to retrieve structurally diverse compounds with similar pharmacophoric features distribution and thus similar activity profile.

It is remarkable to highlight the fact that usually 3D descriptors-based methods are considered to be more appropriate for scaffold hopping, as they are often based in pharmacophore models independent of the compound's structure. SHED descriptors, despite being 2D topological descriptors, are based on pharmacophoric features rather than on atoms and enable a great degree of independence from the compound's structure. On the other hand, scaffold hopping is a term often used on the basis of intuitive criteria after the visual inspection of a hit list. It is thus important to establish a rational unbiased criterion, independent of the similarity method used, to assess it. A very useful approach is to first define the chemical scaffolds of the actives and then to assess the enrichment in terms of novel unique scaffolds or frameworks in the top hits selected [144]. This is the criteria used in our research, where scaffold and framework definition is based on the Bemis and Murcko definition [54] and the detection of repeated or novel structures are done using a unique chemical graph identified based on the Xu and Johnson molecular equivalence indices [44] developed in our laboratory.

In **Chapter III.2** the ability for virtual target profiling of SHED is discussed and proven for the nuclear receptors family. This opens the door towards the full pharmacological profile of compounds for all targets with any ligand information available. This information is useful and can be applied for drug safety assessment, hit progressability assessment and drug repurposing. During the course of this thesis, ligand – target information has been gathered from diverse sources in order to accumulate as much information as we had available to build all possible ligand-based models. Specifically, a commercial annotated chemical library (ACL), the WOMBAT database [71] containing 186,114 annotated compounds and two ACLs developed in our laboratory directed to nuclear receptors [74] with 2,033 compounds and cytochrome P450, with 303 compounds have been used. Additionally, other publicly accessible ACLs, namely BindingDB [145] with 18,450 compounds, DrugBank [66] with 829 drugs, and PDSP Ki database [146] with 215 annotated compounds have been used. This has led to the construction of a final ACL containing 153,511 unique ligands with 426,376 annotations to 1,308 targets.

Two different sets of models have been generated using this information. In the first one, all the activity information below 10  $\mu$ M has been used while, in the second one, just those compounds with activity below 100 nM have been considered. Both these models have been used successfully in several projects, although there is still room for improvement. The usage of negative information has so far been neglected, although giving an alert for compounds with high similarity to other compounds with low activity or proven inactivity could be useful. Moreover, when information is available, the differentiation among agonist and antagonist models could also be useful, as this information is relevant in most drug discovery projects.

Another possible modification of our virtual screening and virtual profiling protocols is to adapt them to quantitative predictions instead of the current qualitative binary prediction (active / non-active). Up to date, the most common approach towards the quantitative prediction of the activity of a compound to a given protein has been the docking scoring functions. Despite this, ligand-based approaches are also suitable for these quantitative predictions, as a biased mean of the activity values of all the compounds previously tested for a target within a confidence range of distances can be used to estimate the activity of a query compound.

In **Chapter III.3**, targeted library design is presented as another of the applications of this developed methodology. After the ability of full profile prediction of large compound libraries has been proven, it can be used to design collections having a limited number of molecules with high probability of displaying the targeted pharmacological profile. This approach is already widely used in pharmaceutical industry as a way to improve cost effectiveness of early phases of the drug discovery process. Traditional high-throughput testing of a large maximally diverse library can be substituted by the testing of a smaller set specifically design for the project of interest.

Furthermore, given the great effort in gathering and organizing information for the generation of the models, we have attempted to fully exploit this information. Beyond considering only columns (in virtual ligand screening) or rows (in virtual target profiling) of this annotations matrix, analyzing



the matrix, or rather the network of connections that could be derived from it, as a whole is a way to get more information and a better perspective of the global process. In **Chapter III.4**, the applicability of the generated models for network pharmacology is proven, although the relevance of the completeness of the ligand – protein data is highlighted.

Hert *et al.* have recently highlighted the difference between sequence-based and ligand-based clustering of proteins. In this study, the different behavior of the networks constructed in the basis of these two approaches is stated. Moreover, the stability of ligand-based target network and its robustness to perturbations in ligand representation are proven [147]. This stresses the interest of careful ligand – target networks analysis, to draw relevant conclusions from those. On the other hand, it would be interesting to assess the differences between the ligand-based and the binding site structure-based classification of proteins. Furthermore, taking the shift in activity value into account when using this information for relating proteins [143] is an important issue that has not been considered in our study and that can be taken into account in following research.

Finally, all these methodologies have been made available through a tool that enables an easy application of any of them. This tool, called ViSCA, is a stand-alone program that can be run in command line and that enables to perform from very simple file management operations to virtual screening, virtual profiling and chemical annotation processes, as will be explained in **Chapter III.5**.



## Chapter II.1 – Future directions of research

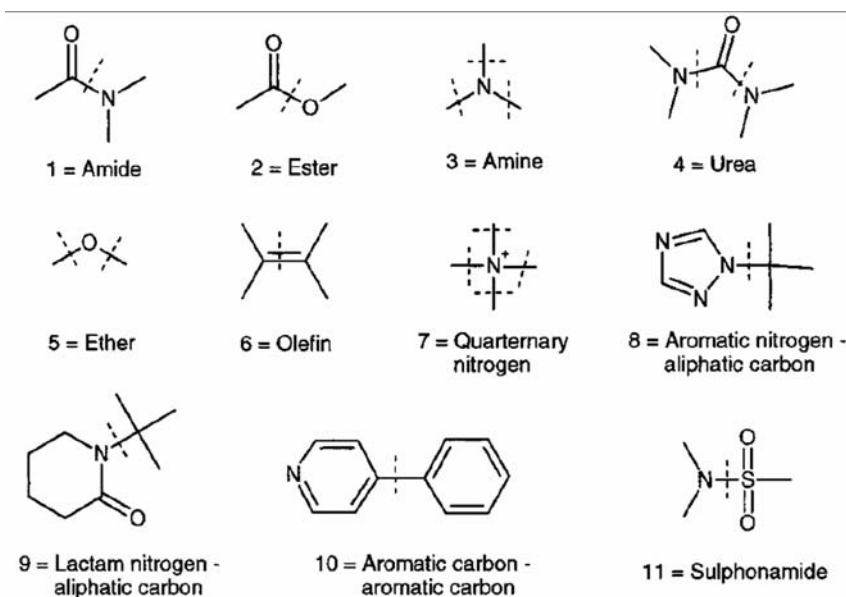
### II.1.1. Chemical structure hopping

The well-known pharmacologist and Nobel laureate Sir James Black is famously quoted to say that “the most fruitful basis for the discovery of a new drug is to start with an old drug”. Inspired by this statement, have begun to be developed SHIFT, a new approach to chemical Structure Hopping by Isosteric Fragment Transformations. In a previous step, a fragments database is generated by systematic fragmentation of a library of over 7 million unique commercially available compounds collected during this thesis. This fragmentation is based on the rules set by Wagener and Lommerse [148], that flag as cleavable all acyclic single bonds not being part of a ring or a functional group and either linking a carbon to a heteroatom or a branching point. To these rules a limitation has been added so no single terminal atom fragment is generated, in order to avoid very small and featureless fragments. Then, the query compound is also fragmented and chemical isosters can be sought in the database for each individual fragment in the molecule, as well as for each combination of two and three contiguous fragments, using SHED descriptors [98].

New molecules will then be built using the top ranked set of isosters to replace the query fragment. A molecule re-ranking with reference to the original molecule using also SHED descriptors will follow these isosteric transformations. In this way, the chemical space around the original molecule will be generated, which can be applied to generating structurally novel potentially active molecules from a reference compound, that being a drug, a natural product a competitor's new chemical entity, or an internal novel hit. The isosteric chemical transformation of this initial bioactive compound will provide a set of structurally different, yet pharmacophorically similar, compounds that are expected to retain (most of) the pharmacological profile of the parent molecule. This ability to move to new scaffolds can be of interest in situations where the natural ligand is known but synthetic inhibitors are not [144]. Alternatively, it can be used to break out of the protected patent space around competitor's compounds or when lead compounds have intractable chemistry, flat SAR or poor physicochemical properties [144]. Combining this chemical space exploration with an *in silico* target profiling method [149] can be ultimately used for prioritising hit series or projecting the pharmacological space relevant to a hit optimisation process.

One of the main issues to be taken into account in these *in silico de novo* ligand design methodologies is the synthetic accessibility of the generated proposals. Once the preferred generated compounds are selected, ideally all of them should be synthesized. The most common approach used currently to address this issue is the RECAP procedure [150], which is based on the splitting of molecules based on a set of 11 bond types pre-defined following a set of reaction

schemes that can be seen in **Figure 1**. These same rules are then used to rebuild the molecule, assuming that these same reactions can then be used in *in vitro* synthesis.



**Figure 1.** Eleven default bond cleavage types. Extracted from [150].

Other fragment replacement-based methods have been already developed, mainly based in 3D structure. Among them, CAVEAT [151] starts from a molecular structure and a selection of at least two outgoing bonds (exit vectors). From this basis, alternative molecular fragments with similar geometric arrangement of exit vectors are searched in a database of geometric relationships of bond pairs. Through the use of experimentally observed conformations, CAVEAT avoids artifacts generated by conformation generation programs. Another approach is Recore [152], which combines fast search with the crystal structure conformations used in CAVEAT. During the preprocessing phase, a database of 3D structures is converted into a fragments database using rules analogous to the RECAP procedure [150]. Then, after a drug-likeness filtering, a geometric rank-searching algorithm is applied based on 3D query, which consists on at least two exit vectors and an optional set of pharmacophoric features, to end up with the final fragments database.

## II.1.2. Systems chemical biology

During this thesis, a first approach to network pharmacology is presented. Ligand-based protein network analysis, as the one presented in **Chapter III.4**, is based on relating proteins in terms of the ligands they share. This type of analysis can be of use in target hopping, re-using chemical expertise from one widely explored target to another protein with highly related ligand-space. This re-using of information is a practice that has historically been used in pharmaceutical industry, although the basis for relating proteins has traditionally been phylogenetic relationships [153].

Moving further in this direction, the integration of chemical and biological spaces can be extended to applications to computational and systems biology, reaching systems chemical biology [27]. This approach seeks to describe all the elements of a biological system, define the biological networks that interrelate the elements of this system and characterize the flow of information that links these elements [154]. In order to construct a network that captures many aspects of how small molecules interact in a cellular context, it is clear to incorporate to ligand-target data the already existing molecular, cellular and organismal data such as microarray experiments and pathways information, as contained in databases like Reactome [155], KEGG [156] and MetaCyc [157]. Biological systems are intrinsically robust, and this property enables them to be resistant to various perturbations. Systems chemical biology will open the possibility of understanding the complex relationship between chemical structures and their effects in living systems. In an integrated interface, biochemical networks, target function and the effects of small molecules could be simulated.

Over the past decade, the entire industry has averaged only two or three small-molecule drugs against “innovative” targets per year [4]. Reverse pharmacology, based on target-based screening, has shown to perform successfully only for well-validated targets. Accordingly, the use of systems chemical biology to return to a forward pharmacology approach taking advantage of current advances in technology and knowledge is expected to be a promising way to increase the number of new chemical entities that reach the market. Rather than considering isolated proteins, the consideration of the biological system as a whole can be of use in a more rational target identification and recognition and avoidance of adverse drug reactions [29,154]. On one hand, the analysis of these networks will enable to identify those proteins not suitable as protein targets, which either constitute essential hubs or whose regulation can be by-passed and compensated by alternative paths. On the other hand, through the comparison of normal and diseased networks, critical proteins can be identified as potential drug targets.



## Part III - Publications

In this part, the results of the research carried out during this thesis and published in peer-reviewed journals are presented.





## Chapter III.1 – Molecular descriptors and ligand-based virtual screening

In this chapter, a new set of topological atom features-based descriptors called Shannon entropy descriptors (SHED) are presented. These molecular descriptors are based on the information-theoretical concept of Shannon entropy [1] applied to quantifying the variability displayed by topological distributions of atom-centered feature pairs in chemical structures. Examples of their possible uses in virtual ligand screening, scaffold hopping and ligand-based virtual target profiling are provided and their performance assessed and compared to well-established methodologies. Their capability of highlighting the common profiles of known actives for a certain target with different scaffolds while differentiating them from actives to other targets is proven.

Papers included in this chapter:

- **Gregori-Puigjané E, Mestres J:** SHED: Shannon entropy descriptors from topological feature distributions. *J Chem Inf Model* 2006, 46:1615–1622.



# SHED: Shannon Entropy Descriptors from Topological Feature Distributions

Elisabet Gregori-Puigjané and Jordi Mestres\*

Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain

A novel set of molecular descriptors called SHED (SHannon Entropy Descriptors) is presented. They are derived from distributions of atom-centered feature pairs extracted directly from the topology of molecules. The value of a SHED is then obtained by applying the information-theoretical concept of Shannon entropy to quantify the variability in a feature-pair distribution. The collection of SHED values reflecting the overall distribution of pharmacophoric features in a molecule constitutes its SHED profile. Similarity between pairs of molecules is then assessed by calculating the Euclidean distance of their SHED profiles. Under the assumption that molecules having similar pharmacological profiles should contain similar features distributed in a similar manner, examples are given to show the ability of SHED for scaffold hopping in virtual chemical screening and pharmacological profiling compared to BCI fingerprints and GRIND descriptors.

---

\* Corresponding author e-mail: [jmestres@imim.es](mailto:jmestres@imim.es)

## Introduction

The generation of mathematical representations for molecules has long been an active line of research in computational drug discovery. As a result, a large number and variety of molecular descriptors reflecting the one-dimensional, two-dimensional, and three-dimensional features of chemical structures have been devised [1]. Once formulated, the relevance of these descriptors is often established according to their ability to reflect the pharmacological properties of molecules [2]. Their potential impact in drug discovery is ultimately assessed when used, for example, in deriving quantitative structure-activity relationships from sets of molecules for which biological data is experimentally available [3], or performing similarity searches of large chemical libraries against a panel of reference active compounds [4].

Within this context, atom-centered feature pairs constitute an attractive family of molecular descriptors [5]. In their various formulations, they have been proven to show a decent performance on a diverse range of computational aspects in drug discovery covering quantitative structure-activity relationships [5], compound selection [2], virtual chemical screening [6-9], and virtual pharmacological profiling [10]. In all these studies, the actual computational encoding of the atom-centered feature pair descriptors attempts to capture their overall distribution within a molecule by storing the occurrence of feature pairs at different distance ranges, either at the topological [2,5-7] or geometrical level [8-10], to form a so-called binned fingerprint representation of each molecule. These molecular fingerprints are then used to assess the degree of resemblance between molecules according to different similarity metrics [11-13].

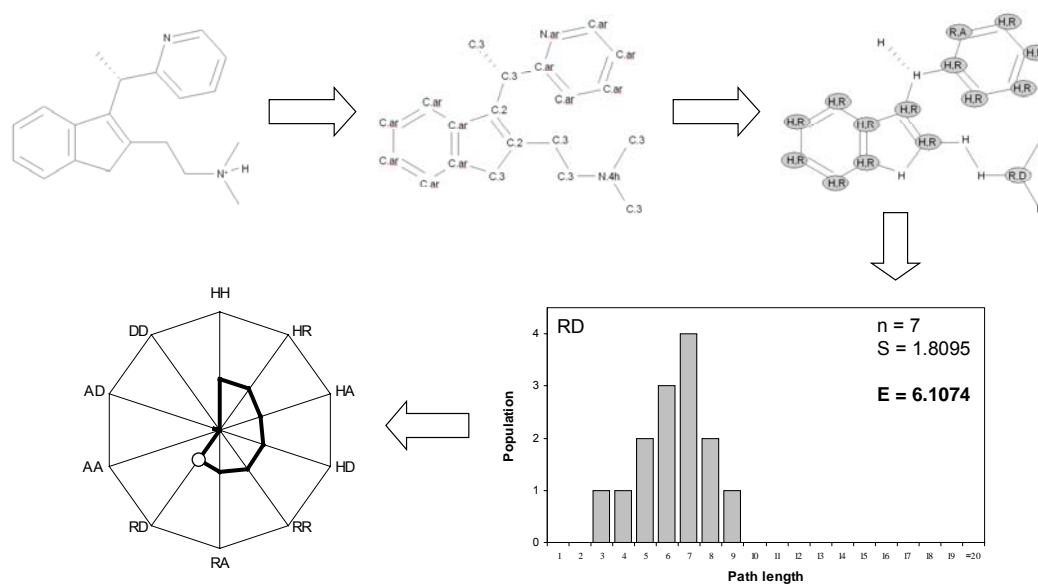
Two aspects are worth emphasizing to understand the scope and limitations of this family of descriptors. On one hand, formulations based on geometric atom pair descriptors require three-dimensional coordinates and thus provide representations which are dependent on the conformation of molecules [8-10]. Consequently, this involves 2D to 3D conversion of atomic coordinates and potentially the generation of multiple conformers for each molecule. On the other hand, formulations based on topological atom pair descriptors do not need three-dimensional coordinates and do not have this conformational dependency [2,5-7]. However, they result in crisp representations of molecules that may not capture some of the essential information present only when using three-dimensional coordinates. A formulation based on a fuzzy description of topological features could be a good balance between the two approaches.

It is along these lines that a novel set of molecular descriptors is introduced. This new formulation takes advantage of the information-theoretical concept of Shannon entropy [14], an approach that is increasingly being applied to process chemical information [15-18]. Accordingly, these new descriptors will be referred to as SHED, for SHannon Entropy Descriptors, and represent a means to quantify the variability displayed by topological distributions of atom-centered feature pairs in molecules. The following sections describe, first, the methodological details for

obtaining SHED and, second, several application examples to assess their potential ability for scaffold hopping in virtual chemical screening and pharmacological profiling.

## Methods

The process of obtaining SHED from chemical structure is illustrated in **Figure 1** for dimetindene, a histamine H1 antagonist. The original input structure should be in MDL's SD file format [19]. From a SD file, each atom in a molecule is first mapped to a Sybyl atom type [20]. Subsequently, each atom type is assigned currently to one or more of four atom-centered features, namely, hydrophobic (H), aromatic (R), acceptor (A), and donor (D). For example, an aliphatic C.3 carbon will be assigned to a hydrophobic feature (H), whereas a protonated N.4h nitrogen will be assigned to both aromatic and donor features (R,D). Then, the shortest path length between atom-centered feature pairs is derived and its occurrence at different path lengths stored to create a feature-pair distribution. A maximum path length of 20 bonds was used. Feature pairs being at distances over 20 bonds are accumulated in the last bin. As an example, the distribution of RD feature pairs within dimetindene is displayed. An equivalent distribution is derived for each of the ten possible feature pairs resulting from all pair combinations of the four features used.



**Figure 1.** Generation of a SHED profile from chemical structure

At this stage, the concept of Shannon entropy [14] is applied to determine the variability of feature-pair distributions. Within this approach, the entropy,  $S$ , of a population,  $P$ , distributed in a certain number of bins (representing in this case the different path lengths),  $N=20$ , is given by

$$S = - \sum_{i=1}^N \rho_i * \ln \rho_i \quad ; \quad \rho_i = p_i / P \quad (1)$$

where  $\rho_i$  and  $p_i$  are, respectively, the probability and the population at each bin  $i$  of the distribution. The values of  $S$  range between 0, reflecting the situation of all population being concentrated in a single bin, and a maximum number,  $S_{max}=\ln N$ , reflecting the situation of a uniformly distributed population among all bins. In the case of dimetindene (**Figure 1**), RD pairs can be found at path lengths occupying 7 bins and the variability in their population gives rise to a distribution with an entropy value of 1.8095. In order to have a more intuitive measure that can be linearly related to the situation of full uniform occupancy, entropy values are transformed into projected entropy values,  $E=e^S$ . Correspondingly,  $E$  values provide a measure of the expected maximum uniform occupancy from the corresponding  $S$  value. Now, for any given population  $P>0$ , the values of  $E$  can vary from 1, reflecting the situation of zero entropy in which the population is totally concentrated in a single bin, to  $N$ , reflecting the situation of maximum entropy in which the population is uniformly distributed among all bins. In the limit case of  $P=0$ , then  $E$  will be assigned to  $E=0$ . For the RD feature pair in dimetindene (**Figure 1**) the maximum achievable  $E$  value for a population occupying uniformly 7 bins would be  $E=7$ . The obtained  $E$  value of 6.1074 reflects a slight deviation from the situation of full uniform occupancy on 7 bins. This  $E$  value will ultimately be the Shannon entropy descriptor (SHED) for the RD feature pair. The set of SHED values obtained for the ten possible feature pairs constitute the SHED profile of a molecule. As illustrated in **Figure 1**, SHED profiles are represented using a wheel chart, the circle in the chart indicating the  $E$  value (SHED) for the RD feature pair in dimetindene.

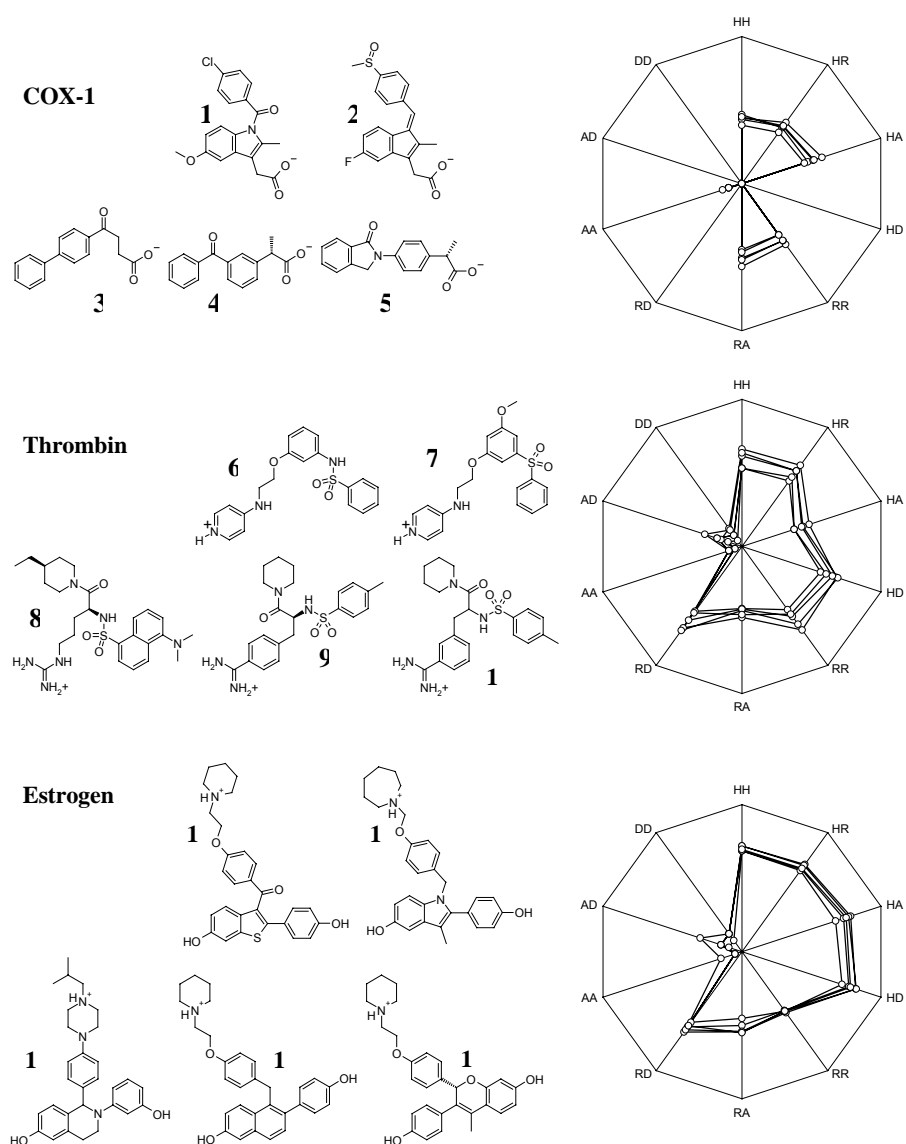
## Results and discussion

The basic assumption is that molecules having similar features arranged in a similar way should display similar SHED profiles. The underlying question is to which extent SHED profiles derived from topology-based atom-centered feature-pair distributions of molecules are well suited to recognize the presence of similar features arranged in a similar way around significantly different molecular scaffolds, an ability usually referred to as “scaffold hopping” [6]. To investigate this issue, an analysis of SHED profiles for molecules directed to different targets is presented next, followed by application examples on the use of SHED profiles for virtual chemical screening and pharmacological profiling.

**Scaffold hopping.** Three diverse sets of molecules containing comparable features arranged similarly around essentially different scaffolds were selected. The structures of the molecules and their corresponding SHED profiles are collected in **Figure 2**. The first set includes a list of five known cyclooxygenase-1 inhibitors (COX-1: EC 1.14.99.1), namely, indomethacin (**1**), sulindac (**2**), fenbufen (**3**), ketoprofen (**4**), and indoprofen (**5**); the second set is a selection of five known thrombin inhibitors (Factor IIa: EC 3.4.21.5), namely, BM14.1224 (**6**), BM51.1047 (**7**), DAPA (**8**), 4-TAPAP (**9**), and 3-TAPAP (**10**); and the third set contains five estrogen receptor subtype  $\alpha$  antagonists (ER $\alpha$ : NR 3.A.1), namely, raloxifene (**11**), bazedoxifene (**12**), a tetrahydroisoquinoline ligand (**13**), LY326315 (**14**), and EM343 (**15**). As can be observed, despite the significant scaffold diversity present in the three families of chemical structures, reasonably equivalent SHED profiles

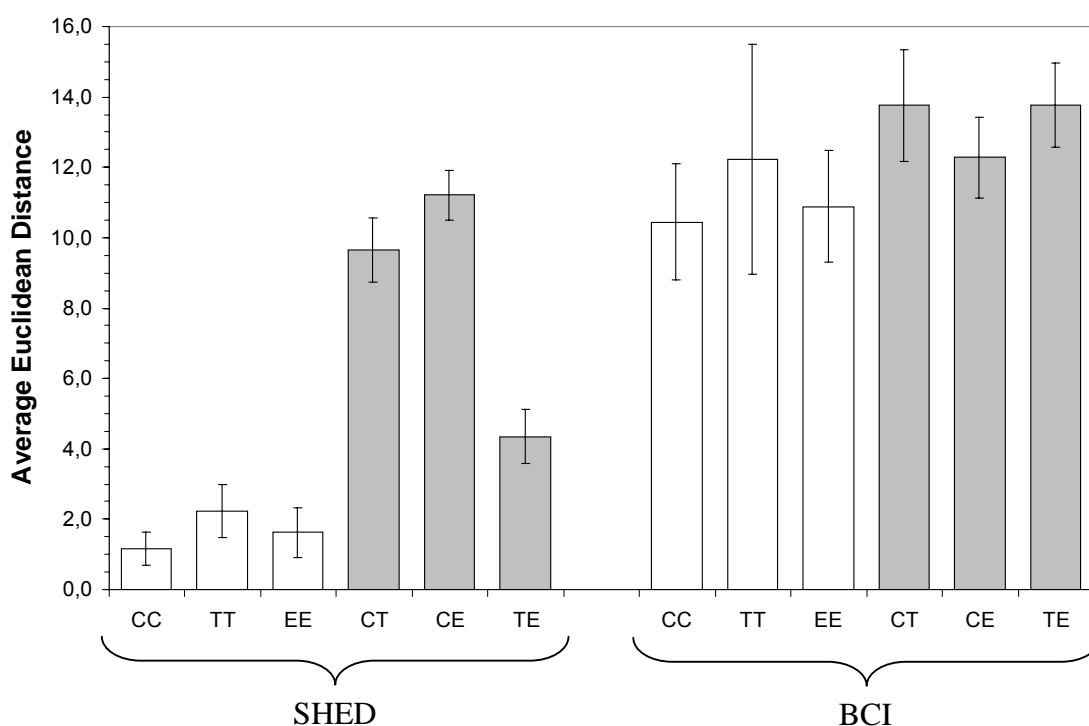
are obtained within each set. At the same time, the various target-directed SHED profiles are found essentially different from one another. Altogether, these results provide evidence of the potential applicability of SHED for identifying molecules containing similar features distributed similarly around diverse molecular scaffolds.

A characteristic worth emphasizing is the fact that size appears to be implicitly accounted for in the SHED profile of molecules, particularly in the SHED values corresponding to feature pairs involving hydrophobic and aromatic centers. For example, average values and standard deviations of SHED for the HH pairs found in COX-1 inhibitors, thrombin inhibitors, and ER $\alpha$  antagonists are, respectively,  $4.40\pm 0.27$ ,  $5.95\pm 0.61$ , and  $7.01\pm 0.17$ , which are indicative of the increasing size of compounds associated with those targets. In general, as compounds become bigger and more complex (in terms of combinations of features), the area filled by their SHED profiles would tend to be larger as well.



**Figure 2.** SHED profiles for three diverse sets of target-directed molecules, namely, cyclooxygenase inhibitors (top), thrombin inhibitors (middle), and estrogen  $\alpha$  antagonists (bottom)

In order to assess quantitatively the degree of discrimination obtained when comparing different types of molecules, average Euclidean distances and standard deviations derived from the SHED profiles of the three sets of ligands are presented in **Figure 3**. It is remarkable the clear separation obtained between comparisons of pharmacophorically similar molecules and comparisons of molecules having essentially different feature distributions. For example, average intra-set distances and standard deviations between COX-1, thrombin, and estrogen ligands are  $1.17\pm 0.46$ ,  $2.23\pm 0.75$ , and  $1.62\pm 0.71$ , respectively. In contrast, average inter-set distances and standard deviations when comparing COX-1/Thrombin, COX-1/Estrogen, and Thrombin/Estrogen ligands are  $9.65\pm 0.92$ ,  $11.21\pm 0.70$ , and  $4.35\pm 0.78$ , respectively. To put these results into perspective, the same exercise was done using BCI fingerprints [21], a representative 2D substructural fingerprint-based method widely used in compound clustering and similarity searching [22]. In this case, discrimination between intra-set and inter-set comparisons is not as clear as with SHED profiles and a more fuzzy (less separation between intra-set and inter-set distances) and less compact (larger values for standard deviations) picture appears. This outcome emphasizes the potential use of SHED in virtual screening applications.



**Figure 3.** Average Euclidean distances and standard deviations obtained when using SHED profiles and BCI fingerprints for the three sets of ligands active to COX-1 (C), thrombin (T), and estrogen (E) shown in **Figure 2**. White bars correspond to intra-set distances (e.g., CC refers to the 10 non-zero Euclidean distances between all pairwise combinations of the five COX-1 inhibitors) and gray bars correspond to inter-set distances (e.g., CT refers to the 25 Euclidean distances between all pairwise combinations of the five COX-1 and the five thrombin inhibitors)

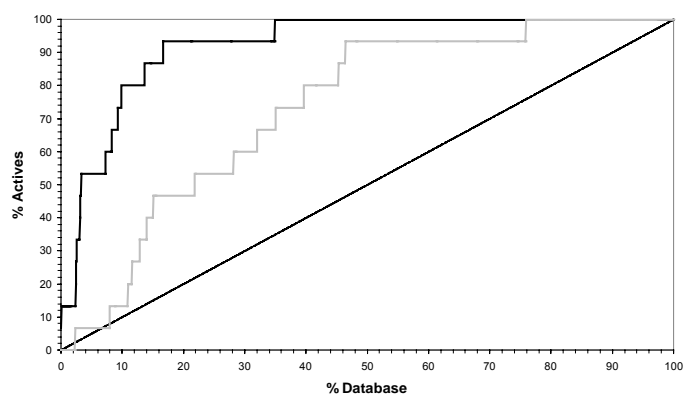
Despite the attractive resemblance observed in the target-directed SHED profiles, both qualitatively in **Figure 2** and quantitatively in **Figure 3**, those compounds represent only a focused subset extracted from the ample diversity of active compounds that could be identified and generated for a given target. In fact, having similar SHED profiles may well be a reflection of



making analogous interactions within similar pockets in their respective targets. In reality, depending largely on the size of the protein binding cavity, its flexibility, and its degree of exposure to the solvent, ligands will bind to different pockets, exploit different interactions, and have a variety of solvent exposed functional groups. In this scenario, a less compact picture will certainly emerge when visualizing the corresponding SHED profiles, but yet each SHED profile will be representative of a particular distribution of features present in a known active compound for a given target.

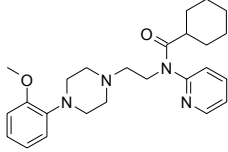
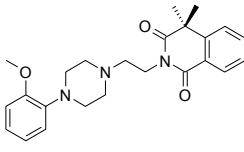
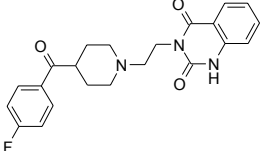
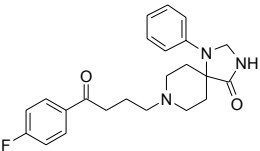
**Virtual chemical screening.** Computational drug discovery of novel chemical modulators for members of the therapeutically-relevant family of G protein-coupled receptors (GPCRs) is still dominated by ligand-based approaches mainly due to, on one hand, the technical difficulties encountered in crystallizing these receptors and, on the other hand, the large amount of biological data available for molecules acting on these receptors. Accordingly, in order to illustrate the applicability and performance of SHED for virtual chemical screening on GPCRs, the structures of a set of 24 highly diverse  $\alpha_{1A}$ -adrenoceptor antagonists with known binding affinities ( $K_i < 300$  nM) were extracted from a recent publication [23]. The SHED profiles derived from those 24 reference compounds were then used to score a database composed of 3033 drugs and a test set of 15  $\alpha_{1A}$ -adrenoceptor antagonists [23]. The scoring of each compound in the database was simply assigned to the minimum value of all Euclidean distances calculated between the SHED profile of the compound and each one of the 24 reference SHED profiles.

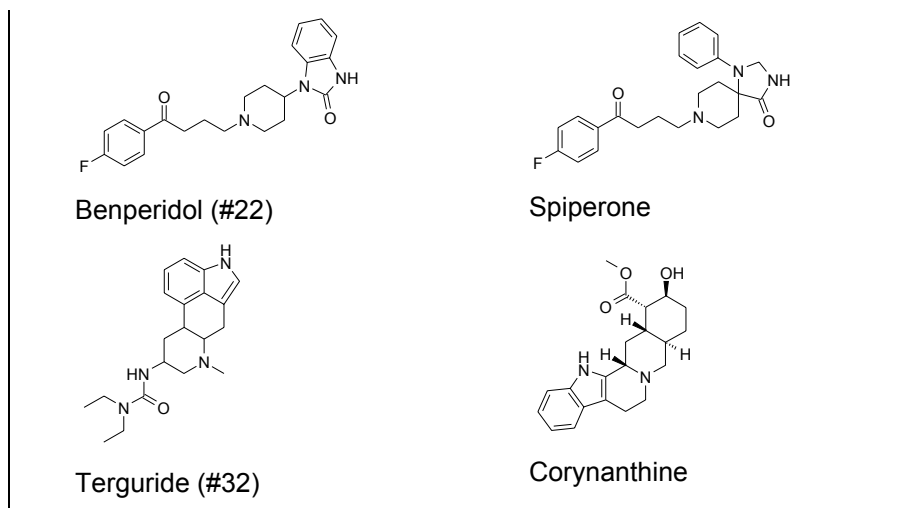
The percentage of actives found with SHED within each percentage of the database is plotted in **Figure 4** (black bold line). After rank ordering, selection of the top-ranked 5% and 10% of compounds in the database would have included 53.3% (8) and 80.0% (12), respectively, of the 15  $\alpha_{1A}$ -adrenoceptor antagonists in the test set. In terms of overall enrichment, in the ideal situation that all 15 actives were found in the 15 top-ranked compounds, the value for the normalized area under the curve (AUC) would be 0.9975 ( $AUC_T$ ), whereas a random identification of actives (symbolized by the thin diagonal line in **Figure 4**) would result in a normalized AUC of 0.5000 ( $AUC_R$ ). Correspondingly, the normalized AUC of the resultant active identification line is 0.9220. From these AUC values, an enrichment factor can be defined as  $E = (AUC - AUC_R) / (AUC_T - AUC_R)$ . This enrichment factor can have values in the range of [-1.0, 1.0]. A value of  $E = -1.0$  would reflect the worst scenario of finding all actives in the database in the last bottom-ranked compounds, whereas a value of  $E = 1.0$  would reflect the ideal situation of finding all actives in the database in the first top-ranked compounds. A random identification of actives would result in  $E = 0.0$ . Based on this definition, the current virtual screening returns an enrichment of  $E = 0.8482$ .



**Figure 4.** Enrichment curve for the retrieval of 15  $\alpha_{1A}$ -adrenoceptor antagonists from a database of 3033 drugs, using a reference set of 24  $\alpha_{1A}$ -adrenoceptor antagonists using SHED (black line) and GRIND (gray line) descriptors

For the sake of comparison, alignment-independent GRIND descriptors derived from three-dimensional molecular interaction fields were also calculated for all compounds using the program ALMOND [24] with three-dimensional structures derived by CORINA [25]. In this case, the scoring of each compound in the database was assigned to the minimum value of all Euclidean distances calculated between the first three scaled principal component analysis scores of the compound and those corresponding to each one of the 24 reference compounds. The percentage of actives found with GRIND within each percentage of the database is plotted in **Figure 4** (gray bold line). After rank ordering, selection of the top-ranked 5% and 10% of compounds in the database would have included 6.7% (1) and 13.3% (2), respectively, of the 15  $\alpha_{1A}$ -adrenoceptor antagonists in the test set. In terms of overall enrichment, the normalized AUC of the resultant active identification line is 0.7340, which corresponds to an enrichment of  $E=0.4704$ . Accounting for more components to calculate the Euclidean distance between compounds did not have an effect on the AUC and  $E$  values presented above.

Drug	$\alpha_{1A}$ -Adrenoceptor Antagonist
 WAY 100635 (#4)	 ARC 239
 Ketanserin (#14)	 Spiperone

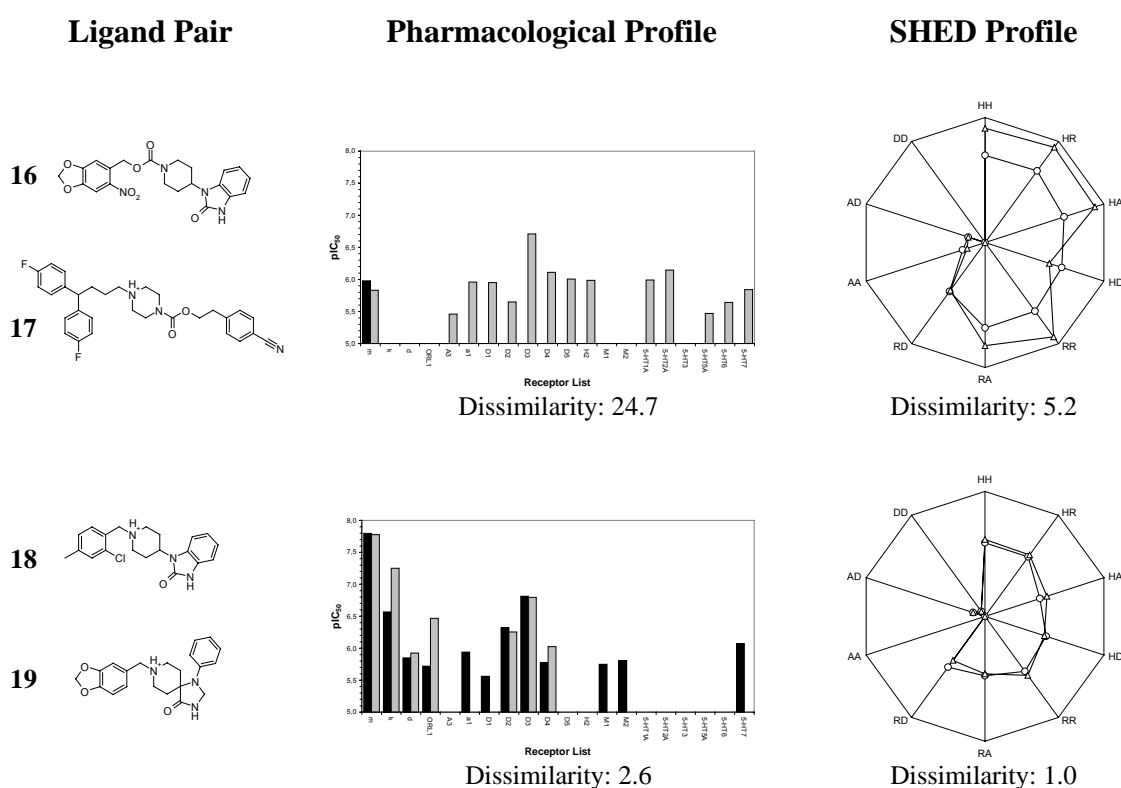


**Figure 5.** Selection of top-ranked drugs identified in the virtual chemical screening for  $\alpha_{1A}$ -adrenoceptor antagonists. Reference  $\alpha_{1A}$ -adrenoceptor antagonist having the closest SHED profile to each drug is also included. Rank position of drugs is given in parenthesis

The performance of SHED for identifying the test set of 15  $\alpha_{1A}$ -adrenoceptor antagonists within the top positions of the rank ordered database may have been masked to some extent by the presence of drugs that might as well have some affinity for the  $\alpha_{1A}$ -adrenoceptor. For instance, droperidol and trazodone were found at ranks 9 and 29, respectively, and were neither in the reference nor in the test set of  $\alpha_{1A}$ -adrenoceptor antagonists. Some examples of drugs that were found within the top ca. 1% of the rank ordered database, together with the reference  $\alpha_{1A}$ -adrenoceptor antagonist with the closest SHED profile to that of each drug, are collected in **Figure 5**. At rank 4 we found WAY 100635, an antagonist of the serotonin 5-HT<sub>1A</sub> receptor. The closest  $\alpha_{1A}$ -adrenoceptor antagonist found in the reference set is ARC 239 ( $pK_i=9.0$ ) [23]. Despite having essentially different scaffolds, the similarities between the structural features of both compounds are remarkable. This result would thus alert on the possibility of WAY 100635 hitting the  $\alpha_{1A}$ -adrenoceptor as an off-target. In fact, evidence can be found in the literature that WAY 100635 induces hypotension in anaesthetized rats and that this effect could be partially explained by antagonism of vascular  $\alpha_1$ -adrenoceptors [26]. Ketanserin, a serotonin 5-HT<sub>2</sub> antagonist, was found at rank 14. The closest  $\alpha_{1A}$ -adrenoceptor antagonist found in the reference set is Spiperone ( $pK_i=8.1$ ), with which Ketanserin shares a similar distribution of pharmacophoric features [23]. Most interestingly, Ketanserin was recently reported to be a potent antagonist for the  $\alpha_{1A}$ -adrenoceptor ( $pK_i=8.0$ ) [27]. Also Benperidol, a dopamine D<sub>2</sub> receptor antagonist, is found at rank 22. As for Ketanserin, Spiperone was the closest reference  $\alpha_{1A}$ -adrenoceptor antagonist to Benperidol. Benperidol is structurally related to Droperidol and shows a striking resemblance with Spiperone, suggesting that the  $\alpha_{1A}$ -adrenoceptor could well be an off-target for Benperidol. A final fourth example was extracted from rank 32, where Terguride, a dopamine D<sub>2</sub> partial agonist, was located. The closest  $\alpha_{1A}$ -adrenoceptor antagonist found in the reference set is Corynanthine ( $pK_i=7.5$ ) [23]. The two compounds present no obvious structural similarities but yet the relative distribution of the different features appears to be remarkably equivalent. Of mention is the fact that a recent study confirmed experimentally that Terguride displays indeed potent antagonist properties at the  $\alpha_{1A}$ -

adrenoceptor ( $pK_b=8.0$ ) [28]. On the basis of these results, the potential use of SHED profiles beyond virtual chemical screening will be investigated next.

**Virtual pharmacological profiling.** The basic assumption sustaining virtual chemical screening activities is that similar compounds are expected to have similar affinities for a given target. However, further than having affinity for a particular target, the ultimate biological effect of compounds is established by their pharmacological profile against a set of biologically relevant targets [29]. Therefore, for virtual pharmacological profiling, the original statement can be extended to assuming that similar compounds should display similar pharmacological profiles. To explore this aspect further, we took a list of 47 compounds for which the experimental pharmacological profile on a panel of 75 targets was known [30]. The majority of compounds had some affinity for the  $\mu$  opiate receptor, but against a panel of diverse targets they displayed essentially different pharmacological profiles. As illustrative examples, two selected pairs of compounds are shown in **Figure 6**. Despite the apparent structural differences, compounds **16** and **17** had similar affinity for the  $\mu$  opiate receptor, with  $pIC_{50}$  values of 5.98 and 5.83, respectively. However, while the pharmacological profile of compound **16** shows a high selectivity toward the  $\mu$  opiate receptor, compound **17** has a poor specificity, with low micromolar affinity for 14 out of 20 GPCR targets. In contrast, compounds **18** and **19** had not only similar high affinities for the  $\mu$  opiate receptor, with  $pIC_{50}$  values of 7.80 and 7.78, respectively, but showed also comparable overall pharmacological profiles against the panel of 20 GPCRs.



**Figure 6.** Structures, pharmacological profiles (extracted from ref. 22), and SHED profiles for two selected pairs of molecules

Having the ability to estimate in advance potential deviations in the pharmacological profiles of a list of hits identified in the early stages of a drug discovery project would be of great value to prioritize further optimization activities on those hits. Recent studies have explored the use of similarity metrics to analyze the relationship between the degree of pharmacophore similarity in a pair of compounds and the similarity of their respective pharmacological profiles [30]. Accordingly, the extent to which the SHED profiles introduced in this work reflect the relative pharmacological profiles of compounds is an aspect worth investigating at this stage. To this aim, the SHED profiles for the two pairs of compounds described above have been included also in **Figure 6**. Comparison of the SHED profiles obtained for compounds **16** and **17** evidences dissimilar feature distributions, the profile for compound **17** covering clearly a wider SHED area consistent with the larger size of this molecule relative to **16**. In contrast, compounds **18** and **19** showed visibly similar SHED profiles, consistent with a pair of compounds of approximately the same size containing similar features arranged similarly.

At a more quantitative level, as proposed in a previous study [30] an activity dissimilarity score of the pharmacological profiles of two compounds can be defined as

$$D(A,B) = \sum \Delta(|\%inh_i(A) - \%inh_i(B)|)$$

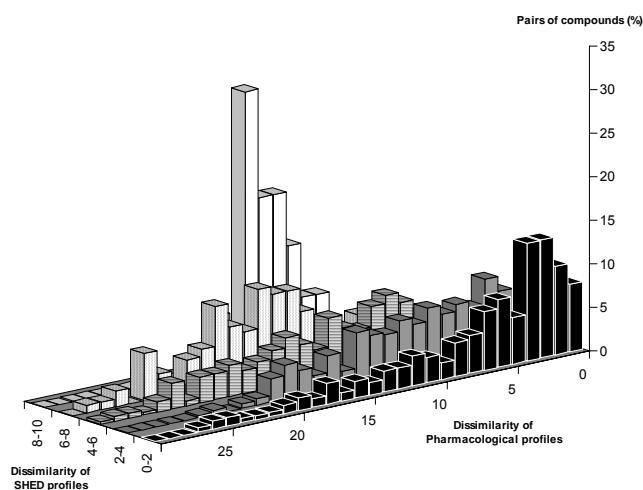
where  $\%inh_i$  stand for the percentages of inhibition (at 10  $\mu$ M) of A and B in the test  $i$  among 75 tests of the profile. The term

$$\Delta (|\%inh_i(A) - \%inh_i(B)|)$$

was originally defined as an empirical measure of how different the two compounds behave with respect to that test:

$$\Delta (x) = \begin{cases} 0 & \text{if } x \leq 30 \\ (x - 30)/40 & \text{if } 30 < x \leq 70 \\ 1 & \text{if } x > 70 \end{cases}$$

As an example, **Figure 6** contains also the values obtained for the dissimilarity of pharmacological profiles and SHED profiles between the two illustrative pairs of compounds. The visual observation that the pharmacological profiles of compounds **16** and **17** differ much more than those of compounds **18** and **19** is reflected in dissimilarity values of 24.7 and 2.6, respectively. Correspondingly, the differences observed in the SHED profiles of compounds **16** and **17** compared to those of compounds **18** and **19** result in dissimilarity values of 5.2 and 1.0, respectively. These results are illustrative of the potential use of SHED profiles as a means for alerting of likely differences or similarities in the pharmacological profiles of series of molecules for which experimental affinities are available only for a particular target.



**Figure 7.** Dissimilarity of pharmacological profiles versus dissimilarity of SHED profiles for all compound pairs derived from a list of 47 compounds for which experimental affinity data on 75 targets are available (extracted from ref. 22)

To investigate this issue further, SHED and pharmacological profile dissimilarities were evaluated for each of the pairwise combinations of the 47 compounds, in an analogous way as reported in the previous study from which the structures and experimental binding affinities of compounds were extracted [30]. Then, all pairs of compounds were sorted into dissimilarity categories. The resultant distribution of the pairs in each category versus the dissimilarity scores obtained for the calculated SHED profiles and the observed pharmacological profiles is given in **Figure 7**. For each pair of compounds, dissimilarity of SHED profiles refers to the Euclidean distance between their SHED profiles, whereas dissimilarity of pharmacological profiles accounts for the differences in their respective percentages of inhibition accumulated over the entire set of biologically relevant targets to which the compounds were tested against, as detailed above [30]. Remarkably, the overall distribution presented in **Figure 7** reproduces qualitatively the results obtained in the earlier parent study (see Figure 2 in ref. 22). In the class of compounds having the most similar feature distributions (dissimilarity of SHED profiles  $\leq 2$ ), the occurrence of pharmacologically similar compound pairs (dissimilarity of pharmacological profiles  $\leq 5$ ) is significantly high. With increasing dissimilarity of SHED profiles, the relative probability of finding pairs of compounds with similar pharmacological profiles decreases considerably. Consequently, pairs of compounds having similar SHED profiles are more likely to have similar pharmacological profiles than any random pair of dissimilar molecules.

## Conclusions

We have introduced SHED as a novel set of molecular descriptors based on the information-theoretical concept of Shannon entropy applied to quantifying the variability displayed by topological distributions of atom-centered feature pairs in chemical structures. Under this new representation, molecules containing comparable features arranged similarly around essentially different scaffolds give rise to similar SHED profiles, illustrated in this work for the cases of

cyclooxygenase-1 inhibitors, thrombin inhibitors, and estrogen  $\alpha$  antagonists. This property was then further assessed in a virtual chemical screening exercise to retrieve 15  $\alpha_{1A}$ -adrenoceptor antagonists from a database of 3033 drugs. Using a reference set of 24 diverse  $\alpha_{1A}$ -adrenoceptor antagonists, selecting the top-ranked 10% of compounds in the database would have included 80.0% of the  $\alpha_{1A}$ -adrenoceptor antagonists in the test set, with an overall enrichment factor of 0.8482. In addition, the use of SHED helped identifying  $\alpha_{1A}$ -adrenoceptor as a potential off-target for several top-ranked drugs. Finally, SHED profiles demonstrated a decent performance for estimating differences in the virtual pharmacological profiling of molecules, with pairs of compounds having similar SHED profiles showing a trend toward having also similar pharmacological profiles.

Given the large amount of experimental data available currently on the affinity of molecules to targets, ligand-based approaches to drug discovery remain still competitive against more sophisticated structure-based methods. In the view of the results presented here, the use of SHED appears as a simple, yet attractive, low-dimensional representation of molecules with promising applicability for the virtual identification and profiling of novel hits at the early stages of drug discovery projects.

## ACKNOWLEDGMENTS

This research is supported by a grant from the Instituto de Salud Carlos III (Ministerio de Sanidad y Consumo), research project reference number 02/3051. We are indebted to Manuel Pastor (Universitat Pompeu Fabra, Spain) for deriving the GRIND descriptors, Xavier Fradera (Organon, Scotland) for obtaining the BCI fingerprints, and Rafael Gozalbes (Cerep, France) for generating **Figure 7**.

**Supporting Information Available:** The mapping of Sybyl atom types to pharmacophoric features, feature-pair distributions and SHED profile for dimetindene, SHED profiles for COX-1, thrombin, and estrogen ligands, and the structures of the 39  $\alpha_{1A}$ -adrenoceptor antagonists used in the virtual chemical screening. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

1. Todeschini R, Consonni V: *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH; 2000.
2. Brown RD, Martin YC: **Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection**. *J. Chem. Inf. Model*. 1996, **36**:572-584.
3. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD: **Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology**. *Chem. Rev*. 2002, **102**:783-812.
4. Maggiora GM, Johnson MA: *Concepts and applications of molecular similarity*. New York: John Wiley & sons; 1990.
5. Carhart RE, Smith DH, Venkataraghavan R: **Atom pairs as molecular features in structure-activity studies: definition and applications**. *J Chem Inf Model* 1985, **25**:64-73.
6. Schneider G, Neidhart W, Giller T, Schmid G: **"Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening**. *Angewandte Chemie International Edition* 1999, **38**:2894-2896.
7. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E: **Similarity metrics for ligands reflecting the similarity of the target proteins**. *J Chem Inf Model* 2003, **43**:391-405.
8. Sheridan RP, Miller MD, Underwood DJ, Kearsley SK: **Chemical similarity using geometric atom pair descriptors**. *J Chem Inf Comp Sci* 1996, **36**:128-136.
9. Makara GM: **Measuring molecular similarity and diversity: total pharmacophore diversity**. *J Med Chem* 2001, **44**:3563-3571.
10. Horvath D, Jeandenans C: **Neighborhood Behavior of in Silico Structural Spaces with Respect to In Vitro Activity Spaces-A Benchmark for Neighborhood Behavior Assessment of Different in Silico Similarity Metrics**. *J. Chem. Inf. Model*. 2003, **43**:691-698.
11. Willett P, Barnard JM, Downs GM: **Chemical Similarity Searching**. *J. Chem. Inf. Comput. Sci*. 1998, **38**:983-996.
12. Maggiora GM, Petke JD, Mestres J: **A general analysis of field-based molecular similarity indices**. *J Math Chem* 2002, **31**:251-270.
13. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to in silico pharmacology: nuclear receptor profiling**. *J Chem Inf Model* 2006, **46**:2725-2736.
14. Shannon CE, Weaver W: *The mathematical theory of communication*. Urbana, IL: University of Illinois Press; 1949.
15. Godden JW, Stahura FL, Bajorath J: **Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations**. *J Chem Inf Comp Sci* 2000, **40**:796-800.
16. Stahura FL, Godden JW, Bajorath J: **Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binaty QSAR calculations**. *J Chem Inf Comp Sci* 2000, **40**:1245-1252.
17. Miller JL, Bradley EK, Teig SL: **Luddite: an information-theoretic library design tool**. *J Chem Inf Comp Sci* 2003, **43**:47-54.



18. Graham DJ: **Information content in organic molecules: aggregation states and solvent effects.** *J Chem Inf Model* 2005, **45**:1223-1236.
19. Elsevier MDL, San Leandro, CA: <http://www.mdl.com/>. (accessed May 2006).
20. Tripos Inc, St. Louis, MO: [http://www.tripos.com/mol2/atom\\_types.html](http://www.tripos.com/mol2/atom_types.html). (accessed May 2006).
21. Barnard Chemical Information Ltd, Leeds, UK: <http://www.bci.gb.com>. (accessed May 2006).
22. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.** *J Chem Inf Model* 2004, **44**:1177-1185.
23. Li MY, Tsai KC, Xia L: **Pharmacophore identification of  $\alpha_{1A}$ -adrenoceptor antagonists.** *Bioorg Med Chem Lett* 2005, **15**:657-664.
24. Pastor M, Cruciani G, McLay I, Pickett SD, Clementi S: **Grid-independent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors.** *J Med Chem* 2000, **43**:3233-3243.
25. Molecular Networks, GmbH, Erlangen, Germany: <http://www.mol-net.de/>. (accessed May 2006).
26. Villalobos-Molina R, Lopez-Guerrero JJ, Gallardo-Ortiz IA, Ibarra A: **Evidence that the hypotensive effect of WAY 100635, a 5-HT<sub>1A</sub> receptor antagonist, is related to vascular alpha 1-adrenoceptor blockade in the adult rat.** *Auton Autacoid Pharmacol* 2002, **22**:171-176.
27. Israilova M, Suzuki F, Tanaka T, Nagamoto T, Taniguchi T, Muramatsu I: **Binding and functional affinity of sarpogrelate, its metabolite m-1 and ketanserin for human recombinant alpha-1-adrenoceptor subtypes.** *Pharmacology* 2002, **65**:69-73.
28. Newman-Tandredi A, Cussac D, Audinot V, Nicolas JP, de Ceuninck F, Boutin JA, Millan MJ: **Differential actions of antiparkinson agents at multiple classes of monoaminergic receptor II. Agonist and antagonist properties ar subtypes of dopamine D2-like receptor and  $\alpha_{1A}$  adrenoceptor.** *J Pharmacol Exp Ther* 2002, **303**:805-814.
29. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan ML, Arkin AP, Davis RW: **Chemogenomic profiling: identifying the functional interactions of small molecules in yeast.** *Proc Natl Acad Sci U S A* 2004, **101**:793-798.
30. Poulain R, Horvath D, Bonnet B, Eckhoff C, Chapelain B, Bodinier MC, Déprez B: **From hit to lead. Analyzing structure-profile relationships.** *J Med Chem* 2001, **44**:3391-3401.

## SUPPORTING INFORMATION

**Table Mapping Sybyl Atom Types and Pharmacophoric Features**

Atom type	Pharmacophore features			
	H	R	A	D
C.1	X	X		
C.2	X	X		
C.3	X			
C.ar	X	X		
C.cat		X		
N.1		X	X	
N.2		X	X	
N.2h		X	X	X
N.3			X	
N.3h			X	X
N.4		X		
N.4h		X		X
N.ar		X	X	
N.arh		X		X
N.am	X	X		
N.amh		X		X
N.pl3		X		
N.plh		X		X
O.2		X	X	
O.3	X			
O.3h			X	X
O.co2		X	X	
Si	X			
P.3			X	
P.3h			X	X
P.o2	X	X		
S.2		X		
S.2h		X		X
S.3	X			
S.3h			X	X
S.o	X	X		
S.o2	X	X		
F	X		X	
Cl	X			
Br	X			
I	X			

## SUPPORTING INFORMATION

### Feature-Pair Distributions and SHED Profile for Dimetindene

<Feature-pair population distributions>

H-H: 19 29 29 28 25 23 19 11 5 2 0 0 0 0 0 0 0 0 0 0  
H-R: 22 29 28 29 29 28 21 15 6 2 0 0 0 0 0 0 0 0 0 0  
H-A: 2 3 3 2 4 3 1 2 0 0 0 0 0 0 0 0 0 0 0 0  
H-D: 3 1 1 2 3 4 3 2 1 0 0 0 0 0 0 0 0 0 0 0  
R-R: 14 17 15 12 13 14 12 6 2 0 0 0 0 0 0 0 0 0 0  
R-A: 2 2 2 2 2 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0  
R-D: 0 0 1 1 2 3 4 2 1 0 0 0 0 0 0 0 0 0 0 0  
A-A: 0  
A-D: 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0  
D-D: 0

<SHED profile>

H-H: 4.2821  
H-R: 4.3533  
H-A: 3.7547  
H-D: 4.0242  
R-R: 4.0974  
R-A: 3.5000  
R-D: 3.0537  
A-A: 0.0000  
A-D: 0.5000  
D-D: 0.0000

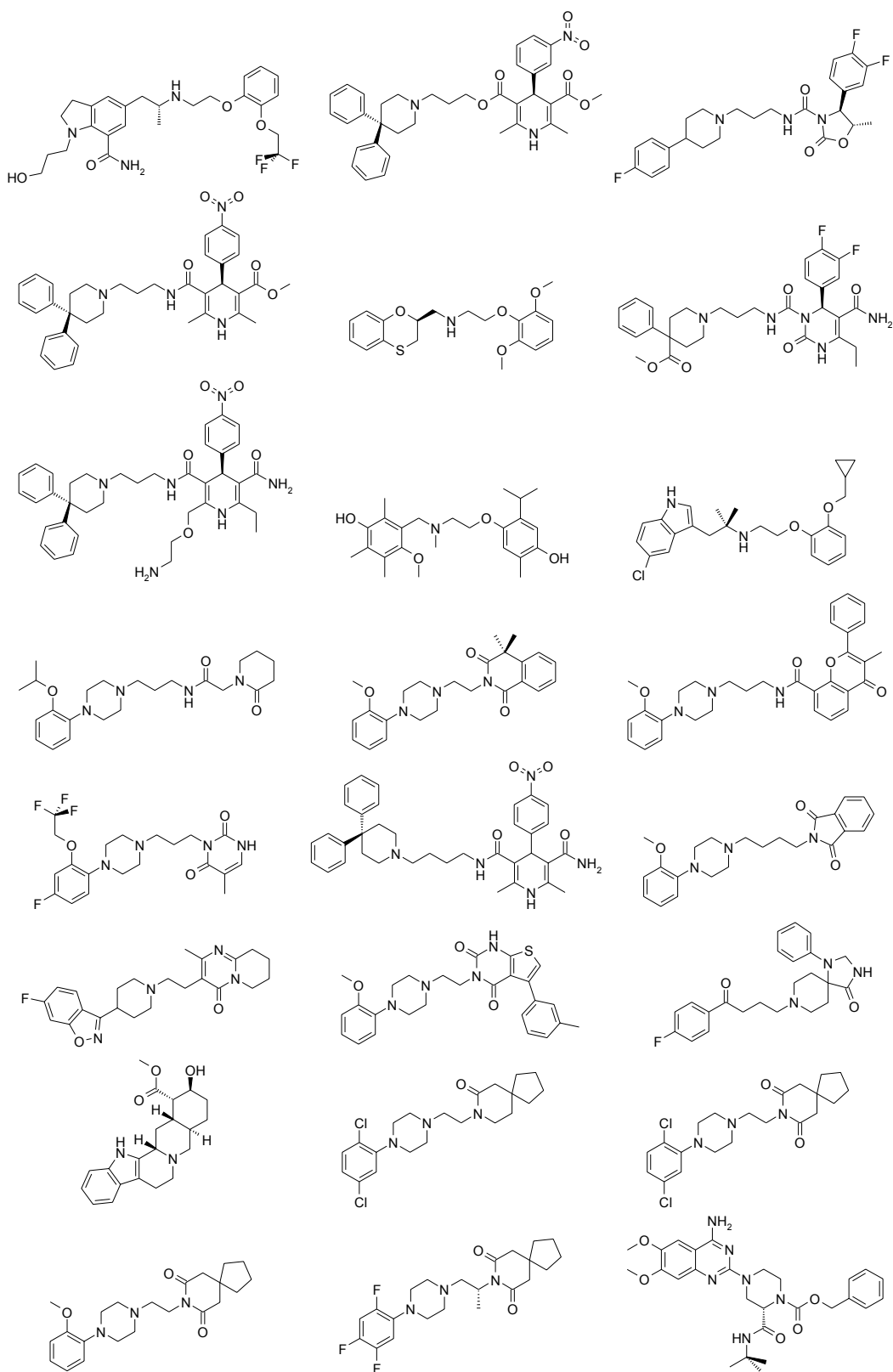
## SUPPORTING INFORMATION

### SHED Profiles for COX-1, Thrombin, and Estrogen Ligands (Ligand Numbering from Figure 2)

Ligand	H-H	H-R	H-A	H-D	R-R	R-A	R-D	A-A	A-D	D-D
1	4.6905	4.6908	4.7615	0.0000	4.2945	4.5000	0.0000	0.9449	0.0000	0.0000
2	4.5336	4.8536	5.1993	0.0000	4.9270	5.1554	0.0000	1.3747	0.0000	0.0000
3	4.4541	5.1367	5.7786	0.0000	5.1141	5.6381	0.0000	0.9449	0.0000	0.0000
4	3.9642	4.3031	4.5004	0.0000	4.3130	4.6657	0.0000	0.9449	0.0000	0.0000
5	4.3645	4.7983	5.1270	0.0000	4.8040	5.0607	0.0000	0.9449	0.0000	0.0000
6	6.6220	6.8085	4.8326	6.7986	7.0193	4.8302	6.8386	0.5000	1.5000	1.5000
7	6.1484	6.3378	4.3840	6.9027	6.5384	4.3565	7.0640	0.5000	1.0000	0.5000
8	6.3634	6.3439	4.2997	6.5151	5.8844	4.2690	6.2042	0.9449	2.6727	1.3747
9	5.3128	5.8198	3.7660	6.0616	5.6920	4.6393	5.5821	0.9449	1.7858	0.9449
10	5.3128	5.5033	3.7660	5.6569	5.3129	4.2297	5.4546	0.9449	1.7858	0.9449
11	7.1872	7.2380	7.6801	8.2332	5.0115	5.4179	6.5613	1.5000	3.0000	1.5000
12	6.9564	6.9914	7.5023	7.8198	4.9978	5.5059	6.2247	0.5000	1.5000	1.5000
13	6.8525	6.8141	6.7470	7.2081	5.0002	4.5425	5.9401	0.5000	0.9449	0.9449
14	7.1816	7.3313	7.8253	8.1830	5.1285	5.4525	6.7486	0.5000	1.5000	1.5000
15	6.8551	6.9267	7.3142	7.6918	4.9353	5.0153	6.3275	0.5000	1.5000	1.5000

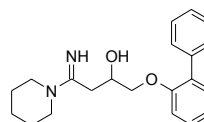
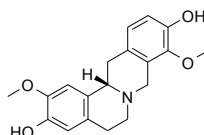
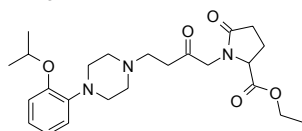
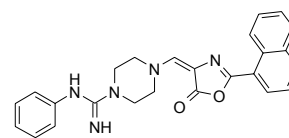
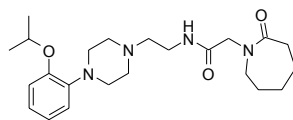
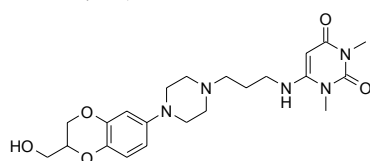
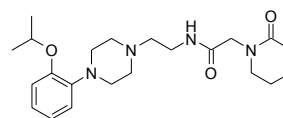
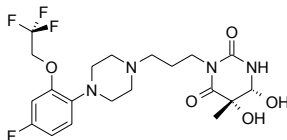
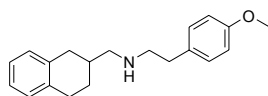
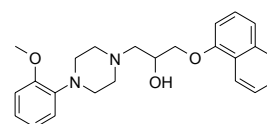
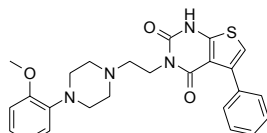
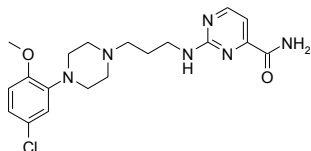
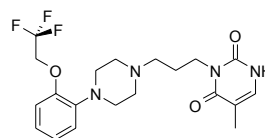
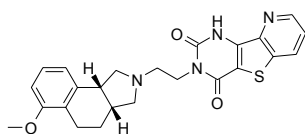
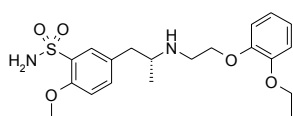
# SUPPORTING INFORMATION

## Reference Set of 24 $\alpha_{1A}$ -Adrenoceptor Antagonists



# SUPPORTING INFORMATION

## Test Set of 15 $\alpha_{1A}$ -Adrenoceptor Antagonists



## Chapter III.2 – Virtual target profiling

In this chapter, an annotated chemical library directed to nuclear receptors developed in-house [74] is used to generate ligand-based models for 25 nuclear receptors using the SHED descriptors presented in the previous chapter. The usefulness of this methodology is proven by presenting the internal validation of the models through leave one out analysis and the external validation using four external chemical libraries of targeted to proteases, kinases, ion channels, and G protein-coupled receptors.

Papers included in this chapter:

- Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to *in silico* pharmacology: nuclear receptor profiling.** *J Chem Inf Model* 2006, **46**:2725-2736.





# A ligand-based approach to *in silico* pharmacology: nuclear receptor profiling

Jordi Mestres<sup>\*,†</sup>, Lidia Martín-Couce,<sup>†</sup> Elisabet Gregori-Puigjané,<sup>†</sup> Montserrat Cases,<sup>†</sup> and  
Scott Boyer<sup>‡</sup>

*Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació  
Mèdica and Universitat Pompeu Fabra, Dr Aiguader 80, 08003 Barcelona, Catalonia, Spain and Safety  
Assessment, AstraZeneca R&D, 43183 Mölndal, Sweden.*

## Abstract

Bioactive ligands are a valuable and increasingly accessible source of information about protein targets. Based on this statement, a list of 25 nuclear receptors was described by a series of bioactive ligands extracted directly from bibliographical sources, stored properly in an annotated chemical library, and mathematically represented using the recently reported SHED molecular descriptors. Analysis of this ligand information allowed for deriving a threshold of nuclear receptor concern. In the case the similarity of one molecule to any of the molecules annotated to one particular nuclear receptor is below that threshold, the molecule receives an alert on the probability of having affinity below 10  $\mu$ M for that nuclear receptor. On this basis, a linkage map was constructed that reveals the interaction network of nuclear receptors from the perspective of their active ligands. This ligand-based approach to nuclear receptor profiling was subsequently applied to four external chemical libraries of 10000 molecules targeted to proteases, kinases, ion channels, and G protein-coupled receptors. The percentage of each library that returned an alert on at least one nuclear receptor was reasonably low and varied between 4.4% and 9.7%. In addition, ligand-based nuclear receptor profiling of a set of 2944 drugs provided an alert for 153 drugs. For some of them, namely, acitretin, telmisartan, phenyltoloxamine, tazarotene, and flumazenil, bibliographical evidence could be found indicating that those drugs may indeed have some potential off-target residual affinity for the nuclear receptor(s) annotated. Overall, the present findings suggest that ligand-based approaches to protein family profiling appear as a promising means towards the establishment of novel tools for *in silico* pharmacology.

---

\* To whom correspondence should be addressed: [jmestres@imim.es](mailto:jmestres@imim.es)

† Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra

‡ AstraZeneca

## Introduction

One of the grand challenges in chemical biology is identifying a small-molecule modulator for each individual function of all human proteins [1]. For pharmaceutical research, this has the potential to provide molecules that may then be used as chemical probes for protein validation and as initial hits for lead generation in target and drug discovery programs, respectively [2]. Vital to this aim is the ability to produce quantitative data on the response of biological systems to the presence of chemical compounds [3]. Pharmacologists have been gathering this type of data for over a century. However, it has not been until recently that the technological advances produced in combinatorial chemistry [4] and high-throughput screening [5] have made possible to collect these data in a more automatic and systematic manner, opening an avenue towards determining experimentally the pharmacological profile of compounds [6-18]. Nonetheless, in spite of the significant progress made towards improving the capacity for chemical synthesis and particularly for biological testing [19], any aspiration of being able to make and store every synthetically feasible molecule and test it on every assayable protein remains to date unreachable and thus complementary strategies for massive pharmacological profiling of large compound collections need to be explored [20,21].

One such complementary approach is the application of *in silico* methods capable of rapidly searching through large virtual chemical spaces for compounds similar to a set of bioactive reference molecules against a panel of multiple targets [22-24]. These methods are based on mathematical representations of molecules [25-29], and capitalize on initiatives aiming at the construction of annotated chemical libraries that incorporate pharmacological data into traditional repositories of chemical structures [30]. Early initiatives focused on gathering biological data for drug molecules. Of mention are the Comprehensive Medicinal Chemistry (CMC) database [31], offering currently biochemical information for over 8400 pharmaceutical compounds, and the Derwent World Drug Index (WDI) [32], containing data on activity and mechanism of action for over 58000 marketed and development drugs worldwide. More recently, those initiatives have extended their scope to capture the increasing amount of pharmacological data available from public sources. Representative examples are the MDL Drug Data Report (MDDR) [33], including information on therapeutic action and biological activity for over 132000 compounds gathered from patent literature, journals, and congresses, and the WOMBAT database [34], offering biological information for 120400 molecules reported in medicinal chemistry journals over the last 30 years. The construction of all these annotated chemical libraries contribute to establishing the knowledge base towards integrating chemical and biological data and thus for gaining a deeper understanding of the properties of molecules associated to the different protein families forming the chemogenomic space [35,36]. Ultimately, the establishment of direct biochemical connections through annotated chemical libraries may contain clues over the existence of apparently unrelated proteins having affinity for similar ligands or the presence of some privileged structures responsible for the activity of ligands in entire protein families [37,38].

The present work aims at introducing a ligand-based approach to *in silico* pharmacology by exploiting publicly available pharmacological data collected in a family-directed annotated chemical library encoded using biologically-relevant molecular descriptors. In particular, the performance of the approach to the family of nuclear receptors is presented. This family of ligand-activated transcription factors is of paramount importance for pharmaceutical research since many of its members are often considered as a double-edged sword. On one hand, they regulate a variety of biological processes, including lipid and glucose homeostasis, detoxification, cellular differentiation, embryonic development and organ physiology. Consistent with these important regulatory roles, mutations in nuclear receptors are associated with many common human diseases like cancer, diabetes, and osteoporosis and thus they are considered highly relevant therapeutic targets [39]. On the other hand, there is increasing evidence that nuclear receptors act as regulators of some cytochrome P450 enzymes, which in turn are responsible for the metabolism of molecules. Accordingly, many nuclear receptors are also regarded as potential therapeutic off-targets [40].

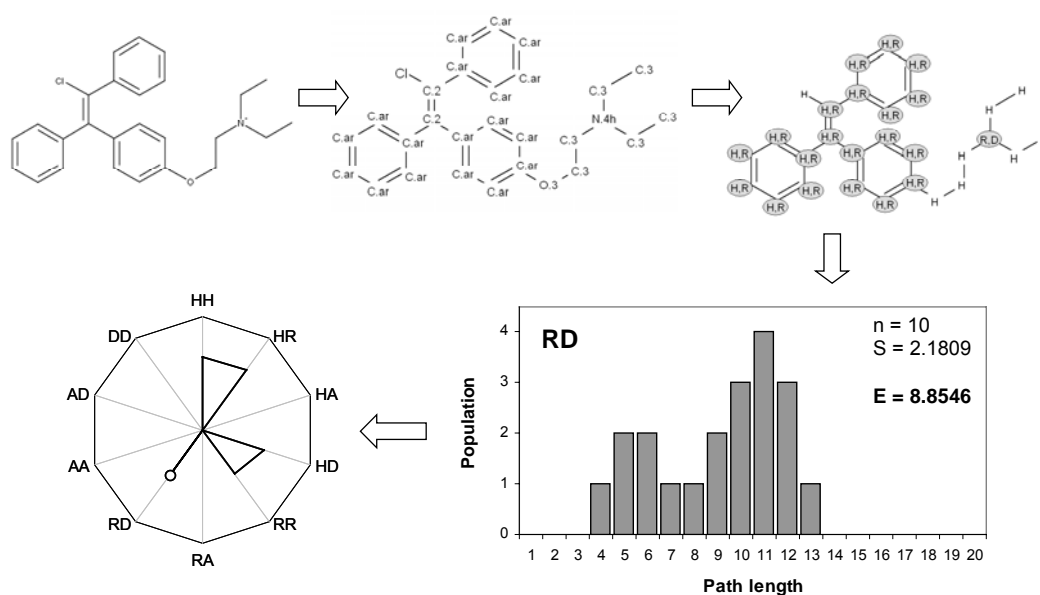
The following sections describe, first, the particular set of molecular descriptors and the annotated chemical library used in this work and, second, the construction of a ligand-based descriptor model as a rapid means for estimating *in silico* the pharmacological profile of compounds across the family of nuclear receptors.

## Methodology

The use of biologically-relevant mathematical representations of molecules and the availability of pharmacological data for a significant number of ligands are the two key elements required to perform the type of analysis presented in this work. Details on the use of Shannon entropy descriptors derived from topological feature-pair distributions and the construction of an annotated chemical library directed to the family of nuclear receptors are provided next.

**Shannon Entropy Descriptors.** A novel set of molecular descriptors called SHED (SHannon Entropy Descriptors) was recently introduced [41]. SHED are derived from distributions of atom-centered feature pairs extracted directly from the topology of molecules. The process of obtaining SHED from chemical structure is illustrated in **Figure 1** for clomifene, a selective estrogen receptor modulator. The original input structure should be in MDL's SD file format [42]. From a SD file, each atom in a molecule is first mapped to a Sybyl atom type [43]. Subsequently, each atom type is assigned to one or more of four atom-centered features, namely, hydrophobic (H), aromatic (R), acceptor (A), and donor (D). For example, an aliphatic C.3 carbon will be assigned to a hydrophobic feature (H), whereas a protonated N.4h nitrogen will be assigned to both aromatic and donor features (R,D). Then, the shortest path length between atom-centered feature pairs is derived and its occurrence at different path lengths stored to create a feature-pair distribution. A maximum path length of 20 bonds was used. Feature pairs being at distances over 20 bonds are accumulated in the last bin. As an example, the distribution of RD feature pairs within clomifene is

displayed. An equivalent distribution is derived for each of the ten possible feature pairs resulting from all pair combinations of the four features used.



**Figure 1.** Generation of a SHED profile from chemical structure (see text for details).

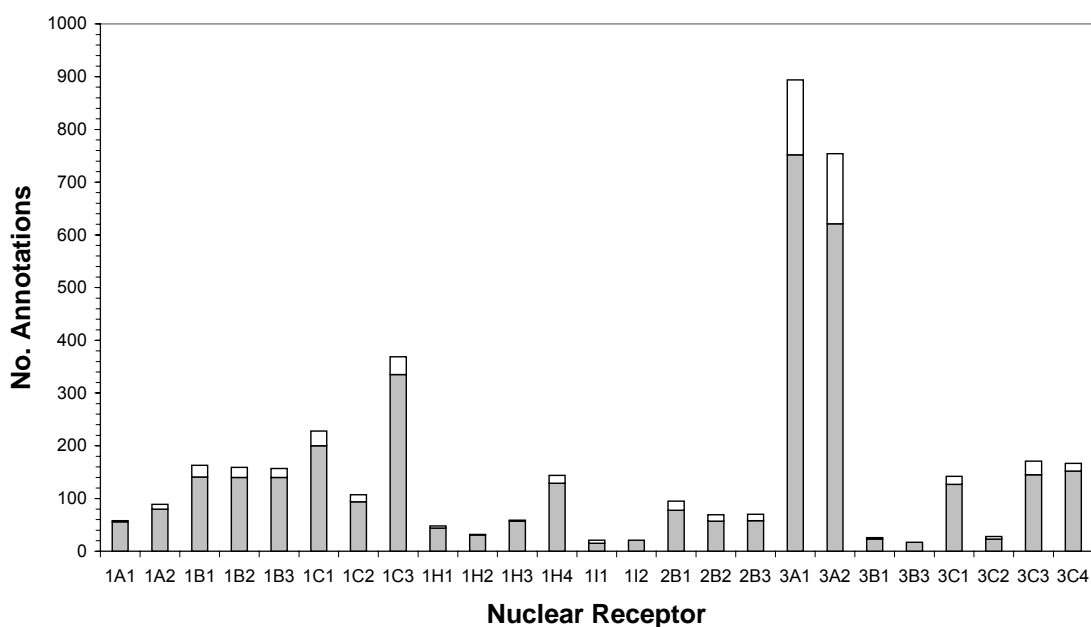
At this stage, the concept of Shannon entropy is applied to determine the variability of feature-pair distributions [41]. Within this approach, the entropy,  $S$ , of a population,  $P$ , distributed in a certain number of bins (representing in this case the different path lengths),  $N=20$ , is given by

$$S = - \sum_{i=1}^N \rho_i * \ln \rho_i \quad ; \quad \rho_i = p_i / P$$

where  $\rho_i$  and  $p_i$  are, respectively, the probability and the population at each bin  $i$  of the distribution. The values of  $S$  range between 0, reflecting the situation of all population being concentrated in a single bin, and a maximum number,  $S_{max} = \ln N$ , reflecting the situation of a uniformly distributed population among all bins. In the case of clomifene (**Figure 1**), RD pairs can be found at path lengths occupying 10 bins and the variability in their population gives rise to a distribution with an entropy value of 2.1809. In order to have a more intuitive measure that can be linearly related to the situation of full uniform occupancy, entropy values are transformed into projected entropy values,  $E = e^S$ . Correspondingly,  $E$  values provide a measure of the expected maximum uniform occupancy from the corresponding  $S$  value. Now, for any given population  $P > 0$ , the values of  $E$  can vary from 1, reflecting the situation of zero entropy in which the population is totally concentrated in a single bin, to  $N$ , reflecting the situation of maximum entropy in which the population is uniformly distributed among all bins. In the limit case of  $P = 0$ , then  $E$  will be assigned to  $E = 0$ . For the RD feature pair in clomifene (**Figure 1**) the maximum achievable  $E$  value for a population occupying uniformly 10 bins would be  $E = 10$ . The obtained  $E$  value of 8.8546 reflects a slight deviation from the situation of full uniform occupancy on 10 bins. This  $E$  value will ultimately be the Shannon entropy descriptor (SHED) for the RD feature pair. The set of SHED values

obtained for the ten possible feature pairs constitute the SHED profile of a molecule. As illustrated in **Figure 1**, SHED profiles can be represented using a wheel chart, the circle in the chart indicating the *E* value (SHED) for the RD feature pair in clomifene.

**Annotated Chemical Library.** An annotated chemical library directed to the nuclear receptor family (NRa1) was recently constructed [44]. All data incorporated in NRa1 were collected from public sources of information, mainly reviews and medicinal chemistry journals of the last 10 years. Each chemical entity within NRa1 contains a set of structural, biological, and bibliographical data. Structural data include a unique identifier for the molecule and its 2D structure representation. Biological data contain the list of nuclear receptors at which the molecule has been reported to be active, identified by their names and corresponding nuclear receptor code, together with the associated pharmacological data ( $K_i$ ,  $IC_{50}$  and/or  $EC_{50}$ ), when available. Finally, bibliographical data collect the list of references from which structural and biological data were extracted.



**Figure 2.** Distribution of all (white bars) and non-redundant (gray bars) chemical annotations present in NRa1 among 25 nuclear receptors. See text for details.

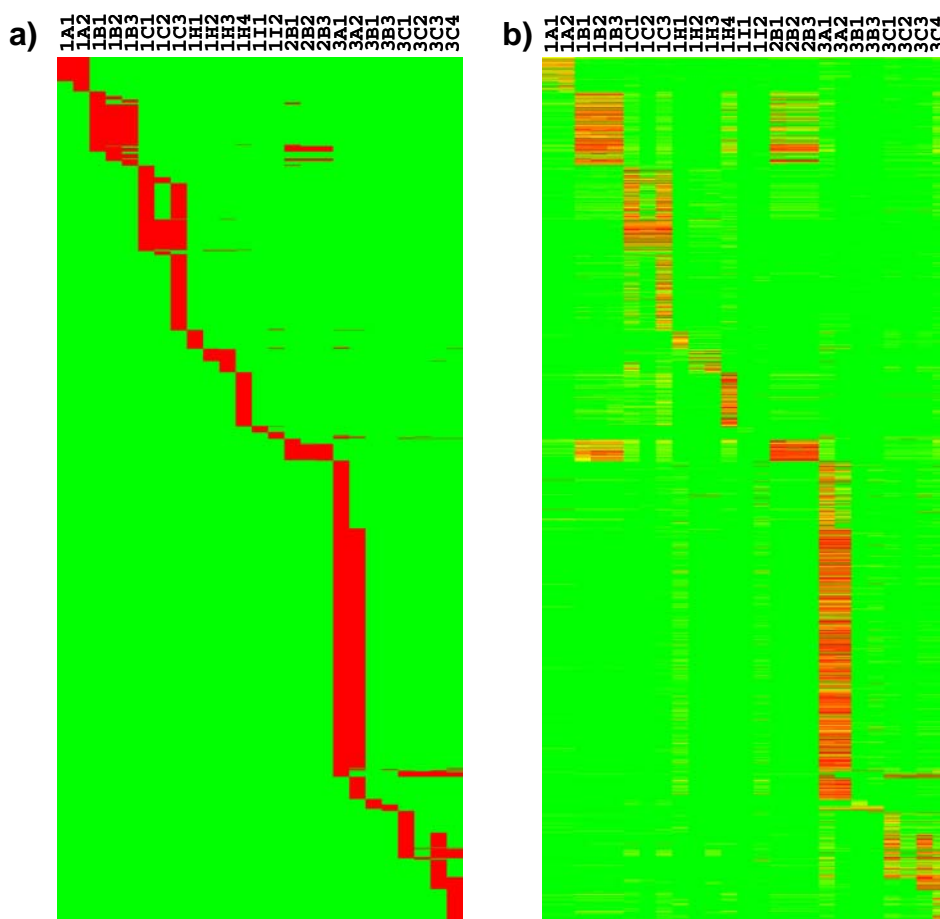
A compound in NRa1 is considered annotated to a given nuclear receptor if its associated pharmacological data ( $K_i$ ,  $IC_{50}$  and/or  $EC_{50}$ ) is under a certain cut-off. In this work, an annotation cut-off of 10  $\mu$ M was considered. Under this cut-off, NRa1 contains currently 4088 annotations to 25 nuclear receptors derived from a total of 2324 molecules, some molecules containing multiple annotations to nuclear receptors. The overall distribution of annotations among all nuclear receptors currently covered by NRa1 is provided in **Figure 2** (white bars). As can be observed, this distribution is a fair reflection of the historical therapeutic relevance of some of the members of this family. For example, the nuclear receptor containing the largest number of chemical annotations is the estrogen receptor subtype alpha ( $ER\alpha$ ; NR3A1), an important target in reproductive medicine and cancer research. Due to its high homology, many compounds binding to  $ER\alpha$  are also reported

to be active to ER $\beta$  thus justifying the large number of annotations present also for the latter. Another nuclear receptor highly populated with annotations in NRa1 is the peroxisome proliferator-activated receptor gamma (PPAR $\gamma$ ; NR1C3), widely recognized as a key regulator in multiple metabolic pathways including fatty acid and carbohydrate metabolism and thus being considered a relevant target in cardiovascular research. In contrast, estrogen-related receptor gamma (ERR $\gamma$ ; NR3B3) and vitamin D3 (VDR; NR1H1) are the only nuclear receptors collecting less than 25 annotations.

All molecules present in NRa1 are unique with respect to their structure. However, in terms of their feature distribution, some molecules may have exactly the same SHED profile, either due to pharmacophorically comparable atom mutations or to topologically equivalent structural symmetries (e.g., stereoisomerism, cis/trans isomerism, or symmetric regioisomerism). For example, having a methyl in a molecule substituted by a chlorine in another will give rise to equivalent SHED profiles since the atom types C.3 and Cl are both assigned to a hydrophobic feature. Accordingly, in order to avoid having descriptor collisions that could bias the significance of any statistics derived subsequently, all SHED-redundant molecules were identified and removed. The final distribution among nuclear receptors of the 3536 annotations remaining from the set of 2033 molecules with non-redundant SHED profiles is also depicted in **Figure 2** (gray bars).

## Results and discussion

The distribution of annotations presently contained in NRa1 is visually illustrated in **Figure 3a**. In the heatmap shown, annotations of molecules (in rows) to nuclear receptors (in columns) are represented as red cells, meaning that the interaction of a particular molecule with a specific nuclear receptor has been positively reported and experimentally quantified in the literature with a pharmacological value below 10  $\mu$ M. In contrast, green cells indicate current lack of information on the possibility of any interaction between a given molecule and a certain nuclear receptor. The extent of the green area denotes the existence of large information gaps, clearly one of the main limitations of dealing with annotated chemical libraries relying on data extracted directly from public sources of information [44]. This is due to the fact that, because of limited time and resources, molecules are usually not screened systematically through a large panel of protein targets for the sake of obtaining the maximum amount of information possible but solely to the target of interest at that point in time. But even if they were screened through multiple targets, habitually only a limited amount of data is made available, since publishing large amounts of negative data is often regarded as not informative. These important, yet often overlooked, aspects lead to a situation of data incompleteness within the interaction matrix depicted as a heatmap in **Figure 3a**.



**Figure 3.** Comparison between the heatmap representing all original annotations extracted from bibliographical sources and stored in NRacl (a) and the heatmap reflecting the minimum SHED Euclidean distances between the SHED profile of each molecule and the set of non-redundant SHED profiles annotated to each nuclear receptor (b). Color coding: (a) red is annotated and green not annotated; (b) red reflects distance values close to 0.0 and as distances increase in magnitude they turn to orange, yellow and finally green at a value of 1.2 and over.

In an attempt to address these limitations, the following section describes a means of filling the gaps in annotated chemical libraries based on deriving ligand-based descriptor models of protein targets. Subsequently, the applicability of these models for the nuclear receptor profiling of chemical libraries is finally tested on a drugs database and a series of targeted libraries designed for proteases, kinases, ion channels, and G protein-coupled receptors.

**Ligand-based Models of Proteins.** The ensemble,  $S$ , of non-redundant SHED profiles,  $s_i^I$ , representing all molecules,  $i=1, M_I$ , annotated to each particular nuclear receptor,  $I=1, N$ , constitutes a mathematical description of the nuclear receptor family from a ligand perspective:

$$S = [\{s_i^I\}_{i=1, M_I}]_{I=1, N}$$

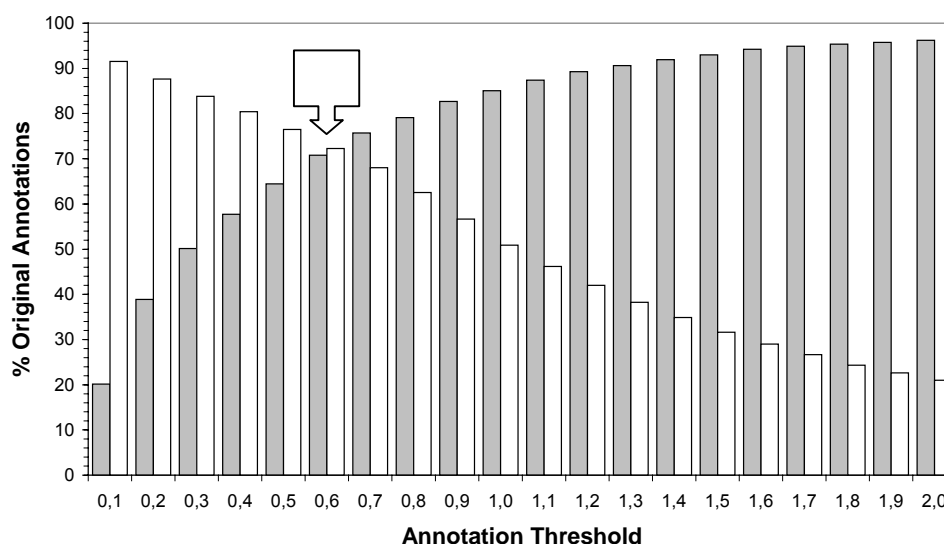
The scoring of each compound in a chemical library,  $d^I$ , with respect to a given nuclear receptor,  $I$ , is then assigned to the minimum value of all Euclidean distances calculated between

the SHED profile of the compound,  $s$ , and each one of the SHED profiles,  $s_i^l$ , describing the molecules annotated to that particular nuclear receptor:

$$d^l = \underset{i=1, M_l}{\text{MINIMUM}} \left( \sqrt{\sum_{f=1}^{10} (s_f - \{s_i^l\}_f)^2} \right),$$

where  $\{s_i^l\}_f$  is the SHED corresponding to the  $f$  feature-pair distribution of molecule  $i$  annotated to nuclear receptor  $l$ . In the context of similarity-based virtual screening, the approach of combining the scores over multiple bioactive reference molecules has been recently referred to as group fusion and proven to give significantly superior results to using data fusion strategies on single reference molecules for a wide variety of protein targets [45].

At this stage, the recall (proportion of original, bibliographically confirmed, annotations stored in NRacl) obtained at different minimum SHED Euclidean distance cutoffs was investigated, with the ultimate aim of identifying the optimum cutoff value to be used as annotation threshold for the nuclear receptor profiling of chemical libraries. Accordingly, calculation of all minimum Euclidean distances was performed between the SHED profile of each molecule in NRacl and the SHED profiles representing each nuclear receptor. In the case the latter contained the SHED profile of the molecule in NRacl being processed, that SHED profile was left out during the calculation of Euclidean distances. The results of this analysis are depicted in **Figure 4** in which gray bars are the percentage of original annotations recovered and white bars are the percentage of original annotations within all annotations assigned at each minimum SHED Euclidean distance cutoff.

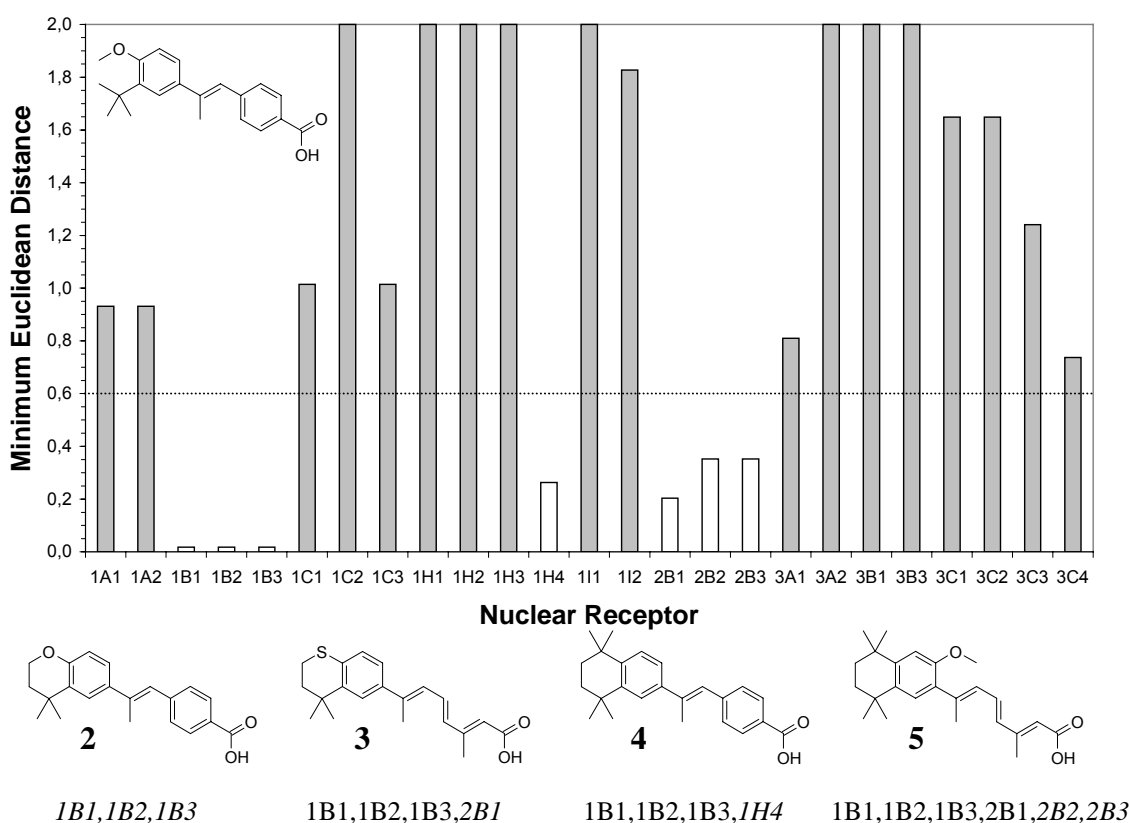


**Figure 4.** Comparison between the percentage of original annotations in NRacl recovered (gray bars) and the percentage of original annotations within all annotations assigned at each annotation threshold (white bars). A value of 0.6 is taken as a threshold of nuclear receptor concern.

As can be observed in **Figure 4**, as the annotation threshold is set to higher minimum SHED distances, a larger percentage of original annotations are being recovered (gray bars) but, in



parallel, they represent also a smaller percentage of all the annotations being assigned to molecules (white bars). An optimal annotation threshold would be one that recovers the majority of the original annotations without adding at the same time too many additional annotations. A minimum SHED Euclidean distance of 0.6 seems to show a satisfactory balance between these two criteria. Under this annotation threshold, a total of 1441 molecules, that is 70.9% of all molecules in NRacl, receive an annotation to at least one nuclear receptor. These 1441 molecules contain a total of 3462 annotations to nuclear receptors, of which 2503 were already present in NRacl and can thus be bibliographically confirmed. These 2503 nuclear receptor annotations represent 70.8% of all annotations present in NRacl (gray bar) and 72.3% of all annotations assigned at this cutoff value (white bar). Since assigning a particular annotation to a molecule effectively reflects a probability for that molecule of having an affinity value under 10  $\mu$ M for the corresponding nuclear receptor, a minimum SHED Euclidean distance of 0.6 will be considered for the remainder of this work as a threshold of nuclear receptor concern when profiling chemical libraries. This strategy follows on recent studies suggesting that similarity to molecules in the reference set is a good criteria for prediction accuracy of external test sets [46].



**Figure 5.** Nuclear receptor profile of molecule 1 based on the minimum SHED Euclidean distance between its SHED profile and the set of non-redundant SHED profiles annotated to each nuclear receptor. Molecules 2 to 5 are the molecules present in NRacl responsible for the annotations (white bars) assigned to molecule 1.

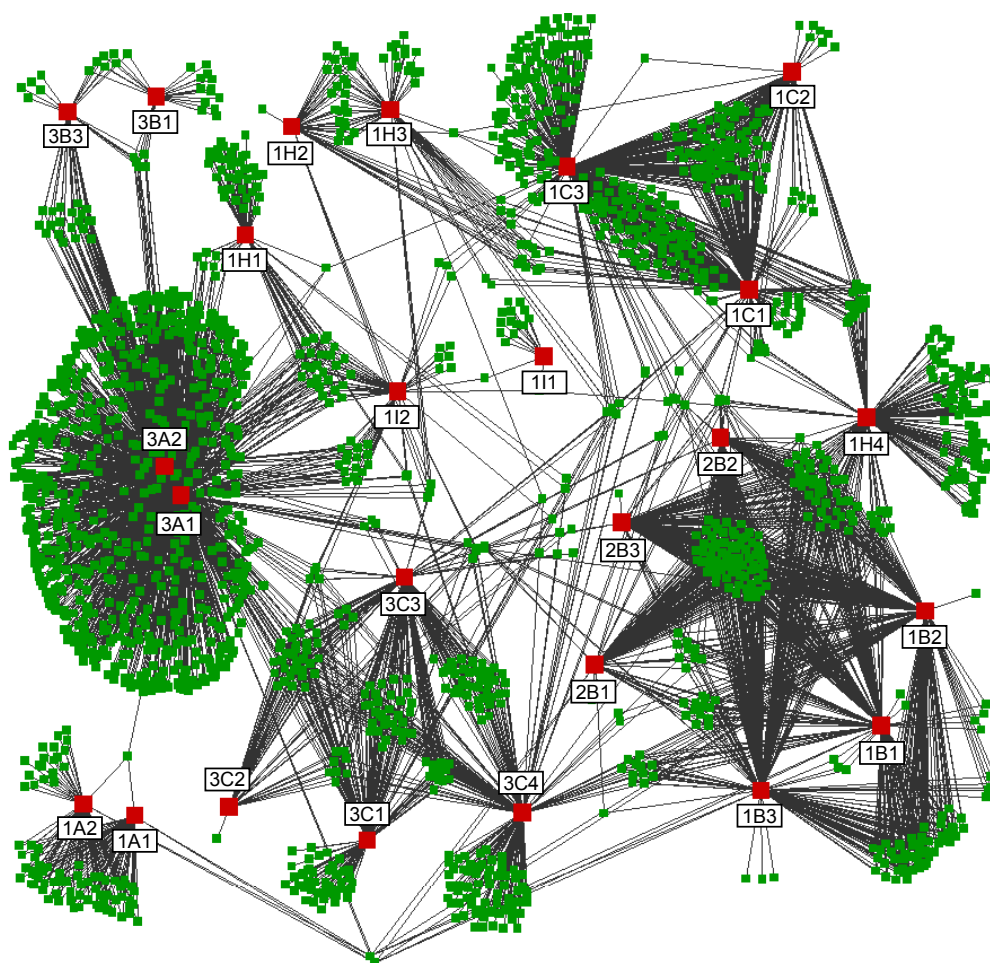
As an illustrative example, **Figure 5** shows the profile of minimum SHED Euclidean distances obtained for molecule 1 over the ligand-based model of 25 nuclear receptors. On the basis of literature data, molecule 1 is annotated in NRacl to the three retinoic acid receptors (RAR $\alpha$ : 1B1,

RAR $\beta$ : 1B2, and RAR $\gamma$ : 1B3) [47]. However, as reflected in **Figure 5**, under the 0.6 annotation threshold established above, molecule **1** can be annotated to seven nuclear receptors (white bars). The lowest values of all minimum Euclidean distances correspond indeed to the three RAR nuclear receptors, thus recovering all annotations assigned originally to molecule **1**. Besides the RAR annotations, Euclidean distances between 0.20 and 0.35 annotate molecule **1** to the three retinoic X receptors (RXR $\alpha$ : 2B1, RXR $\beta$ : 2B2, and RXR $\gamma$ : 2B3) and the farnesoid X receptor (FXR: 1H4).

One of the advantages of using a ligand-based approach for annotating molecules to nuclear receptors is the possibility to examine the ligand(s) responsible for the annotation(s) and, if necessary, go back to the original bibliographical sources stored in NRaCl. The four molecules responsible for the seven annotations to molecule **1** are also collected in **Figure 5**. The full list of annotations assigned to each of those molecules in NRaCl is also included and those annotations furnished by each respective molecule to molecule **1** are given in italics. Among all molecules currently present in NRaCl, molecule **2** is, with a SHED Euclidean distance of 0.0177, the most similar molecule to molecule **1**. It has been reported to have potencies below 10  $\mu$ M for 1B1, 1B2, and 1B3 [48], and is thus responsible for the assignment of those annotations to molecule **1**. Also, molecule **3** has been reported to have potencies below 10  $\mu$ M for 1B1, 1B2, 1B3, and 2B1 [48], and with a SHED Euclidean distance of 0.2033 is responsible for the 2B1 annotation to molecule **1**. In addition, molecule **4** (also known as TTNPB) was the first non-steroidal ligand to be described and is known to be a weak FXR agonist [49]. At a SHED Euclidean distance of 0.2630, molecule **4** is the ultimate responsible for the 1H4 annotation to molecule **1**. Finally, molecule **5** has been reported to have binding affinities below 10  $\mu$ M for all RARs and RXRs [50]. With a SHED Euclidean distance of 0.3527, molecule **5** is responsible for the additional 2B2 and 2B3 annotations to molecule **1**. Besides the reported potencies for the three RARs, we are not able to confirm that molecule **1** has indeed affinities below 10  $\mu$ M for FXR and the three RXRs, but the evident structural similarities with the four molecules responsible for all annotations are an indication that those additional four annotations are not an unreasonable alert.

The process described above for molecule **1** was applied to each one of the 2033 molecules with non-redundant SHED profiles present in NRaCl. The results are given in **Figure 3b**, in which the order of the molecules is exactly the same as the one obtained from the original annotations shown in **Figure 3a**. In contrast to the binary heatmap illustrated in **Figure 3a**, in which red was annotated and green was not annotated, **Figure 3b** presents a color gradation between red and green reflecting the value of the minimum SHED Euclidean distance between the SHED profile of each molecule and the set of non-redundant SHED profiles annotated to each nuclear receptor. Taking the annotation threshold of 0.6 as the center of the color scale, distance values close to 0.0 are represented in red, those close to 0.6 are seen as light orange, and as distances increase in magnitude they turn to yellow and finally green at a value of 1.2 and over. There are two main aspects worth mentioning when comparing the heatmaps of **Figures 3a** and **3b**. On one hand, it is remarkable the fact that the essential pattern observed when plotting the original annotations (**Figure 3a**) is preserved when molecules are processed through the ligand-based descriptor model

of nuclear receptors (**Figure 3b**). This result reveals that the remaining molecules in NRacl are to a great extent representative of the molecule being processed after leaving that molecule out, something that can only be achieved if the annotated chemical space has been sufficiently saturated with as many known bioactive molecules as possible. On the other hand, despite the clear discrimination between nuclear receptor groups, some correlation patterns between them emerge. The most apparent example is the clear correlation observed in **Figure 3b** between RARs (1B1, 1B2, and 1B3) and RXRs (2B1, 2B2, and 2B3), a result that provides an indication of the potential of this approach for understanding side effects through the identification of off-target affinities.



**Figure 6.** Nuclear receptor network derived using the matrix of minimum SHED Euclidean distances (see **Figure 3b**). The set of 2033 molecules from NRacl are given as small green squares and the 25 nuclear receptors as large red squares. A value of 0.6 for the minimum SHED Euclidean distance was used to establish direct links between molecules and nuclear receptors. This linkage map was constructed with Cytoscape [51].

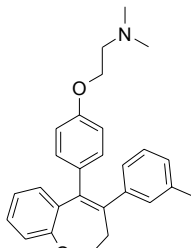
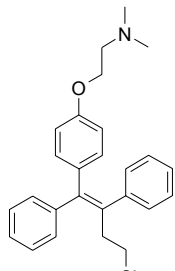
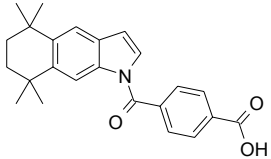
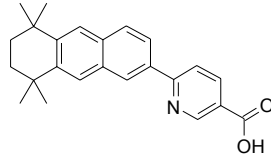
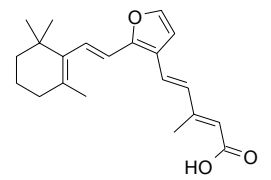
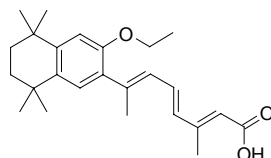
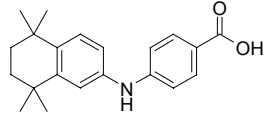
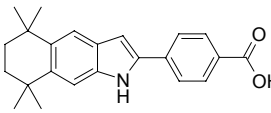
To investigate further any potential links between nuclear receptors from the perspective of their active ligands, we used some graph-based tools to construct an interaction network [51]. A similar method was reported recently to visualize nuclear hormone receptor networks relevant to drug metabolism [52]. **Figure 6** contains the nuclear receptor network obtained on the basis of the matrix of minimum SHED Euclidean distances presented in **Figure 3b**, using the threshold of nuclear receptor concern derived above as the linkage criteria for assigning direct connections

between a molecule and a nuclear receptor. The first observation that can be made from inspection of **Figure 6** is that the essential phylogenetic relationships among nuclear receptors are essentially preserved. The strong interconnections between RARs and RXRs noticed above are also clearly retrieved. But perhaps a more interesting outcome is the existence of several molecules connecting the estrogen receptors (3A1 and 3A2) with the pregnane X receptor (PXR: 1I2) and the ecdysone receptor (1H1). The link observed between PXR and estrogen may have implications related to the metabolism and toxicity of estrogenic compounds [53,54], whereas the link identified between ecdysone and estrogen may suggest that a library designed around estrogenic compounds may be a good starting point for the identification of novel chemical modulators of the ecdysone receptor [55]. Finally, of mention are the links observed between FXR (1H4) and the peroxisome proliferator-activated receptor gamma (PPAR $\gamma$ : 1C3), as well as with members of the RAR and RXR groups, indicating that FXR may be an off-target to take into consideration when developing compounds for PPAR $\gamma$ , RARs, and RXRs [56,57].

**Nuclear Receptor Profiling.** The ligand-based model of nuclear receptors described in the previous section can then be used to profile large chemical libraries on this family of transcription factors of key importance for pharmaceutical research. As a first validation exercise, a small database of 82 diverse molecules, not included in NRacl, was compiled as an external test set from bibliographic sources reporting experimental evidence of activity on nuclear receptors. Profiling of this database on the ligand-based nuclear receptor model provided annotations (that is, a minimum SHED Euclidean distance of 0.6 for at least one nuclear receptor) for 47 molecules (57% of the database). Out of these 47 molecules, 32 molecules had annotations to at least one member of the nuclear receptor group at which the molecules were known to be active. Accordingly, the method provided a correct identification of the target nuclear receptor group for 68.1% of the molecules annotated. This result is in good agreement with the expected performance of the method (70.9%; *vide supra*).

To provide an idea of the level of structural hopping that can be achieved using the current approach, a selection from this 47 molecules, covering the whole range of distance values under the threshold of 0.6, is presented in **Figure 7**. Molecule **6** is a benzoxepin analogue of tamoxifen [58]. Not surprisingly, the closest molecule found when profiled against the ligand-based nuclear receptor model is molecule **7**, toremifene, a chlorine analogue of tamoxifen, annotated in NRacl to have affinity for the estrogen receptor subtype  $\alpha$  (ER $\alpha$ : 3A1) and thus responsible for the 3A1 annotation to molecule **6** [59]. Molecule **6** had also minimum SHED Euclidean distances below 0.6 for other molecules in NRacl that provided additional annotations to the estrogen receptor subtype  $\beta$  (ER $\beta$ : 3A2) and PXR (1I2). Molecule **8**, Am93, is structurally related to Am80, a known potent synthetic retinoid reported to be active to RARs [60]. Nuclear receptor profiling of this compound identifies molecule **9**, with annotations to all RARs [47], as the molecule in NRacl being closest in SHED profile to molecule **8**. Molecule **10** is a conformationally restrained analogue of retinoic acid, with reported nanomolar affinity for all RXRs and hundred-fold selectivity over RARs [61]. Molecule **11** is the closest reference compound found to it in NRacl. Most interestingly, despite being

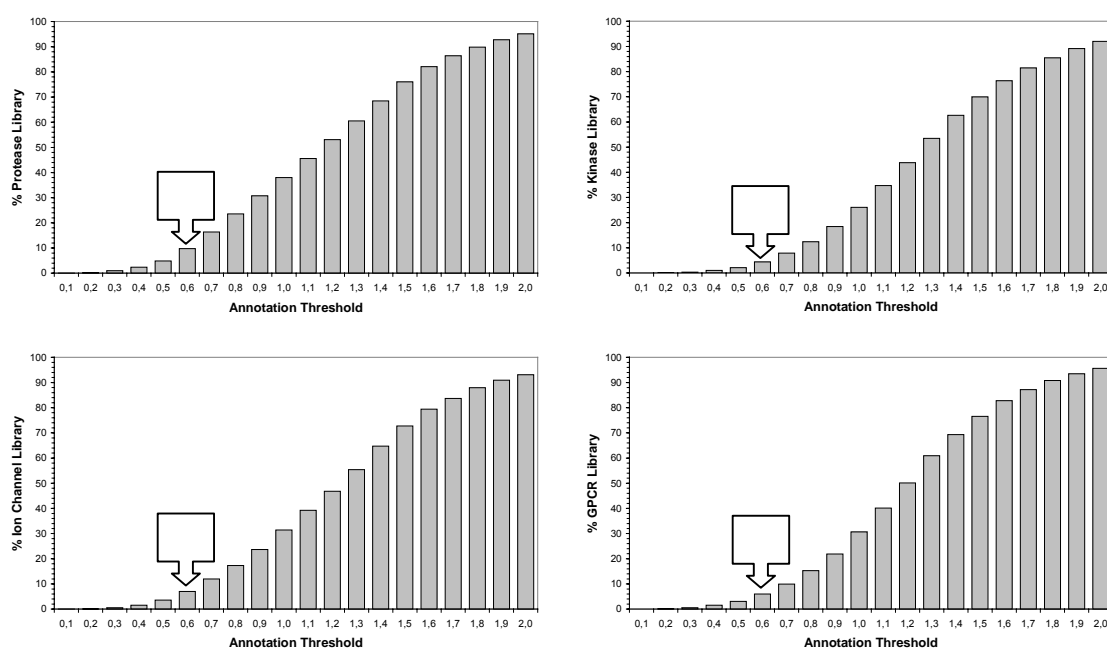
annotated to all RARs and RXRs because of the 10  $\mu$ M cut-off used for the construction of NRaCl, molecule **11** is also being reported to have nanomolar affinities for all RXRs but only micromolar affinities for all RARs, in good agreement with the known profile of molecule **10** [50]. Finally, molecule **12**, DA010, is a known ligand for both RARs and RXRs [62]. In this case, molecule **13** is the only molecule present in NRaCl with a SHED profile under a distance of 0.6 to that of molecule **12** and is thus responsible for the annotation of molecule **12** to all RARs [47]. This is an example where the approach would have missed annotating molecule **12** to all RXRs.

Test compound	Distance	Reference compound
 <p><b>6: 3A1,3A2,1I2</b></p>	0.0597	 <p><b>7: 3A1</b></p>
 <p><b>8: 1B1,1B2,1B3,2B1</b></p>	0.1553	 <p><b>9: 1B1,1B2,1B3</b></p>
 <p><b>10: 1B1,1B2,1B3,2B1,2B2,2B3,1H4</b></p>	0.2598	 <p><b>11: 1B1,1B2,1B3,2B1,2B2,2B3</b></p>
 <p><b>12: 1B1,1B2,1B3</b></p>	0.4093	 <p><b>13: 1B1,1B2,1B3</b></p>

**Figure 7.** Selection of molecules from the test set annotated to nuclear receptors (test compound). The molecule in NRaCl showing the minimum SHED Euclidean distance to each test compound (reference compound) is also included. Besides the annotation(s) provided specifically by the reference compound shown, all annotations assigned to test compounds from reference compounds are also given in italics (see **Figure 5**).

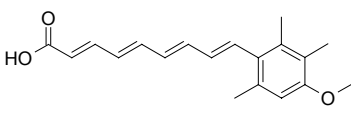
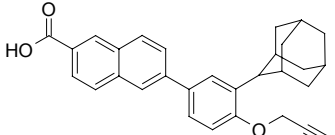
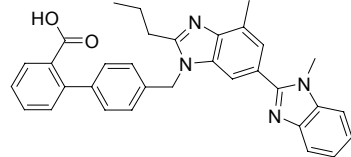
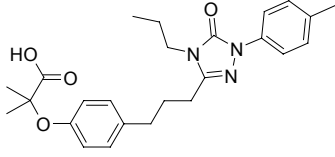
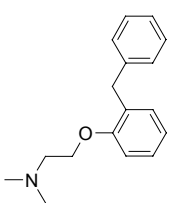
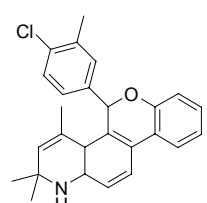
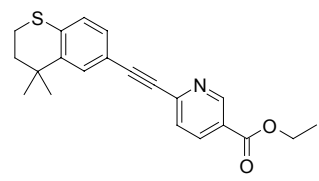
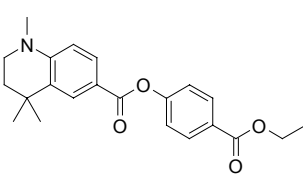
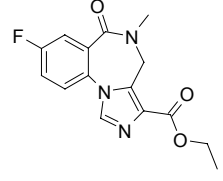
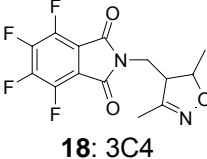
Following current trends in chemogenomic strategies for family-directed drug discovery, a direct application of this approach is in the selection of compounds targeted to nuclear receptors from chemical provider databases and the nuclear receptor profiling of targeted libraries designed specifically for other protein families of high therapeutic relevance. Accordingly, as a second validation exercise, we took four commercially available targeted libraries designed to provide hits for the families of proteases, kinases, ion channels, and G protein-coupled receptors (GPCR)

containing 19649, 515265, 31579, and 110507 molecules, respectively [63]. For the sake of consistency, a set of 10000 molecules was randomly selected from each of them and processed through the ligand-based nuclear receptor model. The results are compiled in **Figure 8**, represented in a form comparable to the gray bars in **Figure 4**. As can be observed, under the annotation threshold of 0.6 established above, only 9.7% of the 10000 molecules contained in the protease library, 4.4% of an equivalent number of molecules present in the kinase library, 7.0% of the ion channel library, and 6.0% of the GPCR library would be alerted on the potential of having affinity to some nuclear receptor. Since molecules designed for these libraries should not be expected to have activity on nuclear receptors, at least to a large extent, these results confirm the validity of the ligand-based approach to nuclear receptor profiling used in this work.



**Figure 8.** Percentage of each targeted library (protease, kinase, ion channel, and GPCR) annotated to nuclear receptors at varying threshold values. Numbers indicate the percentage of each library under the threshold of nuclear receptor concern.

One of the major concerns during the process of drug discovery is the possibility that the compounds being optimized could have residual affinities for some off-targets responsible for highly undesirable side effects. Therefore, the third validation exercise consisted of profiling a dataset of 2944 drugs, none of them present in NRacl, through the ligand-based nuclear receptor model. Interestingly, only 153 drugs, that is 5.2% of the total number contained in the dataset, were identified as having at least a minimum SHED Euclidean distance below 0.6 to a nuclear receptor. Of those, 32 drugs contained a steroidal scaffold and were found to be similar to some of the steroidal hormone receptor ligands present in NRacl. An additional set of 19 drugs was found to have the same scaffold to some molecule present in NRacl, and could thus be considered a close analogue. An illustrative selection of the 102 remaining drugs that were annotated to nuclear receptors is compiled in **Figure 9**, covering the whole range of distances within the 0.6 annotation threshold.

Drug	Distance	Reference compound
 Acitretin	0.1635	 <b>14: 1B1,1B2,1B3</b>
 Telmisartan	0.3902	 <b>15: 1C1,1C2,1C3</b>
 Phenyltoloxamine	0.4234	 <b>16: 3A1,3C1.3C2.3C3.3C4</b>
 Tazarotene	0.4391	 <b>17: 1B1,1B2,2B1,2B2,2B3</b>
 Flumazenil	0.5752	 <b>18: 3C4</b>

**Figure 9.** Selection of drugs annotated to nuclear receptors yet not designed specifically for them. The molecule in NRacl showing the minimum SHED Euclidean distance to each drug (reference compound) is also included.

The first drug collected in **Figure 9** is acitretin, an oral retinoid indicated for the treatment of psoriasis, although its mechanism of action has not been fully elucidated. When profiled against the ligand-based nuclear receptor model, this drug was found at close distance to molecule **14**, a molecule that has been reported to have potencies below 10  $\mu$ M to all three RARs [47]. This result would provide an alert for acitretin having potential affinity for RARs. We have not been able to find direct experimental data confirming this fact, but there is bibliographical evidence suggesting a link between psoriasis and an alteration in the cellular retinoid pathways at which synthetic retinoids, such as acitretin, may interact [64]. The next drug listed is telmisartan, originally designed as an angiotensin II receptor blocker for treating the metabolic syndrome. Nuclear receptor profiling of this drug identified molecule **15** as the closest to its SHED profile, a molecule that has been

reported to have affinity below 10  $\mu\text{M}$  to all peroxisome proliferator-activated receptor subtypes (PPAR $\alpha$ : 1C1, PPAR $\beta$ : 1C2, PPAR $\gamma$ : 1C3) [65]. Recent preclinical studies indicate that telmisartan acts as a PPAR $\alpha$  modulator when tested at concentrations that might be achievable with oral doses recommended for the treatment of hypertension [66]. Having processed telmisartan through some type of nuclear receptor alert system, such as the one presented in this work, this outcome could have been anticipated at an earlier stage. The following drug is phenyltoloxamine, a H1 histamine antagonist used in pain treatment. Molecule **16** was identified as the closest molecule in NRaCl to this drug. This molecule has been reported to show antagonistic activity for ER $\alpha$  (3A1), as well as to the glucocorticoid (GR: 3C1), mineralocorticoid (MR: 3C2), progesterone (PR: 3C3) and androgen (AR: 3C4) receptors [67]. Since the SHED Euclidean distance between phenyltoloxamine and molecule **16** is below 0.6, phenyltoloxamine is alerted for having potential affinity on these five nuclear receptors. We have not been able to find experimental data confirming this result in its totality. But bibliographical evidence has been found on the fact that the antiestrogen binding site is an histamine or histamine-like receptor and that, in particular, phenyltoloxamine has indeed some antagonistic affinity for the estrogen receptor [68]. Next in **Figure 7** is tazarotene, a topically applied retinoid for the treatment of psoriasis. Nuclear receptor profiling of this drug identifies molecule **17**, with annotations to RARs and RXRs, as the molecule in NRaCl being closest in SHED profile to tazarotene [69]. Indeed, indications can be found in the literature that tazarotene does bind to RARs, although selectively with respect to RXRs [70]. Finally, flumazenil is a GABA $_A$ -benzodiazepine receptor antagonist. A SHED Euclidean distance right below 0.6 is obtained with respect to molecule **18**, a known androgen receptor antagonist [71]. Despite no direct evidence of this result could be found, recent indirect reports indicate that flutamide, an androgen receptor antagonist, produced an anticonvulsant effect in common seizure models through a possible interaction with benzodiazepine receptors, thus being indicative of potential cross-pharmacologies between these two receptors [72].

## Conclusions

On the basis of pharmacological data extracted directly from bibliographical sources, a ligand-based descriptor model for the family of nuclear receptors was constructed, offering the possibility to perform *in silico* the profiling of large chemical libraries on 25 nuclear receptors in a fast and efficient manner. It was shown that, provided the annotated chemical space for the protein family of interest is sufficiently well saturated, the model attains a decent degree of both internal consistency and external predictability. The model served also to construct an interaction network from which potential cross-pharmacologies between nuclear receptors emerged. In addition, the approach was proven to be sensible enough to achieve significant discriminative power when applied to external chemical libraries designed for *a priori* unrelated protein families such as proteases, kinases, ion channels, and GPCRs, opening an avenue for its use in the selection and design of targeted libraries. Further external validation of the model was finally provided by the identification of a selected list of drugs for which bibliographical evidence exists, though indirect for some, indicating



that those drugs may indeed have some potential off-target residual affinity for the nuclear receptor(s) annotated.

A vast amount of pharmacological data on many protein targets is becoming available for increasingly large quantities of molecules. The systematic collection of these data in annotated chemical libraries allows for describing proteins from the perspective of ligands, descriptions that develop into a more complete picture as data on new ligands are obtained and the active chemical space stored becomes more saturated. Interestingly, a recent study comparing the performance of some ligand-based and structure-based methods for virtual screening concluded that information about a target derived from knowledge on bioactive ligands can be as valuable as knowledge of the target structures for identifying novel scaffolds by computational means [73]. Functional coverage of the proteome by structures is progressing rapidly but many areas are still devoid of any structural information [74]. In the wait for having at least one representative structure for each target protein, ligand-based representations of proteins may offer a means to move from the traditional virtual chemical screening to the necessary virtual pharmacological profiling.

The ligand-based approach to nuclear receptor profiling presented in this work can be readily extended to other protein families of high therapeutic relevance, such as GPCRs, for which only limited structural information is available but pharmacological data on a significant number of molecules are known. By gathering, properly storing, and maximally exploiting all pharmacological knowledge on ligands available to date, *in silico* pharmacology on a genomic scale may soon become a reality.

**Acknowledgments.** This research was funded by the European Commission (InfoBioMed network of excellence IST-507585) and the Spanish Ministerio de Educación y Ciencia (project reference BIO2005-04171).

## References

1. Schreiber SL: **Small molecules: the missing link in the central dogma.** *Nat Chem Biol* 2005, **1**:64-66.
2. Bredel M, Jacoby E: **Chemogenomics: an emerging strategy for rapid target and drug discovery.** *Nat Rev Genet* 2004, **5**:262-275.
3. Walters WP, Namchuk M: **Designing screens: how to make your hits a hit.** *Nat Rev Drug Discov* 2003, **2**:259-266.
4. Ramström O, Lehn J-M: **Drug discovery by dynamic combinatorial libraries.** *Nat Rev Drug Discov* 2002, **1**:26-36.
5. Sundberg SA: **High-throughput and ultra-high-throughput screening: solution- and cell-based approaches.** *Curr Opin Biotech* 2000, **11**:47-53.
6. Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Golstein A, Bukar R, Bauer KE, Dilley H, Rocke DM: **Predicting ligand binding to proteins by affinity fingerprinting.** *Chem Biol* 1995, **2**:107-118.
7. Weinstein JN, Myers TG, O'Connor TM, Friend SH, Fornace Jr. AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, et al.: **An information-intensive approach to the molecular pharmacology of cancer.** *Science* 1997, **275**:343-349.
8. McBeath G, Koehler AN, Schreiber SL: **Printing small molecules as microarrays and detecting protein-ligand interactions en masse.** *J Am Chem Soc* 1999, **121**:7967-7068.
9. Poulain R, Horvath D, Bonnet B, Eckhoff C, Chapelain B, Bodinier MC, Déprez B: **From hit to lead. Analyzing structure-profile relationships.** *J Med Chem* 2001, **44**:3391-3401.
10. Rabow AA, Shoemaker RH, Sausville EA, Covell DG: **Mining the National Cancer Institute's tumor screening database: identification of compounds with similar cellular activities.** *J Med Chem* 2002, **45**:818-840.
11. Greenbaum DC, Arnold WD, Lu F, Hayrapetian L, Baruch A, Krumrine J, Toba S, Chehade K, Brömme D, Kuntz ID, et al.: **Small molecule affinity fingerprinting: a tool for enzyme family subclassification, target identification, and inhibitor design.** *Chem Biol* 2002, **9**:1085-1094.
12. Kunkel EJ, Dea M, Ebens A, Hytopoulos E, Melrose J, Nguyen D, Ota KS, Plavec I, Wang Y, Watson SR, et al.: **An integrative biology approach for analysis of drug action in models of human vascular inflammation.** *FASEB J* 2004, **18**:1279-1281.
13. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan ML, Arkin AP, Davis RW: **Chemogenomic profiling: identifying the functional interactions of small molecules in yeast.** *Proc Natl Acad Sci U S A* 2004, **101**:793-798.
14. Szakács G, Annereau J-P, Lababidi S, Shankavaram U, Arciello A, Bussey KJ, Reinhold W, Guo Y, Kruh GD, Reimers M, et al.: **Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells.** *Cancer Cell* 2004, **6**:129-137.
15. Roth BL, Sheffler DJ, Kroeze WK: **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.** *Nat Rev Drug Discov* 2004, **3**:353-359.
16. Fabian MA, Biggs III WH, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, et al.: **A small molecule - kinase interaction map for clinical kinase inhibitors.** *Nature Biotech* 2005, **23**:329-336.

17. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: **Analysis of drug induced effect patterns to link structure and side effects of medicines.** *Nature Chem Biol* 2005, **1**.
18. Ramanathan A, Wang C, Schreiber SL: **Profiling of a cell-line model of tumorigenesis by using metabolic measurements.** *Proc Natl Acad Sci USA* 2005, **102**:5992-5997.
19. Haiching M, Horiuchi KY: **Chemical microarray: a new tool for drug screening and discovery.** *Drug Discov Today* 2006, **11**:661-668.
20. Bajorath J: **Integration of virtual and high-throughput screening.** 2002, **1**:882-894.
21. Hopkins AL, Mason JS, Overington JP: **Can we rationally design promiscuous drugs?** *Curr Opin Struct Biol* 2006, **16**:127-136.
22. Krejsa CM, Horvath D, Rogalski SL, Penzotti JE, Mao B, Barbosa F, Migeon JC: **Predicting ADME Properties and Side Effects: the BioPrint Approach.** *Current Opinion Drug Discovery Development* 2003, **6**:470-480.
23. Root DE, Flaherty SP, Kelley BP, Stockwell BR: **Biological mechanism profiling using an annotated compound library.** *Chem Biol* 2003, **10**:881-892.
24. Strausberg RL, Schreiber SL: **From knowing to controlling: a path from genomics to drugs using small molecule probes.** *Science* 2003, **300**:294-295.
25. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E: **Similarity metrics for ligands reflecting the similarity of the target proteins.** *J Chem Inf Model* 2003, **43**:391-405.
26. Horvath D, Jeandenans C: **Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces - A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles.** *J Chem Inf Comp Sci* 2003, **43**:680-690.
27. Poroikov VV, Filimonov DA, Borodina YV, Lagunin AA, Kos A: **Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds.** *J. Chem. Inf. Model.* 2000, **40**:1349-1355.
28. Cleves AE, Jain AN: **Robust Ligand-Based Modeling of the Biological Targets of Known Drugs.** *J. Med. Chem.* 2006, **49**:2921-2938.
29. Nidhi, Glick M, Davies JW, Jenkins JL: **Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases.** *J Chem Inf Model* 2006, **46**:1124-1133.
30. Savchuk NP, Balakin KV, Tkachenko SE: **Exploring the chemogenomic knowledge space with annotated chemical libraries.** *Curr Opin Chem Biol* 2004, **8**:412-417.
31. **MDL comprehensive medicinal chemistry (CMC-3D) database. MDL information systems, Inc. CMC is an updated electronic version of *Comprehensive Medicinal Chemistry*.** Pergamon Press 1990.
32. **World Drug Index.** Derwent Information Ltd.
33. **MDL Drug Data Report.** Edited by. San Leandro, CA: MDL Information Systems, Inc.
34. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity.** In *Chemoinformatics in Drug Discovery*. Edited by Wiley-VCH; 2004:223-239.
35. Vieth M, Sutherland JJ, Robertson DH, Campbell RM: **Kinomics: characterizing the therapeutically validated kinase space.** *Drug Discov Today* 2005, **10**:839-846.

36. Vieth M, Sutherland JJ: **Dependence of molecular properties on proteomic family for marketed oral drugs.** *J Med Chem* 2006, **49**:3451-3453.
37. Koch MA, Waldmann H: **Protein structure similarity clustering and natural product structure as guiding principles in drug discovery.** *Drug Discov Today* 2005, **10**:471-483.
38. Klabunde T, Hessler G: **Drug design strategies for targeting G-protein-coupled receptors.** *ChemBioChem* 2002, **3**:928-944.
39. Gronemeyer H, Gustafsson J-A, Laudet V: **Principles for modulation of the nuclear receptor superfamily.** *Nat Rev Drug Discov* 2004, **3**:950-964.
40. Francis GA, Fayard E, Picard F, Auwerx J: **Nuclear receptors and the control of metabolism.** *Annual Review of Physiology* 2003, **65**:261-311.
41. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.
42. Elsevier MDL, San Leandro, CA: <http://www.mdl.com/>. (accessed May 2006).
43. Tripos Inc, St. Louis, MO: [http://www.tripos.com/mol2/atom\\_types.html](http://www.tripos.com/mol2/atom_types.html). (accessed May 2006).
44. Cases M, Garcia-Serna R, Hettne K, Weeber M, Lei JV, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family.** *Curr Top Med Chem* 2005, **5**:763-772.
45. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.** *J Chem Inf Model* 2004, **44**:1177-1185.
46. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK: **Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.** *J Chem Inf Comp Sci* 2004, **44**:1912-1928.
47. Douguet D, Thoreau E, Grassy G: **Quantitative structure-activity relationship studies of RAR  $\alpha$ ,  $\beta$ , and  $\gamma$  retinoid agonists.** *Quant. Struct.-Act. Relat.* 1999, **18**:107-123.
48. Benbrook DM, Subramanian S, Gale JB, Liu S, Brown CW, Boehm MF, Berlin KD: **Synthesis and characterization of heteroarotinoids demonstrate structure specificity relationships.** *J Med Chem* 1998, **41**:3753-3757.
49. Maloney PR, Parks DJ, Haffner CD, Fivush AM, Chandra G, Plunket KD, Creech KL, Moore LB, Wilson JG, Lewis MC, et al.: **Identification of a Chemical Tool for the Orphan Nuclear Receptor FXR.** *43* 2000.
50. Canan Koch SS, Dardashti LJ, Hebert JJ, White SK, Croston GE, Flatten KS, Heyman RA, Nadzan AM: **Identification of the first retinoid X receptor homodimer antagonist.** *J Med Chem* 1996, **39**:3229-3234.
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003, **13**:2498-2504.
52. Ekins S, Kirillov E, Rakhmatulin EA, Nikolskaya T: **A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism.** *Drug Metab Dispos* 2005, **33**:474-481.

53. Kretschmer XC, Baldwin WS: **CAR and PXR: xenosensors of endocrine disrupters?** *Chem Biol Interact* 2005, **155**:111-128.
54. Ding X, Lichti K, Staudinger JL: **The micoestrogen zearalenone induces CYP3A through activation of the pregnane X receptor.** *Toxicol Sci* 2006, **91**:448-455.
55. Oberdorster E, Clay MA, Cottam DM, Wilmot FA, McLachlan JA, Milner MJ: **Common phytochemicals are ecdysteroid agonists and antagonists: a possible evolutionary link between vertebrate and invertebrate steroid hormones.** *J Steroid Biochem Mol Biol* 2001, **77**:229-238.
56. Fiorucci S, Rizzo G, Antonelli E, Renga B, Mencarelli A, Riccardi L, Morelli A, Pruzanski M, Pellicciari R: **Cross-talk between farnesoid-X-receptor (FXR) and peroxisome proliferator-activated receptor gamma contributes to the antifibrotic activity of FXR ligands in rodent models of liver cirrhosis.** *J Pharmacol Exp Ther* 2005, **315**:58-68.
57. Kassam A, Miao B, Young PR, Mukherjee R: **Retinoid X receptor (RXR) agonist-induced antagonism of farnesoid X receptor (FXR) activity due to absence of coactivator recruitment and decreased DNA binding.** *J Biol Chem* 2003, **278**:10028-10032.
58. Lloyd DG, Hughes RB, Zisterer DM, Williams DC, Fattorusso C, Catalanotti B, Campiani G, Meegan MJ: **Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor.** *J Med Chem* 2004, **47**:5612-5615.
59. Fang H, Tong W, Shi LM, Blair R, Perkins R, Branham W, Hass BS, Xie Q, Dial SL, Moland CL, et al.: **Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens.** *Chem Res Toxicol* 2001, **14**:280-294.
60. Hashimoto Y: **Structural development of synthetic retinoids and thalidomide-related molecules.** *Cancer Chemother Pharmacol* 2003, **52**:S12-23.
61. Vuligonda V, Garst ME, Chandraratna RAS: **Stereoselective synthesis and receptor activity of conformationally defined retinoid X receptor selective ligands.** *Bioorg Med Chem Lett* 1999, **9**:589-594.
62. Ohta K, Tsuji M, Kawachi E, Fukasawa H, Hashimoto Y, Shudo K, Kagechika H: **Potent retinoid synergists with a diphenylamine skeleton.** *Biol Pharm Bull* 1998, **21**:544-546.
63. **Targeted Libraries Generation 1. Enamine: Kiev, Ukraine.** [www.enamine.net](http://www.enamine.net).
64. Saurat J-H: **Retinoids and psoriasis: novel issues in retinoid pharmacology and implications for psoriasis treatment.** *J AM Acad Dermatol* 1999, **41**:S2-6.
65. Xu Y, Mayhugh D, Saeed A, Wang X, Thompson RC, Dominianni SJ, Kauffman RF, Singh J, Bean JS, Bensch WR, et al.: **Design and synthesis of a potent and selective triazole-based peroxisome proliferator-activated receptor alpha agonist.** *J Med Chem* 2003, **46**:5121-5124.
66. Kurtz TW: **Treating the metabolic syndrome: telmisartan as a peroxisome proliferator-activated receptor gamma activator.** *Acta Diabetol* 2005, **42**:S9-16.
67. Zhi L, Tegley CM, Kallel EA, Marschke KB, Mais DE, Gottardis MM, Jones TK: **5-aryl-1,2-dihydrochromeno[3,4-f] quinolines: a novel class of non-steroidal human progesterone receptor agonists.** *J Med Chem* 1998, **41**:291-302.
68. Brandes LJ, McDonald LM, Bogdanovic RP: **Evidence that the antiestrogen binding site is a histamine or histamine-like receptor.** *Biochem Biophys Res Comm* 1985, **126**:905-910.

69. Dhar A, Liu S, Klucik J, Berlin KD, Madler MM, Lu S, Ivey RT, Zacheis D, Brown CW, Nelson EC, et al.: **Synthesis, structure-activity relationships, and RAR gamma-ligand interactions of nitrogen heteroarotinoids.** *J Med Chem* 1999, **42**:3602-3614.
70. Chandraratna RA: **Tazarotene: first of a new generation of receptor-selective retinoids.** *Br J Dermatol* 1996, **135**:18-25.
71. Hashimoto Y, Miyachi H: **Nuclear receptor antagonists designed based on the helix-folding inhibition hypothesis.** *Bioorg Med Chem* 2005, **13**:5080-5093.
72. Ahmadiani A, Mandgary A, Sayyah M: **Anticonvulsant effect of flutamide on seizures induced by pentylenetetrazole: involvement of benzodiazepine receptors.** *Epilepsia* 2003, **44**:629-635.
73. Zhang Q, Muegge I: **Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring.** *J Med Chem* 2006, **49**:1536-1548.
74. Mestres J: **Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery.** *Drug Discov Today* 2005, **10**:1629-1637.

## Chapter III.3 – Targeted library design

The use of relatively small libraries containing compounds designed to have a high probability of activity for a given protein or protein family has recently increased in pharmaceutical industry. In this chapter, a review on new trends in ligand-based techniques for designing targeted libraries, with special focus on nuclear receptors, is provided. Additionally, we discuss on the need for a quality assessment of any targeted library. Actually key features of these libraries, like coverage of the target and chemical spaces or possible biases towards a few particular chemical series or protein targets, are hardly ever assessed and provided.

Papers included in this chapter:

- Gregori-Puigjané E, Mestres J: **Designing chemical libraries directed to nuclear receptors**. In *Nuclear Receptors as Drug Targets*. Edited by Ottow E, Weinmann H: Wiley-VCH: New York; 2008.
- Gregori-Puigjané E, Mestres J: **Coverage and bias in chemical library design**. *Curr Opin Chem Biol* 2008, [doi:10.1016/j.cbpa.2008.03.015](https://doi.org/10.1016/j.cbpa.2008.03.015).





# Designing Chemical Libraries Directed to Nuclear Receptors

Elisabet Gregori-Puigjané and Jordi Mestres

*Chemotargets S.L. and Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain*

## 1 Introduction

In the early 1990s, the advent of high-throughput screening (HTS) increased dramatically the capacity for testing compounds. The implementation of this wonder piece of robotic equipment in pharmaceutical industry was soon perceived as the technological solution to the relatively poor performance of drug discovery in terms of new chemical entities approved per year. Obviously, an increase in the capacity for testing compounds implied immediately increasing the number of compounds available for testing. Within this scenario, the size of the corporate compound collection was perceived as one of the key aspects for having a successful HTS campaign. Accordingly, the high demand for more compounds generated suddenly a strong need for wide compound synthesis and acquisition activities directed mainly at obtaining optimally diverse screening libraries [1].

However, a review of the performance of HTS in its first years of implementation revealed that the number, diversity, and progressability of the hits identified were below original expectations [2]. It became then clear that augmenting the capacity for testing alone was not sufficient for delivering high-quality leads and that more effort was required in carefully balancing the composition of the screening collections with compounds containing features compatible with the nature of the targets or target classes of corporate interest. Therefore, novel strategies were conceived to design chemical libraries focussed to a particular target or directed to entire protein families to enrich corporate collections with a pool of targeted compounds that could complement those selected by diversity means [3]. Some of the main protein families for which chemical libraries have been designed in the last few years include G protein-coupled receptors [4,5], kinases [5], serine proteases [6], and nuclear receptors [7].

The focus of this contribution will be on introducing novel knowledge-based strategies for designing chemical libraries directed to the family of nuclear receptors, with special emphasis on more generic often overlooked aspects such as the degree of expected family coverage and bias by the compounds in the library. As extensively exposed in previous chapters, nuclear receptors form a family of ligand-activated transcription factors that regulate a variety of biological processes, including lipid and glucose homeostasis, detoxification, cellular differentiation, embryonic development and orphan physiology. Consistent with these important regulatory roles, mutations in nuclear receptors are associated with many common human diseases such as cancer, diabetes, and osteoporosis, and thus they are considered highly relevant protein targets [8]. Furthermore, many nuclear receptors play also an important role in mediating the induction of hepatic cytochrome P450s, a class of enzymes involved in drug metabolism and in the toxification and detoxification of xenochemicals prevalent in the environment. Accordingly, many nuclear receptors are also regarded as potential off-targets [9]. Finally, there are still a number of orphan nuclear receptors involved in novel regulatory systems that impact human health for which ligands have yet to be identified and that are likely to lead to the discovery of new drugs in the near future [10]. The combination of these three aspects makes nuclear receptors a protein family of utmost therapeutic

relevance for pharmaceutical industry and, thus, highlights the need for having access to chemical libraries designed especially to cover the entire family.

## 2 Collecting and storing prior knowledge

The design of targeted chemical libraries is an activity that requires the availability of prior knowledge either on bioactive ligands (ligand-based approaches) or on crystallographic data (target-based approaches) for the different members of the protein family of interest. In this respect, one of the current challenges in biomedical research is to collect, store, organize, and connect the increasing amount of data being generated around small molecules, proteins, genes, pathways, and diseases [11]. The efficient access to and linkage of all these data will essentially constitute the stepping stone towards the establishment of novel integrative knowledge-based approaches to drug discovery activities, in particular to the design of targeted chemical libraries [12,13].

A number of initiatives have focused recently on collecting and storing the structures of small molecules for which pharmacological data to a given protein target have been reported in the literature, giving rise to the so-called annotated chemical libraries [14]. Among those, the MDL Drug Data Report (MDDR) includes information on therapeutic action and biological activity for over 132,000 compounds gathered from patent literature, journals, and congresses [15], the WOMBAT database offers currently data on 307,700 biological activities for 154,236 molecules annotated to 1,320 protein targets reported in medicinal chemistry journals over the last 30 years [16], the AurSCOPE databases offer a collection of chemical libraries containing over 323,000 molecules annotated to over 1,300,000 biological activities related to members of therapeutically relevant protein families covered in over 38,000 publications [17], and the MedChem and Target Inhibitor databases compile around 2,000,000 molecules with biological activity, toxicity and pharmacological information for therapeutically relevant protein families extracted from 20,000 publications [18]. All these commercial databases provide an invaluable source of pharmacological data for ligands that can be ultimately exploited for designing targeted libraries.

Under the same spirit, a more modest initiative took place in our laboratory to assemble an annotated chemical library directed to the nuclear receptor family (NRacl) on the basis of public sources of information, mainly reviews and medicinal chemistry journals of the last 13 years [19]. Each entry in NRacl contains information on its topological chemical structure and the connection to nuclear receptors is established through the associated pharmacological data ( $K_i$ ,  $IC_{50}$  and/or  $EC_{50}$ ), as reported. At this stage, only biologically active compounds (activity < 10  $\mu$ M) were entered into NRacl. In total, NRacl includes currently 2718 small molecules connected to 29 nuclear receptors, some molecules containing multiple annotations to nuclear receptors.

Since the ability to extract knowledge from annotated chemical libraries will be highly determined by the way chemical and biological data are stored, when constructing NRacl special emphasis was put in storing both chemical structures and nuclear receptors using appropriate

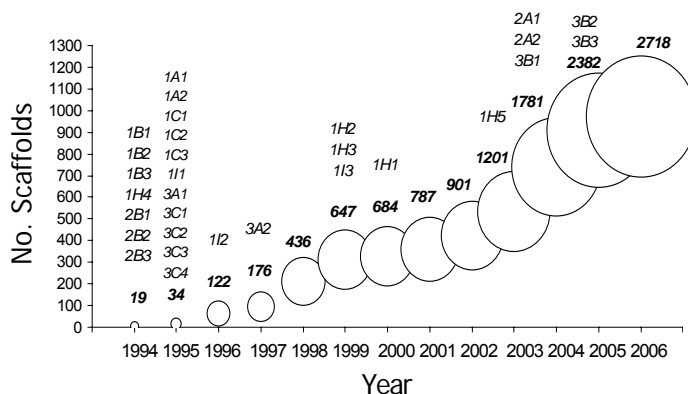
unique identifiers and classification schemes. For chemical structures, we used an in-house proposed Chemical Structure Code (CSC) purely based on topological features of molecules. Accordingly, each molecule in NRacl is identified with a unique hierarchical five-level CSC [19]. The first and second levels are integers specifying, respectively, the number of rings in the largest ring system present in the molecule and the total number of ring systems in the molecule. The third, fourth, and fifth levels are a unique seven-character hash code for the molecular framework, scaffold, and the complete molecular structure, respectively. On the other hand, for the annotation of ligands to nuclear receptors, we avoided using the receptor names directly but using instead the more compact and unified three-character code nomenclature system proposed by the Nuclear Receptors Nomenclature Committee [20]. Within this scheme, the first character is a number that designates the subfamily. There are six main subfamilies, assigned to identifiers 1 to 6. All nuclear receptors in these subfamilies contain a highly conserved zinc-finger DNA-binding domain (DBD) and a less conserved ligand-binding domain (LBD). However, some unusual receptors contain only one of the two conserved domains and thus an additional subfamily, assigned to identifier 0, has been included to account for them. The second character is a capital letter specifying the group within the subfamily, and the third character is a number identifying the individual nuclear receptor within a group. Globally, this classification scheme of nuclear receptors defines at present 7 subfamilies, 25 groups, and 73 receptors. The use of hierarchical classification schemes for both molecules and receptors takes the exploitation of family-directed annotated chemical libraries to another level, the added value coming from the fact that complete flexibility exists for extracting knowledge at all levels of those classification schemes. For the sake of clarity, Table 1 compiles the list of 36 nuclear receptors for which prior knowledge on both small molecules and/or receptor structures is currently available.

Subfamily	Group	Receptor	Name	Abbreviation	BL	PS
1.	A.	1	Thyroid hormone	TR $\alpha$	81	3
		2		TR $\beta$	93	12
	B.	1	Retinoic acid	RAR $\alpha$	107	1
		2		RAR $\beta$	112	2
		3		RAR $\gamma$	118	9
	C.	1	Peroxisome proliferator-activated	PPAR $\alpha$	134	4
		2		PPAR $\beta$	37	8
		3		PPAR $\gamma$	280	25
	F.	1	RAR-related orphan	ROR $\alpha$	0	2
		2		ROR $\beta$	0	3
	H.	1	Ecdysone	ECR	2	3
		2	Liver X	LXR $\beta$	4	6
		3		LXR $\alpha$	30	2
		4	Farnesoid X	FXR	33	3
	I.	1	Vitamin D	VDR	16	18
2		Pregnane X	PXR	1	6	
3		Constitutive androstane	CAR	0	4	
2.	A.	1.	Hepatocyte nuclear factor 4	HNF4 $\alpha$	0	2
		2.		HNF4 $\gamma$	0	1
	B.	1	Retinoid X	RXR $\alpha$	135	18
		2		RXR $\beta$	78	2

Subfamily	Group	Receptor	Name	Abbreviation	BL	PS
		3		RXR $\gamma$	97	1
		4	Ultraspiracle protein	USP	0	5
3.	A.	1	Estrogen	ER $\alpha$	729	39
		2		ER $\beta$	757	21
	B.	1	Estrogen receptor-related	ERR $\alpha$	0	2
		3		ERR $\gamma$	9	11
	C.	1	Glucocorticoid	GR	312	3
		2	Mineralocorticoid	MR	14	10
		3	Progesterone	PR	380	7
		4	Androgen	AR	280	32
4.	A.	1	Nerve growth factor IB	NGFI-B	0	2
		2		NURR1	0	1
		4		DHR38	0	1
5.	A.	1	Steroidogenic factor-1	SF1	0	4
		2	Fetoprotein TF	FTF	0	5

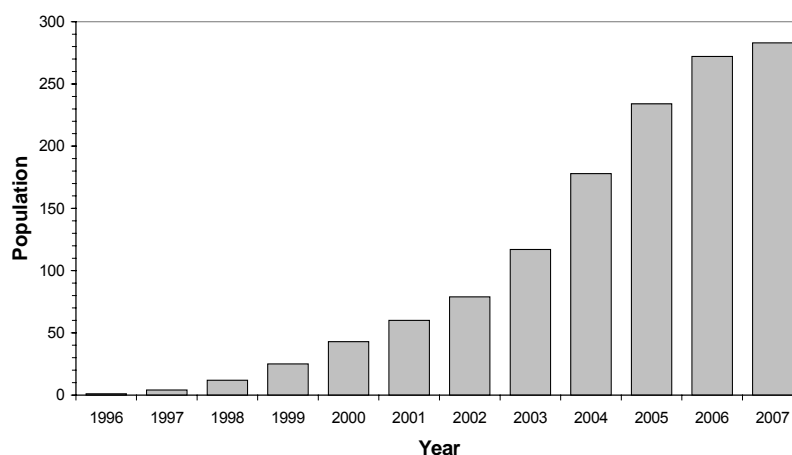
**Table 1.** List of nuclear receptors and the number of bioactive ligands (BL) and protein structures (PS) currently available for each of them.

Recent trends in nuclear receptor medicinal chemistry can be analysed in terms of the number of novel scaffolds representing all small molecules reported to have affinity under 10  $\mu$ M for a nuclear receptor over the last 13 years. As can be observed in **Figure 1**, early efforts in 1994 and 1995 focused primarily on synthesising compounds directed to the groups of retinoic acid receptors (RARs; 1B), retinoic X receptors (RXRs; 2B), thyroid receptors (TRs; 1A), peroxisome proliferator-activated receptors (PPARs; 1C), and estrogen receptors (ERs; 3A). Accordingly, the overall distribution of annotations among all nuclear receptors is a fair reflection of the historical therapeutic relevance of some of the members of these groups (Table I). In particular, the nuclear receptor containing the largest number of chemical annotations is the ER subtype alpha (ER $\alpha$ ; 3A1), an important target in reproductive medicine and cancer research. Due to its high homology, many compounds binding to ER $\alpha$  are also reported to be active to ER $\beta$  thus justifying the large number of annotations present also for the latter. Another nuclear receptor for which vast information on bioactive ligands is available is PPAR gamma (PPAR $\gamma$ ; 1C3), widely recognized as a key regulator in multiple metabolic pathways including fatty acid and carbohydrate metabolism and thus being considered a relevant target in cardiovascular research. In contrast, the groups of hepatocyte nuclear factor 4 receptors (HNF4s; 2A) and estrogen-related receptors (ERRs; 3B) are among the youngest in terms of medicinal chemistry exploration. This will be the main body of ligand-based information used for designing targeted chemical libraries to nuclear receptors.



**Figure 1.** Progress over time in the number of novel scaffolds generated from nuclear receptor medicinal chemistry efforts in the last 13 years. The radius of the circles reflects the cumulative number of molecules entered in NRaL (specific numbers are given above each circle, in bold). When a bioactive molecule to a nuclear receptor was first encountered in the literature, its code (in italics) is added above the circle. (Reproduced with permission of Bentham Science Publishers Ltd.)

Apart from pharmacological data on ligands, the other important source of knowledge generated within protein families is the availability of experimentally determined protein structures. Recent advances in high-throughput methods for protein expression and production, NMR spectroscopy, and X-ray crystallography have led to a significant rise in the number of protein structures solved [21]. Many of these structures are ultimately deposited and made publicly accessible in the Protein Data Bank (PDB), currently containing over 47,000 entries and its size continuing to increase annually at an almost exponential rate [22]. In particular, the first structure of a DBD of a nuclear receptor was deposited in the PDB in 1991 [23], whereas the first LBD structure was not deposited until four years later [24]. Since then, the number of nuclear receptor structures has grown significantly and, as per 27-July-2007, there are 319 entries in the PDB involving 294 separate PDB files, some of which having more than one NR number associated with them [25]. Of them, 36 entries correspond to nuclear receptor DBDs covering 5 subfamilies, 11 groups, and 16 receptors. The remaining 283 entries correspond to nuclear receptor LBDs covering a total of 6 subfamilies, 14 groups, and 37 receptors, which provide essential structural information that can be exploited for designing targeted libraries directed to the entire family of nuclear receptors (Table I). To complement **Figure 1** from a receptor structure viewpoint, the evolution in the number of LBD structures deposited in the PDB over the years is illustrated in **Figure 2**. Overall, beyond the mere increase in the population of structures, it is important to stress again that the general adoption of classification schemes for proteins is an essential aspect for analyzing quantitatively the functional coverage and structural bias of target families in the PDB, and thus ultimately for assessing the applicability of structure-based approaches to targeted chemical library design [26].



**Figure 2.** Growth in the number of nuclear receptor ligand-binding domain structures deposited in the Protein Data Bank

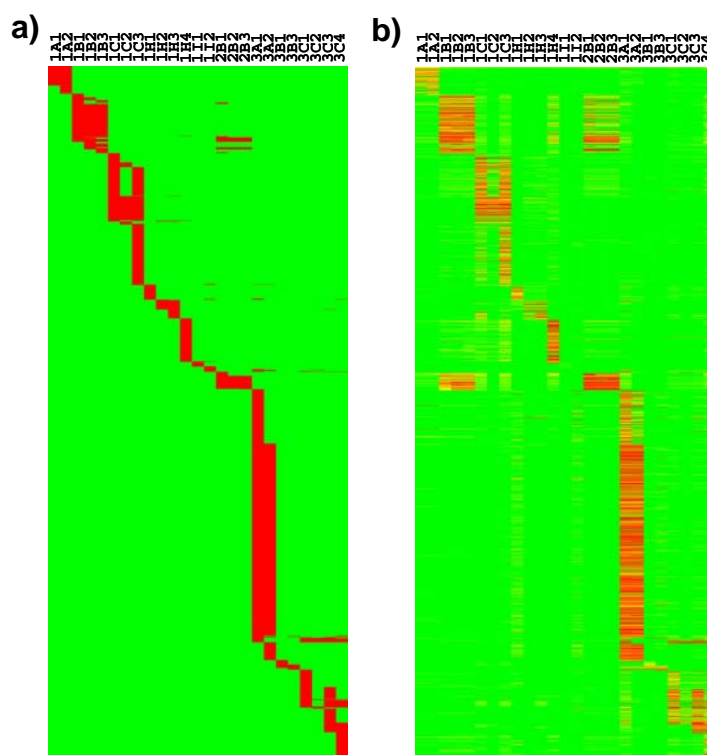
### 3 Nuclear receptor profiling

Beyond having access to prior knowledge, a key aspect for being able to design compounds directed to the protein members of a particular family is the ability to detect that compounds possess the right features arranged in an optimal complementary manner to the protein cavities of interest by properly scoring them on the basis of some predefined metrics. In this respect, the process of scoring and ranking molecules in large chemical libraries according to their likelihood of having affinity for a certain target is generally referred to as virtual screening [27]. The term itself was coined in the late 1990s when computer-based methods reached sufficient maturity to offer an alternative to experimental HTS techniques. In spite of the initial reluctance, over the years pharmaceutical industry has learned to accept that virtual screening methods can indeed be an efficient complement to HTS to the point that they have undoubtedly become an integral part of today's lead generation process [12]. It is worth emphasising again that, in contrast to technology-driven HTS, virtual screening is a knowledge-driven approach that requires structural information either on bioactive ligands for the target of interest or on the target itself. Comparative studies have suggested that information about a target obtained from known bioactive ligands is as valuable as knowledge of the target structures for identifying novel bioactive scaffolds through virtual screening [28]. Therefore, the final choice for a ligand-based or target-based method will ultimately depend on the type and amount of information available without *a priori* having a large impact on performance.

With virtual ligand screening well integrated in the drug discovery process, a wave of new strategies is currently emerging with the aim of exploiting both the ever increasing amount of information and computational power available to add a biological dimension to traditional single-target virtual screening. In this respect, it has been shown recently that these strategies are capable of estimating the pharmacological profile of molecules on multiple targets and promise to have a strong influence in drug discovery as a means for detecting potential side effects of compounds due to off-target affinities earlier on during the optimisation process [29]. As mentioned above, some members of the nuclear receptor family may be considered relevant off-targets to

which it is important to avoid affinity to a safe degree. Therefore, not surprisingly, some ligand-based and target-based approaches to nuclear receptor profiling have been recently described in the scientific literature.

From a ligand-based viewpoint, it is worth understanding that the relative success of ligand-based methods depends to a great extent on the use of biologically-relevant mathematical representations of molecules. In this respect, a novel set of molecular descriptors called SHED (SHannon Entropy Descriptors) was recently introduced [30]. SHED are derived from distributions of atom-centered feature pairs extracted directly from the topology of molecules stored in standard MDL's SD file format. From a SD file, each atom in a molecule is first mapped to a Sybyl atom type. Subsequently, each atom type is assigned to one or more of four atom-centered features, namely, hydrophobic (H), aromatic (R), acceptor (A), and donor (D). For example, an aliphatic C.3 carbon will be assigned to a hydrophobic feature (H), whereas a protonated N.4h nitrogen will be assigned to both aromatic and donor features (R,D). Then, the shortest path length between atom-centered feature pairs is derived and its occurrence at different path lengths stored to create a feature-pair distribution. A maximum path length of 20 bonds was used. Feature pairs being at distances over 20 bonds are accumulated in the last bin. An equivalent distribution is derived for each of the ten possible feature pairs resulting from all pair combinations of the four features used. In summary, each chemical structure is ultimately represented by a SHED profile composed of 10 real numbers reflecting the particular feature-pair distributions present in the molecule.



**Figure 3.** Comparison between the heatmap representing all original annotations extracted from bibliographical sources and stored in NRaCl (a) and the heatmap reflecting the minimum SHED Euclidean distances between the SHED profile of each molecule and the set of non-redundant SHED profiles annotated to each nuclear receptor (b). Color coding: (a) red is annotated and green not annotated; and (b) red reflects distance values close to 0.0 and as distances increase in magnitude they turn to orange, yellow, and finally green at a value of 1.2 and over. (Reproduced with permission of the American Chemical Society)



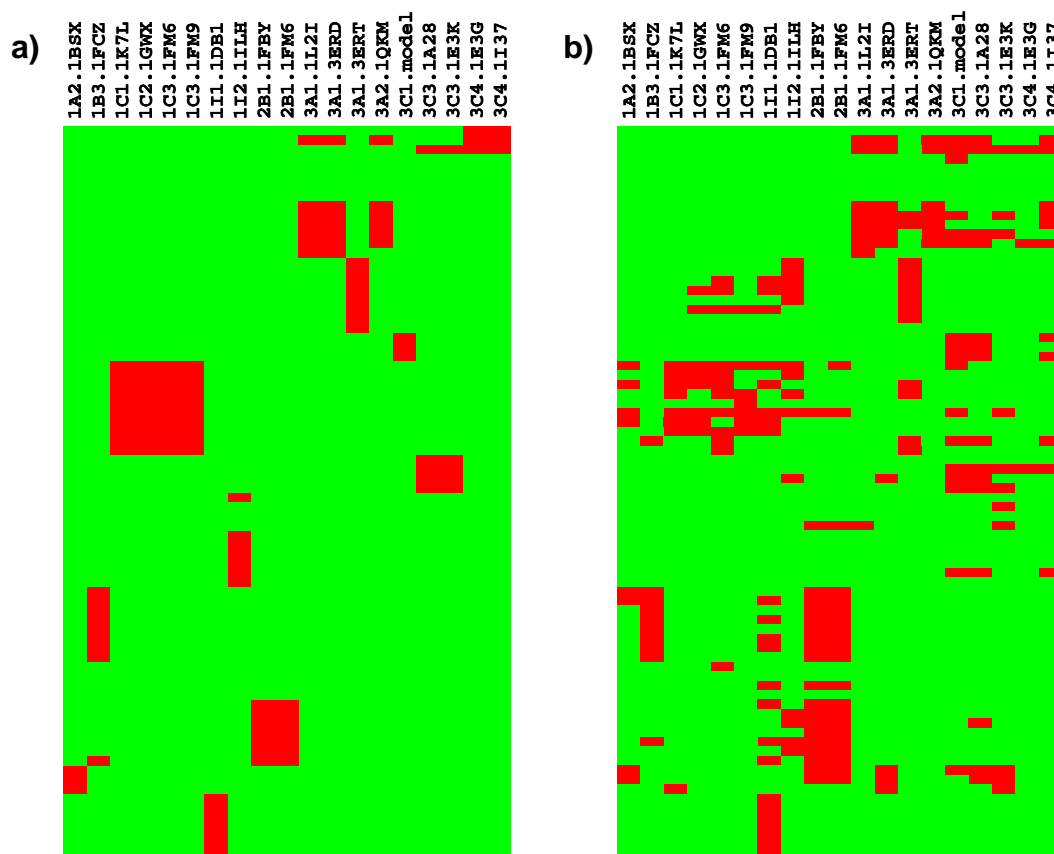
These SHED descriptors were recently used to profile *in silico* a chemical library of 2033 molecules against 25 nuclear receptors [31]. As described above, this annotated chemical library to nuclear receptors (NRaCl) was assembled internally in our laboratory from various medicinal chemistry sources [19]. The distribution of annotations contained at that stage in NRaCl is visually illustrated in **Figure 3a**. In the heatmap shown, annotations of molecules (in rows) to nuclear receptors (in columns) are represented as red cells, meaning that the interaction of a particular molecule with a specific nuclear receptor has been positively reported and experimentally quantified in the literature with a pharmacological value below 10  $\mu$ M. In contrast, green cells indicate current lack of information on the possibility of any interaction between a given molecule and a certain nuclear receptor. The extent of the green area denotes the existence of large information gaps, clearly one of the main limitations of dealing with annotated chemical libraries relying on data extracted directly from public sources of information. This is due to the fact that, because of limited time and resources, molecules are usually not screened systematically through a large panel of protein targets for the sake of obtaining the maximum amount of information possible but solely to the target of interest at that point in time. But even if they were screened through multiple targets, habitually only a limited amount of data is made available, since publishing large amounts of negative data is often regarded as not informative. These important, yet often overlooked, aspects lead to a situation of data incompleteness within the interaction matrix depicted as a heatmap in **Figure 3a**.

The information on bioactive ligands contained in NRaCl was then used to derive a ligand-based model of each nuclear receptor based on the SHED descriptors defined above. Essentially, the scoring of each compound in a chemical library with respect to a given nuclear receptor is assigned to the minimum value of all Euclidean distances calculated between the SHED profile of the target compound and each one of the SHED profiles describing the molecules annotated to that particular nuclear receptor [31]. The result of applying this process to each one of the 2033 molecules with non-redundant SHED profiles present in NRaCl is given in **Figure 3b**, in which the order of the molecules is exactly the same as the one obtained from the original annotations shown in **Figure 3a**. In contrast to the binary heatmap illustrated in **Figure 3a**, in which red was annotated and green was not annotated, **Figure 3b** presents a color gradation between red and green reflecting the value of the minimum SHED Euclidean distance between the SHED profile of each molecule and the set of non-redundant SHED profiles annotated to each nuclear receptor. Taking the annotation threshold of 0.6 as the center of the color scale, distance values close to 0.0 are represented in red, those close to 0.6 are seen as light orange, and as distances increase in magnitude they turn to yellow and finally green at a value of 1.2 and over. There are two main aspects worth mentioning when comparing the heatmaps of **Figures 3a** and **3b**. On one hand, it is remarkable to notice that the essential pattern observed when plotting the original annotations (**Figure 3a**) is preserved when molecules are processed through the ligand-based descriptor model of nuclear receptors (**Figure 3b**). This result reveals that the remaining molecules in NRaCl are to a great extent representative of the molecule being processed after leaving that molecule out, something that can only be achieved if the annotated chemical space has been sufficiently

saturated with as many known bioactive molecules as possible. On the other hand, despite the clear discrimination between nuclear receptor groups, some correlation patterns between them emerge. The most apparent example is the clear correlation observed in **Figure 3b** between RARs (1B1, 1B2, and 1B3) and RXRs (2B1, 2B2, and 2B3), a result that provides an indication of the potential of this approach for understanding side effects through the identification of off-target affinities.

In contrast, from a target-based viewpoint, it is also important to understand that the relative success of a target-based method will depend to a great extent on the availability of representative experimental crystal structures for all members of the protein family of interest, as the performance of these methods tend to degrade depending on whether a holo (ligand bound), an apo, or a model structure of the protein is available [32]. In this respect, the family of nuclear receptors is relatively well covered in terms of structural information available, with representative LBD structures for up to 37 receptors (Table I). Another equally important aspect in target-based methods is the docking procedure used to generate a binding hypothesis of the interaction between the ligand and the protein, which involves conformational sampling and scoring of small molecules into protein cavities [33,34].

Despite the decent amount of structural information available, applications of target-based methods to nuclear receptor profiling have been so far scarce. Perhaps the most comprehensive work in this respect is the recent systematic virtual screening of a library composed of 78 known active ligands against 19 different structures representative of 13 nuclear receptors [35]. Note that for some nuclear receptors, more than one crystal structure was considered to assess the dependency of the results on the particular conformation of the receptor. As mentioned, each one of the 78 ligands is a known binder to certain nuclear receptors (black bars in **Figure 1** of reference [35]). This information is illustrated in **Figure 4a** (the structure-based counterpart of **Figure 3a**), in which annotations of molecules (in rows) to nuclear receptors (in columns) are represented as red cells, meaning that the interaction of that particular ligand with a specific nuclear receptor has been positively reported and experimentally quantified in the literature. Remarkably, **Figure 4a** provides a very crisp picture of the interaction of ligands to nuclear receptors, promiscuity being observed only to a limited degree between some of the steroid hormone receptors. However, interpretation of **Figure 4a** at this stage needs to be taken with the same level of caution highlighted previously when discussing **Figure 3a**, since large information gaps exist in these data (represented by green cells) due to the incompleteness of the experimental information.



**Figure 4.** Comparison between the heatmap representing all original annotations extracted from bibliographical sources on 78 known nuclear receptor binders (a) and the heatmap reflecting the annotation associated to a docking score being above the score threshold to select 1% of a 5000 random molecule database (b). Color coding: red is annotated and green not annotated. (Information derived from *J. Med. Chem.* **2003**, *46*, 3045-3059).

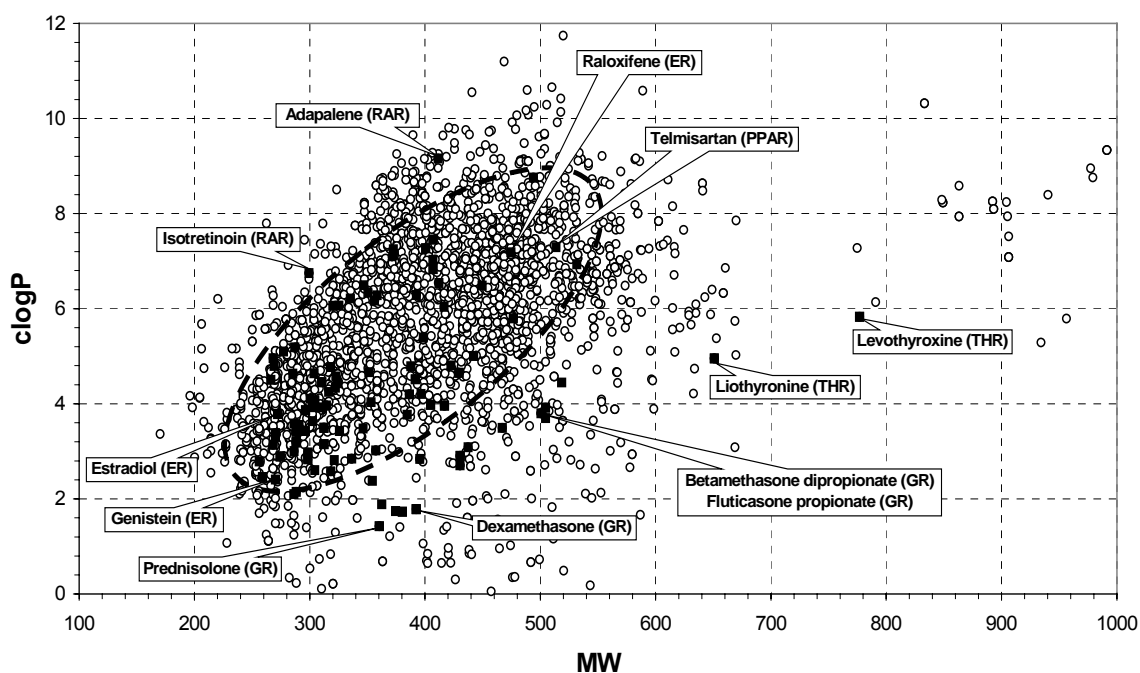
When this set of 78 nuclear receptor ligands was put into a library of 5000 random compounds and scored, the sensitivity of the method for distinguishing between true binders and non-binders could be assessed. In order to perform an analysis comparative to the one presented above when using a ligand-based method, an annotation criterion was selected. Accordingly, a ligand was considered annotated to a given nuclear receptor if its docking score was above the score threshold to select 1% of a 5000 random compound database (the horizontal solid black line in **Figure 1** of reference [35]). The resulting heatmap is depicted in **Figure 4b**. In essence, there are two main aspects worth mentioning when comparing the heatmaps of **Figures 4a** and **4b**. On one hand, it is remarkable how the target-based method is able to identify the correct nuclear receptor target for the majority of true binders. On the other hand, in the same lines as observed previously in **Figure 3b** for a ligand-based method, some stronger correlation patterns between nuclear receptors emerge as a consequence of the extend of the promiscuity profiles predicted. For example, the relatively limited signal shared between steroid hormone receptors in **Figure 4a** is transformed in **Figure 4b** in a strong promiscuity signal among them. But again, the conclusions extracted from **Figure 4b** need to be taken with caution, since a full affinity matrix between the 78 ligands and the 13 receptors is not available.

In summary, examples have been provided in which both ligand-based and target-based methods performed decently when profiling compounds against the family of nuclear receptors. Therefore, it is reasonable to say that these methodologies have reached sufficient level of maturity to be applied sensibly for designing the next generation of chemical libraries directed to entire protein families.

## 4 New trends in designing targeted libraries




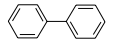
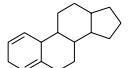

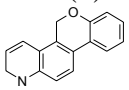
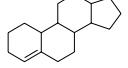
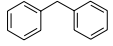
Despite its recognised relevance, it is remarkable to realise that very few reports document recent efforts towards designing chemical libraries particularly directed to the family of nuclear receptors. However, analyses on the characteristics of nuclear receptor ligands have revealed valuable information on the specific molecular properties and topological substructures these ligands possess compared to other family-directed sets of ligands. For example, in terms of molecular descriptors, nuclear receptor drugs seem to be characterised by significantly high mean clogP values (4.1) and low mean counts of oxygen and nitrogen atoms (3.8) compared to drugs designed for primary targets belonging to other protein families [36]. Similar trends were also found when analysing sets of hit-to-lead ligand pairs instead of drugs [37]. In this case, a mean clogP value of 5.0 was found for nuclear receptor ligands, the largest mean clogP value among all compound entries directed to targets belonging to a list of eleven protein families.

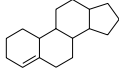
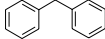

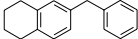


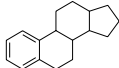


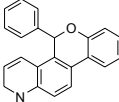
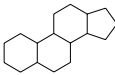

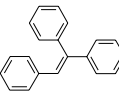
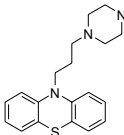

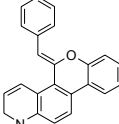
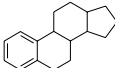
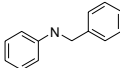
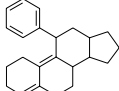
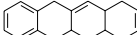

To investigate this aspect further, we took all ligands annotated to nuclear receptors in Wombat having an affinity value (pKi, pIC50, or pEC50) larger than 7.0. This resulted in a list of 2929 molecules containing 3839 annotations to 24 nuclear receptors. **Figure 5** illustrates the distribution of this set of 2929 nuclear receptor bioactive ligands (white circles), together with a set of 135 nuclear receptor drugs (black squares), in the plane defined by two molecular descriptors related to size and hydrophobicity such as molecular weight (MW) and clogP values. It was observed that 65.4% of all bioactive ligands fail to meet the Lipinski criteria for both MW and clogP values and thus, under the rule-of-five, they would receive an alert as having poor oral bioavailability. One could then conclude that high-affinity ligands for nuclear receptors are intrinsically handicapped for oral bioavailability relative to high-affinity ligands for other protein families. In fact, one could delineate an oval-shaped region within the MW vs clogP space that would contain the vast majority of both nuclear receptor bioactive ligands and drugs. The south-west of this oval region appears to be populated by the smaller more compact steroid-like drugs, such as the estrogen receptor agonists estradiol and genistein, whereas the north-east side is occupied by larger hydrophobic compounds, such as the estrogen receptor antagonist raloxifene and the PPAR modulator telmisartan. Outside this region, we find a set of outliers including the iodine substituted thyroid hormone receptor drugs (levothyroxine and liothyronine), a variety of glucocorticoid receptor modulators (such as betamethasone dipropionate, fluticasone propionate, dexamethasone, and prednisolone), and some retinoic-acid receptor ligands (such as isotretinoin and adapalene). Therefore, this region could certainly be used as a fast molecular descriptor filter when designing chemical libraries directed to the nuclear receptor family.



**Figure 5.** Distribution of a set of 2929 nuclear receptor bioactive ligands (white circles) and 135 nuclear receptor drugs (black squares) in the plane defined by the molecular weight (MW) and clogP descriptors. The dashed region defines the nuclear receptor space.

Besides ranges of molecular descriptor values, another strategy often applied for biasing chemical libraries towards particular protein families is to generate and synthesise compounds around so-called privileged substructures [38]. However, the results obtained in a recent substructure-class analysis of ligand sets from five target families, namely, G protein-coupled receptors (GPCRs), nuclear receptors, ligand-gated ion channels, serine proteases, and protein kinases put a question mark on the actual existence of target-family-privileged substructures [39]. For nuclear receptors in particular, the study revealed that nuclear receptor substructure classes were present in 40% of a total of 21620 GPCR ligands, 30% of a total of 3792 ion channel ligands, 17% of a set of 1079 kinase ligands, and 15% of a set of 3015 protease ligands but, most interestingly, 45% of a set of 10000 random ligands. Altogether, these results are an indication that the nuclear receptor substructure classes generated are in fact non-privileged substructure classes for the nuclear receptor family and thus, its use for designing targeted chemical libraries is questionable.

NRactive	Drugs	NCI
 60 (8)	 326	 30560
 43 (6)	 45	 2810
 73 (4)	 44	 2328

 23 (4)	 41	 2279
 20 (4)	 21	 2013
 86 (3)	 17	 1850
 51 (3)	 17	 1786
 39 (3)	 15	 1448
 39 (3)	 15	 1076
 11 (3)	 15	 1037

**Table 2.** List of the 10 most promiscuous scaffolds in a set of nuclear receptor bioactive ligands (NRActive) and 10 most populated scaffolds in a set of drugs and drug candidates (Drugs) and in the open NCI database (NCI). Numbers refer to the population of molecules containing each scaffold in the respective sets. In parenthesis, the level of nuclear receptor promiscuity associated to scaffolds in the NRActive set.

In order to investigate this further, we performed a comparison of the most populated scaffolds in three sets of ligands, namely, the same set of 2929 nuclear receptor bioactive ligands used above (NRActive set), a set of 2900 drugs and drug candidates (Drug set), and the “open NCI database” composed of 250251 compounds (NCI set). The lists of 10 most promiscuous (for NRActive) and 10 most populated (for Drugs and NCI) scaffolds are collected in Table 2, the definition of scaffold in this work being equivalent to that of atomic framework given earlier [40]. As can be observed, phenyl emerges as the most promiscuous scaffold among nuclear receptors, with 60 compounds showing high-affinity for 8 nuclear receptors from 5 different nuclear receptor groups (1B1, 1B3, 1C1, 1H3, 2B1, 2B2, 2B3, 3C4). Comparatively, phenyl is also by far the most populated scaffold in both Drugs and NCI sets. The second most promiscuous scaffold among nuclear receptors is a biphenyl core. It is present in 43 compounds having high-affinity for 6 nuclear receptors from 4 different nuclear receptor groups (1B2, 2B1, 2B2, 2B3, 3A2, 3C4). A recent study suggested that high-throughput screening libraries enriched with biphenyl-containing compounds can be expected to have increased chances of yielding high-affinity ligands for proteins [41]. The

results presented here for nuclear receptors, together with the fact that the biphenyl substructure is found also quite frequently in GPCR ligands, would be supportive of the conclusions reached in that study. Going further down the list of most promiscuous scaffolds present in the NRactive set we notice that the remaining scaffolds have promiscuities below 5 and that, for the majority of these scaffolds, their associated nuclear receptors belong to the group of glucocorticoid-like receptors.

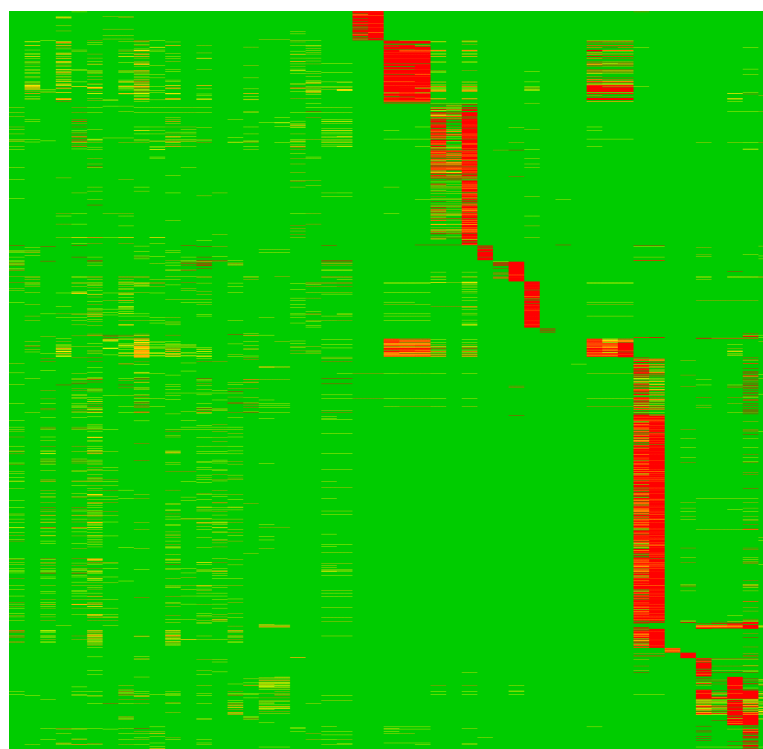
Altogether, this scaffold promiscuity analysis leads to two main conclusions. On one hand, it is remarkable to realize that besides phenyl and biphenyl no other scaffold could be identified that could cover vastly this apparently well conserved protein family, in terms of tertiary structure. In the lines of a recent study [39], the true existence of privileged scaffolds for the entire nuclear receptor family remains thus unclear and, consequently, the use of nuclear receptor substructure classes for targeted chemical library design is also dubious. On the other hand, also in agreement with a previous work [41], the present results would indicate that enriching chemical libraries with compounds containing a variety of substituent decorations around the phenyl and biphenyl cores could be a strategy worth considering to deorphanise nuclear receptors. However, it must be stressed at this stage that these conclusions have been derived from information contained in annotated chemical libraries. As highlighted earlier [19], due to limited time and resources, molecules are usually not screened systematically against the complete panel of proteins forming a family for the sake of generating the maximum amount of information but solely to the target(s) of interest, leading to a data completeness issue. Should the molecules contained in the NRactive set be screened against the entire nuclear receptor family, the existence of privileged scaffolds, currently hidden because of lack of information, could potentially be revealed.

Lately, we have seen a new trend in designing targeted chemical libraries in which not only the descriptor profiles and the presence of particular substructures observed in known bioactive compounds is considered but also the concrete potential coverage of the protein family by the molecules in the chemical library is assessed [31]. This represents adding a biological dimension to the process and that both chemical and biological diversity are included when designing the composition of a targeted chemical library. Using a ligand-based approach to nuclear receptor profiling, **Figure 3b** provides an example of a chemical library covering fully all nuclear receptors under consideration. However, it reveals also that the chemical library is clearly biased with compounds potentially being active to the estrogen receptors and thus it is far from being optimally diverse in terms of projected nuclear receptor pharmacology. Addressing protein family coverage and bias should become standard procedure when designing targeted chemical libraries.

Finally, a new wave of computationally efficient *in silico* pharmacology methods promise to have the ability to profile large chemical libraries against hundreds of protein targets in a reasonably short period of time. These activities may lead to the identification of potential protein family off-targets, defined as those protein targets against which compounds designed for a particular protein family may have some residual affinity. We have profiled the NRacl chemical library used to generate **Figure 3b** against a panel of 674 protein targets covering 411 enzymes,

168 GPCRs, 48 ion channels, 32 nuclear receptors, and 15 transporters. Of those, only 6 targets contained more than 100 annotations from compounds annotated also to any nuclear receptor and only 23 targets had more than 50 annotations. The corresponding heatmap is illustrated in **Figure 6**, in which the family off-target signals (on the left) can be compared against the nuclear receptor profile (on the right) equivalent to **Figure 3b**. In rank #2 of the off-target list we find COX-2. Interestingly, using a cavity site-based similarity searching method, a relationship between the PPAR-gamma agonist binding pocket and the COX-2 binding site was recently identified [42]. Also, in rank #14 of the off-target list we can locate angiotensin II type 1-receptor (ATR1). Again, evidence could be found in the literature of clear cross-pharmacology between ATR1 antagonists

and activation of PPARgamma [43]. Unfortunately, we could not find evidence in the literature relating directly the other off-target names in the list to nuclear receptors. Further investigation is underway in our laboratory.



**Figure 6.** Heatmap reflecting the minimum SHED Euclidean distance between SHED profile of each molecule and the set of non-redundant SHED profiles annotated to each one of the 26 nuclear receptor targets and the 23 off-targets identified. Color coding: red reflects distance values close to 0.0 and as distances increase in magnitude they turn to orange, yellow, and finally green at a value of 0.6 and over

## 5 Conclusions and Outlook

Nuclear receptors are a protein family of utmost importance for pharmaceutical research and thus chemical libraries directed to probe this family exhaustively are required. Lately, a variety of strategies have been applied to designing nuclear receptor chemical libraries. In view of the fresh perspectives novel *in silico* pharmacology methods are offering, it is envisaged that properly addressing coverage and bias during the design process together with the ability to identify



potential protein family off-targets would lead to a new generation of high-content chemical libraries directed to nuclear receptors composed of small molecules exposing a rich diversity of therapeutically-relevant pharmacological profiles. However, recent studies are highlighting the need to go beyond the target level when designing chemical libraries and incorporate information at the pathway level [44]. The relative importance of achieving target selectivity when the target has an intrinsic promiscuity at the pathway level may change the way drug discovery is perceived and smoothly shift from target-focused to systems-oriented research.

## References

1. Blaney JM, Martin EJ: **Computational approaches for combinatorial library design and molecular diversity analysis.** *Curr Opin Chem Biol* 1997, **1**:54-59.
2. Lahana R: **How many leads from HTS?** *Drug Discov Today* 1999, **4**:447-448.
3. Shuttleworth SJ, Connors RV, Fu J, Liu J, Lizarzaburu ME, Qiu W, Sharma R, Wanska M, Zhang AJ: **Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization.** *Curr Med Chem* 2005, **12**:1239-1281.
4. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP: **Property-based design of GPCR-targeted library.** *J Chem Inf Comp Sci* 2002, **42**:1332-1342.
5. Lowrie JF, Delisle RK, Hobbs DW, Diller DJ: **The different strategies for designing GPCR and kinase targeted libraries.** *Comb Chem High Throughput Screen* 2004, **7**:495-510.
6. Lang SA, Kozyukov AV, Balakin KV, Skorenko AV, Ivashchenko AA, Savchuk NP: **Classification scheme for the design of serine protease targeted compound libraries.** *J Comp-Aided Mol Design* 2002, **16**:803-807.
7. Stewart EL, Brown PJ, Bentley JA, Willson TM: **Selection, application, and validation of a set of molecular descriptors for nuclear receptor ligands.** *Comb Chem High Throughput Screen* 2004, **7**:407-412.
8. Gronemeyer H, Gustafsson J-A, Laudet V: **Principles for modulation of the nuclear receptor superfamily.** *Nat Rev Drug Discov* 2004, **3**:950-964.
9. Francis GA, Fayard E, Picard F, Auwerx J: **Nuclear receptors and the control of metabolism.** *Annual Review of Physiology* 2003, **65**:261-311.
10. Shi Y: **Orphan nuclear receptors in drug discovery.** *Drug Discov Today* 2007, **12**:440-445.
11. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, et al.: **The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science* 2006, **313**:1929-1935.
12. Bleicher KH, Bohm H-J, Muller K, Alanine AI: **Hit and lead generation: beyond high-throughput screening.** *Nat Rev Drug Discov* 2003, **2**:369-378.
13. Mestres J: **Computational chemogenomics approaches to systematic knowledge-based drug discovery.** *Curr Opin Drug Discov Devel* 2004, **7**:304-313.
14. Savchuk NP, Balakin KV, Tkachenko SE: **Exploring the chemogenomic knowledge space with annotated chemical libraries.** *Curr Opin Chem Biol* 2004, **8**:412-417.
15. **MDL Drug Data Report.** Edited by. San Leandro, CA: MDL Information Systems, Inc.
16. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity.** In *Chemoinformatics in Drug Discovery*. Edited by Wiley-VCH; 2004:223-239.
17. AurSCOPE databases, Aureus Pharma, France.
18. MedChem and Target Inhibitor databases, GVK Biosciences, India.
19. Cases M, Garcia-Serna R, Hettne K, Weeber M, Lei JV, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family.** *Curr Top Med Chem* 2005, **5**:763-772.

20. Nuclear Receptors Nomenclature Committee: **A unified nomenclature system for the nuclear receptor superfamily.** *Cell* 1999, **97**:161-163.
21. Sali A, Glaeser R, Earnest T, Baumeister W: **From words to literature in structural proteomics.** *Nature* 2003, **422**:216-225.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, H W, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28**:235-242.
23. Luisi BF, Xu WX, Otwinowski Z, Freedman LP, Yamamoto KR, Sigler PB: **Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA.** *Nature* 1991, **352**:497-505.
24. Bourguet W, Ruff M, Chambon P, Gronemeyer H, Moras D: **Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha.** *Nature* 1995, **375**:377-382.
25. Garcia-Serna R, Opatowski L, Mestres J: **FCP: functional coverage of the proteome by structures.** *Bioinformatics* 2006, **22**:1792-1793.
26. Mestres J: **Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery.** *Drug Discov Today* 2005, **10**:1629-1637.
27. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling.** *Br J Pharmacol* 2007, **152**:9-20.
28. Zhang Q, Muegge I: **Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring.** *J Med Chem* 2006, **49**:1536-1548.
29. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry.** *Nat Biotech* 2007, **25**:197-206.
30. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.
31. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to in silico pharmacology: nuclear receptor profiling.** *J Chem Inf Model* 2006, **46**:2725-2736.
32. McGovern SL, Shoichet BK: **Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes.** *J Med Chem* 2003, **46**:2895-2907.
33. Gohlke H, Klebe G: **Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors.** *Angew Chem Int Ed* 2002, **41**:2644-2676.
34. Fradera X, Mestres J: **Guided docking approaches to structure-based design and screening.** *Curr Top Med Chem* 2004, **4**:687-700.
35. Schapira M, Abagyan R, Totrov M: **Nuclear hormone receptor targeted virtual screening.** *J Med Chem* 2003, **46**:3045-3059.
36. Vieth M, Sutherland JJ: **Dependence of molecular properties on proteomic family for marketed oral drugs.** *J Med Chem* 2006, **49**:3451-3453.
37. Morphy R: **The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds.** *J Med Chem* 2006, **49**:2969-2978.
38. Patchett A, Nargund RP: **Privileged structures: an update.** *Annu Rep Med Chem* 2000, **35**:289-298.

39. Schnur DM, Hermsmeier MA, Tebben AJ: **Are target-family-privileged substructures truly privileged?** *J Med Chem* 2006, **49**:2000-2009.
40. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem* 1996, **39**:2887-2893.
41. Hajduk PJ, Bures M, Praestgaard J, Fesik SW: **Privileged molecules for protein binding identified from NMR-based screening.** *J Med Chem* 2000, **43**:3443-3447.
42. Hambly K, Danzer J, Muskal S, Debe DA: **Interrogating the druggable genome with structural informatics.** *Mol Divers* 2006, **10**:273-281.
43. Erbe DV, Gartrell K, Zhang Y-L, Suri V, Kiricich SJ, Will S, Perreault M, Wang S, Tobin JF: **Molecular activation of PPAR $\gamma$  by angiotensin II type 1-receptor antagonists.** *Vascular Pharmacol* 2006, **45**:154-162.
44. Hettne K, Cases M, Boyer S, Mestres J: **Connecting small molecules to nuclear receptor pathways.** *Curr Top Med Chem* 2007, **7**:1530-1536.

# Coverage and bias in chemical library design

Elisabet Gregori-Puigjané and Jordi Mestres\*

*Chemotargets S.L. and Chemogenomics Laboratory, GRIB, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain*

## Summary

The design of chemical libraries directed to target classes is an activity that requires the availability of ligand pharmacological data and/or protein structural data. Based on the knowledge derived from these data, chemical libraries directed mainly to G protein-coupled receptors, kinases, proteases, and nuclear receptors have been assembled. However, current design strategies widely overlook assessing the potential ability of the compounds contained in a focused library to provide uniform ample coverage of the protein family they intend to target. Here we discuss the use of *in silico* target profiling methods as a means to estimate the actual scope of chemical libraries to probe entire protein families and illustrate its applicability in optimizing the composition of compound sets to achieve maximum coverage of the family with minimum bias to particular targets.

---

\* Corresponding author: [jmestres@imim.es](mailto:jmestres@imim.es)

## Introduction

The implementation of high-throughput screening (HTS) within the drug discovery process has increased the traditional capacity for testing compounds to levels sufficient to expect that this technological breakthrough would develop into an endless source of new chemical entities for the pharmaceutical industry. Within this scenario, the size of the screening compound library became one of the key aspects for having a successful HTS campaign. Accordingly, wide compound synthesis and acquisition activities were initiated, mainly aiming at compiling a large, maximally diverse, corporate screening collection [1].

Unfortunately, this early technology-driven diversity-focussed HTS phase showed weaker performance than originally anticipated. In some cases, the progress of hits was limited, as many screening collections were assembled with compounds having poor drug-likeness and the presence of reactive functionalities. In other cases, the number of hits itself was low, as chemical areas relevant for the targets being screened were not properly covered. It then became clear that augmenting the capacity for testing alone was not sufficient for delivering high-quality leads and that more effort was required to carefully balance the composition of diversity-oriented libraries with drug-like non-reactive compounds containing features compatible with the nature of the targets or target classes of corporate interest [2].

In recent years, drug discovery research has been organized gradually around target classes to maximize the efficiency of internal chemistry and biology resources. The adoption of this new chemogenomics paradigm has emphasized more strongly the need for assembling chemical libraries directed to entire protein families [3]. These targeted libraries are expected to contain a diverse set of small molecules, which, as a whole, have a potential ability to probe as many protein members of the family as possible. Therefore, an optimal design of targeted libraries should take into consideration both chemical and target diversity. However, while diversity of chemical space has been investigated thoroughly and is now commonly incorporated in current library designs, diversity within the target space has yet to be properly addressed [4]. In this respect, a wave of recently developed *in silico* target profiling methods is offering a means for assessing the degree of target diversity in chemical libraries, an aspect that is expected to influence the design of the next generation of targeted libraries [5].

## Current design of targeted chemical libraries

The design of targeted chemical libraries is an activity that requires the availability of prior knowledge for the different members of the target family of interest. Recent efforts in collecting, storing, and organizing data on the pharmacology of ligands and on the structure of proteins are facilitating the generation of knowledge on target classes [6,7]. Ultimately, the type and amount of information available for a given protein family will determine whether ligand-based and/or structure-based approaches can be applied when designing targeted libraries. Ideally, enough

pharmacological and structural data on the protein family should be available to allow both methods to be used in a complementary manner. However, despite the fact that the number of protein structures being solved and made available in the Protein Data Bank continues to increase at a remarkable pace, the application of structure-based methods to targeted chemical library design has often been limited by the relatively low functional coverage by experimentally-determined structures within protein families [8]. In addition, this wealth of structural information, with nearly 50,000 entries being currently accessible, is unevenly distributed among the protein members of the main target classes of therapeutic interest. Thus, while enzymes and nuclear receptors have at present a functional coverage of 37% and 51%, respectively, G protein-coupled receptors (GPCRs) are almost devoid of structural information [9]. To complement current low coverage levels by structures within certain target classes, homology modelling techniques offer a matured means to construct computationally-derived structural models [10]. Accordingly, the increase in both numbers and coverage is contributing to make structure-based methods progressively more applicable to targeted compound selection and library design, particularly after the reporting of the first X-ray crystal structure of a human aminergic GPCR [11].

Ligand-based methods face, in contrast, a completely different situation. The vast amount of pharmacological data available for molecules on numerous targets makes them applicable to almost every protein family of therapeutic relevance. One way these data can be exploited is in extracting knowledge about the property differences observed in bioactive molecules depending on the protein family being targeted [12,13]. For example, marketed oral drugs targeting ion channels have been found to be significantly smaller than those targeting proteases, reflected by mean molecular weight values of 306 and 431, respectively, whereas drugs targeting nuclear receptors appear to be significantly more hydrophobic than drugs targeting proteases, with mean counts of oxygen and nitrogen atoms of 3.8 and 7.2, respectively [12]. Similar trends were also found when analysing sets of hit-to-lead ligand pairs instead of drugs [13]. These trends can now be used as simple descriptor-based guidelines to assess or bias the composition of chemical libraries designed to target a given protein family.

Beyond mere molecular property distributions, other ligand-based methods, in combination with a variety of classification schemes, such as recursive partitioning, Bayesian statistics, neural networks or machine-learning methods, have been used to investigate different types of two- and three-dimensional molecular descriptors, features, substructures and pharmacophores for the selection of compounds similar to a set of bioactive reference ligands associated with members of a certain protein target class [14]. Among those, methods aimed at identifying chemical moieties commonly appearing in bioactive ligands have attracted particular attention due to the ease of translation of these privileged structural motifs into compound-library synthesis. However, despite the many evidences of chemical substructures occurring frequently in ligands bioactive across a diverse panel of proteins [15-18], the true existence of target-family selective privileged substructures continues to be a matter of debate [19].

Current design of chemical libraries directed to target classes is focused mainly on GPCRs, kinases, and nuclear receptors [20]. Various strategies have been applied to designing GPCR libraries [21-24], mostly based on ligand information captured in the form of molecular descriptors, pharmacophores, and substructures extracted from a reference set of active compounds among different receptors [25-27]. In contrast, the significant amount of structural information presently available for kinases [28,29] makes structure-based approaches to compound selection and library design [30,31] as applicable to this protein family as ligand-based methods [31-35]. A similar situation is encountered for nuclear receptors, with structure-based and ligand-based information being equally exploited [36,37]. Remarkably, none of the strategies presented above addresses the need to assess the degree of coverage and bias across the protein family offered by the final selection of compounds.

## **Novel approaches to *in silico* target profiling**

Several ligand-based and structure-based approaches to estimate the profile of molecules across a large number of targets have emerged lately [38]. Among those, ligand-based methods have received far more attention than structure-based methods due to their wide applicability to all target classes, but also to their high computational efficiency. All ligand-based target profiling methods developed [39-45] share three common requirements, namely, the availability of a reference database of molecules bioactive with respect to known protein targets, the choice of mathematical descriptors representing the structural features of molecules, and the use of a metric to evaluate the similarity between a target-orphan molecule and all reference target-annotated molecules that will ultimately be used as a score to assign target annotations to molecules. Of mention is the fact that, despite the differences in reference databases, molecular descriptors, and similarity metrics, all methods perform comparably in forecasting the correct target for ligands [40-43].

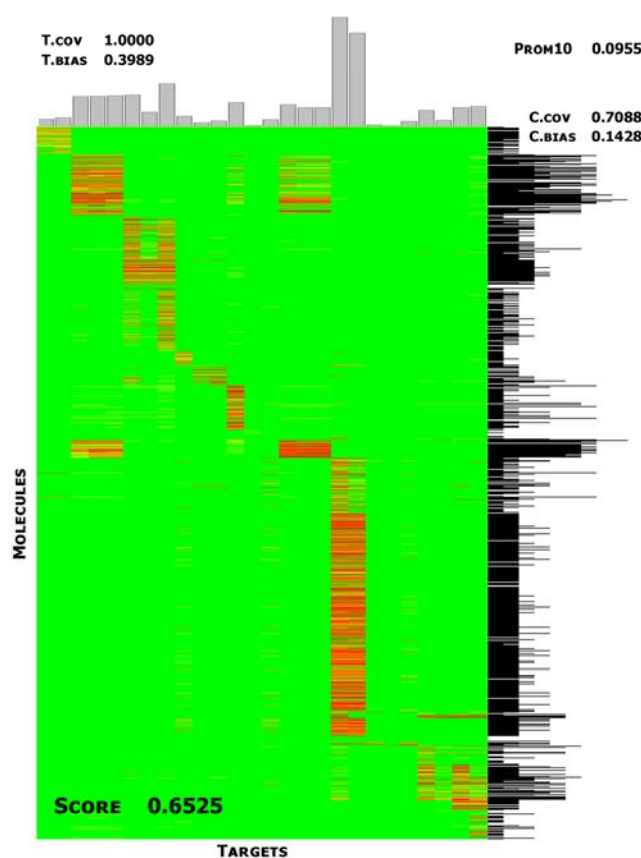
Structure-based target profiling methods have gained interest lately as a result of the considerable advancements made in the functional coverage by the structures of proteins within specific families. Accordingly, it is now possible to generate pharmacophore models from protein cavities and use them to profile compounds against multiple proteins [46], dock a single molecule against thousands of binding sites extracted from the Protein Data Bank [47] and even take advantage of a public web server that allows for the automatic screening of small molecules over a target database of 698 protein structures covering 15 therapeutic areas [48]. At a protein family level, structure-based target profiling has been mainly applied to nuclear receptors and kinases, but also to GPCRs. For nuclear receptors, systematic virtual screening of a library consisting of 5000 random compounds and 78 known active ligands against 19 different protein structures representative of 10 members of this family revealed that it is possible to identify the correct nuclear receptor for a particular active ligand [36]. For kinases, rapid computation of the relative affinity of inhibitors to individual members of this family showed that, on a set of five known kinase inhibitors, the approach is able to identify the correct native kinase target as well as reproduce the



experimental trends in binding affinities [49]. Finally, for GPCRs, it is shown that exploring the boundaries of structure-based methods through protein structure modelling offers, in the lack of experimentally determined structural data, a decent means for estimating the selectivity profile of compounds over a panel of 277 receptors [50]. Given the increasing applicability and overall performance of these *in silico* target profiling methods, projecting the expected target diversity of a compound collection is just one step away.

## Assessing target diversity

The ability to estimate the pharmacological profile of compounds over the members of a given protein family now offers the possibility to account for target diversity when designing chemical libraries directed to protein families. Here, target diversity will be assessed in terms of the degree of coverage and bias within the chemical and protein spaces. To illustrate the different concepts, the estimated ligand-target interaction matrix of a chemical library composed of 2,033 compounds over a set of 26 nuclear receptors is presented in **Figure 1** [43]. The colour gradation used in the heatmap reflects the Euclidean distance of SHED descriptors [51] between each compound in the library (in rows) and the closest reference compound annotated as bioactive to a given target (in columns), with values close to 0.0 being red and turning then into orange, yellow, and green as values approach a predefined distance threshold [43].

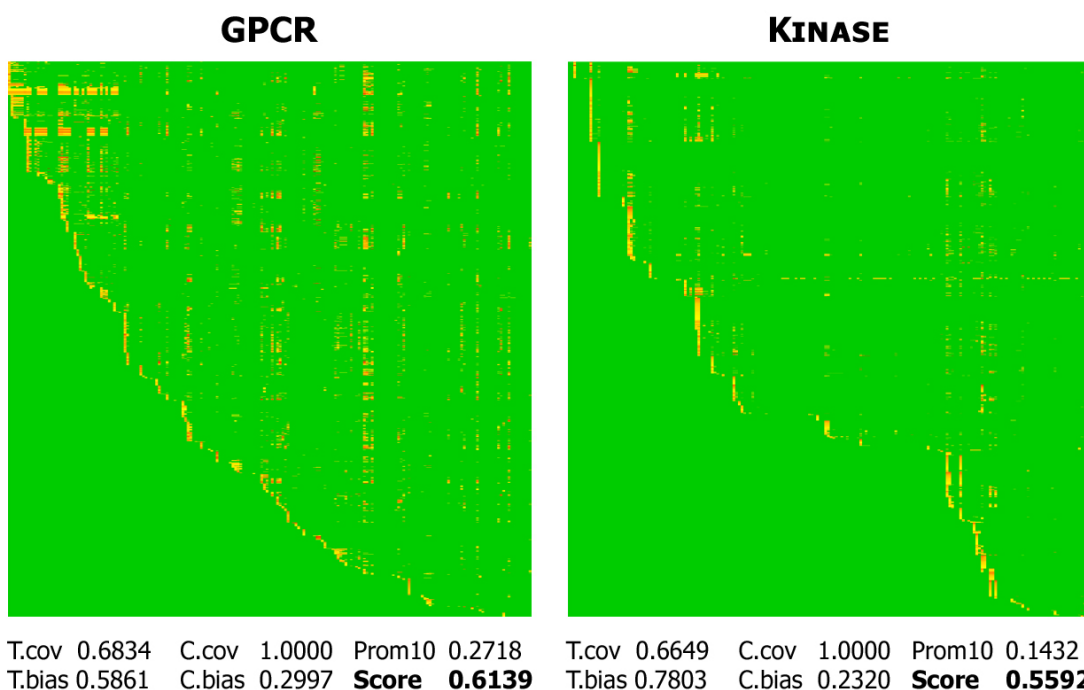


**Figure 1.** Chemical and target coverage (C.cov and T.cov), bias (C.bias and T.bias) and promiscuity (Prom10) to derive an overall diversity Score from an estimated ligand-target interaction matrix (see text for details)

The expected target coverage and bias by the compounds in the library is extracted from the distribution of the number of compounds annotated to each target, depicted on top of the heatmap. Correspondingly, the expected chemical coverage and bias by the targets of the family is derived from the distribution of the number of targets annotated to each molecule, depicted on the right side of the heatmap. Coverage indicates the proportion of targets/ligands with at least one ligand/target annotation (T.cov/C.cov). This particular chemical library offers full target coverage (T.cov = 1.0000) and acceptably high chemical coverage (C.cov = 0.7088), with only 29.12% of the compounds not being annotated to any nuclear receptor under the approach used. In contrast, bias reflects the deviation from uniformity of the distribution of ligand annotations to targets (T.bias) and target annotations to ligands (C.bias), calculated using Shannon entropy [43,51]. As can be observed, this is not a particularly biased library, with target bias (T.bias = 0.3989) being higher than chemical bias (C.bias = 0.1428) as a consequence of the large number of ligand annotations condensed in two individual nuclear receptors. An additional term provides a measure of the average degree of target promiscuity expected from the compounds in the library. In particular, the promiscuity parameter used here focuses on the ten closest annotations to targets (Prom10). Prom10 is bounded between 0 and 1, with values close to 0 reflecting the limit situation of all compounds being selective and as values tend to 1 the selection of compounds becomes systematically more populated with promiscuous compounds over at least 10 targets. This promiscuity parameter is conceptually equivalent to the Gini coefficient suggested recently to assess the selectivity of kinase inhibitors against a panel of kinases [52], the difference lying in the use of the cumulative fraction of total similarity instead of the cumulative fraction of total inhibition. On this basis, an optimal chemical library design to target an entire protein family should exhibit maximum target and chemical coverage, minimum target and chemical bias, and maximum mean promiscuity. Accordingly, a final library score considering both chemical and target diversity within the ligand-target interaction matrix could take the following form:

$$\text{Score} = (\text{T.cov} + \text{C.cov} + (1 - \text{T.bias}) + (1 - \text{C.bias}) + \text{Prom10}) / 5.$$

To put the concepts to work in a real case scenario of compound selection from a particular library, we took two commercially available, yet publicly accessible, chemical libraries designed to target GPCRs [53] and kinases [54] containing 19,533 and 31,882 compounds, respectively. The aim of the exercise was to select an optimal subset of those libraries, containing 10% of the original compounds, according to the diversity score defined above. In addition to the five parameters included in the diversity score, three additional constraints were imposed: first, all compounds containing reactive functionalities will be discarded directly [55]; second, to ensure novelty with respect to prior knowledge of bioactive ligands, compounds selected should not contain a scaffold present in any of the reference bioactive ligands; and third, to avoid potential internal redundancy from having atom variations with the exact same pharmacophoric features, all pairwise Euclidean distances between any two selected compounds should be larger than 0.05. The results of this exercise are presented in **Figure 2**.



**Figure 2.** Examples of compound selections targeted to GPCRs and kinases with optimal coverage, bias, and promiscuity

Using ligand-based models of proteins [43], the original targeted chemical libraries were profiled against 199 GPCRs and 194 kinases, respectively. Starting from an initial set of random selections, a genetic algorithm can be then used to optimize the composition of the selections using the diversity score as fitness function. Interestingly, similar coverage values are obtained for the two selections: on one hand, their chemical composition covers approximately two thirds of all respective protein targets considered and, on the other hand, they both exhibit full chemical coverage, as all compounds contained in the final selections have at least one target annotation. Also, comparable chemical biases are obtained in both libraries, the value for the GPCR subset being slightly higher as a consequence of the level of promiscuity observed among the aminergic GPCRs (top-left region of the heatmap). In contrast, target bias is clearly higher in the kinase subset than in the GPCR selection, reflecting the situation that a small number of kinases concentrates the majority of the compound annotations compared to the more even distribution observed for GPCRs. Finally, the difference in the mean promiscuity value just quantifies the visual perception that the GPCR selection is clearly more promiscuous than the kinase selection. Overall, the higher diversity score obtained for the GPCR subset compared to the kinase subset makes in principle the former selection more suitable than the latter to probe their corresponding protein families.

## Conclusions

Regardless of the strategy used for assembling targeted chemical libraries, it is still uncommon to attempt providing an estimate of the actual coverage and bias that a given selection of compounds is expected to have, as a whole, when tested against an entire target class. This largely overlooked question is crucial to be able to assess objectively the potential scope of a given

chemical library to probe the protein family that was originally intended to target. In this respect, *in silico* target profiling methods provide a means to analyse both chemical and target diversity in terms of the projected pharmacological promiscuity of compound libraries.

As the traditional one drug – one target paradigm is slowly losing acceptance, modern drug discovery is increasingly contemplating the adoption of multitarget strategies for developing the next generation of safer more efficient drugs. Within this context, collecting chemical libraries designed to cover fully and uniformly a panel of proteins is likely to have a strong impact in the identification of novel hits with customised pharmacological profiles for both drugged and orphan targets. Assessment of coverage and bias in chemical libraries targeting multiple proteins may then become an aspect of utmost importance.

## **Acknowledgments**

This research was funded by the Spanish Ministerio de Educación y Ciencia (project reference BIO2005-04171) and the Instituto de Salud Carlos III.

## References and recommended reading

- of special interest
  - of outstanding interest
1. Harper G, Pickett SD, Green DV: **Design of a compound screening collection for use in high-throughput screening.** *Comb Chem High Throughput Screen* 2004, **7**:63-70.
  2. Jacoby E, Schuffenhauer A, Popov M, Azzaoui K, Havii B, Schopfer U, Engeloch C, Stanek J, Acklin P, Rigollier P, Stoll F, Koch G, Meier P, Orain D, Giger R, Hinrichs J, Malagu K, Zimmermann J, Roth HJ: **Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection.** *Curr Top Med Chem* 2005, **5**:397-411.
  3. Miller JL: **Recent developments in focused library design: targeting gene-families.** *Curr Top Med Chem* 2006, **6**:19-29.
  4. Gorse AD: **Diversity in medicinal chemical space.** *Curr Top Med Chem* 2006, **6**:3-18.
  5. Jenwitheesuk E, Horst JA, Rivas KL, van Voorhis WC, Samudrala R: **Novel paradigms for drug discovery: computational multitarget screening.** *Trends Pharmacol Sci* 2008, **29**:62-71.
  6. Balakin KV, Tkachenko SE, Kiselyov AS, Savchuk NP: **Focused chemistry from annotated libraries.** *Drug Discov Today* 2006, **3**:397-403.
  7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
  8. Mestres J: **Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery.** *Drug Discov Today* 2005, **10**:1629-1637.
  9. Garcia-Serna R, Opatowski L, Mestres J: **FCP: functional coverage of the proteome by structures.** *Bioinformatics* 2006, **22**:1792-1793.
  10. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, *et al.*: **MODBASE: a database of annotated comparative protein structure models and associated resources.** *Nucl Acids Res.* 2006, **34**:D291-295.
  11. Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VRP, Sanishvili R, Fischetti RF, Schertler GFX, Weis WI, Kobilka BK: **Crystal structure of the human  $\beta_2$  adrenergic G-protein-coupled receptor.** *Nature* 2007, **450**:383-387.
  - The availability of the first crystal structure of a human aminergic G protein-coupled receptor will be highly influential in future protein modeling by homology
  12. Vieth M, Sutherland JJ: **Dependence of molecular properties on proteomic family for marketed oral drugs.** *J Med Chem* 2006, **49**:3451-3453.
  13. Morphy R: **The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds.** *J Med Chem* 2006, **49**:2969-2978.
  14. Rognan D: **Chemogenomic approaches to rational drug design.** *Br J Pharmacol* 2007, **152**:38-52.
  15. Wilkens SJ, Janes J, Su AI: **HierS: hierarchical scaffold clustering using topological chemical graphs.** *J Med Chem* 2005, **48**:3182-3193.
  16. Cases M, Garcia-Serna R, Hettne K, Weeber M, Lei JV, Boyer S, Mestres J: **Chemical and Biological Profiling of an Annotated Compound Library Directed to the Nuclear Receptor Family.** *Curr Top Med Chem* 2005, **5**:763-772.

17. Yan SF, King FJ, He Y, Caldwell JS, Zhou Y: **Learning from the Data: Mining of Large High-Throughput Screening Databases.** *J Chem Inf Model* 2006, **46**:2381-2395.
18. Müller G: **Medicinal chemistry of target family-directed masterkeys.** *Drug Discov Today* 2003, **8**:681-691.
19. Schnur DM, Hermsmeier MA, Tebben AJ: **Are Target-Family-Privileged Substructures Truly Privileged?** *J Med Chem* 2006, **49**:2000-2009.
  - An interesting substructural analysis of ligands associated with several target classes that adds to the debate on the true existence of privileged substructures within target families
20. Shuttleworth SJ, Connors RV, Fu J, Liu J, Lizarzaburu ME, Qiu W, Sharma R, Wanska M, Zhang AJ: **Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization.** *Curr Med Chem* 2005, **12**:1239-1281.
21. Klabunde T, Hessler G: **Drug design strategies for targeting G-protein-coupled receptors.** *ChemBioChem* 2002, **3**:928-944.
22. Jimonet P, Jäger R: **Strategies for designing GPCR-focused libraries and screening sets.** *Curr Opin Drug Discov Devel* 2004, **7**:325-333.
23. Lowrie JF, Delisle RK, Hobbs DW, Diller DJ: **The different strategies for designing GPCR and kinase targeted libraries.** *Comb Chem High Throughput Screen* 2004, **7**:495-510.
  - This review covers illustrative cases of both ligand-based and structure-based approaches to designing targeted chemical libraries
24. Crossley R: **The design of screening libraries targeted at G-protein coupled receptors.** *Curr Top Med Chem* 2004, **4**:581-589.
25. Balakin KV, Lang SA, Skorenko AV, Tkachenko SE, Ivanshchenko AA, Savchuk NP: **Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries.** *J Chem Inf Comput Sci* 2003, **43**:1553-1562.
26. Lamb ML, Bradley EK, Beaton G, Bondy SS, Castellino AJ, Gibbons PA, Suto MJ, Grootenhuis PDJ: **Design of a gene family screening library targeting G-protein coupled receptors.** *J Mol Graph Model* 2004, **23**:15-21.
27. Bywater RP: **Privileged structures in GPCRs.** *Ernst Schering Found Symp Proc* 2006, **2**:75-91.
28. Noble MEM, Endicott JA, Johnson LN: **Protein kinase inhibitors: insights into drug design from structure.** *Science* 2004, **303**:1800-1805.
29. Marsden BD, Knapp S: **Doing more than just the structure-structural genomics in kinase drug discovery.** *Curr Opin Chem Biol* 2008, doi:10.1016/j.cbpa.2008.01.042.
  - An up-to-date analysis of the current structural coverage of kinases
30. Todorov NP, Buenemann CL, Alberts IL: **Combinatorial ligand design targeted at protein families.** *J Chem Inf Model* 2005, **45**:314-320.
31. Sun D, Chuaqui C, Deng Z, Bowes S, Chin D, Singh J, Cullen P, Hankins G, Lee WC, Donnelly J, Friedman J, Josiah S: **A kinase-focused compound collection: compilation and screening strategy.** *Chem Biol Drug Des* 2006, **67**: 385-394.
32. Xia X, Maliski EG, Gallant P, Rogers D: **Classification of kinase inhibitors using a Bayesian model.** *J Med Chem* 2004, **47**:4463-4470.
33. Prien O: **Target-family-oriented focused libraries for kinases - conceptual design aspects and commercial availability.** *Chembiochem* 2005, **6**:500-505.
34. Briem H, Günther J: **Classifying "kinase inhibitor-likeness" by using machine-learning methods.** *Chembiochem* 2005, **6**:558-566.

35. Aronov AM, McClain B, Moody CS, Murcko MA: **Kinase-likeness and kinase-privileged fragments: toward virtual polypharmacology.** *J Med Chem* 2008, **51**:1214-1222.
36. Schapira M, Abagyan R, Totrov M: **Nuclear hormone receptor targeted virtual screening.** *J Med Chem* 2003, **46**:3045-3059.
37. Stewart EL, Brown PJ, Bentley JA, Willson TM: **Selection, application, and validation of a set of molecular descriptors for nuclear receptor ligands.** *Comb Chem High Throughput Screen* 2004, **7**:407-412.
38. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling.** *Br J Pharmacol* 2007, **152**:9-20.
39. Cleves AE, Jain AN: **Robust ligand-based modeling of the biological targets of known drugs.** *J Med Chem* 2006, **49**:2921-2938.
40. Nidhi, Glick M, Davies JW, Jenkins JL: **Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases.** *J Chem Inf Model* 2006, **46**:1124-1133.
41. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space.** *Nature Biotech* 2006, **24**:805-815.
- This study analyses the ligand-target interaction matrix derived from several annotated chemical libraries to explore the relationship between chemical and biological spaces
42. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M: **Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors.** *J. Med. Chem.* 2006, **49**:6802-6810.
43. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling.** *J Chem Inf Model* 2006, **46**:2725-2736.
44. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry.** *Nature Biotech* 2007, **25**:197-206.
45. Bender A, Young DW, Jenkins JL, Serrano M, Mikhailov D, Clemons PA, Davies JW: **Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprints.** *Comb Chem High Throughput Screen* 2007, **10**:719-731.
46. Steindl TM, Schuster D, Laggner C, Langer T: **Parallel screening: a novel concept in pharmacophore modeling and virtual screening.** *J Chem Inf Model* 2006, **46**:2146-2157.
47. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D: **sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank.** *J Chem Inf Model* 2006, **46**:717-727.
48. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang X, Jiang H: **TarFisDock: a web server for identifying drug targets with docking approach.** *Nucl Acids Res* 2006, **34**:W219-224.
49. Rockey WM, Elcock AH: **Rapid computational identification of the targets of protein kinase inhibitors.** *J Med Chem* 2005, **48**:4138-4152.
50. Bissantz C, Logean A, Rognan D: **High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening.** *J Chem Inf Comput Sci* 2004, **44**:1162-1176.
- This work represents one of the best efforts of the use of structure-based methods to design libraries for the family of GPCRs, traditionally addressed with ligand-based methods
51. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.

52. Graczyk PP: **Gini Coefficient: A New Way To Express Selectivity of Kinase Inhibitors against a Family of Kinases.** *J Med Chem* 2007, **50**:5773-5779.
53. GPCR library, January 2008. Enamine Ltd. [www.enamine.net](http://www.enamine.net)
54. Kinase library, October 2007. Life Chemicals. [www.lifechemicals.com](http://www.lifechemicals.com)
55. Oprea TI: **Property distribution of drug-related chemical databases.** *J Com-Aided Mol Design* 2000, **14**:251-264.



## Chapter III.4 – Network pharmacology

In this chapter, the methodology presented previously is used to build and validate ligand-based models for 684 targets. These models are used in the *in silico* target profiling of 767 drugs and the results analysed in to detect possible cross-pharmacologies among different target classes. A recent work by Yildirim *et al.* [140] highlighted the network behaviour of the relationships among drugs and protein targets, rather than the one drug – one target expected relationship. This conclusion being highly relevant, we here discuss the effect of data completeness in the following conclusions extracted on the nature of these networks.

Papers included in this chapter:

- Gregori-Puigjané E, Mestres J: **A ligand-based approach to mining the chemogenomic space of drugs.** *Comb Chem High Throughput Screen* 2008 (In press).
- Mestres J, Gregori-Puigjané E, Valverde S, Solé RV: **The effect of data completeness on drug-target interaction networks.** *Nature Biotech* 2008 (In press).



# A ligand-based approach to mining the chemogenomic space of drugs

Elisabet Gregori-Puigjané and Jordi Mestres\*

*Chemotargets SL and Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Parc de Recerca Biomèdica, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain*

**Abstract:** The practical implementation and validation of a ligand-based approach to mining the chemogenomic space of drugs is presented and applied to the *in silico* target profiling of 767 drugs against 684 targets of therapeutic relevance. The results reveal that drugs targeting aminergic G protein-coupled receptors (GPCRs) show the most promiscuous pharmacological profiles. The detection of cross-pharmacologies between aminergic GPCRs and the opioid, sigma, NMDA, and 5-HT<sub>3</sub> receptors aggravate the potential promiscuity of those drugs, predominantly including analgesics, antidepressants, and antipsychotics.

**Keywords:** chemogenomics, off-target profiling, drug repurposing, network pharmacology

---

\* Corresponding author. Email: jmestres@imim.es

## Introduction

In recent years, the capacity for synthesising and testing compounds has increased dramatically and screening campaigns of thousands of compounds against a single target are nowadays realistic, both within industrial and academic settings [1]. The total amount of unique compounds available for purchase from chemical providers is estimated at present to be close to eight million [2]. Therefore, with current screening capacities, experimental testing of the entire chemical space synthesised today against a particular target would still be feasible. However, current drug discovery is moving away from the traditional one drug – one target paradigm and focusing more towards identifying molecules that modulate multiple targets simultaneously [3]. Adding a biological dimension to high-throughput screening means that, even with the impressive technological advances made, current capacities are no longer sufficient to tackle the experimental testing of thousands of compounds against hundreds of targets.

Within this new scenario, *in silico* target profiling methods are emerging as efficient alternatives to the currently unaffordable high-throughput *in vitro* target profiling of compounds [4]. All these methods capitalize on the vast amount of prior knowledge available for many targets of therapeutic relevance, either by exploiting all bioactive ligands stored in annotated chemical libraries [5] or by using experimentally-determined crystal structures deposited in the Protein Data Bank [6]. Among some of the recently reported ligand-based methods, Nidhi *et al.* [7] exploited the contents of the WOMBAT annotated chemical library [8] to derive a multiple-category Laplacian-modified naïve Bayesian model that predicted the protein targets for all compounds in the MDDR database with a success rate of 77%, and Mestres *et al.* [9] used an in-house constructed library of 2033 molecules annotated to 25 nuclear receptors [10] to derive a similarity-based model that was able to recover the original annotations with a success rate of 71% and provide estimates of potential off-target affinities to nuclear receptors for several drugs. Among the most recent target-based methods, Rockey and Elcock [11] reported a method for rapidly computing the relative affinity of inhibitors to individual members of the kinase family, Kellenberger *et al.* [12] created a database of 6415 binding sites (sc-PDB) that was used to identify the native target of four unrelated ligands among the top 1% of scored binding sites, and Li *et al.* [13] have recently created a web server for inverse docking that allows for automatically screen small molecules over a panel of 698 protein structures covering 15 therapeutic areas. All these reports provide ample evidence for the possibilities of *in silico* target profiling methods in multitarget drug discovery.

One of the most promising possibilities is in the area of finding new therapeutic uses for approved drugs, an activity often referred to as drug repurposing [14]. Drugs have been traditionally designed to interact with a primary target known to be relevant to the particular disease of corporate interest. During the drug optimisation process, very limited scope was often given to address properly the issue of selectivity, by considering only a handful of additional targets phylogenetically related to the primary target. The ability of *in silico* target profiling methods to identify new targets for old drugs, as demonstrated recently by Keiser *et al.* [15], has direct

implications for using immediately off-patent products in clinical trials but also for alerting of potential secondary effects due to residual affinities for undesired targets [16]. Accordingly, the aim of this work is to probe the chemogenomic space of drugs by profiling *in silico* a set of 767 drugs against a panel of 684 targets representing all therapeutically-relevant protein families.

## Ligand-based prior knowledge

As stated above, *in silico* target profiling methods rely on the availability of prior knowledge on bioactive ligands and protein structures. In this contribution, focus will be given to a ligand-based method that currently exploits pharmacological data stored in three different annotated chemical libraries. The main source of information is the WOMBAT database [8], currently containing 163,102 unique molecules with activity data (125,113 IC<sub>50</sub> values and 91,693 K<sub>i</sub> values) on 677 protein targets collected from medicinal chemistry journals over the last 30 years. WOMBAT is complemented with BindingDB [17], a public database of 12,394 small molecules with activity data (24,772 experimentally determined binding affinities) for 110 protein targets. And these two databases are finally completed with an in-house constructed chemical library composed of 2,718 ligands annotated to 27 nuclear receptors [10]. Integrating all the data from these three different sources and filtering out all compounds having an activity value (K<sub>i</sub>, IC<sub>50</sub>, and/or EC<sub>50</sub>) above 10 μM resulted in a total number of 109,766 unique compounds annotated to 684 targets. e

Class	Number of Targets	Number of Ligands
Proteases (pEC)	74	16,263
Kinases (kEC)	45	9,432
Cytochrome P450 (cEC)	24	1,199
Other Enzymes (oEC)	268	27,392
Aminergic GPCRs (aGR)	40	14,390
Peptidic GPCRs (pGR)	68	17,965
Other GPCRs (oGR)	60	10,505
Ion Channels (IC)	48	5,579
Nuclear Receptors (NR)	30	4,825
Transporters (TC)	15	3,067
Integrins (IN)	8	2,087
Catalytic Receptors (CR)	4	1,232

Table 1. The list of 12 target classes and their respective target size and ligand coverage

Table 1 presents the distribution of these 684 targets among 12 different target classes contained within 7 protein families, namely, enzymes, G protein-coupled receptors (GPCRs), ion channels, nuclear receptors, transporters, integrins, and catalytic receptors. The chemical space

covering the family of enzymes contains 53,702 small molecules annotated to 411 enzymes that, in turn, are subdivided into 4 target classes, namely, proteases, kinases, cytochromes, and the rest, to reflect some of the enzyme subfamilies of main therapeutic relevance. In terms of chemical space coverage, GPCRs come next with 42,064 ligands annotated to 168 receptors which, in turn, are subdivided into 3 target classes, namely, aminergic, peptidic, and the rest, to reflect again some of the therapeutically most relevant GPCR subfamilies. The rest of the protein families are comparably smaller, both in terms of number of targets and chemical coverage. In essence, this will be the body of prior knowledge that will be subsequently used to derive the ligand-based protein models necessary to perform an *in silico* target profiling.

Since the ability to extract knowledge from annotated chemical libraries will be highly determined by the way chemical and biological data are stored, special emphasis was put in storing both chemical structures and protein targets using appropriate unique identifiers and classification schemes [18]. For chemical structures, we used an in-house proposed Chemical Structure Code (CSC) purely based on topological features of molecules. Accordingly, each molecule was identified with a unique hierarchical five-level CSC [9]. The first and second levels are integers specifying, respectively, the number of rings in the largest ring system present in the molecule and the total number of ring systems in the molecule. The third, fourth, and fifth levels are a unique seven-character hash codes for the molecular framework, scaffold, and the complete molecular structure, respectively. For proteins, standard classification schemes were adopted directly when available or derived instead following the phylogenetic relationships among the members of the different protein families. For example, the Enzyme Commission number [19] was used for enzymes and the proposal from the Nuclear Receptor Nomenclature Committee [20] was used for nuclear receptors, whereas the classification scheme for GPCRs had to be derived from its internal phylogeny [21].

## Ligand-based protein models

The set of bioactive ligands collected for a given target provides in fact a complementary description of the target from a ligand viewpoint. In order to be able to process this information efficiently, molecular structures need to be encoded using some sort of mathematical descriptors. In this respect, we use a novel set of low-dimension molecular descriptors called SHED [22]. SHED are derived from distributions of atom-centred feature pairs extracted directly from the topology of molecules. Initially, each atom in a molecule is mapped to a Sybyl atom type and then assigned to one or more of four atom-centred features, namely, hydrophobic (H), aromatic (R), acceptor (A), and donor (D). For example, an aliphatic C.3 carbon will be assigned to a hydrophobic feature (H), whereas a protonated N.4h nitrogen will be assigned to both aromatic and donor features (R,D). Then, the shortest path length between atom-centred feature pairs is derived and its occurrence at different path lengths stored to create a feature-pair distribution. At this stage, Shannon entropies are used to quantify the degrees of occupancy and uniformity of each one of the ten distributions resulting from all pair combinations of the four features. In the end, each chemical structure is

ultimately represented by a vector, referred to as its SHED profile, composed of 10 real numbers reflecting the particular feature-pair distributions present in the molecule. The ensemble of SHED profiles representing all molecules annotated to each particular target constitutes a mathematical description of the target from a ligand perspective.

Ligand-based descriptor models were derived for each one of the 684 targets with information on bioactive ligands. Prior to validating the sensitivity and specificity of the models, the criteria followed for annotating a query molecule to a given target needs to be defined. Therefore, the first step is to calculate the Euclidean distance between the SHED profile of the query molecule and all the target-related SHED profiles. The probability of a compound being active against that target is assumed to be related to the degree of similarity relative to the set of known bioactive ligands. Accordingly, the scoring of each compound in a chemical library with respect to a given target is assigned to the minimum value of all Euclidean distances. In the context of similarity-based virtual screening, the approach of combining the scores over multiple bioactive reference molecules has been recently referred to as group fusion and proven to give significantly superior results to using data fusion strategies on single reference molecules for a wide variety of protein targets [23]. If the minimum Euclidean distance is below a certain value, the molecule is annotated to that target. In this work, following a previous validation analysis [9], the annotation threshold was set to a minimum Euclidean distance value of 0.6. This strategy follows on recent studies suggesting that similarity to molecules in the reference set is a good criteria for prediction accuracy of external test sets [24].

	pEC	kEC	cEC	oEC	aGR	pGR	oGR	NR	IC	TC	IN	CR
pEC	<b>70.0%</b>	7,4%	1,8%	12,5%	1,8%	7,0%	6,6%	1,9%	2,4%	0,5%	0,9%	1,0%
kEC	10,1%	<b>66,9%</b>	3,9%	23,2%	6,9%	9,1%	11,9%	5,4%	7,7%	2,7%	0,5%	9,1%
cEC	18,0%	<b>31,9%</b>	<b>75,8%</b>	<b>59,4%</b>	29,5%	15,4%	29,1%	<b>30,4%</b>	25,6%	10,8%	0,8%	9,7%
oEC	12,6%	14,4%	6,5%	<b>72,0%</b>	8,1%	12,1%	14,4%	8,2%	8,0%	2,6%	0,6%	2,9%
aGR	2,4%	5,8%	4,4%	15,8%	<b>79,4%</b>	23,6%	5,0%	3,6%	13,2%	17,9%	0,1%	0,7%
pGR	8,7%	6,4%	1,6%	14,5%	14,5%	<b>78,0%</b>	9,5%	3,3%	7,6%	4,3%	0,8%	1,3%
oGR	14,5%	14,0%	5,2%	<b>30,8%</b>	8,3%	17,8%	<b>77,6%</b>	9,1%	8,7%	4,2%	0,5%	2,9%
NR	9,4%	15,4%	11,5%	<b>32,8%</b>	10,4%	12,1%	16,5%	<b>83,1%</b>	7,0%	5,4%	0,6%	4,3%
IC	6,7%	14,0%	5,4%	26,6%	17,9%	17,0%	12,4%	5,0%	<b>71,4%</b>	7,1%	0,3%	3,1%
TC	1,8%	8,6%	4,6%	17,9%	<b>48,2%</b>	22,9%	5,9%	5,5%	14,1%	<b>77,9%</b>	0,1%	0,5%
IN	7,4%	2,8%	0,4%	4,9%	0,6%	5,1%	1,2%	0,8%	1,4%	0,1%	<b>68,0%</b>	0,6%
CR	9,7%	<b>45,1%</b>	3,7%	<b>33,3%</b>	5,4%	13,3%	15,6%	6,0%	10,1%	2,2%	0,3%	<b>59,1%</b>

**Table 2.** Confusion matrix of cross-annotations among the 12 target classes. The recall values on the diagonal are highlighted. Also highlighted are the off-diagonal cross-annotations with percentages over 30%

Having clarified the process of target annotation to molecules, the degrees of recall and selectivity of the ligand-based models were assessed at a target class level. Table 2 summarises the confusion matrix resulting from processing all ligand-based target models of a particular target class (in columns) against all bioactive molecules of each target class (in rows). A leave-one-out procedure was enforced to avoid trivial annotations coming from collisions, implying that each

molecule being processed was removed from the respective models. Values on the diagonal reflect the level of target-class recall of each model, that is, the proportion of target-class actives annotated by the target-class model. Off-diagonal values reflect mainly the level of target-class selectivity of each model, that is, the proportion of off-target-class actives annotated by the target-class model.

In general, good recall values are obtained, with percentages over 66% for 11 out of the 12 target classes. Only for the target class of catalytic receptors, with the lowest number of chemical information available, recall falls down to 59%. With respect to selectivity, the ligand-based model for proteases (pEC) is an example of a decent sensitive and selective model, with percentages of annotation to molecules bioactive to target classes other than proteases below 20%. A good selectivity pattern is also observed for the ligand-based model derived for cytochromes (cEC), with all cross-annotation values under 12%. Interestingly, the ligand-based model for kinases (kEC) annotates also 31.9% and 45.1% of all actives to cytochromes and catalytic receptors (CR), although 34.7% of all actives to catalytic receptors were already annotated to kinases originally. In contrast, the ligand-based model for the 268 remaining enzymes (oEC) is clearly the most promiscuous model, with cross-annotation values above 30% for cytochromes, GPCRs other than aminergic and peptidic (oGR), nuclear receptors (NR), and catalytic receptors. Reasonable ligand-based models are also obtained for all three GPCR target classes. For aminergic GPCRs (aGR), the ligand-based model annotates also over 48% of all actives to transporters, an expected cross-annotation signal since 29.4% of them (essentially serotonin transporters) were already annotated to aminergic GPCRs. Also worth stressing is the generally good selectivity profiles obtained for the ligand-based models of both peptidic GPCRs (pGR) and GPCRs other than aminergic and peptidic (oGR). The remaining ligand-based models derived for nuclear receptors, ion channels (IC), transporters (TC), integrins (IN), and catalytic receptors show all low promiscuity levels, the only remark being that the nuclear receptor model annotates also over 30% of all actives to cytochromes. Therefore, overall, the analysis of the confusion matrix indicates that ligand-based protein models can indeed be a promising *in silico* approach to target profiling, with potential applicability in various fields such as targeted library design or drug repurposing.

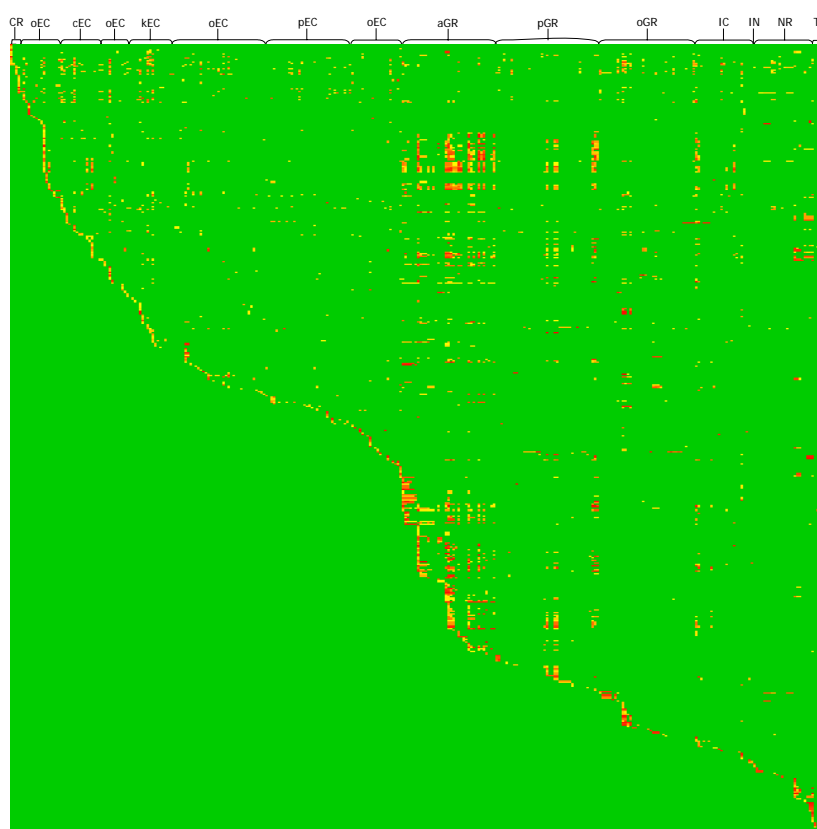
### ***In silico* target profiling**

With validated ligand-based models for 684 targets, we can now attempt estimating *in silico* the pharmacological profile of drugs to illustrate the scope of applicability for drug repurposing. Accordingly, a set of 767 drugs with known affinity for one or more targets was selected. After calculation of their respective SHED profiles, each drug was processed against all 684 ligand-based target models following the procedure described above. In total, we were able to assign at least a target annotation to 592 drugs, corresponding to 77.2% of the total number of drugs processed. Those 592 drugs received 3,728 annotations to 324 targets, which roughly means that on average each drug in this set was annotated to 11 targets. Out of the 3,728 annotations assigned, we were able to reproduce 998 of the original annotations, which represents an



annotation recall of 55.1% (998 out of 1,810). The validity of many of the remaining 2,730 additional target annotations cannot be confirmed at this stage due to the lack of completeness in the activity data and, thus, a wealth of opportunities may be hiding there.

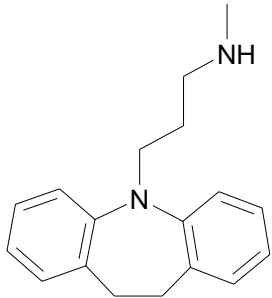
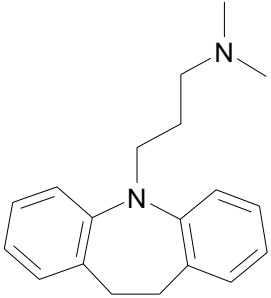
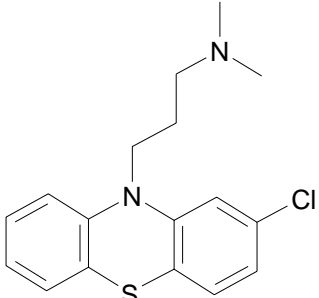
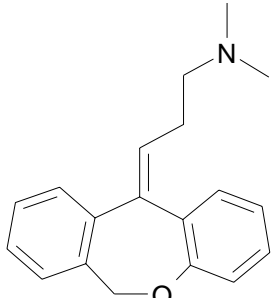
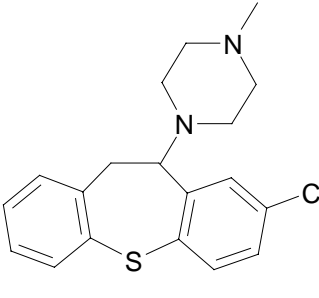
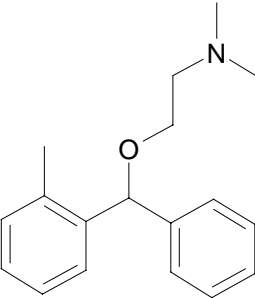
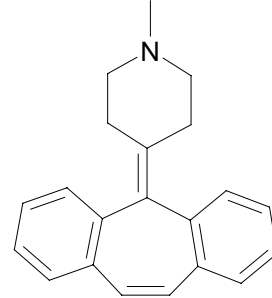
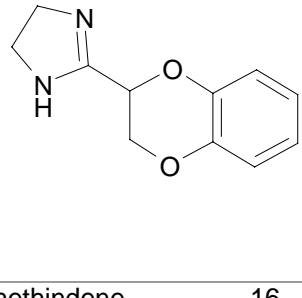
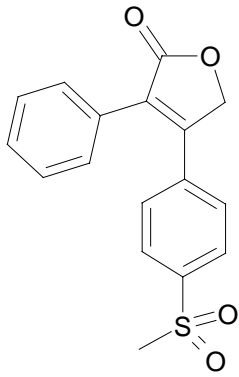
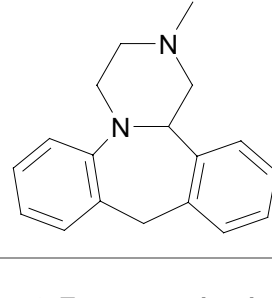
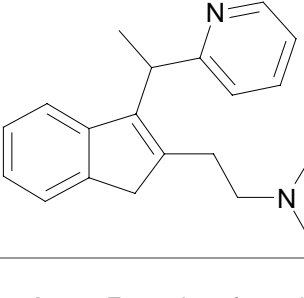
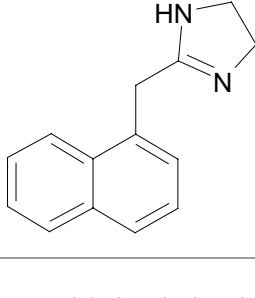
The annotations of each drug (in rows) to each target (in columns) are visually summarised in **Figure 1**. For the sake of clarity, labels have been added in columns to locate the regions covered by the 12 target classes. In the heatmap shown, cells colored within the red-to-yellow spectra mark the drug-target annotations. The colour gradation between red and yellow reflects the value of the minimum SHED Euclidean distance between the SHED profile of each drug and the ensemble of SHED profiles annotated to each target. Accordingly, distance values close to 0.0 are represented in red, those close to 0.3 are seen in orange, and as distance values approach 0.6 they turn into yellow and finally green at values above the annotation threshold.



**Figure 1. The drug-target interaction heatmap.** Annotations between drugs (in rows) and targets (in columns) are denoted by cells colored within the red-to-yellow spectra. Colour gradation reflects distance to the closest bioactive reference: red, orange, and yellow reflect increasing distance values from 0.0 to 0.6; green means distances values above the annotation cut-off. See Table 1 for target class abbreviations.

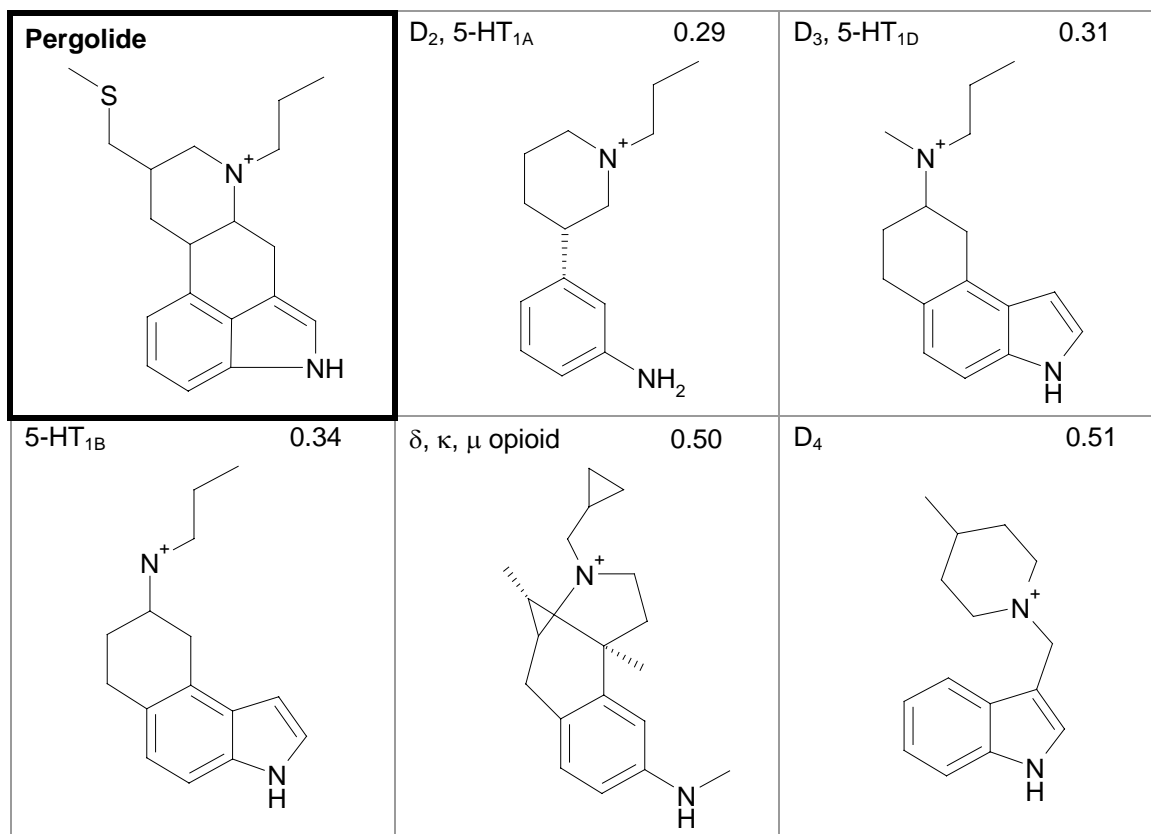
Looking at the heatmap in **Figure 1** from a drug perspective, perhaps the most significant observation is that the drugs that appear to be more promiscuous are by far those being annotated to aminergic GPCRs (aGR). To examine this aspect further, drugs were given a target promiscuity index according to the number of annotations assigned. A selection of the top 100 most promiscuous drugs with indices ranging from 34 to 15 is presented in **Figure 2**. Unfortunately, full assessment of the validity of these results cannot be performed because, as highlighted already

above, activity data available for drugs is not complete. Chlorpromazine is one of the few examples for which we have been able to confirm that it binds to 26 receptors with affinities ( $K_i$ ) below 1  $\mu\text{M}$  [3]. In our *in silico* target profiling exercise, chlorpromazine is found among the most promiscuous drugs (rank #6 of 592) with a promiscuity index of 31. Ranking drugs according to this target promiscuity index could be used as a means to estimate their liability due to residual off-target affinities and thus anticipate potential side effects.

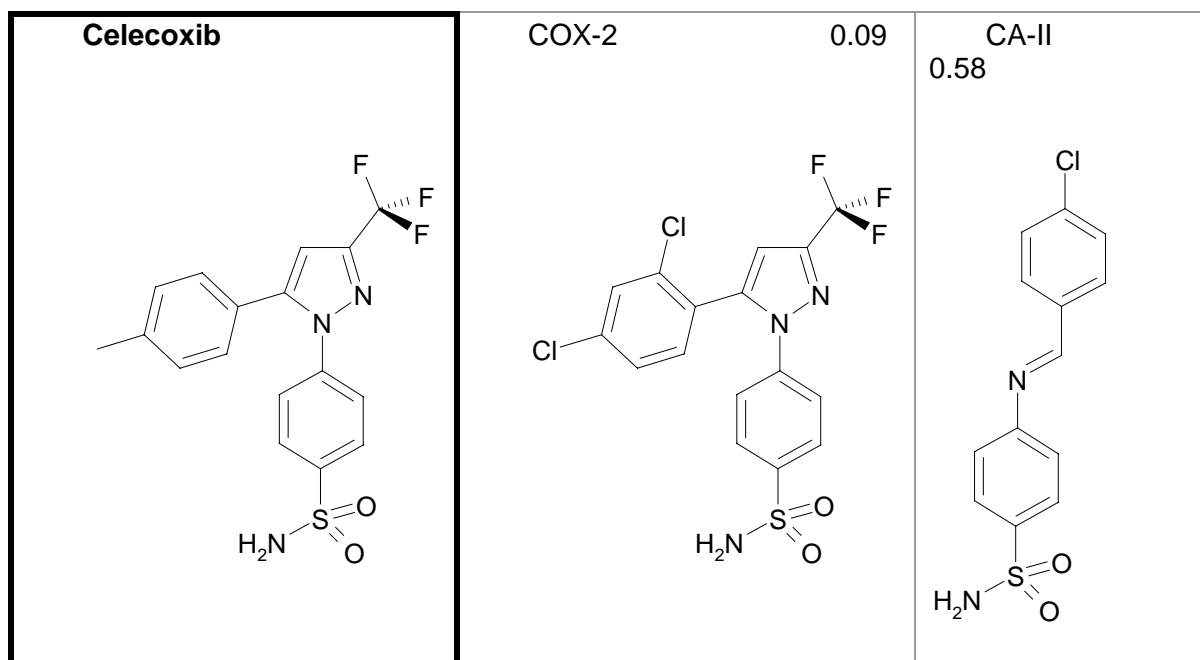
Desipramine 34 	Imipramine 32 	Chlorpromazine 31 
Doxepin 28 	Octochothepin 27 	Orphenadrine 25 
Cyproheptadine 24 	Idazoxan 22 	Rofecoxib 20 
Mianserin 19 	Dimethindene 16 	Naphazoline 15 

**Figure 2. Target promiscuity index for drugs.** Examples of promiscuous drugs and their calculated target promiscuity index.

In this respect, we were able to identify two drugs that can be considered representative of the two possible scenarios. One of them is pergolide, a drug used for the treatment of Parkinson's disease that was recently withdrawn from the market due to latest findings of residual off-target affinity for 5-HT<sub>2B</sub>, a serotonin receptor expressed in cardiac valves, the activation of which is known to be associated with developing drug-induced valvular heart disease [25]. Pergolide was originally designed as a dopaminergic D1/D2 agonist. However, when tested against an extended panel of aminergic GPCRs, it has a rather promiscuous profile with micromolar affinities for D<sub>3</sub>, D<sub>4</sub>, and D<sub>5</sub> dopamine receptors, as well as for the 5-HT<sub>1A</sub>, 5-HT<sub>1B</sub>, 5-HT<sub>1D</sub>, 5-HT<sub>2A</sub>, 5-HT<sub>2B</sub> and 5-HT<sub>2C</sub> serotonin receptors [26]. The results of our *in silico* target profiling of pergolide are summarised in **Figure 3**. Next to the structure of pergolide, the structures of the 5 annotated reference compounds having a SHED Euclidean distance below the annotation cut-off of 0.6 have been collected and their respective literature-extracted annotations shown. For example, the closest compound to pergolide was found at a distance of 0.29 and reported to be active to dopamine D<sub>2</sub> and serotonin 5-HT<sub>1A</sub>. Therefore, these two annotations will be assigned to pergolide and the two of them can be confirmed experimentally. Overall, pergolide received 6 annotations to dopamine D<sub>2</sub>, D<sub>3</sub>, and D<sub>4</sub>, serotonin 5-HT<sub>1A</sub>, 5-HT<sub>1B</sub>, 5-HT<sub>1D</sub> for which experimental evidence could be found [26]. But, in addition to those, 3 plausible annotations to the  $\delta$ ,  $\kappa$ , and  $\mu$  opioids were assigned, for which we were unable to find confirmation in the literature. Regrettably, we were unable to annotate pergolide to 5-HT<sub>2B</sub>. In spite of this, *in silico* target profiling provided a clear signal that pergolide could have residual affinities for at least 3 serotonin receptors. Knowing the intrinsic promiscuity of serotonin receptors, that signal could have triggered an early alert to test pergolide against a panel of serotonin receptors, which could have lead to a much earlier detection of the unwanted 5-HT<sub>2B</sub> agonism. The other drug is celecoxib, a drug used to control the pain and inflammation associated with chronic inflammatory diseases such as rheumatoid- and osteo-arthritis. Celecoxib was originally conceived as a selective COX-2 inhibitor. The results of our *in silico* target profiling of celecoxib are summarised in **Figure 4**. For this drug, only two compounds were found to have a SHED Euclidean distance below the annotation cut-off. The closest one is in fact a structural analogue of celecoxib and is the one responsible for annotating the drug to COX-2. But most interestingly, just under 0.6, a compound active to carbonic anhydrase II (CA-II) is found and, consequently, this annotation is transferred to celecoxib. Providentially, clear evidence could be found in the literature, not only reporting that celecoxib was indeed a nanomolar inhibitor of CA-II but also providing a crystal structure of the interaction of celecoxib in the active site of CA-II [27]. In this case, the CA-II annotation assigned to celecoxib is an example of how *in silico* target profiling can provide new opportunities for old drugs.



**Figure 3. Pergolide's in silico pharmacology profile.** Set of reference compounds responsible for each annotation and the respective SHED Euclidean distance to pergolide.



**Figure 4. Celecoxib's in silico pharmacology profile.** Set of reference compounds responsible for each annotation and the respective SHED Euclidean distance to celecoxib.



The pharmacological network established between the aminergic GPCRs, the opioid and sigma peptidic GPCRs, and the NMDA and 5-HT<sub>3</sub> ligand-gated ion channels is illustrated in **Figure 5**. In order to enhance the target connectivity signal, the network was constructed by linking all targets sharing at least ten drugs. The center and most connected part of the network is composed by all aminergic GPCRs (in dark green). This is in good agreement with a previous analysis suggesting that aminergic GPCRs are among the most promiscuous human proteins [33]. In turn, aminergic GPCRs appear highly connected, on one side, to opioid and sigma receptors (in light green) and, on the other side, to NMDA (in grey) and 5-HT<sub>3</sub> (in white). Although to a much lesser extent, connections between aminergic GPCRs and other GPCRs (light blue), cytochromes (yellow), kinases (red), as well as other enzymes (brown), are also visible. The family of nuclear receptors (dark blue) is the only target class disconnected from the rest, again very much in agreement with previous findings [33].

## Conclusion

The generation of safer and more efficacious drugs for the treatment of diseases is one of the main concerns in current pharmaceutical research. In part, this involves the ability of anticipating the pharmacological profile of drug candidates across multiple targets. In this respect, the capacity of high-throughput screening technologies for testing thousands of small molecules against hundreds of protein targets is still nowadays limited and thus novel *in silico* target profiling methods are emerging as a cost-effective alternative to reduce both the chemical and biological space to explore experimentally. Among those, this work has explored the performance of a ligand-based approach to predict the pharmacological profile of drugs. The results are highly encouraging, as demonstrated for the concrete cases of pergolide and celecoxib. In addition, cross-pharmacologies among proteins belonging to different target classes were detected, some of which could be confirmed in the scientific literature. Although the recall and selectivity rates evidence that there is still room for improvement, a detailed analysis of all the target annotations assigned might bring in the near future interesting new uses for old drugs. Further research in this direction is underway in our laboratory.

## References

1. Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, Austin CP, Auld DS: **High-throughput screening assays for the identification of chemical probes**. 2007, **3**:466-479.
2. Monge AI, Arrault A, Marot C, Morin-Allory L: **Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers**. *Molecular Diversity* 2006, **10**:389-403.
3. Roth BL, Sheffler DJ, Kroeze WK: **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia**. *Nat Rev Drug Discov* 2004, **3**:353-359.
4. Klebe G: **Virtual ligand screening: strategies, perspectives and limitations**. *Drug Discov Today* 2006, **11**:580-594.
5. Savchuk NP, Balakin KV, Tkachenko SE: **Exploring the chemogenomic knowledge space with annotated chemical libraries**. *Curr Opin Chem Biol* 2004, **8**:412-417.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, H W, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucl Acids Res* 2000, **28**:235-242.
7. Nidhi, Glick M, Davies JW, Jenkins JL: **Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases**. *J Chem Inf Model* 2006, **46**:1124-1133.
8. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity**. In *Chemoinformatics in Drug Discovery*. Edited by Wiley-VCH; 2004:223-239.
9. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to in silico pharmacology: nuclear receptor profiling**. *J Chem Inf Model* 2006, **46**:2725-2736.
10. Cases M, Garcia-Serna R, Hettne K, Weeber M, Lei JV, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family**. *Curr Top Med Chem* 2005, **5**:763-772.
11. Rockey WM, Elcock AH: **Rapid computational identification of the targets of protein kinase inhibitors**. *J Med Chem* 2005, **48**:4138-4152.
12. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D: **sc-PDB: an annotated database of druggable binding sites from the protein data bank**. *J Chem Inf Model* 2006, **46**:717-727.
13. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, et al.: **TarFisDock: a web server for identifying drug targets with docking approach**. *Nucl Acids Res* 2006, **34**:W219-224.
14. O'Connor KA, Roth BL: **Finding new tricks for old drugs: an efficient route for public-sector drug discovery**. *Nat Rev Drug Discov* 2005, **4**:1005-1014.
15. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry**. *Nat Biotech* 2007, **25**:197-206.

16. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL: **Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure.** *ChemMedChem* 2007, **2**:861-873.
17. Liu T, Lin Y, Wen X, Jorrisen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.** *Nucleic Acids Research* 2006, **00**:D1-D4.
18. Mestres J: **Mapping the chemogenomic space.** In *Chemogenomics: Knowledge-based Approaches to Drug Discovery*. Edited by Jacoby E: Imperial College Press; 2006:39-57.
19. Biology NCotIUoBaM: *Enzyme nomenclature*. San Diego: Academic Press; 1992.
20. Committee NRN: **A unified nomenclature system for the nuclear receptor superfamily.** *Cell* 1999, **97**:161-163.
21. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR: **On the hierarchical classification of G protein-coupled receptors.** *Bioinformatics* 2007, **23**:3113-3118.
22. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.
23. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.** *J Chem Inf Model* 2004, **44**:1177-1185.
24. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK: **Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.** *J Chem Inf Comp Sci* 2004, **44**:1912-1928.
25. Roth BL: **Drugs and valvular heart disease.** *N Engl J Med* 2007, **356**:6-9.
26. Kvernmo T, Hartter S, Burger E: **A review of the receptor-binding and pharmacokinetic properties of dopamine agonists.** *Clinical Therapeutics* 2006, **28**:1065-1078.
27. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G: **Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition.** *J. Med. Chem.* 2004, **47**:550-557.
28. Werling LL, Lauterbach EC, Calef U: **Dextromethorphan as a potential neuroprotective agent with unique mechanisms of action.** *Neurologist* 2007, **13**:272-293.
29. Minami K, Uezono Y, Ueta Y: **Pharmacological aspects of the effects of tramadol on G-protein coupled receptors.** *Journal of Pharmacological Sciences* 2007, **103**:253-260.
30. Eisensamer B, Rammes G, Gimpl G, Shapa M, Ferrari U, Hapfelmeier G, Bondy B, Parsons C, Gilling K, Zieglgansberger W, et al.: **Antidepressants are functional antagonists at the serotonin type 3 (5-HT<sub>3</sub>) receptor.** *Mol Psychiatry* 2003, **8**:994-1007.



31. Rammes G, Eisensamer B, Ferrari U, Shapa M, Gimpl G, Gilling K, Parsons C, Riering K, Hapfelmeier G, Bondy B, et al.: **Antipsychotic drugs antagonize human serotonin type 3 receptor currents in a noncompetitive manner.** *Mol Psychiatry* 2004, **9**:846-858.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003, **13**:2498-2504.
33. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space.** *Nat Biotech* 2006, **24**:805-815.



# The effect of data completeness on drug-target interaction networks

Jordi Mestres<sup>1</sup>, Elisabet Gregori-Puigjané<sup>1</sup>, Sergi Valverde<sup>2</sup> & Ricard V. Solé<sup>2</sup>

<sup>1</sup>*Chemogenomics Laboratory and* <sup>2</sup>*Complex Systems Laboratory, Research Unit on Biomedical Informatics, Municipal Institute of Medical Research (IMIM) and University Pompeu Fabra, Barcelona, Catalonia, Spain.*  
*email: jmestres@imim.es.*

## To the editor:

The use of network-based approaches to visualise and analyse different types of biologically-relevant interaction data has become increasingly popular in recent years. Topological studies of these interaction networks offer a means to assess the interconnectivity structure established among and between diseases, genes, proteins, and molecules from which influential conclusions and global trends in biology and drug discovery can be derived [1-3]. However, in spite of its indubitable value, currently available interaction data is far from being complete and the portion accessible is often non-homogeneous and biased toward certain areas of interest [4]. This situation results in sample networks that may not be representative of the whole network and thus caution on the conclusions drawn should be highlighted.

Here, we are particularly interested in assessing current levels of completeness in available drug-target interaction data, the potential implications for the topology of the networks derived from them, and the impact that changes in network topology may have on the view of the current status of drug discovery. In this respect, it is widely recognised that, due to limited time and resources, small molecules are usually not screened systematically through a large panel of protein targets for the sake of acquiring knowledge about their complete pharmacological profile but solely to the few targets of interest for the particular project at work. The consequences are that the drug-target interaction data currently available from public bibliographic sources and stored in annotated chemical libraries are largely incomplete and biased toward targets of common therapeutic interest. For example, while in DrugBank [5] the typical antipsychotic drug chlorpromazine is connected to two primary aminergic G protein-coupled receptor (aGPCR) targets (namely, dopamine D<sub>2</sub> and serotonin 5-HT<sub>2A</sub>), and in Wombat [6] it is annotated to another three aGPCRs (namely, dopamine D<sub>1</sub> and D<sub>3</sub>, and serotonin 5-HT<sub>1A</sub>), a more complete receptorome profiling of this drug [7] shows that it actually has sub-micromolar affinity for at least 19 additional aGPCRs.

To gain a deeper insight on the effect that data completeness may have on the topology of drug-target interaction networks, we took a set of 829 small molecule approved drugs from DrugBank and complemented systematically the original interaction data, first, with additional literature-based experimental data available in Wombat and, second, with estimated data obtained

from an *in silico* target profiling method. On the basis of the three sets of interaction data compiled with increased levels of completeness, drug-target (DT) networks were constructed, in which a drug and a protein are connected to each other if the protein is a known target of the drug (**Fig. 1**). The topology of the three DT networks was compared in terms of the fraction of nodes belonging to the largest connected component (nLCC) in the corresponding drug (D) network, in which nodes are drugs connected if they share at least one target, and target (T) network, in which nodes are targets connected if they share at least one drug.

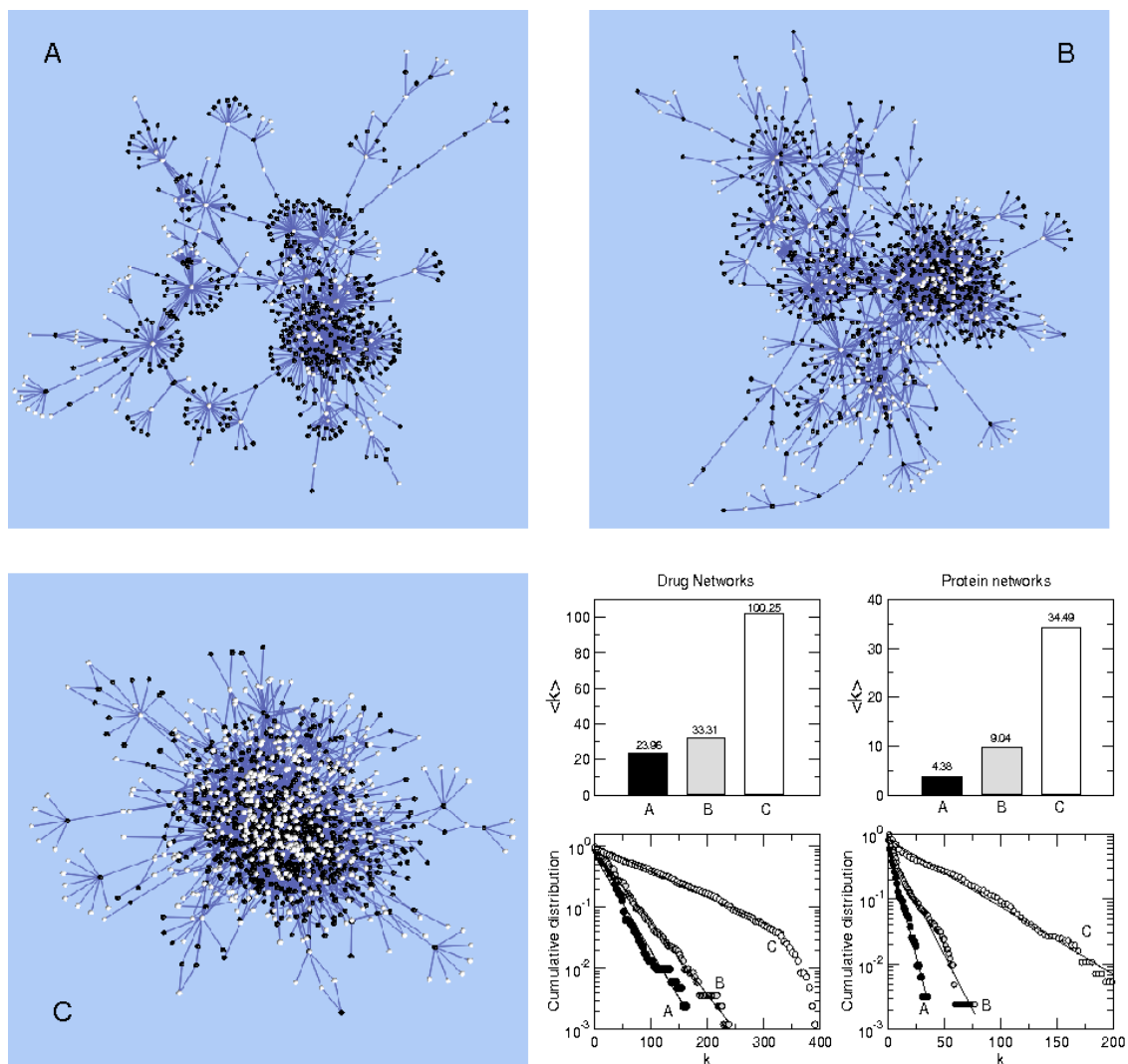
The first DT network was thus based on the 1,445 drug-target interactions available in DrugBank connecting the 829 drugs to 314 protein targets (**Fig. 1A**), resulting in an average number of target proteins per drug of 1.7. The topology of this DT network reveals a well-organised modular structure, with many proteins naturally clustering around phylogenetic families. The nLCC values for the corresponding D and T networks are 0.60 and 0.42, respectively, both numbers being significantly smaller than the values of 0.90 and 0.78 obtained from randomising the networks while keeping the number of nodes and links unchanged. The two nLCC values are also found very much in agreement with those reported recently from networks derived using the same data [8]. However, since DrugBank contains mainly information on the primary targets of drugs (that is, those proteins formally accepted by the originators as being targeted by drugs), its interaction data suffers from the incompleteness and bias issues emphasised above [9]. In fact, a DT network derived solely on the basis of DrugBank interaction data may actually be more representative of the target space explored historically by pharmaceutical industry rather than being a true reflection of drug polypharmacology.

Accordingly, a second DT network was constructed by supplementing the original interaction data in DrugBank with additional literature-based affinity data available in Wombat, resulting in 2,117 drug-target interactions connecting the 829 drugs to 409 protein targets (**Fig. 1B**). This means that the networks derived from these data will account for 672 extra drug-target interactions and 95 proteins relative to the original data present in DrugBank, increasing the average number of target proteins per drug to 2.7. The topology of the DT network is visibly affected, becoming more complex and interconnected. Quantitatively, this is reflected by nLCC values of 0.79 and 0.64 for the corresponding D and T networks, respectively, a situation that becomes much closer to the topology of a randomised network than envisaged from DrugBank data only.

A final third DT network was derived by complementing the literature-based experimental data accumulated from DrugBank and Wombat with annotations assigned using a recently reported ligand-based approach to *in silico* target profiling [10], leading to a total of 5,215 drug-target interactions connecting the 829 drugs to 557 targets (**Fig. 1C**). These results project the average number of target proteins per drug up to 6.3. Even though the number of drug-target interactions obtained at this stage might seem quite large compared to the previous two cases considering literature-based experimental data only, the projection obtained agrees well with the value of 6.8 reported recently for the average number of targets being hit under a 10  $\mu$ M cut-off for a set of 89

drugs tested against more than 60 targets [11]. Back to the example of chlorpromazine, 10 additional annotations to aGPCRs ( $M_1$ ,  $\alpha_{1A}$ ,  $\alpha_{2A}$ ,  $\alpha_{2C}$ ,  $D_4$ ,  $D_5$ ,  $H_1$ ,  $H_2$ ,  $5\text{-HT}_{2C}$ , and  $5\text{-HT}_7$ ) were correctly assigned by this computational approach on top of the 5 connections reported in DrugBank and Wombat, only 9 short from all interactions to aGPCRs confirmed in the literature [7]. Taking this step as an *in silico* estimate to full completion of interaction data involving those 829 drugs, the resulting DT network loses its original modular structure to collapse in an unexpectedly dense topology, with nLCC values for the corresponding D and T networks of 0.96 and 0.92, respectively, very close to the those obtained for the corresponding randomised networks.

In conclusion, it has been shown that systematic completion of drug-target interaction data leads to more complex and disordered network topologies with significantly increased graph density, suggesting that the well-defined separation between hub-related modules obtained when using highly incomplete data is likely to break down with increased completeness. Therefore, any conclusions derived from the analysis of network topologies obtained from incomplete data should be taken with caution. It is thus becoming urgent to support some global coordinated initiatives aiming at generating complete, homogeneous, unbiased drug-target interaction matrices [12] as a means to ensure solid progress in all integrative scientific areas relying on interaction data. In the meantime, the unexpectedly complex picture of the DT network generated from a projection of complete interaction data obtained *in silico* provides an entirely different perspective of the druggable target space. In the last years, the pharmaceutical and biotechnology industries have been alarmed by the fact that as low as 3000 druggable targets may be present in the human genome [13]. Given the high level of interconnectivity observed in DT networks, the question may no longer be how many druggable targets are present in the human genome but rather how many combinations of those druggable targets result in therapeutically-acceptable pharmacological profiles, opening a wealth of possibilities for the future of multitarget drug discovery.



**Figure 1** Changes in network topology under different levels of completeness for drug-target interaction data. **(A)** DrugBank, **(B)** DrugBank+Wombat and **(C)** DrugBank+Wombat+InSilico estimate. Here drugs and proteins are indicated as black and white balls, respectively. Also included (bottom right) are the statistical patterns displayed by the drug (DN) and target (TN) networks (see text). The parameter  $\langle k \rangle$ , measuring the average number of links in the network, grows with completeness. Of mention is the fact that  $\langle k \rangle$  doubles its value for protein targets upon addition of the drug-target interaction data from Wombat. We also estimated the degree distributions  $P(k)$  measuring the probability of finding a node having  $k$  links. In order to have a clean plot, we used the cumulative degree distribution  $P_{>}(k) = P(i)$ . The plots fall off as exponential laws of the form  $P_{>}(k) = N \exp(-k/k_c)$  thus indicating the presence of a characteristic degree  $k_c$  ( $N$  is a normalization constant). Here we estimated the characteristic values  $k_c$  using least squares. For the DN we obtained **(A)**  $k_c = 28.57$ , **(B)**  $k_c = 37.03$  and **(C)**  $k_c = 100.0$  whereas for the TN the cut-off values are **(A)**  $k_c = 6.66$ , **(B)**  $k_c = 12.98$  and **(C)**  $k_c = 43.47$ . These values are consistent with the distribution of links being best described by an exponential shape rather than a scale-free architecture.

## References:

1. Butcher EC, Berg EL, Kunkel EJ: **Systems biology in drug discovery.** *Nat Biotech* 2004, **22**:1253-1259.
2. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space.** *Nat Biotech* 2006, **24**:805-815.
3. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104**:8685-8690.
4. Hakes L, Pinney JW, Robertson DL, Lovell SC: **Protein-protein interaction networks and biology - what's the connection?** *Nat Biotech* 2008, **26**:69-72.
5. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucl. Acids Res.* 2006, **34**:D668-672.
6. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity.** In *Cheminformatics in Drug Discovery.* Edited by Wiley-VCH; 2004:223-239.
7. Roth BL, Sheffler DJ, Kroeze WK: **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.** *Nat Rev Drug Discov* 2004, **3**:353-359.
8. Yildirim MA, Goh K-I, Cusick ME, Barabasi A-L, Vidal M: **Drug-target network.** 2007, **25**:1119-1126.
9. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R: **Network analysis of FDA approved drugs and their targets.** *Mt Sinai J Med* 2007, **74**:27-32.
10. Gregori-Puigjané E, Mestres J: **Mining the chemogenomic space of drugs.** *Comb Chem High Throughput Screen* 2008.
11. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L: **Modeling Promiscuity Based on in vitro Safety Pharmacology Profiling Data.** *ChemMedChem* 2007, **2**:874-880.
12. Taussig MJ, Stoevesandt O, Borrebaeck CA, Bradbury AR, Cahill D, Cambillau C, de Daruvar A, Dübel S, Eichler J, al. e: **ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome.** *Nature Methods* 2007, **4**:13-17.
13. Hopkins AL, Groom CR: **The druggable genome.** *Nat Rev Drug Discov* 2002, **1**:727-730.





## Chapter III.5 – ViSCA

In order to apply all the methodologies developed and presented all along this thesis, a stand-alone application called ViSCA has been developed. It is organized in several modules, each of which enables to performance a set of related functions. This piece of software permits performing from very simple file management operations to the already presented virtual ligand screening and virtual target profiling methodologies.



# ViSCA: Virtual Screening and Chemical Annotation

Elisabet Gregori-Puigjané, Rut Garriga and Jordi Mestres

Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain

## Introduction

We present ViSCA, a software framework presented as a stand-alone application that facilitates managing chemical databases in different formats. Additionally, ViSCA integrates the diverse methodologies developed in our laboratory, enabling to apply them to any input database. Calculating 1D descriptors and applying filters, calculating SHED descriptors for the input database, using them for virtual ligand screening or virtual target profiling and calculating the unique identifiers or hierarchical classification schemes for chemical annotation are other possible uses of ViSCA.

## ViSCA architecture

ViSCA has been implemented as a collection of python modules integrated in a stand-alone application through a user interface. The following sections will present each of the modules and their capabilities.

### *The file management module*

In the first place, ViSCA can be used as a file management tool that is able to parse chemical libraries in SD [1], mol [1], mol2 [2] and PDB [3] formats. It can be used for simple tasks such as counting the number of compounds in the library, extracting a subset of compounds given a list of identifiers and splitting the library in either a fixed number of files or in several files containing a fixed number of compounds.

Other features are specifically related to SD file management, as they deal with the information contained in the information fields that are specific to this format. It enables to remove all or a user-provided set of fields, add information contained in a plain text table to the SD file by adding it into each molecule's fields or put the information contained in these fields to a plain text file so it can be better analysed. It also enables to put the information contained in a field as the molecule identifier or sort the molecules according to the content of a user-defined field.

### ***The properties module***

This module included in the ViSCA suite is the calculation of the one-dimensional properties of each of the molecules in the input library. These include simple and fast to calculate descriptors that can be extracted directly from the atoms count of each molecule like molecular weight, hydrogen bond donors and acceptors and number of halogen atoms, but also more sophisticated descriptors based on the topology of the molecule, like the number of rotatable bonds, rings and ring systems or even the presence and number of reactive groups. All these descriptors can be used to apply diverse filters to the database, like an organic filter that discards molecules containing atoms others than those considered organic (C, N, O, H, S, P, Cl, Br, I, F). Other filters are those for reactive molecules, or descriptors-based drug-like filters [4].

### ***The SHED module***

The calculation of the SHED [5], a set of topological descriptors developed in our laboratory, can be easily done using the ViSCA command line.

### ***The virtual screening module***

For virtual ligand screening, ViSCA enables to prioritise an input database in terms of SHED Euclidean distance in either of the accepted formats with respect to a reference database, which can consist on a single compound or a set of compounds. If the SHED descriptors for both libraries are not provided, they will be calculated on the fly.

### ***The virtual profiling module***

This module enables to build a SHED-based model starting from an annotated chemical library in SD format, where the information on the targets associated to each molecule has to be in a field. Once the model is built, it also enables to profile any input database, in any of the input formats accepted, for all the generated models or for a list of models that can be provided by the user. The output of this profiling are several files, one containing the list of the identifiers of the molecules with some annotation and the list of targets to which they are annotated. Other output files are an SD file containing each annotated molecule followed by the closest reference, a matrix containing the closest distance to each model for each molecule and finally a heatmap to visually inspect the results.

### ***The chemical annotation module***

The chemical annotation module enables to calculate either a unique molecule graph identifier or a hierarchical classification scheme [6], both developed in our laboratory, for the whole input database. These unique codes and structure-based classification schemes can help organizing and analysing the actual contents and diversity of even big databases. Based on these unique codes, the user can extract unique compounds, scaffolds or frameworks from the input database.

## References

1. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *J Chem Inf Comp Sci* 1992, **32**:244-255.
2. Tripos Inc. <http://www.tripos.com/>.
3. <http://www.wwpdb.org/docs.html>.
4. Muegge I, Heald SL, Brittelli D: **Simple selection criteria for drug-like chemical matter.** *J Med Chem* 2001, **44**:1841-1846.
5. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.
6. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to in silico pharmacology: nuclear receptor profiling.** *J Chem Inf Model* 2006, **46**:2725-2736.



This section summarizes the main achievements of the work presented in this thesis:

1. A new set of topological atom-centred feature-based descriptors called Shannon entropy descriptors (SHED) has been developed.
2. On the basis of pharmacological data extracted directly from bibliographical sources, a ligand-based descriptor model has been generated based on SHED descriptors derived from bioactive ligands. This opens the possibility to extend *in silico* the profiling of large chemical libraries against those protein targets for which structural information is not available yet in an efficient manner.
3. This *in silico* profiling towards a single target of interest (virtual ligand screening) has been tested and proven to perform well in terms of absolute enrichment in the top ranked compounds and in terms of novel structures enrichment (scaffold hopping). This has been validated not only with academic exercises but also with “real” drug discovery projects as part of the collaborations that Chemotargets has with several pharmaceutical companies.
4. Ligand-based virtual target profiling examples are provided and their performance assessed and compared to well-established methodologies. It has shown that, provided the annotated chemical space for the protein family of interest is sufficiently well saturated, the model attains a decent degree of both internal consistency and external predictability. Again, this has been validated not only with academic exercises but also with “real” library design projects as part of the collaborations that Chemotargets has with several chemical companies.
5. The virtual target profiling approach developed during this thesis has been proven to be sensible enough to achieve significant discriminative power when applied to external chemical libraries designed for *a priori* unrelated protein families, opening an avenue for its use in the selection and design of targeted libraries.
6. Examples of the applicability of the methodology to targeted library design are provided. Furthermore, the need for a deeper study on the objective quality assessment of the targeted libraries is highlighted. In this respect, *in silico* target profiling methods provide a means to analyse both chemical and target diversity in terms of the projected pharmacological promiscuity of compound libraries.

7. The models developed so far served also to construct an interaction network from which potential cross-pharmacologies between proteins emerge.
8. It has been shown that any conclusions derived from the analysis of network topologies obtained from incomplete data should be taken with caution. Consequently, the systematic completion of ligand-target interaction data has shown to lead to more complex and disordered network topologies with significantly increased graph density.
9. Based on ligand-based protein network analysis, a shift of perspective in drug discovery is proposed. Beyond targeting individual proteins, a wealth of possibilities emerge by combining those druggable proteins that result in a therapeutically-acceptable pharmacological profiles. This opens the way to future systems chemical biology approaches.
10. All the algorithms developed within this thesis have been implemented in an integrative package called ViSCA, which is not only the framework onto which all developments within the Chemogenomics Laboratory are being implemented but it is already being used within pharmaceutical industries.



## References

1. Shannon CE, Weaver W: *The mathematical theory of communication*. Urbana, IL: University of Illinois Press; 1949.
2. Venkatesh S, Lipper RA: **Role of the development scientist in compound lead selection and optimization**. *Journal of Pharmaceutical Sciences* 2000, **89**:145-154.
3. Bredel M, Jacoby E: **Chemogenomics: an emerging strategy for rapid target and drug discovery**. *Nat Rev Genet* 2004, **5**:262-275.
4. Butcher EC: **Can cell systems biology rescue drug discovery?** *Nature Rev Drug Discov* 2005, **4**:461-467.
5. Ghose AK, Herbertz T, Salvino JM, Mallamo JP: **Knowledge-based chemoinformatic approaches to drug discovery**. *Drug Discov Today* 2006, **11**:1107-1114.
6. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling**. *Br J Pharmacol* 2007, **152**:9-20.
7. Ramström O, Lehn J-M: **Drug discovery by dynamic combinatorial libraries**. *Nat Rev Drug Discov* 2002, **1**:26-36.
8. Liu R, Hsieh C-Y, Lam KS: **New approaches in identifying drugs to inactivate oncogene products**. *Seminars in Cancer Biol* 2004, **14**:13-21.
9. Geysen HM, Schoenen F, Wagner D, Wagner R: **Combinatorial compound libraries for drug discovery: an ongoing challenge**. 2003, **2**:222-230.
10. Bleicher KH, Böhm H-J, Müller K, Alanine AI: **Hit and lead generation: beyond high-throughput screening**. *Nat Rev Drug Discov* 2003, **2**:369-378.
11. Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, Austin CP, Auld DS: **High-throughput screening assays for the identification of chemical probes**. *Nat Chem Biol* 2007, **3**:466-479.
12. Schnecke V, Boström J: **Computational chemistry-driven decision making in lead generation**. *Drug Discov Today* 2006, **11**:43-50.
13. Wunberg T, Hendrix M, Hillisch A, Lobell M, Meier H, Schmeck C, Wild H, Hinzen B: **Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits**. *Drug Discov Today* 2006, **11**:175-180.
14. Beno BR, Mason JS: **The design of combinatorial libraries using properties and 3D pharmacophore fingerprints**. *Drug Discov Today* 2001, **6**:251-258.
15. Apic G, Ignjatovic T, Boyer S, Russell RB: **Illuminating drug discovery with biological pathways**. *FEBS Lett Sys Biol* 2005, **579**:1872-1877.
16. Kassel DB: **Applications of high-throughput ADME in drug discovery**. *Curr Opin Chem Biol* 2004, **8**:339-345.
17. Keserü GM, Makara GM: **Hit discovery and hit-to-lead approaches**. *Drug Discov Today* 2006, **11**:741-748.
18. Zhao H: **Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective**. *Drug Discov Today* 2007, **12**:149-155.
19. Whitebread S, Hamon J, Bojanic D, Urban L: **Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development**. *Drug Discov Today* 2005, **10**:1421-1433.
20. Hammer SM, Saag MS, Schechter M, Montaner JSG, Schooley RT, Jacobsen DM, Thompson MA, Carpenter CCJ, Fischl MA, Gazzard BG, et al.: **Treatment for adult HIV infection:**

- 2006 recommendations of the international AIDS society-USA panel.** *JAMA* 2006, **296**:827-843.
21. Mencher SK, Wang LG: **Promiscuous drugs compared to selective drugs (promiscuity can be a virtue).** *BMC Clin Pharmacol* 2005, **5**:3.
  22. Millan MJ: **The role of monoamines in the actions of established and "novel" antidepressant agents: a critical review.** *Eur J Pharmacol* 2004, **500**:371-384.
  23. Charlier C, Michaux C: **Dual inhibition of cyclooxygenase-2 (COX-2) and 5-lipoxygenase (5-LOX) as a new strategy to provide safer non-steroidal anti-inflammatory drugs.** *Eur J Med Chem* 2003, **38**:645-659.
  24. Jimeno A, Hidalgo M: **Multitargeted therapy: Can promiscuity be praised in an era of political correctness?** *Critical Reviews in Oncology/Hematology* 2006, **59**:150-158.
  25. Klabunde T: **Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.** *Br J Pharmacol* 2007, **152**:5-7.
  26. Shorter E: **Looking backwards: a possible new path for drug discovery in psychopharmacology.** *Nat Rev Drug Discov* 2002, **1**:1003-1006.
  27. Oprea TI, Tropsha A, Faulon J-L, Rintoul MD: **Systems chemical biology.** *Nat Chem Biol* 2007, **3**:447-450.
  28. Hopkins AL, Mason JS, Overington JP: **Can we rationally design promiscuous drugs?** *Curr Opin Struct Biol* 2006, **16**:127-136.
  29. Kitano H: **A robustness-based approach to systems-oriented drug design.** *Nat Rev Drug Discov* 2007, **6**:202-210.
  30. Roth BL, Sheffler DJ, Kroeze WK: **Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.** *Nat Rev Drug Discov* 2004, **3**:353-359.
  31. Goldstein DM, Gray NS, Zarrinkar PP: **High-throughput kinase profiling as a platform for drug discovery.** *Nat Rev Drug Discov* 2008, advanced online publication.
  32. Kung C, Kenski DM, Dickerson SH, Howson RW, Kuyper LF, Madhani HD, Shokat KM: **Chemical genomic profiling to identify intracellular targets of a multiplex kinase inhibitor.** *Proc Nat Acad Sci USA* 2005, **102**:3587-3592.
  33. Rausch O: **High content cellular screening.** *Curr Opin Chem Biol* 2006, **10**:316-320.
  34. Korn K, Krausz E: **Cell-based high-content screening of small-molecule libraries.** *Curr Opin Chem Biol* 2007, **11**:503-510.
  35. Haney SA, LaPan P, Pan J, Zhang J: **High-content screening moves to the front of the line.** *Drug Discov Today* 2006, **11**:889-894.
  36. Liptrot C: **High content screening - from cells to data to knowledge.** *Drug Discov Today* 2001, **6**:832-834.
  37. Mestres J: **Computational chemogenomics approaches to systematic knowledge-based drug discovery.** *Curr Opin Drug Discov Devel* 2004, **7**:304-313.
  38. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors.** *J Med Chem* 1996, **39**:3049-3059.
  39. Bajorath J: **Computational analysis of ligand relationships within target families.** *Curr Opin Chem Biol* 2008, In Press, Corrected Proof.
  40. Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko MA: **Chemogenomic approaches to drug discovery.** *Curr Opin Chem Biol* 2001, **5**:464-470.
  41. Mestres J: **Mapping the chemogenomic space.** In *Chemogenomics: Knowledge-based Approaches to Drug Discovery*. Edited by Jacoby E: Imperial College Press; 2006:39-57.

42. Lan N, Montelione GT, Gerstein M: **Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level.** *Curr Opin Chem Biol* 2003, **7**:44-54.
43. Stevens R, Goble CA, Bechhofer S: **Ontology-based knowledge representation for bioinformatics.** *Brief Bioinform* 2000, **1**:398-414.
44. Xu Y-J, Johnson M: **Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries.** *J Chem Inf Comp Sci* 2002, **42**:912-926.
45. Braun J, Gugisch R, Kerber A, Laue R, Meringer M, Rucker C: **MOLGEN-CID - A canonizer for molecules and graphs accessible through the internet.** *J Chem Inf Comp Sci* 2004, **44**:542-548.
46. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y: **Enhancement of the chemical semantic web through the use of InChI identifiers.** *Org Biomol Chem* 2005, **3**:1832-1834.
47. Gregori-Puigjané E, Mestres J: **A ligand-based approach to mining the chemogenomic space of drugs.** *Comb Chem High Throughput Screen* 2008.
48. Frye SV: **Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era.** *Chem & Biol* 1999, **6**:R3-R7.
49. Biology NCotIUoBaM: *Enzyme nomenclature*. San Diego: Academic Press; 1992.
50. Izrailev S, Farnum MA: **Enzyme classification by ligand binding.** *Proteins: Structure, Function, and Bioinformatics* 2004, **57**:711-724.
51. Nuclear Receptors Nomenclature Committee: **A unified nomenclature system for the nuclear receptor superfamily.** *Cell* 1999, **97**:161-163.
52. Jacoby E: **Chemogenomics: drug discovery's panacea?** *Mol BioSystems* 2006, **2**:218-220.
53. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E: **Similarity metrics for ligands reflecting the similarity of the target proteins.** *J Chem Inf Model* 2003, **43**:391-405.
54. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem* 1996, **39**:2887-2893.
55. Wilkens SJ, Janes J, Su AI: **HierS: hierarchical scaffold clustering using topological chemical graphs.** *J Med Chem* 2005, **48**:3182-3193.
56. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H: **The scaffold tree - Visualization of the scaffold universe by hierarchical scaffold classification.** *J Chem Inf Model* 2006, **47**:47-58.
57. Mestres J, Martin-Couce L, Gregori-Puigjané E, Cases M, Boyer S: **Ligand-based approach to *in silico* pharmacology: nuclear receptor profiling.** *J Chem Inf Model* 2006, **46**:2725-2736.
58. Weigelt J, McBroom-Cerajewski LD, Schapira M, Zhao Y, Arrowsmith CH: **Structural genomics and drug discovery: all in the family.** *Curr Opin Chem Biol* 2008, **12**:32-39.
59. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, H W, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28**:235-242.
60. <http://www.ebi.ac.uk/thornton-srv/databases/enzymes/>.
61. Wang R, Fang X, Lu Y, Wang S: **The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.** *J Med Chem* 2004, **47**:2977-2980.
62. <http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp>: (accessed March 2008).
63. Garcia-Serna R, Opatowski L, Mestres J: **FCP: functional coverage of the proteome by structures.** *Bioinformatics* 2006, **22**:1792-1793.
64. Savchuk NP, Balakin KV, Tkachenko SE: **Exploring the chemogenomic knowledge space with annotated chemical libraries.** *Curr Opin Chem Biol* 2004, **8**:412-417.

65. Chen X, Ji ZL, Chen YZ: **TTD: therapeutic target database**. *Nucl Acids Res* 2002, **30**:412-415.
66. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanli M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucl Acids Res* 2008, **36**:D901-D906.
67. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Garcia Urdiales E, Gewiess A, Jensen LJ, et al.: **SuperTarget and Matador: resources for exploring drug-target relationships**. *Nucl Acids Res* 2008, **36**:D919-922.
68. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, et al.: **ChemBank: a small-molecule screening and cheminformatics resource database**. *Nucl Acids Res* 2007, **36**:D351-359.
69. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al.: **Database resources of the National Center for Biotechnology Information**. *Nucl Acids Res* 2008, **36**:D13-21.
70. Fradera X, Mestres J: **Guided docking approaches to structure-based design and screening**. *Curr Top Med Chem* 2004, **4**:687-700.
71. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity**. In *Chemoinformatics in Drug Discovery*. Edited by Wiley-VCH; 2004:223-239.
72. AurSCOPE databases, Aureus Pharma, France.
73. MedChem and Target Inhibitor databases, GVK Biosciences, India.
74. Cases M, Garcia-Serna R, Hettne K, Weeber M, Lei JV, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family**. *Curr Top Med Chem* 2005, **5**:763-772.
75. Evers A, Hessler G, Matter H, Klabunde T: **Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols**. *J Med Chem* 2005, **48**:5448-5465.
76. Zhang Q, Muegge I: **Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring**. *J Med Chem* 2006, **49**:1536-1548.
77. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, et al.: **A critical assessment of docking programs and scoring functions**. *J Med Chem* 2006, **49**:5912-5931.
78. Cavasotto CN, Orry AJW, Murgolo NJ, Czarniecki MF, Kocsi SA, Hawes BE, O'Neil KA, Hine H, Burton MS, et al.: **Discovery of Novel Chemotypes to a G-Protein-Coupled Receptor through Ligand-Steered Homology Modeling and Structure-Based Virtual Screening**. *J. Med. Chem.* 2008, **51**:581-588.
79. Costanzi S: **On the applicability of GPCR homology models to computer-aided drug discovery: a comparison between in silico and crystal structures of the  $\beta_2$ -Adrenergic Receptor**. *J Med Chem* 2008.
80. Alvesalo JKO, Siiskonen A, Vainio MJ, Tammela PSM, Vuorela PM: **Similarity based virtual screening: a tool for targeted library design**. *J Med Chem* 2006, **49**:2353-2356.
81. Johnson M, Maggiora G: *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 2006.
82. Rognan D: **Chemogenomic approaches to rational drug design**. *Br J Pharmacol* 2007, **152**:38-52.
83. Yan SF, King FJ, He Y, Caldwell JS, Zhou Y: **Learning from the data: mining of large high-throughput screening databases**. *J Chem Inf Model* 2006, **46**:2381-2395.
84. Muller G: **Medicinal chemistry of target family-directed masterkeys**. *Drug Discov Today* 2003, **8**:681-691.

85. Schnur DM, Hermsmeier MA, Tebben AJ: **Are target-family-privileged substructures truly privileged?** *J Med Chem* 2006, **49**:2000-2009.
86. Sheridan RP, Kearsley SK: **Why do we need so many chemical similarity search methods?** *Drug Discov Today* 2002, **7**:903-911.
87. Livingstone DJ: **The characterization of chemical structures using molecular properties. A survey.** *J Chem Inf Comp Sci* 2000, **40**:195-209.
88. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.** *J Chem Inf Model* 2004, **44**:1177-1185.
89. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures.** *Org & Biomolec Chem* 2004, **4**:3256-3266.
90. *MACCS-II*. San Leandro, CA: MDL Ltd.; 1992.
91. James C, D W: *Daylight Theory Manual*: Daylight Chemical Information Systems, Inc.; 1995.
92. *UNITY Reference Manual*. St. Louis, MO: Tripos Inc.; 1995.
93. Xue L, Bajorath J: **Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening.** *Comb Chem High Throughput Screen* 2000, **3**:363-371.
94. Swamidass SJ, Baldi P: **Mathematical correction for fingerprint similarity measures to improve chemical retrieval.** *J Chem Inf Model* 2007, **47**:952-964.
95. Renner S, Schneider G: **Scaffold-hopping potential of ligand-based similarity concepts.** *ChemMedChem* 2006, **1**:181-185.
96. Carhart RE, Smith DH, Venkataraghavan R: **Atom pairs as molecular features in structure-activity studies: definition and applications.** *J Chem Inf Model* 1985, **25**:64-73.
97. Schneider G, Neidhart W, Giller T, Schmid G: **"Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening.** *Angewandte Chemie International Edition* 1999, **38**:2894-2896.
98. Gregori-Puigjané E, Mestres J: **SHED: Shannon entropy descriptors from topological feature distributions.** *J Chem Inf Model* 2006, **46**:1615-1622.
99. Pearlman RS, Smith KM: **Novel software tools for chemical diversity.** *Persp Drug Discov Design* 1998, **9/11**:339-353.
100. Cramer RD, Patterson DE, Bunce JD: **Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.** *J Am Chem Soc* 1988, **110**:5959-5967.
101. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S: **Grid-independent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors.** *J Med Chem* 2000, **43**:3233-3243.
102. Hansch C: **Quantitative relationships between lipophilic character and drug metabolism.** *Drug Metab Rev* 1972, **1**:1-13.
103. Roncaglioni A, Benfenati E: **In silico-aided prediction of biological properties of chemicals: oestrogen receptor-mediated effects.** *Chem Soc Rev* 2008, **37**:441-450.
104. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK: **Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.** *J Chem Inf Comp Sci* 2004, **44**:1912-1928.
105. Tetko IV, Luik AI, Poda GI: **Applications of neural networks in structure-activity relationships of a small number of molecules.** *J Med Chem* 1993, **36**:811-814.
106. Schneider G, Wrede P: **Artificial neural networks for computer-based molecular design.** *Progress in Biophysics and Molecular Biology* 1998, **70**:175-222.

107. Burden FR, Winkler DA: **Robust QSAR models using Bayesian regularized neural networks.** *J Med Chem* 1999, **42**:3183-3187.
108. Verma RP, Hansch C: **An approach toward the problem of outliers in QSAR.** *Bioorg Med Chem* 2005, **13**:4597-4621.
109. Polanski J, Bak A, Gieleciak R, Magdziarz T: **Modeling robust QSAR.** *J Chem Inf Model* 2006, **46**:2310-2318.
110. Maggiora GM: **On outliers and activity cliffs - why QSAR often disappoints.** *J Chem Inf Model* 2006, **46**:1535.
111. Johnson SR: **The trouble with QSAR (or how I learned to stop worrying and embrace fallacy).** *J Chem Inf Model* 2008, **48**:25-26.
112. Chen H, Zhou J, Xie G: **PARM: a genetic evolved algorithm to predict bioactivity.** *J Chem Inf Comp Sci* 1998, **38**:243-250.
113. Lin W-Q, Jiang J-H, Zhou Y-P, Wu H-L, Shen G-L, Yu R-Q: **Support vector machine based training of multilayer feedforward neural networks as optimized by particle swarm algorithm: application in QSAR studies of bioactivity of organic compounds.** *J Comput Chem* 2007, **28**:519-527.
114. Klebe G: **Virtual ligand screening: strategies, perspectives and limitations.** *Drug Discov Today* 2006, **11**:580-594.
115. Muegge I, Oloff S: **Advances in virtual screening.** *Drug Discov Today: Technol* 2006, **3**:405-411.
116. Oprea TI, Matter H: **Integrating virtual screening in lead discovery.** *Current Opinion in Chemical Biology* 2004, **8**:349-358.
117. Good AC, Krystek SR, Mason JS: **High-throughput and virtual screening: core lead discovery technologies move towards integration.** *Drug Discovery Today* 2000, **5**:61-69.
118. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kretsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD: **Comparison of topological, shape, and docking methods in virtual screening.** *J Chem Inf Model* 2007, **47**:1504-1519.
119. Hawkins PCD, Skillman AG, Nicholls A: **Comparison of shape-matching and docking as virtual screening tools.** *J Med Chem* 2007, **50**:74-82.
120. Bajorath J: **Integration of virtual and high-throughput screening.** *Nat Rev Drug Discov* 2002, **1**:882-894.
121. Truchon J-F, Bayly CI: **Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem.** *J. Chem. Inf. Model.* 2007, **47**:488-508.
122. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL: **Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure.** *ChemMedChem* 2007, **2**:861-873.
123. O'Connor KA, Roth BL: **Finding new tricks for old drugs: an efficient route for public-sector drug discovery.** *Nat Rev Drug Discov* 2005, **4**:1005-1014.
124. Chong CR, Sullivan DJ: **New uses for old drugs.** *Nature* 2007, **448**:645-646.
125. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry.** *Nat Biotech* 2007, **25**:197-206.
126. Jenwitheesuk E, Horst JA, Rivas KL, Van Voorhis WC, Samudrala R: **Novel paradigms for drug discovery: computational multitarget screening.** *Trends Pharmacol Sci* 2008, **29**:62-71.
127. Rockey WM, Elcock AH: **Rapid computational identification of the targets of protein kinase inhibitors.** *J Med Chem* 2005, **48**:4138-4152.

128. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, et al.: **TarFisDock: a web server for identifying drug targets with docking approach.** *Nucl Acids Res* 2006, **34**:W219-224.
129. Nidhi, Glick M, Davies JW, Jenkins JL: **Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases.** *J Chem Inf Model* 2006, **46**:1124-1133.
130. Martin YC, Kofron JL, Traphagen LM: **Do structurally similar molecules have similar biological activity?** *J Med Chem* 2002, **45**:4350-4358.
131. Lowrie JF, Delisle RK, Hobbs DW, Diller DJ: **The different strategies for designing GPCR and kinase targeted libraries.** *Comb Chem High Throughput Screen* 2004, **7**:495-510.
132. Sun D, Chuaqui C, Deng Z, Bowes S, Chin D, Singh J, Cullen P, Hankins G, Lee W-C, Donnelly J, et al.: **A kinase-focused compound collection: compilation and screening strategy.** *Chem Biol & Drug Design* 2006, **67**:385-394.
133. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Reviews* 1997, **23**:3-25.
134. Schuffenhauer A, Brown N, Selzer P, Ertl P, Jacoby E: **Relationships between Molecular Complexity, Biological Activity, and Structural Diversity.** *J. Chem. Inf. Model.* 2006, **46**:525-535.
135. Rishton GM: **Reactive compounds and in vitro false positives in HTS.** *Drug Discov Today* 1997, **2**:382-384.
136. Oprea TI: **Property distribution of drug-related chemical databases.** *J Comp-Aided Mol Design* 2000, **14**:251-264.
137. Harris CJ, Stevens AP: **Chemogenomics: structuring the drug discovery process to gene families.** *Drug Discov Today* 2006, **11**:880-888.
138. Crossley R: **The design of screening libraries targeted at G-protein coupled receptors.** *Curr Top Med Chem* 2004, **4**:581-589.
139. Rajasethupathy P, Vayttaden SJ, Bhalla US: **Systems modeling: a pathway to drug discovery.** *Curr Opin Chem Biol* 2005, **9**:400-406.
140. Yildirim MA, Goh K-I, Cusick ME, Barabasi A-L, Vidal M: **Drug-target network.** 2007, **25**:1119-1126.
141. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, et al.: **The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science* 2006, **313**:1929-1935.
142. Kuhn M, Campillos M, Gonzalez P, Jensen LJ, Bork P: **Large-scale prediction of drug-target relationships.** *FEBS Letters* 2008, **582**:1283-1290.
143. Paolini GV, Shapland RHB, van Hoornt WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space.** *Nat Biotech* 2006, **24**:805-815.
144. Jenkins JL, Glick M, Davies JW: **A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes.** *J. Med. Chem.* 2004, **47**:6144-6159.
145. Liu T, Lin Y, Wen X, Jorrisen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.** *Nucleic Acids Research* 2006, **00**:D1-D4.
146. Roth BL, Kroeze WK, Patel S, Lopez E: **The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches?** *Neurosci* 2000, **6**:252-262.
147. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK: **Quantifying the Relationships among Drug Classes.** *J. Chem. Inf. Model.* 2008.
148. Wagener M, Lommerse JPM: **The quest for bioisosteric replacements.** *J Chem Inf Model* 2006, **46**:677-685.

149. Gregori-Puigjané E, Mestres J: **A ligand-based approach to mining the chemogenomic space of drugs.** *Comb Chem High Throughput Screen* 2008, In press.
150. Lewell XQ, Judd DB, Watson SP, Hann MM: **RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry.** *J. Chem. Inf. Model.* 1998, **38**:511-522.
151. Lauri G, Bartlett PA: **CAVEAT: a program to facilitate the design of organic molecules.** *J Comp-Aided Mol Design* 1994, **8**:51-66.
152. Maass P, Schulz-Gasch T, Stahl M, Rarey M: **Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations.** *J. Chem. Inf. Model.* 2007.
153. van Gestel S, Schuermans V: **Thirty-three years of drug discovery and reaserach with Dr. Paul Janssen.** *Drug Dev Res* 1986, **8**:1-13.
154. Hood L, Perlmutter RM: **The impact of systems approaches on biological problems in drug discovery.** *Nat Biotech* 2004, **22**:1215-1217.
155. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, et al.: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8**:R39.
156. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucl Acids Res* 2006, **34**:D354-357.
157. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al.: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucl Acids Res* 2008, **36**:D623-631.